



# A seven-gene signature to predict the prognosis of oral squamous cell carcinoma

Ilda Patrícia Ribeiro<sup>1,2,3,4</sup> · Luísa Esteves<sup>1</sup> · Ana Santos<sup>1</sup> · Leonor Barroso<sup>5</sup> · Francisco Marques<sup>2,4,6,7</sup> · Francisco Caramelo<sup>2,8</sup> · Joana Barbosa Melo<sup>1,2,3,4</sup> · Isabel Marques Carreira<sup>1,2,3,4</sup>

Received: 23 July 2020 / Revised: 31 March 2021 / Accepted: 20 April 2021 / Published online: 10 May 2021  
© The Author(s), under exclusive licence to Springer Nature Limited 2021

## Abstract

The prognosis of oral squamous cell carcinoma (OSCC) patients remains poor without implemented biomarkers in the clinical routine practice to help in the patient's management. With this study we aimed to identify specific prognostic biomarkers for OSCC using a whole genome technology as well as to verify the clinical utility of a head and neck cancer-specific multiplex ligation-dependent probe amplification (MLPA) panel. A genomic characterization of tumor samples from 62 OSCC patients was performed using array comparative genomic hybridization (aCGH) and a more straightforward and cost-effective molecular technology, MLPA. The identification of a genomic signature and prognosis biomarkers was carried out by applying several statistical methods. With aCGH we observed that the chromosomes most commonly altered were 3p, 3q, 5q, 6p, 7q, 8p, 8q, 11q, 15q, 17q, and 18q. The MLPA results showed that the chromosomes with a higher frequency of alterations were 3p, 3q, 8p, 8q, and 11q. We identified a genomic signature with seven genes *OCNL* (3p21.31), *CLDN16* (3q29), *SCRIB* (3q29), *IKBKB* (3q22.3), *PAK2* (8q22.3), *PIK3CB* (3q28), and *YWHAZ* (8q24.3) that together allow to differentiate the patients that developed metastases or relapses after primary tumor treatment, with an overall accuracy of 79%. Amplification of *PIK3CB* as a predictor of metastases or relapses development was validated using TCGA data. This amplified gene showed a reduction in more than 5 years in the median survival of the patients. The identified biomarkers might have a significant impact in the patients' management and could leverage the OSCC precision medicine.

These authors contributed equally: Ilda Patrícia Ribeiro, Luísa Esteves

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41388-021-01806-5>.

✉ Isabel Marques Carreira  
citogenetica@fmed.uc.pt

- <sup>1</sup> University of Coimbra, Cytogenetics and Genomics Laboratory, Institute of Cellular and Molecular Biology, Faculty of Medicine, Coimbra, Portugal
- <sup>2</sup> University of Coimbra, Coimbra Institute for Clinical and Biomedical Research (iCBR) and Center of Investigation on Environment Genetics and Oncobiology (CIMAGO), Faculty of Medicine, Coimbra, Portugal
- <sup>3</sup> University of Coimbra, Center for Innovative Biomedicine and Biotechnology (CIBB), Coimbra, Portugal

## Introduction

Head and neck cancer (HNC) is a significant public health concern worldwide, having severe morbidity and mortality. These tumors can be characterized according to the site of tumor origin in the upper aerodigestive tract. Oral squamous cell carcinoma (OSCC) is the most common subtype within HNC and encompasses tumors of the lips, the front two-thirds of the tongue, the floor of the mouth, the gums, the

- <sup>4</sup> Clinical Academic Center of Coimbra (CACC), Coimbra, Portugal
- <sup>5</sup> Maxillofacial Surgery Department, Coimbra Hospital and University Centre (CHUC), EPE, Coimbra, Portugal
- <sup>6</sup> Department of Dentistry, Faculty of Medicine, University of Coimbra, Coimbra, Portugal
- <sup>7</sup> Stomatology Unit, Coimbra Hospital and University Centre (CHUC), EPE, Coimbra, Portugal
- <sup>8</sup> University of Coimbra, Laboratory of Biostatistics and Medical Informatics, iCBR—Faculty of Medicine, Coimbra, Portugal

buccal mucosa, the retromolar trigone, and the hard palate [1]. OSCC mostly affects males and its' main risk factors are tobacco and alcohol consumption [2]. The prognosis of these tumors remains poor, with 5-year survival rate of about 40–50%, mainly due to diagnosis in advanced stages and high recurrence rates [3]. Even with the availability of oral cavity for direct inspection, more than half of those tumors have regional or systemic metastasis when first diagnosed. Approximately 10–30% of the patients with advanced tumors develop locoregional recurrences and almost 15–25% present distant metastases [3–5]. Nowadays, despite the recent progress in the oncologic therapeutic strategies, surgery followed by postoperative radiation or chemoradiation remains the standard treatment for OSCC, with treatment selection being mainly based on the stage and histopathological characteristics of the tumors [6, 7]. The posttreatment development of recurrences and metastases is very frequent in OSCC patients. Similarly to other cancers, OSCC is characterized by the accumulation of genomic alterations that compromise many cellular processes and contribute to the establishment of the cancer hallmarks [8]. The lack of OSCC biomarkers to predict clinical outcomes and the risk of recurrence after treatment is a major concern [9]. The identification, and translation to routine clinical practice, of specific and sensitive biomarkers that could relate OSCC with pathophysiological characteristics would improve prognosis, treatment selection, and prophylactic screening of patients [10]. Currently, although many genes related to OSCC have been hypothesized with biomarker potential, they have not been implemented in the clinical routine yet [8]. There are many challenges to overcome in the molecular clinical management of OSCC patients, and consequently in the implementation of precision medicine to improve patient's

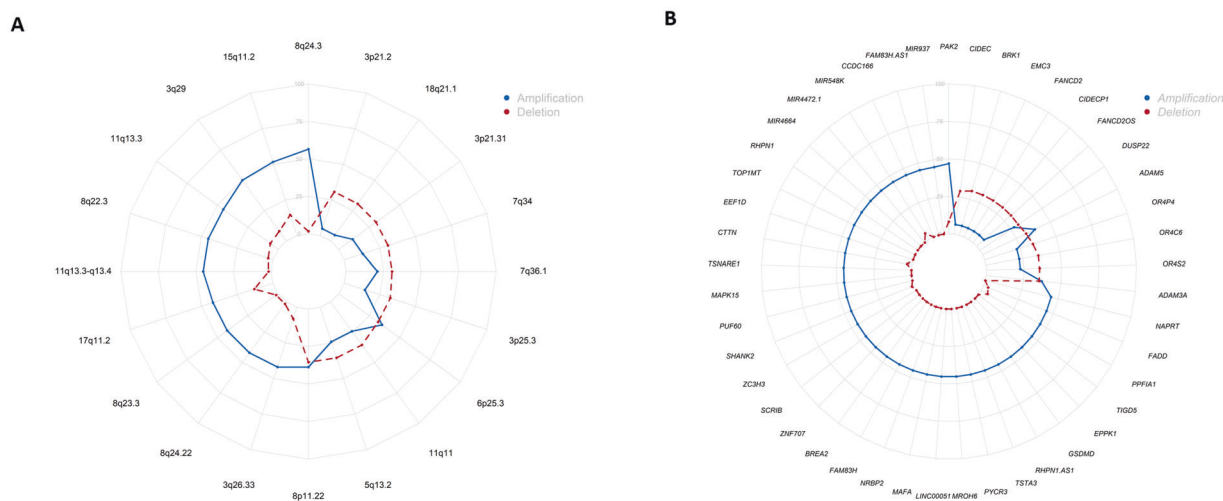
survival and quality of life. In this study, we established the potential of copy number alterations in the prognosis of OSCC, applying statistical methods to identify a seven-gene signature to predict the development of relapses and metastasis. Additionally, we explored the possibility of using a specific MLPA panel that could, in a sensitive, rapid, and cost-effective manner, be used in a routine clinical practice. The *PIK3CB* gene was validated in an independent OSCC dataset from The Cancer Genome Atlas (TCGA) and showed also an association with reduction in patients' survival.

## Results

### Characterization of aCGH data

A total of 8507 chromosomal regions with copy number alteration were detected by aCGH in the 62 OSCC patients. The most frequently altered chromosome bands and genes were highlighted in Fig. 1A and B, respectively. The most frequently amplified chromosome bands were 8q24.3, 15q11.2, 3q29, 11q13.3, 8q22.3, 11q13.3-q13.4, 17q11.2, 3q26.33, 8q23.3, and 8q24.22, all showing alterations in more than 42% of patients and 8q24.3 being altered in 58% of the patients. The most frequently deleted chromosomal regions were 8p11.22, 5q13.2, 11q11, 6p25.3, 3p25.3, 7q36.1, 7q34, 3p21.31, 18q21.1, and 3p21.2, that showed alterations in at least 30% of patients, reaching the highest frequency of alteration in 8p11.22, with 35% of patients showing deletion of genetic material.

The most frequently amplified gene was *PAK2* (47%) located at 3q29 and there were other 33 genes that showed the same frequency of alteration of 45% (Fig. 1B). These



**Fig. 1** Circos plots representing the most frequent alterations detected by array CGH, in the OSCC cohort. **A** Most frequently amplified and deleted chromosome bands detected by aCGH. **B** Most frequently amplified and deleted genes determined by aCGH. Amplifications of genetic material are represented in blue and deletions are represented in red.

genes mostly co-located to the 8q24.3 chromosomal region ( $n = 28$ ), four were mapped to 11q13.3 (*MIR548K*, *CTTN*, *TIGD5*, and *PPFIA1*) and the other, *SHANK2*, at 11q13.3-q13.4. The most amplified genes were all located in chromosome bands that also showed a higher proportion of gains (Fig. 1A).

The most frequently deleted genes were *ADAM3A*, *OR4S2*, *OR4C6*, and *OR4P4* all showed losses in 35% of patients (Fig. 1B). *ADAM3A* is located at 8p11.22 and the others are mapped to 11q11. Next, with deletion in 32% of patients was *ADAM5* (8p11.22). *DUSP22*, *FANCD2OS*, *CIDECP1*, *FANCD2*, *EMC3*, *BRK1*, and *CIDEA* all showed deletion in 30% of patients. *DUSP22* is located at 6p25.3 and the remaining six genes all co-locate to 3p25.3. The most frequently deleted gene was located at the most frequently deleted chromosome band and the other genes were also located in regions with a larger number of detected deletions (Fig. 1A).

Fractions of alteration by chromosomal arm were also calculated and the correlation between them was used to evaluate if there were any concomitant alterations, that is, alterations that could frequently be observed together (Fig. 2).

The strongest relationships in gains were found in the pairs 21p and 22p (Spearman's correlation coefficient = 0.87), 8q and 9q (Spearman's correlation coefficient = 0.84), and 9q and 10q (Spearman's correlation coefficient = 0.8) (Fig. 2A).

When considering losses of genetic material, the strongest correlation was observed for the pair 19p and 20p (Spearman's correlation coefficient = 0.88) (Fig. 2B). This

means that there seems to be a strong enough probability that alterations in those chromosome arms are related, appearing concomitantly in our OSCC cohort.

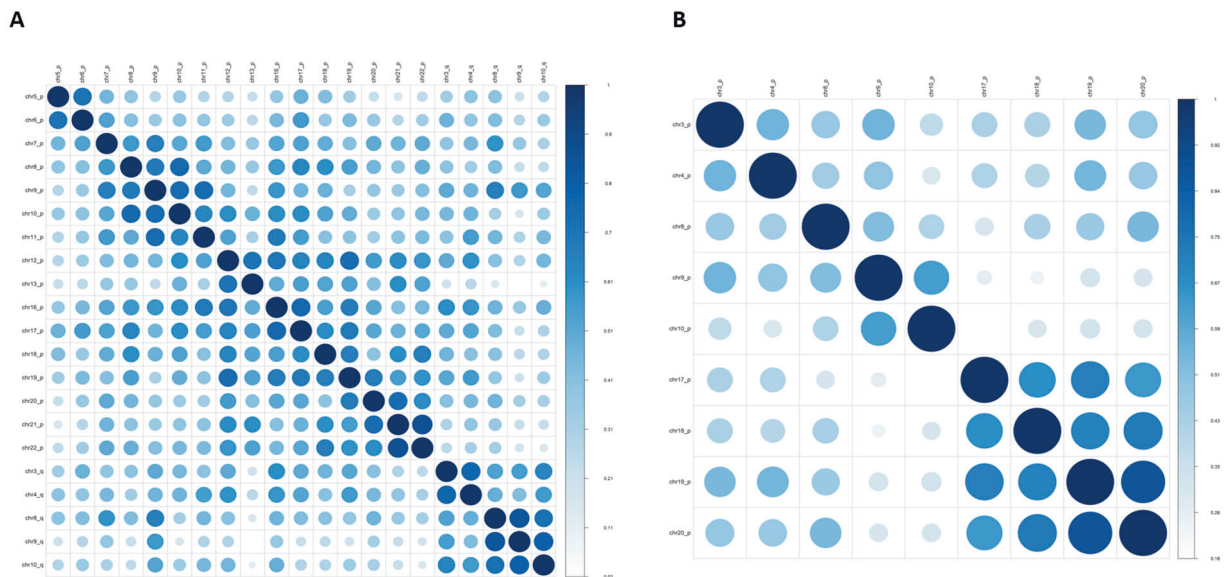
## Data reduction and multivariate analysis

With the objective of finding new relevant genes with prognosis value in the whole genome aCGH data, a 35% frequency filter was established for the number of alterations per gene. From the initial 21,939 mapped genes in the 8507 regions detected by aCGH, 472 genes are left after filtering (Supplementary Fig. 1).

A signaling pathway analysis was performed, resulting in 15 signaling pathways that are overrepresented ( $p < 0.05$ ) in our cohort (Supplementary Fig. 2). A total of 49 genes that were associated with those pathways were considered in further analysis.

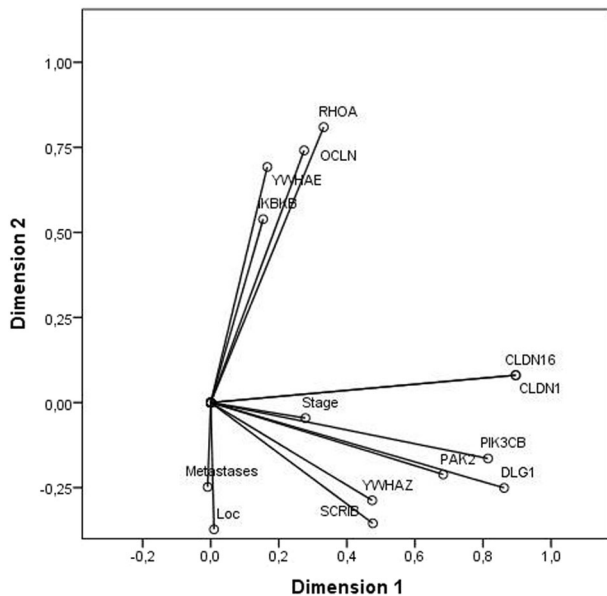
As a following step to reduce the amount of data to be analyzed, only genes present in at least three of these pathways were maintained. The 11 genes that remained were *OCN* (3p21.31), *CLDN16* (3q29), *DLG1* (8p11.21), *SCRIB* (3q29), *IKBKB* (3q22.3), *RHOA* (17p13.3), *PAK2* (8q22.3), *PIK3CB* (3q28), *YWHAE* (3q28), *YWHAZ* (8q24.3), and *CLDN1* (5q13.2).

Subsequent multivariate analyses of the 11 genes and three clinical-pathological parameters (anatomical location of the primary tumor, presence of metastases or relapse, and tumor staging) revealed the different contributions of the tested variables to two dimensions that accounted for 48.4% of variance (Cronbach's alpha coefficient = 0.918) (Fig. 3).



**Fig. 2** Correlation between alterations in chromosomal arms presenting a fraction of alteration higher, in average, than 0.1. **A** Correlation between gain fractions; 21p and 22p (Spearman's correlation coefficient = 0.87), 8q and 9q (Spearman's correlation coefficient = 0.84), and 9q and 10q (Spearman's correlation coefficient =

0.8) were found to show the strongest correlations. **B** Correlation between loss fractions; the strongest correlation was observed for 19p and 20p (Spearman's correlation coefficient = 0.88). Darker and large dots mean larger values of the coefficient of correlation, whereas lighter and smaller dots mean smaller values.



**Fig. 3** PCA using the selected 11 genes identified with aCGH technique. The first dimension includes the *CLDN16*, *CLDN1*, *DLG1*, *PIK3CB*, *PAK2*, *SCRIB*, *YWHAZ* genes and the clinical stage variable. The second dimension comprises *RHOA*, *OCLN*, *YWHAE*, *IKBKB* and the anatomical location of the primary tumor as well as the presence of metastasis or relapse.

The first dimension (Cronbach's alpha coefficient = 0.823) encompassed the *CLDN16*, *CLDN1*, *DLG1*, *PIK3CB*, *PAK2*, *SCRIB*, *YWHAZ* genes and the clinical stage. The second dimension (Cronbach's alpha coefficient = 0.652) was comprised of the *RHOA*, *OCLN*, *YWHAE*, *IKBKB* and the anatomical location of the primary tumor as well as the presence of metastasis or relapse.

The 11 genes were used to build logistic regression models, using the clinical-pathological features as dependent variables. The variables to be included in the model were selected in a stepwise manner, where only the most significant genes for the model are automatically chosen. The stepwise method uses a quality measure of the model to select the most suited variables, incorporating the set of variables that maximizes the model performance. When using the primary location of the tumor or the clinical stage of the patients, no association was found between these features and the selected genes.

However, when considering the presence of metastases or relapse, a statistical model with an overall accuracy of 79% was generated ( $\chi^2 = 29.657$ ,  $p = 0.003$ ). The null model's accuracy was 58.1%. The variables selected by the stepwise algorithm were the genes *OCLN*, *CLDN16*, *SCRIB*, *IKBKB*, *PAK2*, *PIK3CB*, and *YWHAZ*.

According to the Hosmer–Lemeshow goodness of fit test, the model has a good fit for the data with a  $p$  value greater than 0.05 ( $p = 0.981$ ). When taking together the seven independent variables in the logistic model, they

account for 51.1% of the variance in the tumors tested (Nagelkerke  $R^2 = 0.511$ ), meaning that they explain 51.1% of the reason why a patient develops a relapse/metastasis or not. The model has a good separation ability, since the area under the ROC curve was 0.849, with a 95% CI [0.754; 0.945], meaning that there is an 84.9% probability of the model correctly predicting the patient's class.

Both *SCRIB* (OR = 0.11,  $p = 0.015$ ) and *PIK3CB* (OR = 0.041,  $p = 0.04$ ) showed significant predictive value associated with the development of metastasis or relapse.

*SCRIB* presented only two states of alteration in the cohort: normal or amplified; the results show that, when this gene was normal i.e., not altered, the patients had a lower probability to develop metastasis or relapse. If the gene was amplified, the probability to develop metastasis or relapse later on was about nine times higher.

*PIK3CB* showed both amplifications and deletions, however it showed amplification in 32% and deletion in only 6% of patients. When comparing the probability to develop metastasis or relapse in these patients, only the amplification versus the normal state of *PIK3CB* were determined to have statistical significance, and, like *SCRIB*, when the gene was amplified the odds to progress into metastatic cancer or relapse after the primary diagnosis and treatment were higher. In fact, the risk increased 24.4 times.

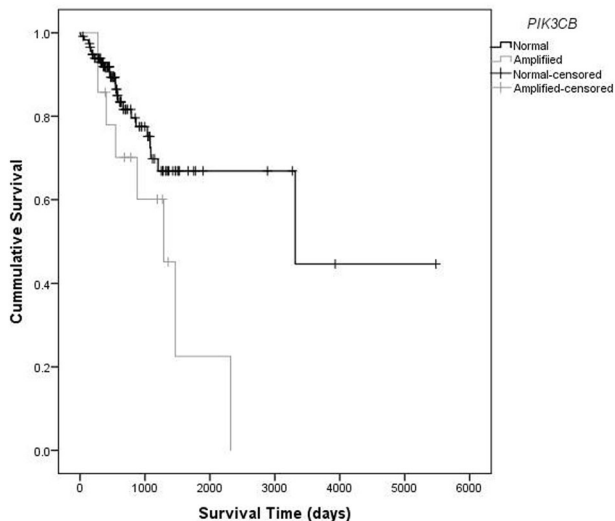
### Validation using TCGA data

A logistic regression model was applied to the TCGA data to establish the relevance of the genes used in the logistic regression model created for the cohort data, to the development of metastasis or relapse. It was found that *PIK3CB* was the only gene that showed significance to the distinction between the two classes with  $p = 0.004$ . *PIK3CB* was amplified in about 21% of patients and was normal in the remaining 79%. The odds of a patient developing metastasis or relapse if the gene was amplified were about seven times higher than when the gene was normal.

A Kaplan–Meier survival analysis was also performed for the TCGA data in order to evaluate individually the impact of the genes included in the classification model on the survival of the patients. In this analysis, only *PIK3CB* showed a statistically significant difference (Log-rank  $p = 0.04$ ) in survival between the patients where the gene was amplified or normal. In median, patients that had a normal *PIK3CB* gene survived 9.08 years, as opposed to the 3.53 years patients with *PIK3CB* amplification survived, in median (Fig. 4).

### Characterization of MLPA data and comparison with aCGH results

The frequency of gene imbalances for all probe panel genes (41 probes covering 36 genes in 12 different chromosomal



**Fig. 4** Kaplan–Meier survival curves relating to the *PIK3CB* gene, in the TCGA cohort. Patients with a normal gene survived, in median, around 5.5 years longer than those where the gene was amplified.

locations) exhibiting alterations was measured directly, except for the reference probes which were used as a quality control checkpoint. Genetic imbalances were observed in all 36 analyzed genes: 27 showed both gains and losses of genetic material, four presented only gains and five were exclusively deleted. A higher frequency of loss of genetic material was observed in 3p (*RARB*, *RASSF1*, and *FHIT*) and in 8p (*CSMD1*, *GATA4*, and *MTUS*). Gain of genetic material was primarily found in 3q (*CCND1*, *PIK3CA*, *MCCCI*, and *TP63*), chromosome 7 (7p, *EGFR*, and 7q, *ABCBI*, *CDK6*, and *MET*), in 8q (*MYC*, *WISP1*, and *PTK2*), and 11q (*CCND1*, *FGF4*, *FADD*, and *CTTN*). A higher frequency of gains was observed when compared to losses. The chromosomes that presented a higher frequency of gains were 11q, 3q, and 8q and chromosomes 8p and 3p presented a higher frequency of losses. The frequency of these imbalances is represented in Fig. 5.

Major losses were present in *GATA4* (24%), *RARB* (23%), *FHIT* (23%), and *MTUS* (23%), and major gains were present in *FADD* (52%), *CTTN* (45%), *CCND1* (45%), *FGF4* (44%), and *MYC* (40%). Some genes also showed low levels of alteration: *MTUS1*, *DEPDC1B*, and *RASSF1* were altered in least than 10% of patients. These genes were mostly deleted in the cohort, with *MTUS1 ex2* being deleted in 3.2% of patients, *MTUS1 ex1* in 4.8%, *DEPDC1B* and *RASSF1* in 9.7%.

The genes included in the MLPA panel were evaluated on the data obtained by aCGH, in order to evaluate their relevance to the panel. The most frequently amplified genes in aCGH from those evaluated in MLPA probe panel were *CTTN* (45%), *FADD* (45%), *FGF4* (44%), *CCND1* (44%), and *WISP1* (40%) (Fig. 6). As for losses, the most

frequently observed were in genes *FHIT* (24%), *CSMD1* (23%), *RASSF1* (21%), *CDK6* (19%), *GATA4* (19%), *H2AFX* (19%), and *ATM* (19%). These results are in line with the frequencies of alteration calculated from the data obtained by MLPA.

To evaluate the agreement between results obtained by aCGH and by MLPA techniques, Cohen's kappa inter-rater agreement index was calculated for each gene included in the MLPA panel. These indexes are represented in Table 1.

The highest agreement was obtained for *FGF4* ( $k = 0.86$ ), *MCCCI* ( $k = 0.81$ ), and *CTTN* ( $k = 0.81$ ), which, according to Altman [11] allows to classify them as having “very good agreement”. Since the values that are nearest to 1 are less likely to be due to chance occurrence, these genes have a better agreement in both MLPA and aCGH techniques. As the values decreasingly approach 0, the probability that there is no agreement between techniques becomes higher. In 19 genes, a very good to moderate agreement were shown and 18 genes show fair to poor agreement (Table 1).

One gene—*WHSC1*—present in the MLPA probe panel was not found in the aCGH results, due to the lack of probe coverage in the region where the gene is located.

## Discussion

High-throughput genomic technologies have the potential to have a major impact on preclinical and clinical drug development and the selection and stratification of patients in clinical trials. High-throughput genomic technologies have the potential to help and guide in the clinical management and in the patient's risk stratification. With aCGH technique we identified the genes and chromosomal regions commonly altered in OSCC. We observed a spectrum of alterations distributed across the chromosomes 3p, 3q, 5q, 6p, 7q, 8p, 8q, 11q, 15q, 17q, and 18q, being 8p and 8q the chromosomes arms most frequently deleted and amplified, respectively. These alterations are in line with the literature [12] and our own previous works [9, 13]. In these altered chromosomal regions several signaling pathways are disrupted and some known and novel putative oncogenes and tumor suppressor genes that could guide and help in the development of new target therapies and patient's management are mapped, such as *PAK2* and *ADAM3A* genes. These two genes are the ones most frequently altered in our cohort. *PAK2* is a member of PAK family of serine/threonine kinases and plays a role in several cellular functions, such as chromatin and cytoskeletal remodeling, histone modifications, proliferation, and apoptosis [14]. The over-expression of *PAK2* seems to be a key player in head and neck progression [14]. Interestingly, a homozygous deletion of *ADAM3A*, a member of ADAM family of genes that



**Table 1** Cohen's kappa calculated for each gene in the HNC-specific probe panel.

Gene	kappa	Gene	kappa
<i>FGF4</i>	0.86	<i>SMAD2</i>	0.43
<i>MCCC1</i>	0.81	<i>FHIT</i>	0.42
<i>CTTN</i>	0.81	<i>CCNL1 ex2</i>	0.39
<i>FADD ex1</i>	0.77	<i>ATM</i>	0.39
<i>CCND1</i>	0.76	<i>KCNRG</i>	0.39
<i>GATA4</i>	0.74	<i>WDR36</i>	0.38
<i>GALR1</i>	0.70	<i>RARB</i>	0.35
<i>MYC</i>	0.65	<i>RASSF1</i>	0.33
<i>PTK2</i>	0.64	<i>FADD ex2</i>	0.32
<i>PIK3CA</i>	0.62	<i>EGFR ex28</i>	0.28
<i>EGFR</i>	0.61	<i>BRCA2</i>	0.27
<i>ABCB1</i>	0.59	<i>DEPDC1B</i>	0.22
<i>MET</i>	0.57	<i>CD38</i>	0.17
<i>MTUS1 ex3</i>	0.57	<i>RB1</i>	0.16
<i>WISP1</i>	0.54	<i>WFS1</i>	0.15
<i>TP63</i>	0.52	<i>MTUS1ex2</i>	0.11
<i>CCNL1 ex1</i>	0.52	<i>BTNL3</i>	0.10
<i>CDK6</i>	0.50	<i>CHEK1</i>	0.04
<i>SMAD4</i>	0.49	<i>MTUS1ex1</i>	0.02
<i>CSMD1</i>	0.46	<i>H2AFX</i>	-0.03

1.0-0.8	Very good agreement
0.8-0.6	Good Agreement
0.6-0.4	Moderate agreement
0.4-0.2	Fair agreement
< 0.2	Poor agreement

Kappa's values are colored from light brown to dark brown as the agreement increases. *FGF4*, *MCCC1*, *CTTN*, and *FADD* exon 1 genes showed higher agreement, while *BTNL3*, *CHEK1*, *MTUS* exon1, and *H2FX* showed the lowest. Also, *H2FX* gene was the only gene with a negative value of  $-0.03$ . Agreement values in accordance with Altman [11].

plays several potential functions related to cell-to-cell communication and adhesion and are also thought to play a role in cancer, was already described in gliomas [15].

The results obtained with whole genome aCGH analysis revealed the great genomic complexity that underlies oral carcinogenesis, which could explain the frequent therapeutic failure and drug resistance and consequently the development of recurrences. It is important to mention that, nowadays, only the anatomic location and tumor stage guide treatment selection, even though patients with similar clinic-pathological characteristics may differ in their clinical outcome, as a possible consequence of the presence of different genomic entities with distinct prognosis and tumor behavior.

Regarding these results, we also observed some correlated alterations between chromosome arms, e.g., in our OSCC cohort some joint losses and gains were simultaneously observed in specific chromosomes. The simultaneous loss of 19p and 20p and the simultaneous gain of 21p and 22p, 8q and 9q, and 9q and 10q were identified. These results could show the importance of these chromosomal regions together for the OSCC carcinogenic process, having possibly a joint driver effect. Further studies of the architecture of these chromosome arms are warranted, in order to try to understand these specific co-alterations.

Around at least one-third of locally advanced HNC patients develop loco-regional relapse and almost 20% of HNC patients develop distant metastasis after treatment [16, 17]. So, the identification of validated and robust biomarkers to predict the risk of relapse and metastasis development is urgent and vital to improve OSCC patient's survival and quality of life.

In this study, we identified a genomic signature with seven genes *OCN* (3p21.31), *CLDN16* (3q29), *SCRIB*

(3q29), *IKBKB* (3q22.3), *PAK2* (8q22.3), *PIK3CB* (3q28), and *YWHAZ* (8q24.3) that together allow to distinguish the patients that present metastases or relapses, with an overall accuracy of 79%. Moreover, both the amplification of *SCRIB* and *PIK3CB* genes showed significant predictive value to identify the patients at risk of metastasis or relapse development, being this risk 9 and 24.4 times higher, respectively. Using data from the TCGA data portal, we validated the amplification of *PIK3CB* as predictor of metastases or relapses in OSCC patients, having also this amplified gene shown a reduction in more than 5 years in the median survival of these patients. Therefore, with our study, we identified two potentially independent biomarkers of prognosis, *SCRIB* and *PIK3CB* genes.

The phosphoinositide 3-kinase (PI3K) pathway is commonly altered in HNC, disrupting several cellular functions related to tumor behavior, such as cell growth, local invasion, metastasis, survival, and resistance to therapy [18]. PI3K signaling pathway is frequently deregulated in human cancers through different mechanisms, such as loss or inactivation of the tumor suppressor *PTEN*, mutation, or amplification of *PI3K*, and activation of tyrosine kinase growth factor receptors or oncogenes upstream of PI3K [19]. In 2014, Al-Rawi et al. [20] described the activation of PIK3CB signaling pathway in advanced stage of OSCC; however, the authors did not find any significant correlation between this protein and the clinic-pathological features of the 31 studied patients. In 2015, Lin et al. [21] observed that in human cancer, the most common alteration in *SCRIB* gene is amplification and increased expression, namely in breast, lung adenocarcinoma, and hepatocellular carcinoma. However, in HNC the *SCRIB* mRNA is not significantly altered, suggesting a different type of regulation [21]. The

*OCLN* gene, responsible for producing the protein occludin that is a tight junction integral protein and structurally connected to the TGF- $\beta$  receptor is implicated in differentiated epithelial cells and in cellular phenotype changes through epithelial-to mesenchymal transformation [22]. The *CLDN16* is a protein coding gene, being the protein a member of the claudins family, a tight junction protein that plays roles in the maintenance of cell polarity, cellular arrangement, adhesion, paracellular transport, and ionic permeability [23]. The *IKBKB* gene encodes a Ser/Thr kinase that is a protein capable to phosphorylate a wide range of substrates and consequently to regulate cellular processes, such as cell cycle, metabolism, and differentiation, promoting or hampering tumoral transformation according to yet unknown cellular or mutational events [24]. The *YWHAZ* gene is frequently amplified in cancer, playing a role in tumor cell proliferation, migration and invasion with its overexpression being associated with the lower expression of miR-375 and seems to could help in the identification of poor prognosis patient's [25]. To our knowledge, our study is the first to report this seven-gene signature related to OSCC recurrence development, highlighting the *PIK3CB* amplification as predictor of metastasis and relapses development in OSCC, with a strong effect on patients' survival.

It is important to stress that large-scale genomic profiling in OSCC has been challenging due to the great intra and intertumoral heterogeneity presented in solid tumors [26]. Additionally, from the clinical point of view, the presence of genetically heterogeneous subclones within the tumor comprises also a challenge for treatment design and patient's management [27]. In this sense, a global and more comprehensive picture of molecular landscape of tumors obtained with high-throughput technologies is vital to identify new prognostic biomarkers for the stratification of patients with high risk of metastasis and relapse and consequently to help in the design and selection of targeted therapies. Nevertheless, since our capability to interpret and use these big data for the patients' benefit is still scarce, the development of a targeted probe panel with specific biomarkers seems to be cheaper and faster in both the laboratory procedures and interpretation, being more suitable for routine clinical practice. In this sense, we include in our study a targeted analysis using a specific MLPA probe panel for HNC in order to verify its utility for genomic characterization of our OSCC cohort. Taking into account the results obtained with this MLPA probe panel, we verified that the chromosomes with a higher frequency of alterations were 3p, 3q, 8p, 8q, and 11q. These results are in agreement with our previous study using also this probe panel in only 30 of these patients [28] as well as the results obtained by studying the whole genome using aCGH technique. In both techniques, aCGH and MLPA, a higher frequency of gains was observed in comparison with losses.

It is also relevant to highlight the genes with probes in this MLPA probe panel specific for HNC that present a low rate of alterations, namely *MTUS1*, *DEPDC1B*, and *RASSF1*, which may be indicative that their relevance for oral carcinogenesis is low. Also, in the aCGH results, we verified the low level of alteration for these genes, ranging from 0% of amplification in *RASSF1* to the maximum being reached by the percentage of deletion in the same gene (20.97%). In this sense, thinking about the clinical utility of these data, replacement of some of these genes could be an adequate strategy, exchanging them with new genes with a strong association with OSCC prognosis.

Our study has some potential limitations, first, the sample size is relatively small; second, we are analyzing as a single tumor entity, sampling from two different anatomic locations in the oral cavity, from different tumor stages and different treatment modalities.

In conclusion, our whole genome results showed a complex genomic profile of cancer-related genes and genomic regions that contribute to OSCC carcinogenesis, supporting its heterogeneous genomic nature. Some common genomic alterations identified were chromosomes 3p, 3q, 5q, 6p, 7q, 8p, 8q, 11q, 15q, 17q, and 18q, where are mapped several cancer related genes, being *PAK2* and *ADAM3A* the two genes most frequently altered in our cohort. We identified a genomic signature with seven genes *OCLN* (3p21.31), *CLDN16* (3q29), *SCRIB* (3q29), *IKBKB* (3q22.3), *PAK2* (8q22.3), *PIK3CB* (3q28), and *YWHAZ* (8q24.3) that together allow to differentiate the patients in risk of metastases or relapses development during and after treatment. Particularly, amplifications of *SCRIB* and *PIK3CB* genes demonstrated an independent significant predictive value to identify those patients in risk of metastasis or relapse development, being the predictive value of *PIK3CB* gene validated in an independent cohort from TCGA and also correlated with decreased survival. The MLPA technique targeting specific diagnostic and prognostic biomarkers showed to be a fast and cost-effective approach for application in routine clinical practice. The present study gave a step forward in the understanding of the molecular basis of oral tumorigenesis as well as in the identification of prognosis biomarkers, which might have a great impact in the patients' management and consequently lead the patients to be treated sooner, with economic and social benefits, especially regarding their longevity and quality of life.

## Materials and methods

### Study population

The study was approved by the Committee on Ethics in Research of the Faculty of Medicine of the University of



Coimbra. All the procedures were conducted according to the Declaration of Helsinki and patients provided written informed consent. The analyzed cohort is comprised of tumor tissue from 62 OSCC patients who underwent tumor resection. They were recruited between October 2010 and May 2015 from the Maxillofacial Surgery and Stomatology Units, of the Coimbra Hospital and University Centre, CHUC, EPE, Portugal. Diagnosis and staging were performed in accordance with the American Joint Committee on Cancer's tumor–node–metastasis staging system. The detailed characterization of the cohort in the study is represented in Table 2. The follow-up periods ranged from 36 to 91 months. For the control group, gingival tissue samples from healthy donors subjected to wisdom teeth removal were used. Tissue samples were preserved in RNAlater and stored at  $-20^{\circ}\text{C}$  until use.

### DNA extraction

Genomic DNA from fresh frozen tissues of patients and controls were extracted using a High Pure PCR Template Preparation Kit (Roche GmbH, Mannheim, Germany), according to the manufacturer's instructions. The DNAs

were quantified using a Nanodrop 1000 Spectrophotometer (Thermo Scientific, USA).

### Array comparative genomic hybridization (aCGH) analysis

Whole genome analyses through aCGH technique were performed using Agilent SurePrint G3 Human Genome microarray 180 K (Agilent Technologies, Santa Clara, CA, USA), according our previous work [9]. DNA from tumor and controls samples was labeled with Cy5 and Cy3, respectively, by random primer labeling. Results were analyzed using Agilent Genomic Workbench v6.5 software with ADM1 as aberration algorithm, threshold of 6.0, moving average 2 Mb. The results are according to Human Genome build 19 and include imbalances with at least three consecutive probes with abnormal  $\log_2$  ratios. This cohort of 62 OSCC patients had already been analyzed by aCGH and the results were included in a larger cohort of HNC patients and published together by our group in 2017 [9].

### Multiplex ligation-dependent probe amplification (MLPA) analysis

MLPA analysis was performed using P428-B1-lot1111 panel, specific for HNSCC. This MLPA assay was designed and optimized at MRC-Holland (Amsterdam, the Netherlands) in order to include genes that are suggested to be of diagnostic and/or prognostic relevance in this neoplasm. This panel includes 41 probes targeting 36 different genes, located on chromosomes 3, 4, 5, 7, 8, 11, 13, and 18. All MLPA reactions were performed according our previous work [28]. Three controls without CNAs as well as a negative control (without DNA) were included in each MLPA assay. Binning of the raw data and comparative analyses were performed using Coffalyser.NET software. For each probe we determined the specific cut-off values for gain and loss, being a copy number gain scored when a value exceeded 1.2 and a copy number loss scored when a value was lower than 0.8. These cut-off values were based on our previous works [13, 28]. From our cohort of 62 OSCC patients, 30 patients had already been analyzed using this MLPA probe panel and the results were published by our group in 2014 [28].

### Statistical analysis

#### Characterization of the aCGH data

Array-CGH was performed on 62 tumor samples from patients with OSCC and the regions with copy number alterations were determined. The genes contained in these regions were determined using the Bioconductor's *Homo*.

**Table 2** Clinic-pathologic characteristics of study population.

Patients ( $n = 62$ )			
$n$ (%)		$n$ (%)	
Gender		Differentiation	
Male	53 (85)	Well	42 (68)
Female	9 (15)	Moderate	15 (24)
Age at diagnosis (years)		Poor	1 (2)
<60	30 (48)	NA	4 (6)
$\geq 60$	32 (52)		
Tobacco		HPV	
Yes	39 (63)	Positive	1 (2)
No	10 (16)	Negative	49 (79)
NA	13 (21)	NA	12 (19)
Alcohol		Margin status	
Yes	25 (40)	R0	38 (61)
No	9 (15)	R1	12 (19)
NA	28 (45)	NA	19 (30)
TNM stage		Anatomic subsite	
I	9 (15)	Tongue	36 (58)
II	15 (24)	Floor of mouth	26 (42)
III	11 (18)		
IV	16 (26)	Invasion peri(neural)	
NA	11 (18)	Yes	25 (40)
Metastasis		No	13 (21)
Yes	26 (42)	NA	34 (39)
No	36 (58)		

*sapiens* package. The frequency of alteration for the previously determined genes contained in the regions detected by aCGH was found. The most frequently altered genes and regions were determined. This analysis was done using R, by counting the number of patients that had alterations in a given gene.

In order to better characterize the genetic alterations detected by aCGH present in the tumor tissue samples from our cohort, a fraction of alteration was calculated for each chromosomal arm. This proportion calculated for each type of copy number change (gain and loss) is defined as the ratio between the number of altered base pairs and the size of the chromosomal arm in base pairs as well. This fraction varies between 0 (when there are no changes in the chromosomal arm) and 1 (when the entire arm is changed). These gain and loss fractions were used to analyze the correlation between chromosomes in order to detect concomitant changes and, for this, Spearman's correlation coefficient was calculated for chromosome arms that showed, at least, a 0.1 average fraction of alteration.

### Data reduction and multivariate analysis

Due to the vast volume of results obtained by aCGH data reduction was required. A 35% frequency filter was established for the number of alterations per gene. The genes contained in the MLPA panel under study were removed from this analysis, since the objective was to find relevant alterations in genes other than those present in the MLPA panel. A signaling pathways analysis was also performed using *limma* and the most significantly overrepresented pathways ( $p < 0.05$ ) were selected. Only the genes that were present in, at least, three of those pathways were kept and used in further analyses. A principal component analysis (PCA) was performed using the selected genes in association with clinic-pathological variables (anatomical location of the tumor, presence of metastasis/relapse, and tumor staging) in order to reduce even further the number of variables and to observe the possible relationships between them on a chart representing the first two principal components. The variables considered in this analysis were all nominal, since the copy number alteration data were discretized, so a nonlinear PCA using the CATPCA algorithm available in IBM SPSS Statistics was performed. A more tuned assessment of the relationships between variables was achieved by applying a logistic regression model, using the alteration status of the genes as independent variable.

### Validation using TCGA data

Array-CGH data from a cohort of 131 HNSCC patients that showed alterations in, at least, one of the genes that were used in the logistic regression model established with the

aCGH data obtained from our cohort, was extracted from TCGA. A logistic regression was performed using the development of metastasis or relapse as a dependent variable, as well as a survival analysis using the Kaplan–Meier method.

Statistical analyses were carried out using Excel, R version 3.5.2, MATLAB release R2019b, and IBM SPSS Statistics version 24. The significance level adopted was 0.05.

### Characterization of MLPA data and comparison with aCGH results

The frequency of alteration was determined for each gene present in the MLPA panel. Cohen's kappa inter-rate agreement statistic was calculated to verify whether there was agreement between the results obtained by aCGH and MLPA, assessing the reliability of gene imbalances detected by the two techniques, for the same gene. To achieve that, the data were discretized into three states (normal, deleted, and amplified), and the comparison was performed on the type of alteration detected in each patient by gene.

### Data availability

The data from array-CGH analyses that support the findings of this study are openly available at <https://apps.uc.pt/myapa/ge/faculty/fcaramelo/en/oscc>.

**Acknowledgements** We thank MRC-Holland for the MLPA P428-B1 probe panel.

### Compliance with ethical standards

**Conflict of interest** The authors declare no competing interests.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### References

- Gilyoma JM, Rambau PF, Masalu N, Kayange NM, Chalya PL. Head and neck cancers: a clinico-pathological profile and management challenges in a resource-limited setting. *BMC Res Notes*. 2015;8:772.
- Chatterjee A, Laskar SG, Chaukar D. Management of early oral cavity squamous cancers. *Oral Oncol*. 2020;104:104627.
- Braakhuis BJ, Brakenhoff RH, Leemans CR. Second field tumors: a new opportunity for cancer prevention? *oncologist*. 2005;10:493–500.
- Leemans CR, Tiwari R, Nauta JJ, van der Waal I, Snow GB. Regional lymph node involvement and its significance in the development of distant metastases in head and neck carcinoma. *Cancer*. 1993;71:452–6.
- Leemans CR, Tiwari R, Nauta JJ, van der Waal I, Snow GB. Recurrence at the primary site in head and neck cancer and the significance of neck lymph node metastases as a prognostic factor. *Cancer*. 1994;73:187–90.

6. Li Y, Cao X, Li H. Identification and validation of novel long non-coding RNA biomarkers for early diagnosis of oral squamous cell carcinoma. *Front Bioeng Biotechnol.* 2020;8:256.
7. Pantvaitya G, Rao K, D'Cruz A. Management of the neck in oral cancers. *Oral Oncol.* 2020;100:104476.
8. Ribeiro IP, Barroso L, Marques F, Melo JB, Carreira IM. Early detection and personalized treatment in oral cancer: the impact of omics approaches. *Mol Cytogenetics.* 2016;9:85.
9. Ribeiro IP, Caramelo F, Esteves L, Menoita J, Marques F, Barroso L, et al. Genomic predictive model for recurrence and metastasis development in head and neck squamous cell carcinoma patients. *Sci Rep.* 2017;7:13897.
10. Zhong L, Liu Y, Wang K, He Z, Gong Z, Zhao Z, et al. Biomarkers: paving stones on the road towards the personalized precision medicine for oral squamous cell carcinoma. *BMC Cancer.* 2018;18:911.
11. Altman DG. *Practical statistics for medical research.* Chapman & Hall/CRC; London, 1991.
12. Vincent-Chong VK, Salahshourifar I, Woo KM, Anwar A, Razali R, Gudimella R, et al. Genome wide profiling in oral squamous cell carcinoma identifies a four genetic marker signature of prognostic significance. *PLoS ONE.* 2017;12:e0174865.
13. Ribeiro IP, Marques F, Caramelo F, Pereira J, Patricio M, Prazeres H, et al. Genetic gains and losses in oral squamous cell carcinoma: impact on clinical management. *Cell Oncol.* 2014;37:29–39.
14. Gupta A, Ajith A, Singh S, Panday RK, Samaiya A, Shukla S. PAK2-c-Myc-PKM2 axis plays an essential role in head and neck oncogenesis via regulating Warburg effect. *Cell Death Dis.* 2018;9:825.
15. Barrow J, Adamowicz-Brice M, Cartmill M, MacArthur D, Lowe J, Robson K, et al. Homozygous loss of ADAM3A revealed by genome-wide analysis of pediatric high-grade glioma and diffuse intrinsic pontine gliomas. *Neuro Oncol.* 2011;13:212–22.
16. Jeong S, Yoo EJ, Kim JY, Han CW, Kim KJ, Kay CS. Re-irradiation of unresectable recurrent head and neck cancer: using Helical Tomotherapy as image-guided intensity-modulated radiotherapy. *Radiat Oncol J.* 2013;31:206–15.
17. Jung AC, Job S, Ledrappier S, Macabre C, Abecassis J, de Reynies A, et al. A poor prognosis subtype of HNSCC is consistently observed across methylome, transcriptome, and miRNome analysis. *Clin Cancer Res: Off J Am Assoc Cancer Res.* 2013;19:4174–84.
18. Lui VW, Hedberg ML, Li H, Vangara BS, Pendleton K, Zeng Y, et al. Frequent mutation of the PI3K pathway in head and neck cancer defines predictive biomarkers. *Cancer Discov.* 2013;3:761–9.
19. Yang J, Nie J, Ma X, Wei Y, Peng Y, Wei X. Targeting PI3K in cancer: mechanisms and advances in clinical trials. *Mol Cancer.* 2019;18:26.
20. Al-Rawi N, Merza M, Ghazi A. PIK3CB and K-ras in oral squamous cell carcinoma. A possible cross-talk! *J Orolfac Sci.* 2014;6:99–103.
21. Lin WH, Asmann YW, Anastasiadis PZ. Expression of polarity genes in human cancer. *Cancer Inform.* 2015;14:15–28.
22. Pierucci-Alves F, Yi S, Schultz BD. Transforming growth factor beta 1 induces tight junction disruptions and loss of transepithelial resistance across porcine vas deferens epithelial cells. *Biol Reprod.* 2012;86:36.
23. Kuo SJ, Chien SY, Lin C, Chan SE, Tsai HT, Chen DR. Significant elevation of CLDN16 and HAPLN3 gene expression in human breast cancer. *Oncol Rep.* 2010;24:759–66.
24. Page A, Navarro M, Suarez-Cabrera C, Bravo A, Ramirez A. Context-dependent role of IKKbeta in cancer. *Genes.* 2017;8:376.
25. Nishimura Y, Komatsu S, Ichikawa D, Nagata H, Hirajima S, Takeshita H, et al. Overexpression of YWHAZ relates to tumor cell proliferation and malignant outcome of gastric carcinoma. *Br J Cancer.* 2013;108:1324–31.
26. McGranahan N, Swanton C. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell.* 2017;168:613–28.
27. Gabusi A, Gissi DB, Montebugnoli L, Asioli S, Tarsitano A, Marchetti C, et al. Prognostic impact of intra-field heterogeneity in oral squamous cell carcinoma. *Virchows Arch.* 2020;476:585–95.
28. Ribeiro IP, Marques F, Caramelo F, Ferrao J, Prazeres H, Juliao MJ, et al. Genetic imbalances detected by multiplex ligation-dependent probe amplification in a cohort of patients with oral squamous cell carcinoma—the first step towards clinical personalized medicine. *Tumour Biol.* 2014;35:4687–95.