

Research Article

Metabolic and Immune Heterogeneity Defines Distinct Tumor States in Esophageal Squamous Cell Carcinoma

Shehzad Khalil^{1*}, Rizwan Ullah², Farwa Anam³, Huma Amir⁴, Muhammad Wasim⁵¹ Institute of Biotechnology and Genetic Engineering, The University of Agriculture, Peshawar, Khyber Pakhtunkhwa, Pakistan² Department of Microbiology, University of Swabi, Khyber Pakhtunkhwa, Pakistan³ Institute of Biochemistry and Biotechnology, University of Veterinary and Animal Sciences, Lahore, Pakistan⁴ Department of Microbiology, University of Swat, Swat, Khyber Pakhtunkhwa, Pakistan⁵ Specialist Physician, Gastroenterology, Sheikh Tahnoon Bin Mohammed Medical City, Al Ain, UAE

*Corresponding author: Shehzad Khalil, Shehzadkhalil911@gmail.com

"Cite this Article" | Received: 12 April 2026; Accepted: 17 June 2026; Published: 30 June 2026.

ABSTRACT

Background: Esophageal squamous cell carcinoma (ESCC) is transcriptionally heterogeneous, but the reproducibility of expression-defined molecular states across cohorts and microarray platforms remains uncertain. This study evaluated whether biologically coherent ESCC expression states could be identified in a pooled cross-platform dataset and whether their boundaries remained portable across cohorts. **Methods:** Publicly available microarray data from three GEO cohorts were analyzed, comprising 77 ESCC tumors and 77 matched normal tissues from GSE23400, GSE77861, and GSE20347. RMA-normalized expression matrices were harmonized to 13,039 common HGNC symbols and adjusted for dataset/platform-associated variation using ComBat. From the 1,200 most variable tumor genes, 264 genes strongly associated with the tumor-normal axis were excluded, leaving 936 features for consensus clustering. Candidate solutions from k=2 to k=6 were evaluated using PAC, silhouette width, CDF behavior, and concordance with non-negative matrix factorization. The resulting states were characterized by limma differential-expression analysis and ssGSEA, followed by two cross-dataset centroid-assignment robustness analyses. **Results:** A three-state working solution was selected, comprising S1 (n=30), S2 (n=27), and S3 (n=20), although cluster-number diagnostics were not fully concordant and NMF agreement at k=3 was modest (ARI=0.267). State distribution was not detectably associated with dataset (Fisher-Freeman-Halton p=0.335). S1 showed metabolic and detoxification-associated programs, S2 complement/B-cell and innate immune-associated programs, and S3 interferon/inflammatory/antiviral programs. Cross-dataset portability was limited: 13/24 assignments were ambiguous in the first robustness analysis (54.2%; 95% CI, 35.1%-72.1%) and 8/17 in the second (47.1%; 95% CI, 26.2%-69.0%). **Conclusion:** ESCC exhibits biologically coherent metabolic and immune-associated expression states, but the exact patient-level boundaries of these states are weakly reproducible across cohorts. The three-state model should therefore be considered hypothesis-generating rather than a validated molecular classifier and requires confirmation in larger, independently processed cohorts using a fully locked validation framework. **Keywords:** Esophageal squamous cell carcinoma; transcriptomic heterogeneity; molecular subtyping; consensus clustering; ssGSEA; cross-platform integration; reproducibility

DECLARATIONS

Author Contributions: Concept: SK, RU; Design: MW, FA; Data Collection: RU, HA; Analysis: MW, FA; Drafting: SK**Ethical Approval:** Not required because the study used secondary data.**Informed Consent:** Written informed consent was obtained from all participants.**Conflict of Interest:** The authors declare no conflict of interest; **Funding:** No external funding; **Data Availability:** Available from the corresponding author on reasonable request; **Acknowledgments:** Not applicable.

INTRODUCTION

Esophageal cancer remains a major global health burden and is associated with substantial mortality. According to GLOBOCAN 2022, an estimated 511,054 new cases and 445,391 deaths occurred worldwide, placing esophageal cancer among the leading causes of cancer-related death (1). Esophageal squamous cell carcinoma (ESCC) constitutes the predominant histologic subtype in several high-incidence regions, particularly in Eastern Asia and parts of sub-Saharan Africa, and outcomes remain poor for patients with advanced disease (2-4). These epidemiologic and clinical challenges have increased interest in defining the molecular heterogeneity of ESCC and identifying biologically meaningful tumor states that may improve understanding of disease mechanisms.

Large-scale genomic and transcriptomic studies have demonstrated that ESCC is molecularly heterogeneous rather than a single biologically uniform disease. The Cancer Genome Atlas identified

distinct molecular features within esophageal carcinoma, while subsequent transcriptome-based analyses have proposed ESCC subgroups characterized by differing signaling, metabolic, immune, and microenvironmental programs (5,6). However, identifying clusters within a single dataset does not establish that those groups represent reproducible molecular classes. For expression-defined states to have broader biological or translational relevance, their defining patterns should remain recognizable when patient populations, laboratory conditions, microarray platforms, and preprocessing pipelines differ across cohorts.

Cross-cohort molecular class discovery is particularly challenging because technical and biological sources of variation are often intertwined. Dataset identity may simultaneously reflect differences in microarray platform, laboratory processing, geographic origin, ancestry, recruitment strategy, and specimen characteristics. Empirical-Bayes approaches such as ComBat can substantially reduce dataset-associated variation, but statistical batch adjustment cannot establish that all technical variation has been removed or that cohort-specific biological information has been fully preserved (7,8). Consequently, apparent clustering after data integration must be interpreted cautiously, particularly when dataset and biological characteristics cannot be completely disentangled.

A further challenge is that unsupervised clustering of cancer transcriptomes may be dominated by strong expression axes that are not specific to stable tumor states. In particular, tumor-versus-normal differences, tissue composition, stromal content, and immune-cell infiltration can contribute substantially to transcriptomic variance and may obscure subtler among-tumor structure. The concept of selecting intrinsic expression features has been influential in molecular subtyping, particularly in breast cancer, where genes informative for variation among tumors were prioritized to reduce the influence of unrelated expression differences (9-11). At the same time, consensus clustering and related class-discovery methods can produce apparently stable partitions even when the underlying class structure is weak or sensitive to methodological choices, emphasizing the importance of multiple complementary criteria for evaluating cluster number and robustness (12).

Against this background, the present study evaluated whether a pooled, cross-platform ESCC transcriptomic dataset supports reproducible expression-defined tumor states after dataset-associated adjustment and tumor-focused feature selection. Three publicly available GEO microarray cohorts were harmonized at the gene level and subjected to consensus class discovery, with the number of states evaluated using complementary clustering diagnostics rather than a single optimization criterion. The resulting states were characterized using differential-expression and gene-set enrichment analyses and were subsequently stress-tested using cross-dataset centroid assignment. The primary objective was therefore not to establish a clinical molecular classifier, but to determine whether biologically coherent ESCC expression states identified in pooled data remain sufficiently portable across cohorts to support a stronger claim of reproducible molecular subtyping.

MATERIALS AND METHODS

Study design and data sources

This secondary transcriptomic analysis used publicly available, de-identified microarray expression data from the NCBI Gene Expression Omnibus (GEO), a public repository for functional genomics datasets (13). Three ESCC cohorts were included: GSE23400, GSE77861, and GSE20347, contributing a total of 77 primary tumor samples and 77 matched normal samples. GSE23400 contributed 53 tumor-normal pairs generated on the GPL96 Affymetrix HG-U133A platform; GSE77861 contributed 7 pairs generated on GPL570 Affymetrix HG-U133 Plus 2.0; and GSE20347 contributed 17 pairs generated on GPL571 Affymetrix HG-U133A 2.0 (14-16). The present study was designed as a focused cross-platform analysis rather than an exhaustive systematic review of all publicly available ESCC transcriptomic cohorts.

Expression-matrix acquisition and cross-platform harmonization

Already processed, RMA-normalized microarray expression matrices were obtained from GEO; raw CEL-file preprocessing was not repeated in the present analysis. Robust multi-array average normalization is an established approach for background adjustment, normalization, and summarization of high-density

oligonucleotide array data (17). Expression values were linked to sample-level tumor/normal and dataset metadata, and probes were mapped to HGNC gene symbols using platform-specific annotation information.

Table 1. Public ESCC cohorts included in the analysis.

GEO series	Platform used in this analysis	Tumor / matched normal	Population/context	Primary source
GSE23400	GPL96, Affymetrix HG-U133A*	53 / 53	China; matched tumor/normal ESCC	Su et al. (14)
GSE77861	GPL570, Affymetrix HG-U133 Plus 2.0	7 / 7	African American patients, USA	Erkizan et al. (16)
GSE20347	GPL571, Affymetrix HG-U133A 2.0	17 / 17	High-risk region of China	Hu et al. (15)

*GSE23400 contains data generated on both GPL96 and GPL97; only the 53-pair GPL96/HG-U133A component was used in the present analysis. GSE20347 corresponds to GPL571/HG-U133A 2.0 and was not treated as an HG-U133B dataset.

The three datasets were then harmonized to the intersection of genes represented across all platforms, yielding 13,039 common HGNC symbols. The harmonized matrices were combined into a 13,039-gene × 154-sample expression matrix comprising 77 tumors and 77 matched normal tissues. Probe-to-gene harmonization can be lossy when multiple probe sets map to the same gene symbol; therefore, the resulting gene-level representation was treated as a practical cross-platform approximation rather than as a uniquely defined biological measurement of each gene.

Adjustment of dataset/platform-associated variation

Dataset/platform-associated variation was adjusted using ComBat, an empirical-Bayes batch-adjustment method, with the combined dataset/platform identifier specified as the batch variable and without inclusion of a biological outcome model (7,8). Because each cohort contained matched tumor and normal samples, tissue status was not completely confounded with dataset; however, dataset identity also captured differences in platform, population, geography, laboratory processing, and other cohort-specific factors. ComBat adjustment was therefore interpreted as attenuation of dataset-associated structure rather than proof of selective removal of technical variation alone.

To quantify dataset association before and after adjustment, principal-component analysis was performed on the pooled expression matrix. For each of the first two principal components, dataset association was summarized using the coefficient of determination (R^2) from a linear model in which the principal-component score was the outcome and dataset was entered as a categorical predictor. This analysis was conceptually analogous to variance-component assessment and was used as a descriptive diagnostic of dataset-associated structure rather than as evidence that all batch effects had been eliminated (18).

Tumor-focused intrinsic feature selection

Feature selection was designed to emphasize among-tumor heterogeneity while reducing domination by the substantially stronger tumor-versus-normal expression axis. This strategy was informed by the intrinsic-feature concept used in earlier molecular-subtyping studies, in which features informative for heterogeneity among tumors were preferentially retained for class discovery (9-11).

First, genes were ranked according to median absolute deviation (MAD) across the 77 tumor samples, and the 1,200 most variable tumor genes were retained as candidate features. A tumor-normal expression proxy was then derived from PC1 calculated using the 1,000 most variable genes across all 154 samples. Candidate genes were excluded if either of two criteria was met: an absolute Pearson correlation of at least 0.50 with the tumor-normal PC1 proxy, or at least 50% of gene-expression variance explained by tumor/normal status in a gene-wise linear model. Of the 1,200 candidate genes, 264 met at least one exclusion criterion, leaving 936 features for unsupervised class discovery.

These thresholds were used operationally to reduce the influence of strong tissue-status-associated genes while retaining sufficient among-tumor variation for clustering. Because feature filtering can influence unsupervised partitions, the resulting classes were subsequently evaluated using multiple clustering

diagnostics and cross-dataset robustness analyses rather than being interpreted as intrinsically validated disease classes.

Consensus clustering and selection of the working number of states

Unsupervised class discovery was performed on the 936×77 tumor expression matrix using ConsensusClusterPlus (19). Pairwise dissimilarity was defined as 1 minus the Pearson correlation coefficient, and hierarchical clustering was performed using Ward linkage. Candidate solutions from $k=2$ to $k=6$ were evaluated using 1,000 resampling iterations with 80% sample subsampling.

Because consensus clustering can produce apparently stable partitions even when the underlying class structure is uncertain, cluster number was not selected using a single diagnostic (12). Four complementary criteria were considered: the proportion of ambiguous clustering (PAC), for which lower values indicate less ambiguous consensus; average silhouette width, for which higher values indicate greater within-cluster cohesion relative to between-cluster separation (20); cumulative distribution function (CDF) area and changes in CDF area across successive k values; and concordance with an independently parameterized non-negative matrix factorization (NMF) partition, summarized using the adjusted Rand index (ARI). NMF used multiplicative-update factorization, a standard approach for non-negative matrix decomposition and metagene-based class discovery (21,22).

The primary k value was selected by triangulating these diagnostics rather than by treating any one metric as definitive. NMF-ARI was interpreted as an algorithmic concordance check rather than as independent validation because both clustering procedures were applied to the same expression matrix. Accordingly, the selected partition was considered a working expression-state solution rather than evidence of a uniquely determined natural class structure.

Differential-expression and pathway characterization

For biological characterization, each discovered expression state was compared with all remaining tumors using a two-group linear model implemented in limma (23,24). For each state, the design was specified as: expression ~ subtype indicator

where the subtype indicator was coded as 1 for tumors assigned to the state of interest and 0 for all other tumors. The resulting subtype coefficient therefore represented the subtype-versus-rest log2 fold change. Dataset was not included as an additional model covariate because dataset/platform had already been used as the batch factor during ComBat adjustment. Differentially expressed genes were defined using a Benjamini-Hochberg false-discovery rate (FDR) <0.05 together with an absolute log2 fold change ≥ 1 (25).

Per-sample pathway activity was estimated using single-sample gene-set enrichment analysis (ssGSEA) as implemented in GSVA (26). Gene sets were drawn from the MSigDB Hallmark collection, Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, and C7 immunologic-signature collections (27-30). Across-state differences in ssGSEA scores were evaluated using Kruskal-Wallis tests, followed by Benjamini-Hochberg correction for multiple testing (31). Pairwise Wilcoxon rank-sum tests with BH adjustment were used for follow-up comparisons where applicable (32). Pathway analyses were used for biological characterization rather than for treatment assignment or clinical prediction.

Because both state discovery and subsequent differential-expression/pathway characterization were performed using the same underlying transcriptomic dataset, these inferential statistics were interpreted as post hoc characterization of the discovered states rather than as independent confirmatory tests. This distinction was particularly important for the highly overlapping C7 immunologic-signature collection, for which multiple significant gene sets may reflect correlated representations of the same underlying immune processes.

Cross-dataset robustness analyses

Two centroid-assignment analyses were performed to assess the portability of the discovered expression states across cohorts. In robustness analysis A, subtype centroids were estimated using the 53 GSE23400 tumors, and the remaining 24 tumors from GSE77861 and GSE20347 were assigned to the state centroid with which they showed the highest Pearson correlation. In robustness analysis B, centroids were

estimated from the combined GSE23400 and GSE77861 tumors (n=60), and the 17 GSE20347 tumors were assigned using the same procedure.

Each tumor received a best-correlating centroid assignment. Assignment confidence was quantified as the difference between the highest and second-highest centroid correlations. A margin <0.05 was operationally classified as ambiguous, whereas assignments with a margin ≥ 0.05 were classified as confident. Thus, ambiguous samples retained a nominal best-correlating state label but were explicitly flagged as having limited separation from the next-best centroid.

These analyses were treated as cross-dataset robustness checks rather than independent external validation. The assignment cohorts had contributed to upstream pooled harmonization and/or class discovery, and the global feature-selection and k-selection procedures were not locked independently of all target cohorts. In addition, pathway-profile concordance after assignment was calculated using the same transcriptomic data from which centroid correlations were derived. Assignment confidence was therefore regarded as the more stringent portability measure, whereas correlations between discovery-state and assigned-group pathway profiles were considered descriptive and potentially optimistic.

Statistical analysis and computational reproducibility

Because the subtype-by-dataset contingency table contained small expected cell counts and a zero cell, association between dataset and subtype was evaluated using a two-sided Fisher-Freeman-Halton exact test rather than Pearson's chi-square test. Wilson 95% confidence intervals were calculated for proportions of ambiguous cross-dataset assignments. Continuous pathway-score comparisons were assessed using Kruskal-Wallis and Wilcoxon tests as described above (31,32). t-distributed stochastic neighbor embedding (t-SNE) was used only for low-dimensional visualization and did not contribute to cluster definition or selection (33).

Analyses were conducted in R 4.6.0 using packages including limma, sva, ConsensusClusterPlus, GSVA, msigdb, clusterProfiler, org.Hs.eg.db, platform-specific Affymetrix annotation resources, pheatmap, Rtsne, reshape2, and ggplot2. A fixed random seed (20260825) was used for resampling, NMF initialization, and t-SNE to improve computational reproducibility. The Fisher-Freeman-Halton test and Wilson intervals were recalculated from the observed subtype-by-dataset and assignment counts.

RESULTS

Cross-platform integration markedly attenuated dataset-associated expression structure

After gene-level harmonization, the combined dataset comprised 13,039 common HGNC symbols across 154 samples, including 77 ESCC tumors and 77 matched normal tissues from GSE23400, GSE77861, and GSE20347. Before dataset/platform adjustment, principal-component analysis demonstrated marked cohort-associated structure, particularly along PC2. Dataset identity explained 6.2% of the variance in PC1 and 72.0% of the variance in PC2. Following ComBat adjustment, the corresponding dataset-associated R^2 values decreased to approximately 6.98×10^{-6} for PC1 and 1.66×10^{-5} for PC2.

The post-adjustment PCA demonstrated substantially greater overlap among samples from the three cohorts than was observed before adjustment. These findings indicate marked attenuation of dataset-associated structure in the leading principal components. Because dataset identity also reflects population, platform, laboratory processing, and other cohort-specific characteristics, this reduction was not interpreted as evidence that exclusively technical variation had been removed.

Tumor-focused feature selection retained 936 genes for class discovery

To reduce domination of the clustering solution by the tumor-versus-normal expression axis, the 1,200 genes with the highest median absolute deviation across tumor samples were initially retained as candidate features. Of these, 264 genes met at least one tumor-normal-proxy exclusion criterion and were removed, leaving 936 genes for unsupervised class discovery.

The retained feature set therefore preferentially represented among-tumor transcriptional variation while reducing the contribution of genes strongly associated with tissue status. Because feature filtering can

influence unsupervised partitions, the resulting class structure was subsequently assessed using multiple internal clustering diagnostics and cross-dataset robustness analyses.

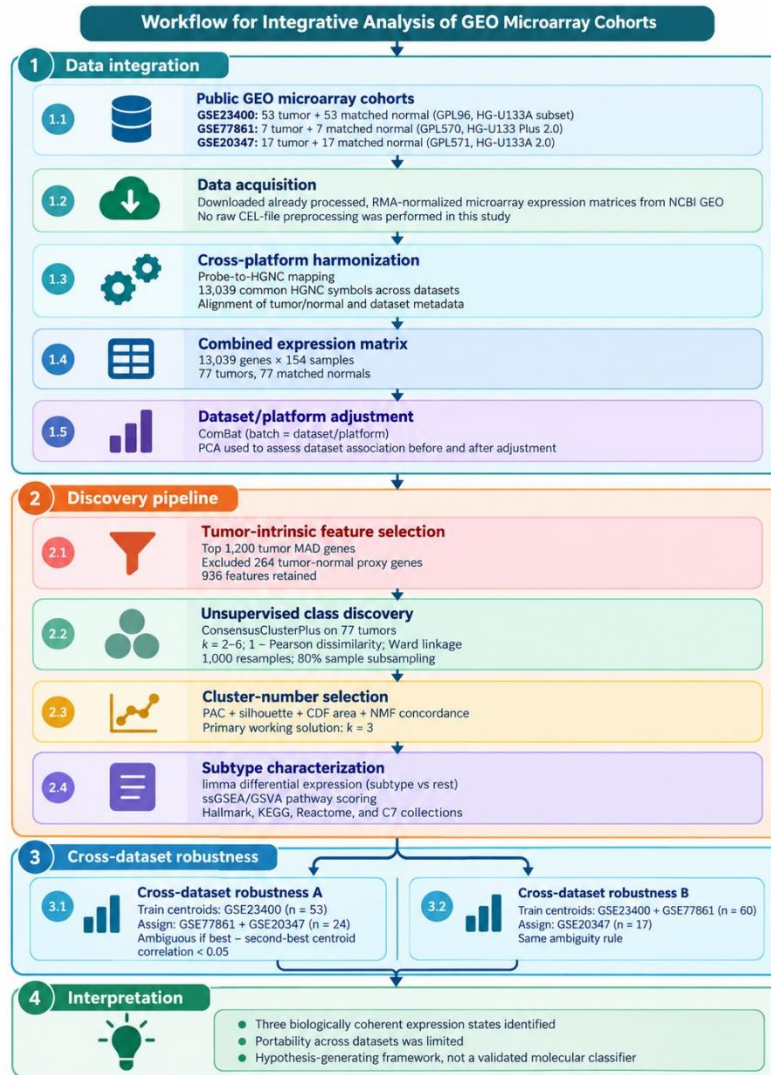


Figure 1. Tumor-focused feature selection for unsupervised class discovery. Candidate genes are displayed according to tumor-only median absolute deviation and their association with the tumor-normal expression axis. Genes meeting at least one prespecified tumor-normal-proxy exclusion criterion are distinguished from retained genes. Of 1,200 candidate genes, 264 were excluded and 936 were retained for consensus clustering. The displayed association represents one component of the filtering procedure; the complete exclusion rule also incorporated the proportion of gene-expression variance explained by tumor/normal status.

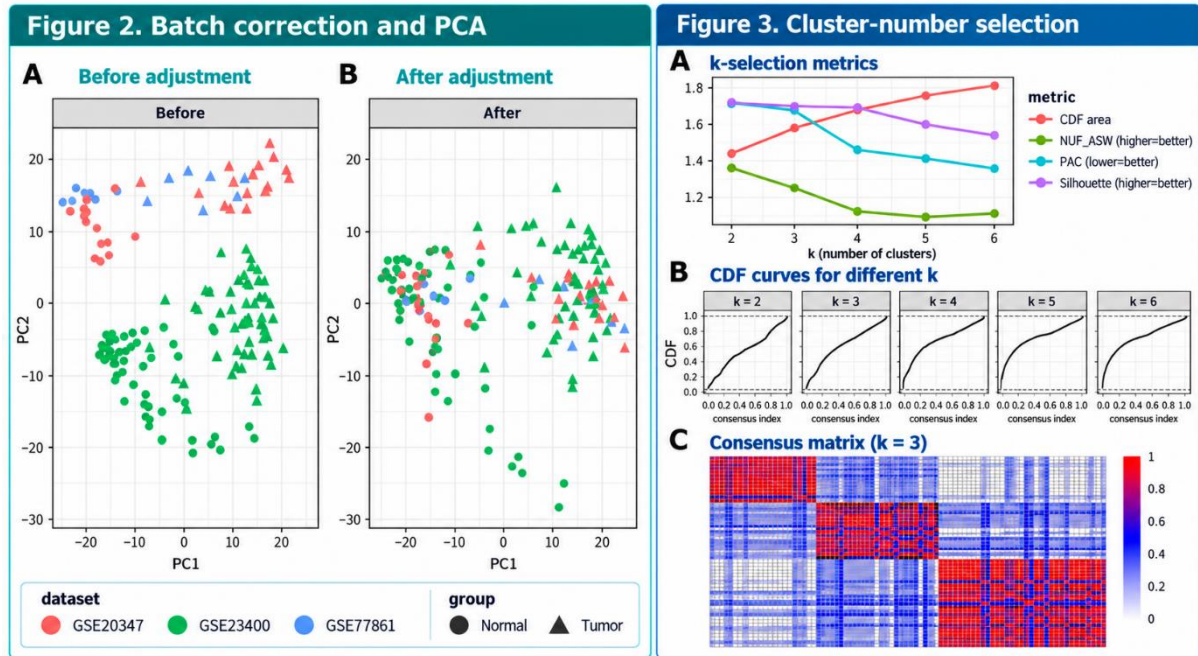
Consensus clustering solutions from $k=2$ to $k=6$ were evaluated using PAC, average silhouette width, CDF-area behavior, and concordance with NMF-derived partitions. The diagnostic criteria did not identify a single unequivocal optimum. Average silhouette width was highest at $k=2$ (0.7303) and remained similar at $k=3$ (0.7127) and $k=4$ (0.7110) before decreasing at higher k values. PAC declined from 0.7266 at $k=2$ to 0.3693 at $k=6$, without a sharply defined optimum at $k=3$. CDF area increased progressively across k values, with the largest increment occurring from $k=2$ to $k=3$ ($\Delta\text{CDF}=0.1536$). Agreement with NMF-derived partitions was modest across all candidate solutions; the NMF ARI was 0.2673 at $k=3$.

Table 2. Cluster-number selection metrics

k	PAC*	Average silhouette width	CDF area	ΔCDF to next k†	NMF ARI
2	0.7266	0.7303	0.4547	0.1536	0.3762
3	0.7010	0.7127	0.6083	0.0810	0.2673
4	0.4775	0.7110	0.6893	0.0633	0.1093
5	0.4306	0.6135	0.7526	0.0471	0.0607
6	0.3693	0.5461	0.7997	—	0.0741

*Lower PAC indicates less ambiguous consensus clustering. † Δ CDF represents the increase in CDF area from the current k to the next k . Primary working solution: $k=3$. The absence of a sharp PAC optimum and the modest NMF concordance are important qualifications of this selection.

On balance, $k=3$ was selected as the primary working solution because it retained near-maximal silhouette width, captured the largest increase in CDF area, and showed greater NMF concordance than $k=4$. The solution was therefore treated as a pragmatic working partition rather than as evidence that three represents a uniquely determined natural number of ESCC molecular classes.



Figures 2 and 3. Dataset/platform adjustment and selection of the primary $k=3$ expression-state solution. (A-B) Principal-component analysis before and after ComBat adjustment, with samples colored by dataset and shaped by tissue group. Dataset-associated separation was markedly attenuated after adjustment, although this does not establish complete removal of technical effects or preservation of all cohort-specific biological variation. (C) Cluster-number diagnostics across $k=2-6$, including PAC, average silhouette width, CDF-area behavior, and NMF concordance measured by the adjusted Rand index, with $k=3$ selected as the primary working solution. (D) Consensus cumulative distribution function curves across candidate k values. (E) Consensus matrix for $k=3$. Collectively, these analyses show substantial reduction in dataset-associated structure and support a usable three-state working partition, while also indicating that the clustering diagnostics do not identify a single sharply defined optimum.

Table 3. Distribution of the three expression states across datasets

Expression state	GSE23400	GSE77861	GSE20347	Total
S1	20	3	7	30
S2	16	4	7	27
S3	17	0	3	20
Total	53	7	17	77

Two-sided Fisher-Freeman-Halton exact test: $p=0.335$.

The $k=3$ partition comprised three expression states without a detectable overall association with dataset

The selected partition classified the 77 tumors into S1 ($n=30$), S2 ($n=27$), and S3 ($n=20$). The distribution of the three states across the source datasets is shown in Table 3. GSE23400 contributed 20 S1, 16 S2, and 17 S3 tumors; GSE77861 contributed 3 S1 and 4 S2 tumors, with no S3 tumors; and GSE20347 contributed 7 S1, 7 S2, and 3 S3 tumors.

No statistically detectable overall association between dataset and state assignment was identified ($p=0.335$). However, the contingency table contained small cell counts and no GSE77861 tumors assigned to S3; therefore, the absence of statistical significance was not interpreted as evidence that state composition was independent of cohort.

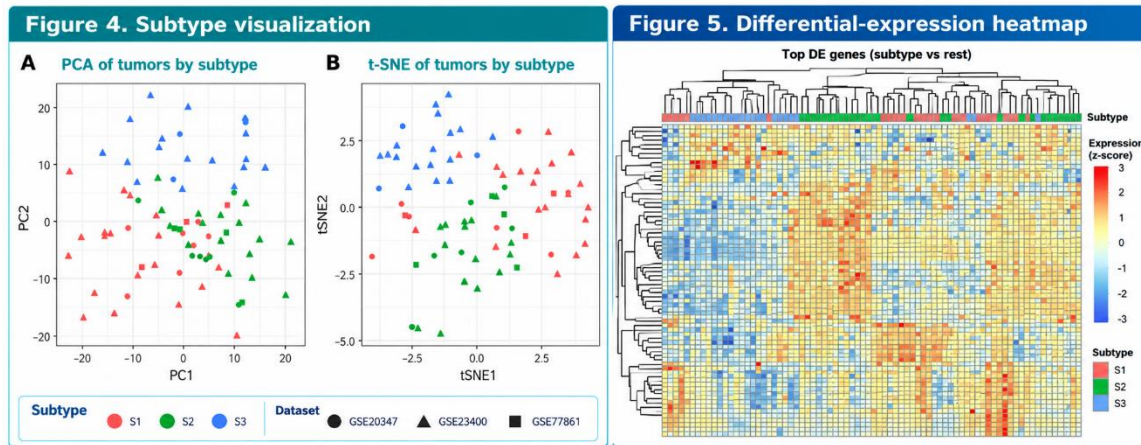
PCA and t-SNE visualizations demonstrated partial separation of S1, S2, and S3, although the states were not geometrically discrete. t-SNE was used only for descriptive visualization and did not contribute to cluster definition or selection of k .

The three expression states exhibited distinct transcriptional programs

Subtype-versus-rest differential-expression analysis identified 61 genes for S1, 51 genes for S2, and 88 genes for S3 that satisfied the prespecified criteria of BH-adjusted FDR < 0.05 and absolute log2 fold change ≥ 1 .

S1 exhibited a predominantly metabolic and detoxification-associated expression pattern. S2 demonstrated stronger expression of genes associated with complement, B-cell, and innate immune processes, whereas S3 showed a more prominent interferon, inflammatory, and antiviral-associated transcriptional pattern.

The differential-expression heatmap demonstrated coherent expression blocks associated with the three states, although overlap among individual tumors remained evident. Thus, transcriptional differences were biologically structured but did not represent completely discrete expression classes.



Figures 4 and 5. Low-dimensional visualization and differential-expression patterns of ESCC tumors according to the $k=3$ working partition. (A) Principal-component analysis and (B) t-distributed stochastic neighbor embedding of the 77 ESCC tumors. Point color indicates expression state (S1, S2, or S3), whereas point shape indicates source dataset. The visualizations show partial but incomplete separation of the three expression states; t-SNE was used solely for visualization and did not contribute to class discovery or selection of k . (C) Heatmap of selected subtype-versus-rest differentially expressed genes across the 77 tumors. Columns represent individual tumors annotated by S1, S2, or S3 assignment, and rows represent selected differentially expressed genes shown as gene-wise standardized expression values. The heatmap demonstrates coherent but overlapping state-associated transcriptional patterns rather than completely discrete molecular classes.

Pathway-level ssGSEA analysis further differentiated the three expression states. S1 showed enrichment of metabolic and detoxification-associated programs, including pentose and glucuronate interconversions, ascorbate and aldarate metabolism, glucuronidation, xenobiotic and drug metabolism by cytochrome P450, retinol metabolism, porphyrin metabolism, fatty-acid metabolism, and NFE2L2-associated antioxidant/detoxification pathways. S2 demonstrated greater activity of pathways associated with complement and humoral/innate immunity, including classical and alternative complement activation, C3/C5 and terminal complement pathways, B-cell receptor signaling, CD22-mediated BCR regulation, Fc-gamma receptor activation, heme scavenging, and phagocytosis-associated signaling. S3 was characterized by interferon- and inflammatory-associated activity, including Hallmark interferon-alpha and interferon-gamma responses, OAS antiviral programs, interferon-stimulated host-response pathways, interleukin-1 processing, CASP4-associated noncanonical inflammasome programs, interleukin-10 signaling, antigen processing, and chemokine-related pathways.

Collectively, these findings support interpretation of S1 as a metabolic/detoxification-enriched state, S2 as a complement/B-cell and innate immune-enriched state, and S3 as an interferon/inflammatory/antiviral-

enriched state. Because state discovery and pathway characterization used the same underlying transcriptomic dataset, these results were regarded as biological characterization of the discovered states rather than independent validation.

Table 4. Biological characteristics of the three ESCC expression states

State	n	DEGs*	Dominant biological programs
S1	30	61	Metabolic/detoxification: pentose-glucuronate and ascorbate-aladarate metabolism; CYP450 xenobiotic/drug metabolism; glucuronidation; NFE2L2-associated antioxidant programs; fatty-acid metabolism
S2	27	51	Complement/B-cell and innate immunity: complement activation; B-cell receptor signaling; Fc-gamma receptor activation; phagocytosis-associated programs
S3	20	88	Interferon/inflammatory/antiviral: IFN- α/γ responses; OAS antiviral programs; IL-1 processing; CASP4-associated inflammasome activity; IL-10 and chemokine signaling

*DEGs were defined using BH-adjusted FDR <0.05 and absolute log2 fold change ≥ 1 .

Collectively, these findings support interpretation of S1 as a metabolic/detoxification-enriched state, S2 as a complement/B-cell and innate immune-enriched state, and S3 as an interferon/inflammatory/antiviral-enriched state. Because state discovery and pathway characterization used the same underlying transcriptomic dataset, these results were regarded as biological characterization of the discovered states rather than independent validation.

C7 immunologic signatures further differentiated the immune-associated states

Among 5,190 C7 immunologic signatures, 2,855 (55.0%) showed state-associated differences at BH-adjusted FDR <0.05 in the post hoc characterization analysis. S2 demonstrated stronger complement- and humoral-associated patterns, whereas S3 showed more prominent interferon- and inflammatory-associated signatures.

The direction of these results was consistent with the Hallmark and Reactome pathway analyses. However, because C7 contains extensive overlap among immunologic gene sets, the number of statistically significant signatures does not represent an equivalent number of independent immune processes. Furthermore, the bulk-tissue expression data could not distinguish tumor-cell-intrinsic immune signaling from variation in infiltrating immune-cell composition.

Boxplots display representative C7 immunologic-signature ssGSEA scores across S1, S2, and S3. The patterns illustrate differences along complement/humoral and interferon/inflammatory immune axes. Because C7 gene sets are highly overlapping and the analysis is based on bulk-tissue expression, the findings should be interpreted as descriptive evidence of immune-state differentiation rather than direct estimates of specific immune-cell abundance or tumor-cell-intrinsic immune signaling. Cross-dataset centroid assignment demonstrated limited portability of the three-state boundaries. Cross-dataset robustness analyses demonstrated that a substantial proportion of tumors could not be assigned confidently to a single state centroid. In robustness analysis A, centroids were estimated from the 53 GSE23400 tumors and applied to the 24 tumors from GSE77861 and GSE20347. Each tumor was assigned nominally to the state with the highest centroid correlation; however, 13 of 24 assignments were ambiguous, corresponding to 54.2% (95% Wilson CI, 35.1%-72.1%). Nominal assigned counts were GSE20347: S1=7, S2=7, and S3=3; and GSE77861: S1=2, S2=4, and S3=1. In robustness analysis B, centroids derived from the combined GSE23400 and GSE77861 tumors (n=60) were applied to the 17 GSE20347 tumors. Eight of 17 assignments were ambiguous, corresponding to 47.1% (95% Wilson CI, 26.2%-69.0%). Nominal assignments were S1=7, S2=7, and S3=3.

Table 5. Cross-dataset robustness of the three-state expression model

Robustness scheme	Training cohort(s), n	Test cohort(s), n	Ambiguous assignments	Hallmark profile r^*	Immune profile r^*
A: GSE23400 $\rightarrow$ other datasets	53	24	13/24, 54.2% (95% CI 35.1-72.1)	0.985-0.997	0.979-0.996
B: GSE23400 + GSE77861 $\rightarrow$ GSE20347	60	17	8/17, 47.1% (95% CI 26.2-69.0)	0.994-0.998	0.994-0.997

*Range of state-specific Pearson correlations between mean discovery-state and assigned-group ssGSEA profiles. These correlations are descriptive and potentially optimistic because state assignments and pathway scores were obtained from the same underlying transcriptomic data.

Mean pathway profiles of nominally assigned groups remained highly correlated with their corresponding discovery-state profiles. In robustness analysis A, Hallmark correlations ranged from 0.985 to 0.997 and immune-profile correlations from 0.979 to 0.996. In analysis B, both Hallmark and immune-profile correlations ranged approximately from 0.994 to 0.998.

These high profile correlations were interpreted cautiously because centroid assignment and pathway scoring arose from the same transcriptomic measurements. In contrast, assignment-margin analysis demonstrated substantial uncertainty at the individual-tumor level. Together with the modest NMF concordance observed for $k=3$, the high frequency of ambiguous assignments indicates that the broad biological programs were more stable than the precise patient-level boundaries of the three expression states.

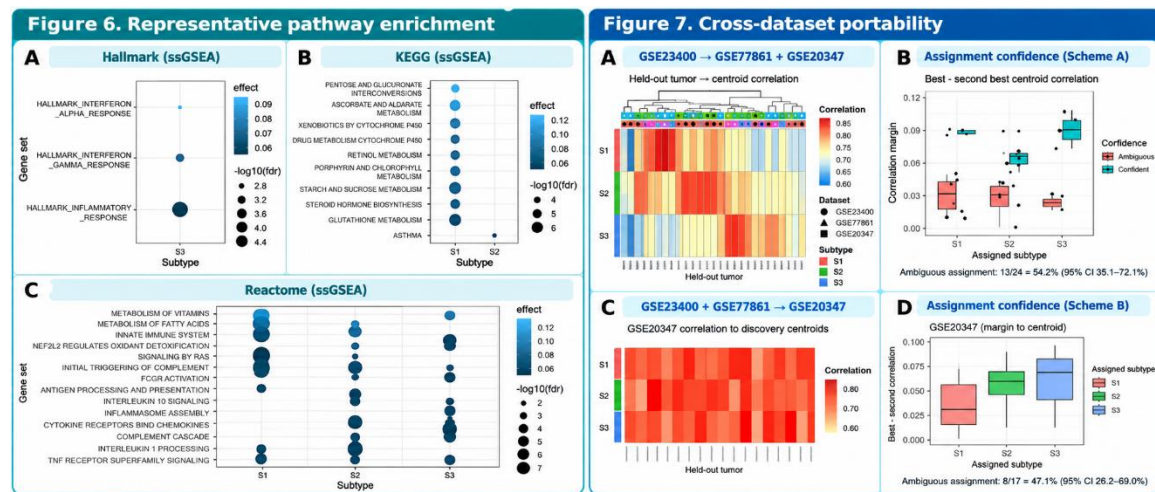


Figure 6. Representative pathway enrichment across ESCC expression states. Representative ssGSEA results from (A) Hallmark, (B) KEGG, and (C) Reactome collections. Point size represents $-\log_{10}(\text{BH-adjusted FDR})$ and point color represents the pathway effect-size measure. S1 is predominantly metabolic/detoxification-enriched, S2 complement/B-cell and innate immune-enriched, and S3 interferon/inflammatory/antiviral-enriched. Related gene sets may overlap and should not be interpreted as independent biological processes. Figure 7. Cross-dataset portability of the three-state ESCC expression model. (A-B) GSE77861 and GSE20347 tumors were assigned to centroids derived from GSE23400; 13/24 (54.2%) assignments were ambiguous using $\Delta r < 0.05$. (C-D) GSE20347 tumors were assigned to centroids derived from GSE23400 plus GSE77861; 8/17 (47.1%) were ambiguous. All tumors retained their highest-correlation state label. The frequent low-margin assignments indicate limited portability of the exact state boundaries across cohorts.

DISCUSSION

This cross-platform transcriptomic analysis identified three biologically interpretable ESCC expression states characterized predominantly by metabolic/detoxification, complement/B-cell and innate immune, and interferon/inflammatory/antiviral programs. The more consequential finding, however, was the contrast between relatively coherent biological patterns within the pooled dataset and limited confidence in transferring the exact state boundaries across cohorts. Although the $k=3$ working solution showed clear transcriptional and pathway-level differences, cluster-number diagnostics were not fully concordant, NMF agreement was modest, and approximately half of cross-dataset centroid assignments were classified as ambiguous. These findings support the existence of recurring biological expression programs in ESCC but do not justify interpretation of the three states as a validated molecular classifier.

The biological characteristics of the three states are broadly plausible in the context of known ESCC biology. The S1 metabolic/detoxification-enriched state included xenobiotic and drug metabolism, glucuronidation, cytochrome P450-associated pathways, fatty-acid metabolism, and NFE2L2-linked antioxidant programs. Such pathways are relevant to the exposure-related and oxidative-stress biology

implicated in esophageal carcinogenesis and have also been observed in prior transcriptomic analyses of ESCC (3,4,16). In contrast, S2 and S3 separated two distinct immune-associated patterns: S2 was enriched for complement activation, B-cell receptor signaling, Fc-receptor pathways, and phagocytosis-related processes, whereas S3 showed stronger interferon-alpha/gamma, antiviral, inflammasome, and chemokine-associated activity. Large-scale molecular profiling has likewise demonstrated substantial immune and microenvironmental heterogeneity in esophageal carcinoma, supporting the biological plausibility of these divergent immune-associated states (5).

At the same time, the present findings should not be interpreted as evidence that S1, S2, and S3 represent stable, lineage-pure tumor classes. The analysis was performed on bulk-tissue transcriptomes, and immune-related expression patterns may therefore reflect differences in leukocyte infiltration and stromal composition in addition to tumor-cell-intrinsic signaling. This consideration is particularly important for S2 and S3, whose defining programs are strongly immune-associated. The very large number of significant C7 immunologic signatures should also be interpreted cautiously because many gene sets in this collection are highly overlapping and correlated; consequently, the number of significant signatures does not correspond to an equivalent number of independent biological mechanisms.

The cluster-number analysis further illustrates why identification of the same nominal number of groups across studies is insufficient evidence of subtype replication. A previous large-scale ESCC transcriptional study also reported three molecular subtypes, but its dominant biological programs included Wnt signaling, glycogen metabolism, and neutrophil degranulation rather than the exact metabolic, complement-associated, and interferon-dominant pattern observed here (6). Thus, concordance at the level of k does not demonstrate recovery of the same biological classes. More generally, consensus clustering can yield apparently stable partitions even when class structure is weak or dependent on analytical choices, and alternative algorithms may produce only partial agreement (12,19-22). In the present study, silhouette width remained similar from $k=2$ to $k=4$, PAC continued to improve toward higher k values without a sharp minimum, and NMF concordance was low across candidate solutions. The selection of $k=3$ is therefore best regarded as a practical working model rather than evidence that ESCC contains exactly three naturally discrete transcriptomic classes.

The cross-dataset assignment analyses provide the clearest evidence of this limitation. Although state-level pathway profiles remained highly correlated after centroid transfer, 54.2% of assignments in the first robustness analysis and 47.1% in the second had correlation margins below the prespecified confidence threshold. These apparently discordant findings are informative rather than contradictory. High average pathway-profile correlation indicates that broad biological programs remain recognizable across cohorts, whereas the frequent low-margin assignments indicate that individual tumors often lie near the boundaries between states. The biological programs may therefore be more reproducible than the precise patient-level classifications. This distinction is important because a clinically useful molecular classifier requires not only coherent average profiles but also reliable, reproducible assignment of individual samples.

The terminology of validation is consequently important. The present cross-dataset analyses constitute internal robustness or transferability tests rather than independent external validation. The target cohorts contributed to pooled preprocessing, global feature selection, and selection of the primary clustering solution before centroid assignment. In addition, pathway-profile concordance was calculated from the same underlying expression data used for state assignment. A definitive validation study would require the complete discovery procedure—including feature set, preprocessing strategy, centroids, decision rules, and ambiguity threshold—to be locked before application to an independently processed cohort. Such a design should preferably include larger RNA-sequencing datasets with harmonized clinical information on stage, grade, treatment, survival, and ancestry to permit evaluation of prognostic, predictive, and population-specific relevance.

The study nevertheless has several strengths. It integrates three independent ESCC cohorts generated on different Affymetrix platforms, explicitly quantifies dataset-associated principal-component structure before and after adjustment, and attempts to reduce domination of clustering by the tumor-versus-normal

axis through tumor-focused feature selection. Cluster number was evaluated using several complementary criteria rather than a single statistic, an alternative NMF partition was used as an algorithmic concordance check, and the cross-dataset analyses retained ambiguous assignments rather than excluding poorly classified samples. Importantly, the analysis reports the weak and discordant robustness findings directly instead of interpreting visually coherent clusters as evidence of validated molecular subtypes.

Several methodological considerations nonetheless constrain interpretation. The effective sample size was modest, with 77 tumors divided among states of 30, 27, and 20 cases, and the cross-dataset test sets contained only 24 and 17 tumors. Dataset identity also represented a mixture of technical and biological factors including platform, geography, ancestry, specimen handling, and recruitment context so ComBat adjustment could not separate technical batch from cohort-specific biology with certainty. Probe-to-gene harmonization across platforms is inherently lossy when multiple probe sets map to the same gene, and the selected feature-filtering thresholds may influence the resulting partition. In addition, Ward linkage with correlation-derived dissimilarity is a pragmatic rather than geometrically ideal combination, and sensitivity analyses using alternative distance-linkage specifications would further characterize the stability of the solution. Finally, because differential-expression and pathway analyses were performed after data-driven state discovery using the same transcriptomic matrix, their statistical significance should be interpreted primarily as post hoc characterization of the discovered states rather than as independent confirmatory inference.

Taken together, the findings support a model in which ESCC contains recurring metabolic and immune-associated transcriptomic programs, but the boundaries between these programs are not sufficiently stable across cohorts to support a validated subtype classifier. The most defensible interpretation is therefore not that three definitive ESCC molecular subtypes have been established, but that biologically coherent expression states can emerge reproducibly at the pathway level while individual tumors remain difficult to assign consistently across heterogeneous datasets. Future studies should determine whether these programs persist under a fully locked discovery-to-validation design and whether single-cell or spatial transcriptomic approaches can distinguish tumor-intrinsic signaling from microenvironmental contributions.

CONCLUSION

This cross-platform transcriptomic analysis supports a biologically coherent three-state working model of ESCC, comprising a metabolic/detoxification-enriched state, a complement/B-cell and innate immune-enriched state, and an interferon/inflammatory/antiviral-enriched state. These states were distinguishable at both gene-expression and pathway levels, indicating that recurring metabolic and immune-associated programs contribute substantially to transcriptional heterogeneity in ESCC.

However, the exact boundaries of these states were not consistently reproducible across cohorts. Cluster-number diagnostics were not fully concordant, agreement with NMF-derived partitions was modest, and approximately half of cross-dataset centroid assignments had low assignment margins. The findings therefore support reproducible biological expression programs more strongly than reproducible patient-level molecular classes.

Accordingly, the three-state solution should be interpreted as a hypothesis-generating framework rather than a validated molecular classifier. The present study does not establish prognostic, predictive, treatment-selection, or other clinical utility. Future studies should apply a fully locked discovery-to-validation pipeline to larger, independently processed cohorts with standardized clinical annotation and should incorporate single-cell or spatial transcriptomic approaches to distinguish tumor-intrinsic expression programs from microenvironmental contributions.

REFERENCES

- 1 International Agency for Research on Cancer. Global Cancer Observatory: Cancer Today. Oesophagus fact sheet, GLOBOCAN 2022. Lyon: International Agency for Research on Cancer; 2022. Accessed August 27, 2026.
- 2 Lagergren J, Smyth E, Cunningham D, Lagergren P. Oesophageal cancer. *Lancet*. 2017;390(10110):2383-96. doi:10.1016/S0140-6736(17)31462-9.
- 3 Pennathur A, Gibson MK, Jobe BA, Luketich JD. Oesophageal carcinoma. *Lancet*. 2013;381(9864):400-12. doi:10.1016/S0140-6736(12)60643-6.
- 4 Uhlenhopp DJ, Then EO, Sunkara T, Gaduputi V. Epidemiology of esophageal cancer: update in global trends, etiology and risk factors. *Clin J Gastroenterol*. 2020;13(6):1013-23.
- 5 Cancer Genome Atlas Research Network. Integrated genomic characterization of oesophageal carcinoma. *Nature*. 2017;541(7636):169-75. doi:10.1038/nature20805.
- 6 Wang D, Dai J, Suo C, Wang S, Zhang Y, Chen X. Molecular subtyping of esophageal squamous cell carcinoma by large-scale transcriptional profiling: characterization, therapeutic targets, and prognostic value. *Front Genet*. 2022;13:1033214. doi:10.3389/fgene.2022.1033214.
- 7 Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118-27.
- 8 Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012;28(6):882-3. doi:10.1093/bioinformatics/bts034.
- 9 Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature*. 2000;406(6797):747-52.
- 10 Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*. 2001;98(19):10869-74.
- 11 Parker JS, Mullins M, Cheang MCU, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*. 2009;27(8):1160-7.
- 12 Senbabaoglu Y, Michailidis G, Li JZ. Critical limitations of consensus clustering in class discovery. *Sci Rep*. 2014;4:6207.
- 13 Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30(1):207-10.
- 14 Su H, Hu N, Yang HH, et al. Global gene expression profiling and validation in esophageal squamous cell carcinoma and its association with clinical phenotypes. *Clin Cancer Res*. 2011;17(9):2955-66. doi:10.1158/1078-0432.CCR-10-2724.
- 15 Hu N, Clifford RJ, Yang HH, et al. Genome wide analysis of DNA copy number neutral loss of heterozygosity (CNNLOH) and its relation to gene expression in esophageal squamous cell carcinoma. *BMC Genomics*. 2010;11:576. doi:10.1186/1471-2164-11-576.
- 16 Erkizan HV, Johnson K, Ghimbovschi S, et al. African-American esophageal squamous cell carcinoma expression profile reveals dysregulation of stress response and detox networks. *BMC Cancer*. 2017;17(1):426. doi:10.1186/s12885-017-3423-1.
- 17 Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003;4(2):249-64.
- 18 Bushel PR. pvca: Principal Variance Component Analysis. Bioconductor package.

- 19 Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*. 2010;26(12):1572-3.
- 20 Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. 1987;20:53-65.
- 21 Lee DD, Seung HS. Algorithms for non-negative matrix factorization. In: *Advances in Neural Information Processing Systems* 13. 2000. p. 556-62.
- 22 Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A*. 2004;101(12):4164-9.
- 23 Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47.
- 24 Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 2004;3:Article3.
- 25 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B*. 1995;57(1):289-300.
- 26 Hanzelmann S, Castelo R, Guinney J. GSEA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*. 2013;14:7.
- 27 Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst*. 2015;1(6):417-25.
- 28 Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 2000;28(1):27-30.
- 29 Fabregat A, Jupe S, Matthews L, et al. The Reactome Pathway Knowledgebase. *Nucleic Acids Res*. 2018;46(D1):D649-55.
- 30 Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 2011;27(12):1739-40.
- 31 Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. *J Am Stat Assoc*. 1952;47(260):583-621.
- 32 Wilcoxon F. Individual comparisons by ranking methods. *Biometrics Bull*. 1945;1(6):80-3.
- 33 van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9:2579-605.