Frontiers in Bioscience Elite

Front. Biosci. (Elite Ed) **2026**; 18(3): 49059
https://doi.org/10.31083/FBE49059

Original Research

Codon Usage Bias Analysis in Oral Cancer: Comprehensive Molecular and Evolutionary Insights

Manal Abouelwafa[1], John J Georrge[2], Supriyo Chakraborty[3,*]

[1]School of Biotechnology, Badr University in Assiut (BUA), Assiut, Egypt
[2]Department of Bioinformatics, University of North Bengal, Raja Rammohunpur, District Darjeeling, 734013 West Bengal, India
[3]Department of Biotechnology, Assam University, 788011 Silchar, Assam, India
*Correspondence: supriyoch_2008@rediffmail.com (Supriyo Chakraborty)

Academic Editor: Gaetano Isola
Submitted: 13 December 2025 Revised: 19 March 2026 Accepted: 13 April 2026 Published: 20 September 2026

Abstract

Background: Oral squamous cell carcinoma (OSCC) is represented as a major global health concern with few treatment options. Codon usage bias (CUB) is a non-random selection of synonymous codons. It offers information on gene expression patterns and molecular evolution. **Methods:** This study examined CUB patterns in 1328 differentially expressed genes (651 up-regulated, 677 down-regulated) from oral cancer tissues. Nucleotide composition analysis, effective number of codons (ENC), Relative Synonymous Codon Usage (RSCU) analysis, neutrality plot analysis, and gene expression analysis were performed to assess codon usage patterns and their relationship to gene expression levels and pathway classification. **Results:** Nucleotide composition analysis indicated a clear GC-rich tendency with base frequencies arranged as C > G > A > T for up-regulated genes and C > A > G > T for down-regulated genes. The average ENC values were 47.21 (up-regulated) and 49.08 (down-regulated), suggesting a low codon usage bias. RSCU analysis revealed 25 more frequently used (mean RSCU >1.0) and 34 less frequently used (mean RSCU <1.0) codons in up-regulated genes, while down-regulated genes had 30 more frequently used and 29 less frequently used codons. Neutrality plot analysis showed that both gene groups were predominantly shaped by selective constraints rather than mutational pressure (slopes of 0.202 and 0.228 for up- and down-regulated genes, respectively), though these slopes provide only indirect, qualitative evidence and should not be interpreted as exact quantitative contributions. Gene expression analysis revealed 1328 differentially expressed genes with functional enrichment in metabolic pathways and cellular components. **Conclusions:** These discoveries offer molecular insights into the genetic pathways that support oral cancer development. It implies that codon optimization could affect the translation efficiency of cancer-promoting genes.

Keywords: codon usage bias; oral cancer; ENC; RSCU; mutational pressure; natural selection; gene expression; translational efficiency

1. Introduction

The incidence of numerous diseases has increased significantly during the previous decade. One of the most common diseases is cancer. The increase in cancer patients is mostly due to the interaction of lifestyle factors and genetic predisposition [1]. This higher prevalence could be attributed to the inheritance of cancer-causing genes, as well as exposure to a variety of environmental variables. Head and neck cancers make up about 5% of all tumor types, with oral cancer accounting for almost half of these instances [2]. Oral cancer is the sixth most frequent cancer globally. It affects the lips, mouth, tongue, gingiva, palate, buccal mucosa, and alveolar mucosa [3]. Each year, more than 500,000 new cases are reported [4]. The epidemiology of oral cancer shows varied patterns in different demographic groups. Regarding gender distribution, oral cancer shows a significantly higher incidence and fatality rate in men compared to women [5]. Age is another key risk factor, with approximately 90% of oral and pharyngeal cancer cases occurring in people over the age of 40. Individuals aged 65 and older account for almost 50% of all oral cancer instances [6,7]. Oral cancer incidence varies significantly by geography, especially in low- and middle-income nations. The disease load is significantly higher. According to Shrestha et al. [8], South and Southeast Asia has the four countries with the greatest incidence of oral cancer. In these countries, oral and lip cavity cancer is the second most prevalent malignancy. This increased incidence in these locations is largely due to the extensive use of gutkha (processed betel nut), pan masala, pan tobacco, and other types of chewing tobacco. According to GLOBOCAN estimates, approximately 657,000 new cases of oral cavity and pharyngeal cancer are diagnosed each year globally, resulting in over 330,000 deaths [3].

Oral microbiota is essential for sustaining proper oral physiological conditions and homeostasis. Furthermore, poor oral hygiene can lead to periodontal disease and tooth loss. However, in some cases, changes in the oral microbiome might result in the development of oral cancer or chronic oral illnesses. The mouth cavity contains around 700 different microbial species, including bacteria, fungi, and viruses that can influence cellular systems. Tobacco chewing can alter the microbial equilibrium and encourage malignant transformation. For example, Gholizadeh et al.

[9] found a well-established link between Helicobacter pylori and stomach cancer. Candida species cause one of the most prevalent fungal infections in the oral cavity. However, under specific clinical situations, Candida infection can lead to the development of oral cancer [10]. Herpes Simplex Virus (HSV) is another major risk factor for mouth cancer. It can cause oral squamous cell carcinomas by integrating with the cell cycle machinery. This leads to the production of viral oncoproteins. The molecular mechanism underpinning this process includes HSV1, which can induce both ocular and oral infections. However, HSV2 is largely linked to genital infections [11]. Although these multiple risk factors might work together to increase the likelihood of developing oral cancer, the specific mechanisms remain incompletely understood.

This study's main goal was to thoroughly examine codon usage bias (CUB) patterns in genes that are differently expressed in oral squamous cell carcinoma (OSCC) in order to clarify the molecular processes that underlie the development of oral cancer. The main goals are to: (1) describe CUB patterns in 651 up-regulated and 677 down-regulated genes; (2) measure the relative contributions of natural selection and mutational pressure to CUB; (3) investigate the connection between CUB and gene expression levels; and (4) ascertain whether pathway classification or gene expression status is a better predictor of codon usage bias in oral cancer.

In order to accomplish these goals, we used a variety of analytical techniques, such as nucleotide composition analysis, effective number of codons (ENC) computation, and Relative Synonymous Codon Usage (RSCU), to examine the coding sequences of 651 up-regulated and 677 down-regulated genes from oral cancer tissues.

In this study, we investigated both up- and down-regulated genes involved in oral cancer metastasis to lymph nodes. We obtained the coding sequences of 651 up-regulated and 677 down-regulated genes from the National Center for Biotechnology Information (NCBI). To gain a thorough understanding of the molecular biology of oral cancer, the degree of CUB of genes, pattern of codon usage, level of gene expression, and the impact of various CUB-causing events were explored in both gene groups.

2. Materials and Methods

2.1 Data Acquisition

The coding sequences (CDS) of 651 up-regulated and 677 down-regulated genes associated with oral cancer were extracted from the GenBank database of the NCBI (https://www.ncbi.nlm.nih.gov) in FASTA format. Transcriptomic data for OSCC were retrieved from the NCBI Gene Expression Omnibus (GEO) database under the accession number GSE25099 (BioProject: PRJNA142727) [12]. The dataset consists of 79 samples, comprising 57 primary OSCC tumor tissues and 22 healthy oral mucosa controls. The samples were profiled using the Affymetrix Human Exon 1.0 ST Array [GPL5175] platform.

2.2 Codon Usage Bias Analysis

2.2.1 Effective Number of Codons (ENC)

The ENC is a simple metric of codon bias that ranges between 20 and 61. The degree of CUB in a gene is represented by this number. When the value of ENC is 20, it means that each amino acid is encoded by a single codon (severe bias), and when the value is 61, it means that all codons are used equally (no bias) [13]. ENC value was calculated with the help of the following formula:

$$ENC = 2 + 9/F_2 + 1/F_3 + 5/F_4 + 3/F_6$$

Where, F (a = 2, 3, 4 or 6) is the average of the F values of the amino acids with a-fold degeneracy.

2.2.2 Relative Synonymous Codon Usage (RSCU)

The RSCU metric of a codon is used to investigate how a codon differs in usage from other synonymous codons within a family. The ratio of the observed frequency of a codon to its expected frequency, considering all synonymous members, is referred to as the RSCU value of the codon.

2.2.3 Nucleotide Composition Analysis

The occurrence frequency of the four nucleotide bases (A, T, C, and G) in the 651 up-regulated genes of oral cancer was calculated. These bases at the third position of the codon (A3%, T3%, C3%, and G3%), as well as GC percent at the three positions of synonymous codons (GC1%, GC2%, and GC3%), were computed to explore the link between base composition and CUB [14].

2.2.4 Correspondence Analysis (COA)

The trend of CUB across all genes is reported via correspondence analysis (COA). All 651 up-regulated and 677 down-regulated genes of oral cancer have been studied using RSCU values of synonymous codons. Coding sequences were evaluated to establish the key factors affecting CUB, as well as the pattern of codon usage variation, according to COA [15].

2.2.5 Parity Plot Analysis

A parity plot was created to determine the effect of evolutionary forces affecting the CUB of genes. The Parity Rule 2 (PR2) figure was produced with the midpoint equal to 0.5 for the 2-fold, 4-fold, and 6-fold degenerate codon families. The earlier study has shown that deviations from parity rule 2 at third codon positions can indicate unequal mutation and selection pressures affecting codon usage [16]. The x-axis was represented by G3/(G3 + C3), and the y-axis by A3/(A3 + T3).

2.2.6 Neutrality Plot Analysis

The neutrality plot depicts the relationship between the G+C concentration of genes at the codon's third position (GC3) and the average of GC at the first and second positions (GC12). The most widely used method for estimating the balance between mutation pressure and natural selection is the neutrality plot. If the slope equals 1, the association between GC3 and GC12 could be quantitatively impacted by mutation pressure [17,18].

2.2.7 Mutation Responsive Index (MRI) and Translational Selection Index (P2)

The degree of CUB, which is altered by mutational pressure, can be determined using the mutation responsive index (MRI). The value of MRI can be either positive or negative. The role of mutational pressure on genes is revealed by a positive MRI value, while the involvement of translational selection is revealed by a negative MRI value [19]. The translational selection index (P2) is calculated to determine the role of translational selection on codon usage bias.

2.2.8 Protein Properties and Nucleotide Skew Properties

There are a few protein properties that might be related to CUB, namely hydrophilicity and aromaticity. Hydrophilicity score measures the degree of hydrophilicity of proteins. The number of aromatic amino acids (tryptophan, phenylalanine, and tyrosine) in a protein is measured by its aromaticity. In the protein, the grand average of hydropathicity (GRAVY) possesses a value ranging from –2 to +2, with the positive value indicating the hydrophobic nature of the protein encoded by the CDS [20,21]. Nucleotide skewness can be estimated to determine the compositional constraints of the coding sequences made up of codons. There are different types of nucleotide skews: AT skew (A–T/A+T), GC skew (G–C/G+C), purine skew (A–G/A+G), pyrimidine skew (T–C/T+C), amino skew (A–C/A+C), and keto skew (T–G/T+G). The positive value of any skew (e.g., AT skew) means the superiority of A over T and vice versa [22].

2.3 Gene Expression and Functional Analysis

2.3.1 Data Preprocessing and Normalization

Raw intensity data were processed using the Robust Multi-array Average (RMA) algorithm via the oligo or affy packages in R. This procedure included background correction, quantile normalization, and log2 transformation to ensure comparability across samples. Principal component analysis (PCA) of codon usage was performed on the Relative Synonymous Codon Usage (RSCU) values of 59 synonymous codons across 1328 differentially expressed genes, comprising 651 up-regulated and 677 down-regulated genes. Each gene was treated as an observation, and the RSCU values were used as variables to characterize variation in codon-usage patterns [23,24].

2.3.2 Differential Expression Analysis

To identify differentially expressed genes (DEGs) between tumor and normal tissues, linear modeling was applied using the limma (Linear Models for Microarray Data) package [24]. A design matrix was constructed to compare the two experimental groups. Empirical Bayes statistics were utilized to moderate the standard errors, improving the reliability of the results. Genes were considered significantly differentially expressed if they met the criteria of an adjusted p-value (Benjamini-Hochberg FDR) <0.05 and an absolute log2-fold change (|log2FC|) >1 [25,26].

2.3.3 Functional Enrichment Analysis

To elucidate the biological significance of the identified DEGs, Gene Ontology (GO) enrichment analysis was performed focusing on Biological Processes (BP), Cellular Components (CC), and Molecular Functions (MF) [27]. Additionally, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis was conducted to identify over-represented signaling pathways. All enrichment analyses were executed using the clusterProfiler package, with a significance threshold set at an adjusted p-value < 0.05 [28,29].

2.4 Statistical Analysis

Pearson correlation coefficients were calculated to assess relationships between variables. Statistical significance was determined at $p < 0.05$ (*) and $p < 0.01$ (**) levels. All statistical analyses were performed using R version 4.6.0 (R Foundation for Statistical Computing, Vienna, Austria).

3. Results

3.1 Nucleotide Composition of Up- and Down-Regulated Genes

Base C (28.47%) had the largest percentage among up-regulated genes, followed by G (27.11%), A (24.13%), and T (20.28%), with GC content of 55.58% and AT content of 44.41% (the four base percentages sum to 99.99% rather than 100.00% due to independent rounding). At the third codon position, C3 was most frequent (35.38%), followed by G3 (28.02%), T3 (20.03%), and A3 (16.57%); accordingly, GC3 was 63.40% and AT3 was 36.60%. Mean GC1s%, GC2s%, and GC3s% (CodonW) were 59.44%, 43.91%, and 63.40%, respectively. The nucleotide composition of the 677 down-regulated genes showed base frequency C (27.11%) > A (25.90%) > G (25.40%) > T (21.59%), with GC content of 52.51% and AT content of 47.49%. At the third codon position, C3 was most frequent (31.64%), followed by G3 (26.16%), T3 (22.64%), and A3 (19.56%); accordingly, GC3 was 57.80% and AT3 was 42.20%. Mean GC1s%, GC2s%, and GC3s% were 57.33%, 42.38%, and 57.80%, respectively. These findings indicate a GC-rich nucleotide composition in both up-regulated and down-regulated gene groups (Fig. 1, **Supplementary Table 1**).

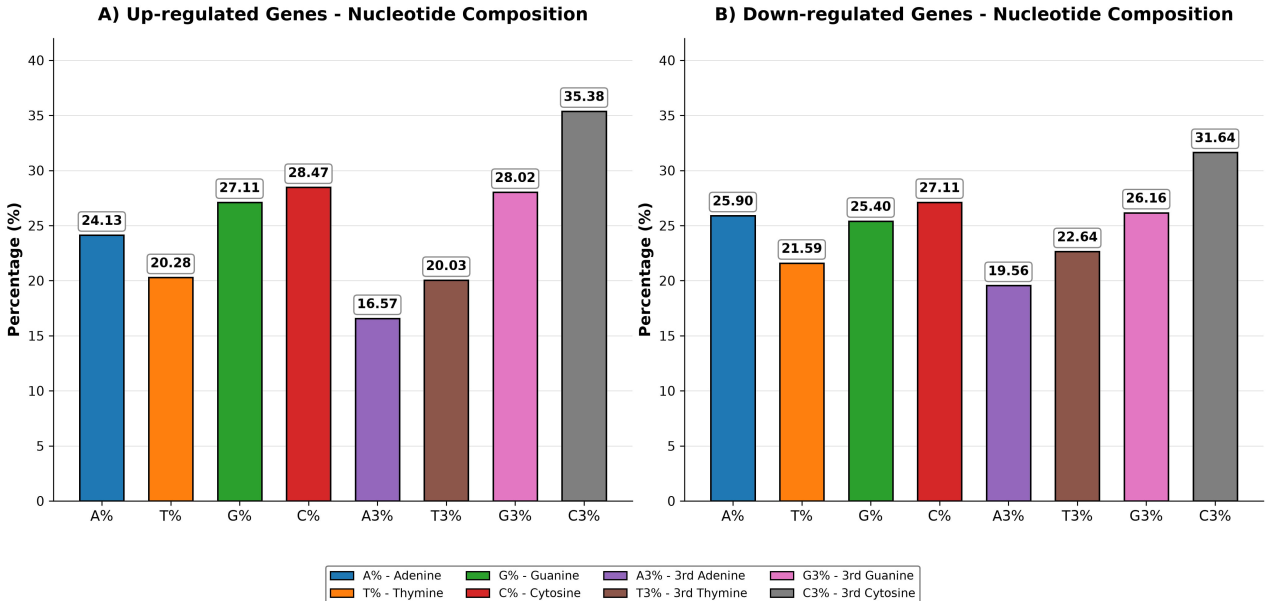


Fig. 1. Nucleotide composition analysis. Nucleotide composition of up-regulated and down-regulated genes in oral cancer. (A) Up-regulated genes (n = 651) show base C (28.47%) as the largest percentage, followed by G (27.11%), A (24.13%), and T (20.28%), with GC content of 55.58% and AT content of 44.41% (the four base percentages sum to 99.99% rather than 100.00% due to independent rounding). At the third codon position, C3 was highest (35.38%), followed by G3 (28.02%), T3 (20.03%), and A3 (16.57%); GC3 and AT3 were 63.40% and 36.60%, respectively. Mean GC1s%, GC2s%, and GC3s% were 59.44%, 43.91%, and 63.40%, respectively. (B) Down-regulated genes (n = 677) showed base frequency C (27.11%) > A (25.90%) > G (25.40%) > T (21.59%), with GC content of 52.51% and AT content of 47.49%. At the third codon position, C3 (31.64%) > G3 (26.16%) > T3 (22.64%) > A3 (19.56%); GC3 and AT3 were 57.80% and 42.20%, respectively. Mean GC1s%, GC2s%, and GC3s% were 57.33%, 42.38%, and 57.80%, respectively. The comparison reveals differential nucleotide bias patterns between the two gene groups, with both showing a GC-rich composition overall.

3.2 Detection of Codon Usage Bias Based on ENC

The ENC is a commonly used criterion for determining the CUB of genes. For the up-regulated genes, the mean ENC value was estimated for 651 genes and ranged from 14.9 to 60.0, showing high variation in the pattern of CUB among them. One up-regulated gene showed an ENC value of 14.9, slightly below the conventional theoretical minimum of 20; such values can arise from Wright's Nc estimator under conditions of extreme codon usage skew in individual genes and were retained in the analysis rather than excluded, as this point is a genuine data value rather than a calculation artifact (Fig. 2). As the mean ENC value was 47.21, which was >35, it suggests that overall, CUB was low in the up-regulated oral cancer genes. To evaluate the parameters responsible for determining CUB, an ENC-GC3 plot was created between the effective number of codons and GC content at the third position of the codon for the 651 up-regulated genes. The scatter diagram between ENC and GC3 depicted the solid line with a few dots near the expected curve (Fig. 2), suggesting there might be an effect of mutational pressure, while a few dots were far from the expected curve, suggesting there might also be a role of natural selection.

For the down-regulated genes, the ENC value was estimated to detect the degree of CUB for different genes related to oral cancer. The ENC value was found in the range of 27 to 60, suggesting a high degree of genetic variations in the usage of codons. However, as the mean ENC value was 49.08, which was >35, we could infer that the overall CUB was low. Further, an ENC-GC3 plot was generated with ENC value on the Y-axis and GC3 contents on the X-axis. The solid line represents the expected curve where a few dots were closer to the curve, indicating there might be an impact of mutational pressure, while a few other dots were discretely distributed, suggesting that natural selection might also have an impact on the CUB of genes (Fig. 2).

3.3 Codon Usage Patterns

3.3.1 Up-Regulated Genes

For all up-regulated genes, the RSCU values of all synonymous codons were determined. Notably, 59 synonymous codons encode 18 amino acids, and the study of these codons could help in understanding the CUB of genes (Fig. 3). The codons were divided into four categories: (a) overrepresented codons (RSCU value >1.6), (b) underrepresented codons (RSCU value <0.6), (c) more frequently used codons (RSCU value >1), and (d) less frequently used codons (RSCU value <1). The overrepresented codons were AGC, CTG, ATC, ACC, GTG, and GCC, which were mostly GC-ended. The underrepresented

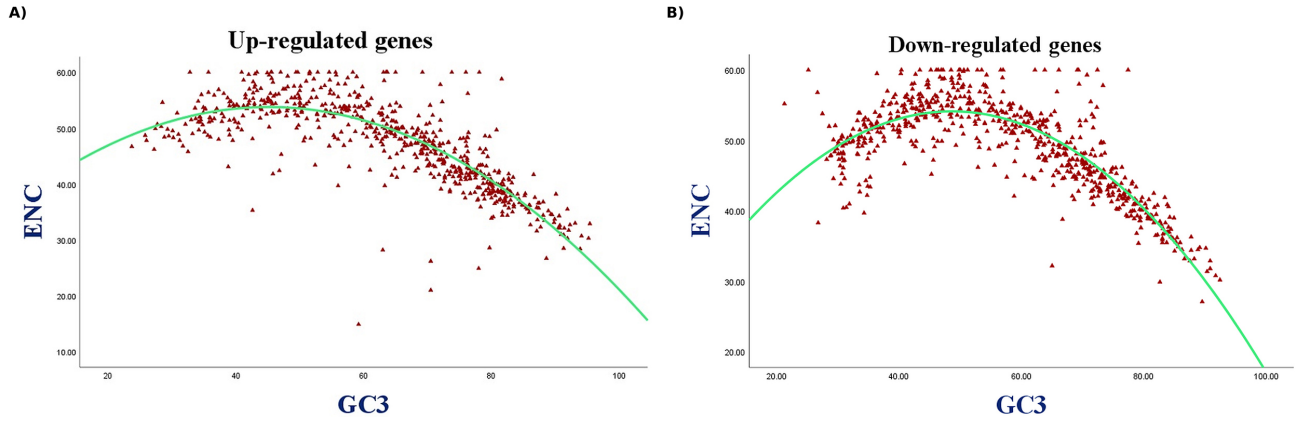


Fig. 2. Detection of codon usage bias based on ENC and GC3. Detection of codon usage bias based on ENC and GC3 content. (A) For up-regulated genes (n = 651), the mean ENC value was 47.21, suggesting low overall CUB since the value was >35. The ENC-GC3 plot depicts the expected curve (solid line) with dots near and far from the curve, suggesting both mutational pressure and natural selection effects. (B) For down-regulated genes (n = 677), the ENC value ranged from 27 to 60 with a mean value of 49.08, indicating low overall CUB. The dispersion of genes around the expected curve suggests that nucleotide composition and selective constraints may both contribute to codon-usage bias. ENC, effective number of codons; GC3, GC content at the third codon position; CUB, codon-usage bias.

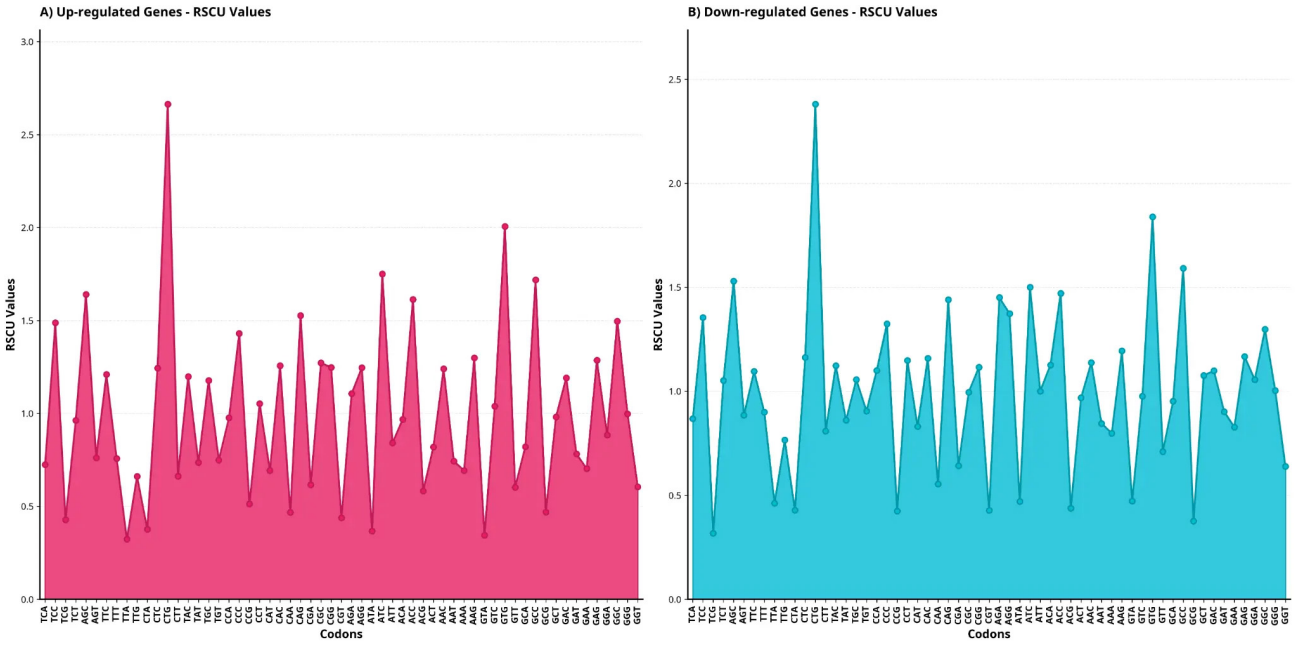


Fig. 3. Relative Synonymous Codon Usage (RSCU) analysis. Relative Synonymous Codon Usage (RSCU) analysis of differentially expressed genes in oral cancer. (A) For up-regulated genes (n = 651), 59 synonymous codons encode 18 amino acids. Overrepresented codons (RSCU >1.6) were AGC, CTG, ATC, ACC, GTG, and GCC (mostly GC-ended), while underrepresented codons (RSCU <0.6) were TCG, TTA, CTA, CCG, CAA, CGT, ATA, ACG, GTA, and GCG (mostly AT-ended). The analysis revealed 25 more frequently used codons (mean RSCU >1.0) and 34 less frequently used codons (mean RSCU <1.0). (B) For down-regulated genes (n = 677), overrepresented codons were CTG and GTG, while underrepresented codons were TCG, TTA, CTA, CCG, CAA, CGT, ATA, ACG, GTA, and GCG. There were 30 more frequently used codons (mean RSCU >1.0) and 29 less frequently used codons (mean RSCU <1.0). The comparison reveals distinct codon usage preferences between up-regulated and down-regulated genes.

codons were TCG, TTA, CTA, CCG, CAA, CGT, ATA, ACG, GTA, and GCG, which were mostly AT-ended as shown in (Fig. 4). The study revealed 25 more frequently used codons and 34 less frequently used codons. GC content is a major contributor to CUB. Higher GC contents might facilitate complex patterns of gene expression. Fur-

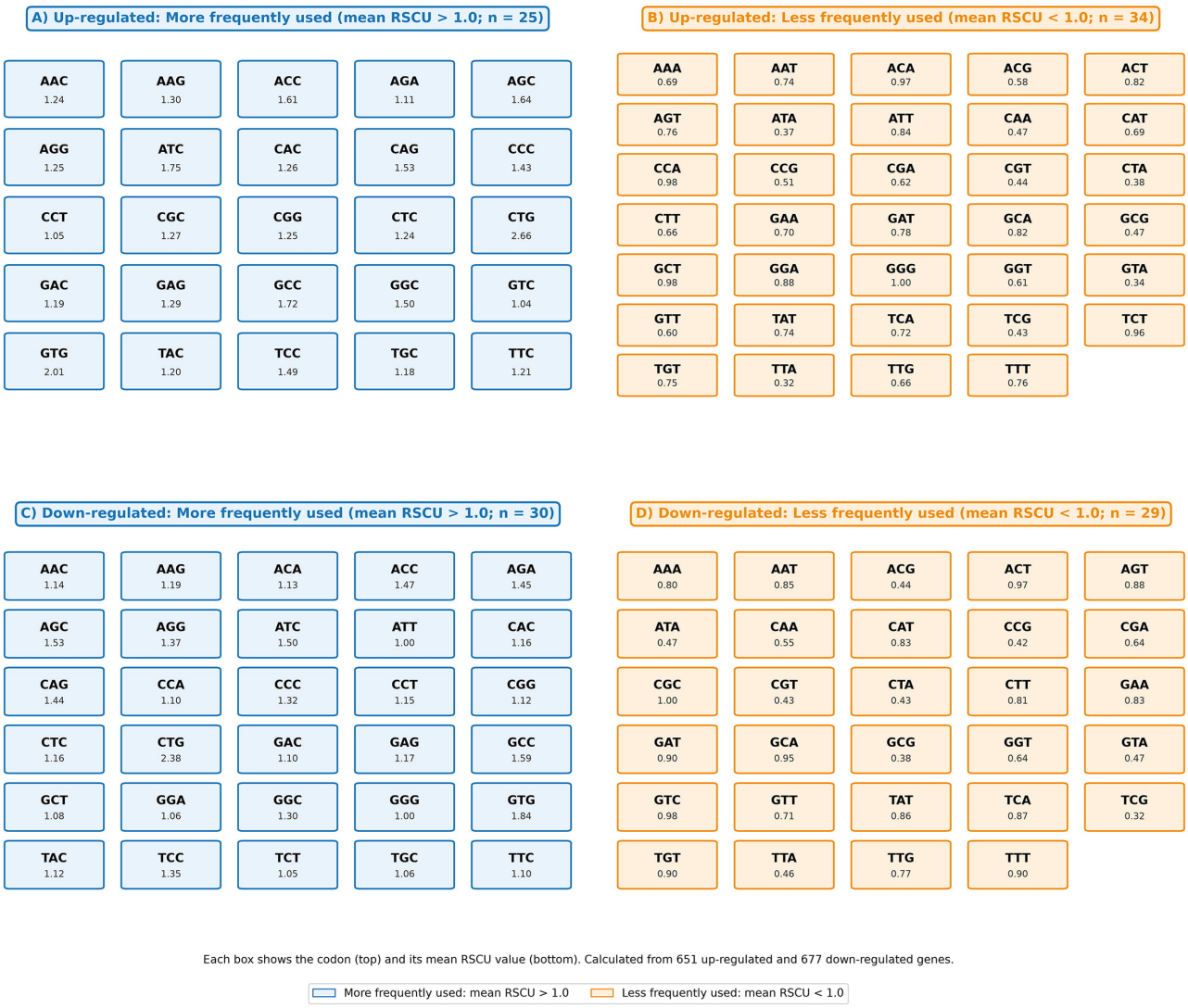


Fig. 4. Codon preference analysis. Codon preference analysis of differentially expressed genes in oral cancer. (A) Up-regulated genes with 25 more frequently used codons (mean RSCU >1.0; blue). (B) Up-regulated genes with 34 less frequently used codons (mean RSCU <1.0; orange). (C) Down-regulated genes with 30 more frequently used codons (mean RSCU >1.0; blue). (D) Down-regulated genes with 29 less frequently used codons (mean RSCU <1.0; orange). The overrepresented codons in up-regulated genes (AGC, CTG, ATC, ACC, GTG, GCC) are mostly GC-ended, while underrepresented codons (TCG, TTA, CTA, CCG, CAA, CGT, ATA, ACG, GTA, GCG) are mostly AT-ended. In down-regulated genes, CTG and GTG are overrepresented, while TCG, TTA, CTA, CCG, CAA, CGT, ATA, ACG, and GTA are underrepresented. The clear separation between preferred and avoided codons demonstrates significant codon usage bias in oral cancer.

ther, we plotted a heatmap using the correlation coefficients between RSCU values of 59 codons and GC3 contents to identify the GC bias in them. All AT-ended codons were found to be negatively correlated with GC3 contents (highlighted in green), whereas most GC-ended codons were found to be favorably correlated with GC3 contents (highlighted in red) (Fig. 5).

3.3.2 Down-Regulated Genes

We calculated the RSCU values of synonymous codons for all 677 down-regulated genes. In fact, 18 amino acids are encoded by 59 synonymous codons, except for two amino acids, methionine (Met) (ATG) and tryptophan (Trp) (TGG), which are encoded by a single codon each (Fig. 3). Based on RSCU values, the codons were categorized into four groups: (i) overrepresented codons (>1.6), (ii) underrepresented codons (<0.6), (iii) more frequently used codons (>1), and (iv) less frequently used codons (<1). The overrepresented codons were CTG and GTG, while the underrepresented codons were TCG, TTA, CTA, CCG, CAA, CGT, ATA, ACG, and GTA. There were 30 more frequently used codons and 29 less frequently used codons. Further, it was observed that the codons of leucine and serine were frequently used (Fig. 4). The relationship between

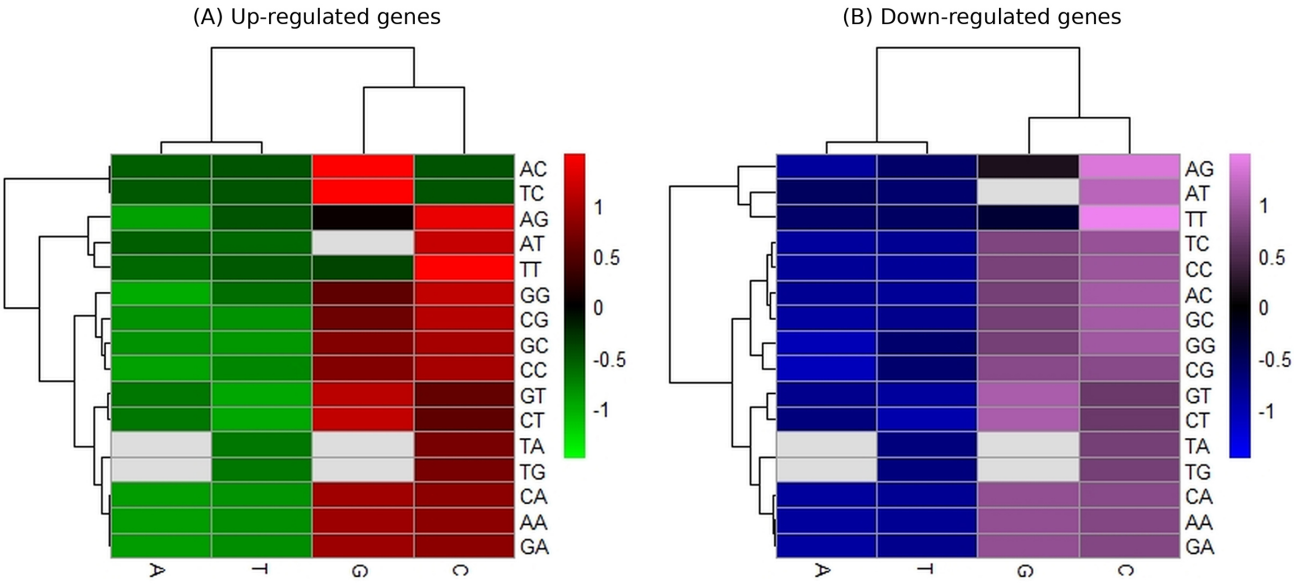


Fig. 5. Correlation between RSCU values and GC3 content in up-regulated and down-regulated genes. Correlation between RSCU values and GC3 content in up-regulated and down-regulated genes was visualized using R version 4.0.0 (R Foundation for Statistical Computing, Vienna, Austria). The heatmaps illustrate the correlation between the Relative Synonymous Codon Usage (RSCU) values of 59 codons and the GC content at the third codon position (GC3s). The color scale indicates the strength and direction of the correlation, with red/purple indicating a positive correlation and green/blue indicating a negative correlation. (A) Up-regulated genes: A heatmap was plotted using the correlation coefficients between RSCU values and GC3 contents. All AT-ended codons were found to be negatively correlated with GC3 contents (highlighted in green), whereas most GC-ended codons were favorably correlated with GC3 contents (highlighted in red). (B) Down-regulated genes: Similarly, the relationship between RSCU values of codons and GC3 contents was determined. The figure shows that GC-ended codons have a positive correlation with GC3 contents, while AT-ended codons have a negative correlation with GC3 contents.

RSCU values of codons and GC3 contents was determined using correlation coefficients and framed in a heatmap. The figure shows that GC-ended codons have a positive correlation with GC3 contents, while AT-ended codons have a negative correlation with GC3 contents (Fig. 5).

3.4 Correspondence Analysis of Codon Usage Patterns

COA plots were generated to obtain the differences in the usage pattern of codons in both up-regulated and down-regulated genes of oral cancer. The plots were framed with RSCU values of 59 codons. The results of the COA plots depict the trends of codon usage, and it was observed that a few dots were closer to the axis, while others were at discrete positions in both gene sets. This finding leads to the conclusion that both mutational pressure and natural selection may have influenced the codon usage pattern of genes, as the pattern of codon usage varied across genes and both evolutionary forces might have influenced the CUB of genes (Fig. 6).

3.5 PR2 Bias Analysis of Nucleotide Preferences

A parity plot is a graph that shows the relationship between AT and GC contents. The plots were drawn individually for 6-fold, 4-fold, and 2-fold degenerate codon families with [A/(A+T)] on the y-axis and [G/(G+C)] on the x-axis. In the PR2 plot, if GC and AT are used equally, then CUB may occur due to mutational pressure, while if GC and AT are not equally used, then CUB may occur due to natural selection. In our PR2 plots, the use of GC and AT was not equal. From this finding, we could conclude that both mutational pressure and natural selection may have an impact on the CUB of genes in oral cancer, suggesting that the CUB of both up-regulated and down-regulated genes may have evolved under the joint impact of mutation and natural selection (Fig. 7).

3.6 Neutrality Plot Assessment of Mutation and Selection

Neutrality-plot analysis examined the relationship between GC content at the first and second codon positions (GC12) and GC content at the third codon position (GC3), providing indirect evidence of the relative influence of mutational pressure and selective constraints on codon-usage bias. In both gene groups, GC12 was significantly and positively correlated with GC3: for up-regulated genes, the regression slope was 0.202 ($r = 0.492$, $p < 0.01$), and for down-regulated genes, the slope was 0.228 ($r = 0.605$, $p < 0.01$). Because both slopes are much closer to 0 than to 1, selective constraints appear to have exerted a greater rela-

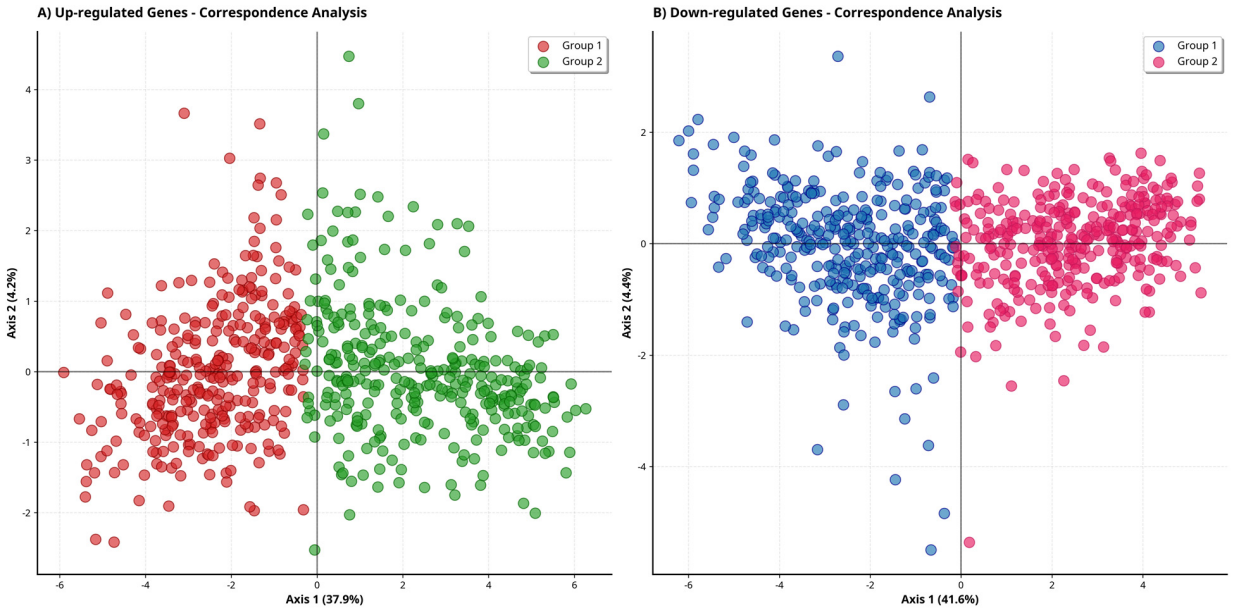


Fig. 6. Correspondence analysis of codon usage patterns. Correspondence analysis (COA) of codon usage patterns in differentially expressed genes. (A) COA plot for up-regulated genes (n = 651) showing the distribution of genes along Axis 1 and Axis 2, with dots closer to the axis and others at discrete positions, indicating variation in codon usage patterns. (B) COA plot for down-regulated genes (n = 677) showing similar distribution patterns. The results depict trends of codon usage and reveal that both mutational pressure and natural selection may have influenced the codon usage pattern of genes. The discrete positioning of dots across both gene sets suggests that both evolutionary forces might have influenced the CUB of genes in oral cancer.

tive influence than mutational pressure on codon-usage bias in both gene groups. However, neutrality-plot slopes provide only indirect, qualitative evidence of this relative influence and should not be converted into exact quantitative contributions of these evolutionary forces (Fig. 8).

It should be noted that the neutrality plot does not directly quantify the contribution of natural selection; rather, it evaluates the relationship between GC12 and GC3 content. The regression slope is therefore interpreted as an indirect indicator of the relative influence of selection versus mutational pressure, and these proportions should be understood as approximations rather than precise measurements.

3.7 Protein Properties and Skew Properties

3.7.1 Protein Properties

Earlier studies reported that a few protein properties are involved in the determination of CUB. A highly significant positive correlation was found between ENC value and aromaticity index **Supplementary Table 2** when correlation analysis of ENC was performed with different protein properties namely hydrophilicity, aromaticity, and GRAVY. This suggests that aromaticity index may have an impact on CUB generation.

3.7.2 Skew Properties

The skew properties of bases may have an impact on shaping the CUB. We, therefore, performed correlation analysis of ENC with different skew values **Supplementary Table 3**. For the up-regulated genes, a highly significant positive correlation was observed in GC skew, Purine skew, Pyrimidine skew, Keto skew, and Amino skew. For the down-regulated genes, analysis of base skew values revealed a mean AT skew value (0.09) as positive, while GC skew value (–0.03) was negative, suggesting that base A was predominant over T while C was predominant over G. Correlation analysis was performed between skew values and ENC value **Supplementary Table 3**, showing a highly significant positive correlation of ENC with all skew properties (AT skew, GC skew, Purine skew, Pyrimidine skew, Keto skew, and Amino skew). Thus, the skew properties of bases might have affected the pattern of CUB in both up-regulated and down-regulated genes of oral cancer.

3.8 Effect of MRI and P2 on Codon Usage Bias

The MRI and P2 are critical indicators of mutation pressure and translational selection on genes. If the value of MRI is positive, it refers to the role of mutational pressure, while if the P2 value is greater than 0.5, it refers to the role of translational selection. MRI and P2 values were calculated for up-regulated and down-regulated genes in oral cancer. For the up-regulated genes, the mean MRI value was 0.247 and the mean P2 value was 0.015 (<0.5), suggesting that mutation pressure played a significant role in determining the CUB, while translational selection had little impact. For the down-regulated genes, the mean MRI value was 0.470, suggesting the role of mutation, while the

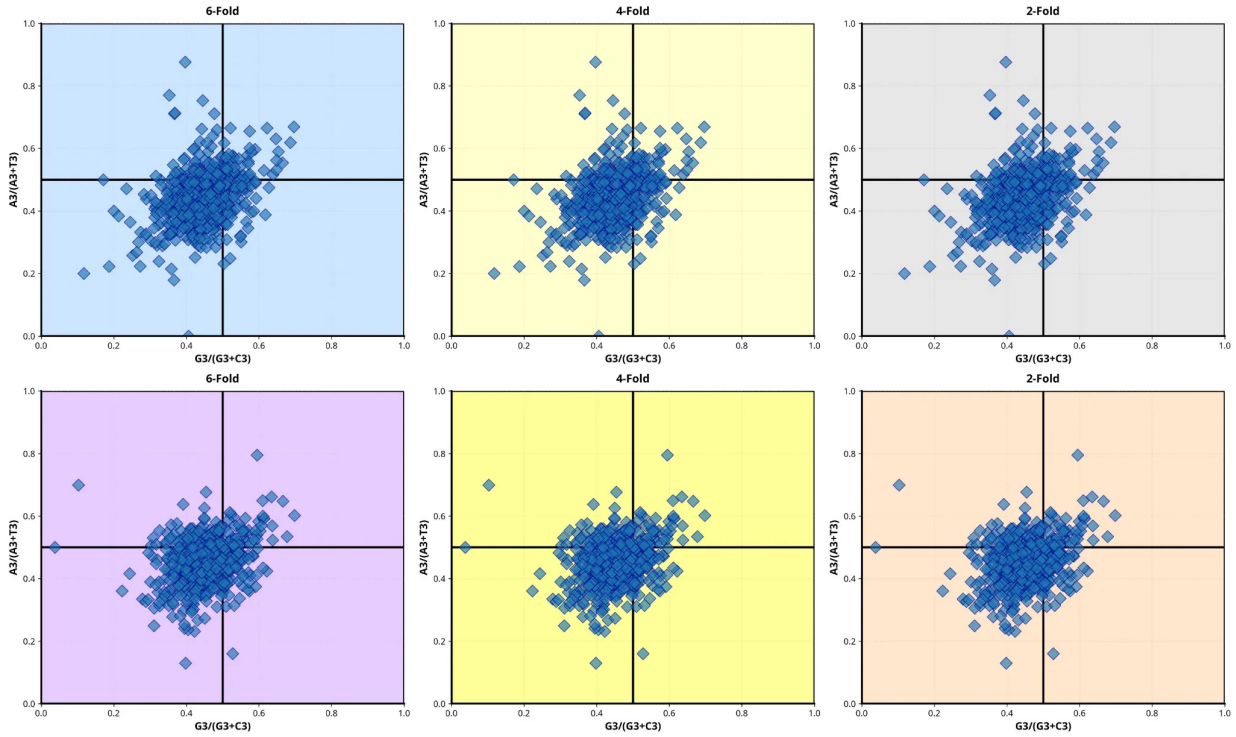
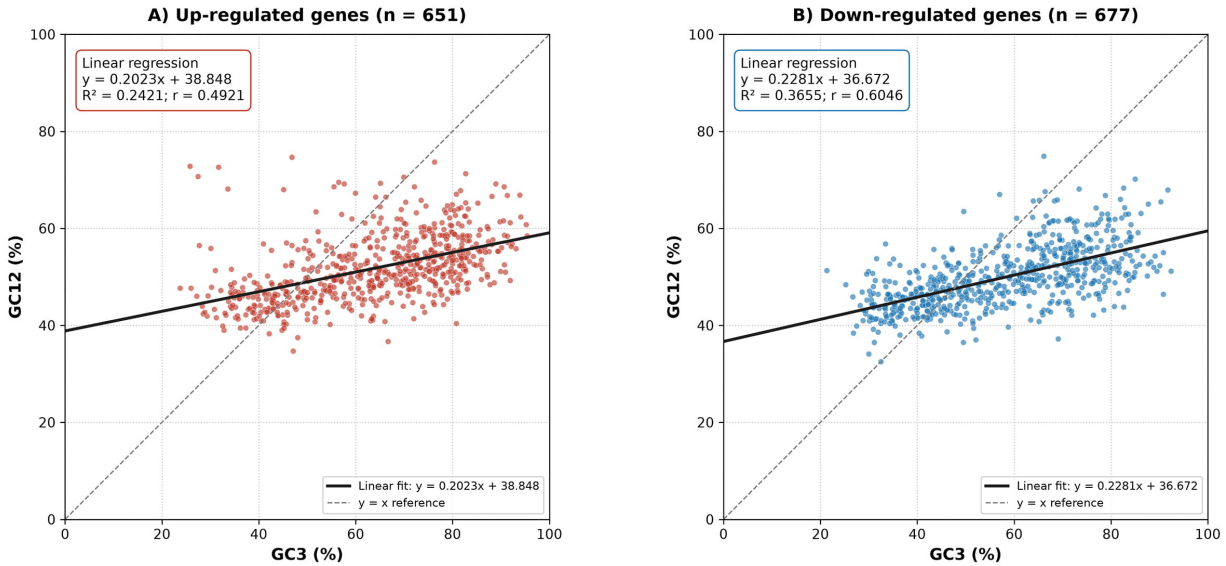


Fig. 7. Parity plots: third position codon bias analysis. Parity plots showing third-position codon bias analysis across different codon degeneracy families. The plots display the relationship between G3/(G3+C3) and A3/(A3+T3) ratios for individual genes in up-regulated (top row, 651 genes) and down-regulated (bottom row, 677 genes) gene groups. Three degeneracy families are shown: 6-fold (left), 4-fold (middle), and 2-fold (right). Each diamond represents a single gene, with reference lines at 0.5 on both axes dividing the plot into four quadrants representing different codon usage bias patterns. The distribution of genes across quadrants reveals the extent and nature of codon usage bias in each gene group.



GC12 represents the mean GC content at the first and second codon positions. The regression slopes provide indirect evidence of the relative influence of mutational pressure and selective constraints.

Fig. 8. Neutrality plot analysis. Scatter plots show the relationship between GC content at the third codon position (GC3; x-axis) and the mean GC content at the first and second codon positions (GC12; y-axis) for up-regulated (A) and down-regulated (B) genes. The regression equations were y = 0.2023x + 38.848 for up-regulated genes and y = 0.2281x + 36.672 for down-regulated genes. The slopes provide indirect evidence of the relative influence of mutational pressure and selective constraints on codon-usage bias.

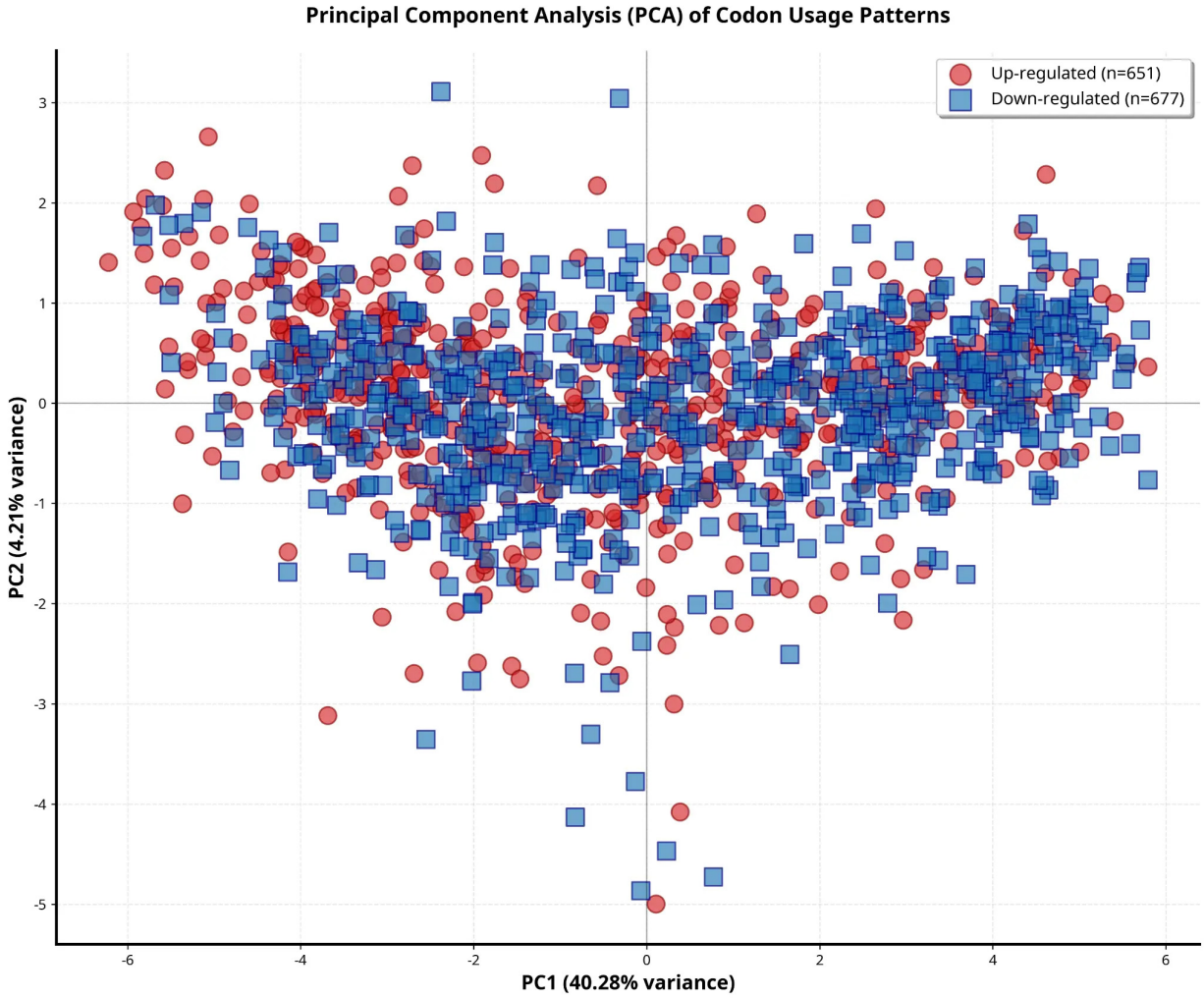


Fig. 9. Principal component analysis of codon-usage patterns. PCA was performed using RSCU values for 59 synonymous codons across 651 up-regulated and 677 down-regulated genes. Each point represents one gene; red circles denote up-regulated genes and blue squares denote down-regulated genes. PC1 and PC2 explained 40.28% and 4.21% of the variance, respectively. The two gene groups show substantial overlap, indicating that codon usage bias alone does not clearly separate up- from down-regulated genes.

P2 value was 0.086 (<0.5), indicating that the role of translational selection was weak. To determine the impact of mutation and translational selection on CUB, we correlated the ENC value with MRI and P2 values and found a highly significant positive correlation between them, suggesting that both mutation and translational selection might have influenced the CUB of genes in oral cancer. This could be due to multiple selective pressures beyond translational efficiency, such as mRNA stability, protein folding, and other evolutionary constraints.

It is important to clarify that P2 values only represent translational selection (codon optimisation for translation efficiency), whereas the neutrality plot evaluates overall natural selection from several sources. The low P2 values (<0.5) indicate weak translational selection, whereas the low neutrality-plot slopes suggest a predominant influence of selective constraints relative to mutational pressure. These slopes provide indirect evidence and should not be interpreted as exact quantitative contributions (**Supplementary Table 4**).

3.9 Gene Expression and Functional Analysis

3.9.1 PCA Plot (Quality Control)

Data preprocessing and normalization: Gene expression data for dataset GSE25099 (Affymetrix Human Exon 1.0 ST Array, GPL5175) were normalized using R version 4.5.2 (R Foundation for Statistical Computing, Vienna, Austria). To examine whether codon usage patterns distinguish up-regulated from down-regulated genes, PCA was performed on RSCU values of all 59 sense codons across the 1328 differentially expressed genes. PC1 and PC2 explained 40.28% and 4.21% of the total variance, respectively. Up-regulated (n = 651) and down-regulated (n = 677) genes showed substantial overlap in this space, indicating that codon usage bias alone does not clearly distinguish the two gene groups (Fig. 9).

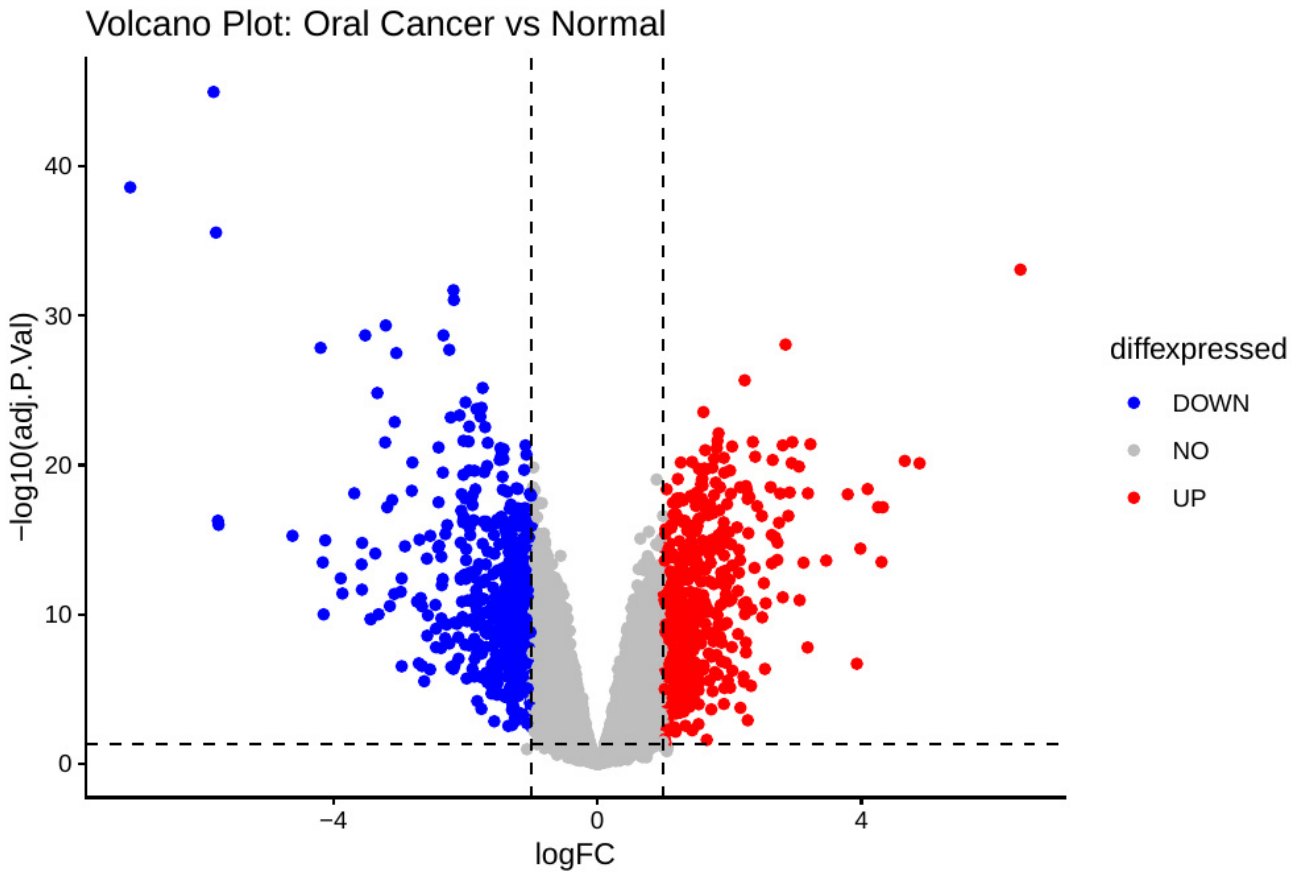


Fig. 10. Volcano plot: differential gene expression. Volcano plot showing differential gene expression between oral cancer and normal tissues. Differential expression analysis was performed with screening conditions of |log2-fold change| >1 and $p < 0.05$. A total of 1328 significantly differentially expressed genes (DEGs) were identified across 4328 genes analyzed, of which 651 were up-regulated and 677 were down-regulated genes. Blue dots represent down-regulated genes (lower expression in oral cancer), red dots represent up-regulated genes (higher expression in oral cancer), and gray dots represent genes with no significant differential expression. Vertical dashed lines indicate fold-change thresholds (logFC = ±1), while the horizontal dashed line indicates the *p*-value significance threshold ($p = 0.05$). Genes positioned in the upper left and upper right regions are significantly differentially expressed with large fold changes and high statistical significance.

3.9.2 Volcano Plot and Differential Expression Analysis

We screened the conditions of |log2-fold change| >1 and $p < 0.05$ for subsequent analysis. A total of 1328 significantly differentially expressed genes (DEGs) across 4328 genes analyzed, of which 651 up-regulated and 677 down-regulated genes (Fig. 10). Descriptive statistics for ENC, Fop, GC3%, base composition, and protein properties across all 1328 differentially expressed genes are provided in **Supplementary Table 5**. A heatmap of 50 selected DEGs in normal and tumor groups illustrated distinct clustering patterns. The heatmap displays 50 selected differentially expressed genes across 22 normal samples and 57 tumor samples. The results show that down-regulated genes were predominantly expressed in normal tissue and suppressed in tumor tissue. Most down-regulated genes *(STAT1*, *IFI6*, *GBP1*, and *COL4A1*) are displayed with red color. Unlike up-regulated genes that were predominantly expressed in tumor tissue and suppressed in normal tissue. Most up-regulated genes (*TYR*, *TYRP1*, *KRT3*, *KRT76*, and *MLANA*) are displayed with blue color. However, there is some heterogeneity in expression pattern (particularly in the middle) (Fig. 11).

3.9.3 Gene Ontology (GO) Enrichment

Gene Ontology enrichment analysis was performed using the differentially expressed genes in oral cancer. The findings show enrichment in three primary categories: biological processes (GO:BP), cellular components (GO:CC), and molecular functions (GO:MF). Biological processes were enriched for metabolic process, biological_process, cellular process, regulation of cellular process, regulation of biological process, biological regulation, response to stimulus, multicellular organismal process, developmental process, and anatomical structure development, with hits ranging from approximately 2% to 3.3% of enriched genes. The KEGG pathway analysis

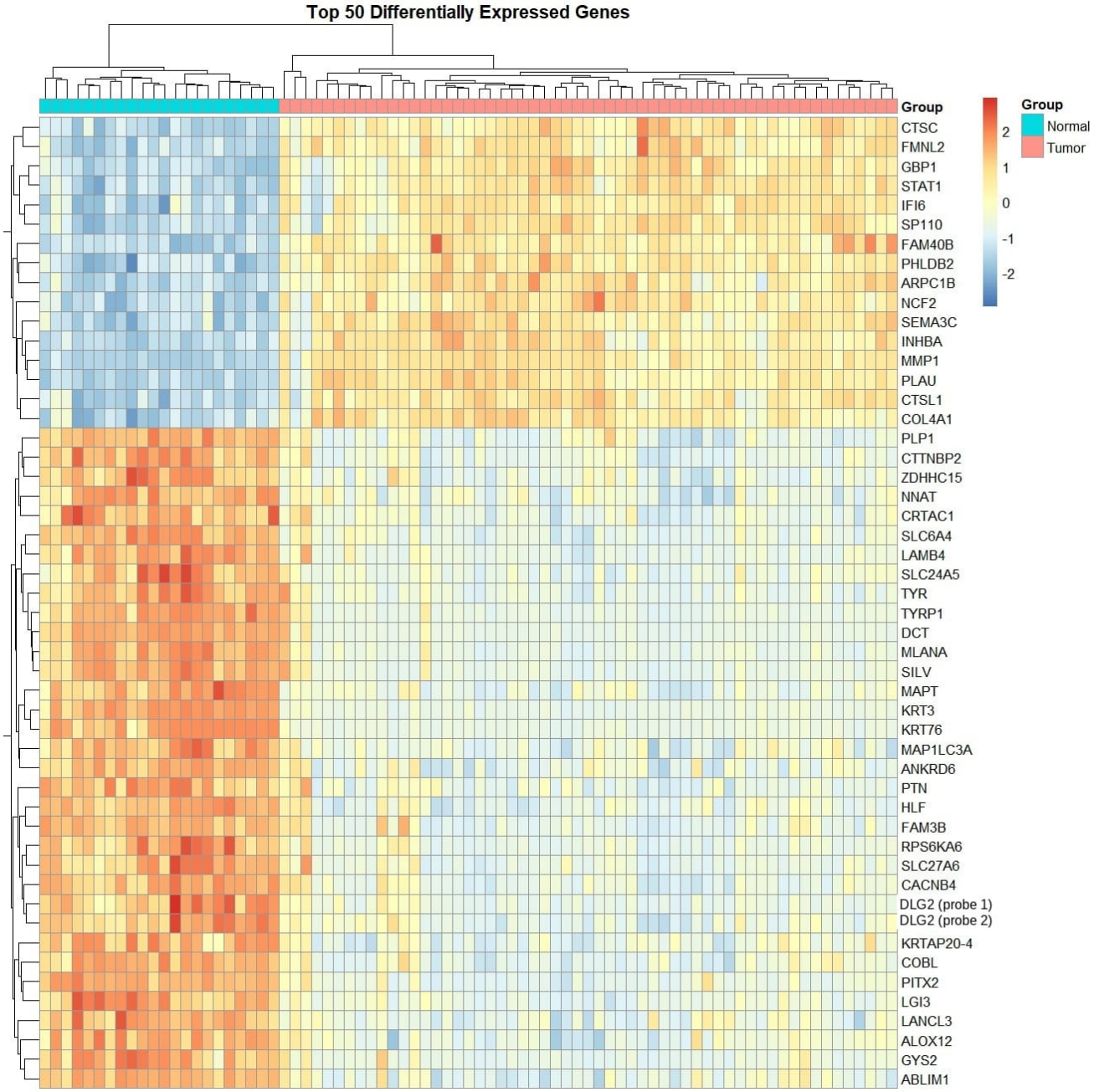


Fig. 11. Hierarchical clustering of top 50 differentially expressed genes. Hierarchical clustering heatmap of the top 50 differentially expressed genes across 22 normal samples and 57 tumor samples. Down-regulated genes were predominantly expressed in normal tissue and suppressed in tumor tissue, with most down-regulated genes (*STAT1*, *IFI6*, *GBP1*, and *COL4A1*) displayed with red color. Up-regulated genes were predominantly expressed in tumor tissue and suppressed in normal tissue, with most up-regulated genes (*TYR*, *TYRP1*, *KRT3*, *KRT76*, and *MLANA*) displayed with blue color. However, there is some heterogeneity in expression pattern, particularly in the middle region. This visualization demonstrates the distinct transcriptomic signature of oral cancer tissues compared to normal tissues.

revealed enrichment in nutritional absorption (hexose absorption ~25% of pathway-enriched genes, intestinal absorption ~25%), lipid metabolism (alpha-linolenic acid metabolism ~17%), and inflammatory processes (Kinin Kallikrein pathway ~14%). Cellular component enrichment included intracellular organelle, intracellular anatomical structure, membrane-bounded organelle, organelle, cellular_component, cellular anatomical structure, cytoplasm, membrane, cell periphery, and extracellular region, with the extracellular region showing the highest enrichment (~4% of hits). KEGG also revealed enrichment in attachment and entry pathways (~25%). Additionally, molecular functions

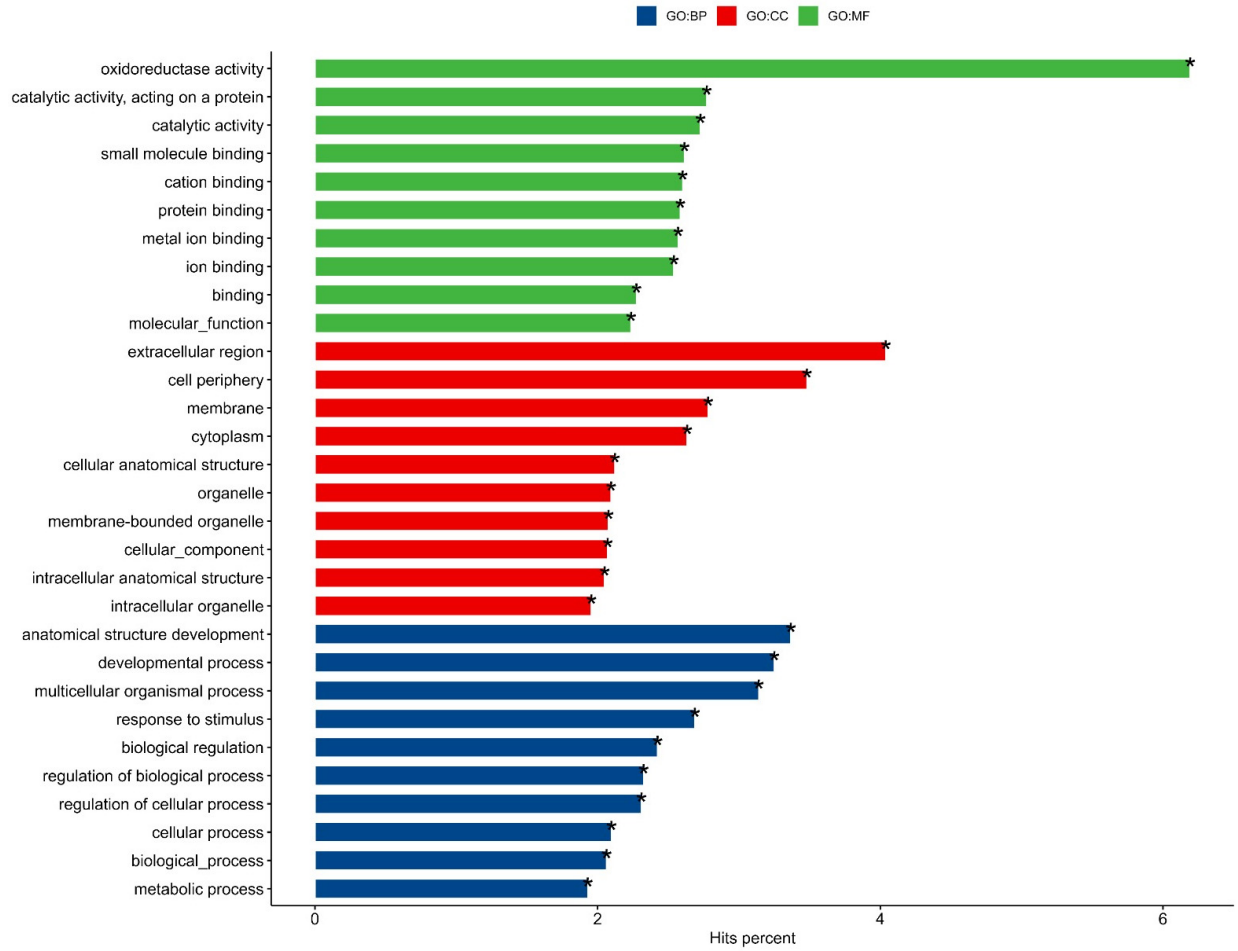


Fig. 12. Gene ontology enrichment analysis. Gene Ontology enrichment analysis of differentially expressed genes in oral cancer, showing the top enriched terms across three categories: biological process (GO:BP, blue), cellular component (GO:CC, red), and molecular function (GO:MF, green). Bars represent the percentage of genes mapping to each term (hits percent); asterisks indicate statistically significant enrichment. Among GO:BP terms, anatomical structure development and developmental process showed the highest enrichment (~3.3%). Among GO:CC terms, extracellular region showed the highest enrichment (~4%), followed by cell periphery (~3.5%). Among GO:MF terms, oxidoreductase activity showed markedly higher enrichment (~6%) than all other terms, most of which clustered between 2% and 3%.

showed notable enrichment in oxidoreductase activity, catalytic activity acting on a protein, catalytic activity, small molecule binding, cation binding, metal ion binding, protein binding, ion binding, binding, and molecular_function, with oxidoreductase activity showing the highest enrichment (~6% of hits). Additionally, KEGG enrichment was noted in metabolic pathways and growth factor signaling, including signaling by ligand-responsive EGFR variants in cancer and constitutive signaling by ligand-responsive EGFR cancer variants (approximately 6% each), as well as signaling by KIT in disease and PI-3K cascade:FGFR4 signaling (approximately 6% each) (Figs. 12,13).

3.9.4 Expression vs. Codon Usage Bias

The relationship between gene expression and codon-usage bias was assessed by examining the association between log2-fold change and ENC values for all differentially expressed genes (Fig. 14). The results showed a statistically significant but weak positive correlation ($r = 0.197$, $p = 0.00995$). Although statistically significant, this weak correlation ($r^2 = 0.039$) indicates that gene expression explains only approximately 3.9% of the variance in CUB. Multiple CUB metrics (ENC, Fop, and GC3) were examined in relation to expression level (Fig. 14B). The median ENC was 48.4 for up-regulated genes and 50.0 for down-regulated genes; the corresponding mean ENC values were 47.21 and 49.08, respectively (Mann-Whitney U test, $p < 0.001$), indicating a modest but statistically significant difference in codon usage bias between the two gene groups.

3.9.5 CUB Metrics Stratified by Pathway Category

Fig. 15 summarizes codon-usage-bias metrics among genes assigned to the “Other” pathway category. Panel A displays the ENC distribution across all differentially ex-

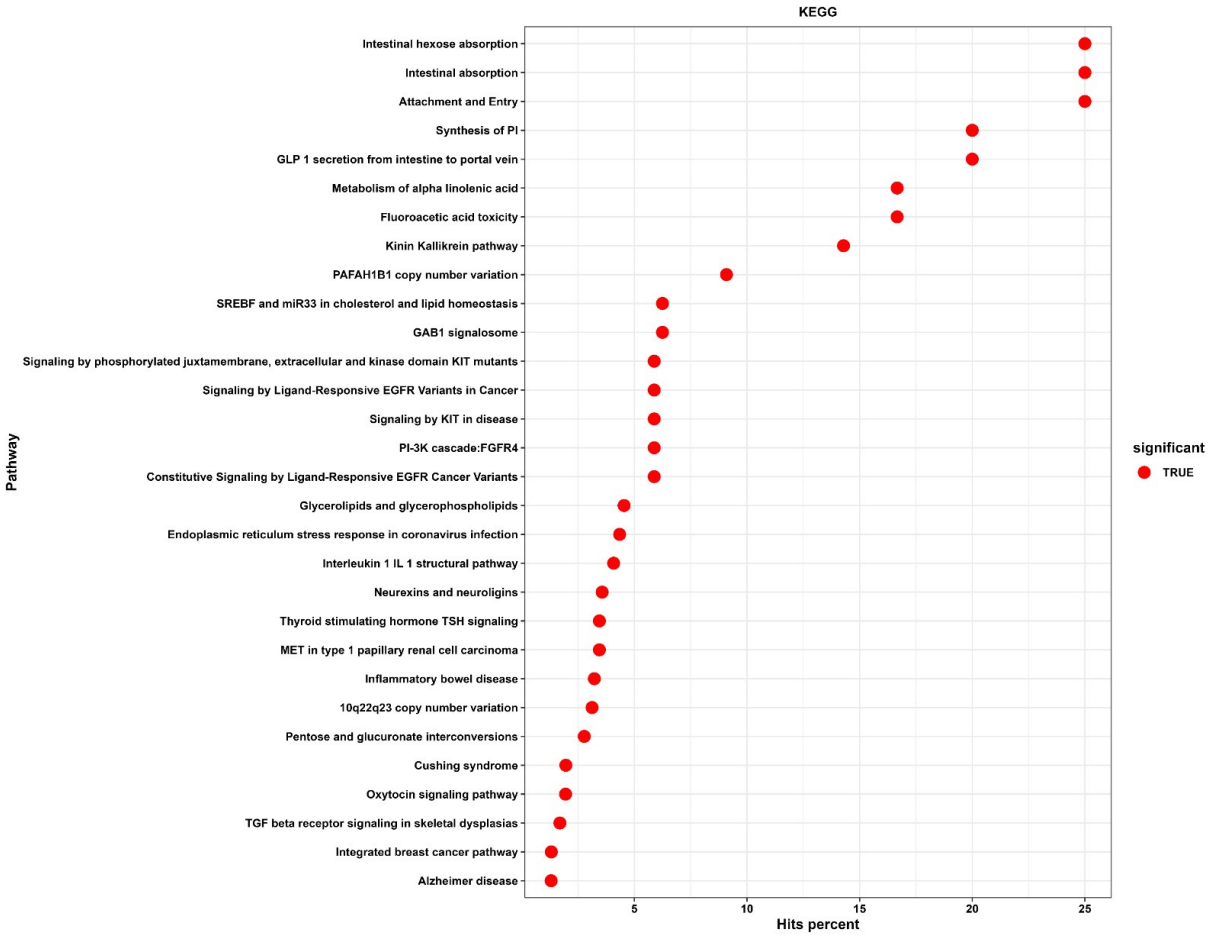


Fig. 13. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis. KEGG pathway enrichment analysis of differentially expressed genes in oral cancer. The scatter plot displays significantly enriched KEGG pathways (red dots, $p < 0.05$) ranked by the percentage of genes mapping to each pathway (hits percent, x-axis). The top enriched pathways were intestinal hexose absorption, intestinal absorption, and attachment and entry (~25% each), followed by synthesis of phosphatidylinositol and GLP-1 secretion from intestine to portal vein (~20% each), and metabolism of alpha-linolenic acid and fluoroacetic acid toxicity (~17% each). Several cancer-associated signaling pathways were also significantly enriched at lower hit percentages (~6%), including signaling by ligand-responsive EGFR variants in cancer, constitutive signaling by ligand-responsive EGFR cancer variants, signaling by KIT in disease, and PI-3K cascade:FGFR4 signaling.

pressed genes in this category. Panel B shows ENC distributions separately for down-regulated and up-regulated genes. Panel C presents the distributions of ENC, frequency of optimal codons (Fop), and GC content at the third codon position (GC3). Panel D shows the corresponding mean values: ENC, 54.58; Fop, 0.44; and GC3, 58%. Because cancer-related pathways are not displayed in Fig. 15, the figure does not support comparisons between cancer-related and other pathways or a conclusion that gene-expression status is more important than pathway classification.

4. Discussion

Cancer is the outcome of the dysfunction of genes led by mutations resulting in cell proliferation that affects cell growth in an irregular manner. There are different oncogene families which can affect cancer proliferation, e.g., RAS in human genes. The specific role of RAS family is not clear but by studying more about the oncogenes it will be easy to stop this disease [30]. Because of changes in genes linked to certain habits, oral cancer is one of the most lethal types of cancer. Several studies reported earlier that gene alteration and silent mutation in oral cancer developed from some wrong habits like smoking, chewing tobacco etc [31]. One of the reasons for the difficulty in studying the genetic characteristics of oral cancer is that it is not stable [32]. As there are no specific biomarkers for oral cancer, genetic study could help in prognosis and diagnosis of the disease.

In this study, we analyzed both up-regulated and down-regulated genes implicated in oral cancer metastasis to lymph nodes [33]. We retrieved the coding sequences

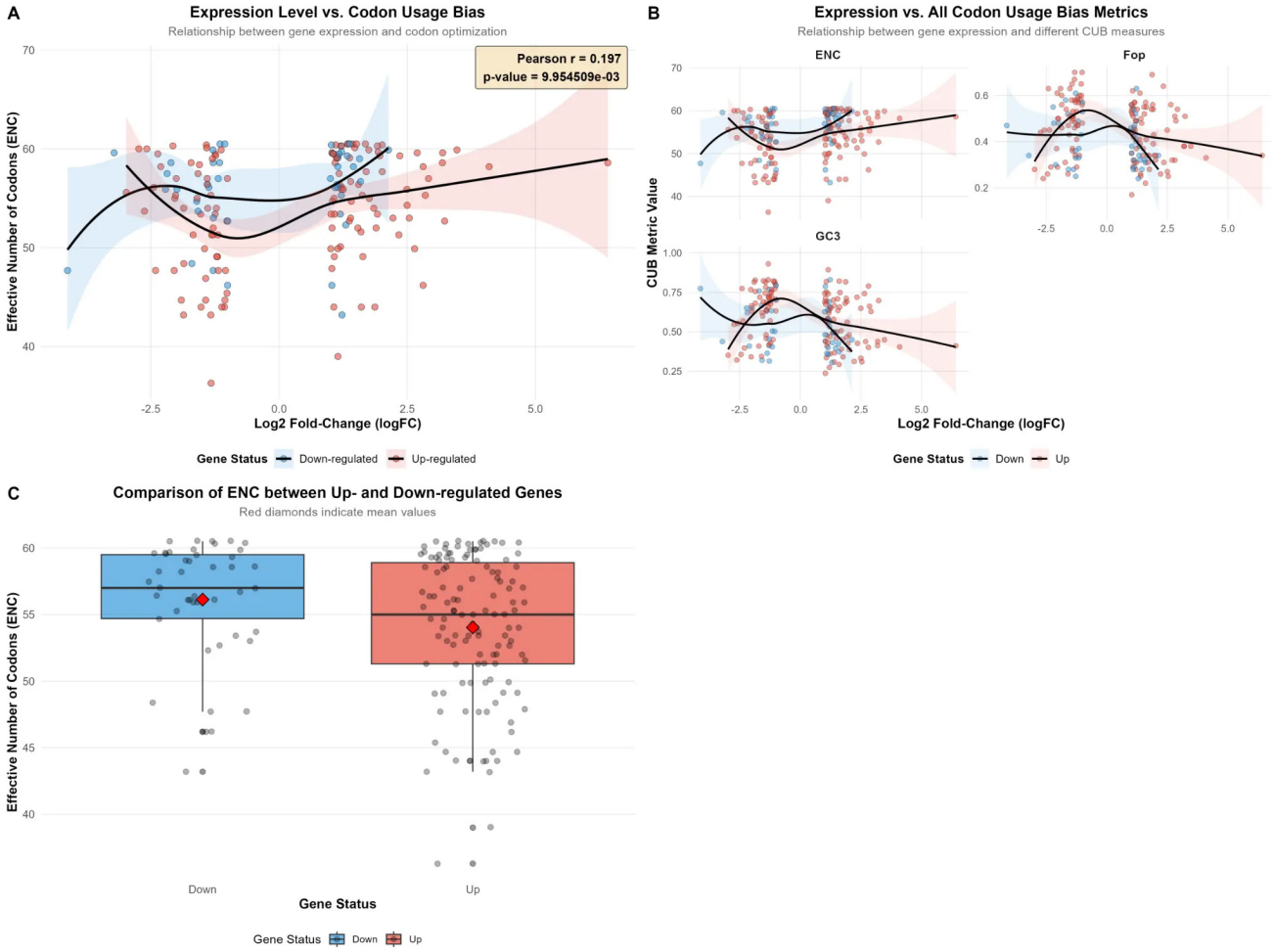


Fig. 14. Relationship between gene expression and codon usage bias. Relationship between gene expression level and codon usage bias in oral cancer. (A) Scatter plot showing the association between log2-fold change (logFC) in gene expression and Effective Number of Codons (ENC), a measure of codon usage bias. Blue dots represent down-regulated genes and red dots represent up-regulated genes. A non-linear loess regression curve (black line) reveals a weak but significant positive correlation (Pearson r = 0.197, *p*-value = 9.95 × 10^{-3}) between expression level and codon usage bias. The shaded blue and red regions indicate 95% confidence intervals for each gene group. Genes with higher expression levels (positive logFC) tend to show slightly higher ENC values, i.e., weaker codon usage bias, while down-regulated genes show greater variability in codon usage patterns. (B) Multi-panel analysis displaying the relationship between logFC and three different codon usage bias metrics: ENC (top), Fop (middle), and GC3 content (bottom). Each panel shows scatter plots with loess regression curves for both up-regulated (red) and down-regulated (blue) genes. The ENC plot demonstrates a non-linear relationship with expression, while Fop shows a declining trend with increasing expression. GC3 content shows a more complex relationship with expression level, suggesting that different aspects of codon bias are differentially associated with gene expression. (C) Box plot comparison of ENC values between up-regulated and down-regulated genes. The median ENC was 48.4 for up-regulated genes and 50.0 for down-regulated genes. The corresponding mean ENC values were 47.21 and 49.08, respectively. Individual data points (gray dots) are overlaid on the box plots to show heterogeneity in codon-usage bias within each gene group. Collectively, these analyses demonstrate that while gene expression level is weakly associated with codon usage bias in oral cancer, factors beyond codon usage bias likely influence gene expression levels in this context.

of 651 up-regulated genes and 677 down-regulated genes from NCBI. The degree of CUB of genes, pattern of codon usage, level of gene expression, and the impact of different factors responsible for CUB were investigated in both gene groups for an in-depth understanding of the molecular biology of oral cancer.

The nucleotide bases in the genes were in the order of C > G > A > T for up-regulated genes and C > A > G > T for down-regulated genes, and both gene groups were GC-rich. Since synonymous mutation has an effect on the third position of codons, the nucleotide bases in the third codon position are differentiated as GC3 (63.40% in up-regulated genes and 57.80% in down-regulated genes) and

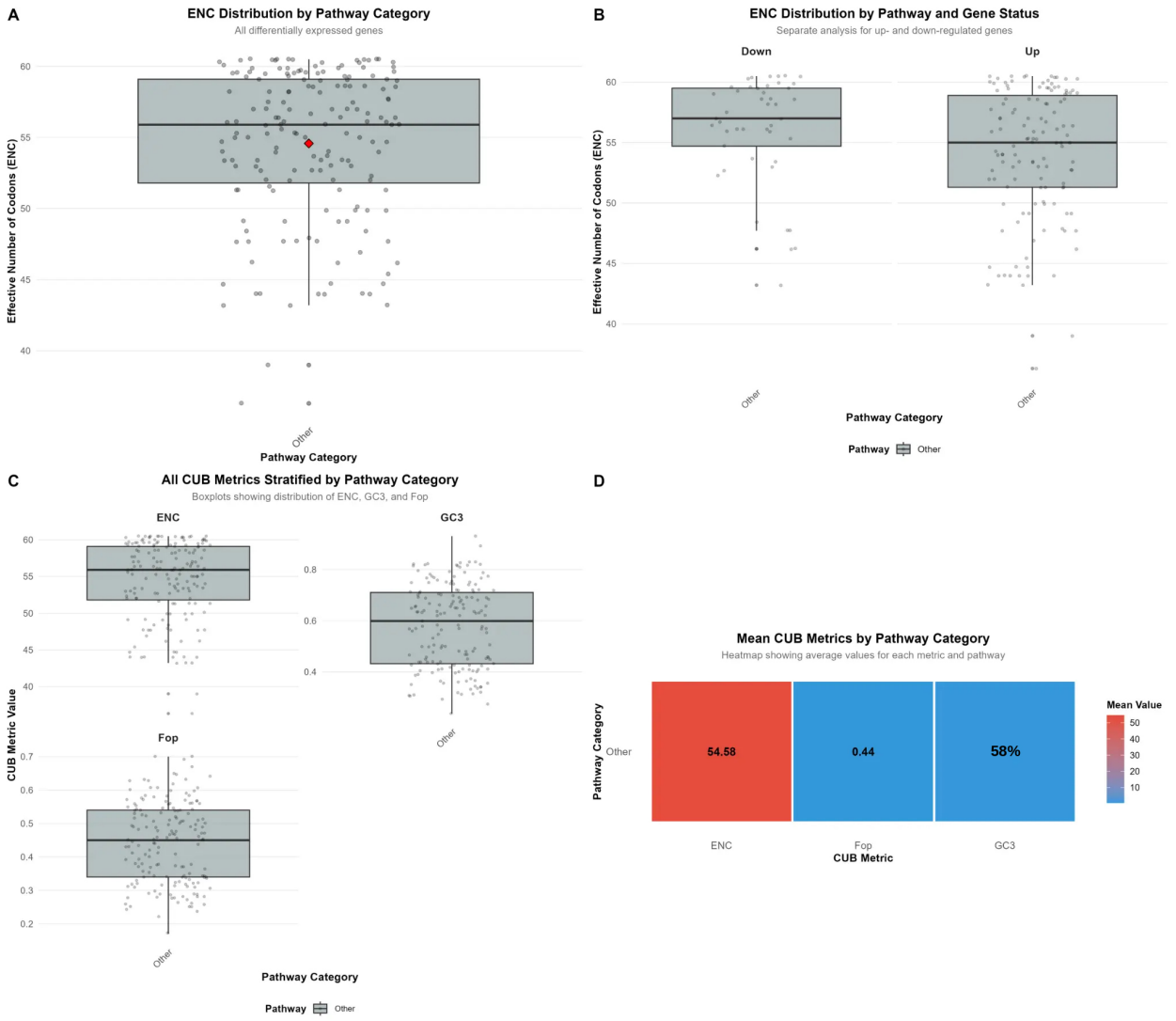


Fig. 15. Codon-usage-bias metrics among genes assigned to the "Other" pathway category. (A) Distribution of Effective Number of Codons (ENC) values for all differentially expressed genes. (B) ENC distributions shown separately for down-regulated and up-regulated genes. (C) Distribution of ENC, GC content at the third codon position (GC3), and frequency of optimal codons (Fop). (D) Heatmap of the mean ENC, Fop, and GC3 values. Individual points represent genes, and box plots summarize the distributions.

AT3 (36.60% and 42.20%, respectively). Furthermore, C3 > G3 > T3 > A3 was the order of the nucleotides at the third codon position in both gene groups, with C being the highest and A being the lowest. In the case of esophageal cancer genes, a similar finding was recorded [34]. The ENC values of up-regulated genes of oral cancer were in the range of 14.9 to 60.0, and down-regulated genes ranged from 27.1 to 60.0, both showing substantial variation in synonymous codon usage. The fact that the mean ENC score was greater than 35 in both gene groups indicated that the genes had a low codon usage bias. In the analysis of CUB for Flaviviridae virus, the ENC value varied from 48.75 to 57.83, i.e., more than 35. Our finding of low CUB was similar to their observation [35].

RSCU values of 59 codons were calculated for both gene groups to identify which codons were overrepresented and underrepresented. In our study, up-regulated genes showed 6 GC-ended overrepresented codons, 10 AT-ended underrepresented codons, 25 more frequently used codons, and 34 less frequently used codons. Down-regulated genes displayed 2 overrepresented codons, 10 underrepresented codons, 30 more frequently used codons, and 29 less frequently used codons. From the overrepresented codons, we observed that leucine and serine were the more preferred amino acids in the encoded proteins, similar to findings from studies on leukemia and pancreatic cancer genes [36].

Some other parameters were studied to identify the effect of mutational pressure and natural selection in both gene groups of oral cancer. Two major evolutionary forces that influence CUB are mutational pressure and natural selection. An ENC-GC3 relationship was plotted in a scattered diagram with the expected curve. As the dots were

not organized and a few dots were closer to the curve while the others were at discrete position, it indicated that the two forces shaped the CUB during evolution. Genes that are susceptible to esophageal cancer have shown similar patterns of codon use bias. The dominance of natural selection over mutation pressure in regulating codon usage patterns was validated by correspondence analysis, neutrality plot, and parity plot analysis [34]. By using the RSCU values of codons for both gene groups, correspondence analysis was performed with two axes (axis-1 and axis-2). The dots represented the trends for the CUB and also reported the role of both evolutionary forces in the formation of the CUB. A similar observation was depicted through parity plots for 6-fold, 4-fold, 2-fold synonymous codon families. The usage of the codons for particular amino acids was not equal indicating again the two forces might have differentially influenced the CUB. Unlike in the study of microsporidium genome, there were some other factors that influenced the codon bias [37]. Further, neutrality plot analysis was performed to examine the relationship between GC content at the three codon positions (GC12 vs. GC3). For up-regulated genes, the slope was 0.202, and for down-regulated genes, the regression coefficient was 0.228; because both slopes are much closer to 0 than to 1, selective constraints appear to have exerted a greater relative influence than mutational pressure on CUB in both gene groups, although these slopes provide only indirect, qualitative evidence and should not be interpreted as exact quantitative contributions. The CancerCoCoPUTs database shows comparable patterns of cancer-specific codon use in 32 main tumor types. Different codon and codon pair use signatures were found. by proving that codon use preferences correspond with clinical outcomes and differ significantly across various cancer types [38]. Correlation coefficient was estimated between the ENC values of genes and a few different parameters. Protein properties may affect the CUB and, in our results, there was a significant correlation between aromaticity and CUB in up-regulated genes. Aromatic amino acids (phe, trp and tyr) may have influenced the changes in codon usage. Nucleotide skew values (AT skew, GC skew, Purine skew, Amino skew, Pyrimidine skew and Keto skew) were examined for their relationship with CUB. There was a highly significant correlation in all the skew properties except AT skew in up-regulated genes, while down-regulated genes showed significant correlation in skew values, suggesting base skewness might also have contributed its impact in CUB of both gene groups. Thyroid cancer genes have been shown to have codon use bias due to protein and nucleotide skews. RSCU analysis indicated particular codons that are crucial to the development of cancer, and compositional analyses showed different groupings of cancer genes based on GC concentration, highlighting the intricate relationship between codon choices and nucleotide composition in genes linked to cancer [39]. Further, the MRI and P2 values of genes were examined. For up-regulated genes, the mean MRI value was positive (0.470) and the mean P2 value was 0.096 (less than 0.5), suggesting a greater role of mutational bias and a limited role of translational selection in shaping CUB. Translational selection and mutation both affected CUB for down-regulated genes, and there were strong relationships between ENC and MRI and P2 values. As seen by genes linked to cancer and major immunodeficiencies, this pattern of multiple evolutionary factors influencing codon use is consistent across many gene sets. Selection pressure, mutation pressure, and compositional limitations all influence codon usage patterns, according to neutrality analysis, parity analysis, and ENc-GC3 analysis [40].

By investigating the relationship between codon usage bias and gene expression, we analysed the correlation between log2 fold change and ENC for all differentially expressed genes. The results showed a significant positive correlation which indicates that higher expression levels are associated with stronger codon bias. The analysis of multiple CUB matrices revealed that ENC displayed weak positive correlation while Fop and GC3 showed negative correlation with expression. It suggests multidimensional codon optimization.

To determine whether pathway influences codon usage bias. We stratified genes by pathway membership and analysing their ENC distribution. The results indicate that pathway classification alone doesn't substantially determine codon usage bias. These results show that gene expression level is a more determinant of codon usage bias than pathway. This suggests that cancer cells have evolved mechanisms to optimize translation efficiency of highly expressed cancer-promoting genes through coordinated changes in multiple dimensions of codon usage bias.

Notably, important oncogenes and tumor suppressor genes linked to oral cancer were found among the differentially expressed genes examined. Translational selection may be involved in controlling the expression of cancer-causing genes in OSCC, as evidenced by the unique GC-rich codon preferences of genes involved in cell proliferation, apoptosis, and metabolic reprogramming.

5. Limitations

This study has several limitations. First, the analysis was based on a publicly available transcriptomic dataset and did not include experimental validation of gene expression, codon usage, or functional effects. Therefore, the observed associations should not be interpreted as evidence of causality. Second, the findings were derived from a single dataset and may not capture heterogeneity related to patient characteristics, tumor stage, molecular subtype, anatomical site, or exposure-related factors. Third, the analysis focused on differentially expressed genes; thus, the reported codon-usage patterns may not represent the complete oral cancer transcriptome. Finally, ENC–GC3, PR2, and neutrality-plot analyses provide indirect evidence regarding factors asso-

ciated with codon-usage bias. In particular, neutrality-plot slopes should be interpreted as suggesting the relative influence of selective constraints and mutational pressure, rather than as exact quantitative contributions. These limitations warrant validation in independent cohorts and through experimental studies. In addition, no independent external dataset was used to validate the differential-expression and codon-usage findings. Consequently, the generalizability and reproducibility of the results should be evaluated in independent oral cancer cohorts.

6. Conclusions

This extensive research of codon usage bias in oral cancer genes yielded numerous significant discoveries. First, the genes were GC-rich overall; base order was C > G > A > T in up-regulated genes and C > A > G > T in down-regulated genes. Second, the mean ENC values for up-regulated genes (47.21) and down-regulated genes (49.08) showed a modest overall codon usage bias. Third, GC-ended codons are preferred, whereas AT-ended codons are avoided. Fourth, neutrality-plot analysis indicated that selective constraints exerted a greater relative influence than mutational pressure on codon usage bias in both gene groups, although these slopes provide only indirect, qualitative evidence and should not be interpreted as exact quantitative contributions. Fifth, aromaticity and nucleotide skew features correlate significantly with codon usage bias. Sixth, gene expression level shows a statistically significant but weak correlation with codon usage bias ($r = 0.197$, $r^2 = 0.039$). These discoveries provide molecular insights into the genetic processes that drive oral cancer. This implies that codon optimization could affect the translation efficiency of cancer-promoting genes. Future research should validate these findings in larger cohorts and investigate the therapeutic potential of addressing codon usage bias in oral cancer treatment regimens.

Availability of Data and Materials

All data generated and analyzed during this study are included in this published article and its supplementary information files. The raw data for up-regulated and down-regulated genes, nucleotide composition analysis, codon usage metrics (ENC, RSCU), and statistical analyses are provided in the supplementary tables. All figures presenting the results of this study are included in the main manuscript. Any additional data or code used in this analysis are available from the corresponding author upon reasonable request.

Author Contributions

MA conceptualized and designed the study, wrote the original draft of the manuscript, and was responsible for the interpretation of the data. JJG and SC contributed to the analysis and interpretation of data, and reviewed and edited the manuscript. All authors read and approved the final manuscript. All authors have participated sufficiently in the work and agreed to be accountable for all aspects of the work.

Ethics Approval and Consent to Participate

Not applicable.

Acknowledgment

We are thankful to Assam University, Silchar, Assam and Saurashtra University, Rajkot, Gujarat, India for providing the necessary lab facilities for this work. The authors would like to thank Mr. Abdelrahman Atef (School of Artificial Intelligence and Data Management, Department of Bioinformatics, Badr University in Assiut (BUA), Assiut, Egypt) for his valuable contribution to the gene expression analysis and his assistance in the editing of the manuscript.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.31083/FBE49059.

References

[1] Silverman S. Oral cancer. PMPH-USA: Shelton, CT. 2003.

[2] Kademani D. Oral cancer. Mayo Clinic Proceedings. 2007; 82: 878–887. https://doi.org/10.4065/82.7.878

[3] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians. 2018; 68: 394–424. https://doi.org/10.3322/caac.21492

[4] Shah JP, Gil Z. Current concepts in management of oral cancer–surgery. Oral Oncology. 2009; 45: 394–401. https://doi.org/10.1016/j.oraloncology.2008.05.017

[5] Lousada-Fernandez F, Rapado-Gonzalez O, Lopez-Cedrun JL, Lopez-Lopez R, Muinelo-Romay L, Suarez-Cunqueiro MM. Liquid Biopsy in Oral Cancer. International Journal of Molecular Sciences. 2018; 19: 1704. https://doi.org/10.3390/ijms19061704

[6] Rhodus NL, Kerr AR, Patel K. Oral cancer: leukoplakia, premalignancy, and squamous cell carcinoma. Dental Clinics of North America. 2014; 58: 315–340. https://doi.org/10.1016/j.cden.2013.12.004

[7] Dhanuthai K, Rojanawatsirivej S, Thosaporn W, Kintarak S, Subarnbhesaj A, Darling M, et al. Oral cancer: A multicenter study. Medicina Oral, Patologia Oral Y Cirugia Bucal. 2018; 23: e23–e29. https://doi.org/10.4317/medoral.21999

[8] Shrestha AD, Vedsted P, Kallestrup P, Neupane D. Prevalence and incidence of oral cancer in low- and middle-income countries: A scoping review. European Journal of Cancer Care. 2020; 29: e13207. https://doi.org/10.1111/ecc.13207

[9] Gholizadeh P, Eslami H, Yousefi M, Asgharzadeh M, Aghazadeh M, Kafil HS. Role of oral microbiome on oral cancers,

a review. Biomedicine & Pharmacotherapy = Biomedecine & Pharmacotherapie. 2016; 84: 552–558. https://doi.org/10.1016/j.biopha.2016.09.082

[10] Tarapan S, Matangkasombut O, Trachootham D, Sattabanasuk V, Talungchit S, Paemuang W, et al. Oral Candida colonization in xerostomic postradiotherapy head and neck cancer patients. Oral Diseases. 2019; 25: 1798–1808. https://doi.org/10.1111/odi.13151

[11] D'souza S, Addepalli V. Preventive measures in oral cancer: An overview. Biomedicine & Pharmacotherapy = Biomedecine & Pharmacotherapie. 2018; 107: 72–80. https://doi.org/10.1016/j.biopha.2018.07.114

[12] Peng CH, Liao CT, Peng SC, Chen YJ, Cheng AJ, Juang JL, et al. A novel molecular signature identified by systems genetics approach predicts prognosis in oral squamous cell carcinoma. PLoS One. 2011; 6: e23452. https://doi.org/10.1371/journal.pone.0023452

[13] Wright F. The 'effective number of codons' used in a gene. Gene. 1990; 87: 23–29. https://doi.org/10.1016/0378-1119(90)90491-9

[14] East JL. Nucleotide composition of ribonucleic acid by spectral analysis and thin-layer chromatography. Analytical Biochemistry. 1968; 24: 409–418. https://doi.org/10.1016/0003-2697(68)90146-2

[15] Suzuki H, Brown CJ, Forney LJ, Top EM. Comparison of correspondence analysis methods for synonymous codon usage in bacteria. DNA Research : an International Journal for Rapid Publication of Reports on Genes and Genomes. 2008; 15: 357–365. https://doi.org/10.1093/dnares/dsn028

[16] Sueoka N. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. Journal of Molecular Evolution. 1995; 40: 318–325. https://doi.org/10.1007/BF00163236

[17] He B, Dong H, Jiang C, Cao F, Tao S, Xu LA. Analysis of codon usage patterns in Ginkgo biloba reveals codon usage tendency from A/U-ending to G/C-ending. Scientific Reports. 2016; 6: 35927. https://doi.org/10.1038/srep35927

[18] Sueoka N. Near homogeneity of PR2-bias fingerprints in the human genome and their implications in phylogenetic analyses. Journal of Molecular Evolution. 2001; 53: 469–476. https://doi.org/10.1007/s002390010237

[19] Gatherer D, McEwan NR. Small regions of preferential codon usage and their effect on overall codon bias–the case of the plp gene. Biochemistry and Molecular Biology International. 1997; 43: 107–114. https://doi.org/10.1080/15216549700203871

[20] Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, et al. Protein Identification and Analysis Tools on the ExPASy Server. In Walker JM (ed.) The Proteomics Protocols Handbook. Springer Protocols Handbooks. Humana Press. 2005; 571–607. https://doi.org/10.1385/1-59259-890-0:571

[21] Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. Journal of Molecular Biology. 1982; 157: 105–132. https://doi.org/10.1016/0022-2836(82)90515-0

[22] Xia X. Genomic Features: Content Sensors, Nucleotide Skew Plot, Strand Asymmetry, and DNA Methylation. In Bioinformatics and the Cell: Modern Computational Approaches in Genomics, Proteomics and Transcriptomics (pp. 255–268). Springer: New York, NY. 2018. https://doi.org/10.1007/978-3-319-90684-3_11

[23] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics (Oxford, England). 2003; 4: 249–264. https://doi.org/10.1093/biostatistics/4.2.249

[24] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015; 43: e47. https://doi.org/10.1093/nar/gkv007

[25] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological). 1995; 57: 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[26] Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Annals of Statistics. 2001; 29: 1165–1188. https://doi.org/10.1214/aos/1013699998

[27] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nature Genetics. 2000; 25: 25–29. https://doi.org/10.1038/75556

[28] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research. 2000; 28: 27–30. https://doi.org/10.1093/nar/28.1.27

[29] Huang DW, Sherman BT, Zheng X, Yang J, Imamichi T, Stephens R, et al. Extracting biological meaning from large gene lists with DAVID. Current Protocols in Bioinformatics. 2009; 27: 13.11.1–13.11.13. https://doi.org/10.1002/0471250953.bi1311s27

[30] Benisty H, Weber M, Hernandez-Alias X, Schaefer MH, Serrano L. Mutation bias within oncogene families is related to proliferation-specific codon usage. Proceedings of the National Academy of Sciences of the United States of America. 2020; 117: 30848–30856. https://doi.org/10.1073/pnas.2016119117

[31] Hashibe M, Mathew B, Kuruvilla B, Thomas G, Sankaranarayanan R, Parkin DM, et al. Chewing tobacco, alcohol, and the risk of erythroplakia. Cancer Epidemiology, Biomarkers & Prevention : a Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology. 2000; 9: 639–645.

[32] Ali J, Sabiha B, Jan HU, Haider SA, Khan AA, Ali SS. Genetic etiology of oral cancer. Oral Oncology. 2017; 70: 23–28. https://doi.org/10.1016/j.oraloncology.2017.05.004

[33] Zhang L, Liu Y, Zheng HJ, Zhang CP. The Oral Microbiota May Have Influence on Oral Cancer. Frontiers in Cellular and Infection Microbiology. 2020; 9: 476. https://doi.org/10.3389/fcimb.2019.00476

[34] Bordoloi H, Nirmala SR. Codon usage bias analysis of genes linked with esophagus cancer. Bioinformation. 2021; 17: 731–740. https://doi.org/10.6026/97320630017731

[35] Yao H, Chen M, Tang Z, et al. Analysis of Synonymous Codon Usage Bias in Flaviviridae Virus. BioMed Research International. 2019; 2019: 5857285. https://doi.org/10.1155/2019

[36] Parvathy ST, Udayasuriyan V, Bhadana V. Codon usage bias. Molecular Biology Reports. 2022; 49: 539–565. https://doi.org/10.1007/s11033-021-06749-4

[37] Xiang H, Zhang R, Butler RR, 3rd, Liu T, Zhang L, Pombert JF, et al. Comparative Analysis of Codon Usage Bias Patterns in Microsporidian Genomes. PloS One. 2015; 10: e0129223. https://doi.org/10.1371/journal.pone.0129223

[38] Meyer D, Kames J, Bar H, Komar AA, Alexaki A, Ibla J, et al. Distinct signatures of codon and codon pair usage in 32 primary tumor types in the novel database CancerCoCoPUTs for cancer-specific codon usage. Genome Medicine. 2021; 13: 122. https://doi.org/10.1186/s13073-021-00935-6

[39] Pepe D, DE Keersmaecker K. Codon bias analyses on thyroid carcinoma genes. Minerva Endocrinologica. 2020; 45: 295–305. https://doi.org/10.23736/S0391-1977.20.03252-6

[40] Khandia R, Alqahtani T, Alqahtani AM. Genes Common in Primary Immunodeficiencies and Cancer Display Overrepresentation of Codon CTG and Dominant Role of Selection Pressure in Shaping Codon Usage. Biomedicines. 2021; 9: 1001. https://doi.org/10.3390/biomedicines9081001