ORIGINAL ARTICLE

Cross-Cohort Discovery and Independent Validation of Transcriptomic Programs and Candidate Genes in Oral Squamous Cell Carcinoma

Abhimanyu Putra[1*], Hasrita Soleiman[2]

[1] Department of Otorhinolaryngology, Mutiara Hospital, Southwest Papua, Indonesia
[2] Department of Otorhinolaryngology, Phlox Institute, Palembang, Indonesia

ARTICLE INFO

ARTICLE TYPE
Original Article

ARTICLE HISTORY
Received: November 5, 2025
Accepted: December 1, 2025
Published: December 29, 2025

KEYWORDS
Differential gene expression; Independent validation; Microarray; Oral squamous cell carcinoma; Pathway enrichment

CORRESPONDING AUTHOR
Abhimanyu Putra
E-mail: abhimanyu@phlox.or.id
ORCID: 0009-0007-2157-6681

DOI
https://doi.org/10.59345/crown.v3i2.328

ABSTRACT

Background: Cross-cohort validation can distinguish reproducible oral squamous cell carcinoma (OSCC) expression signals from cohort- and platform-specific findings.

Objective: This study aimed to identify genes and pathways that replicate across independent OSCC microarray cohorts.

Methods: GSE30784 (discovery) and GSE25099 and paired GSE37991 (validation) comprised 264 OSCC and 107 non-malignant samples. Cohorts were normalized, annotated, and modeled separately using robust empirical-Bayes linear models, with 12,708 shared genes eligible for testing. Discovery required FDR < 0.05 and $|\log_2$ fold change$| \geq 1.0$. Replication required concordant direction, validation FDR < 0.05, and $|\log_2$ fold change$| \geq 0.5$ in both validation cohorts. Direction-specific GO and KEGG enrichment used the common gene universe; QC-flagged arrays were excluded in sensitivity analyses.

Results: GSE30784 yielded 1,167 DEGs (583 upregulated; 584 downregulated), of which 504 replicated in both validation cohorts (273 upregulated; 231 downregulated). CRISP3, MMP10, MMP13, MMP1, FAM3B, KRT4, TMPRSS11B, TYRP1, MMP12, and SERPINE1 had the largest minimum cross-cohort effects. Replicated upregulated genes were enriched in cytokine, ECM-receptor, integrin, IL-17, focal-adhesion, and PI3K-Akt pathways; downregulated genes were enriched in cornified-envelope and metabolic pathways. Sensitivity analysis retained 488 of 504 primary calls, including all top 20 candidates.

Conclusion: Independent validation identified reproducible OSCC transcriptional programs centered on matrix remodeling, adhesion, inflammation, and reduced epithelial-metabolic functions. These expression-based candidates require orthogonal and functional validation.

1. Introduction

Oral squamous cell carcinoma (OSCC) is the principal histological form of oral cavity malignancy and remains a major source of cancer morbidity and mortality, particularly in populations with substantial tobacco, alcohol, and areca-nut exposure. Global estimates continue to show a considerable burden of lip and oral cavity cancer, while outcomes are constrained by late presentation, nodal dissemination, field cancerization, and molecular heterogeneity[1]. Comprehensive genomic profiling has identified recurrent alterations in cell-cycle control, squamous differentiation, growth-factor signaling, and immune regulation in head and neck squamous cell carcinoma[2]. Nevertheless, DNA-level alterations do not fully specify the transcriptional programs operating in tumors and their microenvironment.

Public transcriptomic repositories make it possible to discover expression changes in one cohort and test them in independent datasets. Microarray cohorts differ in platform, probe annotation, tissue composition, control selection, and exposure profile; direct sample pooling can therefore confound biological effects with study structure. A discovery-validation design preserves each cohort's experimental scale and asks whether predefined genes reproduce across different platforms and control designs. Recent multi-omic, spatial, and single-cell studies demonstrate that OSCC programs vary across tumor regions, stromal states, microbial niches, metabolic environments, and premalignant-to-invasive transitions[3–7].

Previous investigators used GSE25099, GSE30784, and GSE37991 to identify common differentially expressed genes by intersection and to explore immune and prognostic signals[8]. A later large-scale analysis integrated multiple OSCC cohorts with a gene-pair framework for diagnostic and prognostic prediction[9]. Those approaches address commonality or prediction, whereas the present analysis evaluates gene-level replication through one discovery cohort and two separate validation cohorts, with a shared measurable-gene universe, stage-specific false-discovery control, and direction and effect-size requirements in both validation cohorts.

OSCC develops through the interaction of epithelial genomic alterations, exposure-related injury, chronic inflammation, stromal remodeling, and immune selection. These processes are not represented by a single molecular lesion. Instead, they produce coordinated changes in differentiation, proliferation, extracellular-matrix organization, metabolism, and host-response pathways. Transcriptome profiling is therefore useful because it captures the aggregate activity of malignant and non-malignant compartments at the time of tissue sampling. This systems-level view complements genomic characterization, which defines recurrent alterations but does not by itself reveal the direction or magnitude of downstream transcriptional activity in a specific specimen[2].

The interpretation of OSCC expression studies is complicated by biological and technical heterogeneity. Tumors arise at different oral subsites, vary in keratinization and stromal content, and are collected from

populations with different mixtures of tobacco, alcohol, areca-nut, viral, nutritional, and socioeconomic exposures. Control tissues may be obtained from healthy volunteers, anatomically adjacent mucosa, or surgical margins, each of which answers a somewhat different biological question. Microarray generations also differ in probe chemistry, transcript coverage, annotation age, and signal range. A signal observed in only one dataset may consequently reflect true context specificity, sampling variation, platform behavior, or an analytic artifact rather than a stable feature of OSCC.

Independent validation addresses this problem more directly than simply increasing the sample count through naïve pooling. When cohorts are combined before modeling, cohort identity can become entangled with disease status, especially if a platform contains mostly tumors and another contains mostly controls. Batch-correction methods cannot recover information that is absent from the design, and an apparently precise pooled estimate may conceal substantial between-study disagreement. An alternative is to analyze each cohort on its own scale and require a discovery signal to satisfy defined statistical, directional, and effect-size criteria in external datasets. This approach treats reproducibility as an empirical property of each gene rather than as an assumption imposed by a combined model.

The three GEO series used here provide a demanding validation setting. GSE30784 offers the largest independent tumor-control comparison; GSE25099 supplies a second independent-control cohort on a different Affymetrix platform; and GSE37991 provides patient-matched tumor and adjacent non-tumor tissue on an Illumina platform[8,9]. Agreement across these datasets therefore requires a signal to persist despite differences in platform, control definition, sample structure, and source population. Conversely, failure to replicate cannot be attributed to a single cause and should not automatically be interpreted as evidence that a discovery result is false. The design is intended to identify the most transportable signals, not to deny context-dependent biology.

Earlier analyses established that the same public cohorts contain shared OSCC expression abnormalities and potentially informative immune and prognostic features[8]. More recent integration work has emphasized platform-robust gene-pair predictors across a broader collection of OSCC datasets[9]. These contributions demonstrate the value of public transcriptomes, but they leave room for a deliberately staged test in which one cohort defines the hypothesis set and two other cohorts independently adjudicate it. Such a design can distinguish three outcomes—replication in both validation cohorts, replication in one cohort, or replication in neither—while maintaining the statistical identity of each study.

A second unresolved issue concerns the scale on which candidates should be prioritized. Ranking exclusively by the discovery effect or P value favors signals that are extreme in the largest cohort, even if their magnitude attenuates elsewhere. Ranking by a pooled estimate can likewise reward genes with high precision in one dataset. The minimum absolute cross-cohort effect used here is intentionally conservative: a gene is ranked according to its weakest observed effect among the discovery and validation cohorts. This criterion emphasizes consistency under unfavorable conditions and is particularly suitable when the purpose is to nominate candidates for expensive orthogonal validation rather than to maximize the apparent strength of a discovery list.

The novelty of this study lies in its cohort-preserving, locked gene-level replication framework: one discovery cohort defines the hypothesis set, two structurally different cohorts independently test each gene, and full replication requires stage-specific FDR control, concordant direction, a minimum validation effect, and stability to QC-based exclusions without pooling expression values or statistics. The primary aim of the study was to identify OSCC genes that are differentially expressed in GSE30784 and independently replicate in both GSE25099 and paired GSE37991. Secondary objectives were to rank candidates by their smallest cross-cohort effect, characterize direction-specific GO and KEGG programs, and assess whether the replication calls remained stable after array-level quality exclusions.

2. Methods

Study design and cohort eligibility

This retrospective secondary analysis included human microarray datasets with clearly identifiable primary OSCC and non-malignant oral tissue, accessible processed expression matrices, and sufficient sample annotation for tumor-control modeling. GSE30784, GSE37991, and GSE25099 met these criteria. Oral dysplasia samples in GSE30784 were excluded to retain a direct invasive-tumor versus healthy-control contrast. GSE37991 contained patient-matched tumor and adjacent non-tumor epithelium; the other two cohorts used independent healthy controls. The cohort designs, platforms, sample counts, control definitions, and mapped-gene totals are detailed in Table 1.

The unit of analysis was an individual microarray. The primary contrast was invasive OSCC tissue versus non-malignant oral tissue. Eligibility was evaluated at the cohort level before gene-wise testing so that dataset selection did not depend on the significance of any candidate gene. Cohorts were required to provide enough metadata to distinguish tumor from control tissue and to support the appropriate independent or paired design. The analysis did not treat dysplasia as either tumor or control because premalignant tissue occupies a biologically intermediate state and would blur the primary invasive-carcinoma contrast.

Cohort roles were fixed before the replication results were interpreted. GSE30784 was assigned to discovery because it contained the largest number of invasive OSCC samples and independent controls. GSE25099 and GSE37991 were retained as validation cohorts rather than rotated through multiple discovery assignments. This single discovery path prevents post hoc selection of the cohort configuration that yields the largest replicated set. The paired structure of GSE37991 was preserved because within-patient matching controls stable person-level characteristics that would otherwise contribute to residual variance.

Table 1. Characteristics of the included OSCC microarray cohorts.

Cohort	Platform	Array	Control	Paired	OSCC n	Ctrl n	Mapped genes
GSE30784	GPL570	Affymetrix U133 Plus 2.0	Independent healthy mucosa	No	167	45	20,848
GSE37991	GPL6883	Illumina HumanRef-8 v3	Matched adjacent non-tumor	Yes	40	40	17,544
GSE25099	GPL5175	Affymetrix Human Exon 1.0 ST	Independent healthy tissue	No	57	22	14,651

Data source: GEO accession records and the processed matrices reported in the source analysis. OSCC, oral squamous cell carcinoma; Ctrl, control.

Data acquisition and provenance

Series-matrix files and platform annotations were obtained from NCBI GEO[10]. GSE37991 detection P values were reconstructed from its supplementary non-normalized file. For GPL5175, probe-to-Entrez mappings were obtained from the huex10sttranscriptcluster annotation resource and converted to approved HGNC symbols. GSE30784 originated from a biomarker-discovery study of 167 OSCC, 17 dysplasia, and 45 healthy oral tissues; GSE25099 contained 57 OSCC and 22 normal tissues; and GSE37991 contained 40 matched pairs from men reporting alcohol, areca-nut, and cigarette exposure.

GEO accession identifiers served as stable provenance anchors linking expression matrices, platform annotations, sample records, and the originating publications. Files were handled at the cohort level and were not concatenated into a common sample matrix. Accession-level provenance was retained throughout preprocessing so that every fitted effect could be attributed to a specific source series and platform. This separation is important for auditability because the same gene symbol can be represented by different probe sets, dynamic ranges, and annotation conventions across arrays.

The analysis relied on deposited expression resources and their associated metadata rather than on newly generated biospecimens. Consequently, the reliability of phenotype labels and platform mappings was assessed from the information available in the repository. Where a field conflicted with better-supported sample descriptors, the conflict was documented and the ambiguous field was not used for modeling. No undocumented clinical characteristic was inferred from an accession title, publication narrative, author affiliation, or population stereotype.

Metadata curation and quality audit

Phenotypes were determined from GEO sample titles and source_name fields. In GSE37991, Sample_description was reversed relative to title and source_name and was excluded from modeling to avoid label inversion. A GEO processing note reports that 28 pairs were analyzed, but the deposited detection data did not identify that subset: all 40 pairs had at least 50% detected probes at $P < 0.05$ and $P < 0.01$, and no inclusion flag was present. All 40 complete pairs were therefore retained rather than applying an irreproducible order-based selection.

Array-level QC included missingness, median expression, interquartile range, median sample-to-sample correlation, robust z scores, and principal-component analysis (PCA). Arrays with extreme distributional or correlation metrics were flagged for sensitivity analysis rather than removed automatically from the primary model.

Metadata curation was performed before differential-expression modeling. Tumor-control labels were cross-checked across sample titles, source descriptors, and the paired identifiers available for GSE37991. The reversed Sample_description field in GSE37991 was treated as a metadata defect because its coding contradicted the concordant title and source_name information. Excluding that field from phenotype assignment avoided a complete inversion of the matched contrast. The decision was applied uniformly to all samples and did not depend on gene-expression results.

Quality assessment was intended to characterize unusual arrays without allowing automated outlier removal to determine the primary conclusion. Distributional summaries captured location and spread, sample-to-sample correlations assessed concordance with the broader cohort, robust z scores reduced sensitivity to extreme values, and PCA provided a multivariate view of sample structure. These diagnostics are complementary: an array can have a typical median yet show low correlation, or it can appear separated in PCA because of genuine biology rather than poor measurement. For this reason, flags defined a sensitivity set instead of functioning as unconditional exclusion rules.

PCA was calculated within each normalized cohort using the 1,000 most variable mapped genes. It was used as an unsupervised diagnostic and not as a classifier, hypothesis test, or basis for sample deletion. The orientation and numerical scale of principal components are arbitrary within each cohort; only the relative configuration of samples inside a panel was interpreted. Visual tumor-control separation was therefore considered evidence that disease status contributed to major expression variation, while residual overlap was retained as a feature of the observed data rather than forced into complete separation.

Expression preprocessing and gene annotation

Deposited matrices were inspected for scale. GSE37991 intensities were $\log_2$-transformed because upper quantiles remained on a raw-intensity scale; GSE30784 and GSE25099 were already log-scaled. Quantile normalization was then applied within each cohort. The root-mean-square change introduced by this step was 0.0527, 0.0018, and 0.1170 for GSE30784, GSE37991, and GSE25099, respectively. Probes mapping to an approved symbol were collapsed by the arithmetic mean. Expression matrices remained separate throughout the analysis.

Scale inspection preceded normalization because quantile normalization assumes that input values are represented on a comparable analytic scale within a cohort. The raw-intensity appearance of GSE37991 warranted log2 transformation, whereas applying another log transformation to already log-scaled matrices would compress genuine differences. Quantile normalization was

then performed separately for each series to reduce distributional variation without imposing a shared distribution across platforms. The reported root-mean-square changes indicate that this step altered the deposited matrices to different degrees, which is expected when repositories contain data processed under different original pipelines.

Probe-level measurements were mapped to approved gene symbols to establish a cross-platform unit of inference. When multiple probes mapped to the same symbol, their values were summarized by the arithmetic mean. This creates one estimate per gene per sample and prevents genes with many probes from entering the hypothesis set multiple times. The procedure does not resolve transcript isoforms or alternative promoters, and discordant probes can be averaged. It was selected as a transparent gene-level harmonization strategy appropriate to the study question, which concerns reproducible gene programs rather than transcript-specific regulation.

The intersection of 12,708 mapped symbols defined the common measurable-gene universe. This restriction was applied before discovery testing. It ensures that every gene eligible for discovery could be evaluated in both validation datasets and that functional enrichment used a background consistent with the actual testing opportunity. Genes present on only one or two platforms were excluded from cross-cohort inference even if they showed large effects locally. The resulting findings should therefore be understood as reproducible signals among shared measurable genes, not as a complete catalogue of all OSCC-associated transcripts.

Differential-expression analysis

Differential expression was estimated in R using limma linear models with trend and robust empirical-Bayes moderation[11]. Independent tumor-control contrasts were fitted for GSE30784 and GSE25099, while GSE37991 included patient-specific blocking for its matched design. Benjamini-Hochberg adjustment controlled the false-discovery rate (FDR). Positive effects represent higher expression in OSCC. Analyses were restricted to 12,708 gene symbols measured in all three cohorts so that every discovery hypothesis was testable in both validation cohorts.

Limma was chosen because empirical-Bayes moderation stabilizes gene-wise variance estimates across thousands of parallel tests and is well established for microarray analysis[11]. The robust option reduces the influence of genes with atypical variance behavior, while the trend option permits the prior variance to depend on average expression when such dependence is present. Each cohort was modeled separately so that platform-specific measurement distributions and residual variances were estimated from the appropriate samples. Effect signs were standardized to represent OSCC relative to non-malignant tissue.

For GSE30784 and GSE25099, the design matrix estimated an independent-group tumor-control contrast. For GSE37991, patient-specific blocking represented the matched nature of the 40 tumor/non-tumor pairs. Ignoring this pairing would treat correlated observations as independent and could inflate residual variability or distort uncertainty. The model did not include harmonized clinical covariates because comparable patient-level variables were not available across all three datasets. Adding different covariate sets to different cohorts would have changed the estimand and complicated replication interpretation.

Benjamini-Hochberg adjustment controlled the expected false-discovery proportion among rejected hypotheses within each analysis stage. FDR thresholds were not interpreted as probabilities that individual genes were true or false. Effect-size requirements supplemented statistical significance because large cohorts can detect small differences that may have limited practical relevance, while a gene with a large but imprecise effect may fail multiplicity control. The joint use of FDR, direction, and magnitude therefore defined replication more stringently than nominal P values alone.

Discovery and independent validation

GSE30784, the largest independent-control cohort, was designated as the discovery cohort in the locked analysis; GSE25099 and GSE37991 served as validation cohorts. Discovery DEGs required FDR < 0.05 and $|\log_2 FC| \geq 1.0$. Only these genes entered validation. In each validation cohort, two-sided limma P values were converted to direction-specific one-sided values according to the discovery direction; an opposite-direction estimate therefore received a P value greater than 0.5. Benjamini-Hochberg adjustment was applied separately within the discovery-gene set in each validation cohort. Full replication required a discovery-consistent direction, validation FDR < 0.05, and $|\log_2 FC| \geq 0.5$ in both validation cohorts. No expression values, effect sizes, standard errors, or P values were pooled across cohorts.

Fully replicated candidates were ranked by the minimum absolute $\log_2 FC$ observed across GSE30784, GSE25099, and GSE37991, followed by the worst stage-specific FDR. This ranking favors signals that remain large in their weakest cohort without creating a pooled summary estimate.

The validation test was conditional on discovery. Only the 1,167 genes meeting the discovery FDR and absolute-effect threshold were carried forward, thereby defining a finite family of replication hypotheses. In each validation cohort, the discovery direction specified the alternative of interest. A validation estimate in the opposite direction was not allowed to count as replication even if its two-sided P value was small. This protects the analysis from labeling contradictory biological effects as confirmation.

Multiplicity adjustment was performed separately within the carried-forward gene family for GSE25099 and GSE37991. A gene was called fully replicated only when it met the validation FDR, direction, and effect-size criteria in both cohorts. Passing one cohort was recorded as partial replication rather than combined with the other cohort through meta-analysis. This decision makes the evidence requirement transparent: neither a very strong GSE25099 result nor a very strong GSE37991 result can compensate for failure in the other validation dataset.

The minimum-effect ranking was calculated after replication status had been established. For each fully replicated gene, the absolute fitted log2 fold change was compared across all three cohorts, and the smallest value became the primary ranking quantity. The worst stage-specific FDR served as a secondary criterion. Because the rank is driven by the weakest cohort, it is resistant to a

single unusually large effect. The ranking is nevertheless statistical and does not incorporate protein abundance, assay availability, cellular localization, essentiality, safety, or clinical utility.

Functional enrichment

Upregulated and downregulated genes that replicated in both validation cohorts were analyzed separately with g:Profiler. GO Biological Process, Molecular Function, and Cellular Component and KEGG pathways were queried using the common set of tested genes as the custom background. Enrichment results were retained at adjusted $P < 0.05$. Disease- or infection-labeled terms were interpreted as gene-set overlap rather than evidence that individual tumors harbored the named disease or infection.

Functional enrichment was performed after gene-level replication and was stratified by direction. Separating upregulated from downregulated genes prevents opposing biological patterns from being combined into a single over-representation result. g:Profiler supplied the enrichment framework, while GO and KEGG provided complementary ontological and pathway vocabularies. The 12,708 shared tested genes were used as the custom universe so that enrichment probabilities reflected the genes that could have entered the replicated set on all three platforms.

Enrichment terms are overlapping summaries rather than independent mechanistic experiments. A single gene can contribute to several related pathways, and broad categories can be driven by partially shared membership. Adjusted P values identify over-representation relative to the selected universe but do not measure pathway activation in an individual patient, establish causal ordering, or indicate therapeutic dependence. Results were therefore synthesized into higher-level biological themes and interpreted alongside the direction and coherence of contributing genes.

Sensitivity analysis

The discovery-validation workflow was refitted after removal of QC-flagged arrays; for GSE37991, both members of an affected patient pair were removed. Thirteen GSE30784 arrays and six GSE37991 arrays representing three complete pairs were excluded; GSE25099 had no flagged arrays. Retention of the primary replication call was used to quantify stability.

The sensitivity analysis repeated the complete discovery-validation sequence after removal of QC-flagged arrays. Thresholds, cohort roles, gene universe, and ranking logic were unchanged. When a GSE37991 array was flagged, its paired counterpart was also removed so that the matched design remained intact. Stability was evaluated by asking whether primary fully replicated genes retained their classification, not by selecting whichever analysis produced more significant results.

3. Results

Cohort composition and sample structure

The primary analysis comprised 371 samples: 264 OSCC and 107 non-malignant controls. The three platforms contributed 54,675, 24,526, and 17,881 measured features before gene-level mapping. Following preprocessing, 20,848, 17,544, and 14,651 gene symbols were analyzed within GSE30784, GSE37991, and GSE25099, respectively. These cohort-level quantities are detailed in Table 1.

PCA of the 1,000 most variable mapped genes showed visible tumor-control structure in each cohort, strongest in GSE25099 and GSE37991 and more heterogeneous in GSE30784 as shown in Figure 1. Axes were computed within cohort and are not directly comparable across panels. Residual overlap is compatible with biological heterogeneity, varying stromal content, and the use of independent rather than matched controls in two cohorts.

The distribution of samples and mapped genes differed materially across cohorts, reinforcing the decision to avoid direct pooling. GSE30784 contributed most of the invasive tumors and therefore supplied the greatest discovery information, whereas GSE37991 contributed a balanced matched comparison. Despite platform-specific feature counts, the intersection strategy yielded a substantial shared gene set for formal cross-cohort testing. No single platform's larger probe count was allowed to increase its representation in the final hypothesis universe.

The PCA panels showed that disease status was a prominent but not exclusive source of expression variation. GSE25099 displayed a visually distinct tumor-control pattern, GSE37991 showed separation within a matched-tissue context, and GSE30784 retained broader dispersion among tumors. The more heterogeneous GSE30784 configuration is compatible with its larger sample and a wider mixture of tumor biology and tissue composition. Because PCA was unsupervised, these patterns were descriptive and did not enter the DEG or replication thresholds.

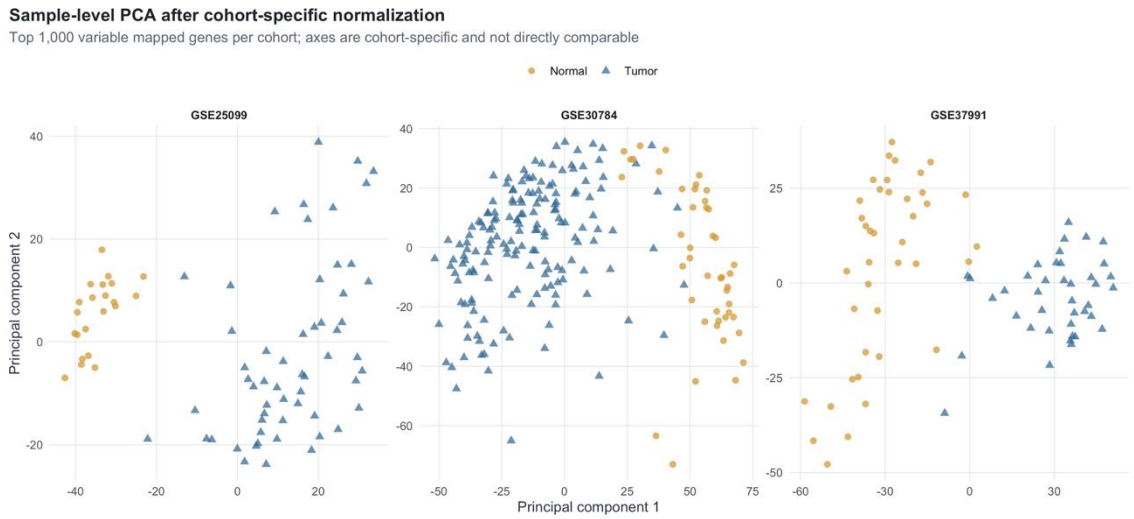

Figure 1. Sample-level principal-component analysis after cohort-specific normalization. PCA used the 1,000 most variable mapped genes within each cohort; colors and shapes distinguish non-malignant and OSCC tissue. Axes are cohort-specific.

Discovery and independent validation

Within the 12,708-gene common hypothesis space, GSE30784 yielded 1,167 discovery DEGs at FDR < 0.05 and $|\log_2 FC| \geq 1.0$: 583 were upregulated and 584 downregulated. Among these, 504 passed the full replication criteria in both GSE25099 and GSE37991: 273 were upregulated and 231 downregulated. A further 508 genes passed one validation cohort, whereas 155 passed neither. The discovery volcano plot displays these outcomes without combining statistics across cohorts as detailed in Figure 2.

The discovery result was nearly balanced by direction, with 583 upregulated and 584 downregulated genes. Full replication preserved a similar bidirectional structure: 273 genes were consistently higher and 231 consistently lower in OSCC. Thus, the reproducible signature was not dominated solely by activated oncogenic or inflammatory programs; it also contained a large component reflecting reduced epithelial and metabolic functions. The 508 genes that validated in only one cohort illustrate the selectivity of requiring agreement across two distinct validation designs.

The volcano plot positions discovery effect magnitude against discovery FDR while encoding subsequent replication outcomes. Fully replicated genes occur on both sides of the plot and across a range of discovery effects, indicating that validation was not limited to one directional class. Genes replicated in only one cohort and genes failing both validations remain visible, which prevents the figure from presenting the discovery list as uniformly reproducible. The plot is a staged evidence display rather than a pooled significance analysis.

Prioritized candidate genes

Ranking by the smallest absolute effect across all three cohorts prioritized CRISP3, MMP10, MMP13, MMP1, FAM3B, KRT4, TMPRSS11B, TYRP1, MMP12, and SERPINE1 as detailed in Table 2. Their weakest observed absolute $\log_2 FC$ ranged from 2.62 to 3.36, and every top-20 candidate retained full replication after QC exclusions. Cross-cohort effects for the top 40 candidates are detailed in Figure 3.

The top-ranked set combined four matrix metalloproteinases with genes linked to epithelial differentiation, secretory biology, proteolysis, pigmentation-related pathways, and fibrinolytic or stromal signaling. Five of the top ten candidates were upregulated and five were downregulated, reflecting the two-sided nature of OSCC remodeling. CRISP3 led the ranking because its weakest absolute effect remained larger than that of the other candidates, not because the analysis established it as the most causal or clinically useful gene.

All top-20 candidates passed the sensitivity analysis, which strengthens confidence that their replication classifications were not dependent on the QC-flagged arrays. This stability does not remove platform or tissue-composition biases, but it shows that the highest-ranked signals survived a meaningful perturbation of the included sample set. The 40-gene heat-map display further demonstrates that the direction of fitted effects was concordant across cohorts even when magnitudes differed.

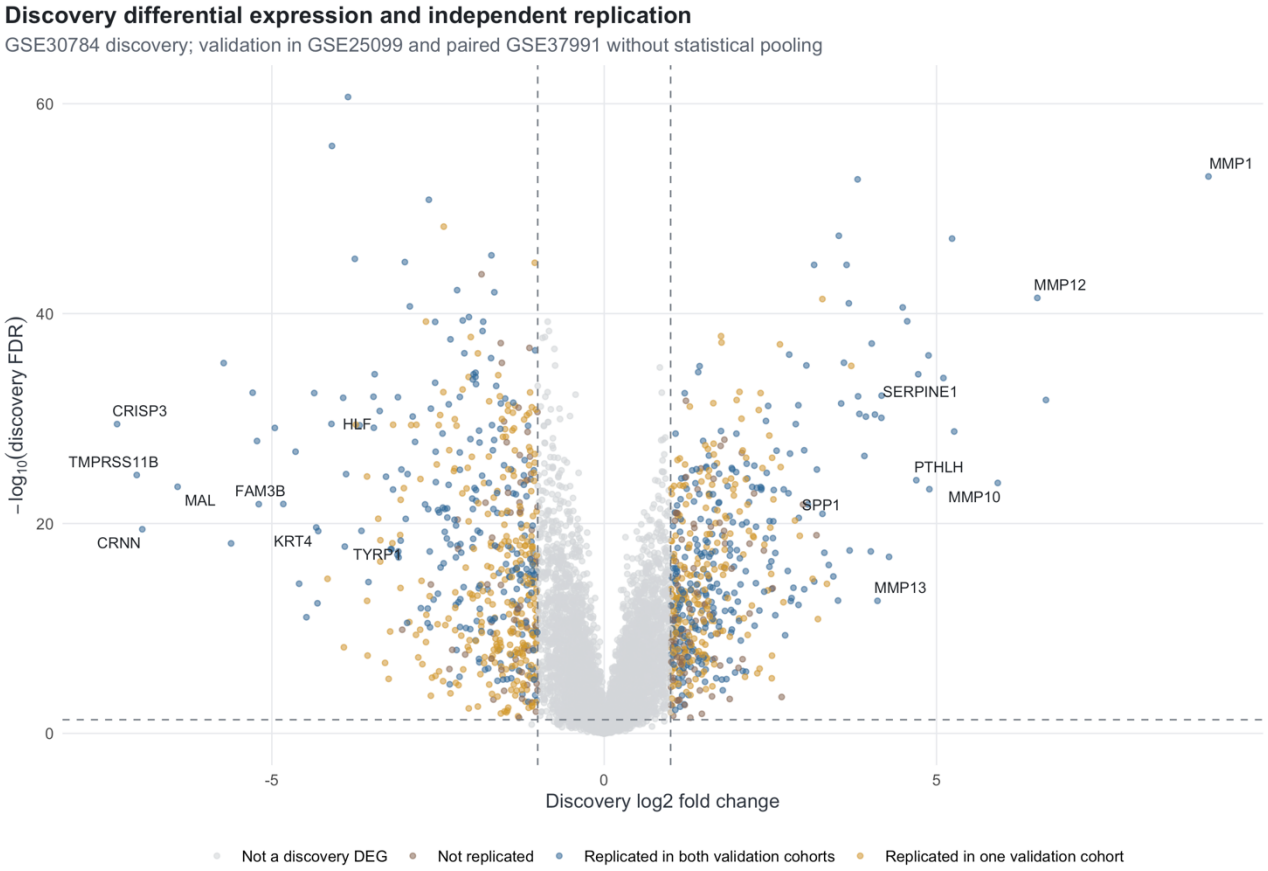

Figure 2. Discovery differential expression and independent replication. The x-axis and FDR are from GSE30784. Full replication required discovery-consistent direction, validation FDR < 0.05, and $|\log_2 FC| \geq 0.5$ in both validation cohorts.

Functional enrichment

Upregulated replicated genes were enriched for viral protein interaction with cytokines and cytokine receptors (adjusted P=2.99 × 10^{-7}), cytokine-cytokine receptor interaction (3.14 × 10^{-7}), ECM-receptor interaction (3.14 × 10^{-7}), integrin signaling (7.01 × 10^{-7}), IL-17 signaling (7.03 × 10^{-7}), focal adhesion (3.10 × 10^{-6}), and PI3K-Akt signaling (4.08 × 10^{-6}). GO results reinforced immune and defense responses, extracellular-matrix structural constituents, receptor-ligand activity, and chemokine activity.

Downregulated replicated genes were enriched for cornified-envelope formation (adjusted P=5.34 × 10^{-7}), tyrosine metabolism (4.76 × 10^{-3}), and arachidonic-acid metabolism (4.76 × 10^{-3}). GO terms emphasized lipid and fatty-acid metabolism and oxidoreductase and monooxygenase activities.

The upregulated enrichment results converged on interacting themes rather than isolated pathways. Cytokine and chemokine activity describes inflammatory communication; ECM-receptor and integrin categories describe cell-matrix engagement; focal adhesion links those extracellular cues to the cytoskeleton; and PI3K-Akt provides a downstream survival and growth axis. Because these categories share genes and biological relationships, their concordance is more informative than treating each adjusted P value as evidence for an independent mechanism.

The downregulated results formed a complementary theme of diminished specialized epithelial function. Cornified-envelope genes reflect terminal differentiation and barrier formation, while lipid, fatty-acid, arachidonic-acid, tyrosine, oxidoreductase, and monooxygenase categories describe metabolic activities of mature mucosa. Their coordinated reduction is consistent with dedifferentiation and altered tissue composition. The results do not determine whether each change arises within malignant cells or from differences in the proportion of normal epithelial, stromal, and immune compartments.

Sensitivity to QC-based exclusions

The sensitivity analysis retained 199 of 212 GSE30784 samples, 74 of 80 GSE37991 samples (37 complete pairs), and all 79 GSE25099 samples. It identified 515 fully replicated genes; 488 of the 504 primary calls retained full replication (96.8%). All top 20 candidates retained their direction, statistical classification, and validation effect threshold.

Removing flagged arrays modestly changed the total number of fully replicated genes from 504 to 515, while 488 primary calls retained full replication. The appearance of additional genes in the sensitivity analysis is compatible with changes in fitted effects, variance estimates, and multiplicity adjustment after sample exclusion. The primary analysis remained the main inferential result because it followed the inclusive design, whereas the refitted analysis tested robustness. The 96.8% retention rate indicates substantial stability without implying that the two analyses were identical.

Table 2. Top 10 independently replicated candidate genes ranked by minimum cross-cohort effect.

Rank	Gene	Direction	GSE30784 log_2FC	GSE25099 log_2FC	GSE37991 log_2FC	Minimum $\lvert log_2FC \rvert$	Worst FDR	Sensitivity
1	CRISP3	Down	−7.33	−3.36	−3.78	3.36	1.98×10^{-10}	Yes
2	MMP10	Up	5.92	4.73	3.33	3.33	2.20×10^{-17}	Yes
3	MMP13	Up	4.11	3.27	5.28	3.27	2.05×10^{-9}	Yes
4	MMP1	Up	9.09	6.51	3.22	3.22	1.11×10^{-18}	Yes
5	FAM3B	Down	−4.83	−3.20	−3.56	3.20	1.55×10^{-19}	Yes
6	KRT4	Down	−4.33	−3.52	−3.10	3.10	5.31×10^{-9}	Yes
7	TMPRSS11B	Down	−7.03	−5.73	−2.93	2.93	8.71×10^{-9}	Yes
8	TYRP1	Down	−3.06	−4.20	−2.91	2.91	6.82×10^{-12}	Yes
9	MMP12	Up	6.52	4.16	2.74	2.74	1.87×10^{-12}	Yes
10	SERPINE1	Up	5.10	2.62	2.83	2.62	2.51×10^{-14}	Yes

FDR, false-discovery rate; log_2FC, log_2 fold change. Sensitivity indicates retention of the full-replication call after QC-based exclusions. Candidates are expression-based hypotheses, not clinically validated therapeutic targets.

4. Discussion

Principal findings and relation to prior work

This study identified 504 genes that met predefined criteria in one discovery cohort and two independent validation cohorts. The replicated genes defined coordinated extracellular-matrix remodeling, integrin signaling, focal adhesion, PI3K-Akt activity, and inflammatory/immune responses, accompanied by depletion of cornified-envelope and lipid-related metabolic programs. These processes map to central capabilities of cancer[12]. Requiring successful validation in both cohorts reduced the influence of platform-specific findings without statistical pooling.

The findings overlap biologically with an earlier common-DEG analysis of the same three datasets, which also highlighted extracellular-matrix and immune programs[8]. They also complement a broader gene-pair integration study that prioritized predictive generalization across multiple platforms[9]. The present analysis differs by preserving each cohort's scale and testing a locked gene-level replication rule with separate multiplicity control, direction concordance, effect-size thresholds, and a QC-exclusion sensitivity analysis.

The principal contribution is not merely the size of the replicated list. It is the demonstration that a broad OSCC expression program persists across two independent validation contexts when cohort identity, platform scale,

and pairing are respected. The simultaneous presence of increased matrix/inflammatory activity and reduced epithelial-metabolic activity supports a model in which tumor progression involves both acquisition of invasive microenvironmental interactions and loss of differentiated mucosal functions. This dual pattern is consistent with contemporary views of cancer as an ecosystem rather than a collection of autonomous malignant cells[12].

Comparison with Xie et al. is important because that study used the same three GEO series and also identified common DEGs, candidate targets, and immune-cell signals[8]. Agreement at the thematic level provides external plausibility for the prominence of matrix and immune biology. The present design asks a narrower but more stringent question: which discovery-defined genes satisfy separate validation FDR, direction, and effect thresholds in both an independent-control and a paired-control cohort? The difference is methodological and inferential; it should not be interpreted as invalidating the earlier intersection approach.

The gene-pair framework described by Li et al. addresses another complementary goal: prediction that is less sensitive to between-platform scale differences[9]. Pairwise ordering can be advantageous for classifiers because relative expression within a sample may transfer across measurement systems. The current study instead retains gene-level effect estimates and does not build a diagnostic or prognostic model. It therefore prioritizes interpretable, directionally reproducible biological candidates while accepting that additional work is needed to determine whether they support accurate individual-level classification.

ECM-integrin-focal-adhesion remodeling

The convergence of integrin signaling, ECM-receptor interaction, focal adhesion, extracellular-matrix structural constituents, and extracellular-region terms is biologically coherent. The tumor-associated ECM provides biochemical and mechanical cues that regulate proliferation, survival, invasion, angiogenesis, and immune behavior[13–18]. Integrins translate these cues into cytoskeletal and signaling changes that support migration and metastatic progression[19,20]. The signal was supported by multiple independently replicated genes rather than a single extreme observation.

This axis also provides a plausible bridge to PI3K-Akt enrichment. Nevertheless, bulk-tissue microarrays cannot determine whether the signal originates from malignant epithelial cells, cancer-associated fibroblasts, endothelial cells, or mixed compartments. Single-cell or spatial profiling is required before assigning cellular causality.

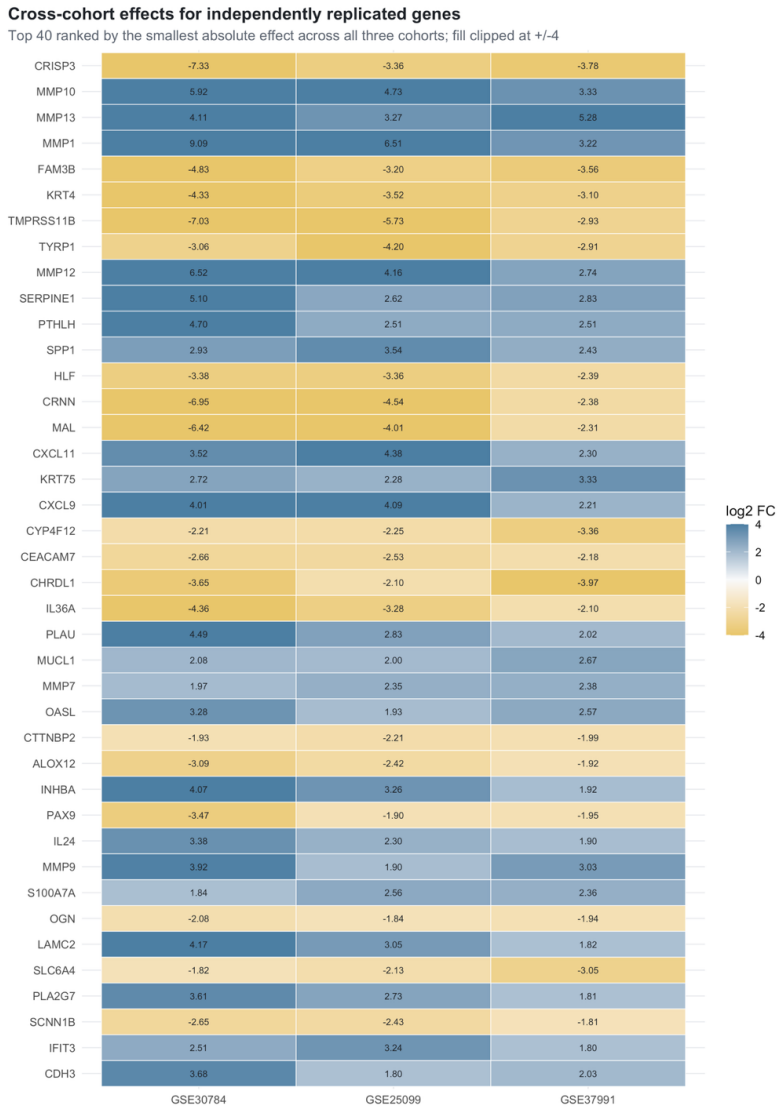

Figure 3. Cross-cohort log_2 fold changes for 40 independently replicated genes. Rows are ordered by the smallest absolute effect across cohorts; the color scale is clipped at ±4 while cell labels show cohort-specific fitted effects.

Extracellular matrix is an active regulator of cancer behavior. Changes in matrix composition, cross-linking, stiffness, and proteolysis affect receptor signaling, cell polarity, migration, vascular behavior, and access of immune cells to tumor regions[13–18]. The replicated elevation of several MMP genes is compatible with enhanced matrix turnover, while enrichment of structural-constituent and extracellular-region terms indicates that the signature extends beyond a single protease family. Because bulk expression measures transcript abundance, it cannot specify whether these genes are produced by carcinoma cells, fibroblasts, inflammatory cells, or combinations of these compartments.

Integrins provide a mechanistic bridge between altered matrix and intracellular responses. They organize adhesion complexes, interact with growth-factor receptors, and transmit mechanical information to cytoskeletal and survival pathways. Their effects depend on receptor composition, ligand context, cellular state, and tissue mechanics; increased expression of pathway members is therefore not equivalent to uniform integrin activation. Nevertheless, replicated integrin and focal-adhesion enrichment across platforms supports the interpretation that cell-matrix communication is a stable component of the OSCC transcriptomic phenotype[19,20].

The accompanying PI3K-Akt enrichment should be interpreted at the program level. PI3K-Akt signaling can be influenced by genomic alterations, receptor tyrosine kinases, integrin engagement, cytokines, and stromal cues. The present data do not distinguish these upstream routes or demonstrate phosphorylation-dependent pathway activity. They show that the replicated upregulated gene set overlaps a pathway vocabulary connected to survival and growth. Protein-level assays, phosphoproteomics, and perturbation experiments would be required to establish functional dependence or therapeutic sensitivity.

Angiogenic and immune candidates

VEGFC and ANGPT2 met the full replication criteria and retained their calls in the sensitivity analysis. Recent OSCC experiments have independently linked hypoxia-responsive extracellular vesicles and circRNA programs to angiogenesis and metastatic behavior[21,22], providing orthogonal biological context for the vascular signal. However, the present analysis does not test whether VEGFC or ANGPT2 is essential for tumor growth or whether pharmacological inhibition would benefit patients.

OAS3 and NLRC5 also replicated in both cohorts and connect the result to interferon and antigen-presentation biology. Recent OSCC studies have documented macrophage-linked mesenchymal programs, IL-6-inflammasome signaling, immune-gene signatures, and tumor-cell mechanisms that shape immunosuppressive phenotypes[23–28]. Higher bulk NLRC5 expression may reflect tumor-cell interferon signaling, increased immune-cell content, or compensatory antigen-presentation activity; OAS3 likewise marks an interferon-responsive program but does not establish active viral infection.

Angiogenic and inflammatory processes are closely linked in the tumor microenvironment. VEGFC and ANGPT2 can participate in vascular remodeling, endothelial activation, and communication between malignant and stromal compartments. Contemporary OSCC experiments showing hypoxia-exosome- and circRNA-associated vascular programs support the biological relevance of this theme without validating the present transcript-based candidates as prognostic markers[21,22]. Differences in assay platform, tissue localization, clinical stage, and endpoint definition preclude a direct translation from replicated expression to patient risk.

The immune-related signature contains both signaling and antigen-presentation components. Cytokine-receptor and chemokine enrichment may reflect active immune recruitment, tumor-cell cytokine production, stromal responses, or changes in the abundance of infiltrating leukocytes. An elevated immune program can coexist with immune evasion because tumors may induce inflammatory signals while disabling effective antigen recognition or effector function. Bulk microarrays cannot resolve these possibilities, and an enrichment label should not be interpreted as evidence of a uniformly inflamed or treatment-responsive tumor state.

NLRC5 is particularly relevant because it regulates transcription of MHC class I pathway genes. Higher bulk NLRC5 expression could represent preserved antigen-presentation capacity, interferon stimulation, compensatory response, or increased immune-cell content. The range of immune phenotypes reported in recent OSCC tissue and experimental studies underscores why bulk expression cannot be equated with effective antitumor immunity[23–28]. OAS3 similarly indicates an interferon-responsive state but cannot establish the presence of a specific virus. These candidates merit cell-resolved analysis that separates malignant epithelium from infiltrating and stromal populations and measures both transcript and protein-level antigen-presentation machinery.

Microbial context offers another plausible contributor to inflammatory and metabolic heterogeneity. Recent OSCC studies have associated intratumoral microbial composition with stage and immune signatures, linked Fusobacterium nucleatum to GLUT1-driven lactate production, and identified host-microbe proteomic pathways relevant to cancer progression[29–31]. These observations support mechanistic follow-up but do not establish microbial causation or active infection in the microarray specimens analyzed here.

Loss of epithelial differentiation and metabolic homeostasis

Cornified-envelope formation was the strongest downregulated KEGG term, while lipid, fatty-acid, arachidonic-acid, oxidoreductase, monooxygenase, and tyrosine metabolic functions were also depleted. This pattern is consistent with loss of mature mucosal differentiation and altered epithelial homeostasis, but it may also reflect tissue-composition differences. Replication across adjacent and independent control designs is informative without eliminating confounding.

The downregulated program is biologically important because cancer studies often emphasize induced pathways and understate the loss of normal tissue identity. Cornified-envelope and keratinization-related functions contribute to the structural barrier of stratified squamous epithelium. Reduced expression can reflect dedifferentiation within malignant cells, replacement of normal epithelium by tumor and stroma, or both. The prominent reduction of KRT4 and TMPRSS11B among prioritized genes is consistent with loss of mature mucosal features, although their cellular sources were not directly measured.

Metabolic depletion should also be interpreted in relation to tissue identity. Reduced lipid, fatty-acid, arachidonic-acid, tyrosine, oxidoreductase, and monooxygenase annotations may reflect altered substrate use by tumor cells, diminished representation of specialized epithelial functions, or compositional differences between tumor and control specimens. These alternatives are not mutually exclusive. Metabolomics, isotope tracing, spatial transcriptomics, and cell-type-specific expression analysis would be needed to determine whether the observed transcript pattern represents a tumor-intrinsic metabolic dependency.

Replication across healthy-control and adjacent-tissue designs strengthens the inference that the downregulated program is not restricted to one comparator. However, adjacent mucosa may carry field effects related to exposure or tumor proximity, while healthy controls may differ from cases in age, behavior, or sampling context. A signal that persists across both comparisons is transportable across these designs, but the magnitude of the fitted effect can still include different mixtures of biological contrast and confounding.

Independent contemporary OSCC studies have nominated expression-based prognostic candidates and described ferroptosis, glycolysis, mitochondrial respiration, hypoxia response, and autophagy-related mechanisms involving NCBP2/TFRC, FTO, CYTOR, circFOXK2, PER1, and NUPR1[32–37]. Their convergence with the present metabolic and differentiation themes provides biological context, while differences in model systems and endpoints prevent direct equivalence with the cross-cohort replication results.

Interpretation of candidate ranking

The conservative ranking emphasized effects that remained large in the weakest cohort. MMP1, MMP10, MMP12, and MMP13 form a coherent matrix-remodeling module, while SERPINE1, PTHLH, and SPP1 connect invasion, stroma, and signaling. CRISP3, FAM3B, KRT4, TMPRSS11B, and TYRP1 were among the most strongly reduced genes. Statistical rank is not equivalent to target tractability or causal importance and does not account for protein localization, tissue specificity, survival association, dependency, or druggability.

The minimum-effect ranking is useful for choosing a compact validation panel because it penalizes candidates whose effects collapse in any cohort. A candidate with a spectacular discovery effect but a modest validation effect ranks below a gene with consistently large effects. This behavior is intentional and differs from ranking by average effect, maximum effect, or smallest P value. It aligns the statistical priority with the operational question of which signals are most likely to remain detectable when transferred to a new assay or population.

The matrix metalloproteinase cluster offers internal biological coherence, yet multiple members of one pathway may provide redundant information in a small assay panel. Conversely, strongly downregulated candidates such as CRISP3, FAM3B, KRT4, TMPRSS11B, and TYRP1 may capture complementary loss-of-identity signals. A future panel-selection strategy should balance magnitude, pathway coverage, assay performance, tissue specificity, correlation among candidates, and incremental value. Those criteria cannot be optimized from the reported group-level effects alone.

Candidate status should be separated from biomarker status. A clinically useful diagnostic biomarker requires a locked measurement procedure, a prespecified decision threshold, blinded evaluation, and performance estimates with uncertainty in an appropriate target population. A prognostic biomarker additionally requires outcome follow-up and adjustment for established clinical factors. A therapeutic target requires evidence of functional dependency and a feasible intervention. None of these standards is met by cross-sectional differential expression alone, even when replication is strong.

Strengths and limitations

Strengths include separation of discovery from validation, validation across two control designs, preservation of the paired GSE37991 structure, platform-specific preprocessing, a common gene hypothesis space, stage-specific false-discovery control, explicit direction and effect-size requirements, a custom enrichment background, and QC-exclusion sensitivity analysis. Expression values and statistics were not pooled across platforms.

Several limitations remain. Processed series matrices were used, so raw CEL- or image-level QC was not recomputed. The independent-control cohorts may retain confounding by age, exposure, anatomic subsite, sampling, and batch. The reversed GSE37991 description field and an unreconstructable note regarding a 28-pair subset introduce metadata uncertainty. Probe-to-gene summarization and historical annotation can obscure transcript isoforms. Over-representation results depend on database definitions and overlapping gene sets. Bulk expression cannot localize signals to tumor, stromal, or immune cells. No survival, treatment-response, mutation, copy-number, methylation, or experimental dependency data were analyzed. Finally, the code and locked analysis archive were not available in the material reviewed for this revision, so the reported numerical results were not independently reproduced.

A major strength is the explicit separation of hypothesis generation from validation. The discovery cohort determined which genes were tested, while the validation cohorts determined whether those signals reproduced. Directional testing and stage-specific multiplicity control reduced the opportunity for contradictory or nominally significant findings to be labeled as confirmation. The minimum-effect ranking and sample-exclusion sensitivity analysis added two distinct robustness perspectives: consistency of magnitude across cohorts and stability to influential arrays.

A second strength is preservation of design structure. GSE37991 was analyzed as matched pairs, whereas the independent-control cohorts were modeled as independent groups. This respects the sampling process of each series and avoids the false simplicity of a single pooled matrix. The common gene universe also ensures that replication and enrichment were conditioned on comparable measurement opportunity. Together, these choices favor interpretable transportability over maximal sample size.

Several limitations arise from secondary use of public microarrays. Processed matrices restrict the ability to repeat image-level diagnostics, background correction, and probe-level normalization from raw files. Historical annotations may map probes imperfectly to contemporary

gene symbols, and mean collapsing can conceal discordant probe behavior or isoform-specific regulation. Platform intersection removes genes that are not jointly measurable, which improves comparability but may exclude important OSCC biology represented only on newer or denser arrays.

Clinical and pathological heterogeneity could not be modeled comprehensively. Harmonized information on age, sex, exposure intensity, anatomic subsite, stage, grade, HPV status, stromal content, treatment, and outcome was not available for all cohorts. Unmeasured imbalance can influence both discovery and validation effects. The paired design reduces stable person-level confounding within GSE37991, but adjacent non-tumor tissue can exhibit field cancerization and cannot be equated with tissue from a cancer-free individual.

The GSE37991 metadata conflict is a specific source of uncertainty. Sample titles and source_name fields supported one phenotype assignment, whereas Sample_description was reversed. The deposited note concerning a 28-pair subset could not be reconstructed from an explicit inclusion flag, so retaining all 40 complete pairs was more reproducible than choosing samples by order. This decision increases transparency but may not reproduce an undocumented subset used by the original investigators. The issue is disclosed because alternative subset definitions could affect effect estimates.

Statistical replication does not eliminate winner's curse, residual batch effects, or selective availability of public datasets. Discovery effects may be larger than effects expected in a future cohort, and the three selected series do not represent every geographic, behavioral, anatomical, or molecular context of OSCC. The study evaluates transportability across the included datasets, not universal invariance. External validation in newly collected specimens remains necessary.

Functional enrichment adds biological organization but has limitations. Ontology terms overlap, pathway databases are incomplete and version-dependent, and over-representation treats genes as members of predefined sets without accounting for pathway topology or cell type. Adjusted enrichment significance does not indicate that every component is active or that a pathway is required for tumor maintenance. The reported themes should guide mechanistic experiments rather than substitute for them.

Finally, the source material did not include the stated locked code, analysis notebook, executable environment, or checksum archive. The editorial workflow could verify internal consistency, accession metadata, citation metadata, and document structure, but it could not rerun the numerical analysis independently. This limitation is material to computational reproducibility and should be resolved by archiving the analysis package before publication whenever it becomes available.

Implications

The immediate priority is orthogonal validation of a compact panel spanning ECM/adhesion, angiogenesis, immune/interferon response, and epithelial metabolism. qRT-PCR should be performed in an independent, prospectively defined tumor-normal cohort with subsite, stage, HPV status, tobacco, alcohol, and areca-nut exposure recorded. Protein-level and spatial assays should establish abundance and cellular localization, while genomic and perturbation studies are needed before diagnostic, prognostic, or therapeutic claims.

The first translational step should be technical and biological confirmation rather than immediate classifier construction. A compact panel should include representatives of matrix remodeling, epithelial differentiation, immune/interferon signaling, angiogenesis, and metabolism. Assays should be optimized for linearity, detection limits, repeatability, tissue input, normalization, and pre-analytic handling. Candidate selection should be finalized before outcome evaluation to avoid optimistic performance estimates.

Validation specimens should be collected under a protocol that defines oral subsite, tumor content, control-tissue source, stage, grade, HPV status, and major behavioral exposures. Matched tumor-normal sampling can control person-level factors, whereas an independent healthy-control group can test diagnostic contrast; both designs answer useful but different questions. Analyses should report effect sizes and uncertainty and should evaluate whether candidate signals add information beyond routine clinical and pathological variables.

Cellular localization is essential for interpreting the replicated pathways. Spatial transcriptomics, multiplex immunohistochemistry, in situ hybridization, or single-cell approaches could determine whether matrix, cytokine, antigen-presentation, and metabolic signals arise from malignant epithelium, fibroblasts, endothelium, or immune populations. Protein and phosphoprotein measurements would establish whether transcript changes correspond to abundance and pathway activity. Perturbation experiments would then be required to distinguish correlates from functional dependencies.

A future diagnostic study should use a locked model and threshold and should report discrimination, calibration, sensitivity, specificity, predictive values, and decision consequences in the intended-use population. A prognostic study should define the time origin, endpoint, follow-up, censoring, and adjustment variables. Therapeutic development would require target engagement, dependency, selectivity, and safety evidence. These stages should remain conceptually separate so that promising replicated genes are not advanced through unsupported claims.

The translational pathway should also be informed by recent OSCC trials and assay studies. Neoadjuvant anti-PD-1/antiangiogenic and immunochemotherapy studies demonstrate the feasibility of tissue-rich clinical evaluation, including single-cell and stromal-niche analyses[38–41]; fluorescence-guided margin imaging and tumor glycoprofiling illustrate complementary routes for orthogonal tissue validation[42,43]. These advances justify prospective testing but do not confer clinical validity on the present candidates.

5. Conclusion

A cohort-preserving discovery-validation analysis of 371 oral tissue microarrays identified 504 OSCC genes that replicated independently in two validation cohorts. The reproducible signature combined ECM-integrin-focal-adhesion remodeling with PI3K-Akt and immune programs and reduced epithelial and metabolic functions. These results provide expression-based candidates and mechanistic hypotheses for orthogonal validation but do

not establish mutation, cellular origin, prognostic utility, or therapeutic actionability.

By requiring concordant statistical and biological evidence in two validation cohorts, the analysis narrows a broad discovery signature to genes and pathways that are more resistant to platform and control-design differences. The resulting candidates provide a rational foundation for focused laboratory and clinical validation. Their principal value at this stage is as reproducible hypotheses about OSCC tissue biology, not as completed biomarkers or therapeutic targets.

Declarations

Ethics approval: Ethical approval was granted by the CMHC Ethics Committee, Indonesia (No. 004/CMHC/Ethics/08/2026). The present study used publicly available, de-identified GEO datasets and involved no new participant recruitment or access to identifiable information.

Consent for publication: Not applicable to the de-identified public data analyzed in this study.

Data availability: The primary data are publicly available from NCBI GEO under accession numbers GSE30784, GSE37991, and GSE25099.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of interest: The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this article.

Author contributions (CRediT): AP: Conceptualization, methodology, formal analysis, investigation, data curation, visualization, and writing - original draft. HS: Conceptualization, methodology, validation, supervision, and writing - review and editing. Both authors reviewed and approved the final manuscript and accept responsibility for the integrity of the work.

Use of artificial intelligence: The authors declare that no generative artificial intelligence tools were used in the conceptualization, analysis, writing, editing, or preparation of this manuscript.

References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229-263. doi: 10.3322/caac.21834.
2. Wu CS, Li HP, Hsieh CH, et al. Integrated multi-omics analyses of oral squamous cell carcinoma reveal precision patient stratification and personalized treatment strategies. Cancer Lett. 2025;614:217482. doi: 10.1016/j.canlet.2025.217482.
3. Arora R, Cao C, Kumar M, et al. Spatial transcriptomics reveals distinct and conserved tumor core and edge architectures that predict survival and targeted therapy response. Nat Commun. 2023;14(1):5029. doi: 10.1038/s41467-023-40271-4.
4. Galeano Niño JL, Wu H, LaCourse KD, et al. Effect of the intratumoral microbiota on spatial and cellular heterogeneity in cancer. Nature. 2022;611(7937):810-817. doi: 10.1038/s41586-022-05435-0.
5. Zhi Y, Wang Q, Zi M, et al. Spatial transcriptomic and metabolomic landscapes of oral submucous fibrosis-derived oral squamous cell carcinoma and its tumor microenvironment. Adv Sci (Weinh). 2024;11(12):2306515. doi: 10.1002/advs.202306515.
6. Liu Z, Zhang Z, Zhang Y, et al. Spatial transcriptomics reveals that metabolic characteristics define the tumor immunosuppression microenvironment via iCAF transformation in oral squamous cell carcinoma. Int J Oral Sci. 2024;16(1):9. doi: 10.1038/s41368-023-00267-8.
7. Sun L, Kang X, Wang C, et al. Single-cell and spatial dissection of precancerous lesions underlying the initiation process of oral squamous cell carcinoma. Cell Discov. 2023;9(1):28. doi: 10.1038/s41421-023-00532-4.
8. Xie P, Wu S, Guo L, et al. Identification of candidate target genes and immune cells in oral squamous cell carcinoma. Comput Math Methods Med. 2021;2021:5802110. doi: 10.1155/2021/5802110.
9. Li N, Hu Z, Zhang N, et al. Pairwise analysis of gene expression for oral squamous cell carcinoma via a large-scale transcriptome integration. J Cell Mol Med. 2024;28(20):e70153. doi: 10.1111/jcmm.70153.
10. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207-210. doi: 10.1093/nar/30.1.207.
11. 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. doi: 10.1093/nar/gkv007.
12. Liu Y, Yang Z, Pu JJ, et al. Proteogenomic characterisation of primary oral cancer unveils extracellular matrix remodelling and immunosuppressive microenvironment linked to lymph node metastasis. Clin Transl Med. 2025;15(3):e70261. doi: 10.1002/ctm2.70261.
13. Hu S, Lu H, Xie W, et al. TDO2+ myofibroblasts mediate immune suppression in malignant transformation of squamous cell carcinoma. J Clin Invest. 2022;132(19):e157649. doi: 10.1172/jci157649.
14. Ding L, Fu Y, Zhu N, et al. OXTR-high stromal fibroblasts control the invasion pattern of oral squamous cell carcinoma via ERK5 signaling. Nat Commun. 2022;13(1):5124. doi: 10.1038/s41467-022-32787-y.
15. Zhang JY, Zhu WW, Wang MY, et al. Cancer-associated fibroblasts promote oral squamous cell carcinoma progression through LOX-mediated matrix stiffness. J Transl Med. 2021;19(1):513. doi: 10.1186/s12967-021-03181-x.
16. Liu X, Li J, Yang X, et al. Carcinoma-associated fibroblast-derived lysyl oxidase-rich extracellular vesicles mediate collagen crosslinking and promote epithelial-mesenchymal transition via p-FAK/p-paxillin/YAP signaling. Int J Oral Sci. 2023;15(1):32. doi: 10.1038/s41368-023-00236-1.
17. Zhao H, Li R, Chen Y, et al. Stromal nicotinamide N-methyltransferase orchestrates the crosstalk between fibroblasts and tumour cells in oral squamous cell carcinoma: evidence from patient-derived assembled organoids. Oncogene. 2023;42(15):1166-1180. doi: 10.1038/s41388-023-02642-5.
18. Li Y, Ma Z, Li W, et al. PDPN-positive cancer-associated fibroblasts facilitate OSCC-cell motility by inhibiting ferroptosis via exosomal lncRNA FTX. Cell Death Dis. 2023;14(11):759. doi: 10.1038/s41419-023-06280-3.
19. Jang TH, Huang WC, Tung SL, et al. MicroRNA-485-5p targets keratin 17 to regulate oral cancer stemness and chemoresistance via the integrin/FAK/Src/ERK/beta-catenin pathway. J Biomed Sci. 2022;29(1):42. doi: 10.1186/s12929-022-00824-z.
20. Wen Z, Zhang Y, Wang X, et al. THBS1-mediated degradation of collagen via the PI3K/AKT pathway facilitates metastasis and poor prognosis of OSCC. Int J Mol Sci. 2023;24(17):13312. doi: 10.3390/ijms241713312.
21. Capik O, Gumus R, Karatas OF. Hypoxia-induced tumor exosomes promote angiogenesis through the miR-1825/TSC2/mTOR axis in oral squamous cell carcinoma. Head Neck. 2023;45(9):2259-2273. doi: 10.1002/hed.27460.
22. Li X, Wang C, Zhang H, et al. circFNDC3B accelerates vasculature formation and metastasis in oral squamous cell carcinoma. Cancer Res. 2023;83(9):1459-1475. doi: 10.1158/0008-5472.can-22-2585.
23. You Y, Tian Z, Du Z, et al. M1-like tumor-associated macrophages cascade a mesenchymal/stem-like phenotype of oral squamous cell carcinoma via the IL6/STAT3/THBS1 feedback loop. J Exp Clin Cancer Res. 2022;41(1):10. doi: 10.1186/s13046-021-02222-z.
24. Xiao L, Li X, Cao P, et al. Interleukin-6-mediated inflammasome activation promotes oral squamous cell carcinoma progression via JAK2/STAT3/SOX4/NLRP3 signaling. J Exp Clin Cancer Res. 2022;41(1):166. doi: 10.1186/s13046-022-02376-4.

25. Chen Y, Feng Y, Yan F, et al. An immune-related gene signature to identify the tumor microenvironment and prognosis in oral squamous cell carcinoma using ssGSEA: a bioinformatics and biological validation study. Front Immunol. 2022;13:922195. doi: 10.3389/fimmu.2022.922195.
26. Wang S, Wang X, Sun J, et al. Down-regulation of FEN1 inhibits OSCC growth by affecting immunosuppressive phenotypes via IFN-gamma/JAK/STAT1. Int J Oral Sci. 2023;15(1):17. doi: 10.1038/s41368-023-00221-8.
27. Shi Y, Ren X, Cao S, et al. TP53 gain-of-function mutation modulates the immunosuppressive microenvironment in non-HPV-associated oral squamous cell carcinoma. J Immunother Cancer. 2023;11(8):e006666. doi: 10.1136/jitc-2023-006666.
28. Zhang Z, Sun D, Tang H, et al. PER2 binding to HSP90 enhances immune response against oral squamous cell carcinoma by inhibiting IKK/NF-kappaB signaling and PD-L1 expression. J Immunother Cancer. 2023;11(11):e007627. doi: 10.1136/jitc-2023-007627.
29. Pratap Singh R, Kumari N, Gupta S, et al. Intratumoral microbiota changes with tumor stage and influences the immune signature of oral squamous cell carcinoma. Microbiol Spectr. 2023;11(4):e04596-22. doi: 10.1128/spectrum.04596-22.
30. Sun J, Tang Q, Yu S, et al. Fusobacterium nucleatum facilitates oral squamous cell carcinoma progression via GLUT1-driven lactate production. EBioMedicine. 2023;88:104444. doi: 10.1016/j.ebiom.2023.104444.
31. Muñoz-Grez CP, Vidal MA, Rojas TB, et al. Host-microbe computational proteomic landscape in oral cancer revealed key functional and metabolic pathways between Fusobacterium nucleatum and cancer progression. Int J Oral Sci. 2025;17(1):1. doi: 10.1038/s41368-024-00326-8.
32. Arora R, Haynes L, Kumar M, et al. NCBP2 and TFRC are novel prognostic biomarkers in oral squamous cell carcinoma. Cancer Gene Ther. 2023;30(5):752-765. doi: 10.1038/s41417-022-00578-8.
33. Wang Z, Li H, Cai H, et al. FTO sensitizes oral squamous cell carcinoma to ferroptosis by suppressing ACSL3 and GPX4. Int J Mol Sci. 2023;24(22):16339. doi: 10.3390/ijms242216339.
34. Zhu W, Wang J, Liu X, et al. lncRNA CYTOR promotes aberrant glycolysis and mitochondrial respiration via HNRNPC-mediated ZEB1 stabilization in oral squamous cell carcinoma. Cell Death Dis. 2022;13(8):703. doi: 10.1038/s41419-022-05157-1.
35. Cui Y, Liu J, Liu L, et al. m6A-modified circFOXK2 targets GLUT1 to accelerate oral squamous cell carcinoma aerobic glycolysis. Cancer Gene Ther. 2023;30(1):163-171. doi: 10.1038/s41417-022-00526-6.
36. Yang Y, Tang H, Zheng J, et al. The PER1/HIF-1alpha negative feedback loop promotes ferroptosis and inhibits tumor progression in oral squamous cell carcinoma. Transl Oncol. 2022;18:101360. doi: 10.1016/j.tranon.2022.101360.
37. Fan T, Wang X, Zhang S, et al. NUPR1 promotes proliferation and metastasis of oral squamous cell carcinoma cells by activating TFE3-dependent autophagy. Signal Transduct Target Ther. 2022;7(1):130. doi: 10.1038/s41392-022-00939-7.
38. Ju WT, Xia RH, Zhu DW, et al. A pilot study of neoadjuvant anti-PD-1 camrelizumab plus the VEGFR2 inhibitor apatinib for locally advanced resectable oral squamous cell carcinoma. Nat Commun. 2022;13(1):5378. doi: 10.1038/s41467-022-33080-8.
39. Huang Y, Sun J, Li J, et al. Neoadjuvant immunochemotherapy for locally advanced resectable oral squamous cell carcinoma: a prospective single-arm trial. Int J Surg. 2023;109(8):2220-2227. doi: 10.1097/js9.0000000000000489.
40. Xiang Z, Wei X, Zhang Z, et al. Efficacy, safety and single-cell analysis of neoadjuvant immunochemotherapy in locally advanced oral squamous cell carcinoma: a phase II trial. Nat Commun. 2025;16(1):3968. doi: 10.1038/s41467-025-59004-w.
41. Liu YT, Liu HM, Ren JG, et al. Immune-featured stromal niches associate with response to neoadjuvant immunotherapy in oral squamous cell carcinoma. Cell Rep Med. 2025;6(3):102024. doi: 10.1016/j.xcrm.2025.102024.
42. de Wit JG, Vonk J, Voskuil FJ, et al. EGFR-targeted fluorescence molecular imaging for intraoperative margin assessment in oral cancer patients: a phase II trial. Nat Commun. 2023;14(1):4952. doi: 10.1038/s41467-023-40324-8.
43. Carnielli CM, Melo de Lima Morais T, Malta de Sá Patroni F, et al. Comprehensive glycoprofiling of oral tumors associates N-glycosylation with lymph-node metastasis and patient survival. Mol Cell Proteomics. 2023;22(7):100586. doi: 10.1016/j.mcpro.2023.100586.