Mathematical Biosciences and Engineering

MBE, 19(12): 13949–13966.
DOI: 10.3934/mbe.2022650
Received: 11 August 2022
Revised: 07 September 2022
Accepted: 12 September 2022
Published: 22 September 2022

http://www.aimspress.com/journal/MBE

Research article

An explorative study for leveraging transcriptomic data of embryonic stem cells in mining cancer stemness genes, regulators, and networks

Jihong Yang[1,2,*], Hao Xu[3], Congshu Li[2], Zhenhao Li[1,2] and Zhe Hu[3,*]

[1] College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China
[2] BoYu Intelligent Health Innovation Laboratory, Hangzhou 311121, China
[3] Department of Anesthesiology, Affiliated Hospital of Guangdong Medical University, Guangdong 524001, China

* **Correspondence:** Email: 11219021@zju.edu.cn, biohuzhe@hotmail.com.

Abstract: Due to the exquisite ability of cancer stemness to facilitate tumor initiation, metastasis, and cancer therapy resistance, targeting cancer stemness is expected to have clinical implications for cancer treatment. Genes are fundamental for forming and maintaining stemness. Considering shared genetic programs and pathways between embryonic stem cells and cancer stem cells, we conducted a study analyzing transcriptomic data of embryonic stem cells for mining potential cancer stemness genes. Firstly, we integrated co-expression and regression models and predicted 820 stemness genes. Results of gene enrichment analysis confirmed the good prediction performance for enriched signatures in cancer stem cells. Secondly, we provided an application case using the predicted stemness genes to construct a breast cancer stemness network. Mining on the network identified CD44, SOX2, TWIST1, and DLG4 as potential regulators of breast cancer stemness. Thirdly, using the signature of 31,028 chemical perturbations and their correlation with stemness marker genes, we predicted 67 stemness inhibitors with reasonable accuracy of 78%. Two drugs, namely Rigosertib and Proscillaridin A, were first identified as potential stemness inhibitors for melanoma and colon cancer, respectively. Overall, mining embryonic stem cell data provides a valuable way to identify cancer stemness regulators.

Keywords: cancer; cancer stem cell; stemness gene; embryonic stem cell; stemness inhibitor

1. Introduction

Cancer stemness refers to the stem-cell-like phenotypes, such as self-renew and differentiation, of cancer cells [1], which is a crucial property of cancer stem cells (CSCs) [2]. CSCs contribute to tumor initiation, progression, metastasis, and therapeutic resistance through stem cell programs [3,4]. In particular, the aberrant expression of central stemness transcription factors, such as NANOG, OCT4, KLF4, SOX2, and c-MYC, is associated with the initiation and progression of poorly differentiated aggressive human tumors [5]. KLF4 is involved in the brain metastasis of breast cancer stem-like cells [6]. MYC targeted CDK18-mediated PARP inhibitor resistance in glioblastoma by promoting ATR and homologous recombination [7]. Therefore, understanding and targeting cancer stemness might provide a significant cornerstone for the cancer therapy [8–12].

To better understand cancer stemness, it is necessary to identify stemness genes. The wet experiment is the most accurate way to identify stemness genes. Genetic manipulation using knockdown, knockout, and overexpression strategies are commonly used to discover stemness genes, contributing to identifying YAP/TAZ as glioblastoma stemness master [13], and Alk, Bclaf3, and Prkra as gastric stemness genes [14]. On the other hand, computational analysis provides a cost-effective alternative to exploring stemness genes via systematic investigation of omics data. The current computational strategies for predicting stemness genes mainly rely on transcriptome analysis [15–19] and epigenetic analysis [15]. Notably, mRNA expression-based (transcriptome level-based) stemness index shows a positive correlation with DNA methylation-based (epigenetic level-based) stemness index, and both indices show good correspondence for a majority of tumors [15]. Thus, information extracted from transcriptome data could greatly help stemness gene identification. Combined with known stemness markers, co-expression analysis on transcriptome data could further investigate gene-stemness marker interaction, thus identifying novel stemness genes.

Cancer and embryonic stem cells share genetic fingerprints [20]. Abundant signaling pathways overlap in cancer and embryonic stem cells [21]. Furthermore, more and more recent studies showed that embryonic stemness gene-coding proteins such as Oct4/Pou5f1 [22], Klf4 [23], Sox2 [24], Myc [25], and Nanog [26], also played regulatory roles in cancer stemness, which supported the possibility of mining embryonic stem cell data in identifying novel cancer stemness genes.

Given these, we developed an approach called Transcriptome-based Stemness Regulator Prediction (TSRP), which integrated co-expression and regression analysis on embryonic stem cell data, and accurately predicted cancer stemness genes, regulators, and networks.

2. Materials and methods

2.1. Stemness markers curation

In this study, we curated 21 well-known stemness markers from the Cell paper [15].

2.2. Transcriptome profiles for 73 different knockout and wildtype mouse embryonic stem cell lines

We curated 319 samples from GSE145653 [27]. These mouse embryonic stem cell samples were cultured in either 2i (2i for two inhibitors, CHIR99021 and PD0325901) medium to maintain the naïve pluripotency state, or N2B27 medium to go through differentiation. Read counts obtained

from GSE145653 were normalized with the Trimmed Mean of M-values (TMM) method [28] using edgeR [29] for further analysis.

2.3. Public RNA-seq data

Dabrafenib data were downloaded from GSE162045 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE162045), and the processed abundance values were used directly for gene set enrichment analysis (GSEA) [30]. Data of Rigosertib, Triptolide, and Chaetocin were downloaded from GSE149737 [31], GSE157927 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE157927), and GSE97215 [32], respectively. The downloaded count data were then normalized with the TMM method as described above for GSEA. Proscillaridin A data were downloaded from GSE89154 [33], and the processed expression values were used directly for GSEA.

2.4. Differentially expressed genes following small molecule perturbation

The "LINCS L1000 CMAP Signatures of Differentially Expressed Genes for Small Molecules" dataset, consisting of 31,028 differentially expressed downregulated gene sets by small-molecule perturbation, was curated from the Harmonizome database [34,35].

2.5. Stemness marker clustering

Since different stemness markers could regulate stem-like features through different pathways, we investigated the co-expression relationship among these markers. Pearson correlation coefficient (Pcc) was computed using Eq (1) for each pair of makers:

$$Pcc_{i,j} = \frac{\mathrm{cov}\left(v_i, v_j\right)}{\sigma_{v_i}\sigma_{v_j}} \tag{1}$$

Here, with v_i and v_j as the vector containing expression values of marker i and j under all conditions. Then, K-means clustering was conducted to group these 21 stemness markers into three clusters. Stemness owns the ability of a cell to perpetuate its lineage, and the adult mammalian organs are developed from three lineages during embryogenesis [2]. Thus, we set the K as three in the present study.

2.6. Transcriptome-based stemness regulator prediction (TSRP)

The TSRP prediction algorithm was divided into two parts, namely stemness gene prediction and stemness inhibitor prediction. The stemness gene prediction algorithm consists of two steps. In step one, 21 known stemness markers and transcriptome data were used as input, and the correlation of a new gene with all 21 known stemness markers was evaluated by StemScore (Stemness Score). In step two, high correlation genes with StemScores passing the threshold were predicted as stemness genes. When a gene is correlated with different groups of stemness markers, we use a regression forecasting model to determine whether the gene is a potential stemness gene. In detail, the StemScore of a new gene x, which is not included in the list of 21 stemness markers, was computed with the overall effect

on known marker genes by the following formulas (Eqs (2)–(5)):

$$StemScore_x = a * Score_{c1,x} + b * Score_{c2,x} + c * Score_{c3,x} \tag{2}$$

$$Score_{c1,x} = \frac{\sum_{m=1}^{n} Pcc_{m.x}}{n}, m \in cluster\ 1 \tag{3}$$

$$Score_{c2,x} = \frac{\sum_{o=1}^{p} Pcc_{o.x}}{p}, o \in clsuter\ 2 \tag{4}$$

$$Score_{c3,x} = \frac{\sum_{s=1}^{t} Pcc_{s,x}}{t}, s \in cluster\ 3 \tag{5}$$

Here, clusters 1–3 were defined by the above clustering analysis. $Score_{c1,x}$, $Score_{c2,x}$, and $Score_{c3,x}$ represent relevance scores of the gene x with three clusters, respectively. The relevance scores were further weighted by a, b, and c to comprehensively consider their relevance with three maker clusters. To obtain a, b, and c, we solved the following matrix equation (Eq (6)) of 21 known stemness markers:

$$\begin{pmatrix} Score_{c1,1} & Score_{c2,1} & Score_{c3,1} \\ \vdots & \vdots & \vdots \\ Score_{c1,21} & Score_{c2,21} & Score_{c3,21} \end{pmatrix} * \begin{bmatrix} a & b & c \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 0 \\ \vdots \\ -1 \end{bmatrix} \tag{6}$$

We assigned StemScore 1, 0, and -1 to three clusters of 21 known stemness markers. For stemness markers in the same cluster, they were assigned the same score. In addition, the Pcc of a known stemness marker with itself was assigned to 0. We changed the order of assigned StemScores, and different sets of [a b c] were calculated to evaluate the prediction errors using L1-Norm distances between predicted and assigned StemScores. When the model achieved the lowest average prediction error, the used set was taken for the following study.

After obtaining the StemScore for each gene, these genes were divided into two groups, namely genes with StemScore large than 0 or small than 0. For each group, z-score normalization was used, and genes with z-score large than 1.65 or small than -1.65 were predicted as stemness genes.

2.7. Stemness gene prediction evaluation

To evaluate the performance of stemness gene prediction algorithm, we computed the overlap between the predicted stemness genes with curated chemical and genetic perturbation gene sets in MsigDB using the “Investigate Gene Sets” function [36].

2.8. Stemness inhibitor prediction of TSRP

To further predict the stemness inhibitors, we compared the overlap between differentially expressed downregulated genes by a small molecule perturbation and the stemness marker gene set. The hypergeometric test was used to test whether the perturbated genes were enriched with stemness markers, and p-value was obtained using the following formula (Eq (7)):

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}} \tag{7}$$

N is the array size, n is the number of perturbated genes, m is the number of stemness marker gene sets, k is the number of perturbated genes with the annotation as the stemness markers.

The p-value was further adjusted using a two-stage step-up method of Benjamini, Krieger and Yekutieli in Prism (version 9) to obtain a q-value. Only chemicals with a q-value less than 0.05 were used for the following analysis. For stemness inhibitor-like chemical prediction, the ratio of targeting stemness markers in downregulation signature was required to be larger than 0.1. For stemness inducer-like chemicals, the percentage of targeting stemness markers in the upregulation signature was required to be larger than 0.1. Then, we removed the overlapped chemicals from each group to obtain predicted stemness inhibitors and inducers, respectively.

2.9. Drug combination prediction of TSRP

For all predicted stemness inhibitors, they were randomly combined into two-inhibitor combinations. The signature for the combination was the union of downregulation signatures from both drugs. Then, we calculated the p-value for enrichment significance of stemness marker genes in the combination signature and the ratio of targeting stemness markers in the combination signature. Only combinations with ratio > 0.05 and p-value < 0.05 were used for following analysis. Next, we removed the drug combinations that either had repeated mechanisms of action (MOA) or unknown MOA since drugs from most synergistic drug combinations target different targets [37], and the combinations that passed the above requirements were predicted as stemness inhibitor drug combinations.

2.10. The overall workflow

TSRP employed co-expression analysis on prior knowledge of stemness markers and transcriptome data to predict different clusters of stemness genes. Furthermore, perturbation signatures of small molecules were used to select monotherapy and drug combinations based upon enrichment analysis (Figure 1).

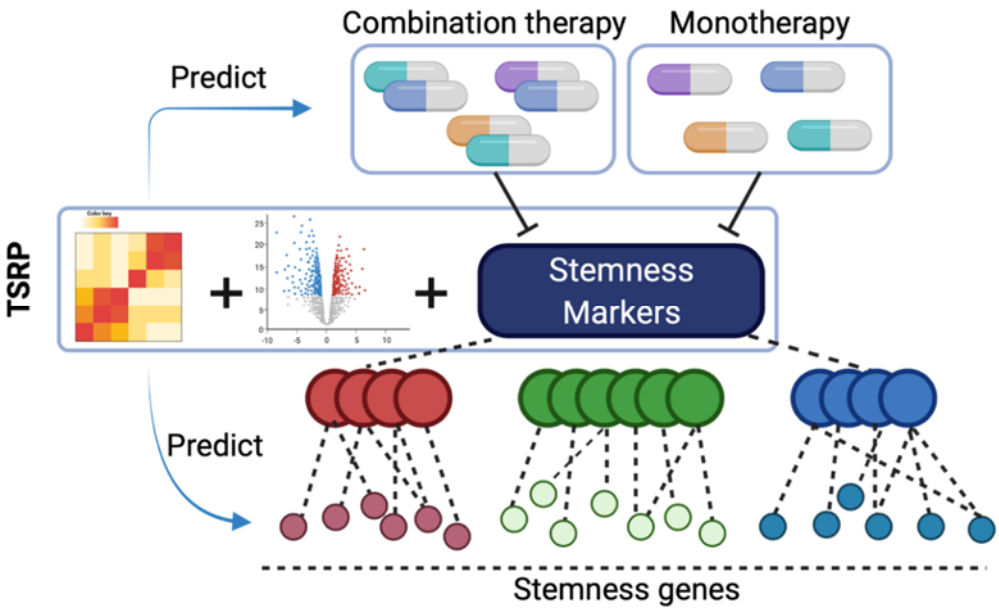

Figure 1. The overall workflow of TSRP.

3. Results

3.1. Three stemness marker clusters

We obtained a list of 21 stemness markers as defined previously in cancer and stem cells [15]. A gene co-expression analysis combined K-means clustering using RNA-seq data [27] in mouse embryonic stem cells identified three clusters (Figure 2A, the reason we chose three was described in the Method section). The first cluster (cluster 1) consisted of five genes, namely Nanog, Sox2, Klf4, Zfp42, and Notch1; the second cluster (cluster 2) consisted of seven genes, namely Cd44, Bmi1, Zscan4c, Cd34, Nes, Prom1, and Twist1; the third cluster (cluster 3) consisted of nine genes, namely Kdm5b, Pou5f1, Ctnnb1, Lgr5, Epas1, Myc, Abcg2, Ezh2, and Hif1a (Figure 2A). Based on these three clusters, we tried to find stemness genes similar to one of them. The co-expression analysis on stemness markers (Figure 2A) indicates that it may be excluded or likely to exhibit properties of the other two types of stemness genes when a stemness gene is similar to one of these three types. Such a situation may lead to genes (such as Prom1, as also indicated by the ward. D2 clustering result, Figure A1) located at the cluster's boundary. Therefore, we designed a simple regression formula (Eq (2)). The weights a, b, and c can be used to reflect this and help find these stemness genes that may locate at boundaries. To further understand the biological roles of the co-expressed genes of the above three clusters, we performed cell type enrichment analysis using Enrichr [38]. For cluster 1, the top 200 (around 1% of all genes) co-expressed genes (correlation coefficient ranged from 0.734 to 0.812, Figure 2B and Table S1) were enriched with cell type terms like gastrointestinal and intestinal cancer stem cells (Figure 2C and Table S2). Indeed, Notch1 from cluster 1 (Figure 2A) has been shown to play important roles in the gastric and intestinal stem cells [39]. The top 200 co-expressed genes of cluster 2 (correlation coefficient ranged from 0.449 to 0.540, Figure 2B and Table S1) were enriched with cardiovascular progenitor and tumor-propagating cell genes (Figure 2D and Table S2). The top 200 co-expressed genes of cluster 3 (correlation coefficient ranged from 0.437 to 0.516, Figure 2B and Table S1) were enriched with neural stem cell and pluripotent stem cell genes (Figure 2E and Table

S2). Overall, the above results perfectly reflect the stemness feature difference originating from three-lineage differentiation in embryogenesis. Of note, all these three groups were enriched in tumor-propagating cells (Figure 2F), which are well known for their role in delivering cancer stemness [40–43].

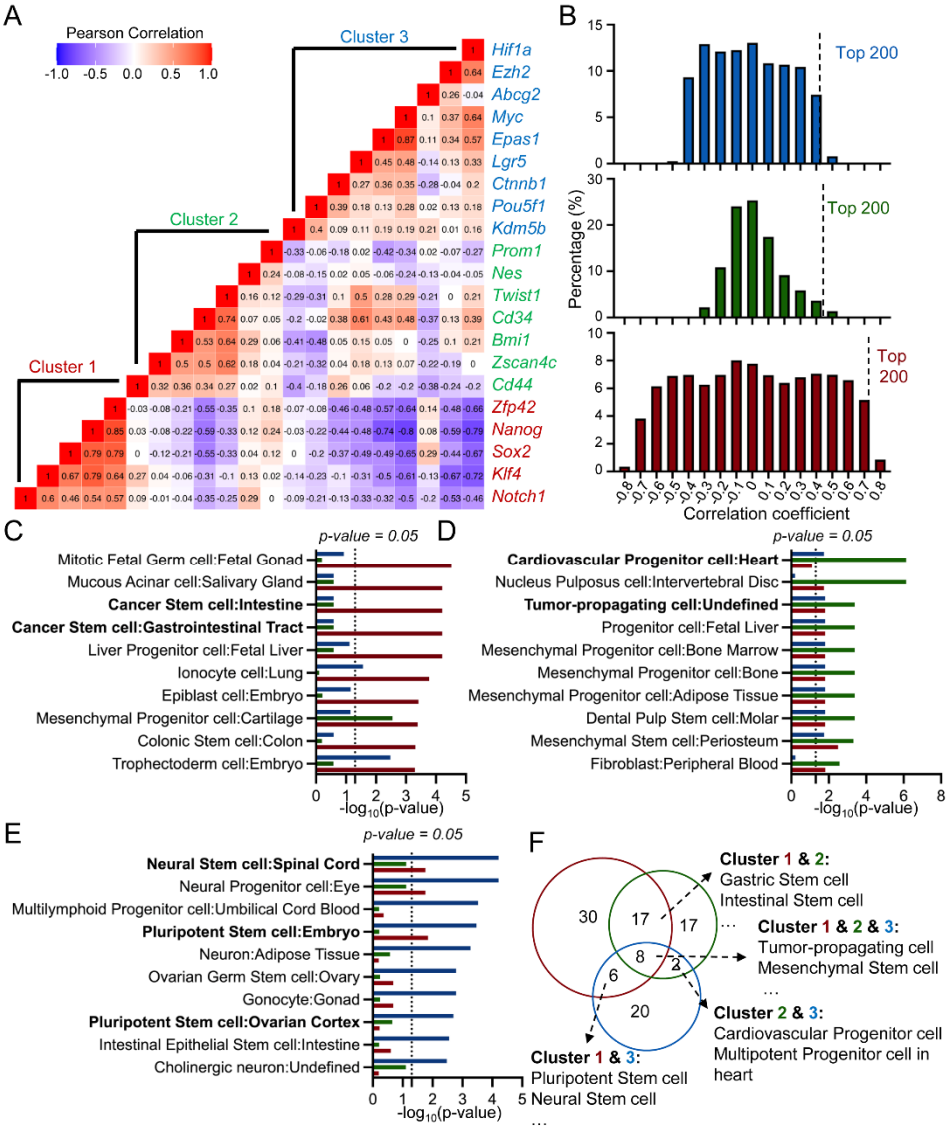

Figure 2. Three stemness marker clusters. (A). Clustering of 21 stemness markers; (B). Co-expressed genes of three stemness marker clusters; (C). Cell types enriched by cluster 1 co-expressed genes; (D). Cell types enriched by cluster 2 co-expressed genes; (E). Cell types enriched by cluster 3 co-expressed genes; (F). Veen diagram of cell types enriched by three groups of co-expressed genes.

3.2. Stemness gene prediction with TSRP

We trained a regression model based on the above-calculated co-expression relationships to obtain the group information of predicted stemness genes (see Methods for details). Briefly, we assigned StemScores 1, 0, and -1 to represent three groups as indicated above, and then compared three models with StemScores set as below: model 1:0 for cluster B, and either 1 or -1 for cluster

A/C (Figure 3A left part); model 2:0 for cluster A, and either 1 or -1 for cluster B/C (Figure 3A middle part); model 3:0 for cluster C, and either 1 or -1 for cluster A/B (Figure 3A right part). With the L1-Norm distance to evaluate the model prediction performance, model 1 was shown to have the lowest prediction error (Figure 3A). This model was used for the following prediction. With model 1, each gene's StemScore (see Method for detail) was calculated and normalized into a z-score. The absolute z-score reflects a gene's relatedness to stemness. After the filtering with absolute z-score > 1.65 (equivalent to obtaining genes with confidence level high than 90%), 242, 280, and 298 genes were predicted as cluster 1, cluster 2, and cluster 3-like stemness genes, respectively (Figure 3B,C and Table S3). To evaluate the predicted stemness genes, enrichment analysis was conducted to compare the overlap between predicted genes and upregulated genes in stem cell vs. non-stem cell samples using the "Investigate Gene Sets" function from the MsigDB website (https://www.gsea-msigdb.org/gsea/msigdb/index.jsp). As expected, stemness-related gene sets were enriched by all three predicted cluster-like stemness genes (Figure 3D–F, see the detailed description of these data sets in Table S4–S6). The above results suggest good prediction performance of the TSRP approach in predicting stemness genes.

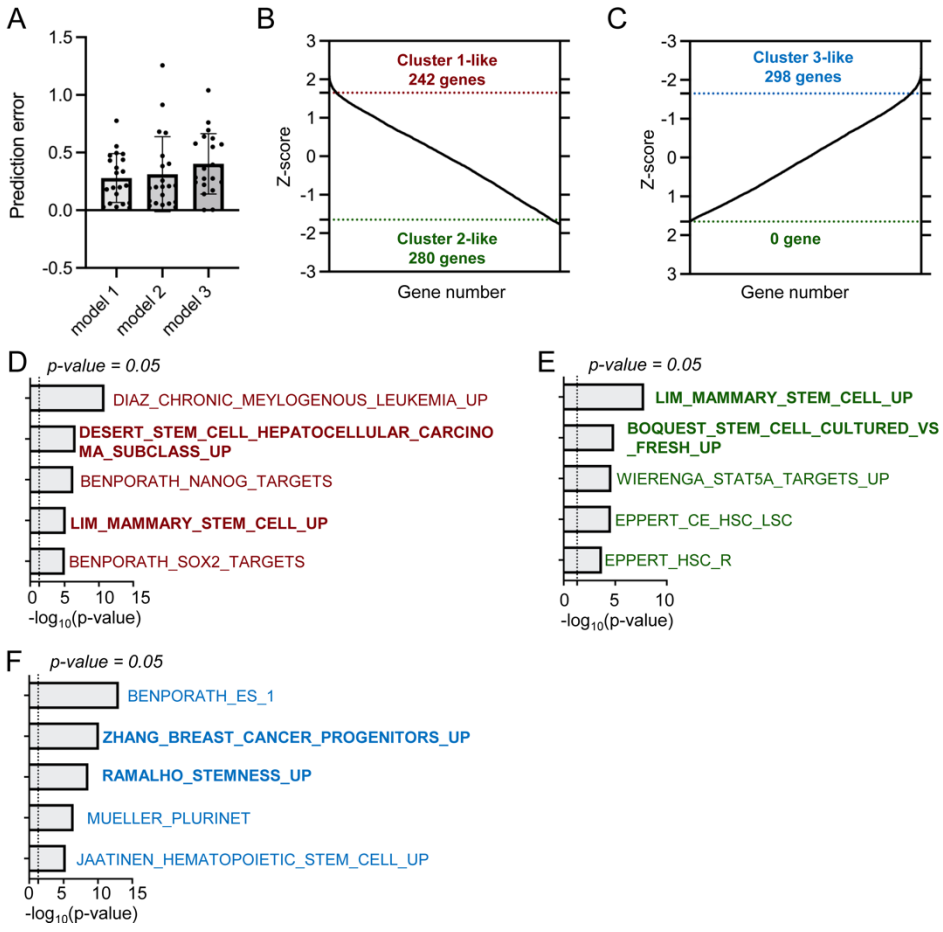

Figure 3. Stemness gene prediction and bioinformatic validation. (A). Evaluation of different regression models; (B). Prediction of cluster 1 and 2-like stemness genes; (C). Prediction of cluster 3-like stemness genes; (D–F). Enriched gene signature terms by predicted cluster 1 (D), 2 (E), 3 (F) -like stemness genes.

3.3. Stemness inhibitor and inducer prediction via TSRP

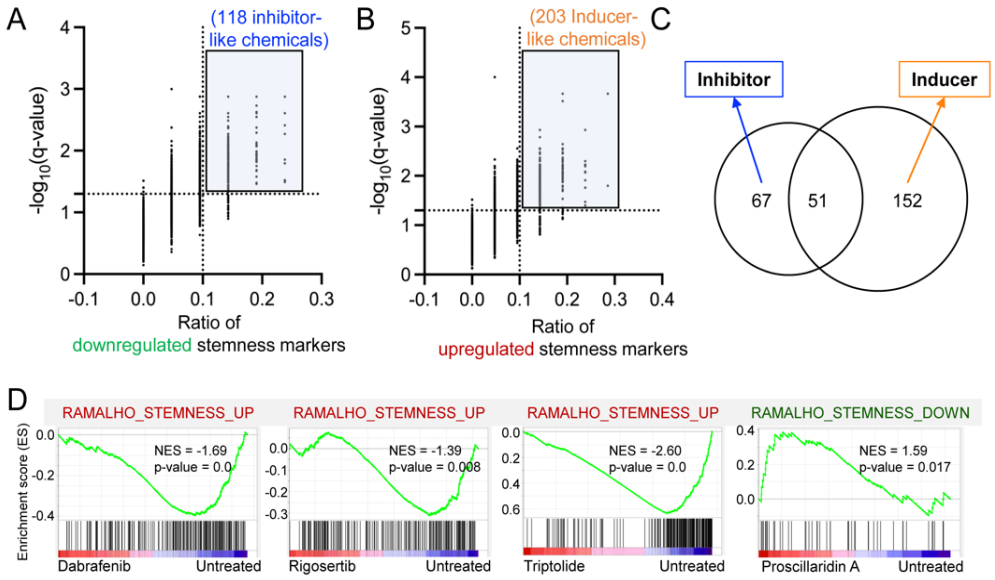

Figure 4. Stemness inhibitor prediction and validation. (A). Prediction of inhibitor-like chemicals by Transcriptome-based Stemness Regulator Prediction (TSRP); (B). Prediction of inducer-like chemicals by TSRP; (C). Overlap of predicted inhibitor- and inducer-like chemicals in B and C; (D). Reduced stemness signatures enriched by Dabrafenib, Rigosertib, Triptolide, and Proscillaridin A. NES for the normalized enrichment score in the gene set enrichment analysis.

To further test whether TSRP could identify the chemicals that can downregulate or upregulate stemness markers, we collected the "LINCS L1000 CMAP Signatures of Differentially Expressed Genes for Small Molecules" dataset from the Harmonizome database [34,35], which was then used for the prediction using TSRP. As a result, we predicted 118 stemness inhibitor-like chemicals (Figure 4A) and 203 stemness inducer-like chemicals (Figure 4B). The overlapped (i.e., both inhibitor as well as inducer; n = 51) chemicals were then removed, resulting in 67 predicted stemness inhibitors (Table S7) and 152 predicted stemness inducers (Table S8, Figure 4C). To evaluate the prediction performance, we searched the literature for above 67 predicted stemness inhibitors and only assigned a "TRUE" label for those that had been confirmed by published literature. Literature evaluation showed that 52 out of 67 chemicals had been validated as potential stemness inhibitors (Table S7). On the other hand, we found that a lot of phase 2 or 3 trial and approved drugs were predicted as stemness inducers, including seven glucocorticoid receptor agonists (Budesonide, Fluocinonide, Diflorasone Diacetate, Fluorometholone, Medrysone, Desoximetasone, and Prednicarbate), two mTOR inhibitors (Dactolisib, and Sapanisertib), two histone deacetylase inhibitors (Mocetinostat and Entinostat), and two MAPKK1/2 (MEK1/2) inhibitors (CI-1040 and Selumetinib). Following literature support, glucocorticoid receptor agonists [44–46], mTOR inhibitors [47,48], HDAC inhibitors [49], and MEK inhibitors [50] did show potential roles in enhancing the stemness. To prove that the identified drugs target stemness, we further mined the 67 predicted stemness inhibitors in the GEO database to collect available RNA-seq data of drug treatment and corresponding untreated control transcriptional profiles. We collected data for Dabrafenib, Rigosertib, Triptolide, Chaetocin, and Proscillaridin A. GSEA was conducted to compare drug-treated vs. untreated profiles. Four (Dabrafenib, Rigosertib, Triptolide,

and Proscillaridin A) out of the five above drugs investigated showed reduced stemness signature (Figure 4D). More importantly, Rigosertib and Proscillaridin A were first suggested by our study to inhibit the stemness of melanoma and colon cancer, respectively. Overall, TRSP performed very well in predicting the chemical inhibitors and inducers for stemness and thus shows its usability as a tool for investigating the relatedness of certain chemicals with stemness.

3.4. Stemness inhibitor combination design via TSRP

Monotherapy may be limited in its effectiveness, but drug combinations are supposed to overcome drug resistance and induce a more durable response in patients [37]. We, therefore, used the TSRP approach to identify whether this approach could be used to predict the outcomes of chemical combinations. We used the LINCS dataset [34] and performed chemicals-combinatorial analysis and monitored which chemical combinations mostly affected the expression of stemness genes. We identified 52 chemical combinations that affected the expression of stemness genes the most (Figure 5A). Next, we removed the drug combinations that either had repeated mechanisms of action (MOA) or MOA was unknown since drugs from most synergistic drug combinations target definite targets [37] and obtained 31 stemness inhibitor drug combinations (Figure 5B, Table S9). 14 out of these 31 combinations (45%) had the theoretical support from literature where one drug could help to overcome drug resistance of the second drug (Table S9). For example, the TSRP identified a combination of EGFR inhibitor (Allitinib) with SRC inhibitor (Bosutinib). The previous finding supports that the SRC inhibitor can overcome the EGFR inhibitor resistance in lung cancer cells [51,52]. Similarly, the TSRP identified a combination of B-raf inhibitor (Dabrafenib) and SRC inhibitor (Bosutinib), which is also supported by the finding that inhibiting SRC family kinase signaling overcomes BRAF inhibitor resistance in melanoma [53]. Furthermore, the combinatorial effects from the drug-combination analysis also reveal the common stemness gene targets of both drugs compared with monotherapy (Figure 5C,D). Taken together, TSRP can help rationally design drug combinations and reveal potential mechanisms for combating cancer stemness.

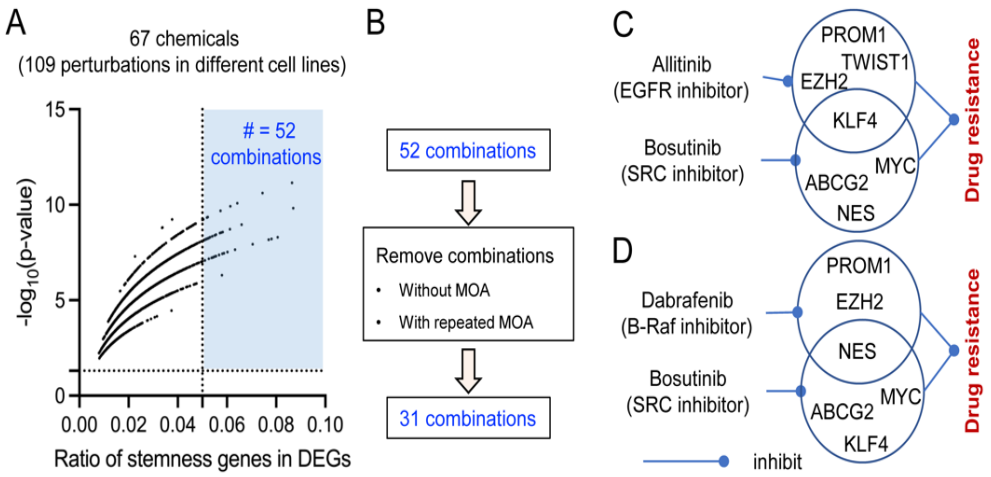

Figure 5. Stemness inhibitor combination prediction and validation. (A). Prediction of stemness inhibitor combinations by TSRP; (B). Filtering strategy in the combination prediction of TSRP; (C,D). The potential mechanism of combination Allitinib+Bosutinib (C) and Dabrafenib+Bosutinib (D) in combating drug resistance.

3.5. MCF-7 stemness network

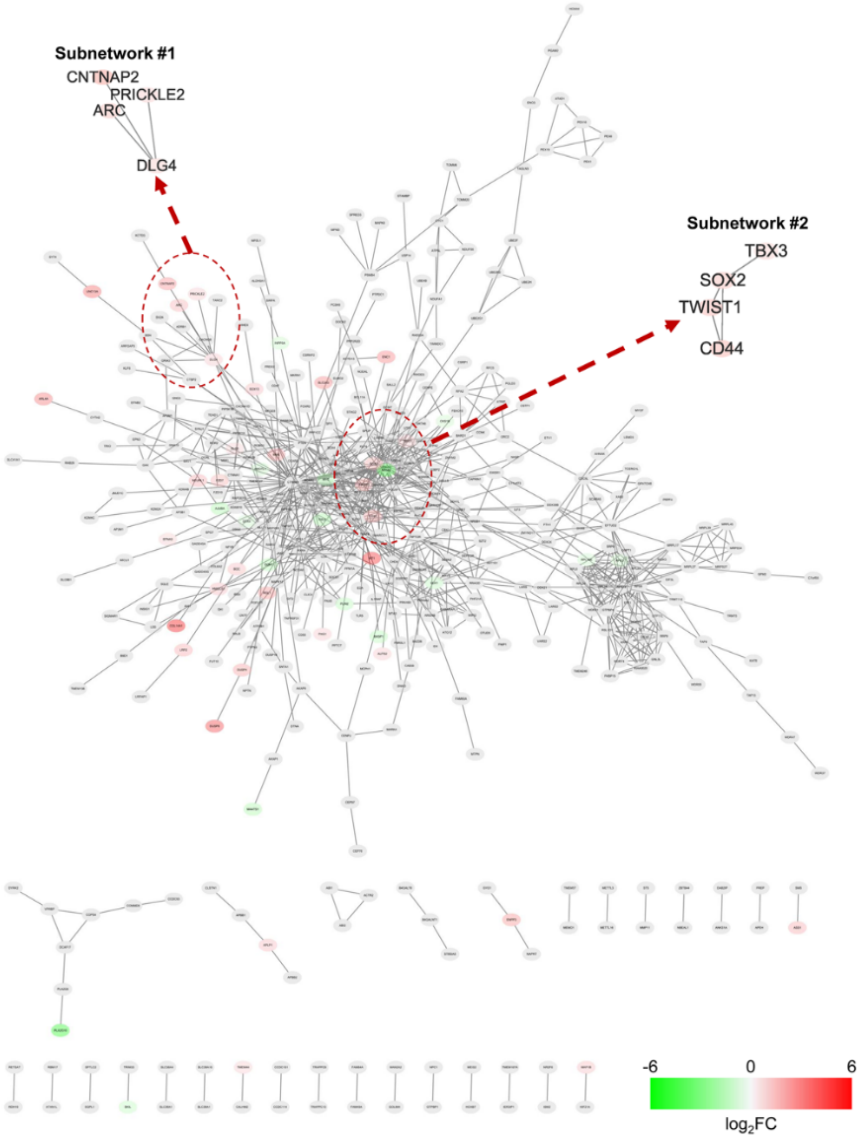

Figure 6. MCF-7 stemness network. The network is visualized by the Cytoscape software. Red nodes for significantly upregulated genes, green nodes for significantly downregulated genes, and grey nodes for no significantly changed genes. Two subnetworks, as indicated by arrows, are identified for connecting red nodes.

Using the 820 predicted stemness genes and 21 known marker genes, we took coding proteins of these genes as input of STRING database [54]. From the database, we extracted high confidence protein-protein interactions (PPIs) with combined scores (defined by STRING database) > 0.7 and constructed a stemness PPI network consisting of 427 proteins and 937 PPIs (Figure 6 and Table S10). We mapped transcriptional profiles (GSE182532 [55]) of MCF-7-derived mammospheres and parental breast cancer cells onto the above network, and only significantly changed genes (log2FC > 1 or < -1, p-value < 0.05, and FC for fold change of expression values in mammospheres vs. parental cells) were mapped with colors using Cytoscape [56] (red for upregulation, and green for downregulation). The network showed a global upregulation of stemness genes with the network stemness score, namely the average log2FC of all significantly changed genes, increased to 0.786 while the baseline score is 0. To

further identify key stemness regulators in the MCF-7 stemness network, we extracted subnetworks consisting of more than four proteins. As a result, two subnetworks of significantly upregulated stemness gene-coding proteins were identified but 0 for significantly downregulated stemness gene-coding proteins (Figure 6). CD44, SOX2, TWIST1, and DLG4 have more than two interactions with other upregulated stemness partners (Figure 6) and were suggested as key regulators in breast cancer stemness. In line with the above speculation, previous studies have demonstrated the roles of CD44 [57], TWIST1 [58], SOX2 [59] in promoting breast cancer stemness. There is no published report for DLG4 in breast cancer stemness, which might need more exploratory studies. Therefore, the case study provides a good example of our predicted genes to study cancer stemness.

4. Discussion

Cancer stemness plays roles in metastasis, recurrence, and resistance to treatment, and targeting stemness has been proposed as a promising approach for the complete treatment of cancer [4,11,12,60,61]. However, cancer stemness is not governed by a single gene or pathway but by a complex network [12]. Thus, it is necessary to uncover the components within the network systematically. In this study, we explored the usage of embryonic stem cell data in identifying stemness genes. We further mapped transcriptome data of breast cancer to these genes to construct a breast cancer stemness network as a case study, thus identifying critical regulators for cancer stemness regulation.

With the urgent need to develop therapies for combating cancer stemness, we further provide a computational strategy to predict stemness inhibitors that can significantly combat but not enhance stemness markers at the transcriptome level. Meanwhile, the strategy can exclude potential stemness-inducing chemicals for identifying true stemness inhibitors. On the other hand, the current strategy might overlook some stemness-related chemicals due to their interactions with unknown stemness markers. Considering the advantage of combinatorial drugs in enhancing efficacy and reducing drug resistance induced by monotherapy, we provided a computational strategy to design stemness inhibitor combinations. In this strategy, we comprehensively investigate the ratio of downregulated stemness markers by two drugs compared to single drugs to priorate drug combinations. Thus, TSRP can simultaneously predict monotherapy and combination therapy, while most prediction approaches focus on either one [62–65]. TSRP can predict therapeutic drugs with good accuracy and owns the potential in unrevealing mechanisms, which is comparable with our previously developed drug repositioning approach, namely causal inference-probabilistic matrix factorization (CI-PMF) [66]. However, the CI-PMF approach can only be applied to drugs with known target records. In terms of this, TSRP can be easily used on chemicals with transcriptome data, even for chemicals without a clear mechanism of action, such as in Traditional Chinese Medicine. In summary, the TSRP approach allows predicting stemness genes and discovering stemness inhibitors for combating cancer stemness.

In the present study, we only chose the 21 known stemness markers that contribute to normal and cancer cells' stemness [15]. In addition to these known stemness markers, targets reported with stemness-promoting roles can also be used, which might lead us to find novel stemness genes. Also, more advanced clustering, classification, and regression models, such as DBSCAN [67], Gaussian mixture model [68], BIRCH [69], SVM [70], Multinomial Logistic Regression [71] could be tested on larger-scale data sets to get a better stemness gene prediction performance. In addition, for the drug combinations study, drug-drug interactions are not considered. We would like to improve our approach further to incorporate the drug-drug interaction information and more comprehensive stemness markers to improve the prediction performance in our future work.

Acknowledgments

This research was funded by the National Natural Science Foundation of China, grant number 81870222 and also by the Natural Science Foundation of Guangdong Province, grant number 2018A030307069 and 2021A1515012167.

Conflict of interest

The authors declare no conflict of interest.

References

1. Y. M. Tsui, L. K. Chan, I. O. Ng, Cancer stemness in hepatocellular carcinoma: Mechanisms and translational potential, *Br. J. Cancer*, **122** (2020), 1428–1440. https://doi.org/10.1038/s41416-020-0823-9
2. P. M. Aponte, A. Caicedo, Stemness in cancer: Stem cells, cancer stem cells, and their microenvironment, *Stem Cells Int.*, **2017** (2017), 5619472. https://doi.org/10.1155/2017/5619472
3. A. Z. Ayob, T. S. Ramasamy, Cancer stem cells as key drivers of tumour progression, *J. Biomed. Sci.*, **25** (2018), 20. https://doi.org/10.1186/s12929-018-0426-4
4. T. Huang, X. Song, D. Xu, D. Tiek, A. Goenka, B. Wu, et al., Stem cell programs in cancer initiation, progression, and therapy resistance, *Theranostics*, **10** (2020), 8721–8743. https://doi.org/10.7150/thno.41648
5. I. Ben-Porath, M. W. Thomson, V. J. Carey, R. Ge, G. W. Bell, A. Regev, et al., An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors, *Nat. Genet.*, **40** (2008), 499–507. https://doi.org/10.1038/ng.127
6. H. Okuda, F. Xing, P. R. Pandey, S. Sharma, M. Watabe, S. K. Pai, et al., miR-7 suppresses brain metastasis of breast cancer stem-like cells by modulating KLF4, *Cancer Res.*, **73** (2013), 1434–1444. https://doi.org/10.1158/0008-5472.CAN-12-2037
7. J. F. Ning, M. Stanciu, M. R. Humphrey, J. Gorham, H. Wakimoto, R. Nishihara, et al., Myc targeted CDK18 promotes ATR and homologous recombination to mediate PARP inhibitor resistance in glioblastoma, *Nat. Commun.*, **10** (2019), 2910. https://doi.org/10.1038/s41467-019-10993-5
8. Y. Li, H. A. Rogoff, S. Keates, Y. Gao, S. Murikipudi, K. Mikule, et al., Suppression of cancer relapse and metastasis by inhibiting cancer stemness, *Proc. Natl. Acad. Sci. U. S. A.*, **112** (2015), 1839–1844. https://doi.org/10.1073/pnas.1424171112
9. C. Saygin, D. Matei, R. Majeti, O. Reizes, J. D. Lathia, Targeting cancer stemness in the clinic: From hype to hope, *Cell Stem Cell*, **24** (2019), 25–40. https://doi.org/10.1016/j.stem.2018.11.017
10. A. Kreso, P. van Galen, N. M. Pedley, E. Lima-Fernandes, C. Frelin, T. Davis, et al., Self-renewal as a therapeutic target in human colorectal cancer, *Nat. Med.*, **20** (2014), 29–36. https://doi.org/10.1038/nm.3418
11. S. Prasad, S. Ramachandran, N. Gupta, I. Kaushik, S. K. Srivastava, Cancer cells stemness: A doorstep to targeted therapy, *Biochim. Biophys. Acta Mol. Basis Dis.*, **1866** (2020), 165424. https://doi.org/10.1016/j.bbadis.2019.02.019

12. L. Yang, P. Shi, G. Zhao, J. Xu, W. Peng, J. Zhang, et al., Targeting cancer stem cell pathways for cancer therapy, *Signal Transduct. Target. Ther.*, **5** (2020), 8. https://doi.org/10.1038/s41392-020-0110-5
13. M. Castellan, A. Guarnieri, A. Fujimura, F. Zanconato, G. Battilana, T. Panciera, et al., Single-cell analyses reveal YAP/TAZ as regulators of stemness and cell plasticity in Glioblastoma, *Nat. Cancer*, **2** (2021), 174–188. https://doi.org/10.1038/s43018-020-00150-z
14. K. Murakami, Y. Terakado, K. Saito, Y. Jomen, H. Takeda, M. Oshima, et al., A genome-scale CRISPR screen reveals factors regulating Wnt-dependent renewal of mouse gastric epithelial cells, *Proc. Natl. Acad. Sci. U. S. A.*, **118** (2021), e2016806118. https://doi.org/10.1073/pnas.2016806118
15. T. M. Malta, A. Sokolov, A. J. Gentles, T. Burzykowski, L. Poisson, J. N. Weinstein, et al., Machine learning identifies stemness features associated with oncogenic dedifferentiation, *Cell*, **173** (2018), 338–354. https://doi.org/10.1016/j.cell.2018.03.034
16. K. Borziak, J. Finkelstein, Identification of liver cancer stem cell stemness markers using a comparative analysis of public data sets, *Stem Cells Cloning*, **14** (2021), 9–17. https://doi.org/10.2147/SCCAA.S307043
17. C. Huang, C. G. Hu, Z. K. Ning, J. Huang, Z. M. Zhu, Identification of key genes controlling cancer stem cell characteristics in gastric cancer, *World J. Gastrointest. Surg.*, **12** (2020), 442–459. https://doi.org/10.4240/wjgs.v12.i11.442
18. H. D. Suo, Z. Tao, L. Zhang, Z. N. Jin, X. Y. Li, W. Ma, et al., Coexpression network analysis of genes related to the characteristics of tumor stemness in triple-negative breast cancer, *Biomed. Res. Int.*, **2020** (2020), 7575862. https://doi.org/10.1155/2020/7575862
19. Z. Wang, D. Wu, Y. Xia, B. Yang, T. Xu, Identification of hub genes and compounds controlling ovarian cancer stem cell characteristics via stemness indices analysis, *Ann. Transl. Med.*, **9** (2021), 379. https://doi.org/10.21037/atm-20-3621
20. M. Baker, Cancer and embryonic stem cells share genetic fingerprints, *Nat. Rep. Stem Cells*, **2008** (2008), 1. https://doi.org/10.1038/stemcells.2008.62
21. O. Dreesen, A. H. Brivanlou, Signaling pathways in cancer and embryonic stem cells, *Stem Cell Rev.*, **3** (2007), 7–17. https://doi.org/10.1007/s12015-007-0004-8
22. H. Lu, Y. Xie, L. Tran, J. Lan, Y. Yang, N. L. Murugan, et al., Chemotherapy-induced S100A10 recruits KDM6A to facilitate OCT4-mediated breast cancer stemness, *J. Clin. Invest.*, **130** (2020), 4607–4623. https://doi.org/10.1172/JCI138577
23. K. Ganguly, S. R. Krishn, S. Rachagani, R. Jahan, A. Shah, P. Nallasamy, et al., Secretory mucin 5AC promotes neoplastic progression by augmenting KLF4-mediated pancreatic cancer cell stemness, *Cancer Res.*, **81** (2021), 91–102. https://doi.org/10.1158/0008-5472.CAN-20-1293
24. M. A. Mamun, K. Mannoor, J. Cao, F. Qadri, X. Song, SOX2 in cancer stemness: Tumor malignancy and therapeutic potentials, *J. Mol. Cell Biol.*, **12** (2020), 85–98. https://doi.org/10.1093/jmcb/mjy080
25. Y. Liu, C. Zhu, L. Tang, Q. Chen, N. Guan, K. Xu, et al., MYC dysfunction modulates stemness and tumorigenesis in breast cancer, *Int. J. Biol. Sci.*, **17** (2021), 178–187. https://doi.org/10.7150/ijbs.51458
26. J. Zhang, L. A. Espinoza, R. J. Kinders, S. M. Lawrence, T. D. Pfister, M. Zhou, et al., NANOG modulates stemness in human colorectal cancer, *Oncogene*, **32** (2013), 4397–4405. https://doi.org/10.1038/onc.2012.461

27. A. Lackner, R. Sehlke, M. Garmhausen, G. Stirparo, M. Huth, F. Titz-Teixeira, et al., Cooperative genetic networks drive embryonic stem cell transition from naive to formative pluripotency, *EMBO J.*, **40** (2021), e105776. https://doi.org/10.15252/embj.2020105776
28. M. D. Robinson, A. Oshlack, A scaling normalization method for differential expression analysis of RNA-seq data, *Genome Biol.*, **11** (2010), R25. https://doi.org/10.1186/gb-2010-11-3-r25
29. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: A Bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics*, **26** (2010), 139–140. https://doi.org/10.1093/bioinformatics/btp616
30. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, et al., Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, *Proc. Natl. Acad. Sci. U. S. A.*, **102** (2005), 15545–15550. https://doi.org/10.1073/pnas.0506580102
31. C. Yan, N. Saleh, J. Yang, C. A. Nebhan, A. E. Vilgelm, E. P. Reddy, et al., Novel induction of CD40 expression by tumor cells with RAS/RAF/PI3K pathway inhibition augments response to checkpoint blockade, *Mol. Cancer*, **20** (2021), 85. https://doi.org/10.1186/s12943-021-01366-y
32. A. Mathison, A. Salmonson, M. Missfeldt, J. Bintz, M. Williams, S. Kossak, et al., Combined AURKA and H3K9 methyltransferase targeting inhibits cell growth by inducing mitotic catastrophe, *Mol. Cancer Res.*, **15** (2017), 984–997. https://doi.org/10.1158/1541-7786.MCR-17-0063
33. N. J. Raynal, E. M. Da Costa, J. T. Lee, V. Gharibyan, S. Ahmed, H. Zhang, et al., Repositioning FDA-approved drugs in combination with epigenetic drugs to reprogram colon cancer epigenome, *Mol. Cancer Ther.*, **16** (2017), 397–407. https://doi.org/10.1158/1535-7163.MCT-16-0588
34. Q. Duan, C. Flynn, M. Niepel, M. Hafner, J. L. Muhlich, N. F. Fernandez, et al., LINCS Canvas Browser: Interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures, *Nucleic Acids Res.*, **42** (2014), W449–W460. https://doi.org/10.1093/nar/gku476
35. A. D. Rouillard, G. W. Gundersen, N. F. Fernandez, Z. Wang, C. D. Monteiro, M. G. McDermott, et al., The harmonizome: A collection of processed datasets gathered to serve and mine knowledge about genes and proteins, *Database*, **2016** (2016), 1–16. https://doi.org/10.1093/database/baw100
36. A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdottir, P. Tamayo, J. P. Mesirov, Molecular signatures database (MSigDB) 3.0, *Bioinformatics*, **27** (2011), 1739–1740. https://doi.org/10.1093/bioinformatics/btr260
37. J. Jia, F. Zhu, X. Ma, Z. Cao, Z. W. Cao, Y. Li, et al., Mechanisms of drug combinations: Interaction and network perspectives, *Nat. Rev. Drug Discov.*, **8** (2009), 111–128. https://doi.org/10.1038/nrd2683
38. Z. Xie, A. Bailey, M. V. Kuleshov, D. J. B. Clarke, J. E. Evangelista, S. L. Jenkins, et al., Gene set knowledge discovery with enrichr, *Curr. Protoc.*, **1** (2021), e90. https://doi.org/10.1002/cpz1.90
39. E. S. Demitrack, L. C. Samuelson, Notch regulation of gastrointestinal stem cells, *J. Physiol.*, **594** (2016), 4791–4803. https://doi.org/10.1113/JP271667
40. S. Boumahdi, G. Driessens, G. Lapouge, S. Rorive, D. Nassar, M. Le Mercier, et al., SOX2 controls tumour initiation and cancer stem-cell functions in squamous-cell carcinoma, *Nature*, **511** (2014), 246–250. https://doi.org/10.1038/nature13305
41. K. Rycaj, D. G. Tang, Cell-of-Origin of cancer versus cancer stem cells: Assays and interpretations, *Cancer Res.*, **75** (2015), 4003–4011. https://doi.org/10.1158/0008-5472.CAN-15-0798

42. F. Papaccio, F. Paino, T. Regad, G. Papaccio, V. Desiderio, V. Tirino, Concise review: Cancer cells, cancer stem cells, and mesenchymal stem cells: Influence in cancer development, *Stem Cells Transl. Med.*, **6** (2017), 2115–2125. https://doi.org/10.1002/sctm.17-0138
43. S. Floor, W. C. van Staveren, D. Larsimont, J. E. Dumont, C. Maenhaut, Cancer cells in epithelial-to-mesenchymal transition and tumor-propagating-cancer stem cells: Distinct, overlapping or same populations, *Oncogene*, **30** (2011), 4609–4621. https://doi.org/10.1038/onc.2011.184
44. H. Y. Lee, X. Gao, M. I. Barrasa, H. Li, R. R. Elmes, L. L. Peters, et al., PPAR-alpha and glucocorticoid receptor synergize to promote erythroid progenitor self-renewal, *Nature*, **522** (2015), 474–477. https://doi.org/10.1038/nature14326
45. K. N. Grise, N. X. Bautista, K. Jacques, B. L. K. Coles, D. van der Kooy, Glucocorticoid agonists enhance retinal stem cell self-renewal and proliferation, *Stem Cell Res. Ther.*, **12** (2021), 83. https://doi.org/10.1186/s13287-021-02136-9
46. H. Karvonen, M. Arjama, L. Kaleva, W. Niininen, H. Barker, R. Koivisto-Korander, et al., Glucocorticoids induce differentiation and chemoresistance in ovarian cancer by promoting ROR1-mediated stemness, *Cell Death Dis.*, **11** (2020), 790. https://doi.org/10.1038/s41419-020-03009-4
47. P. Agrawal, J. Reynolds, S. Chew, D. A. Lamba, R. E. Hughes, DEPTOR is a stemness factor that regulates pluripotency of embryonic stem cells, *J. Biol. Chem.*, **289** (2014), 31818–31826. https://doi.org/10.1074/jbc.M114.565838
48. S. Wang, P. Xia, B. Ye, G. Huang, J. Liu, Z. Fan, Transient activation of autophagy via Sox2-mediated suppression of mTOR is an important early step in reprogramming to pluripotency, *Cell Stem Cell*, **13** (2013), 617–625. https://doi.org/10.1016/j.stem.2013.10.005
49. L. Mousazadeh, E. Alizadeh, N. Zarghami, S. Hashemzadeh, S. F. Aval, L. Hasanifard, et al., Histone deacetylase inhibitor (Trapoxin A) enhances stemness properties in adipose tissue derived mesenchymal stem cells, *Drug Res.*, **68** (2018), 450–456. https://doi.org/10.1055/s-0044-102007
50. T. Zhan, G. Ambrosi, A. M. Wandmacher, B. Rauscher, J. Betge, N. Rindtorff, et al., MEK inhibitors activate Wnt signalling and induce stem cell plasticity in colorectal cancer, *Nat. Commun.*, **10** (2019), 2197.https://doi.org/10.1038/s41467-019-09898-0
51. A. Robles-Perez, J. Dorca, I. Castellvi, J. M. Nolla, M. Molina-Molina, J. Narvaez, Rituximab effect in severe progressive connective tissue disease-related lung disease: Preliminary data, *Rheumatol. Int.*, **40** (2020), 719–726. https://doi.org/10.1007/s00296-020-04545-0
52. Y. Murakami, K. Sonoda, H. Abe, K. Watari, D. Kusakabe, K. Azuma, et al., The activation of SRC family kinases and focal adhesion kinase with the loss of the amplified, mutated EGFR gene contributes to the resistance to afatinib, erlotinib and osimertinib in human lung cancer cells, *Oncotarget*, **8** (2017), 70736–70751. https://doi.org/10.18632/oncotarget.19982
53. M. R. Girotti, M. Pedersen, B. Sanchez-Laorden, A. Viros, S. Turajlic, D. Niculescu-Duvaz, et al., Inhibiting EGF receptor or SRC family kinase signaling overcomes BRAF inhibitor resistance in melanoma, *Cancer Discov.*, **3** (2013), 158–167. https://doi.org/10.1158/2159-8290.CD-12-0386
54. D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, et al., The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets, *Nucleic Acids Res.*, **49** (2021), D605–D612. https://doi.org/10.1093/nar/gkaa1074
55. J. Verigos, D. Kordias, S. Papadaki, A. Magklara, Transcriptional profiling of tumorspheres reveals trpm4 as a novel stemness regulator in breast cancer, *Biomedicines*, **9** (2021), 1368. https://doi.org/10.3390/biomedicines9101368

56. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, et al., Cytoscape: A software environment for integrated models of biomolecular interaction networks, *Genome Res.*, **13** (2003), 2498–2504. https://doi.org/10.1101/gr.1239303
57. P. Huang, A. Chen, W. He, Z. Li, G. Zhang, Z. Liu, et al., BMP-2 induces EMT and breast cancer stemness through Rb and CD44, *Cell Death Discov.*, **3** (2017), 17039. https://doi.org/10.1038/cddiscovery.2017.39
58. Y. Liang, J. Hu, J. Li, Y. Liu, J. Yu, X. Zhuang, et al., Epigenetic activation of TWIST1 by MTDH promotes cancer stem-like cell traits in breast cancer, *Cancer Res.*, **75** (2015), 3672–3680. https://doi.org/10.1158/0008-5472.CAN-15-0930
59. J. M. Yu, W. Sun, Z. H. Wang, X. Liang, F. Hua, K. Li, et al., TRIB3 supports breast cancer stemness by suppressing FOXO1 degradation and enhancing SOX2 transcription, *Nat. Commun.*, **10** (2019), 5720. https://doi.org/10.1038/s41467-019-13700-6
60. P. R. Dandawate, D. Subramaniam, R. A. Jensen, S. Anant, Targeting cancer stem cells and signaling pathways by phytochemicals: Novel approach for breast cancer therapy, *Semin. Cancer Biol.*, **40–41** (2016), 192–208. https://doi.org/10.1016/j.semcancer.2016.09.001
61. J. A. Clara, C. Monge, Y. Yang, N. Takebe, Targeting signalling pathways and the immune microenvironment of cancer stem cells—a clinical update, *Nat. Rev. Clin. Oncol.*, **17** (2020), 204–232. https://doi.org/10.1038/s41571-019-0293-2
62. H. Liu, W. Zhang, Y. Song, L. Deng, S. Zhou, HNet-DNN: Inferring new drug-disease associations with deep neural network based on heterogeneous network features, *J. Chem. Inf. Model.*, **60** (2020), 2367–2376. https://doi.org/10.1021/acs.jcim.9b01008
63. P. Ding, C. Shen, Z. Lai, C. Liang, G. Li, J. Luo, Incorporating multisource knowledge to predict drug synergy based on graph co-regularization, *J. Chem. Inf. Model.*, **60** (2020), 37–46. https://doi.org/10.1021/acs.jcim.9b00793
64. H. Iwata, R. Sawada, S. Mizutani, M. Kotera, Y. Yamanishi, Large-scale prediction of beneficial drug combinations using drug efficacy and target profiles, *J. Chem. Inf. Model.*, **55** (2015), 2705–2716. https://doi.org/10.1021/acs.jcim.5b00444
65. F. Cheng, I. A. Kovacs, A. L. Barabasi, Network-based prediction of drug combinations, *Nat. Commun.*, **10** (2019), 1197. https://doi.org/10.1038/s41467-019-09186-x
66. J. Yang, Z. Li, X. Fan, Y. Cheng, Drug-disease association and drug-repositioning predictions in complex diseases using causal inference-probabilistic matrix factorization, *J. Chem. Inf. Model.*, **54** (2014), 2562–2569. https://doi.org/10.1021/ci500340n
67. M. Ester, H. P. Kriegel, J. Sander, X. Xu, Density-based spatial clustering of applications with noise, in *Int. Conf. Knowledge Discovery and Data Mining*, 1996.
68. X. He, D. Cai, Y. Shao, H. Bao, J. Han, Laplacian regularized gaussian mixture model for data clustering, *IEEE Trans. Knowl. Data Eng.*, **23** (2010), 1406–1418. https://doi.org/10.1109/TKDE.2010.259
69. T. Zhang, R. Ramakrishnan, M. Livny, BIRCH: An efficient data clustering method for very large databases, *ACM Sigmod Rec.*, **25** (1996), 103–114.
70. S. Yue, P. Li, P. Hao, SVM classification: Its contents and challenges, *Appl. Math. A J. Chin. Univ.*, **18** (2003), 332–342.
71. C. Kwak, A. Clayton-Matthews, Multinomial logistic regression, *Nurs. Res.*, **51** (2002), 404–410.

Appendix

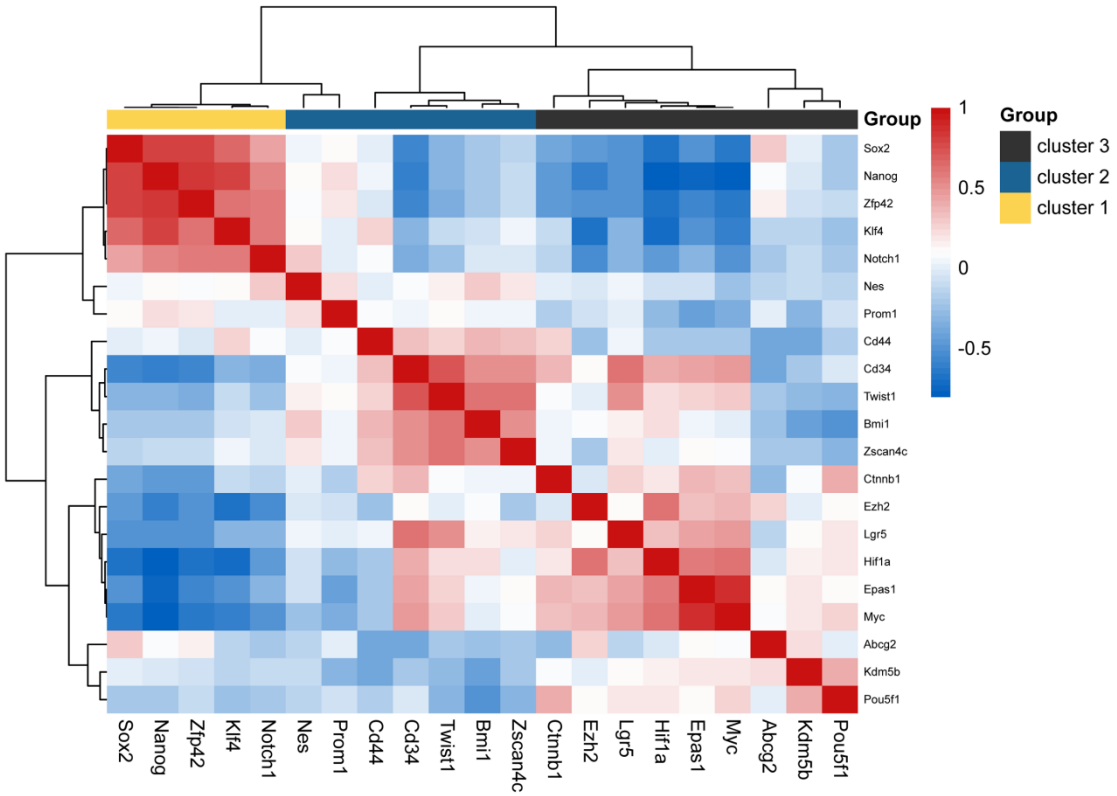

Figure A1. The heatmap of 21 stemness markers. The co-expression matrix is used for clustering 21 markers using the ward. D2 method. The clustering results are compared with K-means clustering, as shown by clusters 1–3. Nes and Prom1 are located at the boundary of cluster 2.