Metastasis Progression Through the Interplay Between the Immune System and Epithelial-mesenchymal-transition in Circulating Breast Tumor Cells

Samane Khoshbakht

Laboratory of Systems Biology and Bioinformatics (LBB), Department of Bioinformatics, Kish International Campus, University of Tehran, Kish Island, Iran https://orcid.org/0000-0003-3253-7577

Sadegh Azimzadeh Jamalkandi (azimzadeh@bmsu.ac.ir)

Chemical Injuries Research Center, Systems Biology and Poisonings Institute, Tehran, Iran
https://orcid.org/0000-0003-3403-3700

Ali Masudi-Nejad

Laboratory of Systems Biology and Bioinformatics (LBB), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran

Research article

Keywords: breast cancer, single / cluster CTC, metastasis, co-expression, epithelial-mesenchymal transition, immune response

Posted Date: September 21st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-74787/v1

Metastasis progression through the interplay between the immune system and Epithelial-Mesenchymal-Transition in circulating breast tumor cells

Samane Khoshbakht[1], Sadegh Azimzadeh Jamalkandi[2*], Ali Masudi-Nejad[1,3*]

1. Laboratory of Systems Biology and Bioinformatics (LBB), Department of Bioinformatics, Kish International Campus, University of Tehran, Kish Island, Iran, samane.khoshbakht@ut.ac.ir
2. Chemical Injuries Research Center, Systems Biology and Poisonings Institute, Tehran, Iran, amasoudin@ut.ac.ir
3. Laboratory of Systems Biology and Bioinformatics (LBB), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran, azimzadeh.jam.sadegh@gmail.com, azimzadeh@bmsu.ac.ir

Abstract

Background

Circulating tumor cells (CTCs) are the critical initiators of distant metastasis formation. In which, the reciprocal interplay among different metastatic pathways which promote survival of CTCs, is not well introduced, using network approaches. CTC cells include single and cluster cells, in which cluster cells revealed 23-50 fold more metastatic potentials.

Here, to investigate the unknown pathways of single/cluster CTCs, the co-expression network reconstructed, using WGCNA (Weighted Correlation Network Analysis) method. Having used the hierarchical clustering, we detected the Immune-response and EMT subnetworks. The metastatic potential of genes was assessed and validated through the support vector machine (SVM), neural network, and decision tree methods on two external datasets. To identify the active signaling pathways in CTCs, we reconstructed a casual network. The Log-Rank test and Kaplan-Meier curve were applied to detect prognostic gene signatures for metastasis-free survival. Finally, a predictive model was developed for metastasis risk of patients, using VIF-stepwise feature selection.

Results

Our results showed the crosstalk among EMT, the immune system, menstrual cycles, and the stemness pathway in CTCs. In which, fluctuation of menstrual cycles is a new detected pathway in breast cancer CTCs. The reciprocal association between immune responses and EMT was identified in single/cluster CTCs. The SVM model indicated a high metastatic potential of EMT subnetwork (accuracy, sensitivity, and specificity scores were 87%). The distant-metastasis-free-survival model was identified to predict patients' metastasis risks. (c-index=0.8). Finally, novel metastatic biomarkers including PTCRA, F13A1, ICAM2, and SNRPC were detected in breast cancer.

Conclusions

In conclusion, the reciprocal interplay among critical pathways in CTCs enhances their survival and metastatic potentials. Such findings may help to develop more precise predictive metastatic-risk models or detect novel biomarkers.

KEYWORDS

breast cancer, single / cluster CTC, metastasis, co-expression, epithelial-mesenchymal transition, immune response

1 | Background

Metastasis is the leading cause of death among women with breast cancer [1, 2]. Cancer progression and metastasis are of the critical and even controversial aspects of cancer studies [2]. There are two arguable metastasis models, including parallel progression and linear progression, which try to explain the dark side of the tumor developments [3]. In the linear model, the tumor initiates by genetic or epigenetic alternations, grows, spreads, and gains metastasis potentials to disseminate ectopic sites; Contrarily, in the parallel model, the metastasis ability initiates early-onset and evolves by circulating tumor cells (CTCs), parallelly [3, 4]. CTCs, which negatively relate to the high rise of mortality rates in cancer, are rare disseminated tumor cells in the peripheral blood of patients [5]. Of note, they appear even in the early stages and are prominent and leading components in metastasis [2, 6]. Therefore, the detection of CTCs in metastatic and non-metastatic breast cancer patients implies their leading role in cancer progression [7]; Moreover, their physical characteristics as single CTC or CTC clusters play a crucial role in metastasis propensity [5]. They borrow the morphologic features of their primary tumors and gain new features to survive in blood [8]. CTC clusters, which consist of 2-50 cancer cells, can transit through the circulation of patients and increase the potential of metastasis to 23- to 50-fold [5]. They overcome many hurdles to colonize distant

organs including intravasation into circulation, evading immune bulwarks, extravasation to distant sites, and eventually replacing the microenvironment of host tissue [9, 10].

Of note, the signaling pathways or intrinsic molecular characteristics of the single/cluster CTCs are not well recognized. Therefore, fully realize the CTCs' cellular features, using network approaches, will guide us to unknown metastasis concepts and more precise therapeutic decisions; In which, the reversal phenotypic of Epithelial-Mesenchymal-Transition (EMT) or immune system are two prominent components in cancer progression [1, 11, 12]. EMT mechanism, which helps cancer cells lose their cell adhesion and gain mesenchymal phenotype, accelerates metastasis through immunosuppression in primary tumors [12-14]. Accordingly, assessment of the role of EMT and immune responses in CTCs as well as intermediate pathways is essential in cancer biology.

In this study, we implemented the co-expression network reconstruction for CTCs isolated from advanced patients' blood, using WGCNA method. We extracted metastasis relevant subnetworks that enriched in the immune system and EMT pathways. The preservation of subnetworks was assessed in GSE51827. The metastasis-free survival analysis and Kaplan-Meier curve of genes were implemented in GSE7390 (external data). Concerning a better understanding of signaling pathways inside CTCs, we also extracted a signaling subnetwork from the KEGG database. To determine the metastasis potentials of identified subnetworks, we carried out the SVM, neural network, and decision tree classifications on GSE7390, and the selected model was validated in GSE9195. We also developed a metastasis-free-survival-risk model to predict patients' risk, using the VIF-stepwise feature selection and cox-PH model. Finally, an article review was implemented to detect novel metastatic biomarkers in breast cancer.

2 | Results

2.1 | Pre-Processing of CTCs and DEA

After pre-processing, 74 out of 77 cells were included in the downstream analyses. The excluded cells revealed low quality. The differential expression analysis was implemented after the normalization step (FDR<0.05) [15]. The adjusted p-values and logarithm of fold changes were reported in Additional file 2: Table S1 and Additional file 3: Table S2. As it is shown in Figure 1, the downregulation of immune subnetwork was detected in CTC clusters (light purple color in the heatmap); and the genes of the EMT subnetworks implied upregulation (dark purple color in the heatmap) in CTC clusters. Moreover, the single and cluster CTCs grouped well in both subnetworks (Figure 1a-b). The immune-related subnetwork represented a stronger expression difference between clustered and single cells.

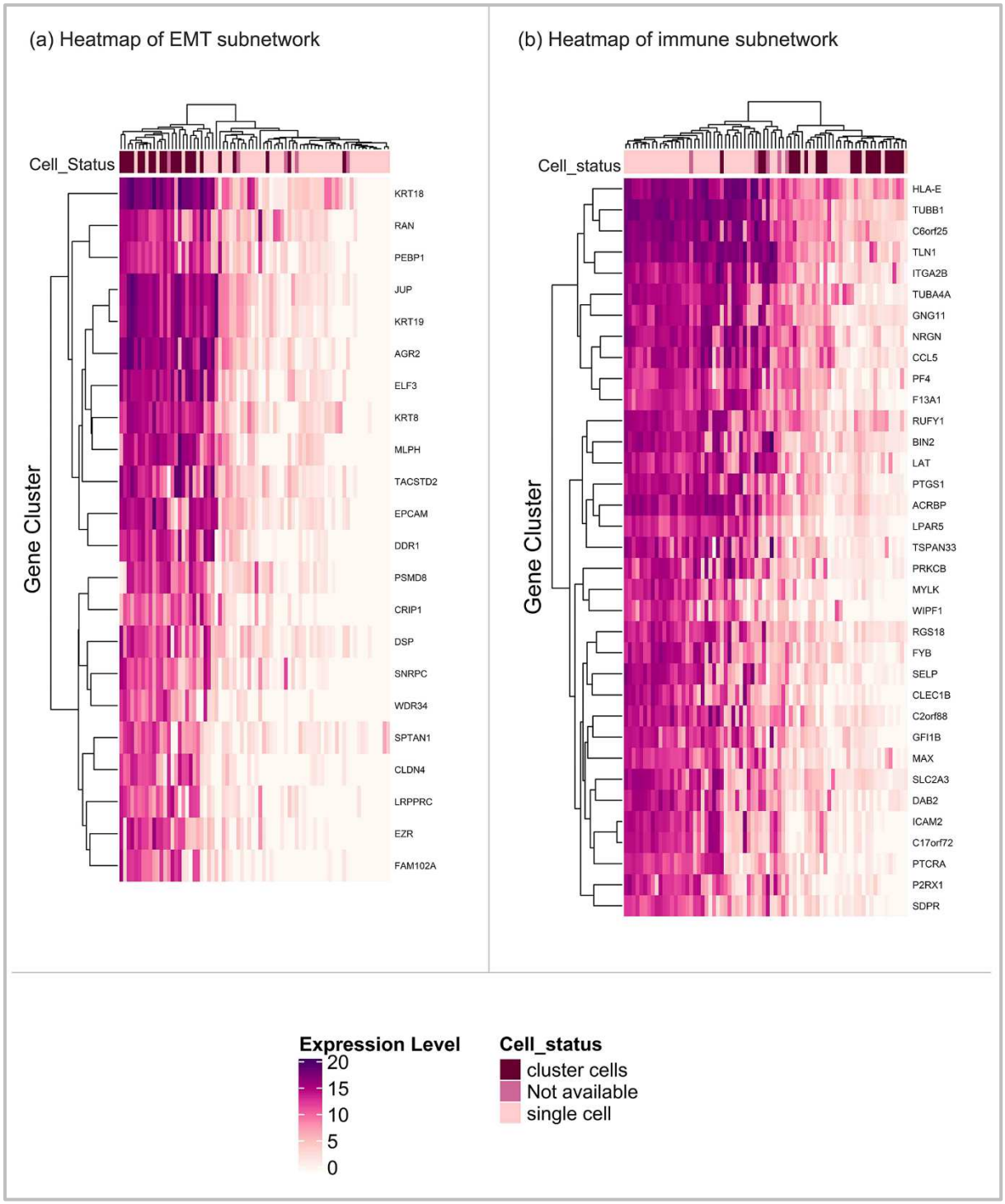

Figure 1 CTC cluster vs. single CTCs gene expression change

(a) Gene expression changes between CTC clusters and single CTCs for EMT-related subnetwork

(b) Gene expression changes between CTC clusters and single CTCs for Immune-related subnetwork

2.2 | Metastasis Associated Subnetworks

Metastasis-associated subnetworks were determined, using co-expression analysis and hierarchical clustering [16]. We detected 16 subnetworks. The first principle component (in PCA analysis) of subnetworks and the trait (cluster CTCs vs. single CTCs) association was assessed by correlation analysis. The two top significant subnetworks, which had the highest correlation with the trait, nominated for enrichment analysis ($|\text{midnightblue correlation}|=0.57$, $|\text{turquoise correlation}|= 0.51$) (Additional file 1: Figure S1). The sizes of midnightblue and turquoise subnetworks were 35 and 22 genes, and they were enriched for immune responses and EMT pathways, respectively (Figure 2).

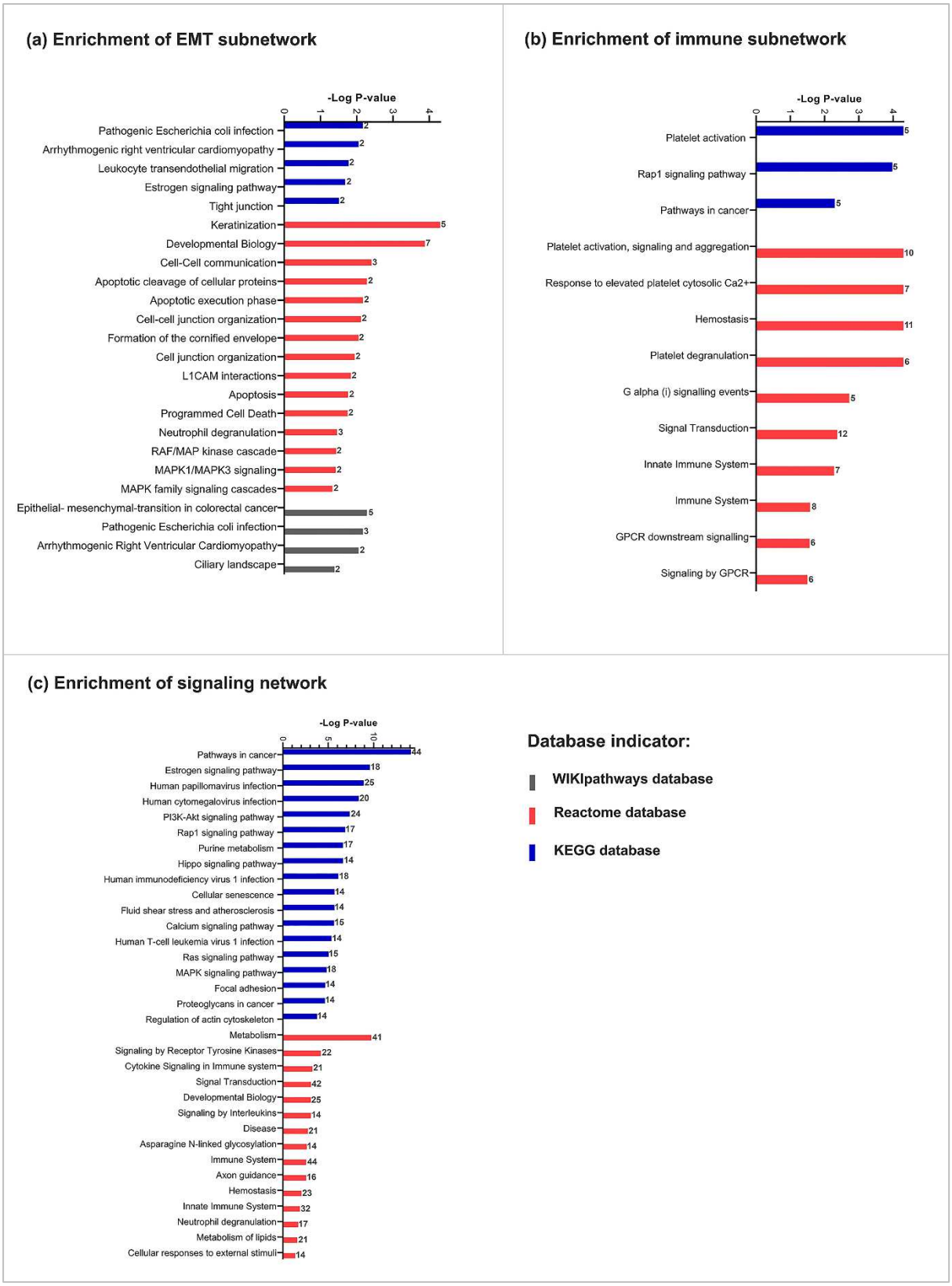

Figure 2 Gene set enrichment analysis.
The numbers for each bar indicate the number of significant genes.

(a) Significant pathways of EMT subnetwork(size=22) (q-value<0.05)
(b) Significant pathways of Immune subnetwork (size=35) (q-value<0.05)
(c) Significant pathways of signaling network of CTCs (q-value<0.05)

To have a biological concept for subnetworks, we addressed the midnightblue and the turquoise subnetworks, the Immune and EMT subnetworks, respectively, The EMT subnetwork included cancer-related pathways such as 'cell-cell communication', 'tight junction', 'keratinization', and 'estrogen signaling pathway'. The Immune subnetwork included pathways such as 'platelet activation', 'immune system', and 'innate immune system'. After having reviewed the literature, we detected the novel metastatic biomarkers in breast cancer. The immune-related novel metastatic biomarkers in breast cancer were PTCRA, F13A1, LAT, GNG11, ICAM2, NRGN, P2RX1, CLEC1B, BIN2, LPAR5, CCL5, SELP, RUFY1, C6orf25, TUBB1, GFI1B, C2orf88, ACRBP, and C17orf72. Module membership and gene significance of Immune subnetwork were reported in Additional file 2: Table S1.

The EMT related genes were LRPPRC, AGR2, CLDN4, CRIP1, DSP, ELF3, JUP, KRT8, KRT18, KRT19, FAM102A, TACSTD2, EPCAM, PEBP1, PSMD8, RAN, SNRPC, SPTAN1, EZR, DDR1, MLPH, and WDR34. In which, gene SNRPC, upregulated in CTC clusters, is a metastatic novel biomarker in breast cancer. Module membership and gene significance for EMT subnetwork were summarized in Additional file 3: Table S2.

The preservation of all subnetworks was assessed in the external dataset (GSE51827). The two combined statistics $Z_{summary}$ and $Median_{rank}$ were calculated to assess subnetworks preservation in the second dataset (Immune subnetwork: $Z_{summary} = 14$, $Median_{rank} = 9$ and EMT subnetwork: $Z_{summary} = 31, Median_{rank} = 6$). The $Z_{summary}$ values for both subnetworks were more than 10, and $Median_{rank}$ values were low; Subsequently, these statistics indicated the Immune and EMT subnetwork preservations in the second dataset. The statistics for all subnetworks reported in Additional file 4: Table S3).

2.3 | Directed Network Reconstruction

The signaling crosstalk between two selected subnetworks, Immune and EMT, was investigated by mapping them to KEGG and extracting induced subnetwork. A directed subnetwork of size 255 genes was extracted and illustrated in Figure 3. The network density was 0.5 and it included 40 components, in which, PLCG1 showed the highest betweenness value; and MYC, MYLK, and MRAS showed the highest closeness in the subnetwork.

We could detect 12 gene categories based on biological processes, including 'Hormonal regulation', 'Immune responses', 'Ion metabolism', 'Nucleobase metabolism', 'Oxidative responses', 'Protein localization', 'Protein topology response', 'STAT singling pathway', 'Vitamin metabolism', 'Cell differentiation', 'Circulation in blood regulation', and 'Energy metabolism' (Figure 3). These categories were illustrated by colors on the network nodes, and the genes with no category remained grey. The nodes with multiple colors indicated different biological processes. The node size was illustrated by node degrees (Figure 3). PLCG1 and ENTPD8 participated in 'Energy Metabolism' and 'Nucleobase Metabolism', were two hub nodes in our detected directed network. The ClueGO results were reported in Additional file 5: Table S4.

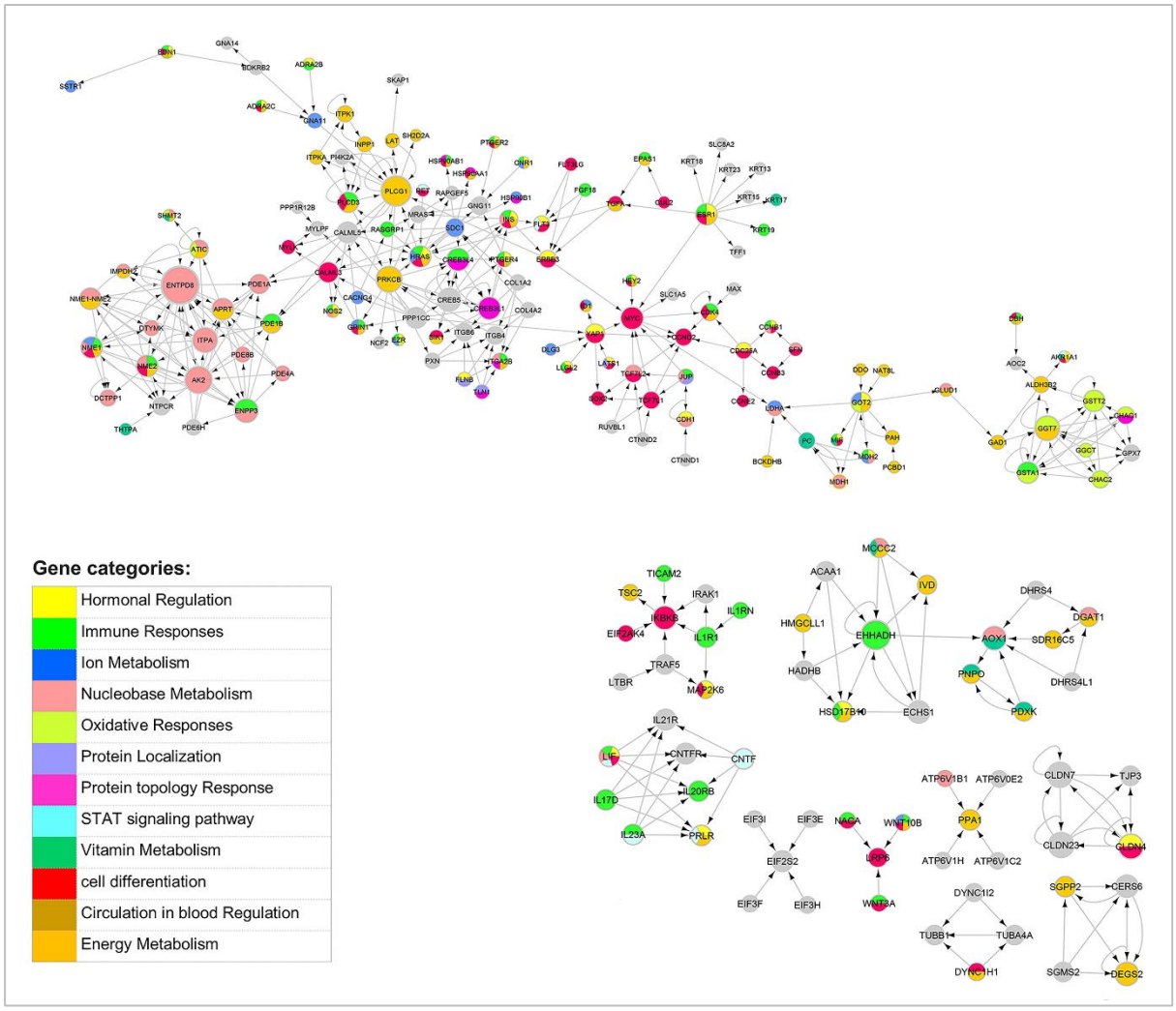

Figure 3 Signaling network of CTCs

The node size indicates the node degree. The direction among genes is based on KEGG directions.

2.4 | Distant Metastasis Classification Model

The distant metastasis potential of two nominated subnetworks was assessed using SVM, neural network, and decision tree classification methods in GSE7390. The accuracy, sensitivity, and specificity scores of the SVM model for EMT subnetwork were 79%, 78%, and 21%, respectively. The neural network accuracy, sensitivity, and specificity scores were 18%, 18%, and 80%, respectively. Eventually, the decision tree accuracy, sensitivity, and specificity scores were 60%, 60%, and 30%, respectively. These results refer to a full model (all genes in

the subnetwork included). Comparing three models, the SVM model was the strongest method in classifying metastatic and non-metastatic patients, but the specificity score was too low.

The SVM accuracy, sensitivity, and specificity scores for immune-related subnetwork were 78%, 78%, and 78%, respectively. The neural network accuracy, sensitivity, and specificity scores were 85%, 85%, and 14%, respectively. Eventually, the decision tree accuracy, sensitivity, and specificity scores were 71%, 71%, and 36%, respectively. These results refer to a full model (all genes of subnetwork included). The specificity of the neural network and decision tree methods was low compare to the SVM model. Due to the results, the SVM model was the most powerful method in classifying metastatic and non-metastatic patients for the immune-related subnetwork. The SVM model accuracy, sensitivity, and specificity for the Immune subnetwork were superior to EMT subnetwork.

The feature selection algorithms, for the SVM model, were implemented for both subnetworks. The WCC introduced 13 and 15 genes, and GA introduced 12, 17 genes for EMT- and immune-related subnetworks, respectively. The WCC introduced HLA-E, MYLK, WIPF1, TLN1, F13A1, NRGN, ICAM2, PTGS1, SELP, PF4, ITGA2B, GFI1B, TUBB1, PTCRA, RUFY1, BIN2, and CLEC1B and the GA introduced CCL5, MYLK, WIPF1, TLN1, NRGN, GNG11, PTGS1, SELP, ITGA2B, MAX, GFI1B, P2RX1, PTCRA, RUFY1, and BIN2 for the immune subnetwork.

The SVM model, full model, for immune subnetwork validated in GSE9195. The accuracy, sensitivity, and specificity were 0.868; surprisingly, the validation scores were superior to GSE7390. The results confirmed that the immune-related genes detected in this study can classify metastatic and non-metastatic samples more precisely compared to the neural network and decision tree models, using two data sets. We implemented the classification methods to assess the metastasis potential of two nominated CTC-related subnetworks.

198 **2.5 | Distant Metastasis-Free-Survival and Overall Survival Analyses**

199 The association between gene expression and distant metastasis-free survival /overall survival
200 was implemented to detect metastatic potential genes in two selected subnetworks. Overall
201 survival and distant metastasis-free survival of JUP, KRT18, and KRT19 were significant (
202 Log-rank test p-value<0.05; the exact p-values were reported in Figures) (Figure 4a-f). These
203 three genes belonged to EMT subnetwork. The upregulation of JUP, KRT18, and KRT19 was
204 associated with more metastases; Therefore, the lower overall survival of patients (Figure 4a-
205 f). Moreover, JUP, KRT18, and KRT19 were upregulated in CTC clusters (Figure 1a). The
206 lower distant metastasis-free survival and lower overall survival curves confirmed the
207 importance of selected genes in metastasis. Therefore, they are important gene signatures in
208 CTCs.

209 We fitted a metastasis free-survival Cox-PH regression model for EMT and Immune
210 subnetworks to assess patients'metastasis risks through a predictive model. The Immune Cox-
211 PH model included RUFY1 and P2RX1 variables (Likelihood ratio test p-value = 0.0295); and
212 the EMT Cox-PH model included of RAN, PEBP1, KRT8, DSP, DDR1, and CLDN4 variables
213 (Likelihood ratio test p-value = 0.0001016). The coefficients of variables and p-values were
214 reported in Additional file 6: Table S5 and Additional file 7: Table S6. All the significant genes
215 in the model had VIF<2 to avoid multicollinearity problems, using the VIF-feature selection
216 method (Additional file 8: Table S7 and Additional file 9: Table S8). The proportional hazard
217 assumptions for two model variables were assessed by the Schoenfeld residuals (Additional
218 File 10: Figure S2 and Additional File 11: S3). The predictive Cox-PH models for distant
219 metastasis-free survival for two subnetworks were illustrated in Figure 4g,h. The concordance
220 index, as a performance evaluation measure, for EMT and Immune predictive Cox-PH models
221 were 0.7 and 0.6, respectively. Therefore, the EMT model is more powerful in discriminating

patients into low, medium, and high metastasis risk groups comparing to the Immune model (higher concordance index indicates more power in discrimination).

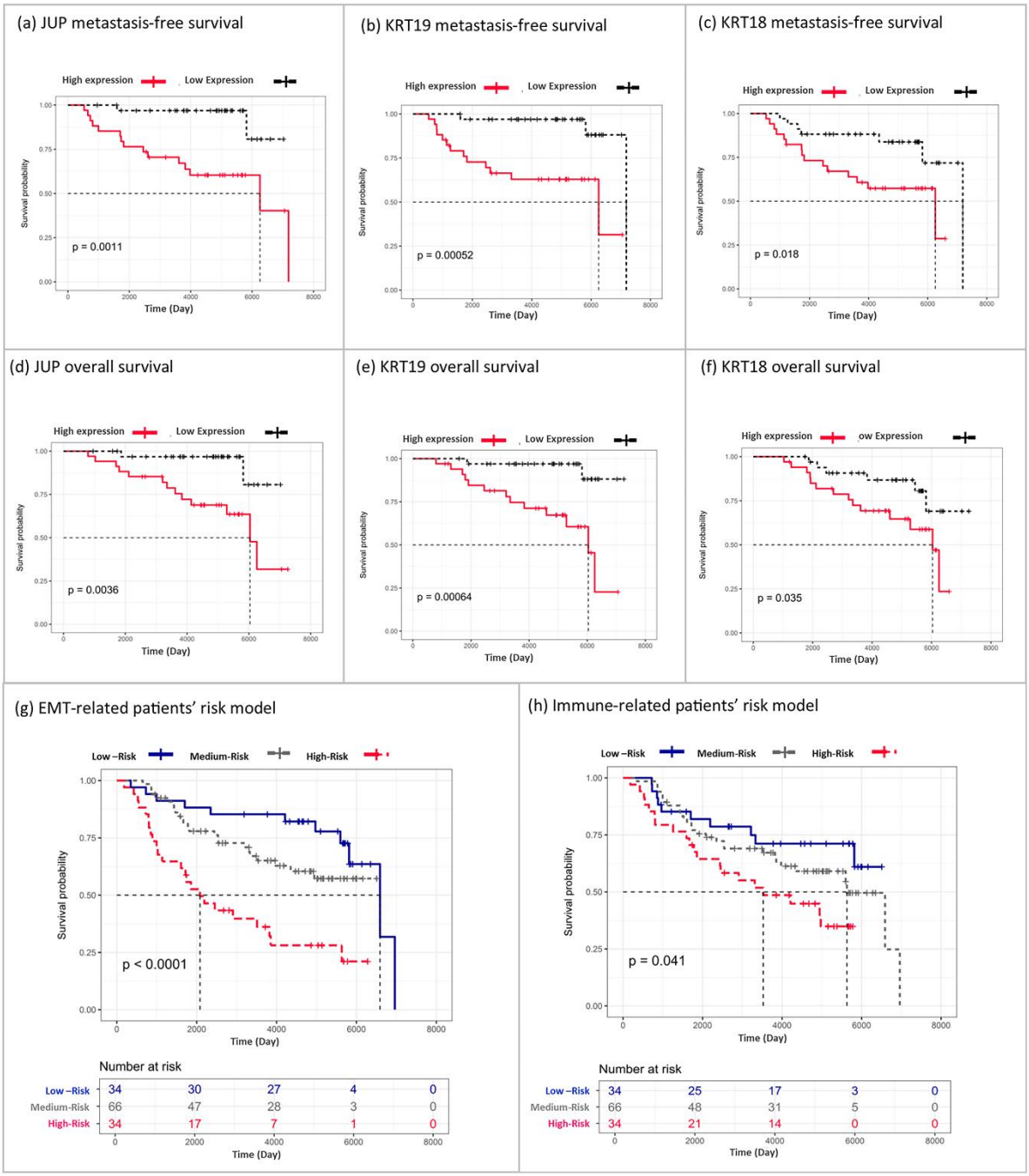

Figure 4 Metastasis free survival and overall survival.
p indicate p-value of Log-Rnk test in a,b,c,d,e,f,g, and h section.
(g) a predictive metastasis risk model for EMT subnetwork. High risk indicates upper-quartile of gene expression; low-risk indicates lower-quartile of gene expression.
(h) a predictive metastasis risk model for the immune subnetwork. High risk indicates upper-quartile of gene expression; low-risk indicates lower-quartile of gene expression.

3 | Discussion

Whereas multiple studies on circulating tumor cells (CTCs) as single CTCs or metastatic microemboli (CTC clusters) have been conducted, the molecular mechanisms of such rare cells are insufficiently characterized. CTCs bear several undiscovered metastatic potentials to overcome many restrictions, including extravasation of the primary tumor microenvironment, survival in the bloodstream, and successfully colonize secondary organs; Therefore, a better understanding of the biological mechanisms of different types of CTCs, single/cluster, is essential.

This study was aimed to explore metastasis-related mechanisms within CTCs. We have implemented the co-expression analysis to detect subnetworks discriminating single/cluster CTCs (Figure 1). Two of subnetworks indicated a significant correlation to the trait (single/cluster status of CTCs). Our detected subnetworks illustrated immune- and EMT-related pathways (Figure 2a-b). Due to previous studies, the immune-associated mechanisms and EMT pathways are of two major arms in breast cancer progression and metastasis, but investigating them in CTCs is not thoroughly studied [11]. To prepare cancer cells for intravasation in early stages, the keratin family, claudins, and cadherins must be downregulated through the EMT process in primary tumors; still, due to the surviving urgency of CTCs in the bloodstream and avoiding anoikis, a small number of tumor cells must be attached and break off from the primary site [17, 18]. Therefore, the keratins, claudins, and cadherins should be upregulated in CTC clusters to survive shear forces in blood circulation. The plakoglobin (JUP), KRT8, KRT18, KRT19, CLDN4, and EPCAM, which are EMT-related biomarkers, are such essential genes that their role in breast cancer metastasis demonstrated in several studies [5, 18, 19]. The KRT8, KRT18, KRT19 are a group of cytoskeleton genes within the cellular

cytoplasm called keratins. Although they are extensively used as diagnostic tumor markers, several studies have demonstrated their involvement in cancer cell invasion and metastasis, as well as in treatment responsiveness [19, 20]. Keratins are the intermediate filament-forming proteins of epithelial cells that organize the internal three-dimensional cellular structure; In fact, they act in cell shape maintenance for bearing tensions. [20]. Moreover, the plakoglobin, the upregulation of which in breast cancers CTC clusters in comparison to single cells was demonstrated in Aceto, Nicola, et al. study, is one of the cell junction genes that hold tumor cells together to leave primary tumor as CTC clusters [5]. In our study, the overexpression of JUP, KRT18, and KRT19, as well as the overall survival and metastasis-free survival were significant (Figure 4a-f). Therefore, they may play an essential role in the integrity of CTC clusters in the bloodstream shear forces. Meanwhile, EPCAM and cytokeratins have been reported as detection markers in the enrichment of CTCs [21]. Of note, few markers guide scientists to detect metastatic patients in the clinic therefore the rest of the genes detected in EMT subnetwork were not investigated in CTC studies; Consequently, they could be new targets in experimental studies for CTCs.

Several types of immune cells ambiguously reveal anti- and pro-tumor behaviors [22]. The immunosuppressive microenvironment of tumors protects the primary tumor cells. Nevertheless, while tumor cells extravasate and enter circulation, they lose their tumor protection; Therefore, they must adapt themselves to escape immune surveillance [1, 22]. The interplay between immune cells and cytokeratins may contribute to evasion of CTCs from immune surveillance. The cytotoxic T lymphocytes (CTLs) were recruited by recognizing tumor antigens presented by major histocompatibility class I (MHCI) [23, 24]. The under-expression of MHCI in tumor cell surface guides them to hide from CTLs and thereby survive in circulation. Moreover, the overexpression of cytokeratins such as KRT8, and together with

heterodimeric partners KRT18 and KRT19 inhibit MHCI interactions with CTLs [22, 23]. All these findings, overexpression of KRT8, KRT18, KRT19, and under-expression of HLA-E, are consistent with our results which highlight the CTC cluster potential to evade the immune system; consequently, longer survival (Figure 1a-b).

Of note, several studies supported the association between EMT and immune cell escape of cancer cells [25, 26]. Moreover, a plethora of genes and signals support stemness pathways such as Wnt, TGF-β, and NOTCH in CTCs [10]. Downregulation of DAB2, which is a putative tumor suppressor and involves in TGF-β pathway and promotes EMT, was reported in breast cancer tumors [27, 28]. Therefore, the under-expression of DAB2 in CTC clusters might be related to the stemness phenotype which helps CTC clusters to escape the immune system. In our study, DAB2 the expression of which was downregulated in CTC clusters, the logarithm of fold change = -5.7, was detected in our immune subnetwork (Figure 1b). Therefore, our findings may indicate the survival potential of CTC clusters in circulation, which were consistent with previous studies on cancer biology. Whereas the CTC clusters have higher metastatic potential due to less frequency in metastatic patients, but the single cells contribute metastasis either. Hereof, several studies such as Szczerba, Barbara Maria, et al. indicated more single-cell, about 88.0%, detected in metastatic patients [29]. Therefore, either single CTCs or CTC clusters have metastasis potential with different molecular mechanisms.

Not only the immune system and EMT but also the intermediate pathways are important in the progression of CTCs. The crosstalk between the signaling pathways of immune response and hormonal regulations, such as the fluctuation of menstrual cycles, were investigated by Atashgaran, Vahid, et al. in breast cancer [30]. Furthermore, they demonstrated the dis-regulation of hormonal factors affecting genome instability and the decrease of immune surveillance in breast cancer [31]. We know EMT is a complex process through which tumor

303 cells facilitate their dissemination and acquire stemness characteristics [14, 32]. Not only
304 signaling pathways of stemness but also in the stimulation of self-renewing pathways in tumor
305 cells are essential in embryogenesis [14, 33]. Such several metastatic prone pathways and the
306 interplay among all of them activated in circulating tumor cells (Figure 2c and Figure 3). The
307 CTCs are tumor cells that reflect characteristics of primary tumors and likewise more additional
308 metastatic propensity to survive in blood-stream and extravasate secondary site. These pro-
309 metastatic cells need to recruit different signaling pathways, the interrelationship of which
310 leads to creating multi-role cells that reflect great metastatic and survival potentials. As a result,
311 characterizing multiple aspects of CTC, involved in cancer progression is essential; Moreover,
312 useful in finding novel biomarkers or patients' treatment strategies.

313 # 4 | Conclusions

314 In summary, although CTCs, which are cancer-related biomarkers, are applied in the clinic, the
315 molecular mechanisms were not investigated well. The unknown crosstalks among multiple
316 pathways including EMT and immune responses improve the survival of CTCs in the patients'
317 blood. Therefore, they may contribute to therapeutic resistance and metastasis. Computational
318 investigations on CTCs suggest novel metastatic biomarkers which could be new targets for
319 experimental studies or therapeutic aims.

320 # 5 | Methods

321 ## 5.1 | Data sets and Metadata Information:

322 The single-cell RNA-seq data related to advanced ER+ breast cancer patients were
323 downloaded from the NCBI data repository (GSE86978). The data consist of 77 cells which
324 47 of them were CTC clusters, 22 cells were single CTCs, and the rest of the cells were not

categorized. The GSE51827, which consists of 29 cells (15 single CTCs and 14 CTC clusters), was used for subnetwork preservation analysis. The gene expression of GSE7390, which consists of 198 untreated breast cancer patients, was used for assessment of metastasis potential of subnetworks. The GSE9195, which consists of 77 breast cancer patients, was used for validation of the classification model, overall survival, and metastasis-free survival analyses.

5.2 | Pre-Processing, Normalization, and Differential Analysis

To have more precise downstream analysis and remove non-biological variations, we implemented several pre-process steps on genes and also cells. At the first step, we filtered out low abundance genes, then the small count cells omitted subsequently. In the last step, the expression data were normalized to reduce technical effects, using the scatter package in R [34]. The differential expression analysis (DEA), which compares clustered cells' expression to the single cells' expression, was implemented using the limma package in R (FDR<0.05) [15].

5.3 | Co-Expression Network (CEN) Reconstruction and Subnetwork Extraction

The co-expression network reconstructed using a weighted correlation network analysis (WGCNA) method [16]. The pairwise relation among genes was estimated using the Pearson correlation among genes. Concerning to have more connected subnetworks, we carried out the topological overlap matrix (TOM) and connectivity gene filtering (connectivity values less 0.1 were omitted) [16]. Higher connectivity values indicate more considerable co-expressed subnetworks [16, 35]. Eventually, we used hierarchical clustering to extract subnetworks. The trait used in this study was the cluster and single status of the cells captured in blood. The subnetworks, which have strong correlations between their first principle component and the biological trait, were selected as trait related (metastasis) subnetworks. The gene significance and module membership were used to filter essential genes in selected subnetworks. The gene

significance is the correlation between gene expression and the metastasis trait. The module membership is the correlation between gene expression and module representative (first principle component in the principal component analysis (PCA)). The preservation of subnetworks were assessed in the external dataset (GSE51827)

5.4 | Signaling Network Reconstruction

We downloaded all homo sapiens pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database resource and merged them [36]. Furthermore, the KEGG ids were annotated to gene symbols. At the last step, to have casual relations among genes, we extracted a directed induced subnetwork from the KEGG database, using two detected significant subnetworks ($|\text{correlation}|>0.5$ was considered significant) [37]. The genes were categorized based on biological process (BP of gene ontology) terms, using the ClueGO plug-in in Cytoscape [38, 39]. The network visualization was implemented by the Cytoscape and the Gephi software [39, 40].

5.5 | Gene Set Enrichment Analysis and Subnetwork Preservation Analysis

The significant trait-related, metastatic, subnetworks ($|\text{correlation}|>0.5$) were enriched, using ConsensusPathDB webserver (q-value<0.05) [41]. The GSE51827, which has CTC gene expression, downloaded from NCBI to implement preservation analysis of subnetworks in the external dataset in R [16, 35, 42]. The scatter package, which is suitable for single-cell RNA-seq data, was used to preprocess, normalized, and merged expression data [34].

The combined statistics for preservation assessment, which includes $Z_{summary}$ and $Median_{rank}$, were used to check the reproducibility of subnetworks [35]. The $Z_{summary}$, which shows the interaction pattern among genes in subnetworks, evaluate connectivity and density in the external dataset. $Z_{summary} < 2$ indicates not preserved subnetworks. If $2 < Z_{summary} < 10$, the subnetwork is semi preserved, and if $Z_{summary} > 10$, the subnetwork is

preserved. Moreover, a higher $Median_{rank}$ indicates more preservation of subnetworks in the external dataset [35]. The $Z_{summary}$ and $Median_{rank}$ were assessed for our subnetworks.

5.6 | Assessment of Distant Metastasis Potential of Subnetworks

To evaluate the importance of selected subnetworks and the metastasis potential of genes, we implemented the classification algorithms on two individual datasets (GSE7390 and GSE9195). The GSE7390 (Affymetrix platform, HG-U133A) downloaded using the GEOquerry package in R [42]. The ER+ patients (134 patients out of 198 ones) filtered, and the expression data normalized using the RMA method [43]. In this section, we learned three classifiers, including support vector machine (SVM), artificial neural network (ANN), and decision tree on metastatic and non-metastatic patients [44]. The classification algorithms were run with and without feature selection algorithms, including the genetic algorithm (GA) and the world competitive contest (WCC) algorithm. The SVM was implemented with 5-fold cross-validation, linear kernel, and 80 percent of cells as the training set. Finally, the accuracy, precision, and specificity were checked to select a better classifier for metastasis prediction; Furthermore, to identify the most metastatic-related subnetwork. To assess the reproducibility of our results, the selected model was validated in another dataset (GSE9195).

5.7 | Distant Metastasis-Free Survival Analysis

The Kaplan-Meier curve, distant metastasis-free survival, and overall survival analyses were implemented using GSE7390 in R [45]. The patients were stratified due to quartiles. The expression values lower than the first quartile were labeled low expression, and expression values higher than the third quartile were labeled high expression. The stepwise Cox proportional hazard ratio (Cox-PH) was implemented for selected subnetworks [45]. The concordance index was calculated to evaluate model performance. The Variance Inflation Factor (VIF) lower than two was used as the variable selection criteria. The first and third

quartiles of the predicted hazard ratio were used for stratifying patients into three groups, including low-risk, medium-risk, and high-risk groups.

List of abbreviations

CTC: circulating tumor cell

EMT: epithelial-mesenchymal-transition

SVM: Support vector machine

PTCRA: Pre T Cell Antigen Receptor Alpha

F13A1: Coagulation Factor XIII A Chain

LAT: Linker For Activation Of T Cells

ICAM2: Intercellular Adhesion Molecule 2

OS: overall survival

PCA: principal component analysis

WCC: world competitive contest

GA: genetic algorithm

ANN: artificial neural network

Cox-PH: Cox proportional hazard ratio

DEG: differentially expressed gene

VIF: Variance Inflation Factor

FDR: false discovery rate

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The public data were used in this article including GSE51827, GSE9195, GSE7390, and GSE86978.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Authors' contributions

Conceptualization, Khoshbakht and Azimzadeh; Methodology, Khoshbakht and Azimzadeh; Validation: Khoshbakht; Formal Analysis, Khoshbakht; writing—Original draft preparation, Khoshbakht; writing, Review, and editing, Khoshbakht, Azimzadeh and Masudi-Nejad; Visualization, Khoshbakht; Supervision, Masudi-Nejad and Azimzadeh. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

Not applicable.

References

1. Leone, K., C. Poggiana, and R.J.D. Zamarchi, *The interplay between circulating tumor cells and the immune system: from immune escape to cancer immunotherapy.* 2018. **8**(3): p. 59.
2. Weigelt, B., J.L. Peterse, and L.J.J.N.r.c. Van't Veer, *Breast cancer metastasis: markers and models.* 2005. **5**(8): p. 591-602.
3. Klein, C.A.J.N.R.C., *Parallel progression of primary tumours and metastases.* 2009. **9**(4): p. 302-312.
4. Ghajar, C.M. and M.J.J.N. Bissell, *Metastasis: pathways of parallel progression.* 2016. **540**(7634): p. 528-529.
5. Aceto, N., et al., *Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis.* 2014. **158**(5): p. 1110-1122.
6. Lang, J.E., et al., *RNA-Seq of circulating tumor cells in stage II–III breast cancer.* 2018. **25**(8): p. 2261-2270.
7. Barneh, F., et al., *Valproic acid inhibits the protective effects of stromal cells against chemotherapy in breast cancer: Insights from proteomics and systems biology.* Journal of cellular biochemistry, 2018. **119**(11): p. 9270-9283.
8. Yang, C., et al., *Circulating tumor cells in precision oncology: clinical applications in liquid biopsy and 3D organoid model.* Cancer Cell International, 2019. **19**(1): p. 341.
9. Dasgupta, A., A.R. Lim, and C.M.J.M.o. Ghajar, *Circulating and disseminated tumor cells: harbingers or initiators of metastasis?* 2017. **11**(1): p. 40-61.
10. Massagué, J. and A.C.J.N. Obenauf, *Metastatic colonization by circulating tumour cells.* 2016. **529**(7586): p. 298-306.
11. Santisteban, M., et al., *Immune-induced epithelial to mesenchymal transition in vivo generates breast cancer stem cells.* 2009. **69**(7): p. 2887-2895.
12. Pastushenko, I. and C.J.T.i.c.b. Blanpain, *EMT transition states during tumor progression and metastasis.* 2019. **29**(3): p. 212-226.
13. Kudo-Saito, C., et al., *Cancer metastasis is accelerated through immunosuppression during Snail-induced EMT of cancer cells.* 2009. **15**(3): p. 195-206.
14. Barneh, F., et al., *Integrated use of bioinformatic resources reveals that co-targeting of histone deacetylases, IKBK and SRC inhibits epithelial-mesenchymal transition in cancer.* Briefings in bioinformatics, 2019. **20**(2): p. 717-731.
15. Smyth, G.K., *Limma: linear models for microarray data*, in *Bioinformatics and computational biology solutions using R and Bioconductor*. 2005, Springer. p. 397-420.
16. Langfelder, P. and S.J.B.b. Horvath, *WGCNA: an R package for weighted correlation network analysis.* 2008. **9**(1): p. 559.
17. Fabisiewicz, A. and E.J.M.O. Grzybowska, *CTC clusters in cancer progression and metastasis.* 2017. **34**(1): p. 12.
18. Joosse, S.A., et al., *Changes in keratin expression during metastatic progression of breast cancer: impact on the detection of circulating tumor cells.* 2012. **18**(4): p. 993-1003.
19. Tőkés, A.-M., et al., *Claudin-1,-3 and-4 proteins and mRNA expression in benign and malignant breast lesions: a research study.* 2005. **7**(2): p. R296.
20. Karantza, V.J.O., *Keratins in health and cancer: more than mere epithelial cell markers.* 2011. **30**(2): p. 127-138.

21. Deng, G., et al., *Enrichment with anti-cytokeratin alone or combined with anti-EpCAM antibodies significantly increases the sensitivity for circulating tumor cell detection in metastatic breast cancer patients.* 2008. **10**(4): p. R69.
22. Mohme, M., S. Riethdorf, and K.J.N.r.C.o. Pantel, *Circulating and disseminated tumour cells—mechanisms of immune surveillance and escape.* 2017. **14**(3): p. 155.
23. Wu, M.-S., et al., *Cytokeratin 8-MHC class I interactions: a potential novel immune escape phenotype by a lymph node metastatic carcinoma cell line.* 2013. **441**(3): p. 618-623.
24. Joosten, S.A., L.C. Sullivan, and T.H.J.J.o.i.r. Ottenhoff, *Characteristics of HLA-E restricted T-cell responses and their role in infectious diseases.* 2016. **2016**.
25. Jia, D., et al., *Quantifying cancer epithelial-mesenchymal plasticity and its association with stemness and immune response.* 2019. **8**(5): p. 725.
26. Terry, S., et al., *New insights into the role of EMT in tumor immune escape.* 2017. **11**(7): p. 824-846.
27. Bagadi, S.A.R., et al., *Frequent loss of Dab2 protein and infrequent promoter hypermethylation in breast cancer.* 2007. **104**(3): p. 277-286.
28. Martin, J., B.-S. Herbert, and B.J.B.j.o.c. Hocevar, *Disabled-2 downregulation promotes epithelial-to-mesenchymal transition.* 2010. **103**(11): p. 1716-1723.
29. Szczerba, B.M., et al., *Neutrophils escort circulating tumour cells to enable cell cycle progression.* Nature, 2019. **566**(7745): p. 553-557.
30. Atashgaran, V., et al., *Dissecting the biology of menstrual cycle-associated breast cancer risk.* 2016. **6**: p. 267.
31. Atashgaran, V., et al., *Dissecting the biology of menstrual cycle-associated breast cancer risk.* Frontiers in oncology, 2016. **6**: p. 267.
32. Takebe, N., R.Q. Warren, and S.P.J.B.c.r. Ivy, *Breast cancer growth and metastasis: interplay between cancer stem cells, embryonic signaling pathways and epithelial-to-mesenchymal transition.* 2011. **13**(3): p. 211.
33. Takebe, N., R.Q. Warren, and S.P. Ivy, *Breast cancer growth and metastasis: interplay between cancer stem cells, embryonic signaling pathways and epithelial-to-mesenchymal transition.* Breast cancer research, 2011. **13**(3): p. 211.
34. McCarthy, D.J., et al., *Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R.* 2017. **33**(8): p. 1179-1186.
35. Langfelder, P., et al., *Is my network module preserved and reproducible?* 2011. **7**(1).
36. Zhang, J.D. and S.J.B. Wiemann, *KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor.* 2009. **25**(11): p. 1470-1471.
37. Piran, M., et al., *Can we assume the gene expression profile as a proxy for signaling network activity?* 2020. **10**(6): p. 850.
38. Bindea, G., et al., *ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks.* 2009. **25**(8): p. 1091-1093.
39. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* 2003. **13**(11): p. 2498-2504.
40. Bastian, M., S. Heymann, and M. Jacomy. *Gephi: an open source software for exploring and manipulating networks*. in *Third international AAAI conference on weblogs and social media*. 2009.
41. Kamburov, A., et al., *ConsensusPathDB: toward a more complete picture of cell biology.* 2011. **39**(suppl_1): p. D712-D717.
42. Davis, S. and P.S.J.B. Meltzer, *GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor.* 2007. **23**(14): p. 1846-1847.

532 43. Gautier, L., et al., *affy—analysis of Affymetrix GeneChip data at the probe level.*
533 2004. **20**(3): p. 307-315.
534 44. Masoudi-Sobhanzadeh, Y., H. Motieghader, and A.J.B.b. Masoudi-Nejad,
535 *FeatureSelect: a software for feature selection based on machine learning*
536 *approaches.* 2019. **20**(1): p. 170.
537 45. Therneau, T., *A Package for Survival Analysis in S. version 2.38*. 2015.

538 Additional Files.

539 Additional file 1:

540 Figure s1. subnetwork-trait correlation p-values.

541 The metastasis-related subnetwork was selected, using subnetwork/trait correlation.

542 (|correlation |> 0.5)

543 Additional file 2:

544 Table S1. midnightblue (immune) subnetwork.

545 The list of genes, module membership, the logarithm of fold change, differential analysis
546 adj-p.value, and gene-significance statistics.

547 Additional file 3:

548 Table S2. turuise (EMT) subnetwork.

549 The list of genes, module membership, the logarithm of fold change, differential analysis
550 adj-p.value, and gene-significance statistics.

551 Additional file 4:

552 Table S3. The preservation statistics.

553 This table includes $Z_{summary}$ and $Median_{rank}$ statistics.

554 Additional file 5:

555 Table S4. Biological pathways in the directed network.

556 We categorized genes of the detected directed network, using ClueGO plugin in Cytoscape.
557 The Biological pathways and p-values were reported in Table S4.

Additional file 6:

Table S5. the EMT subnetwork cox-PH results.

The cox-PH analysis was implemented for the EMT subnetwork genes. The selected genes, coefficients, and p-values were reported in Table S5.

Additional file 7:

Table S6. the Immune subnetwork cox-PH results

The cox-PH analysis was implemented for the immune subnetwork genes. The selected genes, coefficients, and p-values were reported in Table S6.

Additional file 8:

Table S7. EMT VIF values.

To investigate multicollinearity in the cox-PH model, we calculated the Variance Inflation Factor (VIF). The VIF < 10 indicates no multicollinearity. The VIF of immune genes was reported in Table S7.

Additional file 9:

Table S8. immune VIF values.

To investigate multicollinearity in the cox-PH model, we calculated the Variance Inflation Factor (VIF). The VIF < 10 indicates no multicollinearity. The VIF of immune genes was reported in Table S8.

Additional file 10:

Figure S1. The Schoenfeld residuals for EMT genes.

The proportional hazard ratio investigated, using the Schoenfeld residuals. The residuals (red dots) must be between the curves.

Additional file 11:

Figure S2. The Schoenfeld residuals for immune genes.

582 The proportional hazard ratio investigated, using the Schoenfeld residuals. The residuals (red
583 dots) must be between the curves.
584
585
586
587
588
589

Figures

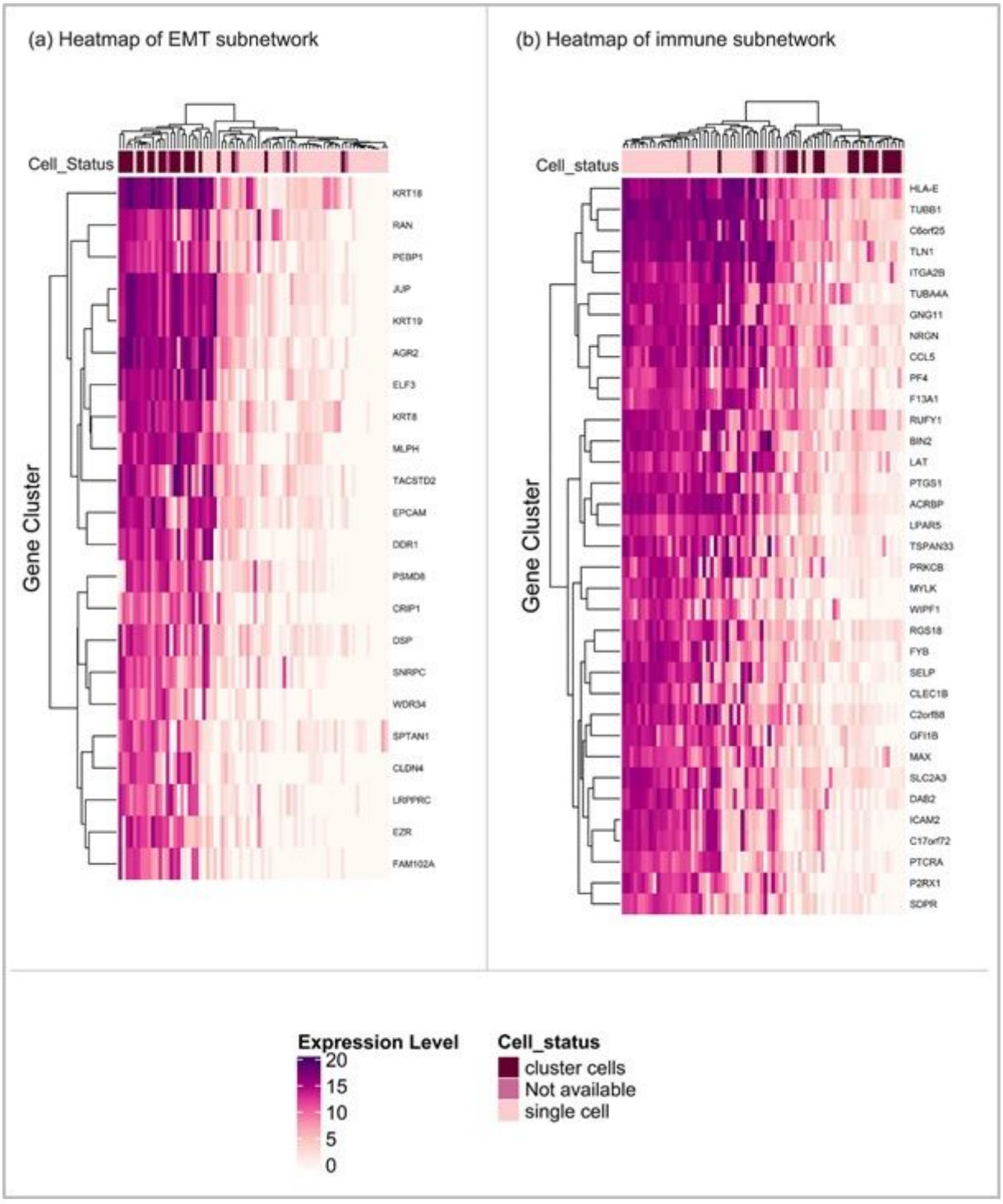

Figure 1

CTC cluster vs. single CTCs gene expression change (a) Gene expression changes between CTC clusters and single CTCs for EMT-related subnetwork (b) Gene expression changes between CTC clusters and single CTCs for Immune-related subnetwork

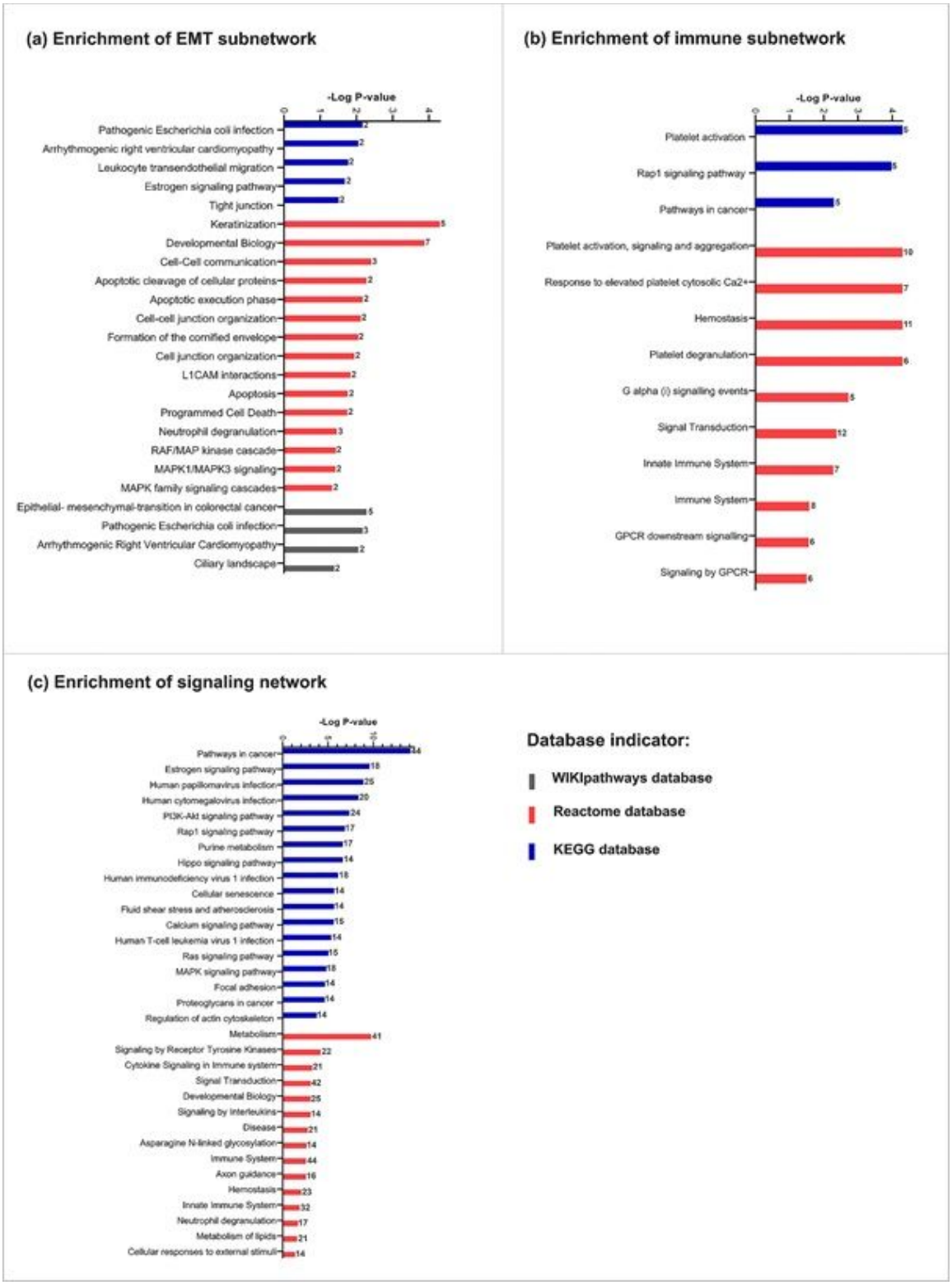

Figure 2

Gene set enrichment analysis. The numbers for each bar indicate the number of significant genes. (a) Significant pathways of EMT subnetwork(size=22) (q-value<0.05) (b) Significant pathways of Immune subnetwork (size=35) (q-value<0.05) (c) Significant pathways of signaling network of CTCs (q-value<0.05)

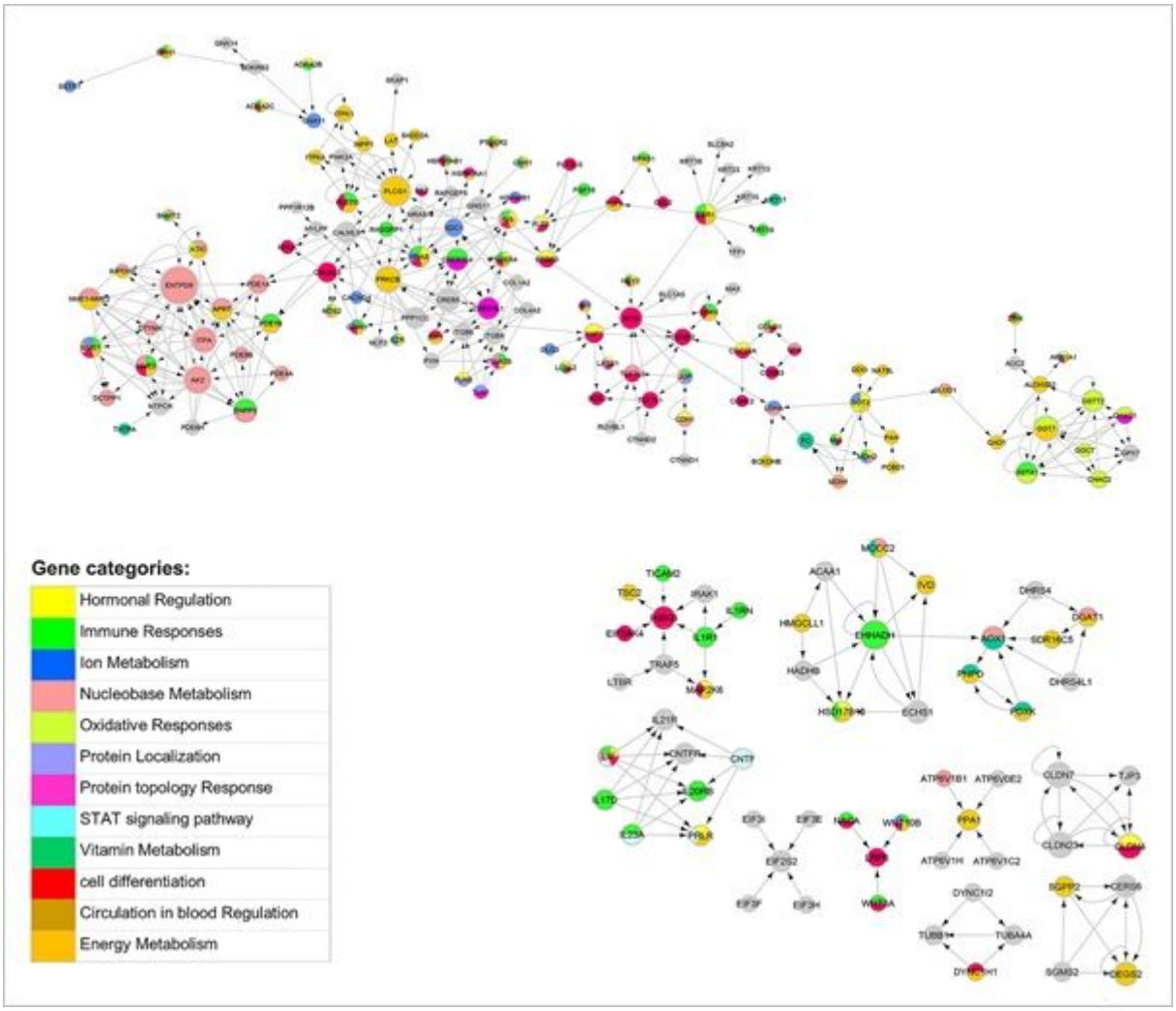

Figure 3

Signaling network of CTCs The node size indicates the node degree. The direction among genes is based on KEGG directions.

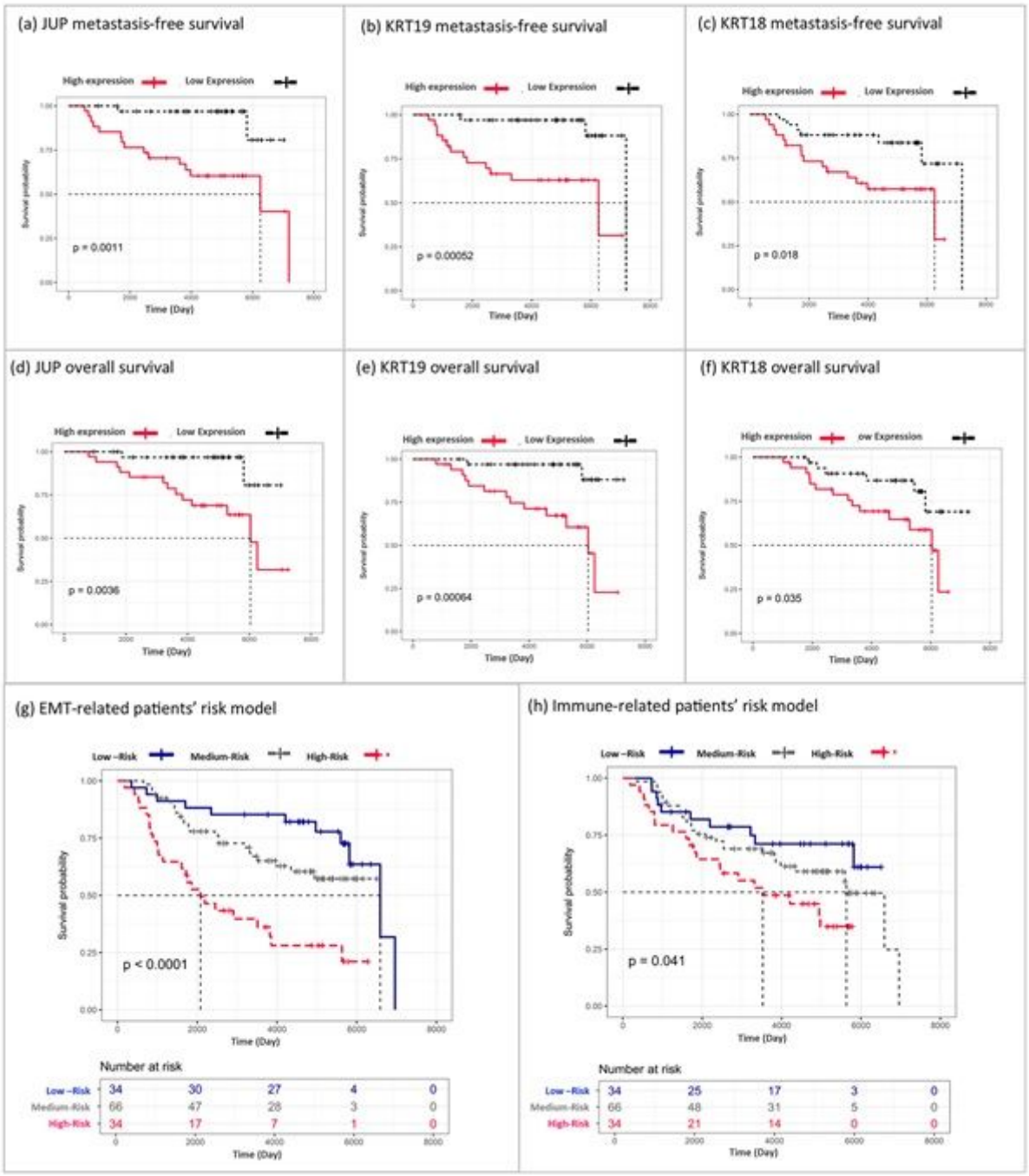

Figure 4

Metastasis free survival and overall survival. p indicate p-value of Log-Rnk test in a,b,c,d,e,f,g, and h section. (g) a predictive metastasis risk model for EMT subnetwork. High risk indicates upper-quartile of gene expression; low-risk indicates lower-quartile of gene expression. (h) a predictive metastasis risk model for the immune subnetwork. High risk indicates upper-quartile of gene expression; low-risk indicates lower-quartile of gene expression.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementary.docx
- TableS1.xlsx
- TableS2.xlsx
- TableS3.xlsx
- TableS4.xlsx
- TableS5.xlsx
- TableS6.xlsx
- TableS7.xlsx
- TableS8.xlsx
- FigureS1.tif
- FigureS2.tif
- FigureS3.tiff