Discover Oncology

Discover

Research

Transcriptomic analysis predicts the risk of progression of premalignant lesions in human tongue

Tuo Zhang[3] · David Kutler[2] · Theresa Scognamiglio[4] · Lorraine J. Gudas[1] · Xiao-Han Tang[1]

Received: 2 December 2022 / Accepted: 12 February 2023
Published online: 23 February 2023

Abstract

The 5-year survival rate for patients with oral squamous cell carcinomas (SCC), including tongue SCC, has not significantly improved over the last several decades. Oral potentially malignant disorders (OPMD), including oral dysplasias, are oral epithelial disorders that can develop into oral SCCs. To identify molecular characteristics that might predict conversion of OPMDs to SCCs and guide treatment plans, we performed global transcriptomic analysis of human tongue OPMD (n = 9) and tongue SCC (n = 11) samples with paired normal margin tissue from patients treated at Weill Cornell Medicine. Compared to margin tissue, SCCs showed more transcript changes than OPMDs. OPMDs and SCCs shared some altered transcripts, but these changes were generally greater in SCCs than OPMDs. Both OPMDs and SCCs showed altered signaling pathways related to cell migration, basement membrane disruption, and metastasis. We suggest that OPMDs are on the path toward malignant transformation. Based on patterns of gene expression, both OPMD and tongue SCC samples can be categorized into subclasses (mesenchymal, classical, basal, and atypical) similar to those seen in human head and neck SCC (HNSCC). These subclasses of OPMDs have the potential to be used to stratify patient prognoses and therapeutic options for tongue OPMDs. Lastly, we identified a gene set (ELF5; RPTN; IGSF10; CRMP1; HTR3A) whose transcript changes have the power to classify OPMDs and SCCs and developed a Firth logistic regression model using the changes in these transcripts relative to paired normal tissue to validate pathological diagnosis and potentially predict the likelihood of an OPMD developing into SCC, as data sets become available.

Keywords Tongue squamous cell carcinomas · Oral potentially malignant disorders · RNA-seq · Firth logistic regression

Tuo Zhang and David Kutler have contributed to this work equally to this work

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12672-023-00629-y.

✉ Lorraine J. Gudas, ljgudas@med.cornell.edu; ✉ Xiao-Han Tang, xit2001@med.cornell.edu | [1]Department of Pharmacology, Weill Cornell Medical College of Cornell University, 1300 York Avenue, New York, NY 10065, USA. [2]Division of Head and Neck Surgery in the Department of Otolaryngology at New York Presbyterian Hospital/Weill Cornell Medical Center, New York, NY 10065, USA. [3]Genomics Resources Core Facility, Weill Cornell Medical College of Cornell University, New York, NY 10065, USA. [4]Division of Anatomic Pathology, New York Presbyterian Hospital, Department of Pathology and Laboratory Medicine, Weill Cornell Medical College of Cornell University, New York, NY 10065, USA.

1 Introduction

The American Cancer Society estimates 34,730 new oral cancer diagnoses in 2022, including 17,860 new tongue cancer cases within the United States [1]. More than 90% of oral cancer cases are squamous cell carcinomas (SCCs), a type of head and neck SCC (HNSCC) [2, 3]. The 5-year survival rate for patients with oral SCC (OSCC) has not significantly improved, despite various new treatment options, in the last several decades [4]. Current therapies for the majority of oral SCC patients include surgery, cytotoxic chemotherapy, and radiation therapy [2]. Even after treatment about half of the patients relapse, with a subsequent median survival time to 8–10 months [2, 5]. Oral potentially malignant disorders (OPMDs) are a group of oral epithelial disorders that has an increased risk of developing into OSCCs as compared to clinically normal oral mucosa [6]. OPMDs include leukoplakia, erythroplakia, erythroleukoplakia, oral lichen planus, oral submucous fibrosis, and oral dysplasia [7, 8]. An objective system that would predict the tendency of OPMD to develop into OSCC could be helpful both to manage clinical care and to understand the progression of OPMDs to OSCCs.

The current clinical staging system of oral SCCs that determines the basic characteristics and prognosis for patients is based on assessments of primary tumors (Tx, T0, T1, T2, T3, T4a, and T4b), regional lymph nodal metastases (Nx, N0, N1, N2a, N2b, N2c, and N3), and distant metastasis (Mx, M0, M1) [9]. Although this system is useful for determining clinical treatments for OSCC, this arbitrary staging system is not informative in terms of predicting SCC risks for individual patients with OPMDs [10–12]. Therefore, in addition to histopathological assessment for OPMDs, it is important to identify molecular markers that can predict the risks of OPMDs advancing to SCCs and guide treatment plans. These markers could also be very useful in human populations with a high risk of oral cancer, such as people who smoke and/or drink heavily [13].

There have been studies to identify markers to distinguish OPMD and malignant oral lesions [10, 14, 15], however, these studies did not assess global transcript levels during multi-step OSCC carcinogenesis. Researchers have used genome-wide transcriptomics analyses to assess cancer risks, to distinguish human oral leukoplakia subtypes (low and high grade dysplasia) [16–18], to assess OSCCs [19–25], and to predict the clinical outcome of OPMD [26]. Because the patient samples analyzed in these studies [16–26] were obtained from a variety of sites in the human oral cavity, including the tongue, palate, lower/upper gingiva, floor of mouth, buccal mucosa, and sinus, the results may not reflect the genome-wide RNA profiles at specific sites, such as the tongue. Additionally, most of these earlier studies [16–21] used cDNA microarray technology, which has now been replaced by RNA-seq technology because it has a lower background noise, higher specificity, greater dynamic range for quantifying gene expression levels, and the ability to distinguish different transcript isoforms [27]. Although there are studies using RNA-seq [28–30] to characterize human HNSCC, including oral SCC, and identify different molecular events at pathological stages, there are not many studies focusing on tongue OPMD and SCC, a subtype of head and neck cancer. One recent study [22] using RNA-seq technology on brush biopsies of human oral cancers does not have the capacity to show the changes in genome-wide mRNA levels in human OSCC, especially in invasive tumors. Other studies that used RNA-seq focused on miRNA and long non-coding RNAs in human oral cancers [23–25].

Because histopathological characteristics of OPMD do not always accurately predict the clinical outcomes of these lesions, we sought to develop a signature gene set in which transcript level changes in an individual patient with tongue OPMDs could be used to predict the risk of developing SCC. Comparative analysis of samples between OPMDs and SCC is a major challenge because of the variabilities among individual patient transcript changes in these lesions relative to normal tissue samples from the same patient [22]. This variability could be a major reason that there have been few successes in using RNA-based diagnosis or prognosis for oral cancer since many prior studies [10, 14, 19, 21, 30] grouped all normal and lesion samples separately for comparison.

In this report we conducted genome-wide RNA-seq analyses of human tongue OPMDs and SCCs from individual patients compared to the normal healthy tongue epithelia in the same patient. We then developed a gene set that informs the risk of lesions advancing to SCCs. The results from this comparison provide novel molecular markers that may predict the risk of OPMDs becoming OSCCs and have the potential to improve early diagnosis and treatment decisions for OSCCs.

2 Materials and Methods

2.1 Human OPMD and SCC samples

Human patient tongue lesion samples and their corresponding margin samples (far away from the lesions) were surgically resected (See Supplementary Information).

2.2 Pathological diagnosis

Histopathologic examination was performed for all lesions (See Supplementary Information).

2.3 RNA-seq analysis of transcriptome

We prepared total RNA from the human margin, OPMD, and SCC samples. Subsequent steps for RNA-seq were carried out at the Genomics Resources Core Facility of WCMC (See Supplementary Information).

2.4 Pathway and gene ontology analysis

We performed pathway and gene ontology analysis using Enrichr (See Supplementary Information).

2.5 SCC subclass correlation studies

We carried out Pearson and Spearman correlation studies using GraphPad Prism software (See Supplementary Information).

2.6 Firth logistic regression analysis

We built a logistic regression model to classify OPMD and SCC samples based on gene expression changes (See Supplementary Information).

2.7 Data and code availability

The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database. A web application to predict the SCC risk of a tongue OPMD is available at https://freshtuo.shinyapps.io/sccpred/ (See Supporting Information).

3 Results

3.1 Gene expression profiles of human tongue lesions at different pathological stages indicate transcripts involved in human oral SCC progression

We performed RNA-seq analysis on the OPMD and SCC samples and the corresponding margin (normal) samples from the same patients. Our analysis revealed that 1439 transcript levels (differentially expressed genes (DEGs)) were altered significantly in the tongue OPMD samples (n = 9) compared to the corresponding margin samples (n = 9), including increases in 729 and decreases in 710 transcripts (adjusted p-value < 0.05) (Fig. 1A, C, Table S1). We identified a total of 6281 transcripts, including 3061 transcripts increased and 3220 transcripts decreased significantly (adjusted p-value < 0.05), in tongue SCC samples (n = 11) compared to the corresponding margin samples (n = 11) (Fig. 1A, C, Table S1). Fold changes in 161 transcripts increased and 76 transcripts decreased were |Log2|> 1, respectively, in OPMD samples vs margin (normal). In SCC samples, 1470 increased and 1416 decreased transcripts showed fold changes of |Log2|> 1, respectively, compared to margin (Fig. 1D). Analysis of transcriptome profiling similarities using

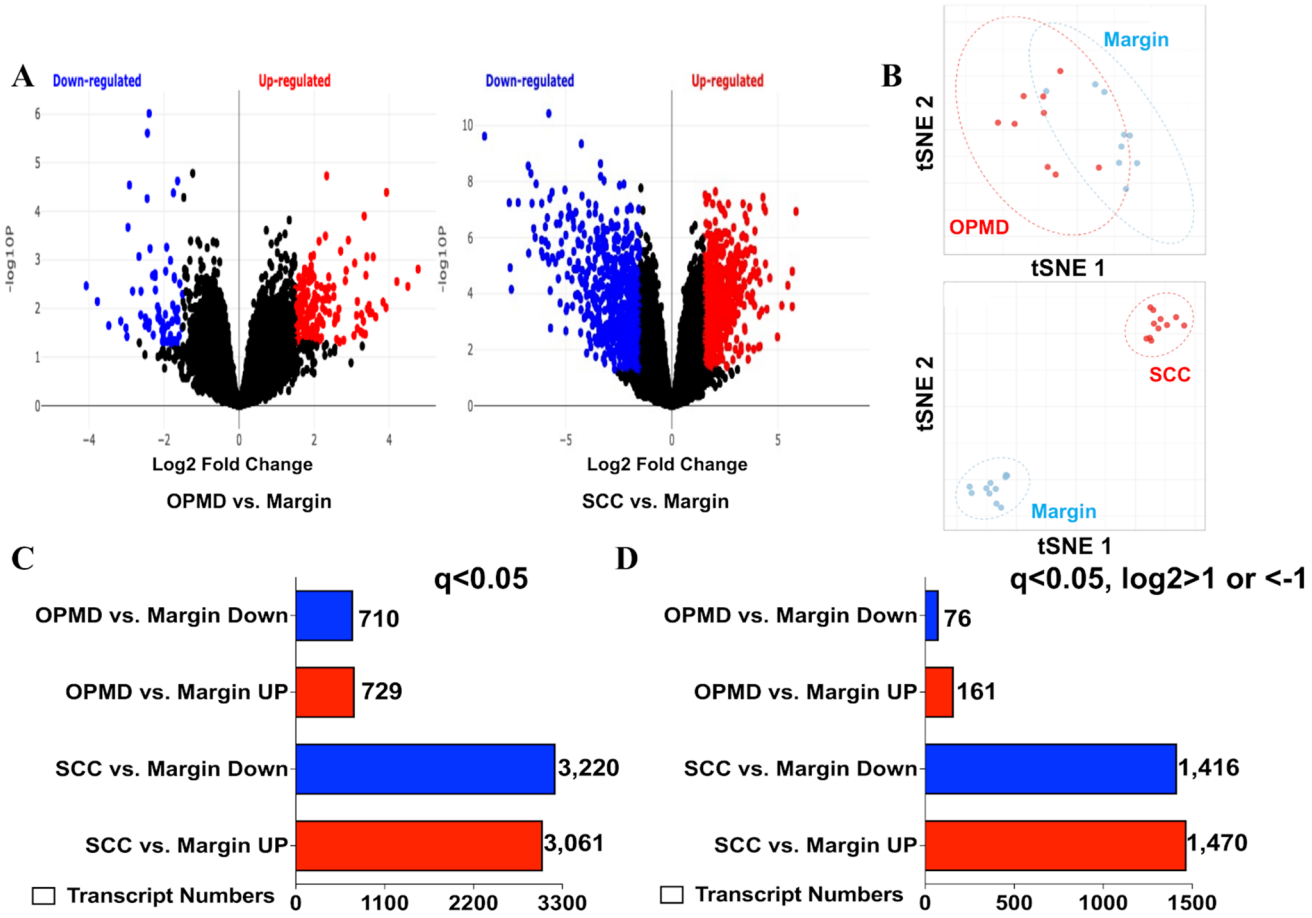


Fig. 1 Comparison of transcriptomes between human tongue oral potentially malignant disorders (OPMD), squamous cell carcinoma (SCC), and margins. **A** Volcano plot showing total numbers of transcripts with statistically significant increases or decreases ($q < 0.05$) in the OPMD ($n = 9$) and SCC ($n = 12$) compared with the margin group. **B** tSNE (t-distributed stochastic neighbor embedding plot showing that OPMD is an intermediate state between margin (healthy state) and SCC (cancer state). **C** Numbers of transcripts significantly ($q < 0.05$) altered OPMD and SCC, vs. margin. **D** Numbers of transcripts significantly ($q < 0.05$) altered OPMD and SCC, vs. margin, with log2 changes > 1 or < − 1

a tSNE (t-distributed stochastic neighbor embedding) plot (Fig. 1B) shows that the margin and SCC groups were well separated; samples from each group formed a unique cluster, and in the OPMD group one-third of the samples ($n = 3$) were mixed with the margin group and two-thirds of the samples ($n = 6$) were close to the margin group (Fig. 1B). These data suggest that OPMD is an intermediate state between margin (healthy state) and SCC (cancer state).

We analyzed the relationship between the transcripts changed in SCC/margin and that in OPMD/margin with fold changes of |Log2|> 1. A Venn diagram (Figure S1A) shows that there were 146 transcripts that overlapped between SCC/margin and OPMD/margin (Table S2). These transcript overlaps were much more significant than expected by random chance, with a p-value = 3.245 e-73 using Fisher's exact test (Figure S1B). The log2 fold changes of these 146 transcripts are shown (Figure S1C). Although 141 of these 146 transcripts showed same change directions in SCC/margin and OPMD/margin, with the average fold changes in SCC/margin greater than that in OPMD/margin (Figure S1C), the transcript level changes of these 146 genes in individual patients exhibited variability (Figure S1D). For the following studies below, we only used transcripts with fold changes of |Log2|> 1.

To ascertain whether our patients' lesions captured the characteristics of human oral SCC, we performed disease signature pathway analysis using "Disease Perturbations from GEO up/down" [31], a platform designed to probe a variety of gene and disease associations. This disease association study verified that both OPMD and SCC samples closely resembled human oral SCC molecular features (Table S3). SCC samples exhibited more human oral SCC transcript markers than OPMD samples, based on the numbers of transcripts, q values, odds ratios, and combined scores (Fig. 2A), indicating that SCC is a more advanced pathologic state than OPMD.

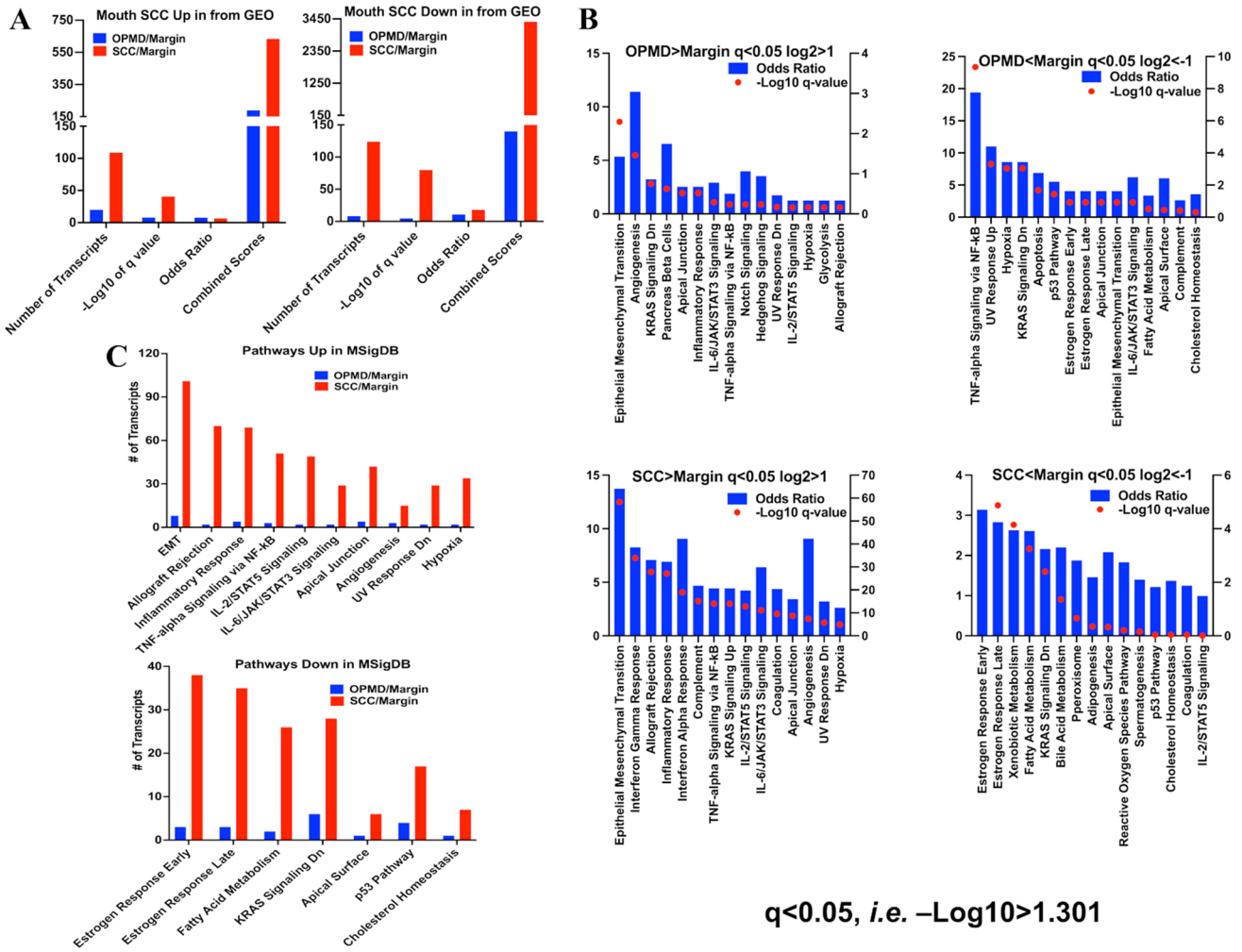


Fig. 2 Pathway analysis transcriptomic changes in human tongue oral potentially malignant disorders (OPMD) and squamous cell carcinoma (SCC). **A** The association between significantly altered transcripts ($q < 0.05$, $\log2 > 1$ or < -1) in OPMD and SCC and human oral SCC using Disease Perturbations from GEO database. **B** Pathway enrichment analysis of significantly altered transcripts ($q < 0.05$, $\log2 > 1$ or < -1) in OPMD and SCC using Molecular Signatures Database (MSigDB). Left Y axis, odds ratio; right Y axis, -Log10 q-value. **(C)** The numbers of altered transcripts ($q < 0.05$, $\log2 > 1$ or < -1) in OPMD and SCC enriched in the top 15 pathways derived from MSigDB analysis

3.2 Transcripts altered in pathways in human tongue lesions at different stages are associated with pathological stages of the tongue lesions

We conducted a pathway enrichment analysis of the transcripts altered in the OPMD and SCC samples using the Enrichr web tool, based on the MsigDB (Molecular Signatures Database) [32]. The top 15 pathways enriched in OPMD and SCC samples are shown (Fig. 2B). In OPMD samples, only pathways "Epithelial Mesenchymal Transition" (EMT) and "Angiogenesis", enriched with increased transcript levels, were statistically significant ($q < 0.05$), while in SCC samples, all top 15 pathways enriched with elevated transcript levels were statistically significant, including "Epithelial Mesenchymal Transition" and "Angiogenesis" (Fig. 2B, Table S4). The top 6 pathways enriched with decreased transcript levels in both OPMD and SCC samples were statistically significant (Fig. 2B, Table S4). We then compared the numbers of increased and decreased transcripts in the top overlapped pathways between OPMD and SCC groups (Fig. 2C). Within the 10 and 7 overlapped pathways enriched with increased and decreased transcripts levels, respectively, SCC samples showed changes in more transcripts than OPMD samples. These data, along with the greater average fold changes in SCC/margin than in OPMD/margin (Figure S1C), indicate that SCC shows more intense perturbations on these

Fig. 3 Correlation studies of the variabilities among the global transcriptomic changes ($q < 0.05$, $\log2 > 1$ or < -1) in individual OPMD or SCC, compared to the margin from the same patient. **A** Spearman and **B** Pearson correlation analyses of the variabilities and association among the OPMD and SCC samples. **C** Human head and neck squamous cell carcinoma (SCC) subclass categorization of the OPMD and SCC samples, based on the global transcriptomic changes ($q < 0.05$, $\log2 > 1$ or < -1) in individual OPMD or SCC. D (disorders), OPMD; M, margin; T (tumor), SCC; D/M, OPMD/margin; T/M, SCC/margin

pathways than OPMD. Importantly, the "Epithelial Mesenchymal Transition" and "Angiogenesis" pathways (Fig. 2B, C) suggest that cancer cell migration and metastasis occur at the pre-SCC stage of human tongue carcinogenesis.

3.3 Transcripts altered in pathways in human tongue lesions at different stages are associated with pathological stages of the tongue lesions

By using the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway analysis program we discovered that the top 15 pathways significantly enriched with transcripts increased in the SCC group included the cancer-related pathways "Cytokine-cytokine receptor interaction" and "ECM-receptor interaction," while there were no cancer-related pathways with increased transcripts significantly ($q < 0.05$) enriched in the OPMD group (Figure S2, Table S5). In the SCC group with transcripts decreased "Metabolism of xenobiotics by cytochrome P450" and "Fatty acid degradation" were among the top 8 pathways significantly ($q < 0.05$) enriched. The top pathways significantly ($q < 0.05$) enriched with transcripts decreased in the OPMD group, such as "Estrogen signaling pathway" and "Lipid and atherosclerosis," showed no directly cancer-related pathways (Figure S2, Table S5). Thus, our KEGG pathway analysis shows that cancer-related pathways are more highly activated in the SCC than in OPMD samples.

Gene set enrichment analysis (GSEA) using the Gene Ontology (GO) Biological Process Database revealed that the top 15 pathways with transcripts increased in the SCC group included "Extracellular matrix organization (GO:0030198)" and "Inflammatory response (GO:0006954)", and the top pathways enriched with transcripts in the OPMD group showed "Skin development (GO:0043588)" and "Extracellular matrix organization (GO:0030198)" (Figure S3, Table S5). Additionally, the top pathways with transcripts decreased in the SCC and OPMD groups included "Epidermis development (GO:0008544)" and "Cellular heat acclimation (GO:0070370)" (Figure S3, Table S5). The KEGG and GO pathways enriched with transcripts increased in the OPMD and SCC samples suggest: (1) an increase in cell motility potential in SCC and OPMD [9], with greater potential in SCC than OPMD, in line with the data from the MSigDB analysis (Fig. 2B); and (2) perturbation of the immune environment in SCC, *e.g.* increases in interferon responses and cytokine signaling (Fig. 2B).

3.4 The changes in gene expression profiles in individual patients categorize these OPMD and SCC lesions into distinct subclasses

A global heatmap analysis on the RNA-seq data, with each margin sample followed by the lesion sample from the same patient (OPMD or SCC) (Figure S4) demonstrated that there were differences in the basal gene expression in the margins in different patients. It was also clear that there were differences between the margin tissue and the OPMD or SCC samples. Furthermore, the changes in the average levels of individual transcripts in the tongue OPMD and SCC groups vs. the margin groups did not always reflect the changes in individual patients (Table S2, Figure S1C and D, as an example). These data indicate divergent gene expression in human tongue OPMD and SCC samples as well as variations in gene expression in normal tissues. Therefore, we next compared the changes in the transcriptomes in individual tongue lesion vs. margin tissue from the same patient. This comparison could mitigate the possible errors caused by comparing the group average of all margin samples with that of all lesion samples and reveal more consistent changes between OPMD and OSCC. We performed correlation studies (Spearman and Pearson) using the log2 fold changes of transcripts between tongue lesions (OPMD or SCC) vs. their margins in the same patients to examine the similarities in the entire transcriptome among patients. In the OPMD group, samples #5 and #2 were quite similar, and sample #9 showed no correlations with all other samples except a small similarity to #7 (Fig. 3A), shown by both Pearson and Spearman correlation studies. In the SCC group, samples #1, and #2; #3, #4 and #5; and #8 and #9 showed higher similarities ($r > 0.5$) than other sample comparisons using both Spearman and Pearson correlations (Fig. 3A). These data confirm variability by varied transcript level changes among human tongue lesions, including OPMD and SCC.

Previous researchers have discovered four distinct, clinically relevant subclasses of human head and neck SCC based on the gene expression patterns: [1] mesenchymal, [2], classical, [3] basal, and [4] atypical [33, 34]. We next examined whether our individual human tongue OPMD and SCC samples could be categorized into these subclasses, based on the

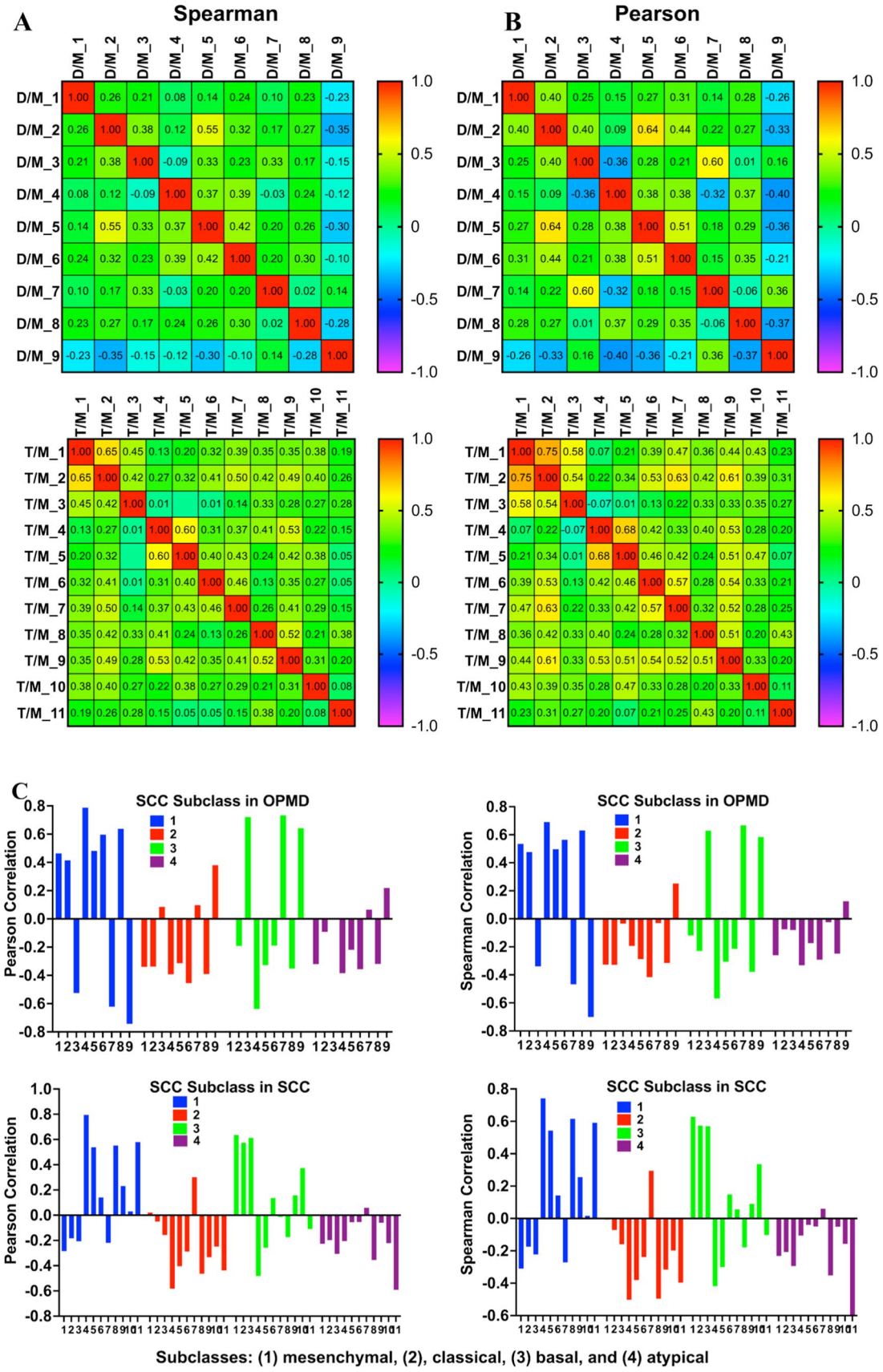

A
Spearman
B
Pearson
D/M_1 D/M_2 D/M_3 D/M_4 D/M_5 D/M_6 D/M_7 D/M_8 D/M_9
T/M_1 T/M_2 T/M_3 T/M_4 T/M_5 T/M_6 T/M_7 T/M_8 T/M_9 T/M_10 T/M_11
1.0 0.5 0 -0.5 -1.0
C
SCC Subclass in OPMD
SCC Subclass in SCC
Pearson Correlation
Spearman Correlation
1 2 3 4
Subclasses: (1) mesenchymal, (2), classical, (3) basal, and (4) atypical

Fig. 4 Transcript level changes of ELF5, RPTN, IGSF10, HTR3A, and CRMP1 in individual OPMD and SCC samples and correlation studies among them. **A** Transcript level changes of these transcripts in individual patients. **B** The probabilities of SCC of individual tongue lesion samples derived from Leave-One-Out Test". **C** Pearson and Spearman correlation studies among ELF5, RPTN, IGSF10, CRMP1, and HTR3A transcripts within the OPMD and SCC groups

transcriptomic changes in an individual tongue lesion vs. margin tissue from the same patient. We performed correlation study using the transcriptomic changes (|Log2|> 1) in each patient (Table S6) and the mRNA markers for each subclass of human head and neck SCC downloaded from the Broad Firehose website (https://gdac.broadinstitute.org/). This study (Fig. 3C) shows that in the OPMD group samples #1, 2, 4, 5, 6, and 8 correlated with subclass 1 (mesenchymal); #3 and 7 correlated with subclass 3, (basal); and #9 correlated with both subclasses 2 (classical) and 3 (basal), indicating that #9 showed features of these two subclasses. In the SCC group samples #1, 2, 3, and 10 correlated with subclass 3 (basal); samples #4, 5, 8, and 11 correlated with subclass 1 (mesenchymal); sample #7 correlated with subclass 2 (classical); and samples #6, 8 did not show a strong correlation with any subclasses (Fig. 3C). The correlations between samples in the OPMD and SCC groups, based on the mRNA markers for each human HNSCC subclass, are shown (Figure S5); this extends data in Fig. 3C. Notably, most of these lesions belonged to the mesenchymal and basal subclasses, while for both OPMD and SCC groups the classical and atypical subclasses were rare.

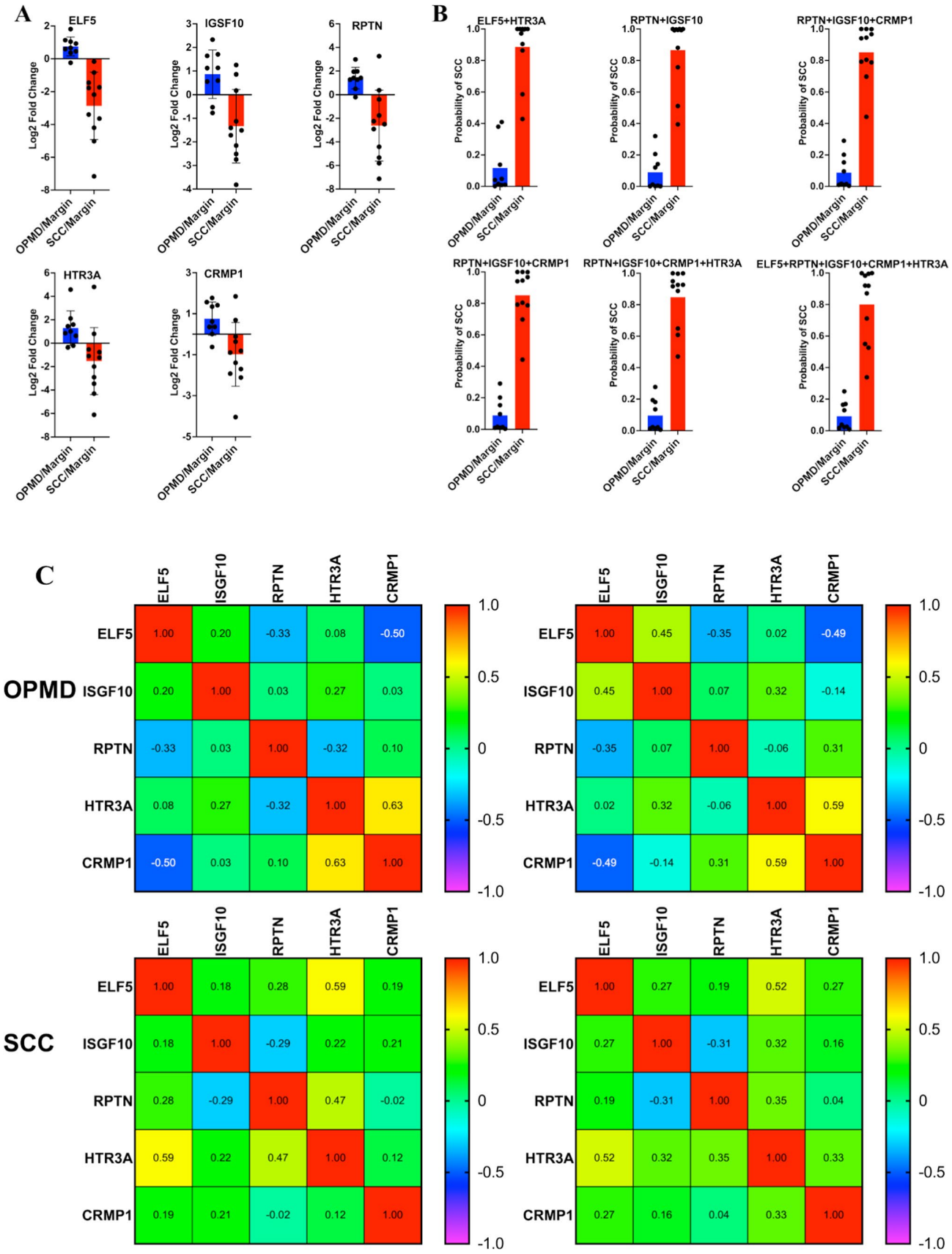
3.5 Firth logistic regression analysis shows the transcript changes in individual patients that differ between OPMD and SCC

Because the majority of OPMD do not progress to SCC [6, 26], we develop an machine learning approach to classify OPMDs and SCCs using a signature gene set whose transcript changes in a patient's tongue lesion, compared to the same patient's margin tissue, and predict the probability of a lesion to progress to SCC. We implemented our analysis using Firth logistical regression to alleviate possible overfitting caused by small sample size [35]. We performed the analysis on the 6,693 transcripts selected by taking a union of the differential expressed genes (adjusted p-value < 0.05) derived from the two groups (OPMD and SCC) and ranking the transcripts by their area under curve (AUC) scores of the Receiver Operator Characteristic (ROC) curve, a graphical plot illustrating the diagnostic ability of a binary classifier system as its discrimination threshold is varied (Table S7). The AUC values of 848 and 149 transcripts were above 0.8 and 0.9, respectively, indicating sufficient power to distinguish OPMD from SCC (Table S7). Then we applied four filters (described in MATERIALS AND METHODS, Supplementary Information) to refine candidate transcript selections.

We screened five candidate transcripts as described in the Materials and Methods: ELF5 (E74 Like ETS Transcription Factor 5), IGSF10 (Immunoglobulin Superfamily Member 10), CRMP1 (Collapsin Response Mediator Protein 1), RPTN (Repetin), and HTR3A (5-Hydroxytryptamine Receptor 3A) (Table S6). ELF5 and IGSF10 mRNA levels are significantly decreased in human head and neck cancers [36, 37]. CRMP1 inhibits prostate cancer cell migration and metastasis by suppressing EMT [38]. Although there are no reports of HTR3A in human tongue SCC, elevated HTR3A expression correlates with increased human lung adenocarcinoma cell proliferation and is associated with aggressive histopathology [39]. RPTN is associated with epithelial differentiation [40]. The log2 changes of these transcripts in each patient's tongue lesion vs. the margin tissue are shown (Fig. 4A). The average changes in these transcript levels in tongue SCC were in line with previous studies described above. Importantly, we show that the average changes of these transcripts in OPMD and SCC occurred in opposite directions, which underscores the potential of the changes in these transcripts to differ OPMDs from SCCs.

We further tested all transcript combinations and their power to distinguish OPMD from SCC, out of which 6 combinations had the highest power (AUC = 1) with "Leave-One-Out" cross-validation (Table S7). The results for individual samples are shown (Fig. 4B and Table S8), indicating that the combinations of these transcript changes have the potential to separate OPMD and SCC samples.

Transcripts whose changes are well correlated are not considered independent predictors of outcome [41]. We evaluated if there was a strong correlation between any two of the five potential markers. Correlation analysis shows no significant correlations in both OPMD and SCC groups ($p > 0.05$) (Fig. 4C and Table S9). Because there is no database specific about tongue SCC, a subtype of HNSCC, we then used the HNSCC TCGA Pan-Cancer Atlas database (https://www.cbioportal.org/) to validate the correlations between the changes of these 5 transcripts in SCC relative to margin. Neither CRMP1 (Figure S6A-D) nor IGSF10 (Figure S6D-G) were correlated with the other 4 markers. ELF5, HTR3A, and RPTN were weakly correlated with each other (Figure S6H-J). Thus, changes in these transcripts can serve as independent predictors of outcome.


A
ELF5
IGSF10
RPTN
HTR3A
CRMP1
Log2 Fold Change
OPMD/Margin
SCC/Margin
B
ELF5+HTR3A
RPTN+IGSF10
RPTN+IGSF10+CRMP1
RPTN+IGSF10+CRMP1
RPTN+IGSF10+CRMP1+HTR3A
ELF5+RPTN+IGSF10+CRMP1+HTR3A
Probability of SCC
C
OPMD
ELF5 ISGF10 RPTN HTR3A CRMP1
ELF5 1.00 0.20 -0.33 0.08 -0.50
ISGF10 0.20 1.00 0.03 0.27 0.03
RPTN -0.33 0.03 1.00 -0.32 0.10
HTR3A 0.08 0.27 -0.32 1.00 0.63
CRMP1 -0.50 0.03 0.10 0.63 1.00
ELF5 ISGF10 RPTN HTR3A CRMP1
ELF5 1.00 0.45 -0.35 0.02 -0.49
ISGF10 0.45 1.00 0.07 0.32 -0.14
RPTN -0.35 0.07 1.00 -0.06 0.31
HTR3A 0.02 0.32 -0.06 1.00 0.59
CRMP1 -0.49 -0.14 0.31 0.59 1.00
SCC
ELF5 ISGF10 RPTN HTR3A CRMP1
ELF5 1.00 0.18 0.28 0.59 0.19
ISGF10 0.18 1.00 -0.29 0.22 0.21
RPTN 0.28 -0.29 1.00 0.47 -0.02
HTR3A 0.59 0.22 0.47 1.00 0.12
CRMP1 0.19 0.21 -0.02 0.12 1.00
ELF5 ISGF10 RPTN HTR3A CRMP1
ELF5 1.00 0.27 0.19 0.52 0.27
ISGF10 0.27 1.00 -0.31 0.32 0.16
RPTN 0.19 -0.31 1.00 0.35 0.04
HTR3A 0.52 0.32 0.35 1.00 0.33
CRMP1 0.27 0.16 0.04 0.33 1.00
1.0 0.5 0 -0.5 -1.0

We then built six SCC prediction models using the six transcript combinations (Table S7) on all samples to predict the risk of an individual patient's tongue lesion to become SCC. The models are listed below:

$$P = 1/(1 + e^{-(-1.5937842 - 3.0567907 * ELF5 + 0.4074382 * HTR3A)}) \quad \mathbf{p = 0.00015971}$$

$$P = 1/(1 + e^{-(-0.5560738 - 1.4883524 * RPTN - 1.7171354 * IGSF10)}) \quad \mathbf{p = 0.0001055603}$$

$$P = 1/(1 + e^{-(-0.9167024 - 1.8009372 * ELF5 - 0.8139859 * CRMP1 + 0.1961076 * HTR3A)}) \quad \mathbf{p = 0.000623414}$$

$$P = 1/(1 + e^{-(-0.7026886 - 0.9277354 * RPTN - 0.9976144 * ISGF10 - 0.6827067 * CRMP1)}) \quad \mathbf{p = 0.0005281575}$$

$$P = 1/(1 + e^{-(-0.76304119 - 0.79095631 * RPTN - 0.98737768 * IGSF10 - 0.62807910 * CRMP1 + 0.06449884 * HTR3A)}) \quad \mathbf{p = 0.0023437143}$$

$$P = 1/(1 + e^{-(-0.8914550 - 0.1397542 * ELF5 - 0.7157705 * RPTN - 0.8221783 * IGSF10 - 0.5300456 * CRMP1 + 0.1286990 * HTR3A)}) \quad \mathbf{p = 0.00582664}$$

We finally chose the gene combination "RPTN, IGSF10" as a diagnostic signature to predict the probability of a OPMD becoming SCC because RPTN and IGSF10 were independent predictors of outcome and had statistically significant correlations with the sample groups (OPMD or SCC), with p-values equal to 1.4E-3 and 2.8E-3, respectively (Table S10). We then applied this model to all lesion samples and calculated their probabilities to become SCC (see Table 1), indicating that our prediction model could be complementary to pathological diagnosis.

The logistic regression model not only distinguishes OPMD from SCC states and validate pathological diagnosis, but also has potential to predict the probability of a sample progressing towards SCC. Therefore, we propose the gene combinations "RPTN and IGSF10" as a diagnostic signature gene set because these transcript changes were independent predictors of outcome. We have prepared an app for physicians to validate pathological assessment and potentially predict SCC risk of a tongue OPMD: https://freshtuo.shinyapps.io/sccpred/. To use these two genes in this model, a physician would do the following: (A) Get RNA-seq of the lesion vs. normal tissue nearby from biopsies; (B) Open the app, input the log2 fold changes in these two transcript levels (lesion vs. normal margin); then press "submit"; (C) See the SCC probability result (See an example in the Appendix). This logistic regression model shows proof of concept that the changes in the transcript levels (lesion vs. margin) has potential to verify pathological diagnosis and potentially predict the SCC risk of a tongue lesion. To validate this prediction, the algorithm should be tested on an independent data set, but currently we do not have such a dataset available.

Table 1 Predicted chance (%) of SCC in our samples using Firth logistic regression model

	RPTN + IGSF10		RPTN + IGSF10
OPD1	1.21	SCC1	99.37
OPD2	0.50	SCC2	100.0
OPD3	0.10	SCC3	76.57
OPD4	0.17	SCC4	73.65
OPD5	0.13	SCC5	90.89
OPD6	22.96	SCC6	99.20
OPD7	14.60	SCC7	99.69
OPD8	8.81	SCC8	81.65
OPD9	10.32	SCC9	100.0
		SCC10	99.14
		SCC11	100.0

4 Discussion

Tongue OPMDs are lesions with a high risk of progressing to SCC [6]. Our genome-wide gene expression profiling of human tongue OPMD and SCC samples identified changes in the transcripts that are associated with the clinical pathological diagnosis: **(1)** Compared to the margin samples, OPMD samples show changes in much fewer transcripts compared to SCC samples; **(2)** The number of changed transcripts resembling human oral SCC molecular features in the OPMD samples is lower than that in the SCC samples. These data indicate that, as expected, OPMDs display early pathological changes that may lead to SCC development. The paired design allows us to remove variations among individuals so that we can focus on variations between OPMD/SCC and normal conditions.

Multiple overlapping pathway analyses suggest that EMT and cell migration start in OPMD and that higher signals in the SCC samples indicate that in EMT cell migration occurs more frequently. These data suggest that "quasi-normal" epithelial cells start to disseminate from pre-neoplastic lesions at an early carcinogenesis stage, such as OPMD. These results are in line with the "parallel progression model" of metastasis, *i.e.*, metastasis is initiated long before the primary tumor is well developed and diagnosed [42, 43]. Notably, clinical studies have shown that oral leukoplakia exhibits basement membrane disruption [44, 45].

We categorized these OPMD and SCC lesions into distinct classes and confirmed the heterogeneity among individual human tongue pre-cancerous and cancerous lesions. Importantly, to our knowledge this is the first report in which, based on the transcriptomic changes in OPMD/margin in individual patients, human OPMD samples are classified into different head and neck SCC subclasses (mesenchymal, basal, classical, and atypical). Previous studies [33, 34, 46–48] suggest that this subclass categorization has clinical relevance and has the potential to serve as a reference for patient prognosis and therapeutic options for tongue SCC, and especially OPMD, so this correlation strengthens the possibility that gene profile analysis could be a useful tool to guide prognosis and therapy.

Various prognostic molecular markers for OPMD, including p53, Ki67 and PCNA, cell cycle proteins, loss of heterozygosity (LOH), and some cell surface and stromal proteins, identified in prior studies, have failed to be used in clinical practice for OPMD prognosis [10, 49] because many studies did not have the adequate follow-up required by the longitudinal design criteria and well-defined diagnostic criteria [49]. Recently, researchers proposed a gene set with 11 genes to predict the risk of OPMD progression to SCC [26]. This research used oral lesion samples, not just tongue lesions, and did not focus on transcript changes in individual patients. Moreover, researchers extracted RNA from formalin-fixed paraffin-embedded (FFPE) tissues, which could produce higher data variation at the single gene level, despite the applicability of FFPE tissue for global gene expression analysis [50]. Thus, our high quality, total RNA samples from fresh human tissue (RIN > 8.5) give a more accurate measurement on a single transcript level, and this is critical in generating a gene set signature.

Here we focused on paired cases of OPMD and SCC, *i.e.*, a tongue lesion versus the tongue margin tissue in the same patient. We think that this type of comparison is optimal for addressing the problem of variability in transcript changes among patients. Additionally, we screened candidate transcripts using the opposite directions of fold changes observed in OPMD and SCC groups. Therefore, our prediction model derived from the screened transcripts has the potential to distinguish OPMD from SCC and predict the risks of OPMDs progressing to SCCs. This will help physicians decide whether to surgically remove an OPMD.

One limitation of our study is the relatively small sample size. Thus, validation with an independent data set is needed. Another point is that we have only analyzed mRNA, and not protein levels of these genes. In future work we plan to examine the levels of these proteins in OPMDs. However, we have shown proof of principle that changes in a set of transcripts between a tongue lesion and normal tongue epithelial tissue from the same patient have the potential to classify OPMDs and SCCs and provide insights to the probability of OPMDs becoming oral SCCs.

5 Conclusion

We have built a Firth logistic regression model using the changes in these transcripts relative to paired normal tissue to validate pathological diagnosis and potentially predict the likelihood of an OPMD developing into SCC, as data sets become available.

Acknowledgements We thank the Gudas laboratory for scientific discussions. We thank Dr. John Wagner for critically reading this manuscript. This research was supported by R01 CA205258 to LJG and by Weill Cornell Medicine funds.

Author contributions TZ: conducted research, data collection, software, and statistical analysis, provided data interpretation, manuscript editing, and critical review; DK: conducted research, provided data interpretation, manuscript editing, and critical review; TS: conducted research, provided data interpretation, manuscript critical review and editing; LJG: designed experiment, manuscript critical review and editing, and funding; XHT: designed and conducted research, data collection, conducted molecular and statistical analysis, provided data interpretation, manuscript writing and critical review. All authors read and approved the final manuscript.

Funding This research was supported by R01 CA205258 to LJG and Weill Cornell Medicine funds.

Declarations

Competing interests The authors declare no competing interests.

Appendix

For example, in one patient, the log2 fold change (lesion vs. margin) in RPTN and IGSF10 transcript levels are 1.5 and 2.5, respectively, then the calculated the SCC risk is 0.084%; in another patient, the log2 fold change (lesion vs. margin) in RPTN and IGSF10 transcript levels are − 0.5 and − 1, respectively, then the calculated the SCC risk is 87.05%. Thus, the second patient has a higher SCC risk than the first patient.

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin. 2022;72:7–33.
2. Chinn SB, Myers JN. Oral cavity carcinoma: current management, controversies, and future directions. J Clin Oncol. 2015;33:3269–76.
3. Montero PH, Patel SG. Cancer of the oral cavity. Surg Oncol Clin N Am. 2015;24:491–508.
4. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin. 2020;70:7–30.
5. Warnakulasuriya S. Global epidemiology of oral and oropharyngeal cancer. Oral Oncol. 2009;45:309–16.
6. Warnakulasuriya S. Oral potentially malignant disorders: a comprehensive review on clinical aspects and management. Oral Oncol. 2020;102:104550.
7. Warnakulasuriya S. Clinical features and presentation of oral potentially malignant disorders. Oral Surg Oral Med Oral Pathol Oral Radiol. 2018;125:582–90.
8. Warnakulasuriya S, Johnson NW, van der Waal I. Nomenclature and classification of potentially malignant disorders of the oral mucosa. J Oral Pathol Med. 2007;36:575–80.
9. Edge SB, Compton CC. The American joint committee on cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. Ann Surg Oncol. 2010;17:1471–4.
10. Nikitakis NG, Pentenero M, Georgaki M, Poh CF, Peterson DE, Edwards P, et al. Molecular markers associated with development and progression of potentially premalignant oral epithelial lesions: current knowledge and future implications. Oral Surg Oral Med Oral Pathol Oral Radiol. 2018;125:650–69.
11. Schepman KP, van der Meij EH, Smeele LE, van der Waal I. Malignant transformation of oral leukoplakia: a follow-up study of a hospital-based population of 166 patients with oral leukoplakia from The Netherlands. Oral Oncol. 1998;34:270–5.
12. Warnakulasuriya S, Reibel J, Bouquot J, Dabelsteen E. Oral epithelial dysplasia classification systems: predictive value, utility, weaknesses and scope for improvement. J Oral Pathol Med. 2008;37:127–33.
13. Hashibe M, Brennan P, Chuang SC, Boccia S, Castellsague X, Chen C, et al. Interaction between tobacco and alcohol use and the risk of head and neck cancer: pooled analysis in the international head and neck cancer epidemiology consortium. Cancer Epidemiol Biomarkers Prev. 2009;18:541–50.
14. Samman M, Wood HM, Conway C, Stead L, Daly C, Chalkley R, et al. A novel genomic signature reclassifies an oral cancer subtype. Int J Cancer. 2015;137:2364–73.
15. Mithani SK, Mydlarz WK, Grumbine FL, Smith IM, Califano JA. Molecular genetics of premalignant oral lesions. Oral Dis. 2007;13:126–33.
16. Kondoh N, Ohkura S, Arai M, Hada A, Ishikawa T, Yamazaki Y, et al. Gene expression signatures that can discriminate oral leukoplakia subtypes and squamous cell carcinoma. Oral Oncol. 2007;43:455–62.
17. Kuribayashi Y, Morita K, Tomioka H, Uekusa M, Ito D, Omura K. Gene expression analysis by oligonucleotide microarray in oral leukoplakia. J Oral Pathol Med. 2009;38:356–61.
18. Odani T, Ito D, Li MH, Kawamata A, Isobe T, Iwase M, et al. Gene expression profiles of oral leukoplakia and carcinoma: genome-wide comparison analysis using oligonucleotide microarray technology. Int J Oncol. 2006;28:619–24.

19. Li G, Li X, Yang M, Xu L, Deng S, Ran L. Prediction of biomarkers of oral squamous cell carcinoma using microarray technology. Sci Rep. 2017;7:42105.
20. Yu YH, Kuo HK, Chang KW. The evolving transcriptome of head and neck squamous cell carcinoma: a systematic review. PLoS ONE. 2008;3:e3215.
21. Choi P, Chen C. Genetic expression profiles and biologic pathway alterations in head and neck squamous cell carcinoma. Cancer. 2005;104:1113–28.
22. Adami GR, Tang JL, Markiewicz MR. Improving accuracy of RNA-based diagnosis and prognosis of oral cancer by using noninvasive methods. Oral Oncol. 2017;69:62–7.
23. Richter GM, Kruppa J, Munz M, Wiehe R, Häsler R, Franke A, et al. A combined epigenome- and transcriptome-wide association study of the oral masticatory mucosa assigns CYP1B1 a central role for epithelial health in smokers. Clin Epigenetics. 2019;11:105.
24. Zheng X, Wu K, Liao S, Pan Y, Sun Y, Chen X, et al. MicroRNA-transcription factor network analysis reveals miRNAs cooperatively suppress RORA in oral squamous cell carcinoma. Oncogenesis. 2018;7:79.
25. Yang M, Xiong X, Chen L, Yang L, Li X. Identification and validation long non-coding RNAs of oral squamous cell carcinoma by bioinformatics method. Oncotarget. 2017;8:107469–76.
26. Sathasivam HP, Kist R, Sloan P, Thomson P, Nugent M, Alexander J, et al. Predicting the clinical outcome of oral potentially malignant disorders using transcriptomic-based molecular pathology. Br J Cancer. 2021;125:413–21.
27. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63.
28. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell. 2017;171:1611-24.e24.
29. Network CGA. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 2015;517:576–82.
30. Conway C, Graham JL, Chengot P, Daly C, Chalkley R, Ross L, et al. Elucidating drivers of oral epithelial dysplasia formation and malignant transformation to cancer using RNAseq. Oncotarget. 2015;6:40186–201.
31. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res. 2013;41:D991–5.
32. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1:417–25.
33. Chung CH, Parker JS, Karaca G, Wu J, Funkhouser WK, Moore D, et al. Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression. Cancer Cell. 2004;5:489–500.
34 Walter V, Yin X, Wilkerson MD, Cabanski CR, Zhao N, Du Y, et al. Molecular subtypes in head and neck cancer exhibit distinct patterns of chromosomal gain and loss of canonical cancer genes. PLoS One. 2013;8:e56823.
35. Rahman MS, Sultana M. Performance of Firth-and logF-type penalized methods in risk prediction for small or sparse binary data. BMC Med Res Methodol. 2017;17:33.
36. Piggin CL, Roden DL, Gallego-Ortega D, Lee HJ, Oakes SR, Ormandy CJ. ELF5 isoform expression is tissue-specific and significantly altered in cancer. Breast Cancer Res. 2016;18:4.
37. Ling B, Liao X, Huang Y, Liang L, Jiang Y, Pang Y, et al. Identification of prognostic markers of lung cancer through bioinformatics analysis and in vitro experiments. Int J Oncol. 2020;56:193–205.
38. Cai G, Wu D, Wang Z, Xu Z, Wong KB, Ng CF, et al. Collapsin response mediator protein-1 (CRMP1) acts as an invasion and metastasis suppressor of prostate cancer via its suppression of epithelial-mesenchymal transition and remodeling of actin cytoskeleton organization. Oncogene. 2017;36:546–58.
39. Tone M, Tahara S, Nojima S, Motooka D, Okuzaki D, Morii E. HTR3A is correlated with unfavorable histology and promotes proliferation through ERK phosphorylation in lung adenocarcinoma. Cancer Sci. 2020;111:3953–61.
40. Farah CS. Molecular landscape of head and neck cancer and implications for therapy. Ann Transl Med. 2021;9:915.
41. Mount DW, Putnam CW, Centouri SM, Manziello AM, Pandey R, Garland LL, et al. Using logistic regression to improve the prognostic value of microarray gene expression data sets: application to early-stage squamous cell carcinoma of the lung and triple negative breast carcinoma. BMC Med Genomics. 2014;7:33.
42. Valastyan S, Weinberg RA. Tumor metastasis: molecular insights and evolving paradigms. Cell. 2011;147:275–92.
43. Klein CA. Parallel progression of primary tumours and metastases. Nat Rev Cancer. 2009;9:302–12.
44. Tamgadge SA, Ganvir SM, Hazarey VK, Tamgadge A. Oral leukoplakia: transmission electron microscopic correlation with clinical types and light microscopy. Dent Res J. 2012;9:S94–104.
45. Zhang Z, Guo W, Zhang Y, Wang X, Liu H, Xu S, et al. Changes in the expression of Col IV, gelatinase and TIMP-1 in oral leukoplakia. Int J Clin Exp Pathol. 2017;10:8535–43.
46. Belbin TJ, Singh B, Barber I, Socci N, Wenig B, Smith R, et al. Molecular classification of head and neck squamous cell carcinoma using cDNA microarrays. Cancer Res. 2002;62:1184–90.
47. Lee DJ, Eun YG, Rho YS, Kim EH, Yim SY, Kang SH, et al. Three distinct genomic subtypes of head and neck squamous cell carcinoma associated with clinical outcomes. Oral Oncol. 2018;85:44–51.
48. De Cecco L, Nicolau M, Giannoccaro M, Daidone MG, Bossi P, Locati L, et al. Head and neck cancer subtypes with biological and clinical relevance: Meta-analysis of gene-expression data. Oncotarget. 2015;6:9627–42.
49. Speight PM, Khurram SA, Kujan O. Oral potentially malignant disorders: risk of progression to malignancy. Oral Surg Oral Med Oral Pathol Oral Radiol. 2018;125:612–27.
50. Wimmer I, Tröscher AR, Brunner F, Rubino SJ, Bien CG, Weiner HL, et al. Systematic evaluation of RNA quality, microarray data reliability and pathway analysis in fresh, fresh frozen and formalin-fixed paraffin-embedded tissue samples. Sci Rep. 2018;8:6351.