INNSPUB

International Journal of Biosciences | IJB |
ISSN: 2220-6655 (Print); 2222-5234 (Online)
Website: https://www.innspub.net
Email contact: info@innspub.net
Vol. 27, Issue: 3, p. 145-157, 2025

RESEARCH PAPER

OPEN ACCESS

Epithelial cell adhesion molecule-centered, bioinformatics and machine learning-based meta-analysis for the identification of pan-cancer epithelial-mesenchymal markers for circulating tumor cells

Shubham Singh[1], **GR Brindha**[2], **Nagarajan Rajendra Prasad**[*1]

[1] *Department of Biochemistry and Biotechnology, Annamalai University, Chidambaram, Tamil Nadu, India*

[2] *School of Computing, SASTRA Deemed University, Thanjavur, Tamil Nadu, India*

Key words: Epithelial cell adhesion molecule, Machine learning, Meta-analysis, Epithelial-mesenchymal transition, Circulating tumor cell

DOI: https://dx.doi.org/10.12692/ijb/27.3.145-157 Published: September 18, 2025

ABSTRACT

Advances in bioinformatics have greatly contributed to the discovery of epithelial–mesenchymal transition (EMT) markers, such as epithelial cell adhesion molecule (EPCAM). This study aimed to conduct an EPCAM-centered meta-analyses of previously RNA-sequencing data for identifying pan-cancer EMT markers in circulating tumor cells (CTCs) utilizing bioinformatics- and machine learning (ML)-based approaches. In this study, the RNA sequencing data of seven different cancer types from two datasets, namely GSE273023 and GSE274442, were analyzed. Gene–gene correlation among included cancer samples and EPCAM-centered gene–gene correlation analysis were performed. The data were subjected to ML-based pathway and gene clustering analysis. Notably, the results showed that most of the cancers presented similar gene expression profile, albeit with some differences, which were primarily attributed to differences in mitochondrial gene expression. Furthermore, gene–gene correlation analysis revealed multiple genes with significantly altered expression, including *CBWD2*, *MED23*, *QRSL1*, *ZNF568*, and *INTU*. Similarly, *TRPS1* was found to be significantly correlated with *EPCAM*. Overall, the findings of this study reveal the association between *EPCAM–TRPS1* and *CBWD2*-associated *MED23–QRSL1–ZNF568–INTU* axes, thereby showing their potential as co-markers and for the development of multiplexed immunoassay for a robust pan-cancer CTC detection approach.

***Corresponding author:** Nagarajan Rajendra Prasad ✉ drprasadnr@gmail.com

* https://orcid.org/0000-0002-3937-8735

First author: https://orcid.org/0009-0009-1324-7698

Co-authors:

GR Brindha: https://orcid.org/0000-0001-5911-8327

INTRODUCTION

Cancer is a complex, multifactorial disease that significantly burdens the public health worldwide (Ferlay *et al*., 2024). Additionally, processes such as metastasis further exacerbate cancer, causing the disease to progress to more advanced and aggressive stages and other locations, thereby significantly increasing disease severity. Metastasis refers to the process that instigates the spread of cancer cells from the primary tumor to distant organs or tissues to forming secondary tumors, and epithelial–mesenchymal transition (EMT) is a crucial part of this process (Dongre and Weinberg, 2019; Gerstberger *et al*., 2023; Yeung and Yang, 2017).

Circulating tumor cells (CTCs) have been critically implicated in metastasis, as these cells travel from the primary tumor following EMT to colonize the other body parts through entry into the bloodstream or lymphatic system (Dongre and Weinberg, 2019). For CTCs to generate following EMT, cancer cells lose their cell–cell adhesion properties and structural polarity, thereby exhibiting increased motility and invasiveness. This transformation allows them to detach from the primary tumor mass, penetrate surrounding tissues, and enter circulation as CTCs (Garg, 2013; Lamouille *et al*., 2014). Owing to this, CTCs have been associated with the early tumor development stage and the complex process of metastasis.

Meta-analysis is a well-recognized approach for exploring and identifying novel markers and pathways for disease prognosis, and various studies have employed meta-analysis for reporting prognostic factors for different cancer types (Borenstein *et al*., 2021; Groot Koerkamp *et al*., 2013; Guven *et al*., 2022; Lv *et al*., 2016). For instance, in a meta-analysis of 18 independent EMT gene expression studies (both cell line and treatment-based), a core set of consistently up- and down-regulated genes were identified. These genes were overlapped with known EMT markers to reveal novel candidates, some of which were associated with poor therapeutic response in breast cancer (Gröger *et al*., 2012). Similarly, in bladder cancer, analysis of integrated networks revealed *CORO1C* and *TMPRSS4* as hub genes, and they were associated with EMT and poor prognosis through bioinformatics approach (Wang *et al*., 2020). Overall, meta-analysis and bioinformatics-based approaches offer notable advantages in the analysis of gene expression data for the exploration of novel marker sets.

In recent times, machine learning (ML) has become a powerful tool in studies on metastasis, facilitating risk prediction, biomarker identification, and understanding of mechanisms. For instance, in a study on colorectal cancer, ML was employed along with experimental validation to screen for metastasis biomarkers, and genes that distinguished primary tumor and liver metastasis were identified (Ahmadieh-Yazdi *et al*., 2023). Similarly, Random Forest models were used on clinical/demographic data of the patients with thyroid cancer to predict bone metastasis (Liu *et al*., 2021). Such studies highlight the applicability of ML in meta-analysis for a robust identification of disease markers.

Hence, this study aimed to conduct an epithelial cell adhesion molecule (EPCAM)-centered meta-analysis of RNA-sequencing (RNA-seq) data of multiple cancers for identifying pan-cancer EMT markers in CTCs utilizing bioinformatics- and ML-based approaches.

MATERIALS AND METHODS

Study selection

In the present study, the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/) was screened for datasets using the following terms and their combinations: “epithelial–mesenchymal transition,” “RNA sequencing,” “RNA-seq,” “metastasis,” “circulating tumor cell,” and “cancer,” and the species was set to “Homo sapiens.” Ultimately, GEO datasets GSE273023 (Liao and Zhou, 2025a) and GSE274442 (Liao and Zhou, 2025b), containing RNA-seq data for lung adenocarcinoma (LUAD), skin cutaneous melanoma (SKCM), liver hepatocellular carcinoma

(LIHC), thymoma (THYM), breast invasive carcinoma (BRCA), prostate adenocarcinoma (PRAD), and sarcoma (SARC), were selected for further analysis.

Pan-cancer gene expression analysis for EMT-driver genes

To identify pan-cancer-specific driver genes among the included datasets, a core set of EMT-associated genes was curated from the GeneCards database (https://www.genecards.org/) (Table 1), and the genes were grouped in mesenchymal-like (such as vimentin [VIM] and N-cadherin) and epithelial-like (namely E-cadherin [CDH1] and EPCAM) gene sets.

The EMT score represented the up- and downregulation of mesenchymal and epithelial components, as follows:

$$EMT\ score = Mesenchymal\ component - Epithelial\ component \qquad \text{(Eq. 1)}$$

where mesenchymal and epithelial components were the means of z-scores for mesenchymal markers (such as VIM, FN1, and ZEB1) and epithelial markers (such as CDH1 and EPCAM), respectively.

To scale differences robustly, z-scoring was performed, with positive EMT denoting more mesenchymal-like characteristic, whereas negative EMT denoting more epithelial-like characteristic.

The z-score standardized gene expression for fair comparison, regardless of the absolute levels of different genes, showing the distance of a value is from the mean in units of standard deviation (Shah and Parveen, 2025).

$$Z_{g,s} = \frac{x_{g,s} - \mu_g}{\sigma_g} \qquad \text{(Eq. 2)}$$

where x(g,s) denotes the raw expression of gene g in sample s; μ(g) is the mean expression of gene g across all samples. Σ(g) denotes the standard deviation of gene g across all samples. The z-scoring was evaluated as follows: z=0: exactly average; if z>0, EMT higher than the average; and z<0, EMT lower than the average.

False discovery rate (FDR)

The FDR denotes the proportion of false positives among the genes declared significant in the gene clusters, controlling “how many mistakes you tolerate” in large-scale testing. It is employed with summarizing pathway-level activity from multiple genes. The p-values were ranked as follows: p(1) ≤ p(2) ≤ …≤ p(m), and q-values were adjusted to ensure their monotonic increase with rank (Rosati *et al.*, 2024).

The Benjamini–Hochberg equation was used for calculating the q-value:

$$q_{(i)} = \frac{p_{(i)} \times m}{i} \qquad \text{(Eq. 3)}$$

where m is the total number of tests (genes) and q(i) is the FDR-adjusted p-value (q-value).

If FDR = 0.05, of all “significant” genes, approximately 5% may be false positives. The results were shown as a graph, and each point represented one cancer sample, plotted by its EMT score from Byers and Creighton EMT-scoring scoring systems. The line represented the best-fit linear regression.

Gene–gene correlation analysis

In the present study, gene–gene correlation analysis was performed to quantify the variation in the expression of two genes across samples or time through Pearson correlation analysis (r range: −1, perfect anti-correlation, to +1, perfect co-expression). Based on the per-gene z-scored expression for the top-40 EMT-driving candidates across samples, hub genes that represented strongest positive and negative gene–gene pairs were identified within the correlation networks.

EPCAM-centered gene–gene correlation analysis

Following the identification of the top 40 EMT-related genes across the cancer types included in this study, the gene–gene association of EMT- and metastasis-related markers (Tables 1–2) in regards to EPCAM was analyzed. Based on this analysis, the top 10 positively and and negatively EPCAM-correlated markers were identified.

K-Means clustering of gene expression data and boxplot analysis

The K-Means clustering algorithm was applied to the obtained low-dimensional clusters to identify inherent groupings within the data (Hussain *et al.*, 2024). Notably, the optimal number of clusters was determined using internal validation techniques such as the silhouette score, and the clustered data was visualized in the Uniform Manifold Approximation and Projection (UMAP) 2D space and as scatterplots to facilitate identifying separate clusters and underlying patterns in the dataset. Additionally, boxplot analysis was performed to evaluate the variability in distributional characteristics of gene expression in different cancer types and identify the central tendency (median, Inter Quartile Range [IQR]) across all cancer groups.

Pathway-guided weighted distance (PGWD) K-means analysis for gene expression data

Pathway-guided clustering, an unsupervised machine learning technique that utilizes route information to guide the clustering process, was employed to classify genes based on their activity in pre-established biological pathways. Notably, gene expression data were first converted into pathway activity profiles, which were then clustered to capture coordinated biological processes. To cluster genes contributing to comparable pathways, the algorithm iteratively assigned genes to clusters and updated centroids based on weighted pathway activity. PGWD K-Means analysis reveals biologically meaningful genes by weighting features using pathway knowledge, following which, standard K-means is run in a re-weighted feature space (equivalent to Euclidean distance after feature scaling) (Malla *et al.*, 2024; Sun *et al.*, 2023; Yousef *et al.*, 2023). Proposed Algorithm PGWD–K-Means (including parameters X, A, K, α, β, λ, τ, and q) is discussed below.

For preprocessing, X[:, g] was standardized for each gene g to zero-mean, unit-variance, and low-variance genes were filtered. To determine pathway relevance (r_p), r_p was computed for each pathway p using either unsupervised variance of pathway activity, supervised enrichment/t-score/area under the reciever operating curve if labels exist, and prior knowledge score. Notably, r_p ← r_p/(median_p r_p) was normalized, and r_p was clipped at the upper percentile q to avoid domination. Gene Centrality c_{g,p} (uniform, network, or stability-based) was computed for for each (g, p) with A[g,p] = 1. The value was normalized within the pathway as follows: c_{g,p} ← c_{g,p}/(mean_{g∈G_p} c_{g,p}).

To aggregate gene weights for each gene g, the following process was employed:
s_g ← Σ_{p} A[g,p] · r_p^α · c_{g,p}^β
w_g_raw ← s_g / (1 + τ · (degree_g - 1)), where overlap penalty was (degree_g=Σ_p A[g,p])

Shrinkage/regularization was performed as follows:
w_g ← (1 - λ)·w_g_raw + λ, where λ ∈ [0,1] keeps minimum weight > 0
Next, weights were rescaled and clipped at upper percentile q:
w_g ← w_g / median_g(w_g).

Weighted Feature Space was presented as Construct X_w, where X_w[:, g] ← X[:, g] · √w_g.
For analyzing K-Means in Weighted Space, μ_k (k-means++ on X_w) was initialized and the process was repeated until convergence:
Assignment: z_i ← argmin_k || X_w[i, :] - μ_k ||_2^2
Update: μ_k ← mean of assigned X_w rows in cluster k
Return z, w.

For two samples x_i, x_j, distance induced by PGWD was calculated as follows:

$$d_{PGWD}(x_i,x_j)=\Sigma_{g=1}^{d}(x_{i,g}-x_{j,g})^2 \qquad \text{(Eq. 4)}$$

which is exactly Euclidean distance after scaling each feature by $\sqrt{w_g}$.

Hyperparameters included α∈[0.5,2] denoting pathway relevance emphasis, β∈[0,1] denoting centrality emphasis, λ∈[0,0.3] for weight shrinkage for stability, τ∈[0,0.5] for the overlap penalty for multi-pathway genes, and q∈[0.90,0.99] denoting clipping percentile to prevent domination. Owing to prevalence of overlapping pathways, the overlap penalty (τ) was induced to prevent inflated weights for hub genes. This process included top-weighted pathways and genes, and ablation of α, β, λ to show robustness.

RESULTS

Identification of EMT-driving genes

In the present study, expression patterns of the genes were evaluated to identify EMT-driver genes. For each cancer, the top five overexpressed mesenchymal and suppressed epithelial markers were identified (Table 1). Overexpression meant higher-than-normal gene activity; suppression meant lower activity.

Table 1. Summary of markers and each cancer type

Cancer type	Top mesenchymal driver(s)	Top suppressed epithelial marker(s)	Hypothesized EMT mechanism
Lung adenocarcinoma	VIM, FN1, SNAI2, ZEB1, MMP9	CDH1, OCLN, EPCAM, DSP, S100A9	SNAI2-mediated suppression of adhesion for invasion.
Skin cutaneous melanoma	VIM, CDH2, ZEB2, TWIST1, MMP2	CDH1, OCLN, DSP, EPCAM, S100A9	ZEB2-driven neural crest traits for melanocyte dissemination.
Thymoma	FN1, SNAI1, ZEB1, VIM, MMP9	CDH1, OCLN, DSP, EPCAM, S100A9	SNAI1-induced dedifferentiation in thymic stroma.
Liver hepatocellular carcinoma	VIM, SNAI1, TWIST1, ZEB1, MMP9	CDH1, OCLN, DSP, EPCAM, S100A9	TWIST1-TGFβ axis for matrix remodeling and vascular invasion.
Breast invasive carcinoma	VIM, SNAI1, ZEB1, FN1, MMP2	CDH1, OCLN, DSP, EPCAM, S100A9	SNAI1-ZEB1 repression of adhesion for stemness.
Sarcoma	VIM, FN1, CDH2, TWIST2, MMP2	CDH1, OCLN, DSP, EPCAM, S100A9	Inherent mesenchymal state with FN1-matrix deposition.
Prostate adenocarcinoma	VIM, SNAI2, ZEB2, FN1, MMP9	CDH1, OCLN, DSP, EPCAM, S100A9	SNAI2-androgen independence for bone tropism.

EMT, epithelial–mesenchymal transition.

A moderate-to-strong positive correlation was found between the Byers and Creighton EMT-scoring methods, suggesting that both approaches broadly agree on which samples are epithelial-like and mesenchymal-like. However, the spread around the regression line showed notable variability, indicating that some samples scored differently depending on the method.

Overall, this agreement validates EMT scoring as a reproducible concept across algorithms, with the scatter emphasizing that while EMT is a robust program, methodological differences may shift individual sample classifications (Fig. 1).

Gene-gene correlation analysis

The results of gene–gene correlation showed that per-gene z-scored expression were used to identify the top-40 EMT-driving markers across all samples, revealing a strong epithelial module (Fig. 2A).

Epithelial hub genes *CBWD2*, *MED23*, *ZNF568*, *NBPF11*, *FAM120B*, *INTU*, *QRSL1*, *ZSCAN30*, and *ZNF443* formed a strong association, with high expression in PRAD/THYM and low in LUAD1. In contrast, mesenchymal/immune cluster such as *SERPINI2*/*RASA3*/*MGAT1*/*FKBP11*/*FOLR2* showed relative elevation in mesenchymal samples, with *CYB561D2* standing out as higher in mesenchymal-high versus epithelial-low, consistent with the abovementioned findings of this study. Many epithelial-hub genes exhibited large differences between mesenchymal and epithelial endpoints (≈2–3 z-units), visually evident as deep cool versus warm blocks between LUAD1-like and PRAD/THYM columns, which supported a coordinated switch-off of the epithelial program during EMT. However, these results only emphasize effect sizes and co-expression modules, with n = 10 and approximately 50k tests, none of the single genes crossed FDR < 0.05.

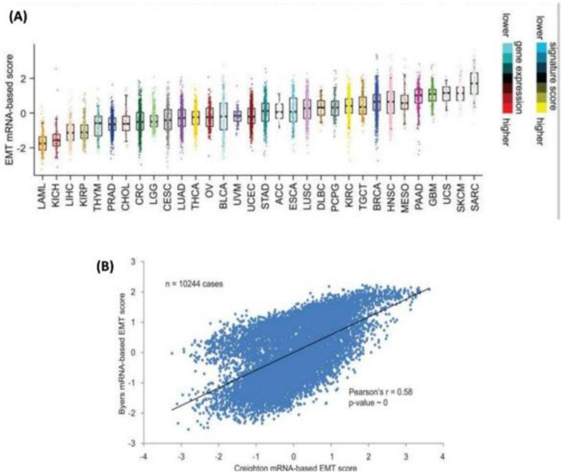

Fig. 1. (A) Statistical summary of gene versus EMT messenger RNA-based score. (B) Byers versus Creighton EMT scores. EMT, epithelial–mesenchymal transition

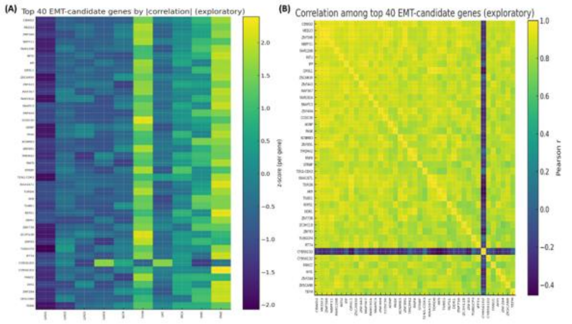

Fig. 2. (A) Top 40 EMT-candidate genes across cancer types. (B) Pearson correlation matrix of top 40 EMT-candidate genes. EMT, epithelial–mesenchymal transition

The matrix (Fig. 2B) visualized pairwise Pearson correlations among the top-40 EMT-tracking genes, showing that linear co-variation of two genes across samples: r≈1 indicated lockstep co-expression, r≈0 independence, and r<0 anti-phasic expression. Module blocks appeared as contiguous high-r regions, while cross-shaped negative bands marked genes that oppose an entire block. Hub genes showed uniformly high correlations across a row/column and had high mean absolute r to others, summarizing module behavior and often being anchor biological programs such as epithelial modules in EMT. Hub metrics placed *CBWD2* at the center of the epithelial module, with *MED23/QRSL1/ZNF568/INTU* close behind. This explained the strong block structure and model where an epithelial program switched off coherently as EMT increases while mesenchymal/immune activation is more heterogeneous.

EPCAM-centered gene–gene analysis

In the present study, an EPCAM-centered analysis was performed considering its established role as an epithelial and CTC marker in the context of EMT and metastasis. Notably, the top 10 most positively and negatively correlated genes with EPCAM were extracted. Furthermore, an EPCAM-centered hub-and-spoke network was visualized, correlating EPCAM and other markers (Table 2).

Table 2. Top 10 mesenchymal- and epithelial-like genes

Mesenchymal-like			Epithelial-like		
Gene	Pearson correlation (r)	FDR	Gene	Pearson correlation (r)	FDR
SERPINI2	0.905	0.232	CBWD2	-0.971	0.103
RASA3	0.898	0.232	MED23	-0.967	0.103
FKBP11	0.887	0.232	ZNF568	-0.962	0.106
MGAT1	0.879	0.232	NBPF11	-0.960	0.106
MATK	0.872	0.232	FAM120B	-0.955	0.145
FOLR2	0.866	0.232	INTU	-0.951	0.155
KCNH3	0.866	0.232	IPP	-0.949	0.155
SNAPC2	0.852	0.232	QRSL1	-0.945	0.155
CD300C	0.849	0.232	ZSCAN30	-0.944	0.155
RNH1	0.845	0.232	ZNF443	-0.944	0.155

FDR, false-discovery rate.

Consistent with its role as an epithelial marker, EPCAM was found to be among the top downregulated genes across multiple tumor types in this study and strongly positively correlated with epithelial-associated markers and both negatively correlated with canonical EMT/metastasis markers. The EPCAM-centered network visualization (Fig. 3A) illustrates this relationship, emphasizing its utility as a CTC marker and its regulatory placement within the EMT spectrum.

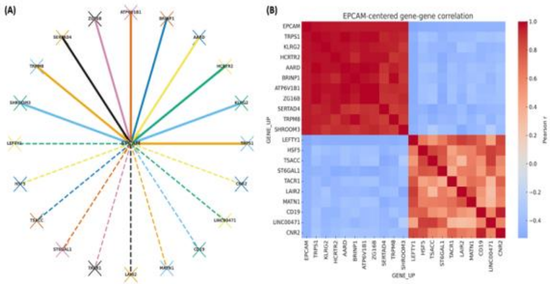

Fig. 3. (A) EPCAM-centered correlation network. Solid lines, positive correlation; dotted lines, negative correlation. (B) Heat map with EPCAM as the hub gene. EPCAM, epithelial cell adhesion molecule

Notably, EPCAM displayed strong bidirectional correlations with EMT- and metastasis-related genes across different cancer types. Among the positively correlated genes, the strongest associations were observed with *TRPS1* (r = 0.975), *KLRG2* (r = 0.963), and *HCRTR2* (r = 0.963), followed by *AARD* (r = 0.960) and *BRINP1* (r = 0.958). These genes represent epithelial or epithelial-like signatures that reinforce the canonical role of EPCAM as an epithelial marker. In contrast, the negatively correlated genes included *LEFTY1* (r = −0.954), *HSF5* (r = −0.954), *TSACC* (r = −0.953), *ST6GAL1* (r = −0.950), and *TACR1* (r = −0.947). These markers are consistent with mesenchymal or EMT-associated programs (Fig. 3B; Table 3).

Table 3. Top 10 positively and negatively correlated genes with epithelial cell adhesion molecule

Gene	Positive correlation (r)	Gene	Negative correlation (r)
TRPS1	0.974	LEFTY1	-0.523
KLRG2	0.963	HSF5	-0.499
HCRTR2	0.962	TSACC	-0.488
AARD	0.958	ST6GAL1	-0.484
BRINP1	0.957	TACR1	-0.478
ATP6V1B1	0.957	LAIR2	-0.472
ZG16B	0.956	MATN1	-0.471
SERTAD4	0.955	CD19	-0.469
TRPM8	0.954	LINC00471	-0.466
SHROOM3	0.954	CNR2	-0.465

K-Means clustering of gene expression data and boxplot analysis

In the present study, the cancer type data were clustered into the groups of two, three, and four (k=2, 3, and 4, respectively). In column-wise k=4 clustering, four main clusters were formed, as follows: Cluster 0, LUAD3, LUAD4, PRAD, and SARC; Cluster 1, LUAD2; Cluster 2, LUAD1; and Cluster 3, BRCA, THYM, SKCM, and LIHC (Fig. 4A–C). The results suggested that data in Cluster 0 and 3 showed considerable intra-cluster similarities, whereas LUAD1 and LUAD2, which were present at distinct points, indicated their significantly different variance patterns. In k=3 clustering, Cluster 0 absorbed LUAD2, resulting in LUAD1 to be a consistent outlier cluster. In k=2 clustering, Cluster 0 included all samples except LUAD1, which comprised a separate Cluster 1, indicating its significantly varying expression profile, even compared with those of the same cancer type (namely LUAD2–4).

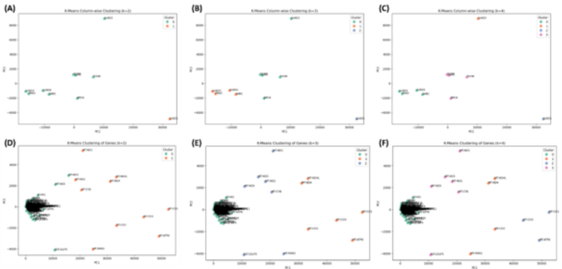

Fig. 4. K-means column-wise sample clustering results at (A) k=2; (B) k=3; (C) k=4. K-means gene clustering results at (D) k=2; (E) k=3; (F) k=4. PC, PC, principal component

Regarding the clustering of genes, k=2, 3, and 4 clustering analyses revealed notable separation between nuclear and mitochondrial gene set (Fig. 4D–F). In k=4 clustering, Cluster 0 contained most genes (including all nuclear) and was densely packed near the origin. In contrast, Clusters 1, 2, and 3 contained various mitochondrial genes, suggesting that functional grouping of mitochondrial genes separated them from nuclear gene expression. In k=3 clustering, Cluster 0 still retained most nuclear genes, and Clusters 1 and 2 comprised the distinct mitochondrial set. Overall, mitochondrial genes were broadly split as follows: ND-type genes (ND1–ND5 + CYB) and CO/ATP-type genes (CO1–CO3 + ATP8 + ND4/ND4L).

In k=2 clustering, some mitochondrial genes were absorbed in Cluster 0; however, they were still considered the major contributor of the expression heterogeneity.

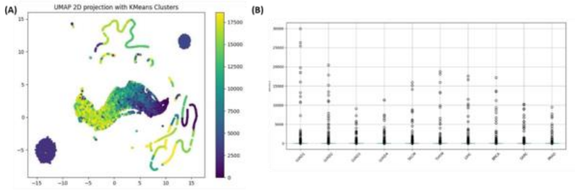

Fig. 5. (A) Uniform Manifold Approximation and Projection 2D plot with K-means clusters. (B) Boxplot analysis result encompassing all features across multiple cancer types and showing pronounced variability in distributional characteristics

The UMAP 2D projection of clustering results suggested that the dataset exhibited both well-defined and loosely structured groups (Fig. 5A). The dense central cluster likely represented the dominant patterns or majority class, whereas the isolated peripheral clusters denoted niche or outlier groups. The elongated and curved cluster shapes further indicated that the data is non-linearly separable in its original feature space, which UMAP effectively captured in this 2D representation. The overlapping regions between clusters implied certain similarity or shared characteristics across those data points, potentially signaling gradual transitions rather than sharp boundaries. This insight may be crucial for downstream tasks such as classification, segmentation, or anomaly detection, to ultimately reveal complex relationships within the dataset that a simple linear model might not capture.

The boxplot effectively encompassed all features across multiple cancer types (Fig. 5B), demonstrating pronounced variability in distributional characteristics. Notably, cancers such as LUAD and BRCA exhibited higher medians and broader IQRs, suggesting heterogeneous gene expression landscapes. In contrast, cancers such as THYM and PRAD exhibited narrower distributions, which indicated greater uniformity across samples. These differences highlight the diverse molecular architectures of the cancers under study, with some showing significant within-group variability while others remain relatively stable.

ML-based pathway-guided clustering of gene expression data

Notably, two GEO datasets GSE273023 and GSE274442, containing RNA-seq data for LUAD (LUAD1–4), SKCM, LIHC, THYM, BRCA, PRAD, and SARC, were included in this study. The results of pathway-guided clustering of gene expression data revealed that most cancer samples (namely LUAD, BRCA, LIHC, THYM, PRAD, and SARC) were grouped under Cluster 0, suggesting overlapping gene expression patterns—driven by shared oncogenic pathways—among these cancers (Fig. 6A). Interestingly, LUAD4 (Cluster 1) and SKCM (Cluster 2) were positioned separately from most samples, reflecting differences in pathway-level expression and transcriptional activity compared with the other cancers. These results highlighted that pathway-guided clustering can detect both common and unique signatures across different cancer types.

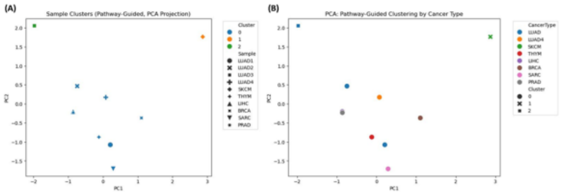

Fig. 6. (A) PCA plot of pathway-guided sample clustering. (B) PCA plot of pathway-guided clustering by cancer type. PCA, principal component analysis; PC, principal component

In clustering based on cancer type, most cancers (namely LUAD, BRCA, LIHC, THYM, PRAD, and SARC) remained closely grouped, indicating the presence of shared molecular characteristics (Fig. 6B). Consistent with the results of pathway-based clustering, SKCM and LUAD4 were observed to be placed differently from the major cluster, further validating their different transcriptional program. Overall, these alignments of clustering with known cancer types demonstrated the biological interpretability of the pathway-weighted approach.

DISCUSSION

This study conducted an EPCAM-centered meta-analyses of RNA-seq data derived from GEO datasets GSE273023 and GSE274442 (Liao and Zhou, 2025a, 2025b) for identifying pan-cancer EMT markers in CTCs utilizing bioinformatics- and ML-based approaches. The integration of pathway-guided clustering, ML approaches, and EMT-focused gene expression profiling showed notable similarities between different cancer types. In the present study, most cancers showed overlapping pathway-level expression signatures, with LUAD subsets and SKCM diverging as distinct clusters with unique transcriptional activity. Moreover, mitochondrial gene expression played a significant role in shaping clustering outcomes. Additionally, EPCAM-centered correlations provided an array of positively and negatively correlated markers.

The pathway-guided clustering results revealed the involvement of shared oncogenic pathways regarding similarities across different cancer types. LUAD, BRCA, LIHC, THYM, PRAD, and SARC were largely grouped into a single dominant cluster, indicating the presence of core transcriptional programs related to proliferation, metabolic reprogramming, and cell survival (Hung *et al.*, 2015; Zhang *et al.*, 2014). Interestingly, LUAD4 and SKCM diverged from the dominant cluster, indicating differences in tissue-specific and subtype-specific transcriptional activity. For instance, SKCM was shown to be driven by melanocyte lineage programs and immune evasion mechanisms, whereas deviations in LUAD4 reflected genomic alterations or microenvironmental influences unique to that sample set. These results of clustering patterns were consistent with those of pathway-weighted approaches, which accounted for raw expression variance and higher-level functional context. Overall, these results signified the use of pathway-guided models in the analysis of heterogeneous cancers owing to their advantages of comprehensive analysis.

K-means clustering revealed intra-type heterogeneity and mitochondrial contributions, complementing pathway-guided findings, particularly within LUAD. LUAD1 consistently emerged as an outlier, suggesting a different molecular subset in their transcriptional programs. Similarly, the divergence of LUAD2 suggested that even cancers classified under the same histological type may harbor distinct transcriptomic landscapes (Allison and Sledge, 2014; Roggli *et al.*, 1985; Wu *et al.*, 2021).

Furthermore, the consistent separation of mitochondrial and nuclear genes was observed. The clustering of ND- and CO/ATP-type mitochondrial genes into distinct groups represented important sources of heterogeneity across tumors. Moreover, UMAP projections underscored the non-linear structure of the data, highlighting that gene expression heterogeneity in cancer cannot be captured by simple linear boundaries, and thus, further justifying the application of advanced dimensionality reduction and ML-based clustering approaches (Lee *et al.*, 2021; Lee *et al.*, 2022; Vera-Yunca *et al.*, 2020). The findings of boxplot analysis provided a complementary distributional perspective, highlighting that cancers showed both converging mechanisms and tissue-specific uniqueness across different types.

The gene–gene correlation analysis identified robust epithelial modules (e.g., *CBWD2*, *MED23*, *ZNF568*), which demonstrated strong positive correlations and hub-like architecture, suggesting that epithelial programs are tightly regulated and switch off in a coordinated fashion during EMT. The EPCAM-centered analysis contextualizes EMT in terms of a well-known epithelial marker with clinical relevance in CTC detection. EPCAM strongly correlated with epithelial-associated markers such as *TRPS1*, *KLRG2*, and *BRINP1*, along with its negative correlations with EMT/mesenchymal-associated genes such as *LEFTY1* and *ST6GAL1*, highlighting its regulatory opposition to EMT programs.

TRPS1 expression has been reported to influence the progression in different tumors, indicating its prognostic role regarding CTCs and EMT (Hong *et al.*, 2013; Stinson *et al.*, 2011). Similarly, studies have implicated the expression of *CBWD2*, *MED23* (Shi *et al.*, 2014), *QRSL1* (Dursun *et al.*, 2022; Wang *et al.*, 2023), *ZNF568* (Han *et al.*, 2024; Wang *et al.*, 2020), and *INTU* (Chan and Chen, 2022) in tumor progression and associated processes across different cancer types. Overall, these studies indicate the involvements of aforementioned genes in various tumors. Combining with the results of the present study, which highlight the significant correlations among these genes, the findings suggest that these genes may serve as prognostic markers while offering a robust prognostic efficacy when combined with EPCAM for metastasis evaluation and CTC detection.

Altogether, these results highlight the potential of *EPCAM–TRPS1* and *CBWD2*-associated *MED23–QRSL1–ZNF568–INTU* axes as potential biomarkers, along with underscoring the importance of hybrid EMT states and the prognostic superiority of ML-based scoring methods. Furthermore, the findings provide a research basis for the future studies on the proposed axes for the development of robust CTC detection methods such as multiplexed immunoassays.

ACKNOWLEDGEMENTS

This study was financially supported by the DST-INSPIRE fellowship, Department of Science and Technology, Government of India, India [Grant number DST/INSPIRE/03/2021/002275].

REFERENCES

Ahmadieh-Yazdi A, Mahdavinezhad A, Tapak L, Nouri F, Taherkhani A, Afshar S. 2023. Using machine learning approach for screening metastatic biomarkers in colorectal cancer and predictive modeling with experimental validation. Scientific Reports **13**, 19426. https://doi.org/10.1038/s41598-023-46633-8

Allison KH, Sledge GW. 2014. Heterogeneity and cancer. Oncology (Williston Park) **28**, 772–778.

Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. 2021. Introduction to meta-analysis (2nd ed.). John Wiley & Sons.

Chan HYE, Chen ZS. 2022. Multifaceted investigation underlies diverse mechanisms contributing to the downregulation of Hedgehog pathway-associated genes INTU and IFT88 in lung adenocarcinoma and uterine corpus endometrial carcinoma. Aging **14**, 7794–7823.
https://doi.org/10.18632/aging.204262

Dongre A, Weinberg RA. 2019. New insights into the mechanisms of epithelial–mesenchymal transition and implications for cancer. Nature Reviews Molecular Cell Biology **20**, 69–84.
https://doi.org/10.1038/s41580-018-0080-4

Dursun F, Genc HM, Mine Yılmaz A, Tas I, Eser M, Pehlivanoglu C, Yilmaz BK, Guran T. 2022. Primary adrenal insufficiency in a patient with biallelic QRSL1 mutations. European Journal of Endocrinology **187**, K27–K32.
https://doi.org/10.1530/EJE-22-0233

Ferlay J, Ervik M, Lam F, Laversanne M, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, Bray F. 2024. Global Cancer Observatory: Cancer today. Lyon, France: International Agency for Research on Cancer.
https://gco.iarc.who.int/today

Garg M. 2013. Epithelial-mesenchymal transition-activating transcription factors: Multifunctional regulators in cancer. World Journal of Stem Cells **5**, 188.https://doi.org/10.4252/wjsc.v5.i4.188

Gerstberger S, Jiang Q, Ganesh K. 2023. Metastasis. Cell **186**, 1564–1579.
https://doi.org/10.1016/j.cell.2023.03.003

Gröger CJ, Grubinger M, Waldhör T, Vierlinger K, Mikulits W. 2012. Meta-analysis of gene expression signatures defining the epithelial to mesenchymal transition during cancer progression. PLoS ONE **7**, e51136.
https://doi.org/10.1371/journal.pone.0051136

Groot Koerkamp B, Rahbari NN, Büchler MW, Koch M, Weitz J. 2013. Circulating tumor cells and prognosis of patients with resectable colorectal liver metastases or widespread metastatic colorectal cancer: A meta-analysis. Annals of Surgical Oncology **20**, 2156–2165. https://doi.org/10.1245/s10434-013-2907-8

Guven DC, Sahin TK, Erul E, Kilickap S, Gambichler T, Aksoy S. 2022. The association between the pan-immune-inflammation value and cancer prognosis: A systematic review and meta-analysis. Cancers **14**, 2675. https://doi.org/10.3390/cancers14112675

Han CW, Jeong MS, Jang SB. 2024. Influence of the interaction between p53 and ZNF568 on mitochondrial oxidative phosphorylation. International Journal of Biological Macromolecules **275**, 133314. https://doi.org/10.1016/j.ijbiomac.2024.133314

Hong J, Sun J, Huang T. 2013. Increased expression of TRPS1 affects tumor progression and correlates with patients' prognosis of colon cancer. BioMed Research International **2013**, 1–6. https://doi.org/10.1155/2013/454085

Hung RJ, Ulrich CM, Goode EL, Brhane Y, Muir K, Chan AT, Marchand LLe, Schildkraut J, Witte JS, Eeles R, Boffetta P, Spitz MR, Poirier JG, Rider DN, Fridley BL, Chen Z, Haiman C, Schumacher F, Easton DF, Landi MT, Brennan P, Houlston R, Christiani DC, Field JK, Bickeböller H, Risch A, Kote-Jarai Z, Wiklund F, Grönberg H, Chanock S, Berndt SI, Kraft P, Lindström S, Al Olama AA, Song H, Phelan C, Wentzensen N, Peters U, Slattery ML; GECCO; Sellers TA; FOCI; Casey G, Gruber SB; CORECT; Hunter DJ; DRIVE; Amos CI, Henderson B; GAME-ON Network. 2015. Cross cancer genomic investigation of inflammation pathway for five common cancers: Lung, ovary, prostate, breast, and colorectal cancer. Journal of the National Cancer Institute **107**, djv246. https://doi.org/10.1093/jnci/djv246

Hussain I, Nataliani Y, Ali M, Hussain A, Mujlid HM, Almaliki FA, Rahimi NM. 2024. Weighted multiview K-means clustering with L2 regularization. Symmetry **16**, 1646. https://doi.org/10.3390/sym16121646

Lamouille S, Xu J, Derynck R. 2014. Molecular mechanisms of epithelial–mesenchymal transition. Nature Reviews Molecular Cell Biology **15**, 178–196. https://doi.org/10.1038/nrm3758

Lee D, Park Y, Kim S. 2021. Towards multi-omics characterization of tumor heterogeneity: A comprehensive review of statistical and machine learning approaches. Briefings in Bioinformatics **22**, bbaa188. https://doi.org/10.1093/bib/bbaa188

Lee JY, Lee K, Seo BK, Cho KR, Woo OH, Song SE, Kim E-K, Lee HY, Kim JS, Cha J. 2022. Radiomic machine learning for predicting prognostic biomarkers and molecular subtypes of breast cancer using tumor heterogeneity and angiogenesis properties on MRI. European Radiology **32**, 650–660. https://doi.org/10.1007/s00330-021-08146-8

Liao Z, Zhou W. 2025a. RNA-seq of vertebral metastatic tumor samples from pan-cancer primary tumors. Gene Expression Omnibus (GEO) database (Accession Number GSE273023).

Liao Z, Zhou W. 2025b. RNA-seq of vertebral metastatic tumor samples from pan-cancer primary tumors II. Gene Expression Omnibus (GEO) database (Accession Number GSE274442).

Liu W, Li Z, Luo Z, Liao W, Liu Z, Liu J. 2021. Machine learning for the prediction of bone metastasis in patients with newly diagnosed thyroid cancer. Cancer Medicine **10**, 2802–2811. https://doi.org/10.1002/cam4.3776

Lv Q, Gong L, Zhang T, Ye J, Chai L, Ni C, Mao Y. 2016. Prognostic value of circulating tumor cells in metastatic breast cancer: A systemic review and meta-analysis. Clinical and Translational Oncology **18**, 322–330. https://doi.org/10.1007/s12094-015-1372-1

Malla SB, Byrne RM, Lafarge MW, Corry SM, Fisher NC, Tsantoulis PK, Mills ML, Ridgway RA, Lannagan TRM, Najumudeen AK, Gilroy KL, Amirkhah R, Maguire SL, Mulholland EJ, Belnoue-Davis HL, Grassi E, Viviani M, Rogan E, Redmond KL, Sakhnevych S, McCooey AJ, Bull C, Hoey E, Sinevici N, Hall H, Ahmaderaghi B, Domingo E, Blake A, Richman SD, Isella C, Miller C, Bertotti A, Trusolino L, Loughrey MB, Kerr EM, Tejpar S; S:CORT consortium; Maughan TS, Lawler M, Campbell AD, Leedham SJ, Koelzer VH, Sansom OJ, Dunne PD. 2024. Pathway level subtyping identifies a slow-cycling biological phenotype associated with poor clinical outcomes in colorectal cancer. Nature Genetics **56**, 458–472. https://doi.org/10.1038/s41588-024-01654-5

Roggli VL, Vollmer RT, Greenberg SD, McGavran MH, Spjut HJ, Yesner R. 1985. Lung cancer heterogeneity: A blinded and randomized study of 100 consecutive cases. Human Pathology **16**, 569–579. https://doi.org/10.1016/S0046-8177(85)80106-4

Rosati D, Palmieri M, Brunelli G, Morrione A, Iannelli F, Frullanti E, Giordano A. 2024. Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A review. Computational and Structural Biotechnology Journal **23**, 1154–1168. https://doi.org/10.1016/j.csbj.2024.02.018

Shah SNA, Parveen R. 2025. Differential gene expression analysis and machine learning identified structural, TFs, cytokine and glycoproteins, including SOX2, TOP2A, SPP1, COL1A1, and TIMP1 as potential drivers of lung cancer. Biomarkers **30**, 200–215. https://doi.org/10.1080/1354750X.2025.2461698

Shi J, Liu H, Yao F, Zhong C, Zhao H. 2014. Upregulation of mediator MED23 in non-small-cell lung cancer promotes the growth, migration, and metastasis of cancer cells. Tumor Biology **35**, 12005–12013. https://doi.org/10.1007/s13277-014-2499-3

Stinson S, Lackner MR, Adai AT, Yu N, Kim H-J, O'Brien C, Spoerke J, Jhunjhunwala S, Boyd Z, Januario T, Newman RJ, Yue P, Bourgon R, Modrusan Z, Stern HM, Warming S, de Sauvage FJ, Amler L, Yeh R-F, Dornan D. 2011. TRPS1 targeting by miR-221/222 promotes the epithelial-to-mesenchymal transition in breast cancer. Science Signaling **4**, ra41. https://doi.org/10.1126/scisignal.2001538

Sun Z, Chung D, Neelon B, Millar-Wilson A, Ethier SP, Xiao F, Zheng Y, Wallace K, Hardiman G. 2023. A Bayesian framework for pathway-guided identification of cancer subgroups by integrating multiple types of genomic data. Statistics in Medicine **42**, 5266–5284. https://doi.org/10.1002/sim.9911

Vera-Yunca D, Girard P, Parra-Guillen ZP, Munafo A, Trocóniz IF, Terranova N. 2020. Machine learning analysis of individual tumor lesions in four metastatic colorectal cancer clinical studies: Linking tumor heterogeneity to overall survival. The AAPS Journal **22**, 58. https://doi.org/10.1208/s12248-020-0434-7

Wang C, Yang Y, Yin L, Wei N, Hong T, Sun Z, Yao J, Li Z, Liu T. 2020. Novel potential biomarkers associated with epithelial to mesenchymal transition and bladder cancer prognosis identified by integrated bioinformatic analysis. Frontiers in Oncology **10**, 931. https://doi.org/10.3389/fonc.2020.00931

Wang H, Shen L, Li Y, Lv J. 2020. Integrated characterisation of cancer genes identifies key molecular biomarkers in stomach adenocarcinoma. Journal of Clinical Pathology **73**, 579–586. https://doi.org/10.1136/jclinpath-2019-206400

Wang X, Li X, Jiang W. 2023. High expression of RTN4IP1 predicts adverse prognosis for patients with breast cancer. Translational Cancer Research **12**, 859–872. https://doi.org/10.21037/tcr-22-2350

Wu F, Fan J, He Y, Xiong A, Yu J, Li Y, Zhang Y, Zhao W, Zhou F, Li W, Zhang J, Zhang X, Qiao M, Gao G, Chen S, Chen X, Li X, Hou L, Wu C, Su C, Ren S, Odenthal M, Buettner R, Fang N, Zhou C. 2021. Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. Nature Communications **12**, 2540. https://doi.org/10.1038/s41467-021-22801-0

Yeung KT, Yang J. 2017. Epithelial–mesenchymal transition in tumor metastasis. Molecular Oncology **11**, 28–39. https://doi.org/10.1002/1878-0261.12017

Yousef M, Ozdemir F, Jaber A, Allmer J, Bakir-Gungor B. 2023. PriPath: Identifying dysregulated pathways from differential gene expression via grouping, scoring, and modeling with an embedded feature selection approach. BMC Bioinformatics **24**, 60. https://doi.org/10.1186/s12859-023-05187-2

Zhang J, Wu L-Y, Zhang X-S, Zhang S. 2014. Discovery of co-occurring driver pathways in cancer. BMC Bioinformatics **15**, 271. https://doi.org/10.1186/1471-2105-15-271