Bioinformatics, 2026, **42**(5), btag286
https://doi.org/10.1093/bioinformatics/btag286
Published: 7 May 2026
Applications Note | Gene expression

OXFORD

EMTscore infers divergent EMT pathways from omics data and enables rapid screening for EMT-associated gene sets

Haimei Wen[1], Leonidas Bleris[2], Tian Hong[1,*]

[1]Department of Biological Sciences, The University of Texas at Dallas, Richarson, TX 75080, United States
[2]Department of Bioengineering, The University of Texas at Dallas, Richarson, TX 75080, United States
*Corresponding author. Department of Biological Sciences, The University of Texas at Dallas, 800 W Campbell Rd, Richarson, TX 75080, United States.
E-mail: hong@utdallas.edu.
Associate Editor: Peter Robinson

Abstract

Motivation: **Quantitative analyses of epithelial-mesenchymal transition (EMT) have been widely used in several areas of biomedical sciences due to its importance in development and cancer progression, but its multi-contextual nature requires standardization and implementation of gene set scoring methods beyond capacities of conventional tools.**

Results: **We developed EMTscore, a package that provides an efficient implementation of unbiased scoring methods for multiple EMT pathways using individual single-cell or bulk omics data, and the package allows rapid screening for cellular processes correlated with EMT.**

Availability and implementation: **EMTscore is available from GitHub https://github.com/wenmm/EMTscore under the GNU General Public License, and is uploaded on Zenodo with a DOI 10.5281/zenodo.19487376.**

1 Introduction

Epithelial-mesenchymal transition (EMT) is the cellular process in which epithelial cells with tight cell junction and apical-basal polarity transform into motile mesenchymal cells. EMT is responsible for key steps of development and is activated in diseases progressions such as fibrosis and metastasis (Nakaya and Sheng 2013). Particularly, cancer cells can exploit the gene regulatory network of EMT and the invasiveness of the mesenchymal (M) cells (Thiery et al. 2009). Experimental and theoretical studies have shown that EMT involves multiple intermediate states rather than a binary switch (Lu et al. 2013, Zhang et al. 2014, Hong et al. 2015, Yang et al. 2020), and partial EMT was shown to be crucial for cancer progression (Pastushenko et al. 2018). In addition, divergent transcriptional programs of EMT can be activated in fate changes toward multiple cell subtypes/states even in the same disease (Groves et al. 2022, Youssef et al. 2024). The importance and complexity of EMT gained significant attention from multiple subfields of biomedical sciences. Over the past decade, several thousand EMT-related research articles have been published each year (Fig. 1A).

Due to the multistate nature of EMT and its diverse physiological contexts, EMT cannot be accurately quantified with a small number of genes (Yang et al. 2020). Gene sets analysis is therefore widely used to generate "EMT scores" to assess the degree of EMT in different settings (Fig. 1A) (Tan et al. 2014, Panchy et al. 2022). Scoring EMT with high-throughput experiments, including single-cell RNA-sequencing (scRNA-seq), enabled both discoveries for fundamental understanding of EMT and development of new therapeutic strategies (Elamin et al. 2022, Cassier et al. 2023, Youssef et al. 2024, Groves *et al*. 2023). Meanwhile, to facilitate characterization of EMT pathways, several databases and tools have been developed to date (Zhao et al. 2015, Zhao et al. 2017, Vasaikar et al. 2021). However, these resources do not meet the increasingly complex gene set analyses for EMT. While general gene set enrichment methods are useful for quantifying EMT progression with predefined EMT gene sets, the research in this field has been challenged by several key issues. First, different EMT gene sets have been used in different studies (Liberzon et al. 2011, Tan et al. 2014, Carbon et al. 2021, Panchy et al. 2022), and these gene sets have very limited overlaps (Fig. 1B); the current gene set analysis tools do not allow easy access to diverse gene sets or fast tests with multiple gene sets. Secondly, discoveries and visualization of divergent activations of subsets of EMT genes require *ad hoc* methods with which it is difficult to reproduce key results (Youssef et al. 2024). Finally, there has not been an easy-to-use tool that allows the mapping from standard cell state labels to biologically meaningful EMT states.

Received: 24 January 2026. **Revised:** 11 April 2026. **Accepted:** 28 April 2026

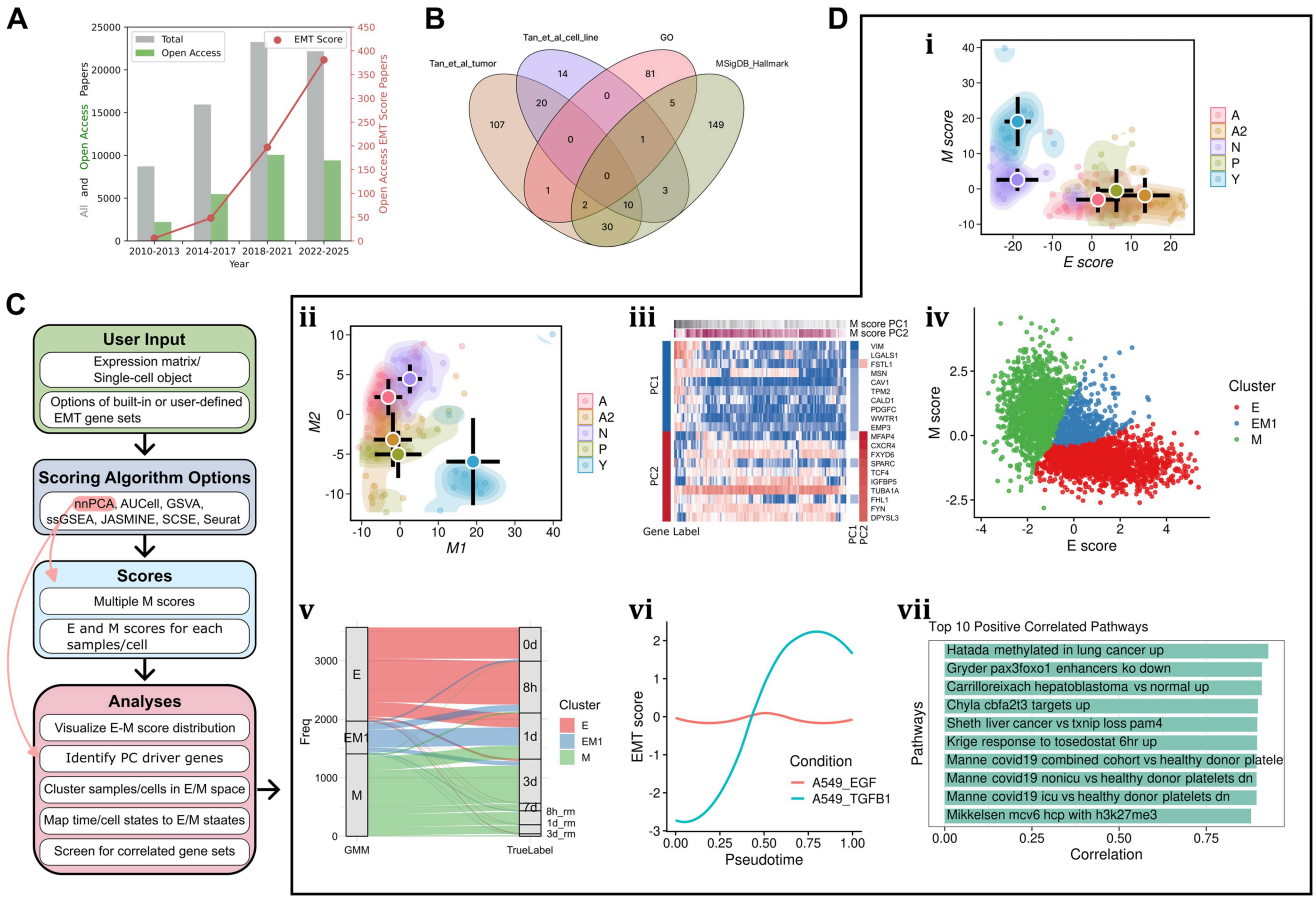

Figure 1 Motivation and overview of EMTscore. (A) EMT-related publication counts with respect to time. Articles with both keywords "epithelial" and "mesenchymal" mentioned in abstracts were extracted from PubMed as of December 31, 2025. Open access, original research articles were identified from this total pool. Among open access articles, papers with keyword "EMT score" were further identified. (B) Venn diagram showing the overlaps of four widely used EMT gene sets. (C) Workflow of EMT score package. (D) Example visualizations from EMScore. (i–iii) Analyses for 120 SCLC bulk RNA-seq samples. (i) nnPCA-based E and M scores. (ii) nnPCA-based M1 and M2 scores. (iii) Expression levels of top genes contributing to M1 and M2 scores. (iv–vii) scRNA-seq data for A549 cells treated with TGF-β or EGF. (iv) GMM based clustering of TGF-β treated cells. (v) Sankey plot showing the mapping of time points and EMT states from GMM clusters. (vi) E and M1 scores as functions of pseudotime for two types of treatments. (vii) Top 10 gene sets whose scores are positively correlated with EMT (M1) scores for TGF-β-treated cells.

We developed an R package named EMTscore that enables easy, simultaneous use of multiple popular EMT gene sets (Taube et al. 2010, Liberzon et al. 2011, Gröger et al. 2012, Tan et al. 2014, Cursons et al. 2015, Du et al. 2016, Foroutan et al. 2017, Cook and Vanderhyden 2020, The Gene Ontology Consortium 2021, Vasaikar et al. 2021, Panchy et al. 2022, Youssef et al. 2024) and scoring algorithms (Sigg and Buhmann 2008, Hänzelmann et al. 2013, Aibar et al. 2017, Pont et al. 2019, Noureen et al. 2022, Hao et al. 2024), as well as in-depth (single-cell) RNA-seq data analyses of multistate EMT and their visualizations with publication-quality plots. We implemented a method based on nonnegative principal component analysis (nnPCA) that enables automatic detection of groups of divergently activated EMT samples and genes in the same dataset. Additionally, the toolbox includes identification of EMT states in low-dimensional functional space and its mapping to user-defined cell clusters. Finally, our package can screen gene sets whose expressions are correlated with EMT progression, and this reveals the relationship between EMT and other biological processes efficiently.

2 Software description

2.1 Implementation of algorithms for computing EMT scores

To perform scoring of epithelial (E) or mesenchymal (M) gene set activities across transcriptomes (bulk or single-cell samples), EMTscore requires an expression dataset provided as a matrix file, a Seurat object, or a SingleCellExperiment (SCE) object. In addition, EMTscore provides several widely used E/M gene sets which can be replaced by custom gene sets provided by users (Fig. 1C, green).

To enable complex downstream analyses, we implemented multiple scoring algorithms that allow flexibility for choices of popular gene set enrichment methods. A default scoring algorithm based on nonnegative PCA (nnPCA) enables automated detection of multiple EMT programs (Sigg and Buhmann 2008, Panchy et al. 2021) (Fig. 1C, purple). Here, we briefly describe the advantages of the implemented methods. nnPCA is a fast method based on variance of data points (see Supplementary

Information, available as supplementary data at *Bioinformatics* online). Its unique advantage of providing multiple axes (e.g. multiple scores) with ranked variances explained makes it suitable for revealing multiple EMT programs in the same dataset. However, nnPCA does not provide a statistical test. In contrast, the widely used single-sample gene set enrichment analysis (ssGSEA) method uses Kolmogorov-Smirnov (KS) test for determining significant enrichment (Barbie et al. 2009), but the method is inefficient to handle large datasets such as scRNA-seq data. Other methods, such as GSVA (Hänzelmann et al. 2013), AUCell (Aibar et al. 2017), SCSE (Pont et al. 2019) and JASMINE (Noureen et al. 2022), are either based on gene ranks or normalized sums of expression (Aibar et al. 2017, Pont et al. 2019, Noureen et al. 2022). They are more efficient than ssGSEA but their statistical rigor varies with context, and they do not allow multiple gene scores from the same dataset.

EMTscore allows users to choose E and/or M gene sets from commonly used sources (Fig. 1B). A default set of E/M genes from Panchy et al. (2022) can be used. This collection combines a data-derived set of EMT genes and additional, widely recognized transcription factors controlling EMT. To enable easy access and application in a unified format, we compiled previously published EMT gene sets into a single GMT file included in this package.

2.2 Analysis of multiple EMT transcriptional programs

Upon completion of nnPCA-based scoring, EMTscore extracts multiple leading principal components (PCs) for M scores. Similar to conventional PCA, these PCs are ranked based on their variances explained. For some analyses, one M axis (i.e. the leading PC based on a M gene set) is used, and each sample/cell receives one M score and one E score (Fig. 1D, i). To detect divergent EMT (M) programs, two or more PCs from M scores are used. Typically, the first M PC (M1) and the second M PC (M2) can be used to describe the two directions of M programs driving the variance of the data points (Fig. 1D, ii). Genes whose expressions contribute to M1 and M2 can be extracted based on their rotations (loadings) (Fig. 1D, iii).

EMTscore visualizes the progression of EMT in the space of E-M for conventional EMT scores, and in the space of M1–M2 for divergent EMT programs. Scatter plots with various options are provided for these visualizations (Fig. 1D, i and ii). Note that our package allows user-defined gene sets with an input GMT file, so the capabilities of detecting divergent transcriptional programs can be generalized to study other pathways.

2.3 Detection of EMT states with Gaussian Mixture Model

In RNA-seq data, samples are sometimes with experimental labels (e.g. time points). In addition, for scRNA-seq datasets that typically contain a large amount of samples, clustering information is readily available from commonly used packages such as Seurat (Stuart et al. 2019). However, the relationship between these cluster labels and EMT states is often unclear. To provide insights into this question, EMTscore uses Gaussian Mixture Model (GMM) to cluster cells in E-M space and the clusters are interpretable biologically. Next, the package detects the extreme E and M subpopulations as well as hybrid EMT subpopulations relative to the dataset (Fig. 1D, iv). A mapping of the samples' original labels to EMT states is visualized (Fig. 1D, v) and the numerical distributions are reported. Single-cell level labels, such as pseudotime inferred from other packages, can be analyzed in relation to EMT scores (Fig. 1D, vi).

2.4 Screening for cellular processes correlated with EMT

EMT has been shown to have crosstalk with other important cellular processes such as alteration of motility and invasiveness (Jolly and Huang 2014, Wong et al. 2014). It is therefore of interest to search for gene set scores that are correlated and/or anti-correlated with EMT scores. Leveraging the high efficiency of scoring methods such as nnPCA, we implemented a screening function for searching through MSigDB (7561 curated gene sets in C2), and ranking gene sets that are highly correlated/anti-correlated with EMT scores (nnPCA-based EMT scores are M1 scores by default) across samples/cells. EMTscore reports the top positively correlated gene sets and top negatively correlated gene sets separately as lists and bar charts illustrate their Pearson correlation coefficients (e.g. Fig. 1D, vii). A gene set overlapping threshold (default 30%) is used to filter out gene sets with too many overlapping genes with the input gene list.

3 An example analysis pipeline

We first analyzed a bulk RNA-seq dataset containing transcriptomes from 120 small cell lung cancer (SCLC) cell lines with various known tumor subtypes (A2, A, N, P, and Y) (Cerami et al. 2012, Groves et al. 2022). To investigate the relationship between the subtypes of SCLC and EMT states, we computed E and M scores with each scoring method, and we found that A2 was an extremely E-like subtype, whereas Y was extremely M-like (Fig. 1D, i and Fig. 1, available as supplementary data at *Bioinformatics* online). The EMT identities for other subtypes were not clear from a single M score. With nnPCA, we found that there was a divergent expression of subsets of M genes between Y and A/N subtypes (Fig. 1D, ii), and that each M direction involves upregulation of a distinct set of M genes. Further analysis with the two leading M PCs showed that M markers such as VIM contributed significantly to PC1 (M1) whereas M2 was supported by other M markers such as CXCR4 (Fig. 1D, iii). Both marker sets are recognized M genes (Mendez et al. 2010, Cheng et al. 2018), but they showed distinct patterns across different SCLC cells. Although VIM is widely used as a key EMT marker due to its role in promoting cell motility (Gilles et al. 1999), it also protects cells from nuclear deformation by increasing nuclear stiffness (Patteson et al. 2019). This latter effect can inhibit cell migration in constricted environments (Patteson et al. 2019). SCLC cells may therefore utilize diverse programs and cell phenotypes to enhance their invasiveness with high adaptiveness.

We next use a scRNA-seq dataset from a time-course experiment profiling EMT induction in 12 911 A549 cells treated with TGF-β (Cook and Vanderhyden 2020). With E-M scoring followed

by GMM-based clustering, we found that cells at the hybrid EMT state are mainly distributed at the transition time points of the experiments (Fig. 1D, iv and v), and that the progression of cell states over time was strongly correlated with the downregulation of E genes and upregulation of M genes with the treatment of TGF-β (Fig. 1D, v). This trend was also consistent with the progression of M scores (M1 from nnPCA) over pseudotime (Fig. 1D, vi, blue), which is more widely available from scRNA-seq datasets compared to true time labels. Interestingly, independent analysis of 12 435 A549 cells treated with EGF from the same study did not show such a trend (Fig. 1D, vi). This observation is consistent between nnPCA and other methods (Fig. 2, available as supplementary data at *Bioinformatics* online). Finally, we identified several gene sets whose scores were strongly positively/negatively correlated with M (i.e. EMT) scores (Fig. 1D, vii and Fig. 3, available as supplementary data at *Bioinformatics* online).

In addition to the examples related to cancer progression, we analyzed a scRNA-seq dataset for trunk neural crest (Soldatov et al. 2019), and we found that the E-M scoring from EMTscore not only had reasonable agreement with development trajectories but also gave insights into potential transient expression of some M genes during development (Fig. 4, available as supplementary data at *Bioinformatics* online).

4 Conclusions

EMTscore provides flexible, versatile and easy-to-use functionality that allows in-depth analyses with gene set scores for EMT progression and publication-quality visualizations. The package is suitable for both bulk and sing-cell omics datasets. A default nnPCA-based method enables identification of multiple, divergent EMT pathways that are supported by different subsets of M genes, and this method is efficient for screening pathways that are associated with EMT. Future developments of this package will provide more biological insights enabled by combination of EMTscore-based cell states and other computational methods such as intercellular communication inference (Jin et al. 2021, Lopez et al. 2025).

Acknowledgements

The authors thank the members of the Tian Hong lab for critical reading of the manuscript.

Author contributions

Haimei Wen (Methodology [equal], Software [equal], Writing—original draft [equal], Writing—review & editing [equal]), Leonidas Bleris (Methodology [equal], Writing—original draft [equal], Writing—review & editing [equal]), and Tian Hong (Project administration [equal], Resources [equal], Supervision [equal], Writing—original draft [equal], Writing—review & editing [equal])

Supplementary material

Supplementary material is available at *Bioinformatics* online.

Conflicts of interest

None declared.

Funding

This work was supported by the National Institutes of Health [R35GM149531 awarded to T.H.], and the National Science Foundation [2243562 awarded to T.H.]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability

EMTscore is available from GitHub https://github.com/wenmm/EMTscore under the GNU General Public License, and is uploaded on Zenodo with a DOI 10.5281/zenodo.19487376.

References

Aibar S, González-Blas CB, Moerman T *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* 2017;**14**:1083–6.

Barbie DA, Tamayo P, Boehm JS *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* 2009;**462**:108–12.

Carbon S *et al.* The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res* 2021;**49**:D325–34.

Cassier PA, Navaridas R, Bellina M *et al.* Netrin-1 blockade inhibits tumour growth and EMT features in endometrial cancer. *Nature* 2023;**620**:409–16.

Cerami E, Gao J, Dogrusoz U *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2012;**2**:401–4.

Cheng Y, Song Y, Qu J *et al.* The chemokine receptor CXCR4 and c-MET cooperatively promote epithelial-mesenchymal transition in gastric cancer cells. *Transl Oncol* 2018;**11**:487–97.

Cook DP, Vanderhyden BC. Context specificity of the EMT transcriptional response. *Nat Commun* 2020;**11**:2142.

Cursons J, Leuchowius K-J, Waltham M *et al.* Stimulus-dependent differences in signalling regulate epithelial-mesenchymal plasticity and change the effects of drugs in breast cancer cell lines. *Cell Commun Signal* 2015;**13**:26.

Du L, Yamamoto S, Burnette BL *et al.* Transcriptome profiling reveals novel gene expression signatures and regulating transcription factors of TGFβ-induced epithelial-to-mesenchymal transition. *Cancer Med* 2016;**5**:1962–72.

Elamin YY, Robichaux JP, Carter BW *et al.* Poziotinib for EGFR exon 20-mutant NSCLC: clinical efficacy, resistance mechanisms, and impact of insertion location on drug sensitivity. *Cancer Cell* 2022;**40**:754–67.e6.

Foroutan M, Cursons J, Hediyeh-Zadeh S *et al.* A transcriptional program for detecting TGFβ-induced EMT in cancer. *Mol Cancer Res* 2017;**15**:619–31.

Gilles C, Polette M, Zahm JM *et al.* Vimentin contributes to human mammary epithelial cell migration. *J Cell Sci* 1999;**112**:4615–25.

Gröger CJ, Grubinger M, Waldhör T *et al.* Meta-analysis of gene expression signatures defining the epithelial to mesenchymal transition during cancer progression. *PLoS One* 2012; **7**:e51136.

Groves SM, Ildefonso GV, McAtee CO *et al.* Archetype tasks link intratumoral heterogeneity to plasticity and cancer hallmarks in small cell lung cancer. *Cell Syst* 2022;**13**:690–710.e617.

Groves SM, Panchy N, Tyson DR *et al.* Involvement of epithelial-mesenchymal transition genes in small cell lung cancer phenotypic plasticity. *Cancers (Basel)* 2023;**15**:1477.

Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* 2013;**14**:7.

Hao Y, Stuart T, Kowalski MH *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* 2024;**42**:293–304.

Hong T, Watanabe K, Ta CH *et al.* An Ovol2-Zeb1 mutual inhibitory circuit governs bidirectional and multi-step transition between epithelial and mesenchymal states. *PLoS Comput Biol* 2015;**11**:e1004569.

Jin S, Guerrero-Juarez CF, Zhang L *et al.* Inference and analysis of cell-cell communication using CellChat. *Nat Commun* 2021; **12**:1088–20.

Jolly MK, Huang B. et al. Towards elucidating the connection between epithelial–mesenchymal transitions and stemness. *J R Soc Interface* 2014;**11**:20140962.

Liberzon A, Subramanian A, Pinchback R *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 2011;**27**:1739–40.

Lopez D, Tyson DR, Hong T. Intercellular signaling reinforces single-cell level phenotypic transitions and facilitates robust re-equilibrium of heterogeneous cancer cell populations. *Cell Commun Signal* 2025;**23**:386.

Lu M, Jolly MK, Levine H *et al.* MicroRNA-based regulation of epithelial-hybrid-mesenchymal fate determination. *Proc Natl Acad Sci USA* 2013;**110**:18144–9.

Mendez MG, Kojima S-I, Goldman RD. Vimentin induces changes in cell shape, motility, and adhesion during the epithelial to mesenchymal transition. *FASEB J* 2010;**24**:1838–51.

Nakaya Y, Sheng G. EMT in developmental morphogenesis. *Cancer Lett* 2013;**341**:9–15.

Noureen N, Ye Z, Chen Y *et al.* Signature-scoring methods developed for bulk samples are not adequate for cancer single-cell RNA sequencing data. *Elife* 2022;**11**:e71994.

Panchy N, Watanabe K, Hong T. Interpretable, scalable, and transferrable functional projection of large-scale transcriptome data using constrained matrix decomposition. *Front Genet* 2021;**12**:719099.

Panchy N, Watanabe K, Takahashi M *et al.* Comparative single-cell transcriptomes of dose and time dependent epithelial-mesenchymal spectrums. *NAR Genom Bioinf* 2022;**4**:lqac072.

Pastushenko I, Brisebarre A, Sifrim A *et al.* Identification of the tumour transition states occurring during EMT. *Nature* 2018; **556**:463–8.

Patteson AE, Pogoda K, Byfield FJ *et al.* Loss of vimentin enhances cell motility through small confining spaces. *Small* 2019; **15**:e1903180.

Patteson AE, Vahabikashi A, Pogoda K *et al.* Vimentin protects cells against nuclear rupture and DNA damage during migration. *J Cell Biol* 2019;**218**:4079–92.

Pont F, Tosolini M, Fournié JJ. Single-cell signature explorer for comprehensive visualization of single cell signatures across scRNA-seq datasets. *Nucleic Acids Res* 2019;**47**:e133.

Sigg CD, Buhmann JM. Expectation-maximization for sparse and non-negative PCA. In: *Proceedings of the 25th International Conference on Machine Learning*. Helsinki, Finland: Association for Computing Machinery, 2008, 960–7.

Soldatov R, Kaucka M, Kastriti ME *et al.* Spatiotemporal structure of cell fate decisions in murine neural crest. *Science* 2019; **364**:eaas9536.

Stuart T, Butler A, Hoffman P *et al.* Comprehensive integration of single-cell data. *Cell* 2019;**177**:1888–902.e21.

Tan TZ, Miow QH, Miki Y *et al.* Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO Mol Med* 2014;**6**:1279–93.

Taube JH, Herschkowitz JI, Komurov K *et al.* Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. *Proc Natl Acad Sci USA* 2010;**107**:15449–54.

The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res* 2021;**49**:D325–34.

Thiery JP, Acloque H, Huang RYJ *et al.* Epithelial-mesenchymal transitions in development and disease. *Cell* 2009; **139**:871–90.

Vasaikar SV, Deshmukh AP, den Hollander P *et al.* EMTome: a resource for pan-cancer analysis of epithelial-mesenchymal transition genes and signatures. *Br J Cancer* 2021;**124**:259–69.

Wong IY, Javaid S, Wong EA *et al.* Collective and individual migration following the epithelial–mesenchymal transition. *Nat Mater* 2014;**13**:1063–71.

Yang J, Antin P, Berx G *et al.*; EMT International Association (TEMTIA). Guidelines and definitions for research on epithelial–mesenchymal transition. *Nat Rev Mol Cell Biol* 2020; **21**:341–52.

Youssef KK, Narwade N, Arcas A *et al.* Two distinct epithelial-to-mesenchymal transition programs control invasion and inflammation in segregated tumor cell populations. *Nat Cancer* 2024;**5**:1660–80.

Zhang J, Tian X-J, Zhang H *et al.* TGF-β -induced epithelial-to-mesenchymal transition proceeds through stepwise activation of multiple feedback loops. *Sci Signal* 2014;**7**:ra91.

Zhao M, Kong L, Liu Y *et al.* dbEMT: an epithelial-mesenchymal transition associated gene resource. *Sci Rep* 2015;**5**:11459.

Zhao Z, Zhou W, Han Y *et al.* EMT-Regulome: a database for EMT-related regulatory interactions, motifs and network. *Cell Death Dis* 2017;**8**:e2872.