



OPEN ACCESS

EDITED BY

Shama Prasada Kabekkodu,
Manipal Academy of Higher Education, India

REVIEWED BY

Ranajit Das,
Yenepoya University, India
Tianxiao Wang,
Nanjing Medical University, China

*CORRESPONDENCE

Weidong Tian
✉ drtwd@sina.com
Bairong Shen
✉ bairong.shen@scu.edu.cn

[†]These authors have contributed equally to this work

RECEIVED 26 August 2025

ACCEPTED 14 October 2025

PUBLISHED 31 October 2025

CITATION

Zhou Y, Wu Y, Shi W, Zhang Y, Liu X, Zhang Y, Zhan C, Chen B, Tian W and Shen B (2025) dbGVOSCC: a comprehensive database of genetic variation for systems genetics research on oral squamous cell carcinoma. *Front. Oncol.* 15:1692732. doi: 10.3389/fonc.2025.1692732

COPYRIGHT

© 2025 Zhou, Wu, Shi, Zhang, Liu, Zhang, Zhan, Chen, Tian and Shen. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

dbGVOSCC: a comprehensive database of genetic variation for systems genetics research on oral squamous cell carcinoma

Yi Zhou^{1†}, Yutao Wu^{1,2†}, Wenjing Shi^{1,3}, Yingbo Zhang¹, Xingyun Liu¹, Yuxin Zhang¹, Chaoying Zhan¹, Bingyue Chen¹, Weidong Tian^{2*} and Bairong Shen^{1*}

¹Department of Breast Surgery and Institute for Systems Genetics, Frontiers Science Center for Disease-Related Molecular Network, West China Hospital, Sichuan University, Chengdu, China, ²State Key Laboratory of Oral Diseases & National Clinical Research Center for Oral Diseases & Engineering Research Center of Oral Translational Medicine, Ministry of Education & National Engineering Laboratory for Oral Regenerative Medicine, West China Hospital of Stomatology, Sichuan University, Chengdu, China, ³01life Institute, Shenzhen, China

Introduction: Oral squamous cell carcinoma (OSCC) is a highly aggressive malignancy of the oral epithelium, marked by a high rate of lymph node metastasis and a profound negative impact on patients' quality of life. Despite its severity, no routine screening program currently exists for OSCC. To address the genetic heterogeneity underlying OSCC, we have developed a database of genetic variation in oral squamous cell carcinoma (dbGVOSCC; <http://www.sysbio.org.cn/dbGVOSCC/>).

Methods: OSCC literature (1991–2024) was queried from PubMed and screened manually and via PubTator, following predefined inclusion/exclusion criteria. Entities and relations were extracted from qualifying articles and organized into tables. The database adopted a browser/server architecture using HTML and XAMPP. Front-end was built with HTML and CSS for web display; server-side used Apache for infrastructure, MySQL for data management, and PHP/JavaScript for backend-frontend integration. Bioinformatics included mapping genes to STRING (confidence >0.9), hub gene identification via PPI degree centrality, and GO/KEGG enrichment with clusterProfiler (FDR-corrected). Usability was assessed using SUS and NPS surveys.

Results: dbGVOSCC comprises 1,788 somatic genetic variation entries from 400 original studies and 106,079 clinical samples, covering epimutations/methylations (329), SNPs (411), point mutations excluding SNP (258), indels (98), CNVs (348), LOH (28), one locus mutation, plus 333 unspecified mutations. We curated 817 biomarker-linked variations (diagnostic n=71, therapeutic n=175, prognostic n=291; 277 multi-application). PPI analysis highlighted 15 key genes (e.g., TP53, CTNNB1, AKT1, EGFR, PIK3CA). Enrichment implicated proliferation, adhesion/migration, p53/DNA damage response, and PI3K–Akt signaling. User testing showed SUS 88.75 (grade A) and NPS 90.

Discussion: dbGVOSCC represents a robust and reliable knowledge base, offering clinicians and researchers an open-source platform for personalized genotype-phenotype association studies and systems genetics research into the mechanisms of OSCC.

KEYWORDS

oral squamous cell carcinoma, dbGVOSCC database, cancer heterogeneity, genetic variations, precision oncology

1 Introduction

OSCC originates from the squamous epithelial cells of the oral cavity and arises in various sites, including the gums, hard palate, tongue, buccal mucosa, lips, and adjoining structures. It represents the most common form of oral cancer, accounting for over 90% of cases globally (1). In 2022, oral cavity cancers—including OSCC—accounted for approximately 188,438 deaths worldwide and an estimated total of 380,000 new cases, as reported by GLOBOCAN (2, 3). With over 650,000 new head and neck cancer cases each year, OSCC is one of the most aggressive subtypes within this category (4).

Although the incidence and mortality of OSCC have gradually declined since the 1970s, the absence of standardized screening programs and reliable blood-based diagnostics remains a major obstacle to early detection. Consequently, cervical lymph node metastases are present in approximately 40% of patients at diagnosis (5). Moreover, the prognosis is markedly poorer for these patients, with a 5-year survival rate of only 25–40%, compared to about 90% in non-metastatic cases (6). While cetuximab, an epidermal growth factor receptor (EGFR)-targeting monoclonal antibody, has been approved for the treatment of head and neck squamous cell carcinoma since 2006 (7, 8). However, its therapeutic efficacy is influenced by the genetic profile of individual patients. Genetic mutations, particularly in driver genes such as TP53, PIK3CA, and KRAS, are critical to OSCC pathogenesis and progression (9). For example, PAIP1 promotes OSCC metastasis (10), while polymorphisms in genes related to vascular endothelial growth factor (VEGF), hypoxia-inducible factor-1 α (HIF-1 α), and T-cell regulation are associated with clinical outcomes (11–13). Additionally, epigenetic modifications, including amplified factors such as RUVBL1 and protein tyrosine kinase 6 (PTK6), play pivotal roles in driving cancer proliferation and influencing immunotherapy responses (14, 15).

Most OSCC patients have a history of long-term smoking or alcohol consumption, whereas cases among non-smokers and non-drinkers are relatively uncommon (16). Studies suggest that certain substances, such as *Catha edulis*, *Cannabis sativa*, and *Areca catechu*, exert stronger carcinogenic effects than traditional tobacco, particularly in regions such as Pakistan, India, Afghanistan, and Iran (17). Although alcohol itself is not directly carcinogenic, it acts as a solvent that enhances the penetration of

carcinogens into the oral mucosa (18). Individuals who both smoke and consume alcohol heavily have been reported to face up to a 30-fold greater risk of developing OSCC compared with abstainers. Notably, genetic mutations—including FAT1, CASP8, CDKN2A, and NOTCH1—further increase susceptibility by disrupting carcinogen metabolism (9). These mutations are more frequently observed in OSCC than in other head and neck cancers, with their prevalence varying across populations exposed to different risk factors.

Complex diseases like OSCC often arise from the cumulative effects of multiple genetic variants, which can only be effectively studied through systems genetics approaches (19, 20). The emergence of the fourth scientific discovery paradigm, driven by precision medicine and data integration, underscores the need for comprehensive genomics and clinical omics data (21). A comprehensive knowledge database is essential for enabling system-level analysis of these variants and understanding the molecular mechanisms underlying OSCC. Existing databases, such as the oral cancer-related gene database OCDB v.2 (22), the Copenhagen OSCC database (23), and dbGENVOC (24), have notable limitations. For example, OCDB v.2 includes only 374 genes without verification from original studies, and its data have not been updated since 2011. The Copenhagen OSCC database lacks mutation data and was not published due to privacy concerns. While dbGENVOC provides genomic data for 325 Indian oral cancer patients and 118 patients from published literature, it lacks detailed clinical information. The Cancer Genome Atlas (TCGA) offers robust genomic data for 172 to 361 cases of lip, oral cavity, and pharyngeal cancers, but its samples are primarily from the US and Canada, limiting its global applicability (25).

To address these gaps, we developed dbGVOSCC, a comprehensive database integrating fragmented OSCC-related genetic variants with corresponding data extracted from the publications. Building upon the initial version of dbGVOSCC, we have conducted bioinformatics analyses, including protein-protein interaction network studies, to investigate these variants at a systems biology level. Biomarker information has also been curated and classified to examine key genes and their roles in OSCC. The database includes manually curated clinical and topographical data from original articles, providing users with real-world clinical insights.

2 Materials and methods

2.1 Data sources

All data for the OSCC Genetic Variation Database were curated from the public PubMed database (www.ncbi.nlm.nih.gov/PubMed). The search was conducted using the following query: (oral squamous cell carcinoma[ti] OR OSCC[ti]) NOT review[ptyp] AND English[LA] AND (1991/01/01[DP]: 2024/12/31[DP]).

This search retrieved 8,657 relevant articles published between January 1, 1991, and December 31, 2024. Abstracts of these articles were reviewed, and the literature was manually selected based on predefined inclusion and exclusion criteria.

Inclusion Criteria:

- Epidemiological studies on genetic variations in OSCC published in PubMed.
- Studies involving patients clinically diagnosed with OSCC and their corresponding control samples.

Exclusion Criteria:

- Articles not focused on genetic variation in OSCC.
- Studies with incomplete or unavailable data.

To facilitate data annotation, PubTator, an automated biomedical annotation tool, was employed to extract concepts such as genes (26) and mutations from the full text (27, 28). Articles mentioning a specific gene or mutation less than once were excluded from further analysis. Subsequently, the full texts of the remaining articles were manually reviewed, and studies with incomplete data or insufficient information were excluded.

Following this rigorous screening process, a total of 400 articles were selected as the original data source for dbGVOSCC. These articles form the foundation of the database, ensuring the inclusion of

high-quality, reliable information for systems genetics studies of OSCC.

2.2 Database implementation

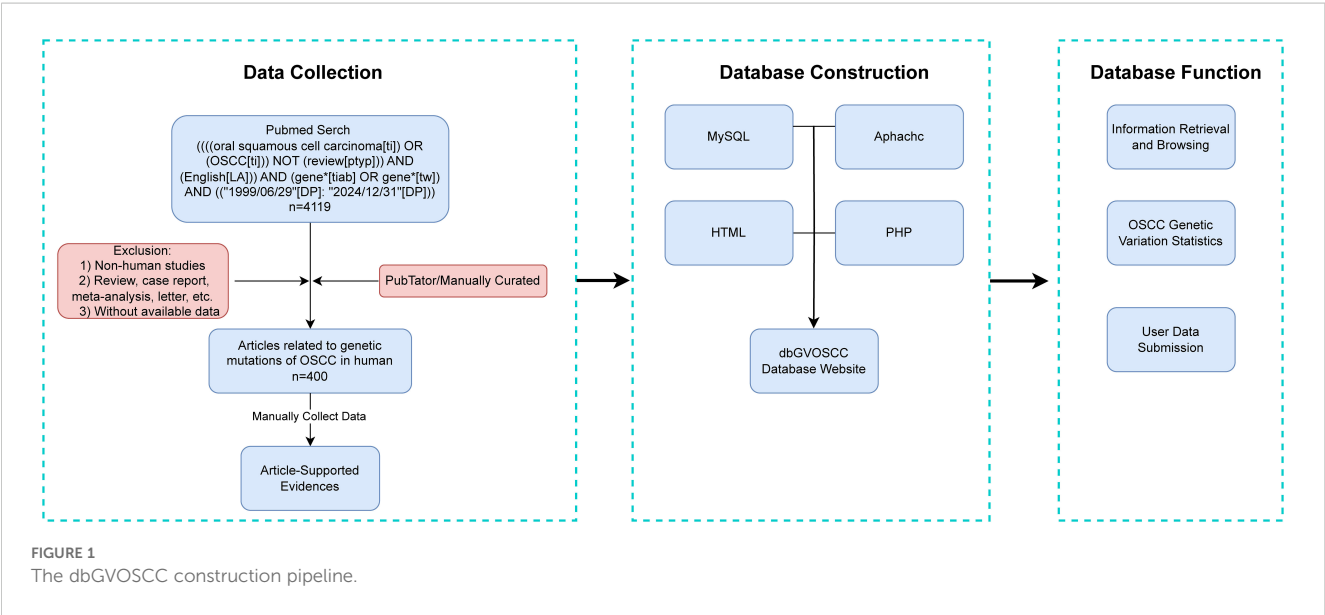
After screening and organizing the research literature, we systematically collected and structured the relevant information. The workflow for dbGVOSCC, including data collection, construction, and functionality, is illustrated in Figure 1.

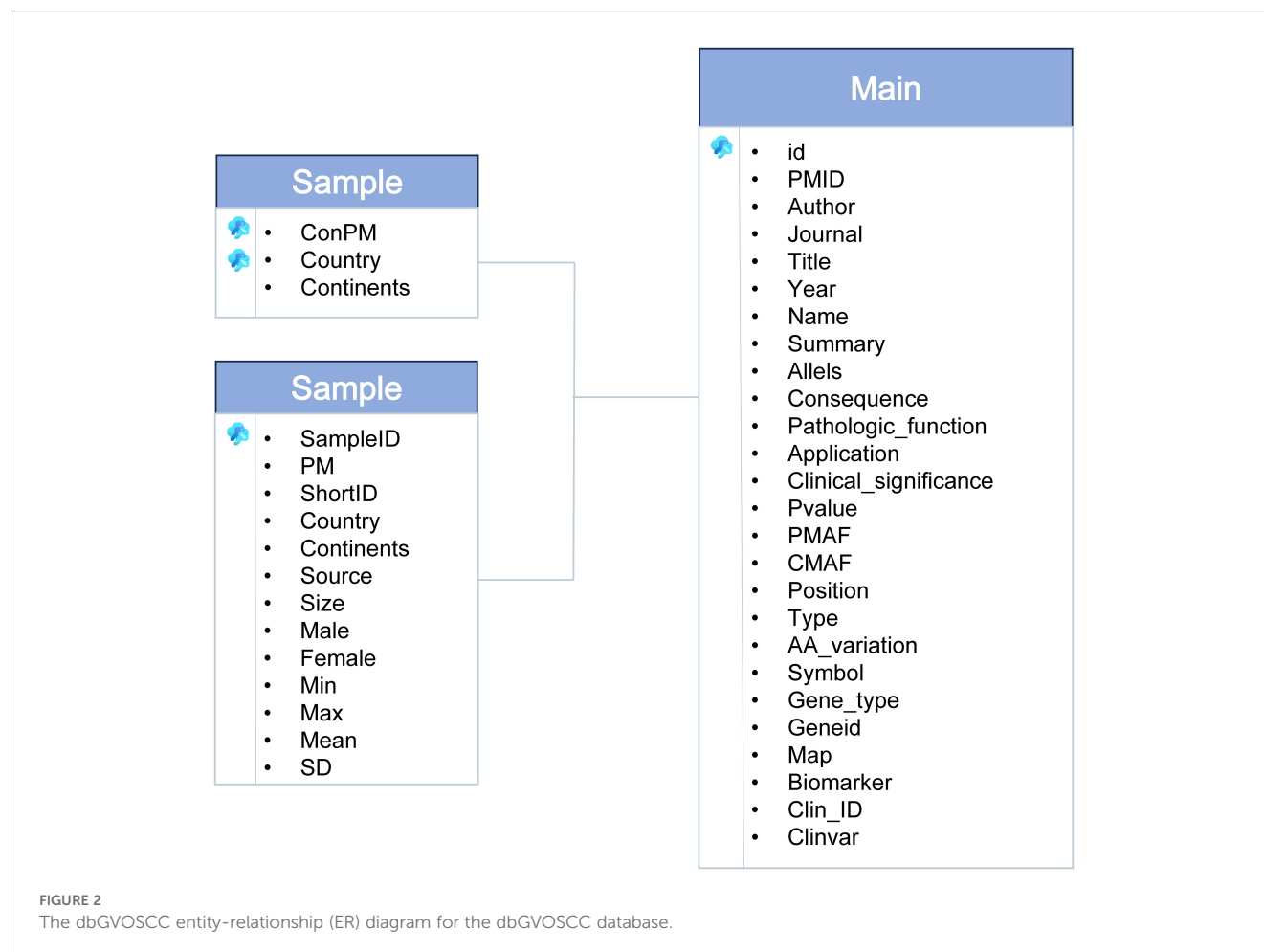
dbGVOSCC adopts a browser/server (B/S) architecture, enabling users to directly access data through a web browser. The front-end was developed using HTML and JavaScript, while the back-end utilizes an Apache server connected to a MySQL database via PHP. Users can explore data through categorized browsing or perform targeted searches using the online interface. A statistics page summarizes all included data, while the submission page allows users to upload new data, which is reviewed and approved by administrators before being added to the database.

2.3 Database structure

The dbGVOSCC database is organized into three entity tables and one relational table, as shown in Figure 2. These tables encompass all data within the database:

1. Sample Table: Contains information on clinical samples, including sample ID, age, country, subject size, sex, and sample source.
2. Reference Table: Includes details of the literature, such as PMID, year, article title, authors, and journal name.
3. Mutation Table: Records mutation-related data, including main ID, pathological function, clinical significance, clinical application, patient mutation frequency, control variant frequency, and a summary.





The relational table integrates these entity tables, facilitating seamless data retrieval and analysis.

2.4 Bioinformatics analysis

Proteins, as the functional products of genes, play pivotal roles in metabolic and signaling pathways. To explore the associations between genetic variants and disease phenotypes, we mapped the genes from dbGVOSCC to the STRING database with a confidence score > 0.9. Using the mapped data, a protein-protein interaction (PPI) network was constructed, and node degree centrality was calculated with Cytoscape. Nodes were ranked in descending order of centrality, and the top 5% were identified as key genes.

Subsequently, functional enrichment analyses were performed using the clusterProfiler package in R (v4.2.0), with the curated OSCC-mutated genes identified in this study as the input set. The background universe was defined as all annotated human genes available in Gene Ontology (GO) enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG). Enrichment significance was assessed using a hypergeometric test with Benjamini-Hochberg false discovery rate (FDR) correction. GO enrichment covered three ontologies (biological process, molecular function, and cellular component), while pathways

were annotated based on KEGG. Terms with FDR < 0.05 and fold enrichment > 2 were considered significantly enriched. For visualization, results were summarized by gene ratio (number of input genes in the term/total number of input genes), fold enrichment, point/bar size (hit gene count), and color scale ($-\log^{10}[\text{FDR}]$), and enriched terms were ranked first by FDR and then by gene ratio.

2.5 Validation of the efficacy and utility of dbGVOSCC

First, we performed a comparative analysis between dbGVOSCC and ChatGPT-5 to assess the effectiveness of dbGVOSCC in retrieving relevant information from scientific literature. Specifically, we submitted genetic variation data from 10 oral squamous cell carcinoma (OSCC) patients to both dbGVOSCC and ChatGPT-5. The returned results were then evaluated based on the reported relationships between genetic variations and OSCC, along with the corresponding source publications, following the evaluation criteria outlined in [Supplementary Table S1](#). The comparison with ChatGPT-5 was conducted using the official GPT-5 model via <https://chat.openai.com> in August 2024, ensuring consistency in system version and output.

Second, we conducted System Usability Scale (SUS) and Net Promoter Score (NPS) surveys to assess the usability and user loyalty of dbGVOSCC. We collected 21 valid responses from 10 clinicians, 8 researchers, and 3 public users.

Last, we compared dbGVOSCC with the OSCC-related databases to highlight its distinct features and enhancements.

3 Results

3.1 Data statistics

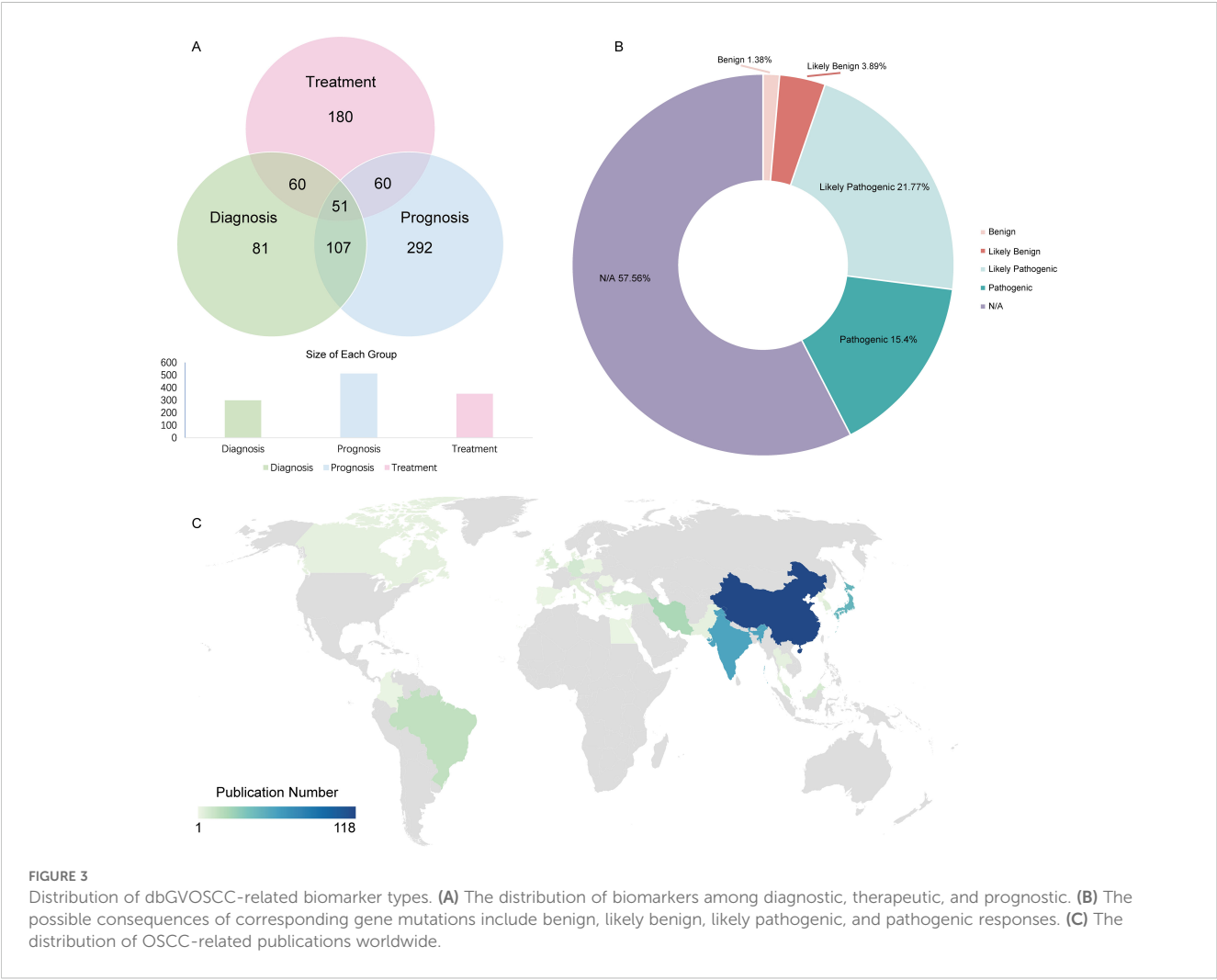
Following the screening and sorting procedures described above, we identified a total of 1,788 genetic variants associated with OSCC, including 329 epimutations/methylations, 411 single nucleotide polymorphisms (SNPs), 258 point-mutations excludes SNP, 98 insertion-deletion mutations (indels), 348 copy number variations (CNVs), 28 instances of loss of heterozygosity (LOH), 1 locus mutation, and 333 unspecified gene mutations. Classified by the types of gene function mutations, including 514 protein-coding DNA genes, 21 miRNAs, and 1 pre-mRNA. The corresponding clinical data encompassed a total of 106,079 clinical samples, of which 59,440

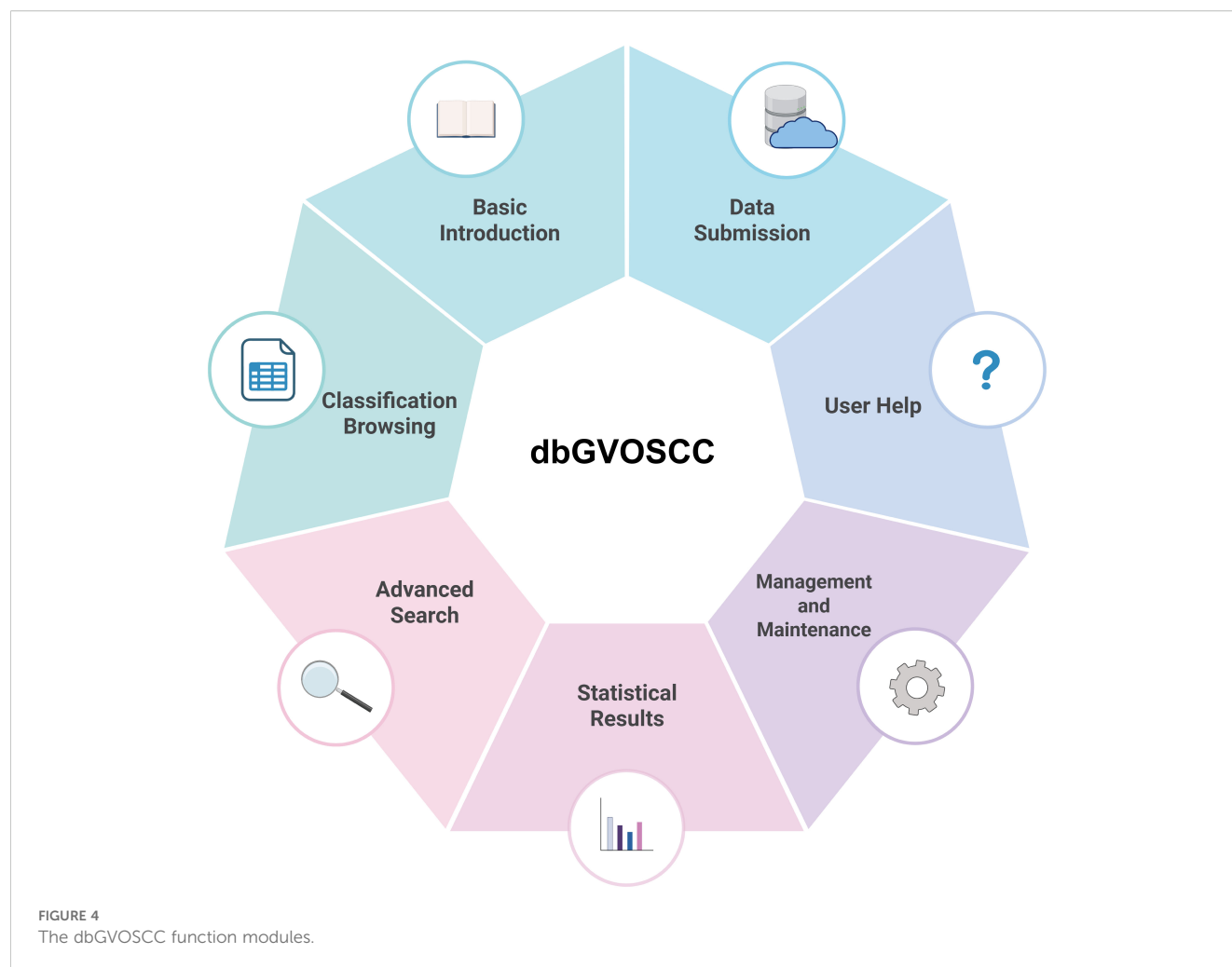
were distinctly identified as male and 20,674 as female (Table 1). Clinical samples are the total sample size aggregated across all studies included in dbGVOSCC. Topographic data are the subset of those samples with the primary OSCC site explicitly annotated in the

TABLE 1 Basic data statistics of the database.

Genetic variant type	Count
CNV	348
LOH	28
Locus mutation	1
Methylation	329
SNP	411
Point mutation excludes SNP	258
Indel mutation	98
Unspecified gene mutation	333

CNV, Copy number variation; LOH, Loss of heterozygosity; SNP, Single nucleotide polymorphism. Locus mutation: Genetic variation identified at a specific genomic locus or region, where the gene responsible for the mutation cannot be confidently determined. Unspecified gene mutation: The mutation is described in relation to OSCC, but the specific gene and/or the exact genetic mutation involved is not clearly identified.





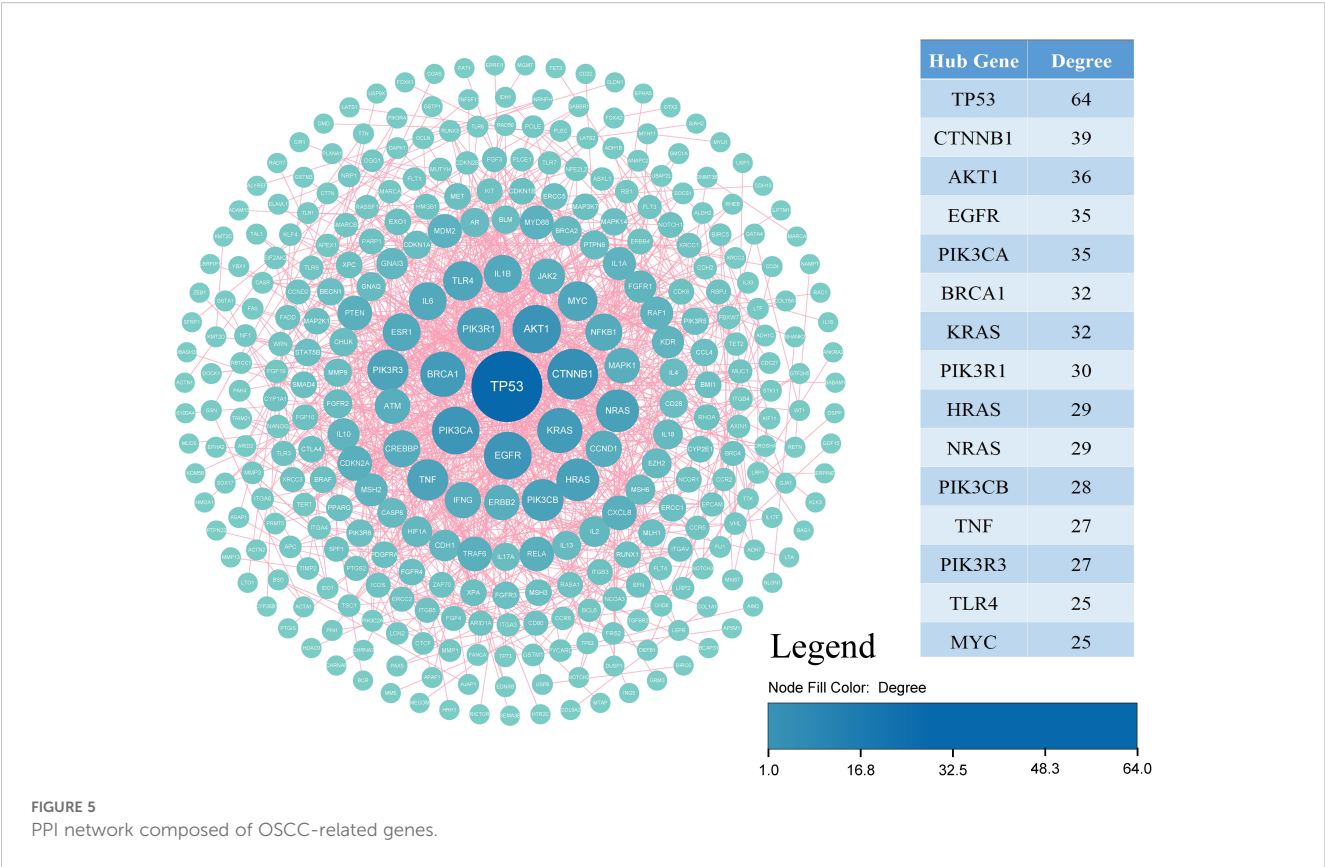
source studies, including 13, 083 samples mainly distributed in 7 different oral topographic sites (Supplementary Figure S1).

All 817 genetic variations were identified as potential biomarkers, with 201 explicitly labeled as such. These genetic variations were categorized into three primary application types –diagnostic, therapeutic, and prognostic –based on their roles in standard biological processes, pathological mechanisms, or pharmacological responses to therapeutic interventions (71, 175, and 291 variations, respectively) (Figure 3A). The distribution of variant clinical significance (benign, likely benign, likely pathogenic, pathogenic) is shown in Figure 3B. Among the 817 genetic variations, 277 were classified as multi-application biomarkers, meaning they were associated with two or more application types, accounting for 33.9% of the total. Based on data from 400 original research articles (Figure 3C), China has published the largest number of OSCC-related studies.

3.2 Website

Based on the structure and application requirements of dbGVOSCC, we developed seven functional modules for online use, as illustrated in Figure 4:

1. Home page: This page provides a comprehensive overview of dbGVOSCC, including references, related information, and resources. It also includes links to associated databases and research institution websites.
2. Browse page: This page allows users to explore data through three classification methods: variation type, clinical sample source, and mutation-related gene classification.
3. Search page: This page enables users to search for specific information using mutation name, data source (location), or gene name as search criteria.
4. Statistics page: This page displays key insights, such as the distribution of mutation types, the top ten genes ranked by frequency, the chromosomal distribution of mutations, and the geographic distribution of clinical sample sources.
5. Submit page: Users can contribute new data by providing four required items: mutation name, reference, contact email, and details. Mutation name and email are mandatory fields. Submitted data is reviewed, and upon approval, it is promptly added to dbGVOSCC.
6. Help page: Offers detailed guidance on navigating and using dbGVOSCC, ensuring that users can efficiently locate the desired information.



3.3 Bioinformatics analysis result

Using STRING, identifying 283 nodes and 954 interactions. Through degree centrality analysis, we pinpointed 15 key genes—TP53, CTNNB1, AKT1, EGFR, PIK3CA, BRCA1, KRAS, PIK3R1, HRAS, NRAS, PIK3CB, TNF, PIK3R3, TLR4, and MYC—as shown in Figure 5 and Table 2.

All included genes underwent Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses. The top ten results in each of the four GO categories were ranked in descending order based on the number of associated genes, as detailed in Supplementary Figure S2.

3.3.1 Gene Ontology biological process enrichment analysis

The Gene Ontology (GO) biological process enrichment analysis revealed significant involvement of genes in several key biological processes associated with oral squamous cell carcinoma (OSCC). As illustrated in Supplementary Figure S2A, epithelial cell proliferation, regulation of epithelial cell proliferation, muscle cell proliferation, regulation of smooth muscle cell proliferation, smooth muscle cell proliferation, gland development, and positive regulation of kinase activity indicate active growth and renewal of epithelial cells, often linked to tissue development or tumor progression. Other processes, ameboidal-type cell migration, tissue migration, and regulation of leukocyte differentiation, indicate dynamic cell motility processes, important for development, immune responses, and metastasis. These biological

processes emphasize cell proliferation, tissue/organ development, migration, and immune regulation, with a strong focus on epithelial and muscle cells. The enrichment highlights pathways that underlie growth, differentiation, motility, and kinase-driven signaling, processes that are central to both normal tissue remodeling and pathological conditions such as cancer progression.

3.3.2 Gene Ontology cellular components enrichment analysis

The cellular component enrichment analysis identified several key structures associated with oral squamous cell carcinoma (OSCC), as shown in Supplementary Figure S2B. cell-substrate junction, focal adhesion, protein complex, cell leading edge, and lamellipodium are involved in cell adhesion and migration, regulating adhesion, migration, and downstream signaling. Extrinsic component of membrane, extrinsic component of plasma membrane, and glutamatergic synapse are primarily linked to the cell surface, typically involved in signal reception and molecular recognition. In addition, transferase complex, transferring phosphorus-containing groups, and phosphatidylinositol-3-kinase (PI3K) complex represent phosphorylation-related enzymatic activity, highlighting enriched protein phosphorylation regulation. These enriched components mainly converge on cell adhesion/migration (linked to invasion and metastasis), membrane-associated signaling complexes, and kinase-driven signal transduction (such as PI3K). This suggests significant involvement in cell motility, receptor-mediated signaling, and phosphorylation regulation, processes highly relevant to tumor progression and abnormal signaling activation.

TABLE 2 Functions of key genes in OSCC.

Gene	Functions	PMID
TP53	DNA damage response, cell cycle arrest, apoptosis	29489434
CTNNB1	Wnt/ β -catenin signaling; cell adhesion and transcription regulation	33531646
AKT1	Regulation of cell proliferation and growth, and participation in apoptosis and glucose metabolism	27889930
EGFR	Cell proliferation, angiogenesis, tumor invasion, metastasis, and apoptosis	33680380
PIK3CA	Tumor growth	31516747
BRCA1	DNA repairment	29896272
KRAS	Regulation of cell growth and activation	20813562
PIK3R1	Cell adhesion and development	32115621
HRAS	Transduce signals that instruct cells to grow and divide	31845386
NRAS	Transduction of signals that instruct cell growth and division	33680380
PIK3CB	Cell growth, proliferation, differentiation, and survival	32115621
TNF	Promotion of inflammation	37505766
PIK3R3	Promotion of cell proliferation, immortalization, dedifferentiation, and transformation	32115621
TLR4	NF- κ B/MAPK activation, immune escape, chemoresistance	38593176
MYC	Transcription factor; proliferation, apoptosis, tumor progression	35890188

TP53, Tumor Protein P53; CTNNB1, Catenin Beta 1; AKT1, AKT Serine/Threonine Kinase 1; EGFR, Epidermal Growth Factor Receptor; PIK3CA, Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha; BRCA1, Breast Cancer 1, DNA Repair Associated; KRAS, KRAS Proto-Oncogene, GTPase; PIK3R1, Phosphoinositide-3-Kinase Regulatory Subunit 1; HRAS, HRAS Proto-Oncogene, GTPase; NRAS, NRAS Proto-Oncogene, GTPase; PIK3CB, Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Beta; TNF, Tumor Necrosis Factor; PIK3R3, Phosphoinositide-3-Kinase Regulatory Subunit 3; TLR4, Toll-Like Receptor 4; MYC: MYC Proto-Oncogene, BHLH Transcription Factor.

3.3.3 Gene Ontology molecular function enrichment analysis

The molecular function enrichment analysis identified significant roles of key molecular activities in oral squamous cell carcinoma (OSCC), as shown in [Supplementary Figure S2C](#). The most prominently enriched functions included DNA-binding transcription factor binding, RNA polymerase II-specific DNA-binding transcription factor binding, catalytic activity, acting on DNA, damaged DNA binding, and p53 binding, indicating roles in DNA metabolism, repair, and stress-response pathways, particularly involving the tumor suppressor p53. Cytokine receptor binding, growth factor binding, protein tyrosine kinase activity, transmembrane receptor protein kinase activity, and transmembrane receptor protein tyrosine kinase activity highlight communication between cells via cytokines and growth factors, critical in immune responses, proliferation, migration and differentiation. The enriched functions center on DNA interaction and repair (including p53-mediated processes) and kinase-mediated signal transduction (via growth factor and cytokine

receptors). Together, these functions suggest strong involvement in gene regulation, DNA damage response, and growth/survival signaling, processes that are pivotal in both normal physiology and cancer progression. These results provide a detailed understanding of the molecular mechanisms underlying OSCC, highlighting potential targets for therapeutic intervention.

3.3.4 KEGG pathway enrichment analysis

The KEGG pathway enrichment analysis highlighted several critical pathways associated with oral squamous cell carcinoma (OSCC), as shown in [Supplementary Figure S2D](#). The most significantly enriched pathways included PI3K-Akt signaling pathway, MicroRNAs in cancer, and proteoglycans in cancer, which are well-known for their roles in tumorigenesis, cellular proliferation, and survival. Other notable enriched pathways involved specific cancer types, including breast cancer, prostate cancer, bladder cancer, and melanoma, reflecting the shared molecular mechanisms between OSCC and these malignancies. Chronic myeloid leukemia, platinum drug resistance, and Central carbon metabolism in cancer pathways were also enriched, indicating potential links to treatment resistance and therapeutic targeting. These findings provide valuable insights into the molecular mechanisms driving OSCC and highlight key pathways for future therapeutic interventions.

3.4 Validation of the efficacy and utility of dbGVOSCC

We compared the effectiveness of dbGVOSCC and ChatGPT-5 in retrieving relevant information from patient data within scientific and clinical research contexts ([Supplementary Material S1](#)). Both systems performed well in matching queries with relevant responses (9.63 vs. 10). However, compared with ChatGPT-5, dbGVOSCC demonstrated superior performance in several key aspects: providing detailed information from research articles (0.41 vs. 9.6), recommending publications with accessible links (0.26 vs. 1.3), achieving a higher authenticity ratio of scientific articles (48.83 vs. 100), identifying original research articles (0.93 vs. 1.3), and delivering standardized output (0 vs. 100) ([Table 3](#), [Supplementary Table S2](#)).

dbGVOSCC achieved a score of 88.75 on the System Usability Scale (SUS), with adjective ratings classified as “Excellent”, a grade scale of “A”, and an acceptability range of “Acceptable”. These results indicate that dbGVOSCC demonstrates strong usability for users. Similarly, in the Net Promoter Score (NPS) assessment, dbGVOSCC scored 90, with 90% of participants identified as promoters, 10% as passive, and none as detractors. This further underscores the high level of user satisfaction with dbGVOSCC ([Figure 6](#)).

We also compared the dbGVOSCC with other databases relevant to oral OSCC, focusing on sample size, population diversity, disease specificity, and clinical annotations. As shown in [Table 4](#), these comparisons underscore the distinctive scope and utility of dbGVOSCC as the most comprehensive, OSCC-specific genomic resource currently available.

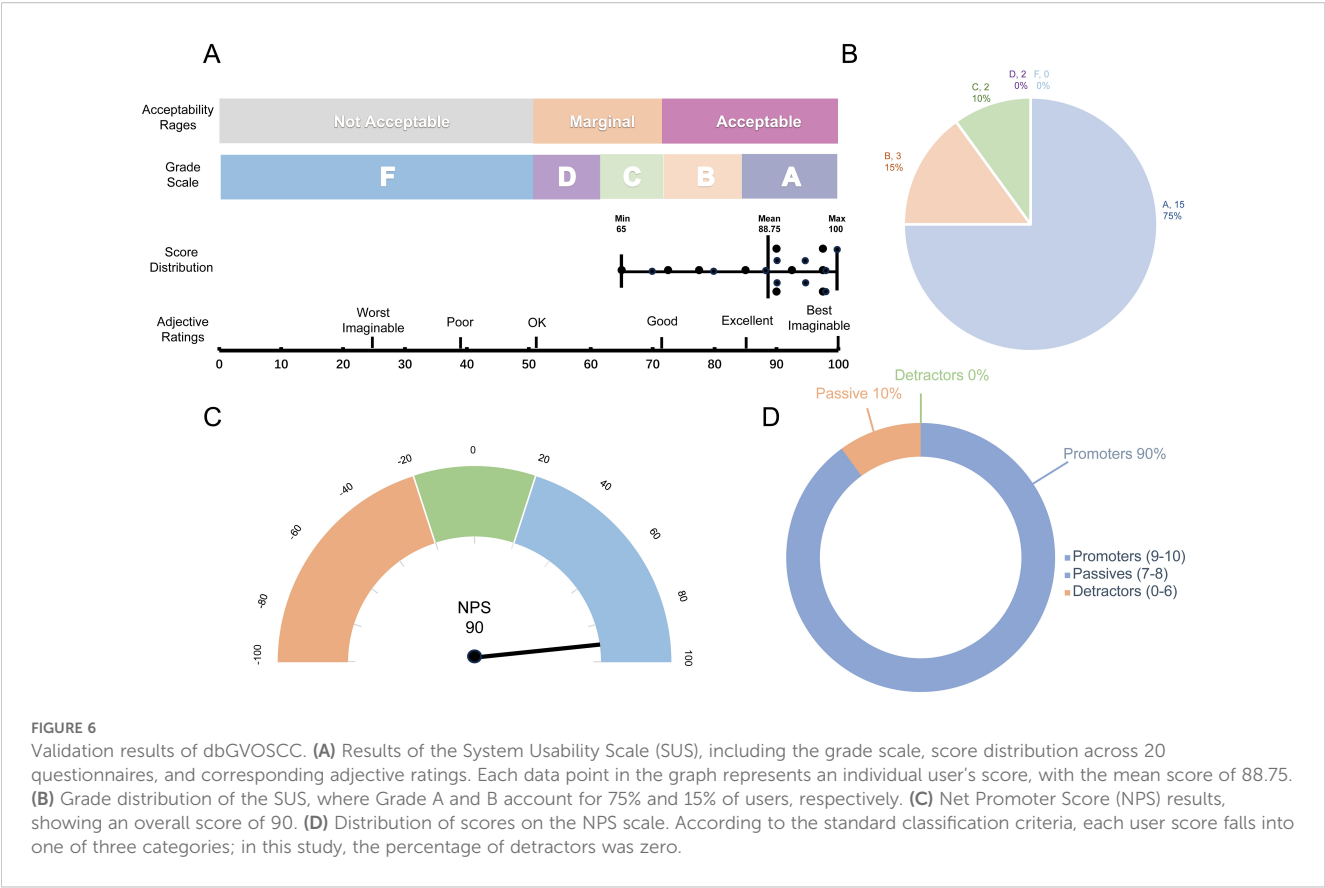
TABLE 3 Comparison between ChatGPT-5 and dbGVOSCC on question response.

Question ID	Matching between questions and responses	Present the details of the research articles	Number of recommended publications	Number of recommended publications with links	Scientific article authenticity ratio (%)	Number of original research	Standardized output results
ChatGPT-5							
1	9.333	1	3.333	0.667	70	2	N
2	10	0	2.667	0.667	50	1	N
3	9.333	0.667	2	0	22.222	0.667	N
4	9.333	0	1	0	22.222	0.667	N
5	9.333	0.667	2	0	72.222	1.333	N
6	10	0.333	3	1	88.889	0.667	N
7	10	0	2.333	0	58.333	1.333	N
8	10	0.333	0.333	0	33.333	0	N
9	9.333	0.667	1	0	22.222	0.667	N
10	10	0	3	0	100	3	N
Mean	9.63 ± 0.33	0.41 ± 0.35	1.96 ± 0.95	0.26 ± 0.37	48.83 ± 27.23	0.93 ± 0.81	0
dbGVOSCC							
1	10	10	1	1	100	1	Y
2	10	10	1	1	100	1	Y
3	10	10	2	2	100	2	Y
4	10	8	1	1	100	1	Y
5	10	10	1	1	100	1	Y
6	10	10	2	2	100	2	Y
7	10	10	1	1	100	1	Y
8	10	10	1	1	100	1	Y
9	10	8	1	1	100	1	Y
10	10	10	2	2	100	2	Y
Mean	10	9.6 ± 0.83	1.3 ± 0.47	1.3 ± 0.47	100	1.3 ± 0.47	100

4 Discussion

In this study, we established dbGVOSCC, a manually curated knowledge base that systematically integrates 1,788 entries of somatic genetic variations from 106,079 clinical samples of 400 peer-reviewed studies. By encompassing a wide spectrum of mutation types, including SNPs, point mutations, indels, CNVs, LOH, and epimutations, dbGVOSCC provides a more comprehensive and structured resource than existing OSCC-related databases. Notably, it demonstrates higher accuracy and more organized output results when extracting information from scientific literature compared with ChatGPT-5, while its high SUS (88.75) and NPS (90) further underscore its reliability and usability. Together, these attributes position dbGVOSCC as a robust platform for systems genetics research and a valuable tool for advancing precision oncology in OSCC.

The dbGVOSCC platform was developed to address the pressing need for accessible, structured, and high-quality data in OSCC research. With its seven functional modules, including advanced search tools, classification browsing, and regular updates, dbGVOSCC not only facilitates the exploration of OSCC-related genetic variations but also bridges the gap between raw genetic data and actionable clinical insights. As a first step toward addressing OSCC heterogeneity, dbGVOSCC already extends beyond SNVs to include CNVs and LOH, and it provides clinical and topographic metadata (tumor subsite, age, sex, sample source, country), enabling stratified analyses relevant to heterogeneity. This user-friendly, open-access platform empowers researchers and clinicians alike, enhancing the translational application of genetic information in diagnostics, therapeutics, and prognostics.



Bioinformatics analyses provided comprehensive insights into the biological processes, cellular components, and molecular functions underlying OSCC pathogenesis. GO enrichment revealed that OSCC-related genes play key roles in tumor growth, immune modulation, and signaling pathways, particularly cell proliferation, leukocyte differentiation, and the MAPK cascade (29). Cellular component enrichment highlighted receptor and adhesion-related protein

complexes essential for signal transduction and tumor-microenvironment interactions. While molecular function analysis emphasized tyrosine kinase activity and p53 binding, pointing to potential therapeutic targets (30). KEGG pathway analysis further identified the PI3K-Akt pathway, general cancer pathways, and other malignancy-specific pathways (e.g., melanoma, bladder cancer, breast cancer) as central to OSCC progression (31). These shared mechanisms

TABLE 4 Comparison of OSCC-related databases.

Database	Year	Sample Size (OSCC-related)	Population diversity	OSCC-specific	Clinical annotation	Key notes
dbGVOSCC (This work)	2024	106,079	Multi-ethnic (Asia, Europe, Americas)	√	√ Full: age, sex, site, geography	Comprehensive, population-diverse, clinically rich OSCC database
dbGENVOC	2021	100 WES + 5 WGS (India) + 220 TCGA + 118 literature = 443	Indian + US (TCGA) + global publications	√ Partial	× Minimal	Mainly Indian WES/WGS + integrated TCGA/literature mutations
TCGA-HNSC	~2015	~315–391 OSCC subset (of 528 HNSC)	North American (mainly Caucasian)	×	√ Survival, pathology, treatment	Subset of HNSC cohort with OSCC tumors, strong clinical annotation
cBioPortal (HNSC)	Ongoing	~528 (matches TCGA-HNSC)	Same as TCGA	×	√	Interactive platform accessing TCGA and other HNSC data
ICGC (OSCC projects)	~2020	57 OSCC samples	South Asian and South American	× Partial	√ Limited	Includes data from Indian and Brazilian OSCC projects
OncoMX	2023	Not defined (meta-database)	Global (TCGA, ICGC, GTEx)	×	√ Aggregated clinical/biomarker data	Meta-resource integrating multi-platform cancer genomics data

√, Feature present/available for that database.

suggest opportunities for cross-cancer therapeutic strategies, with the PI3K/mTOR pathway already showing promise through inhibitors such as BKM120 and BYL719, which act as radiosensitizers in OSCC (32). Disease-gene and phenotype enrichment analyses underscored the broader impact of OSCC-related genetic variations, revealing overlap with thoracic, breast, and gastrointestinal cancers and associations with gonadal, ovarian, and skin neoplasms (33, 34). Tissue expression analysis further demonstrated links to cervical carcinoma cells, regulatory T lymphocytes, and leukemia cells (35), highlighting the dual roles of OSCC-related genes in tumor-specific and immune-related processes. Collectively, these findings emphasize the importance of targeting both tumor biology and the tumor microenvironment to enhance treatment efficacy and reduce resistance.

Despite the rapid advancement of large language models (LLMs) such as ChatGPT in biomedical knowledge extraction, their utility in clinical and translational research remains limited by fundamental challenges, including unverifiable data sources, generalized knowledge representations, and the hallucination phenomenon (36). LLMs are typically trained on heterogeneous, non-curated corpora, leading to compromised data traceability and an elevated risk of generating fabricated or inaccurate information. Just as in the comparison results of ChatGPT-5, our database exhibits superior accuracy and more structured output, further supported by high SUS and NPS. Such limitations are particularly problematic for applications in precision oncology, where the fidelity, granularity, and clinical validity of data are paramount.

In contrast, dbGVOSCC, our curated knowledge base focused on genetic variants in OSCC, was meticulously constructed through manual extraction from peer-reviewed scientific literature, ensuring the authenticity, accuracy, and traceability of every entry. Compared with existing OSCC-related databases such as dbGENVOC and OncoMX (37), dbGVOSCC incorporates a more comprehensive and continuously updated collection of data. Beyond basic variant annotation, dbGVOSCC systematically

captures critical clinical metadata, including anatomical tumor subsite, patient age, sample source, sex, and geographic origin, enabling stratified analyses of genetic variation patterns across diverse populations. Recognizing that genetic predisposition and somatic mutation spectra vary substantially across ethnicities, regions, and lifestyle backgrounds, dbGVOSCC offers a level of contextual specificity not achievable by generalized LLM outputs.

Moreover, its strictly curated and continuously updated architecture circumvents the issue of model collapse that affects LLMs trained on self-generated data, ensuring dbGVOSCC remains a reliable and clinically actionable resource over time. Collectively, these attributes position dbGVOSCC as a more specialized, precise, and trustworthy platform for advancing systems genetics research and facilitating the development of targeted therapies for oral squamous cell carcinoma. Furthermore, dbGVOSCC may serve as a foundation for training specialized machine learning models, leveraging curated, high-fidelity datasets to overcome the hallucination and data contamination challenges associated with general-purpose LLMs (38). By maintaining a rigorous curation standard while embracing computational advances, dbGVOSCC is poised to support future efforts in precision oncology, including biomarker discovery, therapeutic target identification, and personalized treatment strategy development for oral squamous cell carcinoma (39).

4.1 Future directions and challenges

While dbGVOSCC provides a robust platform for exploring OSCC-related genetic variations, some limitations remain. Tumor heterogeneity, a hallmark of malignancies, is not yet fully captured in the database. Intratumoral heterogeneity and genomic instability, which contribute to tumor progression and therapeutic resistance, are areas that require further data integration. Addressing these gaps will be a priority in future updates.

dbGVOSCC update roadmap

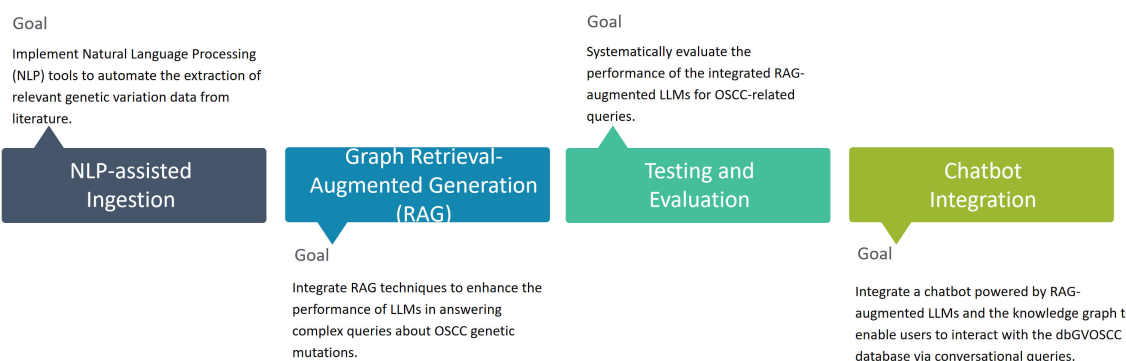


FIGURE 7

The dbGVOSCC update roadmap. This roadmap delineates the principal milestones and timelines for embedding NLP tools and knowledge graphs within the dbGVOSCC database. It encompasses key stages, including knowledge graph construction, NLP-facilitated data ingestion, seamless integration with RAG models to optimize query generation, and culminating in a chatbot deployment tailored for clinical and research support.

We plan to enhance the representation of OSCC complexities and heterogeneity in future updates of dbGVOSCC by: 1) incorporating finer-grained intertumoral architecture through NLP-assisted literature curation; 2) developing an OSCC-specific ontology to standardize heterogeneous features; 3) converting the relational database into a knowledge graph, allowing systematic capture and integration of genomic instability patterns (Figure 7). These advancements will enhance the platform's usability and applicability in translational medicine. Moreover, the integration of a knowledge graph with medical chatbots could facilitate more effective doctor-patient communication, improving clinical decision-making and personalized care.

In conclusion, this study highlights the complex molecular mechanisms underlying OSCC and demonstrates the utility of dbGVOSCC as a resource for advancing research and clinical applications. By integrating genomic, clinical, and functional data, dbGVOSCC not only provides a comprehensive tool for understanding OSCC pathogenesis but also sets the stage for developing novel diagnostic tools, therapeutic strategies, and prognostic models. Continued updates and expansions of the database will ensure its relevance and utility in the rapidly evolving field of precision oncology.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding authors.

Ethics statement

Because dbGVOSCC is based exclusively on publicly available, de-identified data, issues of personal data protection or re-consent do not arise. At the same time, we fully recognize the importance of responsible genomic data sharing. To this end, we (1) maintain traceable links to the original publications to ensure transparent data provenance; (2) adhere to the FAIR principles—findability, accessibility, interoperability, and reusability—to promote data reuse; and (3) plan to align future updates with community guidelines on ethical genomic data sharing and reuse.

Author contributions

YZ: Writing – original draft, Data curation, Software, Writing – review & editing, Methodology, Conceptualization. YW: Conceptualization, Writing – original draft, Data curation, Software, Methodology, Writing – review & editing. WS: Writing – original draft, Methodology, Software, Data curation. YBZ: Methodology, Writing – original draft, Visualization, Software. XL: Software, Writing – original draft, Visualization, Methodology. YXZ: Methodology, Writing – original draft, Software, Visualization. CZ: Software, Writing – original draft, Visualization.

BC: Writing – original draft, Software, Visualization. WT: Conceptualization, Writing – review & editing, Supervision. BS: Supervision, Writing – review & editing, Conceptualization.

Funding

The author(s) declare financial support was received for the research and/or publication of this article. This work was supported by the National Nature Science Foundation of China (No. 32270690) and the Sichuan Provincial Science and Technology Program (No. 2024YFHZ0205).

Conflict of interest

Author WS was employed by company 01life Institute.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2025.1692732/full#supplementary-material>

SUPPLEMENTARY FIGURE 1

Distribution of OSCC samples by primary oral topographic sites. (A) Illustrations of 7 different topographic locations in the oral cavity; (B) 13,083 samples mainly distributed in 7 different oral topographic sites.

SUPPLEMENTARY FIGURE 2

The dbGVOSCC database variant gene enrichment analysis. (A) Biological Process; (B) Cellular Component; (C) Molecular Function; (D) KEGG

References

- Ariyanon T, Klibngern H, Sittitai P, Ruenmarkkaew D, Watcharatsiriyuth W. Survival outcomes of surgically treated oral cavity squamous cell carcinoma patients at tertiary care hospital in Northern Thailand. *J Stomatol Oral Maxillofac Surg.* (2025) 126:102166. doi: 10.1016/j.jormas.2024.102166
- Wu J, Chen H, Liu Y, Yang R, An N. The global, regional, and national burden of oral cancer, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. *J Cancer Res Clin Oncol.* (2025) 151:53. doi: 10.1007/s00432-025-06098-w
- Barsouk A, Aluru JS, Rawla P, Saginala K, Barsouk A. Epidemiology, Risk Factors, and Prevention of Head and Neck Squamous Cell Carcinoma. *Med Sci (Basel).* (2023) 11:42. doi: 10.3390/medsci11020042
- Chow LQM. Head and Neck Cancer. *N Engl J Med.* (2020) 382:60–72. doi: 10.1056/NEJMra1715715
- Luo YH, Mei XL, Liu QR, Jiang B, Zhang S, Zhang K, et al. Diagnosing cervical lymph node metastasis in oral squamous cell carcinoma based on third-generation dual-source, dual-energy computed tomography. *Eur Radiol.* (2023) 33:162–71. doi: 10.1007/s00330-022-09033-6
- Shaikh S, Yadav DK, Bhadrasha K, Rawal RM. Integrated computational screening and liquid biopsy approach to uncover the role of biomarkers for oral cancer lymph node metastasis. *Sci Rep.* (2023) 13:14033. doi: 10.1038/s41598-023-41348-2
- Bonner JA, Harari PM, Giral J, Azarnia N, Shin DM, Cohen RB, et al. Radiotherapy plus cetuximab for squamous-cell carcinoma of the head and neck. *N Engl J Med.* (2006) 354:567–78. doi: 10.1056/NEJMoa053422
- Pickering CR, Zhang J, Yoo SY, Bengtsson L, Moorthy S, Neskey DM, et al. Integrative genomic characterization of oral squamous cell carcinoma identifies frequent somatic drivers. *Cancer Discov.* (2013) 3:770–81. doi: 10.1158/2159-8290.CD-12-0537
- Chai AWY, Lim KP, Cheong SC. Translational genomics and recent advances in oral squamous cell carcinoma. *Semin Cancer Biol.* (2020) 61:71–83. doi: 10.1016/j.semcancer.2019.09.011
- Swarup N, Hong KO, Chawla K, Choi SJ, Shin JA, Oh KY, et al. Effect of PAIP1 on the metastatic potential and prognostic significance in oral squamous cell carcinoma. *Int J Oral Sci.* (2022) 14:9. doi: 10.1038/s41368-022-00162-8
- Kammerer PW, Toyoshima T, Schoder F, Kammerer P, Kuhr K, Brieger J, et al. Association of T-cell regulatory gene polymorphisms with oral squamous cell carcinoma. *Oral Oncol.* (2010) 46:543–8. doi: 10.1016/j.oraloncology.2010.03.025
- Shieh TM, Chang KW, Tu HF, Shih YH, Ko SY, Chen YC, et al. Association between the polymorphisms in exon 12 of hypoxia-inducible factor-1 α and the clinicopathological features of oral squamous cell carcinoma. *Oral Oncol.* (2010) 46:e47–53. doi: 10.1016/j.oraloncology.2010.04.009
- Supic G, Jovic N, Zeljc K, Kozomara R, Magic Z. Association of VEGF-A genetic polymorphisms with cancer risk and survival in advanced-stage oral squamous cell carcinoma patients. *Oral Oncol.* (2012) 48:1171–7. doi: 10.1016/j.oraloncology.2012.05.023
- Lin D, Lin B, Bhanot H, Riou R, Abt NB, Rajagopal J, et al. RUVBL1 is an amplified epigenetic factor promoting proliferation and inhibiting differentiation program in head and neck squamous cancers. *Oral Oncol.* (2020) 111:104930. doi: 10.1016/j.oraloncology.2020.104930
- Zhou L, Xu N, Shibata H, Saloura V, Uppaluri R. Epigenetic modulation of immunotherapy and implications in head and neck cancer. *Cancer Metastasis Rev.* (2021) 40:141–52. doi: 10.1007/s10555-020-09944-0
- Hsiao JR, Chang CC, Lee WT, Huang CC, Ou CY, Tsai ST, et al. The interplay between oral microbiome, lifestyle factors and genetic polymorphisms in the risk of oral squamous cell carcinoma. *Carcinogenesis.* (2018) 39:778–87. doi: 10.1093/carcin/bgy053
- Solhi M, Fattahi E, Barati H, Mohammadi M, Kasmaei P, Rastaghi S. Smokeless Tobacco Use in Iran: A Systematic Review. *Addict Health.* (2020) 12:225–34. doi: 10.22122/ahj.v12i3.274
- Di Credico G, Polesel J, Dal Maso L, Pauli F, Torelli N, Luce D, et al. Alcohol drinking and head and neck cancer risk: the joint effect of intensity and duration. *Br J Cancer.* (2020) 123(9):1456–63. doi: 10.1038/s41416-020-01031-z
- Civelek M, Lusis AJ. Systems genetics approaches to understand complex traits. *Nat Rev Genet.* (2014) 15:34–48. doi: 10.1038/nrg3575
- Romanoski CE, Lee S, Kim MJ, Ingram-Drake L, Plaisier CL, Yordanova R, et al. Systems genetics analysis of gene-by-environment interactions in human cells. *Am J Hum Genet.* (2010) 86:399–410. doi: 10.1016/j.ajhg.2010.02.002
- Shen L, Bai J, Wang J, Shen B. The fourth scientific discovery paradigm for precision medicine and healthcare: Challenges ahead. *Precis Clin Med.* (2021) 4:80–4. doi: 10.1093/pcmedi/pbab007
- Gadewal NS, Zingde SM. Database and interaction network of genes involved in oral cancer: Version II. *Bioinformatics.* (2011) 6:169–70. doi: 10.6026/97320630006169
- Schmidt Jensen J, Jakobsen KK, Mirian C, Christensen JT, Schneider K, Nahavandipour A, et al. The Copenhagen Oral Cavity Squamous Cell Carcinoma database: protocol and report on establishing a comprehensive oral cavity cancer database. *Clin Epidemiol.* (2019) 11:733–41. doi: 10.2147/CLEP.S215399
- Pradhan S, Das S, Singh AK, Das C, Basu A, Majumder PP, et al. dbGENVOC: database of GENomic Variants of Oral Cancer, with special reference to India. *Database.* (2021) 2021:baab034. doi: 10.1093/database/baab034
- Ivanga M, Iroungou BA, Bigoundou-Nzamba S, Bisseye C. Accelerating cancer genomics research in Sub-Saharan Africa. *Front Oncol.* (2025) 15:1531799. doi: 10.3389/fonc.2025.1531799
- Wei CH, Allot A, Leaman R, Lu Z. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res.* (2019) 47:W587–93. doi: 10.1093/nar/gkz389
- Wei CH, Kao HY, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res.* (2013) 41:W518–522. doi: 10.1093/nar/gkt441
- Wei CH, Leaman R, Lu Z. Beyond accuracy: creating interoperable and scalable text-mining web services. *Bioinformatics.* (2016) 32:1907–10. doi: 10.1093/bioinformatics/btv760
- Aoki M, Fujishita T. Oncogenic Roles of the PI3K/AKT/mTOR Axis. *Curr Top Microbiol Immunol.* (2017) 407:153–89. doi: 10.1007/82_2017_6
- Stein Y, Rotter V, Aloni-Grinstein R. Gain-of-Function Mutant p53: All the Roads Lead to Tumorigenesis. *Int J Mol Sci.* (2019) 20:6197. doi: 10.3390/ijms20246197
- Britten CD. PI3K and MEK inhibitor combinations: examining the evidence in selected tumor types. *Cancer Chemother Pharmacol.* (2013) 71:1395–409. doi: 10.1007/s00280-013-2121-1
- Chuang FC, Wang CC, Chen JH, Hwang TZ, Yeh SA, Su YC. PI3k inhibitors (BKM120 and BYL719) as radiosensitizers for head and neck squamous cell carcinoma during radiotherapy. *PLoS One.* (2021) 16:e0245715. doi: 10.1371/journal.pone.0245715
- Tan Y, Wang Z, Xu M, Li B, Huang Z, Qin S, et al. Oral squamous cell carcinomas: state of the field and emerging directions. *Int J Oral Sci.* (2023) 15:44. doi: 10.1038/s41368-023-00249-w
- Tan TZ, Ye J, Yee CV, Lim D, Ngoi NYL, Tan DSP, et al. Analysis of gene expression signatures identifies prognostic and functionally distinct ovarian clear cell carcinoma subtypes. *eBioMedicine.* (2019) 50:203–10. doi: 10.1016/j.ebiom.2019.11.017
- Gan M, Liu N, Li W, Chen M, Bai Z, Liu D, et al. Metabolic targeting of regulatory T cells in oral squamous cell carcinoma: new horizons in immunotherapy. *Mol Cancer.* (2024) 23:273. doi: 10.1186/s12943-024-02193-7
- Yu E, Chu X, Zhang W, Meng X, Yang Y, Ji X, et al. Large Language Models in Medicine: Applications, Challenges, and Future Directions. *Int J Med Sci.* (2025) 22:2792–801. doi: 10.7150/ijms.111780
- Dingerdisen HM, Bastian F, Vijay-Shanker K, Robinson-Rechavi M, Bell A, Gogate N, et al. OncoMX: A Knowledgebase for Exploring Cancer Biomarkers in the Context of Related Cancer and Healthy Data. *JCO Clin Cancer Inform.* (2020) 4:210–20. doi: 10.1200/cci.19.00117
- Sun Y, Cheng G, Wei D, Luo J, Liu J. Integrating omics data and machine learning techniques for precision detection of oral squamous cell carcinoma: evaluating single biomarkers. *Front Immunol.* (2024) 15:1493377. doi: 10.3389/fimmu.2024.1493377
- Jayawickrama SM, Ranaweera PM, Pradeep R, Jayasinghe YA, Senevirathna K, Hilmi AJ, et al. Developments and future prospects of personalized medicine in head and neck squamous cell carcinoma diagnoses and treatments. *Cancer Rep (Hoboken).* (2024) 7:e2045. doi: 10.1002/cnr2.2045