Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

SCMG: A Multi-Omics and Stage-Resolved Single-Cell Framework for Cancer Driver Gene Prioritization

Shumei Zhang

zhangshumei@nefu.edu.cn

Northeast Forestry University

Dejian Xing

Northeast Forestry University

Yang Li

Northeast Forestry University

Weiwei Du

Northeast Forestry University

Yue Zhang

Northeast Forestry University

Yuxuan Wang

Northeast Forestry University

Wen Yang

Shenzhen University General Hospital

Research Article

Keywords: cancer driver gene, bulk multi-omics, single-cell RNA sequencing, graph neural network, gene interaction network, pan-cancer analysis

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11313159/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

SCMG: A Multi-Omics and Stage-Resolved Single-Cell Framework for Cancer Driver Gene Prioritization

Shumei Zhang[1*], Dejian Xing[1], Yang Li[1],Weiwei Du[1], Yue Zhang[1],Yuxuan Wang[1],Wen Yang[2*]

[1] School of Computer Science and Artificial Intelligence, Northeast Forestry University, Harbin, 150040, China,

[2] International Medical Center, Shenzhen University General Hospital, Shenzhen, 518000, China

*Correspondence and requests for materials should be addressed to Shumei Zhang (zhangshumei@nefu.edu.cn), Correspondence may also be addressed to Wen Yang (13159850336@163.com).

Abstract

Background: Cancer driver gene prioritization largely relies on bulk molecular profiles, leaving epithelial transcriptional changes across disease stages underused.

Results: We present Single-Cell Multi-omics Graph (SCMG), a graph-learning framework integrating bulk mutation, DNA methylation and gene expression with stage-resolved epithelial single-cell data and biological interaction networks. SCMG outperformed five comparison methods across six networks and achieved the highest area under the precision–recall curve in nine of 15 cancer types and the highest area under the receiver operating characteristic curve in seven. Ablation analyses supported contributions from single-cell features, network propagation and multimodal integration. External catalogue evaluation showed enrichment for driver genes outside the training positive-label set. Combining prediction scores with a background-corrected measure of stage-associated expression change organized high-scoring candidates into two groups with contrasting expression patterns. Their gene-expression signatures were associated with patient outcomes, while an external cutaneous squamous cell carcinoma cohort supported disease-associated expression of selected candidates.

Conclusions: SCMG provides a framework for integrating stage-resolved epithelial context into cancer driver gene prioritization and interpreting candidate rankings.

Keywords: cancer driver gene; bulk multi-omics; single-cell RNA sequencing; graph neural network; gene interaction network; pan-cancer analysis

Background

Cancer initiation and progression are accompanied by the continued accumulation of molecular abnormalities and the remodeling of cellular states, but not all molecular events provide a selective advantage to a cell or clone. Genes that drive tumor initiation, maintenance, and progression are generally referred to as driver genes. Identifying these genes can help clarify oncogenic mechanisms and is also directly relevant to the prioritization of candidate therapeutic targets and the advancement of precision medicine research[1, 2].

Early computational methods mainly focused on recurrent mutations or signals of positive selection. MutSigCV[3] identifies significantly mutated genes by accounting for differences in background mutation rates among genes. Oncodrive-CLUST[4] identifies candidate driver genes from non-random clustering of mutation positions in protein sequences. dNdScv[5] uses a selection model to assess evidence of positive selection in coding mutations. These methods have relatively clear statistical interpretations, but background mutation rates are affected by factors including gene length, replication timing, chromatin state, and heterogeneity among cancer types. They may therefore have limited sensitivity for low-frequency driver genes or genes that act only in specific molecular contexts.

Network-based methods further use the tendency of cancer genes to act together in common pathways, protein complexes, or functional modules, allowing weak abnormal signals distributed across individual genes to be interpreted in an interaction context. For example, HotNet2[6] uses network diffusion to identify significant subnetworks in which rare somatic mutations cluster across pan-cancer data. Network information can reduce the limitations of

relying only on mutation frequencies in individual genes, although performance remains affected by network coverage, tissue specificity, and interaction noise. MUFFINN[7] integrates mutation information for an individual gene with that of its neighbors in a functional network to identify driver genes.

In recent years, graph neural networks have provided a more flexible representation-learning framework for the joint modeling of multi-omics features and gene relationships. EMOGI[8] inputs multi-omics gene features and a protein interaction network into a graph convolutional network. MTGCN[9] uses multi-task learning to jointly exploit node-classification and network-structure information. SMG[10] applies self-supervised masked graph learning to enhance gene representations. HGDC[11] and SGCD[12] subsequently improved message propagation through heterophilic graph diffusion and simplified heterophilic graph learning, respectively, whereas deepCDG[13] explored deeper graph convolution for multi-omics integration. TGCN[14] combined Transformer-based global representation learning with Chebyshev graph convolution for cancer driver gene identification. More recently, MNGCL[15] and MMGN[16] introduced multi-network graph contrastive learning or multiplex-network representations into pan-cancer driver gene identification. These studies indicate that multi-omics evidence and network topology provide complementary information. Single-cell information has also been incorporated into driver-gene prioritization; for example, scCDX combines scRNA-seq with bulk multi-omics and network-derived features using extreme gradient boosting[17]. SCMG focuses on incorporating stage-associated epithelial summaries and cross-stage differences into a dual-branch graph-learning framework.

Despite the expansion of modeling strategies, existing methods still rely mainly on bulk

molecular features measured at the time of tumor sampling, such as mutation burden, DNA methylation alterations, and differences in gene expression. For epithelial tumors with Normal, Precancer, and Cancer stages, single-cell transcriptomic data can additionally provide information on epithelial-cell expression levels, the proportion of expressing cells, and mean expression among expressing cells within each stage. These stage-resolved gene-level features can serve as auxiliary inputs beyond bulk multi-omics and network topology to organize and interpret high-scoring candidate genes[18–20].

Dataset selection prioritized joint coverage of Normal, Precancer and Cancer states within an epithelial-cell context. We therefore used three datasets meeting these criteria and retained both stage and source identities during feature construction. These data provide gene-level summaries of epithelial expression and cross-stage variation to complement bulk molecular alterations and network relationships.

Based on these considerations, we developed SCMG, a multi-omics graph-learning framework for cancer driver gene prioritization in a pan-cancer setting. SCMG uses bulk multi-omics features and biological networks as its predictive backbone and introduces auxiliary transcriptional context from three-stage epithelial single-cell data. We aimed to assess whether this integration strategy could improve candidate driver gene prioritization without changing the primary task and provide an additional interpretive dimension related to epithelial malignant progression for high-scoring candidates.

Methods

Data sources

We extracted gene-level mutation-frequency, DNA-methylation, and gene-expression features from 16 cancer-type cohorts in The Cancer Genome Atlas (TCGA)[21]. Each omics type formed a 16-dimensional feature block. Before concatenation, features with different scales were min-max normalized across genes for each cancer-type feature column. The three feature blocks were then concatenated into a 48-dimensional continuous feature vector. The detailed construction of the bulk features is described in Additional file 1, Supplementary Methods S1.

Single-cell RNA-sequencing data were obtained from GSE251990, GSE161277 and GSE218170 to provide epithelial expression profiles across Normal, Precancer and Cancer states[19, 22–24]. For GSE161277 and GSE218170, cells with at least 500 detected genes, at least 1,000 UMIs and no more than 20% mitochondrial-gene content were retained[25], and doublets identified by scDblFinder were removed[26]. Harmony was applied to PCA representations for sample integration and clustering[27], and epithelial cells were selected using SingleR[28] with the HumanPrimaryCellAtlasData reference[29]. Harmony-corrected representations were used only for integration, clustering and cell-type annotation, whereas gene-expression matrices were retained for downstream model-input construction. For GSE251990, the supplied processed epithelial expression matrices and stage annotations were used. We randomly sampled 1,000 epithelial cells from each source–stage combination, yielding 9,000 cells for source-aware feature construction[30, 31]. Dataset-specific processing

and stage assignments are described in Supplementary Methods S2, and cell counts are summarized in Supplementary Table S1.

To capture topological information at different levels, including physical interactions, functional associations, pathway co-occurrence, and gene-gene relationships, we compiled and constructed six gene-level interaction networks (GINs): CPDB[32], STRINGdb[33], IRefIndex[34], PCNet[35], PathNet, and GGNet. CPDB and STRINGdb were filtered using database-provided confidence thresholds (CPDB > 0.5 and STRINGdb > 0.85)[8]. IRefIndex was strictly restricted to human binary interactions. Following the EMOGI processing strategy, no further confidence filtering was applied to PCNet. PathNet was constructed from curated pathway relationships in KEGG[36] and Reactome[37]. GGNet was constructed from gene-gene interaction information in the Encyclopedia of RNA Interactomes[38]. All node identifiers were mapped to official HGNC symbols. Unless otherwise stated, STRING, IRef, and PATHNET below refer to STRINGdb, IRefIndex, and PathNet, respectively. The numbers of nodes, edges, and supervised labels in the six networks are summarized in Supplementary Table S2.

For the supervised classification task, we followed the EMOGI pan-cancer all-label construction strategy[8]. Positive examples were obtained from NCG known cancer genes[39] and pan-cancer cancer-associated genes in DigSEE[40] supported by mutation, methylation, or expression evidence, and were restricted to the final aligned gene set. The COSMIC Cancer Gene Census[41] was used as one resource for conservative background exclusion and was not used as an additional source of positive examples. After the positive examples were removed, genes with evidence from OMIM, COSMIC, KEGG, or the NCG candidate cancer

gene set were excluded from the background set. This background set therefore served mainly as a conservative reference for the prioritization task rather than as a set of rigorously validated true negatives.

After model prediction, we further integrated external expression, clinical, and driver-gene resources for functional and patient-level association analyses of candidate genes. Patient-level analyses used TCGA Pan-Cancer gene-expression, clinical-phenotype, and survival-follow-up data provided by UCSC Xena[42], including overall survival (OS), progression-free interval (PFI), and disease-specific survival (DSS) endpoints[43]. In addition, we used the IntOGen[44] driver catalogue as a later driver-gene catalogue to assess the coverage of additional driver genes in the model ranking.

An external cSCC single-cell dataset (GSE193304) was used to assess candidate-gene expression in three tumor-matched-normal donor pairs (P4 – P6), using the authors' Basal1 and Basal2 annotations[45].

Overview of SCMG

As shown in Figure 1, we developed SCMG, a graph-learning framework for cancer driver gene prioritization. The framework integrates bulk multi-omics features, stage-resolved single-cell RNA-seq information, and gene-level interaction networks (GINs) to represent molecular abnormalities in individual genes, stage change during cancer progression, and topological or functional relationships among genes. GINs provide a prior representation of gene relationships and structural support for graph-aware representation learning, allowing the model to capture potential network-propagation signals beyond individual-gene features.

SCMG separately learns complementary evidence from bulk multi-omics data and stage-resolved scRNA data. The bulk branch combines omics-specific molecular features with gene-interaction topology to produce a bulk driver logit, whereas the scRNA branch models gene-level stage-source summaries and expression changes across Normal, Precancer, and Cancer to produce an scRNA driver logit. The two modality-specific branches are first optimized independently using weighted binary cross-entropy. Their parameters are subsequently frozen, and a learnable linear late-logit fusion layer is trained to combine the branch-specific logits into the final cancer driver score.

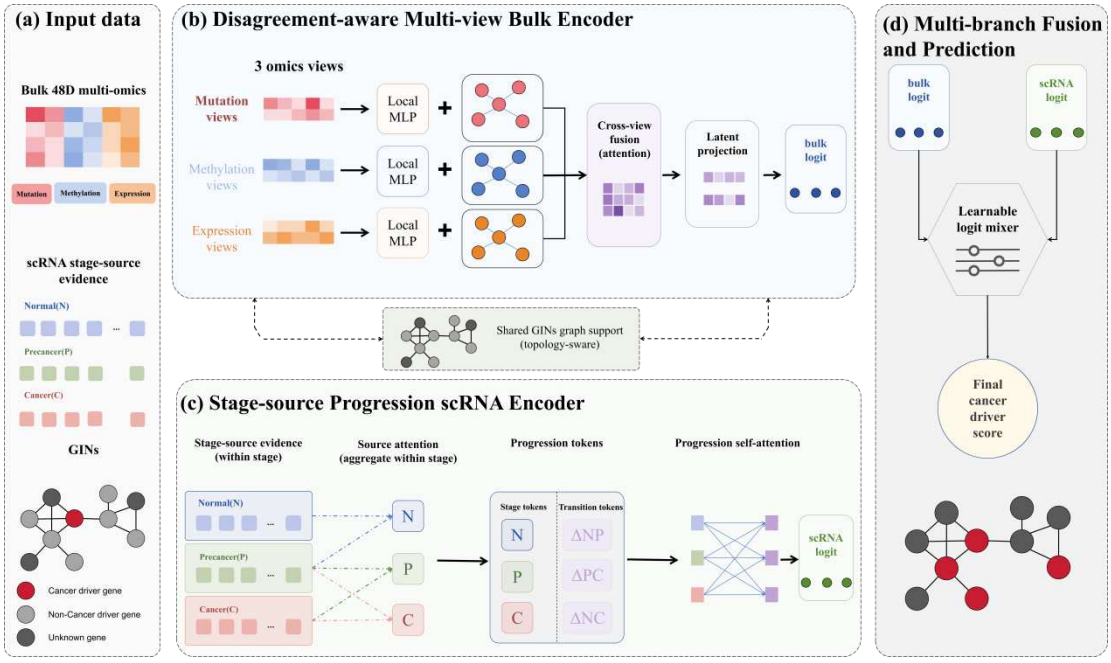


Figure 1. Overview of the SCMG framework for cancer driver gene prioritization. (A) SCMG integrates three types of gene-level evidence: 48-dimensional bulk multi-omics features, stage-source single-cell evidence from Normal, Precancer, and Cancer states, and gene-level interaction networks. (B) In the bulk branch, mutation, methylation, and expression features are processed by view-specific encoders and combined with graph-structure information from the gene interaction network for cross-omics representation learning, producing the bulk-branch prediction. (C) In the scRNA branch, the model uses single-cell source features from the Normal, Precancer, and

Cancer stages, aggregates within-stage expression information, and models changes between adjacent disease stages to produce the single-cell-branch prediction. (D) The bulk- and scRNA-branch predictions are integrated through a learnable fusion module to obtain the final cancer driver gene score for candidate driver gene prioritization.

Disagreement-aware bulk multi-view encoder

For each omics view, SCMG combines gene-local features with graph-propagated information:

$$G_\theta(X,G)=\mathrm{ReLU}(XW_L + b_L)+\exp(a_G)GCN_\theta(X,G). \quad (1)$$

Here, X is the node-feature matrix and $G = (V, E)$ is the gene interaction graph. The local projection has weights W_L and bias b_L; a_G is a learned scalar controlling the graph contribution. GCN_θ denotes graph convolution. For omics view m, the representation of gene i is

$$t_i^m=[G_{\theta_m}(X^m,G)]_i+\lambda_m[G_{\theta_0}(X^m,G)]_i+\rho_m U_m(x_i^m)+e_m \quad (2)$$

$$\lambda_m = \mathrm{sigmoid}(\alpha_m), \rho_m = \tanh(\kappa_m). \quad (3)$$

Here, X^m contains the features of view m, x_i^m is its row for gene i, and $[\cdot]_i$ selects the corresponding encoded row. G_{θ_m} and G_{θ_0} are view-specific and shared encoders, respectively. U_m is a view-specific two-layer adapter, α_m and κ_m are learned scalars, and e_m is a learned view-identity embedding. Signed cross-view features are then constructed and integrated as follows.

$$\Delta_i^{ab}= t_i^a - t_i^b, \ \ (a, b) \in \{(1,2), (1,3), (2,3)\} \quad (4)$$

$$r_i^{view} = R_\theta\left(\frac{\Delta_i^{12} + \Delta_i^{13} + \Delta_i^{23}}{3}\right) \quad (5)$$

$$h_i^B = F_\theta\left(t_i^1 \oplus t_i^2 \oplus t_i^3 \oplus \Delta_i^{12} \oplus \Delta_i^{13} \oplus \Delta_i^{23} \oplus r_i^{view}\right) \quad (6)$$

Here, $\oplus$ denotes concatenation, R_θ is a multilayer perceptron encoding the mean signed difference, and F_θ is a multilayer perceptron integrating the view and relation features. The embeddings have a common dimension and are jointly optimized through the downstream objective; their differences are learned interaction features, not calibrated distances between measurements in the original omics units.

The fused representation is projected using a graph-aware block combining local and graph-derived evidence. A gate receives their concatenation and absolute difference and modulates the graph contribution. The bulk logit combines the graph-supported classifier output with learned auxiliary readouts of the bulk representations and features.

Stage-source scRNA progression encoder

Single-cell data provide epithelial expression information resolved by disease stage and source dataset. For gene i, stage $s \in \{N, P, C\}$ and source $r \in \{1,2,3\}$, SCMG constructs a four-component expression-summary token:

$$p_{i,s,r} = \left[\mu_{i,s,r}^{\log}, d_{i,s,r}, q_{i,s,r}, \log\left(1 + \frac{c_{i,s,r}}{n_{s,r} + \epsilon}\right)\right], \quad \epsilon = 10^{-6} \quad (7)$$

Here, $\mu_{i,s,r}^{\log}$ is the mean log-normalized expression across cells, $d_{i,s,r}$ is the fraction of cells with non-zero expression, and $q_{i,s,r}$ is the mean log-normalized expression among expressing cells, set to zero when no cells express the gene. The quantities $c_{i,s,r}$ and $n_{s,r}$ are the summed expression in the source input layer and the number of cells, respectively, for the indicated

gene, stage and source. N, P and C denote Normal, Precancer and Cancer. The source index identifies GSE251990, GSE161277 and GSE218170, with a consistent ordering across stages.

Each token coordinate was standardized across training genes within each fold, and the fitted transformation was applied to all genes. Below, a tilde denotes a standardized token. Stage and source embeddings preserve token identity during encoding:

$$u_{i,s,r}=W_p\tilde{p}_{i,s,r}+b_p+a_s+b_r, \qquad d_i^T=O_T[\mathrm{MeanPool}\left(\mathrm{Transformer}_{\theta_T}\left(\{u_{i,s,r}\}_{s,r}\right)\right)] \tag{8}$$

W_p and b_p define a learned linear projection, while a_s and b_r are learned stage and source embeddings. MeanPool averages the nine Transformer output tokens. O_T applies layer normalization followed by a linear projection.

To construct stage-level representations, SCMG applies attention across sources within each stage:

$$\omega_{i,s,r}=\mathrm{softmax}_r\left(w^T\phi_p\left(\tilde{p}_{i,s,r}\right)\right), \qquad g_{i,s}=\sum_{r=1}^{3}\omega_{i,s,r}\,\phi_p\left(\tilde{p}_{i,s,r}\right) \tag{9}$$

Here, ϕ_p comprises a linear layer, layer normalization and GELU activation; w is a learned attention vector. Softmax is evaluated across the three sources separately for each gene and stage. Three stage representations and their pairwise differences form six progression tokens:

$$Q_i=\left[g_{i,N},g_{i,P},g_{i,C},g_{i,P}-g_{i,N},g_{i,C}-g_{i,P},g_{i,C}-g_{i,N}\right] \tag{10}$$

$$H_i = Transformer_{\theta_p}(Q_i + E_{\mathrm{prog}}) \tag{11}$$

$$a_{i,k} = softmax_k(v^T H_{i,k}) \tag{12}$$

$$d_i^P = O_P\left(\sum_{k=1}^{6} a_{i,k}\, H_{i,k}\right) \quad (13)$$

E_{prog} contains learned embeddings identifying the six token positions, $H_{i,k}$ is the output for token k, and v is a learned attention query. The weights $a_{i,k}$ define attention pooling over the six tokens; O_P applies layer normalization and a linear projection. The direct and progression Transformers have separate parameters.

The hybrid encoder combines direct stage-source and progression representations through a learned gate:

$$d_i^H = d_i^{T,H} + \gamma A_\theta(d_i^{T,H} \oplus d_i^P) \odot d_i^P, \ \gamma = \text{sigmoid}(\alpha) \quad (14)$$

Here, $d_i^{T,H}$ is produced by the direct encoder within the hybrid branch. This encoder has the same architecture as the standalone direct encoder producing d_i^T, but the two encoders have independently learned parameters.Here, $A_\theta$denotes a learned feature-wise gating network with a sigmoid output, α is a trainable scalar, and $\odot$ denotes element-wise multiplication.

Static-dynamic scRNA integration

The static vector s_i contains 13 expression summaries: three stage-level mean-expression values, three detection rates, three expressing-cell means, three pairwise stage differences in mean expression, and one cross-source direction-consistency summary. Mean-expression and expressing-cell summaries were source-centred before averaging across sources. Direction consistency summarizes agreement in the signs of source-specific changes across the three stage comparisons. The 13 features were standardized using training-fold statistics.

For each scRNA branch $b \in \{T, H\}$, a static encoder and feature-wise linear modulation

(FiLM) combine static and dynamic information:

$$e_i^b = E_{\theta_b}(s_i) \tag{15}$$

$$(\eta_i^b, \beta_i^b) = M_{\theta_b}(e_i^b) \tag{16}$$

$$z_i^b = S_{\theta_b}(e_i^b) + [1 + 0.1tanh(\eta_i^b)] \odot P_{\theta_b}(d_i^b) + 0.1\beta_i^b \tag{17}$$

Here, T and H identify the standalone direct and hybrid branches, respectively. E_{θ_b} is a multilayer perceptron encoding the static vector, S_{θ_b} and P_{θ_b} are linear projections, and M_{θ_b} generates the modulation vectors η_i^b and β_i^b. The two branches have separately learned parameters. Each z_i^b is mapped to a logit by a residual classifier, and the two logits are averaged to obtain ℓ_i^S.

Late-logit fusion and training

SCMG combines the bulk and scRNA logits through a learned linear fusion:

$$\ell_i = w_B \ell_i^B + w_S \ell_i^S + b_0, \qquad \hat{y}_i = \sigma(\ell_i) \tag{18}$$

Here, ℓ_i^B and ℓ_i^S are the bulk and scRNA logits, respectively; w_B, w_S and b_0 are learned fusion parameters; and $\sigma(x) = 1/(1 + e^{-x})$ is the logistic sigmoid. The output $\hat{y}_i$ is the prioritization score for gene i. The bulk and scRNA branches were optimized separately using weighted binary cross-entropy w ith logits. For each cross-validation fold, the positive-class weight was calculated from the ratio of negative to positive training genes and clipped to the range 0.5–5.0. The trained modality-specific branches were subsequently frozen, and the linear late-logit fusion predictor was optimized to integrate their scores. Predictive

performance was evaluated using stratified five-fold cross-validation, with evaluation restricted to the held-out labelled genes in each fold.

Calculation of epithelial alteration score

To characterize stage-associated epithelial expression changes after model prediction, we calculated an epithelial alteration score (EAS) for each gene.EAS was used only for post hoc interpretation and was not included as a model input or training target.

Let $\mu_{g,s}$ and $p_{g,s}$ denote the mean expression and proportion of expressing epithelial cells for gene g at stage $s \in \{N,P,C\}$, where N, P, and C indicate Normal, Precancer, and Cancer, respectively. Expressing-cell proportions were clipped to [0.005, 0.995] before applying $\mathrm{logit}(p) = \log[p/(1-p)]$ to avoid infinite values.

The Normal-to-Precancer and Precancer-to-Cancer stage-change vectors were defined as

$$d_{NP,g} = \left(\mu_{g,P} - \mu_{g,N},\ \mathrm{logit}\left(p_{g,P}\right) - \mathrm{logit}\left(p_{g,N}\right)\right) \tag{19}$$

and

$$d_{PC,g} = \left(\mu_{g,C} - \mu_{g,P},\ \mathrm{logit}\left(p_{g,C}\right) - \mathrm{logit}\left(p_{g,P}\right)\right) \tag{20}$$

EAS was calculated as

$$EAS_g = \log_2\left(\frac{||d_{PC,g}||_2 + \tau}{||d_{NP,g}||_2 + \tau}\right) \tag{21}$$

Here, $\|\cdot\|_2$ denotes the Euclidean norm, and τ is the fifth percentile of the positive norms pooled across both transitions. Positive EAS indicates a larger Precancer-to-Cancer change, whereas negative EAS indicates a larger Normal-to-Precancer change. Total transition extent was defined as

$$(extent_g = \|d_{NP,g}\|_2 + \|d_{PC,g}\|_2.) \quad (22)$$

To reduce variation associated with the overall magnitude of stage change, network connectivity and expression coverage, a matched-background correction was applied to EAS. Genes with a maximum stage-specific expressing-cell proportion of at least 1% and a total transition extent at or above the tenth percentile were considered eligible for matching. Total transition extent, CPDB node degree and maximum expressing-cell proportion were each divided into eight percentile bins.

$$r_g = EAS_g - median_{j \in M(g)}(EAS_j) \quad (23)$$

Here, r_g is residual EAS, and $M(g)$ is the matched background set for gene g. When fewer than five eligible genes were available in a matched bin, matching was progressively relaxed from extent–degree–detection matching to extent–detection matching, extent-only matching, and finally the global background median. Residual EAS measures a gene's stage-transition balance relative to the matched background defined above. Together with the SCMG score, it provides a coordinate system for comparing high-scoring candidates across epithelial disease states.

Post hoc validation and interpretation analyses

For downstream analyses, CPDB prediction scores were obtained by averaging gene-level predictions from the five cross-validation models. These scores were combined with residual EAS for candidate ranking and interpretation.

For external driver-catalog validation, genes overlapping the training positive labels were removed from the later IntOGen catalogue to define new-only drivers. AUROC and average

329 precision were calculated, and Top-K enrichment was evaluated using Fisher's exact tests
330 followed by Benjamini–Hochberg correction.

331 Among eligible genes, the top 5% of the prediction-score distribution were defined as
332 high-scoring candidates. Genes with residual EAS $r_g \geq 0$ were assigned to Q1 and those
333 with $r_g < 0$to Q2. Q1 and Q2 denote operational regions of the joint prediction-score–
334 residual-EAS space rather than temporal classes of driver genes. For functional and
335 patient-level analyses, genes within each region were ranked by prediction score and the first
336 80 genes were retained. Exact prediction-score ties were resolved by residual EAS, in
337 descending order for Q1 and ascending order for Q2, followed by gene symbol. The resulting
338 Q1 and Q2 signatures are listed in Supplementary Table S4. Functional over-representation
339 analysis was performed using Hallmark[46], KEGG[36] and GO[47] Biological Process gene
340 sets, with the model-scored gene universe as background. Hypergeometric-test P values
341 were adjusted within each gene-set collection using the Benjamini–Hochberg procedure.

342 Representative genes for the heatmaps were selected using stage-specific expression patterns,
343 residual EAS and expression-detection criteria; this display selection did not alter the Q1/Q2
344 regions or the 80-gene signatures used in subsequent analyses.

345 Patient-level Q1/Q2 scores were calculated as the mean z-standardized expression of matched
346 signature genes. OS Kaplan–Meier analyses used global-median groups and log-rank tests.
347 Associations with OS, PFI and DSS were estimated separately for each score using
348 unpenalized Cox models stratified by cancer type and adjusted for age and sex. Hazard ratios
349 were reported per within-cancer standard deviation, with Benjamini–Hochberg correction

across six primary tests. Sensitivity analyses and proportional-hazards diagnostics are described in Supplementary Table S6.

External cSCC expression analysis used GSE193304 and all three available tumor–matched-normal donor pairs (P4–P6). The fixed candidate list comprised the 15 highest-scoring eligible non-training-positive genes from each of Q1 and Q2, without replacement for non-estimable genes. Raw RNA counts from author-annotated Basal1 and Basal2 cells were pooled within each sample and analyzed using edgeR expression filtering and TMM normalization followed by limma–voom with a donor-paired design and robust empirical Bayes estimation. Benjamini–Hochberg correction was applied across 120 candidate–contrast tests, comprising 30 genes across four comparisons; non-estimable tests were assigned P = 1 only for adjustment. Donor-level logCPM differences and analyses within each basal subtype were used to describe consistency and assess sensitivity to subtype composition. Candidate selection and the four-contrast multiple-testing family are defined in Supplementary Methods S5. Supplementary Table S7 presents the fixed candidate list, cSCC sample inclusion, estimable cSCC effects and donor-level differences, whereas Supplementary Table S8 presents basal-subtype composition, subtype-specific effects and paired differences.

Implementation details of SCMG

SCMG was implemented in Python 3.9 using PyTorch and PyTorch Geometric. The 48-dimensional bulk multi-omics input was divided into three 16-dimensional omics views. In the bulk branch, each view was processed by a graph-aware encoder with a hidden

dimension of 64. The fused multi-view representation was projected into a 100-dimensional latent space and processed by a graph-based predictor with a hidden dimension of 200 and one output unit. The cross-view relation module used a hidden dimension of 32. In the scRNA branch, the stage-source tokens were represented in a 64-dimensional latent space and processed using two Transformer layers with two attention heads. The scRNA representations were mapped to branch-specific predictions using a residual classifier containing two residual blocks. The two scRNA predictions were averaged to produce the final scRNA-branch logit. The scRNA dropout was set to 0.35, and the initial scRNA weight in the late-logit fusion layer was set to 0.10. SCMG was trained in three stages: bulk-branch pretraining for 1,200 epochs, scRNA-branch pretraining for 140 epochs, and late-logit fusion training for 100 epochs. The corresponding learning rates were 8×10^{-4}, 2.4×10^{-4}, and 1×10^{-3}, respectively. Adam was used with zero weight decay and a gradient-clipping threshold of 5. The bulk input dropout was 0.35, whereas the model dropout and edge dropout were both 0.55. Supervised classification experiments used stratified five-fold cross-validation.

For the ablation and control experiments in Table 3, results were summarized across five seed-level means using base seeds 101, 202, 303, 404 and 505. Each repeat used stratified five-fold cross-validation, with performance averaged arithmetically across the five held-out folds. Fold-specific training seeds were set to the base seed plus the zero-based fold index. Identical cross-validation splits were used across all seeds and model settings, with the split random state fixed at 42. Results are reported as the mean and sample standard deviation across the five seed-level means. Two-sided 95% confidence intervals for the mean were calculated as mean $\pm$ $t_{0.975,4} \times SD/\sqrt{5}$, with the sample SD calculated using four degrees of

freedom. Operational definitions of the ablation and control settings are provided in Additional file 1, Supplementary Table S9.

Results

SCMG performance in pan-cancer driver gene prioritization across heterogeneous network backbones

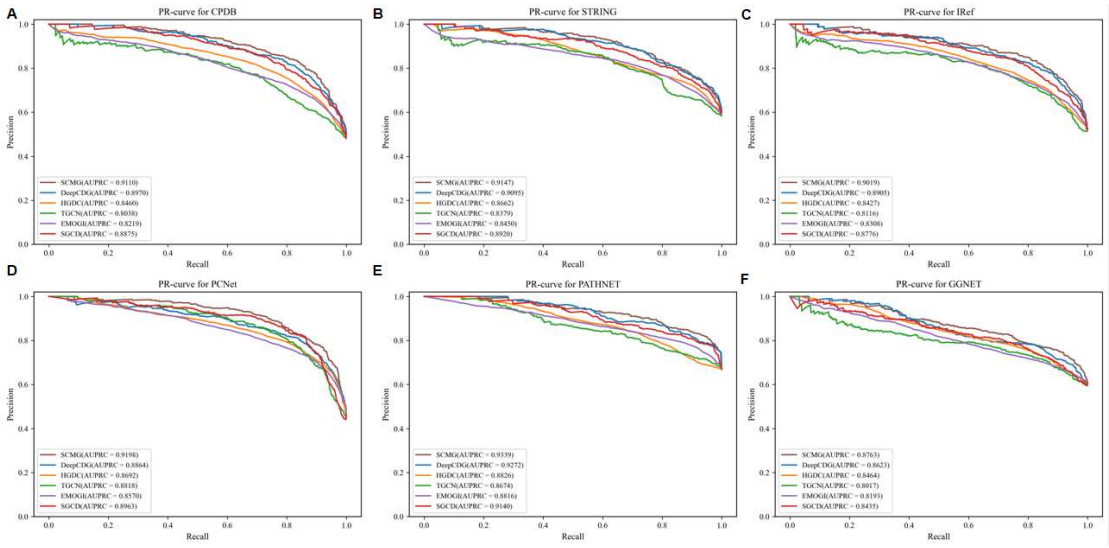


Figure 2. Comparison of precision-recall curves for SCMG and baseline methods across six biological networks. (A) PR curve for CPDB. (B) PR curve for STRING. (C) PR curve for IRef. (D) PR curve for PCNet. (E) PR curve for PATHNET. (F) PR curve for GGNET.

To evaluate SCMG for cancer driver gene prioritization in a pan-cancer setting, we compared it with five representative baseline methods across six network backbones: HGDC, EMOGI, DeepCDG, TGCN, and SGCD. We reimplemented the five baseline methods according to their original publications and evaluated them using the same processed bulk inputs, stratified five-fold splits, and network-specific data settings, with AUPRC as the primary metric. Figure

2 shows the precision-recall (PR) curves, and Table 1 reports the AUPRC values across all networks. SCMG achieved the highest AUPRC on CPDB, STRING, IRef, PCNet, PathNet, and GGNet, with values of 0.9110, 0.9147, 0.9019, 0.9198, 0.9339, and 0.8763, respectively. DeepCDG was the strongest-performing baseline on CPDB, STRING, IRef, PathNet and GGNet, with AUPRC values of 0.8970, 0.9095, 0.8905, 0.9272 and 0.8623, respectively. On PCNet, SGCD was the strongest-performing baseline, with an AUPRC of 0.8963. The 100 highest-scoring candidate genes from the main CPDB analysis are listed in Supplementary Table S5.

These results indicate that the performance of SCMG was not restricted to a single network type and was consistent across conventional protein-interaction networks and networks focused on pathways or gene-gene relationships. Although the magnitude of improvement differed among networks, SCMG consistently prioritized candidate driver genes under the unified evaluation setting, providing a basis for the subsequent interpretation of high-scoring candidates.

Table 1. AUPRC comparison of SCMG and baseline methods across six biological networks.

GINs	CPDB	STRING	IRef	PCNet	PATHNET	GGNET
SCMG	**0.9110**	**0.9147**	**0.9019**	**0.9198**	**0.9339**	**0.8763**
DeepCDG	0.8970	0.9095	0.8905	0.8864	0.9272	0.8623
HGDC	0.8460	0.8662	0.8427	0.8692	0.8826	0.8464
TGCN	0.8038	0.8379	0.8116	0.8818	0.8674	0.8017
EMOGI	0.8219	0.8450	0.8308	0.8571	0.8816	0.8193
SGCD	0.8875	0.8920	0.8776	0.8963	0.9141	0.8435

The best results are highlighted in bold.

Performance of cancer type-specific driver gene prioritization

We next compared SCMG with the other methods across 15 cancer types using the CPDB

network: LUAD, BRCA, BLCA, LIHC, CESC, COAD, ESCA, HNSC, KIRC, KIRP, LUSC, PRAD, STAD, THCA, and UCEC. Cancer-type-specific positive genes were defined using NCG6.0 annotations, and negative genes were drawn from the reference background set mapped to CPDB. For SCMG, only the corresponding mutation, methylation, and expression feature columns were retained for each cancer type, while the single-cell inputs were shared across cancer types. As shown in Table 2, SCMG achieved the highest AUPRC in nine cancer types (BRCA, BLCA, LIHC, CESC, COAD, KIRP, PRAD, STAD, and UCEC) and the highest AUROC in seven (BLCA, LIHC, CESC, COAD, KIRC, KIRP, and PRAD). The largest relative AUPRC improvement over the next-best method was observed for UCEC (0.7137 versus 0.6225; 14.65%), followed by COAD (6.85%) and STAD (5.65%). SCMG ranked highest for both metrics in BLCA, LIHC, CESC, COAD, KIRP, and PRAD. These results support the applicability of SCMG to cancer-type-specific prioritization, with performance gains varying across cancer types.

Table 2. AUPRC and AUROC of SCMG and baseline methods across 15 cancer types.

Ablation and control analyses of SCMG

We evaluated the contributions of bulk multi-omics, scRNA features, network propagation and fusion components (Table 3). On CPDB, incorporating single-cell features increased AUPRC from 0.9018 for the bulk-only model trained for 1,200 epochs to 0.9113 for full SCMG, an absolute gain of 0.0095. The corresponding AUROC values were 0.9076 and 0.9124, respectively. The scRNA-only model achieved an AUPRC of 0.7046 and an AUROC of 0.7291. All bulk feature-subset settings performed below the full model, with GE + METH providing the strongest two-omics combination.

Collapsing the single-cell stages reduced AUPRC to 0.9068. Zero-initializing the trainable stage embeddings yielded an AUPRC of 0.9078, while replacing the progression pathway with mean pooling of projected stage–source tokens yielded 0.9086. Random binary scRNA inputs yielded an AUPRC of 0.9055, while fixed fusion weights and removal of disagreement correction yielded 0.9090 and 0.9091, respectively. These comparisons support the contribution of stage-resolved epithelial information and adaptive integration to driver-gene prioritization alongside bulk multi-omics and network features.

Removing PPI propagation produced the largest reduction among the component ablations, yielding an AUPRC/AUC of 0.8092/0.8154. Degree-preserving network rewiring yielded 0.8786/0.8794, supporting a contribution from network connectivity beyond node degree in this control. The simpler modality-matched MLP+GCN baseline achieved 0.8309/0.8401, further supporting the contribution of SCMG's architecture under the evaluated settings.

Table 3. Ablation and control analyses of SCMG.

Methods	AUPRC mean±SD	AUPRC 95% CI	AUC mean±SD	AUC 95% CI
Full SCMG	0.9113±0.0012	[0.9098, 0.9128]	0.9124±0.0015	[0.9106, 0.9143]
Bulk-only (600 epochs)	0.8990±0.0016	[0.8970, 0.9010]	0.8997±0.0012	[0.8981, 0.9012]
Bulk-only (1200 epochs)	0.9018±0.0018	[0.8995, 0.9041]	0.9076±0.0019	[0.9053, 0.9100]
scRNA-only	0.7046±0.0044	[0.6992, 0.7100]	0.7291±0.0037	[0.7246, 0.7337]
Zero-initialized stage embeddings	0.9078±0.0010	[0.9065, 0.9091]	0.9108±0.0014	[0.9091, 0.9125]
Collapsed scRNA stages	0.9068±0.0009	[0.9057, 0.9079]	0.9082±0.0013	[0.9067, 0.9098]
Mean-pooling scRNA token control	0.9086±0.0008	[0.9076, 0.9095]	0.9100±0.0014	[0.9083, 0.9117]
Random binary scRNA control	0.9055±0.0015	[0.9036, 0.9073]	0.9066±0.0014	[0.9049, 0.9082]
Fixed late-logit fusion	0.9090±0.0010	[0.9077, 0.9102]	0.9101±0.0014	[0.9084, 0.9118]
w/o PPI propagation	0.8092±0.0020	[0.8067, 0.8117]	0.8154±0.0010	[0.8142, 0.8167]
w/o disagreement correction	0.9091±0.0013	[0.9076, 0.9107]	0.9102±0.0017	[0.9081, 0.9123]
GE only	0.8898±0.0010	[0.8885, 0.8911]	0.8875±0.0005	[0.8869, 0.8881]
MF only	0.8588±0.0023	[0.8559, 0.8617]	0.8672±0.0014	[0.8655, 0.8690]
METH only	0.8645±0.0053	[0.8578, 0.8711]	0.8724±0.0047	[0.8666, 0.8783]
GE + MF	0.9013±0.0011	[0.9000, 0.9027]	0.8983±0.0008	[0.8973, 0.8993]
GE + METH	0.9016±0.0008	[0.9006, 0.9025]	0.9003±0.0006	[0.8995, 0.9011]
MF + METH	0.8746±0.0024	[0.8716, 0.8775]	0.8852±0.0019	[0.8828, 0.8876]
Simple modality-matched MLP+GCN	0.8309±0.0026	[0.8277, 0.8341]	0.8401±0.0019	[0.8377, 0.8425]
Degree-preserving random PPI	0.8786±0.0023	[0.8758, 0.8814]	0.8794±0.0018	[0.8772, 0.8816]

Values are mean ± SD across five seed-level means, each obtained by averaging performance across five held-out folds. The 95% confidence intervals are two-sided Student's t intervals across the five seed-level means (df = 4).

Parameter-sensitivity analysis

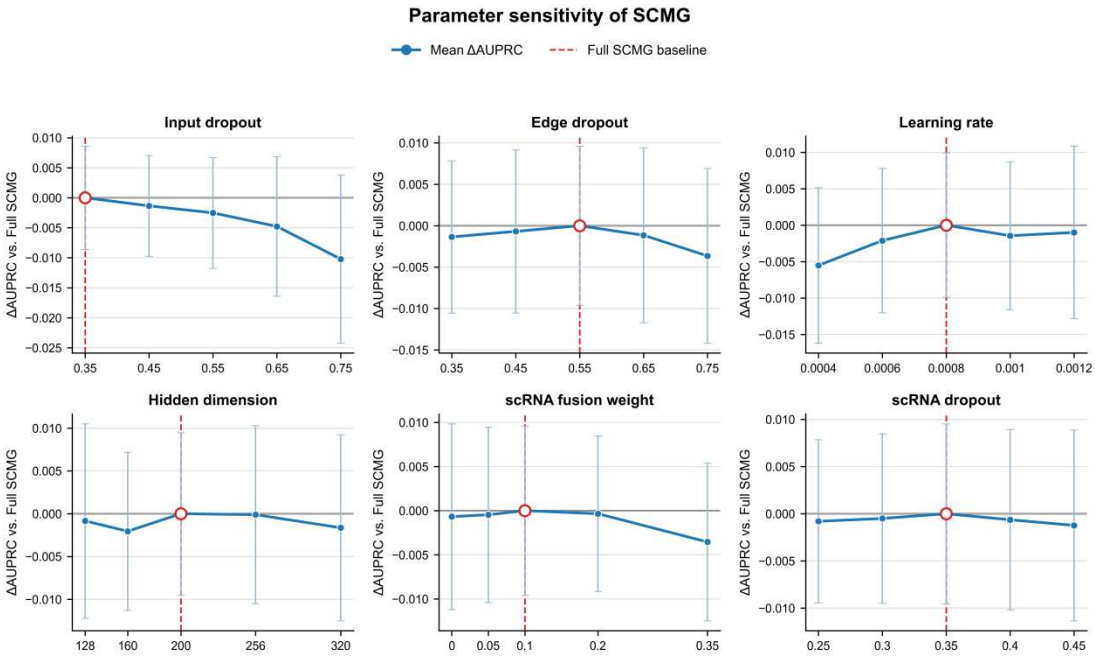


Figure 3. Parameter-sensitivity analysis of SCMG. Performance change is measured by ΔAUPRC relative to the Full SCMG baseline. Each panel varies one hyperparameter while keeping the others fixed. Points indicate the mean ΔAUPRC across five folds, and error bars denote the fold-wise standard deviation. The red dashed line marks the baseline setting used by Full SCMG.

To assess the stability of SCMG with respect to key hyperparameter choices, we performed one-factor-at-a-time sensitivity analyses around the full-model settings (Figure 3). SCMG remained stable across most reasonable parameter ranges, with only small fluctuations in AUPRC, indicating that model performance did not depend on a single specific hyperparameter value. Performance decreased more clearly when some parameters moved outside suitable ranges. For example, high bulk-input dropout and edge dropout weakened the input omics signals and graph-topology information, a low learning rate affected model optimization, and a high initial fusion weight for the scRNA branch could disrupt the balance between bulk and scRNA information.

Overall, the parameter-sensitivity analysis indicated that SCMG was stable under reasonable hyperparameter perturbations and that its performance depended on moderate regularization, a stable optimization process, and balanced cross-modal fusion.

External driver-catalogue validation

SCMG rankings were enriched for IntOGen driver genes outside the model's original positive-label set (Supplementary Table S3). After excluding the 791 positive-label genes from the 12,814-gene SCMG score universe, 12,023 genes remained for evaluation, including 138 IntOGen drivers. These external drivers showed an AUROC of 0.7720 and an average precision of 0.0465. The Top 500 remaining SCMG-ranked genes were significantly enriched for IntOGen drivers (OR = 9.40, 95% CI 6.39–13.81; FDR = 2.28×10^{-20}).

To assess the influence of TCGA-derived evidence represented in IntOGen, we repeated the analysis after excluding driver-call records from the 32 cohorts explicitly labelled as TCGA. Of the 271 cohorts listed in the release metadata, 260 were represented in the final driver-call table and 228 remained after this exclusion. This yielded 108 external drivers among the same positive-label-excluded SCMG evaluation universe. Enrichment remained similar (AUROC = 0.7718; average precision = 0.0402), with a Top-500 OR of 10.79 (95% CI 7.09–16.41; FDR = 2.89×10^{-19}).

In a stricter sensitivity analysis, we additionally excluded PCAWG-labelled cohort records because PCAWG incorporates data contributed by both ICGC and TCGA projects. This retained 101 external drivers supported by the 197 remaining cohorts. The corresponding AUROC was 0.7650 and average precision was 0.0369, with a Top-500 enrichment OR of

10.30 (95% CI 6.65–15.93; FDR = 3.81×10^{-17}). Similar enrichment was observed when the original 12,814-gene score universe was retained as a conservative comparison (Supplementary Table S3). Together, these analyses show that high SCMG rankings were enriched for IntOGen-defined drivers outside the original positive-label set, and that this enrichment was retained after exclusion of explicitly TCGA-labelled and PCAWG-labelled driver-call records.

Prediction score and residual EAS jointly organize high-scoring candidate genes

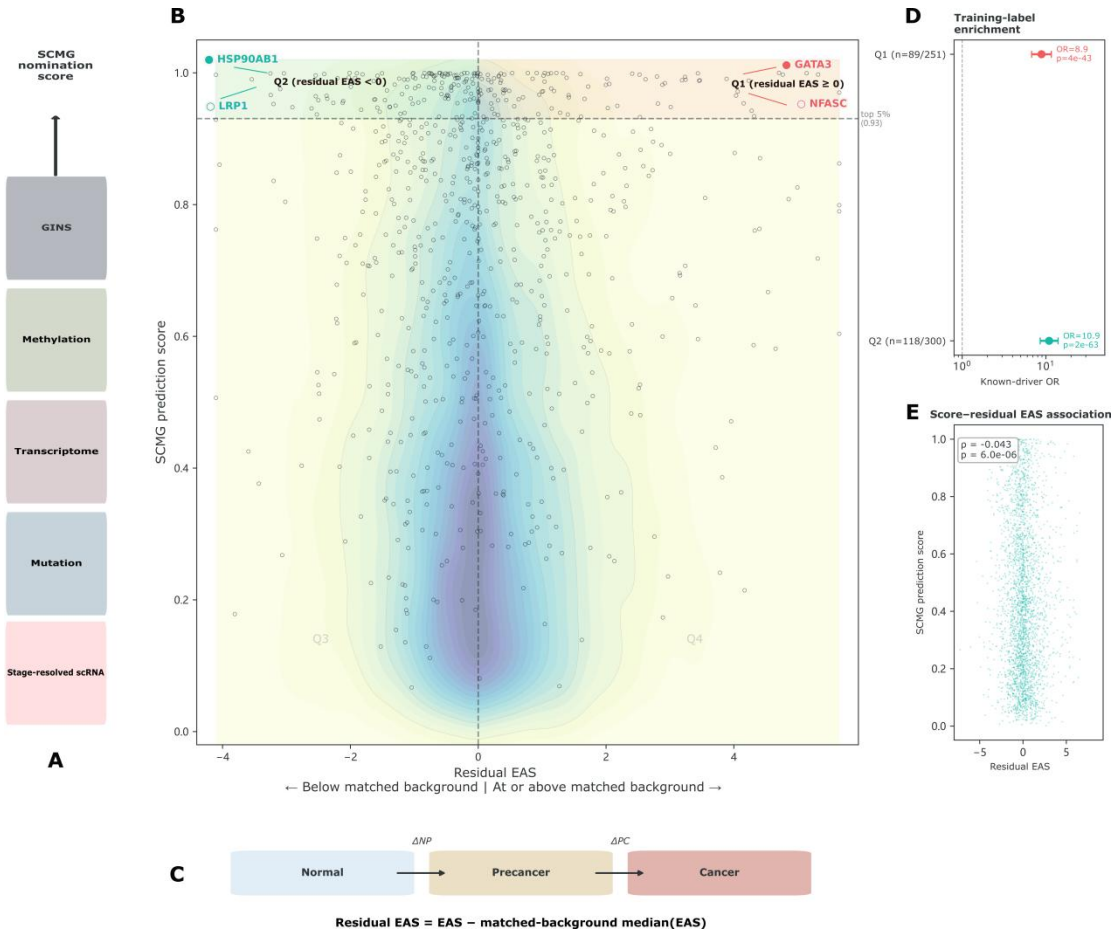


Figure 4. Joint organization of candidate genes by SCMG score and residual EAS. (A) Bulk multi-omics, stage-resolved single-cell expression and gene-interaction networks contribute to SCMG prediction scores. (B)

Distribution of eligible genes in the prediction-score–residual-EAS space. The horizontal dashed line marks the top 5% score threshold, and the vertical dashed line marks residual EAS = 0. High-scoring genes with residual EAS < 0 form Q2 (n = 300), and those with residual EAS ≥ 0 form Q1 (n = 251). (C) Normal–Precancer–Cancer comparisons and matched-background correction underlying residual EAS. (D) Enrichment of training-positive genes in Q1 and Q2 relative to the remaining eligible genes. Points and horizontal lines represent odds ratios and 95% confidence intervals. (E) Association between SCMG prediction score and residual EAS (Spearman ρ = −0.0431).

High-scoring candidates were divided into the Q2 and Q1 operational candidate regions according to residual EAS (Figure 4B). Among genes in the top 5% of prediction scores, Q2 contained 300 genes with residual EAS<0, and Q1 contained 251 genes with residual EAS≥0.

Both regions were enriched for known driver genes among the training labels (Figure 4D). Q2 contained 118/300 training-label genes, with OR=10.95 (95% CI, 8.56–14.00), whereas Q1 contained 89/251, with OR=8.86 (95% CI, 6.76–11.62). Residual-EAS permutation analysis did not support additional enrichment for known driver genes beyond that attributable to high-score selection (Q2 empirical P=0.152; Q1 empirical P=0.884).

The correlation between prediction score and residual EAS was weak (Spearman ρ=−0.0431; Figure 4E). Residual EAS therefore provided a stage-associated coordinate distinct from the model score for organizing and comparing high-scoring candidates.

Q1 and Q2 show different tendencies in stage patterns and functional distributions

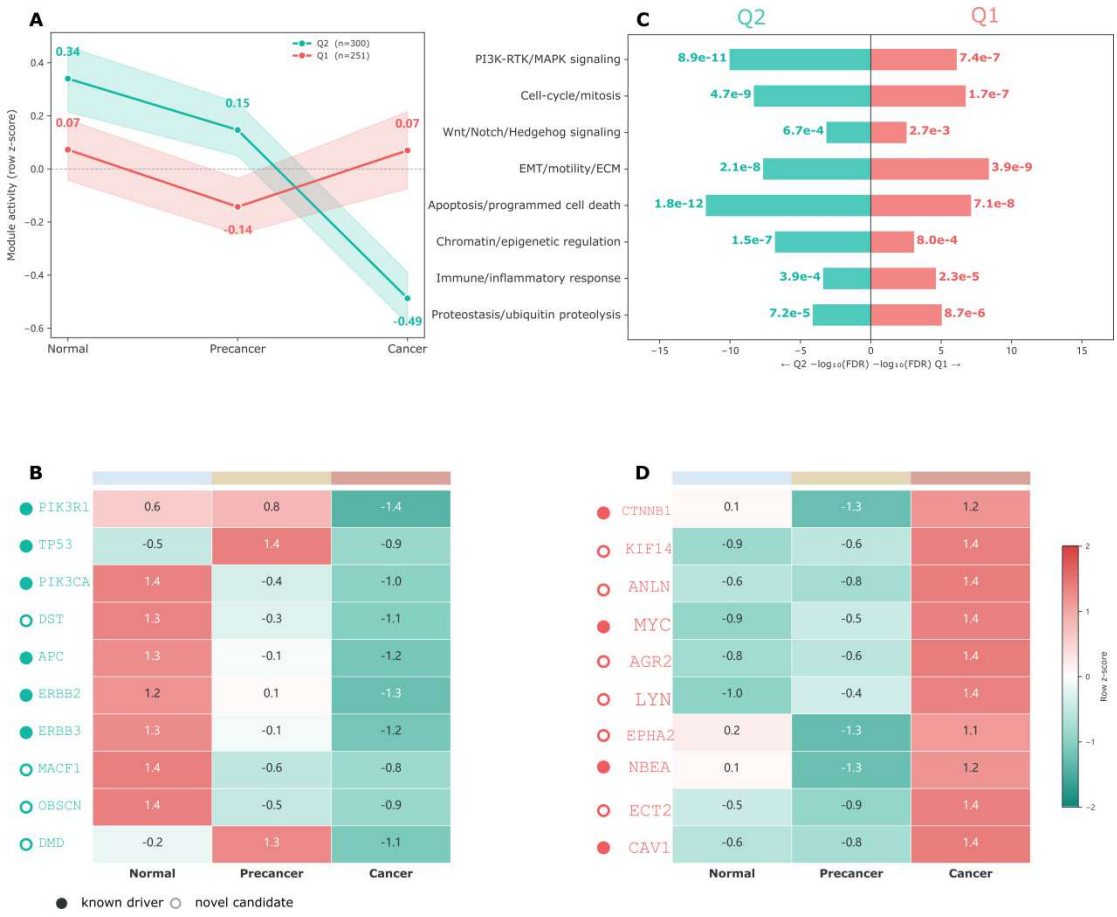


Figure 5. Stage-associated expression patterns and functional enrichment of Q2 and Q1 candidates. Q2 and Q1 comprise high-scoring genes with residual EAS < 0 and residual EAS ≥ 0, respectively. (A) Stage-level mean expression was log1p-transformed and standardized within each gene across Normal, Precancer and Cancer. Lines show the mean within-gene z-score across all eligible Q2 (n = 300) or Q1 (n = 251) genes; shaded areas show 95% intervals from 2,000 gene-level bootstrap resamples. (B) Ten representative Q2 genes selected for relatively lower expression in Cancer. Heatmap values are within-gene z-scores. Filled circles ("known driver") denote genes in the training positive-label set, whereas open circles ("novel candidate") denote prioritized genes outside that set. (C) Over-representation analysis of the predefined Q2 and Q1 80-gene signatures using GO Biological Process, KEGG and MSigDB Hallmark gene sets. Bars show −log10 of the smallest FDR within each displayed functional theme, following Benjamini–Hochberg correction within each gene-set collection. (D) Ten representative Q1 genes selected for relatively higher expression in Cancer using the display criteria described in the Methods.

Q2 and Q1 showed different mean three-stage expression patterns across their complete candidate regions (Figure 5A). The mean within-gene z-score for Q2 decreased from 0.34 in Normal to 0.15 in Precancer and then to −0.49 in Cancer. For Q1, the mean decreased from 0.07 in Normal to −0.14 in Precancer and then returned to 0.07 in Cancer. Thus, Q2 showed a continued decrease, whereas Q1 decreased in Precancer and recovered in Cancer.

Heatmaps displayed representative genes selected for their stage-associated expression patterns. PIK3R1, TP53, PIK3CA, APC and ERBB2 in Q2 had lower within-gene z-scores in Cancer (Figure 5B). CTNNB1, KIF14, ANLN, MYC, EPHA2 and ECT2 in Q1 had relatively higher values in Cancer (Figure 5D).

Both predefined 80-gene signatures were enriched for multiple cancer-related functions (Figure 5C). In both signatures, at least one term reached FDR<0.05 in each of the following themes: PI3K–RTK/MAPK signaling, cell cycle/mitosis, Wnt/Notch/Hedgehog signaling, EMT/motility/ECM, apoptosis/programmed cell death, chromatin/epigenetic regulation, immune/inflammatory response, and proteostasis/ubiquitin proteolysis. Bar length represents $-\log_{10}$ of the minimum FDR within each theme. The two signatures therefore showed broadly overlapping backgrounds of cancer-related functional enrichment.

Q1 and Q2 signature scores are associated with patient outcomes

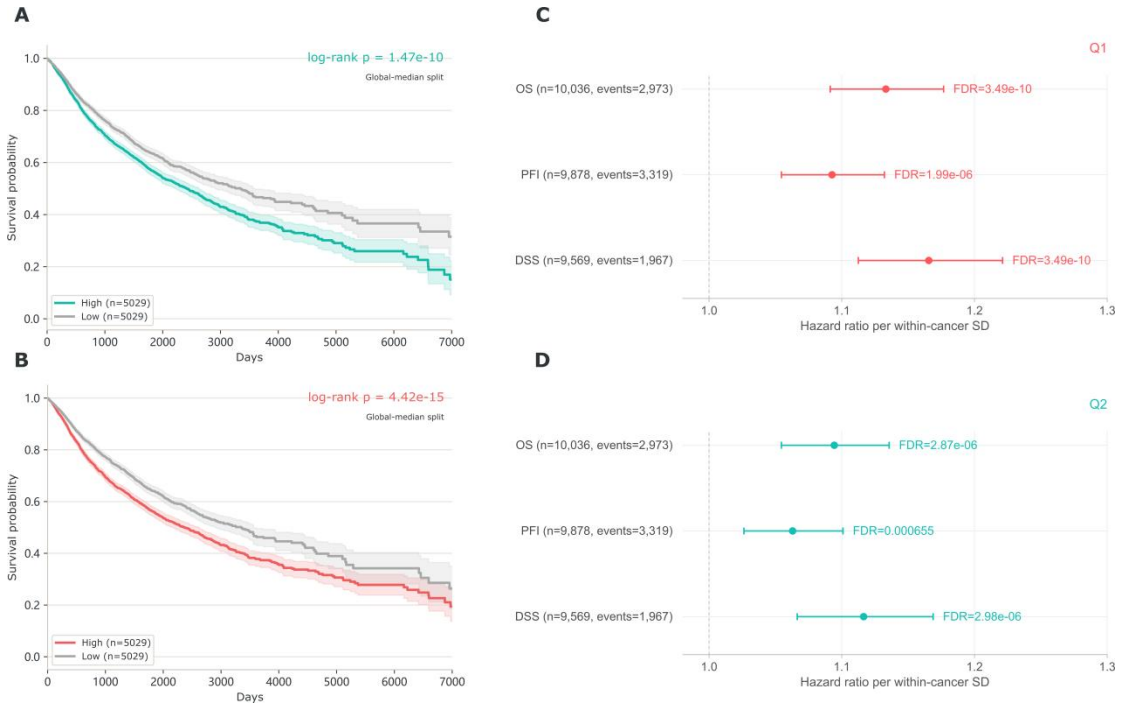


Figure 6. Exploratory associations of Q1 and Q2 signature scores with patient outcomes. (A, B) Pooled, unadjusted overall-survival (OS) Kaplan–Meier curves for Q2 (A) and Q1 (B). High- and low-score groups were defined separately for each score using its global median; the displayed log-rank P values are unadjusted. (C, D) Forest plots of separate single-signature Cox models for Q1 (C) and Q2 (D) across OS, progression-free interval (PFI), and disease-specific survival (DSS). Each unpenalized model included one score standardized within cancer type, age, and sex, with baseline hazards stratified by cancer type. Hazard ratios (HRs) are reported per one within-cancer standard deviation; horizontal lines show 95% confidence intervals. Patients with missing DSS event status were excluded from the DSS analysis. Benjamini–Hochberg false-discovery rates (FDRs) were calculated across the six primary score–endpoint tests. The descriptive Kaplan–Meier panels and Cox models have different eligible populations; the Kaplan–Meier curves are not covariate-adjusted survival predictions.

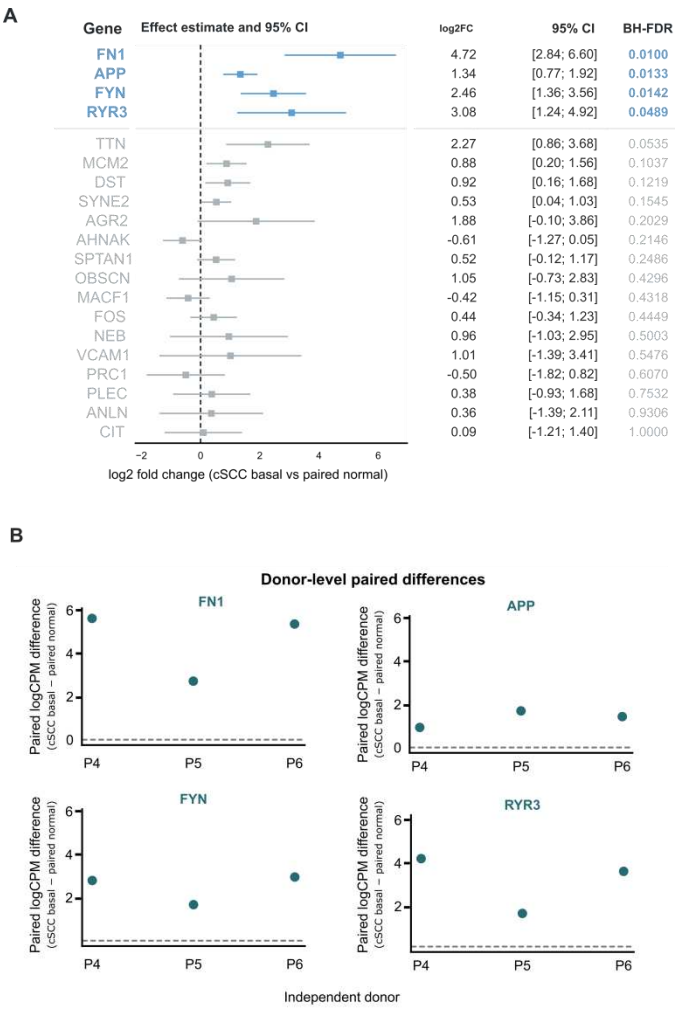
The pooled, unadjusted OS Kaplan–Meier curves showed lower observed survival in the high-score groups for Q2 and Q1 (log-rank P = 1.47×10^{-10} and 4.42×10^{-15}, respectively; Figure 6A,B). In separate Cox models stratified by cancer type and adjusted for age and sex, both scores

were positively associated with OS, PFI and DSS, with all six associations passing FDR correction (Figure 6C,D). For OS, the HR per within-cancer standard deviation was 1.133 for Q1 (95% CI, 1.091–1.177) and 1.094 for Q2 (95% CI, 1.054–1.136). These separate models do not estimate the scores' independent effects relative to one another.

In the stage-complete subset spanning 11 cancer types, only Q1–OS remained significant after FDR correction (HR = 1.115; 95% CI, 1.033–1.202; FDR = 0.030), limiting conclusions about stage-independent associations. Neither score-specific proportional-hazards diagnostics nor time-interaction tests detected departures in the six primary full-signature models.

After excluding training-label genes, Q2 remained associated with all three endpoints, although its association with PFI was weak (HR = 1.036; 95% CI, 1.001–1.073; FDR = 0.0415). Non-label Q1 showed departures from proportional hazards; the corresponding proportional-hazards diagnostics and time-interaction test results are reported in Supplementary Table S6.

Disease-associated expression of selected SCMG candidates in an external cSCC cohort



635

636 Figure 7. Disease-associated expression of selected SCMG candidates in an external cSCC cohort. (A)
637 Tumor-versus-matched-normal log2 fold changes for all 20 estimable genes from the fixed 30-gene candidate list,
638 based on donor-paired pseudobulk analysis of pooled, author-annotated Basal1 and Basal2 cells. Squares indicate
639 estimates and horizontal lines indicate pointwise 95% confidence intervals. The dashed line denotes zero effect;
640 blue indicates FDR < 0.05. FDR values were calculated across 120 candidate‑contrast tests, retaining
641 non-estimable tests as P = 1 for adjustment (Supplementary Methods S5). (B) Tumor-minus-matched-normal
642 logCPM differences for FN1, APP, FYN and RYR3. Each point represents one donor pair (n = 3; P4‑P6), and the
643 dashed line denotes zero difference. Both panels use the same donor pairs. Subtype-composition and sensitivity
644 results are provided in Supplementary Table S8.

We next examined disease-associated expression of genes from the fixed Q1/Q2 candidate list in an external cutaneous squamous cell carcinoma (cSCC) single-cell cohort (GSE193304). All three available tumor–matched-normal pairs (P4–P6) were included. Donor-paired pseudobulk analysis of pooled, author-annotated Basal1 and Basal2 cells yielded estimable results for 20 of the 30 candidates. FN1, APP, FYN and RYR3 showed positive tumor-versus-normal effects and FDR < 0.05 across the defined family of 120 candidate–contrast tests (Figure 7A; Supplementary Table S7). Paired logCPM differences for each of these four genes were positive in all three donors, indicating within-cohort directional consistency (Figure 7B).

The Basal1 proportion within the pooled basal compartment was higher in each tumor than in its matched normal sample. We therefore analyzed the two basal subtypes separately as a sensitivity check for subtype composition. APP and FN1 showed positive estimated effects within both basal subtypes. FYN showed a positive effect in Basal1 but was not retained by the expression filter in Basal2, whereas RYR3 was not retained in either subtype (Supplementary Table S8). These findings provide external expression-level support for a subset of SCMG-prioritized candidates in the author-annotated basal compartment of cSCC.

Discussion

Bulk molecular alterations and gene-interaction networks provide a strong basis for driver-gene prioritization. SCMG extends this representation with epithelial expression levels, detection frequencies and changes across disease states. The ablation results support an incremental predictive contribution from these single-cell features, while the 15-cancer comparison shows that the relative performance of SCMG varies across cancer types.

In the downstream analysis, the same stage-resolved expression resource enabled high-scoring genes to be organized by residual EAS and compared through their expression patterns and functional enrichment. The resulting Q1 and Q2 regions showed different mean stage-associated expression patterns against broadly overlapping backgrounds of cancer-related functional enrichment. Together, these analyses connect candidate ranking with epithelial disease-state context.

Q1 and Q2 signature scores were each associated with patient outcomes in separate cancer-type-stratified models adjusted for age and sex. These findings add a clinical association dimension to the candidate signatures, without estimating their independent effects or establishing associations independent of pathological stage.

The external cSCC analysis provides complementary expression-level support for selected candidates in the author-annotated basal compartment, with interpretation limited by three donor pairs and sparse expression in some subtype-level comparisons.

Several limitations remain. The single-cell inputs represent three available epithelial disease contexts rather than a unified pan-cancer progression trajectory. The negative labels form a conservatively selected background, and the interaction networks provide fixed priors. Clinical associations require external replication and more complete covariate adjustment, while candidate driver activity requires functional testing. Future work could incorporate broader single-cell cohorts and context-dependent gene interactions.

Conclusions

This study indicates that three-stage epithelial single-cell expression features can serve as

complementary inputs beyond bulk multi-omics and gene interaction networks for cancer driver gene prioritization. Through dual-branch late-logit fusion, SCMG jointly uses bulk multi-omics/GIN information and stage-source expression from epithelial cells in the Normal, Precancer, and Cancer stages to produce a gene-level prediction score. In the post-prioritization analysis, high-scoring candidates were organized into the Q2 and Q1 operational candidate regions by combining the SCMG prediction score with residual EAS. Mean stage-level module activity in Q2 decreased progressively from Normal to Cancer, whereas Q1 decreased in Precancer and recovered in Cancer. Both the Q1 and Q2 signatures were significantly enriched for cancer-related functional themes, including PI3K–RTK/MAPK signaling, cell cycle/mitosis, and EMT/motility/ECM. Q1 and Q2 signature scores were associated with patient outcomes when evaluated separately in cancer-type-stratified models adjusted for age and sex.

Thus, SCMG not only generates a ranking of candidate driver genes but also provides a subsequent comparison framework for high-scoring candidates based on three-stage epithelial expression features. This framework supports the selection of candidates for further biological investigation.

Declarations

Ethics approval and consent to participate

This study used only publicly available, de-identified datasets obtained from TCGA, GEO and UCSC Xena. No new participants were recruited and no new biological specimens were collected by the authors. Additional ethics approval and consent to participate were not required for this secondary analysis of publicly available, de-identified data.

Consent for publication

Not applicable.

Availability of data and materials

The SCMG source code and instructions for downloading the model-ready datasets are available at [https://github.com/mantouwang/SCMG]. Bulk molecular and patient-level resources were obtained from TCGA and UCSC Xena. The single-cell datasets are available from NCBI GEO under accession numbers GSE251990, GSE161277, GSE218170 and GSE193304 . The external driver catalogue was obtained from IntOGen. Supplementary methods and supporting results are provided in Additional file 1.

Project name: SCMG

Project home page: https://github.com/mantouwang/SCMG

Archived version: https://github.com/mantouwang/SCMG/releases/tag/v1.0.0 (tag: v1.0.0)

Operating system(s): Windows 10

Programming language: Python 3.9

Other requirements: PyTorch and PyTorch Geometric; additional dependencies are specified in the repository's requirements.txt file. A CUDA-enabled GPU is recommended.

License: MIT License

Any restrictions to use by non-academics: None.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China (No. 62371117 to S.Z.; No. 62302315 to W.Y.; No.62522106 to Y.L.), the Heilongjiang Province Postdoctoral Fund (LBH-Z23004 to S.Z.), the Natural Science Foundation of Heilongjiang Province (No. JQ2026F002 to Y.L.), and the Fundamental Research Funds for the Central Universities (2572025JT05-03 to S.Z.).

Authors' contributions

SZ: Shumei Zhang; DX: Dejian Xing; YL: Yang Li;WD: Weiwei Du; YZ: Yue Zhang; YW: Yuxuan Wang; WY: Wen Yang.

DX conceived and designed the study, collected and processed the datasets, developed and implemented the SCMG model, performed the computational analyses, interpreted the results, and drafted the manuscript. SZ, YL and WY supervised the study and critically reviewed and revised the manuscript. WD, YZ, and YW contributed to discussions of the study design and biological interpretation of the results. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Additional file

Additional file 1 (.docx): Supplementary methods and tables. Supplementary Methods S1 – S5 and Supplementary Tables S1 – S9 describe data processing, network statistics, external driver-catalogue evaluation, gene signatures, survival analyses, external candidate-expression analyses, and the implementation of ablation and control settings.

References

1. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer Genome Landscapes. Science. 2013;339:1546–58. https://doi.org/10.1126/science.1235122.

2. Hanahan D. Hallmarks of Cancer: New Dimensions. Cancer Discov. 2022;12:31–46. https://doi.org/10.1158/2159-8290.CD-21-1059.

3. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499:214–8. https://doi.org/10.1038/nature12213.

4. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics. 2013;29:2238–44. https://doi.org/10.1093/bioinformatics/btt395.

5. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell. 2017;171:1029-1041.e21. https://doi.org/10.1016/j.cell.2017.09.042.

6. Leiserson MDM, Vandin F, Wu H-T, Dobson JR, Eldridge JV, Thomas JL, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet. 2015;47:106–14. https://doi.org/10.1038/ng.3168.

7. Cho A, Shim JE, Kim E, Supek F, Lehner B, Lee I. MUFFINN: cancer gene discovery via network analysis of somatic mutation data. Genome Biol. 2016;17:129. https://doi.org/10.1186/s13059-016-0989-x.

8. Schulte-Sasse R, Budach S, Hnisz D, Marsico A. Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. Nat Mach Intell. 2021;3:513–26. https://doi.org/10.1038/s42256-021-00325-y.

9. Peng W, Tang Q, Dai W, Chen T. Improving cancer driver gene identification using multi-task learning on graph convolutional network. Brief Bioinform. 2022;23:bbab432. https://doi.org/10.1093/bib/bbab432.

10. Cui Y, Wang Z, Wang X, Zhang Y, Zhang Y, Pan T, et al. SMG: self-supervised masked graph learning for cancer gene identification. Brief Bioinform. 2023;24:bbad406. https://doi.org/10.1093/bib/bbad406.

11. Zhang T, Zhang S-W, Xie M-Y, Li Y. A novel heterophilic graph diffusion convolutional network for identifying cancer driver genes. Brief Bioinform. 2023;24:bbad137. https://doi.org/10.1093/bib/bbad137.

12. Li X, Xu J, Li J, Gu J, Shang X. Towards simplified graph neural networks for identifying cancer driver genes in heterophilic networks. Brief Bioinform. 2025;26:bbae691. https://doi.org/10.1093/bib/bbae691.

13. Wu Y, Xu J, Li J, Gu J, Shang X, Li X. Deep graph convolutional network-based multi-omics integration for cancer driver gene identification. Brief Bioinform. 2025;26:bbaf364. https://doi.org/10.1093/bib/bbaf364.

14. Chen Y, Li G, Liu T, Zhou R, Mao Y, Wu J, et al. A Deep Learning Framework for Identifying Cancer Driver Genes Based on Transformer and Graph Convolutional Network. IEEE Trans Comput Biol Bioinforma. 2025;22:1850–60.

https://doi.org/10.1109/TCBBIO.2025.3574337.

15. Peng W, Zhou Z, Dai W, Yu N, Wang J. Multi-Network Graph Contrastive Learning for Cancer Driver Gene Identification. IEEE Trans Netw Sci Eng. 2024;11:3430–40. https://doi.org/10.1109/TNSE.2024.3373652.

16. Li X, Li J, Hao J, Liao X, Li M, Shang X. Multiplex Networks and Pan-Cancer Multiomics-Based Driver Gene Identification Using Graph Neural Networks. Big Data Min Anal. 2024;7:1262–72. https://doi.org/10.26599/BDMA.2024.9020043.

17. Lee K, Kim J, Kim TY. scCDX: enhancing cancer driver gene identification and model interpretability with single-cell RNA sequencing data and extreme gradient boosting. 2024. https://doi.org/10.21203/rs.3.rs-4222777/v1.

18. Baysoy A, Bai Z, Satija R, Fan R. The technological landscape and applications of single-cell multi-omics. Nat Rev Mol Cell Biol. 2023;24:695–713. https://doi.org/10.1038/s41580-023-00615-w.

19. Zheng X, Song J, Yu C, Zhou Z, Liu X, Yu J, et al. Single-cell transcriptomic profiling unravels the adenoma-initiation role of protein tyrosine kinases during colorectal tumorigenesis. Signal Transduct Target Ther. 2022;7:60. https://doi.org/10.1038/s41392-022-00881-8.

20. Sun L, Kang X, Wang C, Wang R, Yang G, Jiang W, et al. Single-cell and spatial dissection of precancerous lesions underlying the initiation process of oral squamous cell carcinoma. Cell Discov. 2023;9:28. https://doi.org/10.1038/s41421-023-00532-4.

21. The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis

821 project. Nat Genet. 2013;45:1113–20. https://doi.org/10.1038/ng.2764.
822 22. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI
823 GEO: archive for functional genomics data sets—update. Nucleic Acids Res.
824 2013;41:D991–5. https://doi.org/10.1093/nar/gks1193.
825 23. Strasser MK, Gibbs DL, Gascard P, Bons J, Hickey JW, Schürch CM, et al. Concerted
826 epithelial and stromal changes during progression of Barrett's Esophagus to invasive
827 adenocarcinoma exposed by multi-scale, multi-omics analysis. 2023.
828 https://doi.org/10.1101/2023.06.08.544265.
829 24. Schütz S, Solé-Boldo L, Lucena-Porcel C, Hoffmann J, Brobeil A, Lonsdorf AS, et al.
830 Functionally distinct cancer-associated fibroblast subpopulations establish a tumor
831 promoting environment in squamous cell carcinoma. Nat Commun. 2023;14:5413.
832 https://doi.org/10.1038/s41467-023-41141-9.
833 25. Chen X, Liu Q, Wu J, Zhou P, Zhao M, Song J. DSS1 inhibits autophagy to activate
834 epithelial-mesenchymal transition in a pro-metastatic niche of renal cell carcinoma.
835 Nat Commun. 2025;16:6769. https://doi.org/10.1038/s41467-025-62135-9.
836 26. Germain P-L, Lun A, Meixide CG, Macnair W, Robinson MD. Doublet identification
837 in single-cell sequencing data using scDblFinder. F1000Research. 2022;10:979.
838 https://doi.org/10.12688/f1000research.73600.2.
839 27. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive
840 and accurate integration of single-cell data with Harmony. Nat Methods.
841 2019;16:1289–96. https://doi.org/10.1038/s41592-019-0619-0.
842 28. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of

lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20:163–72. https://doi.org/10.1038/s41590-018-0276-y.

29. Mabbott NA, Baillie J, Brown H, Freeman TC, Hume DA. An expression atlas of human primary cells: inference of gene function from coexpression networks. BMC Genomics. 2013;14:632. https://doi.org/10.1186/1471-2164-14-632.

30. Tao W, Yu Z, Han J-DJ. Single-cell senescence identification reveals senescence heterogeneity, trajectory, and modulators. Cell Metab. 2024;36:1126-1143.e5. https://doi.org/10.1016/j.cmet.2024.03.009.

31. Jiang J, Chen S, Tsou T, McGinnis CS, Khazaei T, Zhu Q, et al. D-SPIN constructs regulatory network models from scRNA-seq that reveal organizing principles of perturbation response. Cell. 2026;189:4159-4184.e35. https://doi.org/10.1016/j.cell.2026.04.028.

32. Kamburov A, Pentchev K, Galicka H, Wierling C, Lehrach H, Herwig R. ConsensusPathDB: toward a more complete picture of cell biology. Nucleic Acids Res. 2011;39 suppl_1:D712–7. https://doi.org/10.1093/nar/gkq1156.

33. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 2021;49:D605–12. https://doi.org/10.1093/nar/gkaa1074.

34. Razick S, Magklaras G, Donaldson IM. iRefIndex: A consolidated protein interaction database with provenance. BMC Bioinformatics. 2008;9:405. https://doi.org/10.1186/1471-2105-9-405.

865 35. Huang JK, Carlin DE, Yu MK, Zhang W, Kreisberg JF, Tamayo P, et al. Systematic
866 Evaluation of Molecular Networks for Discovery of Disease Genes. Cell Syst.
867 2018;6:484-495.e5. https://doi.org/10.1016/j.cels.2018.03.001.
868 36. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG:
869 integrating viruses and cellular organisms. Nucleic Acids Res. 2021;49:D545–51.
870 https://doi.org/10.1093/nar/gkaa970.
871 37. Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, et al. The
872 reactome pathway knowledgebase 2022. Nucleic Acids Res. 2022;50:D687–92.
873 https://doi.org/10.1093/nar/gkab1028.
874 38. Li J-H, Liu S, Zhou H, Qu L-H, Yang J-H. starBase v2.0: decoding miRNA-ceRNA,
875 miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data.
876 Nucleic Acids Res. 2014;42:D92–7. https://doi.org/10.1093/nar/gkt1248.
877 39. Dressler L, Bortolomeazzi M, Keddar MR, Misetic H, Sartini G, Acha-Sagredo A, et al.
878 Comparative assessment of genes driving cancer and somatic evolution in non-cancer
879 tissues: an update of the Network of Cancer Genes (NCG) resource. Genome Biol.
880 2022;23:35. https://doi.org/10.1186/s13059-022-02607-z.
881 40. Kim J, Kim J, Lee H. An analysis of disease-gene relationship from Medline
882 abstracts by DigSee. Sci Rep. 2017;7:40154. https://doi.org/10.1038/srep40154.
883 41. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer
884 Gene Census: describing genetic dysfunction across all human cancers. Nat Rev
885 Cancer. 2018;18:696–705. https://doi.org/10.1038/s41568-018-0060-1.
886 42. Goldman MJ, Craft B, Hastie M, Repečka K, McDade F, Kamath A, et al. Visualizing

887 and interpreting cancer genomics data via the Xena platform. Nat Biotechnol.
888 2020;38:675–8. https://doi.org/10.1038/s41587-020-0546-8.
889 43. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An
890 Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival
891 Outcome Analytics. Cell. 2018;173:400-416.e11.
892 https://doi.org/10.1016/j.cell.2018.02.052.
893 44. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C,
894 et al. A compendium of mutational cancer driver genes. Nat Rev Cancer. 2020;20:555–
895 72. https://doi.org/10.1038/s41568-020-0290-x.
896 45. Zou D-D, Sun Y-Z, Li X-J, Wu W-J, Xu D, He Y-T, et al. Single-cell sequencing
897 highlights heterogeneity and malignant progression in actinic keratosis and
898 cutaneous squamous cell carcinoma. eLife. 2023;12:e85270.
899 https://doi.org/10.7554/eLife.85270.
900 46. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The
901 Molecular Signatures Database Hallmark Gene Set Collection. Cell Syst. 2015;1:417–25.
902 https://doi.org/10.1016/j.cels.2015.12.004.
903 47. The Gene Ontology Consortium, Carbon S, Douglass E, Good BM, Unni DR, Harris
904 NL, et al. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res.
905 2021;49:D325–34. https://doi.org/10.1093/nar/gkaa1113.
906
907
908

909

910

911

912

Supplementary Files

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

- Additionalfile1.docx