Article

Transcriptomic Biomarker Signatures for Discrimination of Oral Cancer Surgical Margins

Simon A. Fox [1], **Michael Vacher** [2] **and Camile S. Farah** [1,*]

1 Australian Centre for Oral Oncology Research & Education, P.O. Box 285, Nedlands, WA 6909, Australia; simon@oralmedpath.com.au
2 Australian eHealth Research Centre, The Commonwealth Scientific & Industrial Research Organisation, Floreat, WA 6104, Australia; michael.vacher@csiro.au
* Correspondence: camile@oralmedpath.com.au

Abstract: Relapse after surgery for oral squamous cell carcinoma (OSCC) contributes significantly to morbidity, mortality and poor outcomes. The current histopathological diagnostic techniques are insufficiently sensitive for the detection of oral cancer and minimal residual disease in surgical margins. We used whole-transcriptome gene expression and small noncoding RNA profiles from tumour, close margin and distant margin biopsies from 18 patients undergoing surgical resection for OSCC. By applying multivariate regression algorithms (sPLS-DA) suitable for higher dimension data, we objectively identified biomarker signatures for tumour and marginal tissue zones. We were able to define molecular signatures that discriminated tumours from the marginal zones and between the close and distant margins. These signatures included genes not previously associated with OSCC, such as *MAMDC2*, *SYNPO2* and *ARMH4*. For discrimination of the normal and tumour sampling zones, we were able to derive an effective gene-based classifying model for molecular abnormality based on a panel of eight genes (*MMP1*, *MMP12*, *MYO1B*, *TNFRSF12A*, *WDR66*, *LAMC2*, *SLC16A1* and *PLAU*). We demonstrated the classification performance of these gene signatures in an independent validation dataset of OSCC tumour and marginal gene expression profiles. These biomarker signatures may contribute to the earlier detection of tumour cells and complement existing surgical and histopathological techniques used to determine clear surgical margins.

Keywords: oral squamous cell carcinoma; surgical margins; gene expression profiling; gene biomarker signature; multivariate statistics

1. Introduction

Oral squamous cell carcinoma (OSCC) is the most common and deadly head and neck cancer (HNC), affecting more than 750,000 people worldwide [1]. In the last three decades, diagnostic progress in OSCC has been limited, and overall patient survival has improved by less than 10%, with 5-year survival rates of 8–27% reported for patients with metastatic disease [2,3]. Therapeutic protocols for OSCC include surgery, radiotherapy and chemotherapy; however, surgical resection of tumours remains the cornerstone of the initial definitive management of OSCC. The outcomes of surgical resection of OSCC have long been affected by limitations in the evaluation of surgical margins using pre-operative white light visual examination, palpation of the lesion and microscopic evaluation of frozen sections at a histopathological level. Regional failure and distant tumour recurrence continue to account for the high mortality rate of OSCC patients. Local relapse rates of 20% have been reported for oral cancer cases with confirmed tumour-free "non-involved" surgical margins defined by traditional examination [4,5]. This local relapse may occur due to minimal residual disease from incomplete resection that is not detected by routine histopathological investigation or premalignant tissue that gives rise to a later tumour through field cancerisation [6] or through independent multiclonal expansion of precursor tissue [7]. The ability to detect residual abnormality in oral cancer surgical margins

using highly sensitive molecular techniques to assay for specific biomarkers would offer a superior assessment of surgical performance and inform decision-making regarding postoperative treatment. Once confirmed, these markers may be used for real-time in vivo endomicroscopic assessments of tumour margins, enabling live molecular imaging. Such interventions have the potential to guide the surgical removal of tumours and improve patient outcomes by minimising the retention of abnormal tissue within the surgical bed.

Many different statistical methods have been applied to the identification of biomarkers; however, generally, variables are evaluated for statistically significant behaviour between sample groups using classical statistical tests by focussing upon candidate biomarkers individually. When working with highly complex datasets such as those generated by transcriptome analysis, classification techniques can be extremely powerful in the identification of biomarkers [8]. In contrast to classical methods, multivariate methods incorporate the correlation structure of the data and accommodate interactions among candidate biomarkers. This approach has the potential to develop biomarker panels with superior prognostic performance to individual biomarkers with regards to sensitivity, specificity and reliability [8]. Partial least squares-discriminant analysis (PLS-DA) is a multivariate regression method that has been widely used for classification purposes and biomarker discovery in large datasets such as metabolomics [9], proteomics [10] and transcriptomics [11]. Sparse PLS-DA (sPLS-DA) is an extension of the technique that improves the application of classification and biomarker discovery by providing automatic variable selection [12].

Previously, we have reported that the margins determined using Narrow Band Imaging (NBI) harbour less molecular abnormality than those derived using conventional white light techniques [13,14]. In these studies, we have generated global mRNA and microRNA (miRNA) expression data for tumour (T), conventional margins determined by white light examination (M) and a more molecularly distant margin determined by NBI (N). We know from our own data [13] and from clinical evidence of local recurrence [5] that tumour-adjacent tissues may harbour significant aberrations. Here, we use the more distant NBI-defined surgical margins as the control tissue to build a predictive biomarker panel for molecular abnormality in the tumour margin. We apply sPLS-DA to develop classification models for the discrimination of tumour margins and the identification of molecular abnormality in the margins. More importantly, we demonstrate the generalisation of our biomarker statistical models by validation in an external dataset.

2. Materials and Methods

2.1. Patient Group and Sample Collection

The details of the patient group and sample collection have been previously described [13,15]. Briefly, eighteen patients with oral cavity SCC were enrolled prospectively. Patient demographics, tumour characteristics and surgery have been described by us previously [13,15]. The study conformed to the principles outlined in the Declaration of Helsinki (2008), with ethical approval from the Royal Brisbane and Women's Hospital Human Research Ethics Committee (HREC/08/QRBW20 and HREC/10/QRBW336). All subjects provided written informed consent to participate in the study. Prior to surgery, the OSCC sites were visualised using conventional white light illumination and then Narrow Band Imaging (NBI) using an Olympus NBI ENF-VQ nasendoscope with CLV-180 light source and processor (Olympus Medical Systems Corp., Tokyo, Japan). The OSCC resection was taken to ≥5 mm beyond the NBI-defined surgical margin; following which, 4-mm punch biopsies were taken from: (1) 5 mm beyond the limit of tissue abnormality visible by NBI (defined here as normal (N)), 5 mm beyond the limit of tissue abnormality visible by conventional white light (defined here as the margin (M)) and the core of the primary tumour (defined here as T). The relative locations of the sampling sites have been illustrated by us previously (Figure 1 in Reference [13]). The tumour histology was predominantly moderate (10)-to-well (7)-differentiated OSCC, with one case of verrucous carcinoma. Independent histopathology indicated "clear" margins for all but one resection, although only

two had margins with >5 mm separation from disease. One "close" margin had just 1 mm of separation. Applying this approach, five-year disease-free survival (DFS) was 84.21%, and the local recurrence rate (LRR) was 5.26% [15].
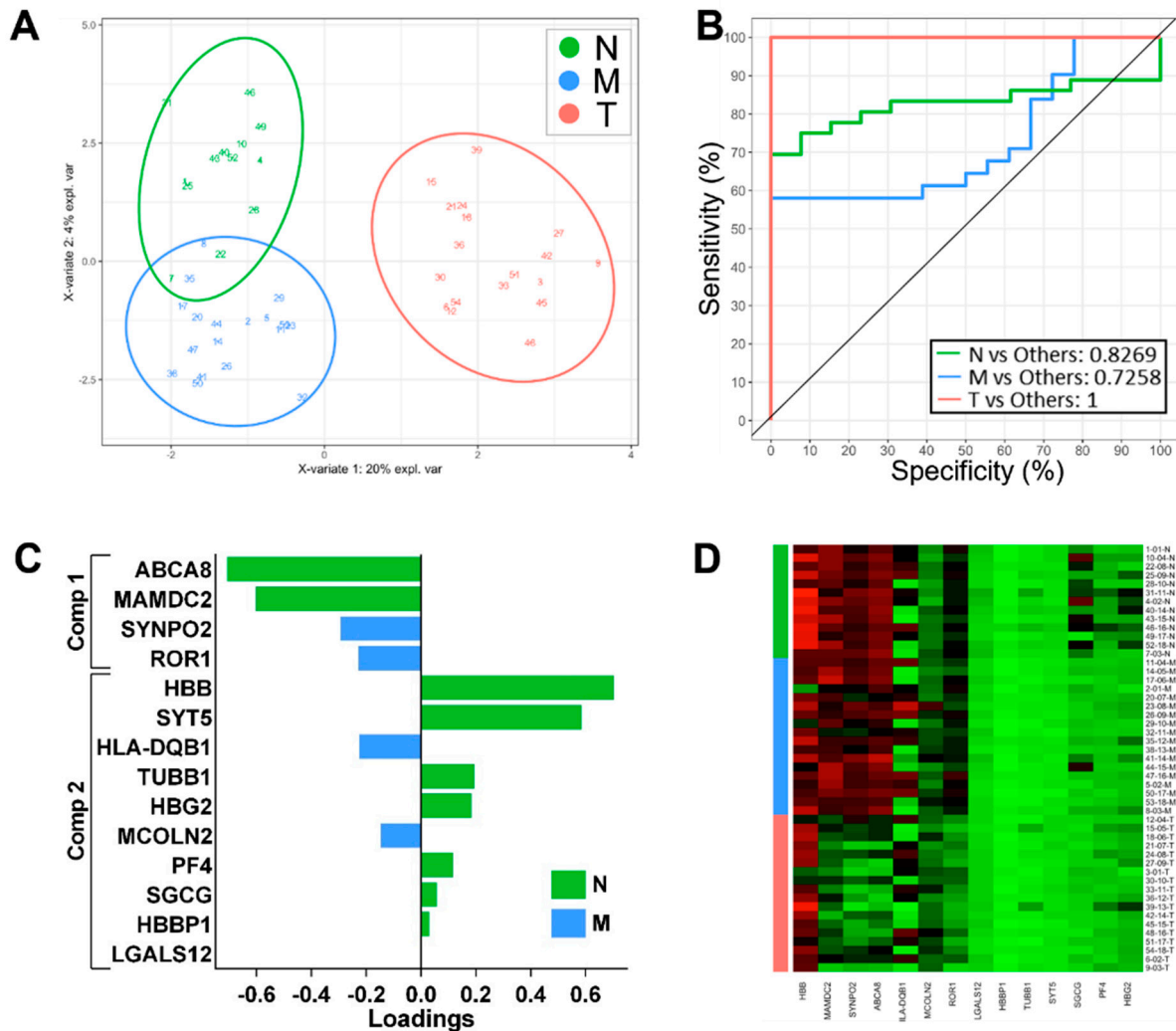

Figure 1. Classification performance sPLS-DA-derived gene expression signatures on all sample groups: N vs. M vs. T. (**A**) Two-dimensional sample plots for the multinomial classifier for 3 groups with confidence ellipse plots. (**B**) Corresponding ROC curves. (**C**) Loading plots for multinomial N vs. M vs. T classifiers showing the contribution of each gene to the signature. (**D**) Heatmap of the gene expression across all samples for the multinomial N vs. M vs. T classifiers.

2.2. Gene Expression Profiling and Bioinformatics

The detailed protocols for gene and miRNA expression profiling have been previously described by us [13,14]. The processed transcriptome and miRNA expression matrices from these studies were used in the present study and are available at ArrayExpress: E-MTAB-5933 and E-MTAB-6470. Briefly, mRNA expression data was obtained using Human Genome U133 Plus 2.0 Arrays (Affymetrix, Santa Clara, CA, USA). Following quality control filtering, the remaining array data was pre-processed using the GCRMA normalisation method [16]. Individual mRNA probes were mapped to genes using the manufacturer's annotation data. In cases where multiple probes were associated with a single gene, the probe with the highest Median Absolute Deviation was kept. Additionally, probes not located within a gene were discarded from the analysis. The functional annotation of the genes was performed using the DAVID [17] and Harmonizome [18] databases.

2.3. Sample Classification and Feature Selection

To identify the classifying gene (or miRNA) expression signatures characteristic of different tissue zones (N, M and T), we performed sPLS-DA [12], as implemented in the mixOmics R package [19]. This method combines sample classification and feature selection in a one-step process, allowing the identification of key variables (i.e., genes) specific to each group (tissue zone). To maximise the classification performance, we identified the optimal number of components and number of variables to select per component using the "*tune*" method implemented in the mixOmics package. Briefly the "*tune*" function uses a grid search approach to test multiple combinations of parameters (number of components and number of variables selected on each component) and identify the set of parameters providing the lowest error rate. Specifically, we used this function to perform a 10-fold cross-validation, repeated 10 times and with a range from 5 to 100 variables per component. The set of parameters providing the lowest Balanced Error Rate (BER) were retained to fit the final models. The performance and discriminative ability of the final classifier were assessed using the Area Under the Curve (AUC) approach, averaged over 10 cross-validations using one vs. all comparison. The plotting functions of the mixOmics package were used to generate all plots except the loading plots, which were generated from the mixOmics output using Prism v 6.07 (GraphPad Software).

2.4. External Validation Dataset

An external gene expression dataset was used for additional validation of the classifying signature biomarkers downloaded from the GEO database (https://www.ncbi.nlm.nih.gov/geo/ accessed on 11 June 2021): GSE31056 [20], from which data was obtained for oral cavity OSCC tumours, surgical margins and adjacent normal tissue for 23 patients from a Canadian population. This dataset was generated using the same Human Genome U133 Plus 2.0 Array platform used in our study.

3. Results

3.1. Differential Gene Expression Profiling

The results of the differential gene and miRNA expression analysis from the three pairwise comparisons between the N, M and T sample groups have been previously reported by us [13,14]. In brief, there were a total of 4794 genes differentially expressed with individual comparisons showing 4387 genes for T vs. N, 3266 genes for T vs. M and 7 genes for N vs. M. Similarly, there were a total of 119 miRNA differentially expressed with individual comparisons showing 109 miRNAs for T vs. N, 81 miRNAs for T vs. M and 7 miRNAs for N vs. M. Thus, as we have previously reported, the N samples were more molecularly distinct from the tumours than the M samples, supporting the use of an adjunctive optical imaging approach for surgical resection of the oral cavity SCC [15].

3.2. Identification of Classifying Biomarkers for Tumour, Margin and Normal Tissues

Initially, we explored whether we could derive classifying features in the gene expression data using sPLS-DA for multinominal discrimination across all three tissue locations to assess the classifying performance and identify genes that may be of biological significance. Following preliminary modelling using the dataset of differentially expressed genes and tuning to identify a classifying subset, we derived a set of genes with optimal performance for distinguishing the three groups. The classifying performance of this model is presented in Figure 1A,B based on the expression of 14 genes, which were independently weighted within the model (Figure 1C). Expression of these genes across all sample zones is illustrated in Figure 1D. Using this model, discrimination of the T samples was highly effective; however, discrimination of the normal (N) and margin (M) samples was less effective (Figure 1A,B). Therefore, we removed the T samples and repeated the analysis to derive a binomial classifier for the remaining samples and reveal molecular differences between the normal and peritumoral tissue. The resulting binomial model consisted of 20 genes (Figure 2C) with good classifying performance for discriminating N

and M samples (Figure 2A,B). This analysis indicates that marginal tissue, while histologically tumour-free, can be readily distinguished from the more distant normal tissue in our dataset using this panel of 20 gene expression biomarkers.
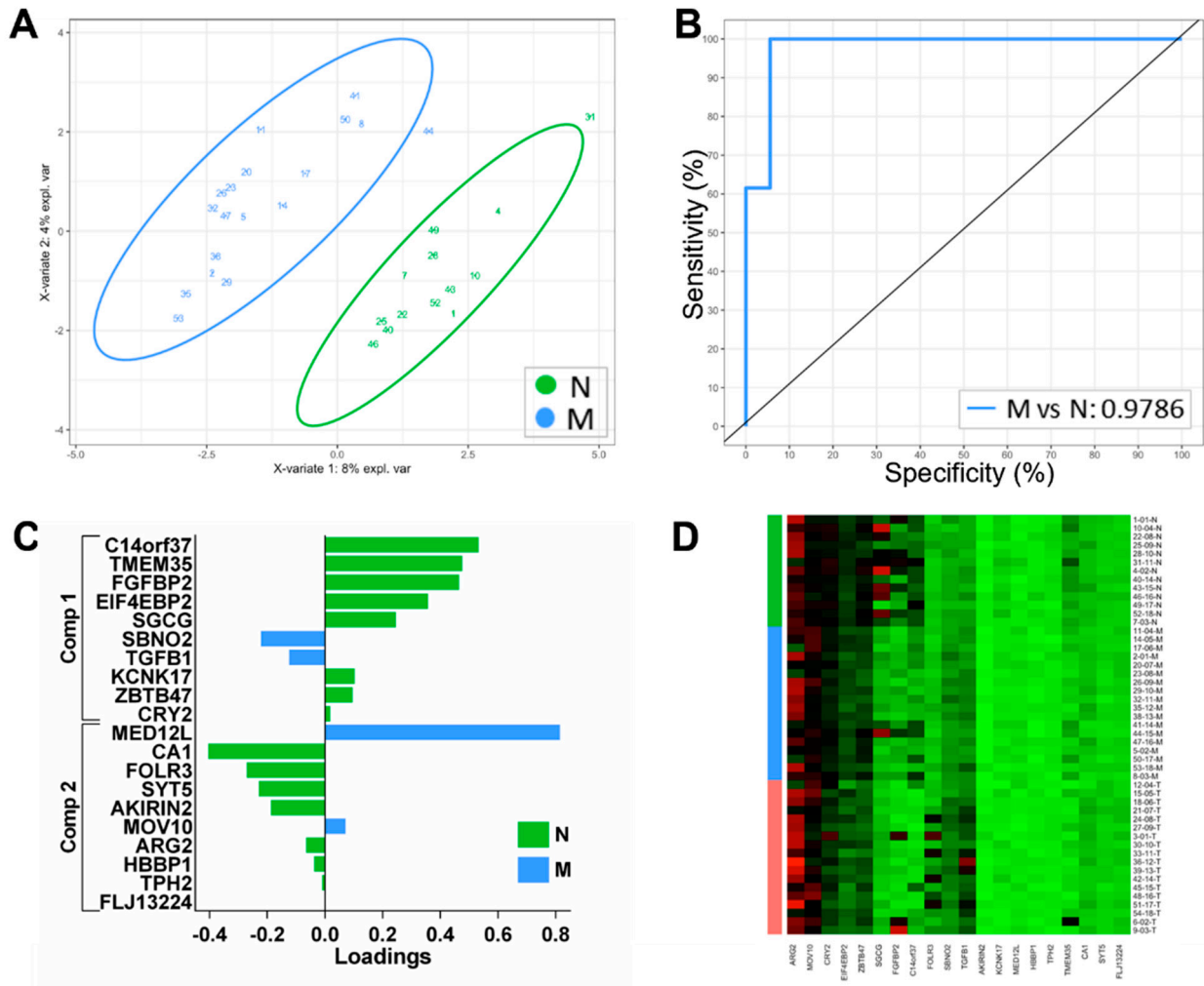

Figure 2. Classification performance sPLS-DA-derived gene expression signatures for binomial comparison: N vs. M. (**A**) Two-dimensional sample plots for the 2 groups with confidence ellipse plots. (**B**) Corresponding ROC curve. (**C**) Loading plots for N vs. M classifiers showing the contribution of each gene to the signature. (**D**) Heatmap of the gene expression across all samples for the N vs. M classifiers.

In addition to the gene expression dataset, we also analysed the comprehensive microRNA expression data available for the same samples that we explored as another source of classifying features across the three tissue groups using sPLS-DA. Using a similar strategy, we identified differentially expressed miRNA by pairwise comparisons between the sample groups and derived a classifying model based upon miRNA expression alone. We found that the classification performance of the miRNA expression-based model (Figure 3A–C) was broadly inferior to the gene expression model, especially for the T and M zones (Figure 1A). We also examined the performance of elastic net regression [21] for the identification of biomarkers and sample classification using the gene expression dataset. When we ran this process for multinomial classification of the N, M and T samples using a training set consisting of two-thirds of the dataset, we were unable to identify a subset of genes optimal for building a prediction model, and we consistently found that elastic net regression gave

an inferior discriminating performance to sPLS-DA. As a consequence, we focussed on the application of sPLS-DA and the use of mRNA-based signatures only.
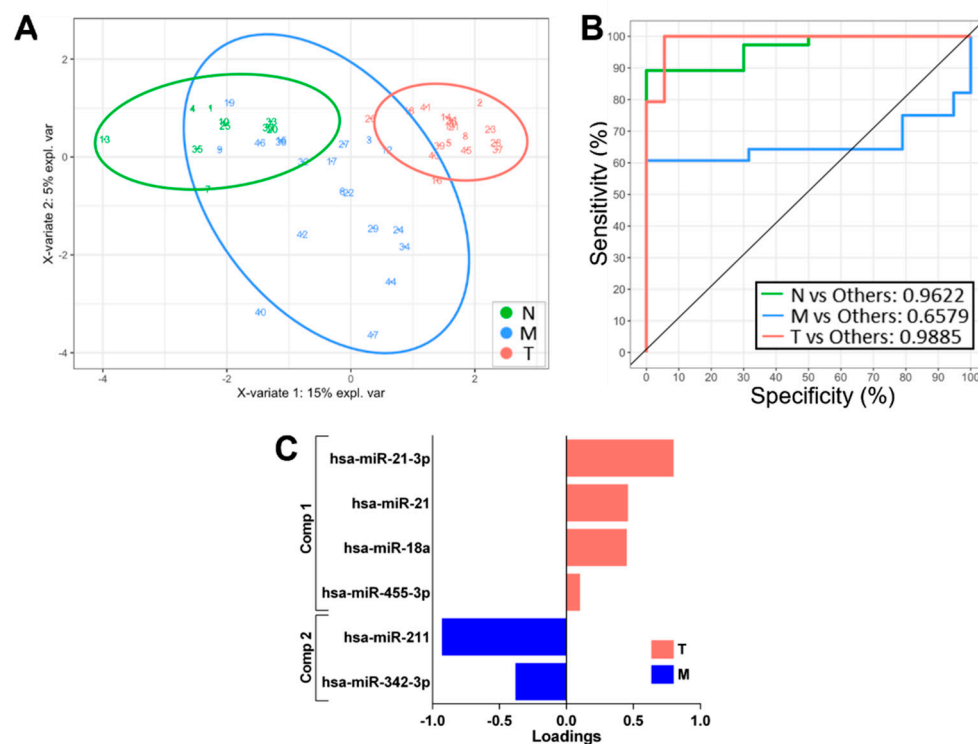

Figure 3. Classification performance of a sPLS-DA-derived miRNA expression signature. (**A**) Two-dimensional sample plots for the 3 groups with confidence ellipse plots. (**B**) Corresponding ROC curve. (**C**) Loading plots for multinomial N vs. M vs. T classifiers showing the contribution of each miRNA to the signature.

3.3. Development of a Biomarker Signature for Classification of Tumour Margins

In order to develop discriminating biomarkers able to identify residual abnormality in OSCC surgical margins, our strategy was to first identify classifier genes that were upregulated in tumour tissue and could reliably be applied to distinguish tumour tissue from normal tissue. These biomarkers could then be used to interrogate the corresponding peritumoral margin (M) biopsy samples.

We used sPLS-DA to derive a classifying gene expression panel that discriminated T from N samples using the subset of genes that were significantly upregulated in tumour tissue. For this T vs. N analysis, cross-validation tuning of the parameters showed that the best discriminating performance using sPLS-DA occurred when eight genes were selected (Figure 4A). Following identification of the optimal values, the methods were applied to the entire dataset of 31 samples (18 T and 13 N), and a list of genes for binomial classification were obtained. For sPLS-DA, the eight genes of the model and their weightings are shown in Figure 4A, and as expected, their classification performance for T vs. N was excellent (Figure 4B). This gene signature sPLS-DA model was applied to the 18 marginal samples, where 13 samples were predicted as normal while five were predicted as tumours (Figure 4C). Some predictions were unambiguous (e.g., samples 53-18-M as T and 50-17-M as normal), while others were more borderline (e.g., samples 35-12-M or 29-10-M as T), as shown in the prediction score plot (Figure 4C). The expression of these classifiers in all samples across the entire dataset is illustrated in Figure 4D. This gene signature and the associated algorithms represent tools that can be applied toward the classification of tissues.
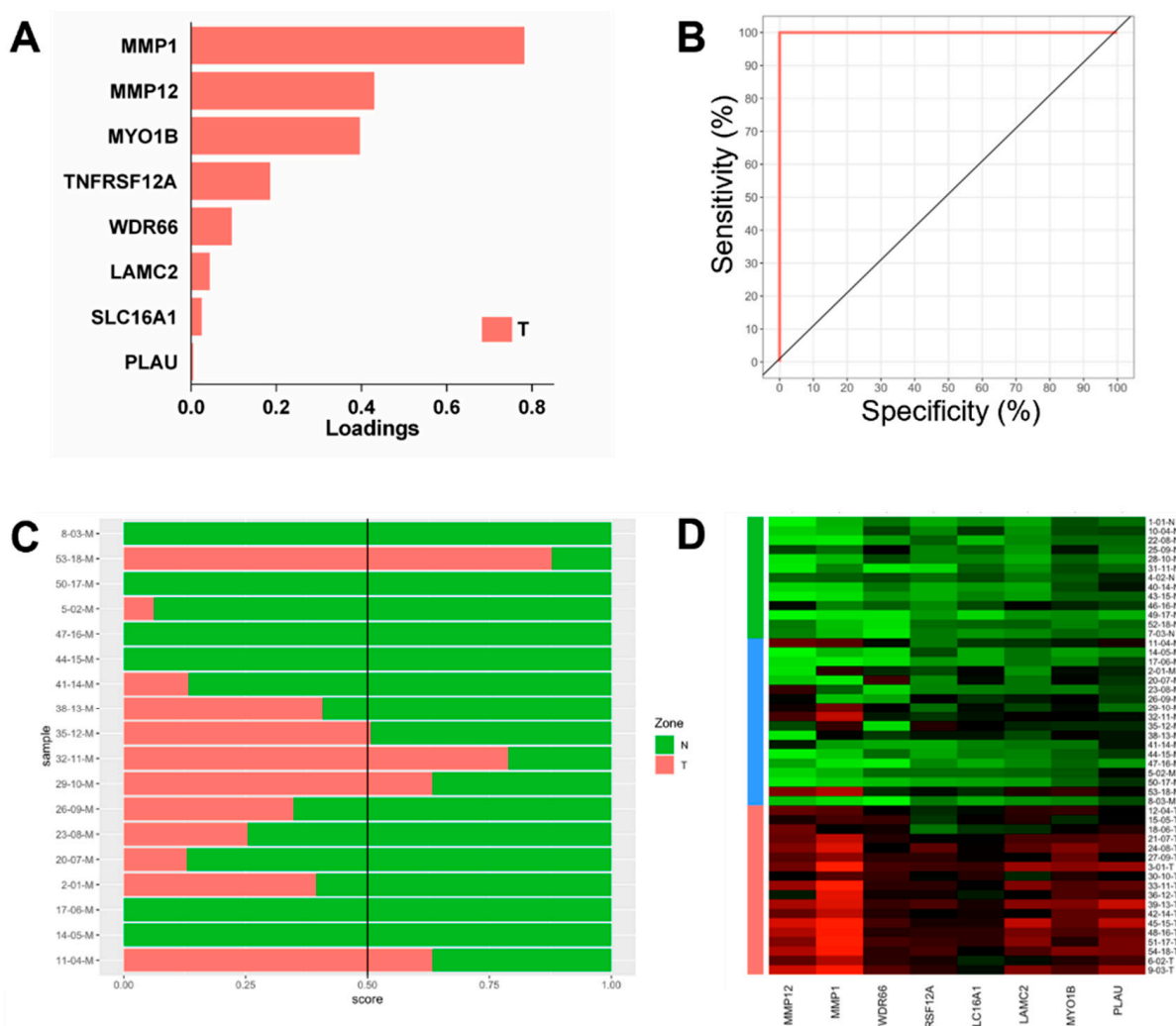
Figure 4. Classification performance of sparse Partial Least-Squares Discriminant Analysis-derived gene signatures for tumour vs. normal (T vs. N) using genes upregulated in tumours. (**A**) Loading plot of a discriminating gene panel. (**B**) ROC curve for tumour vs. normal discrimination. (**C**) Prediction score plot for the prediction of margins classed as T or N. (**D**) Heatmap of gene expression for the T vs. N gene models in all the samples.

3.4. Validation in External Datasets

In order to test the generalisability of our model and investigate its clinical applicability, we performed validation using an independent OSCC dataset of tumour, marginal and normal samples [20]. Initially, we explored this dataset using the model trained with all three zones and all significant expressed genes as derived in Figure 1A. We applied this model to the validation dataset, with the resulting confusion matrix of the classifying performance shown in Figure 5A. For this dataset derived from a Canadian OSCC population, the multinomial three-way classifier performed well for the classification of tumour (T) samples but performed poorly for discrimination between normal and margin samples (Figure 3A). This may reflect difficulties in classifying these samples or differences between the sampling strategies between the studies, since our own study used NBI to define the N zone for sampling [13]. We next used our binomial classifying model for T vs. N, as shown in Figure 4A, and then applied it to all marginal (M) samples in the validation dataset. For this population, there was gene expression data for multiple marginal zone samples for most patients, and we applied our classifier to all these samples. Expression of the classifying gene panel in all the samples in the validation data (N, M and T) is visualised by

a heatmap in Figure 5B. Of 49 M samples from 24 patients, 13 were classified as tumours (T) by our classifier (Figure 5C, prediction score 0.5 or greater). We next examined whether the presence of molecular abnormality in the margins as classified by our model may have value in the prediction of recurrence. We stratified the patients in the independent cohort, with at least one margin predicted as T in our analysis, into a "high-risk" group. We then performed a survival analysis, which showed that the differences between the groups were not statistically significant (Figure 5D). These results demonstrated that, while our classifying signatures were effective in identifying tumour samples in the independent cohort, the statistical model signature of abnormality in the margins could not be used to discriminate the samples associated with recurrence.
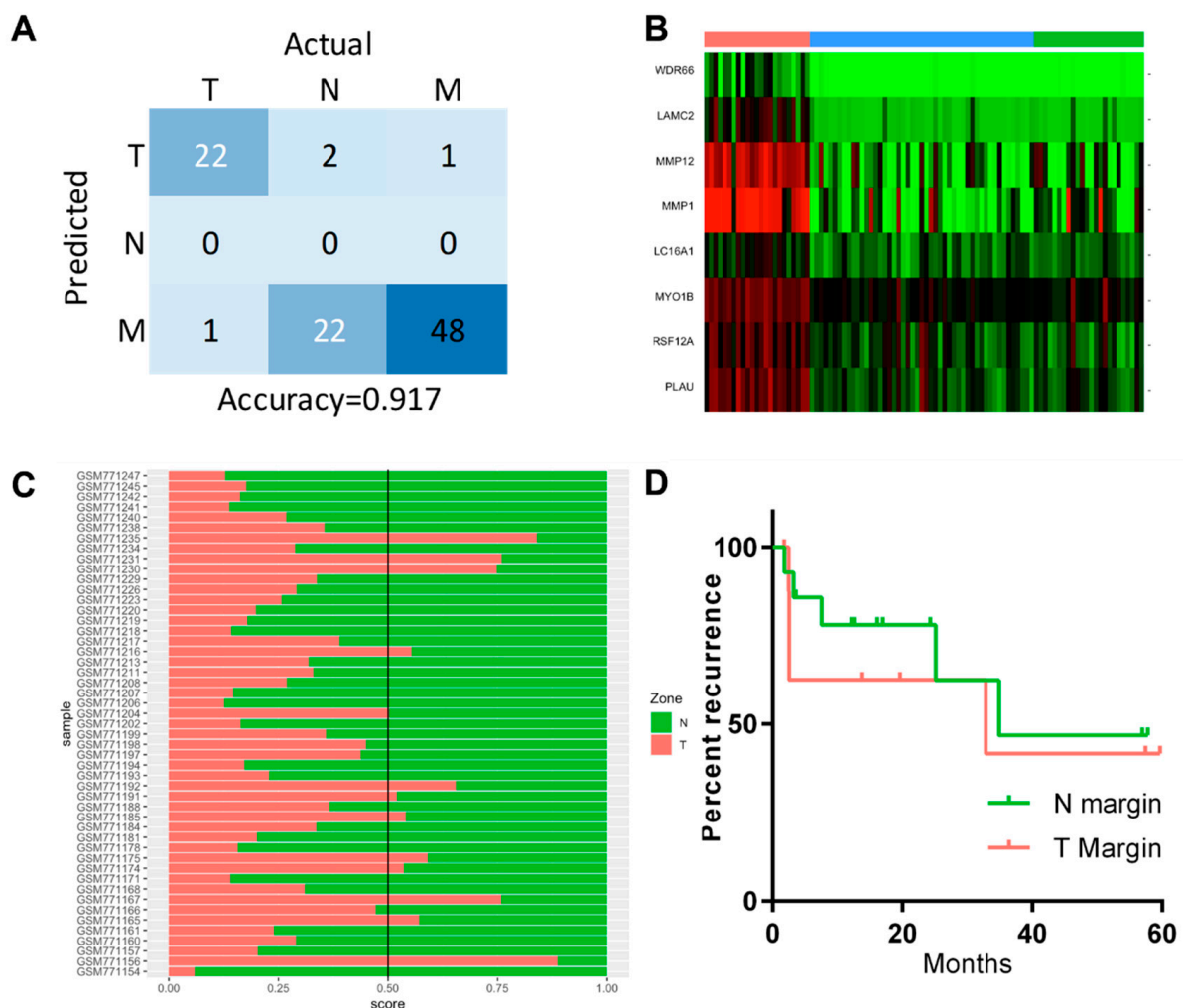

Figure 5. Testing sPLS-DA models in the validation data. (**A**) Confusion matrix of 3 zone predictions of the validation dataset using the classifiers from Figure 1. (**B**) Heatmap of the expression of T vs. N prediction model classifiers in the validation dataset. (**C**) Prediction score plot of the validation margins using the binary classifier for T vs. N in the validation data. (**D**) Survival plot for tumour recurrence for patients with primary margins predicted as T or N.

4. Discussion

Several studies have investigated transcriptional dysregulation and disease-specific signatures in OSCC in the context of investigating carcinogenic processes, prognostic signatures and molecular targets [22,23]. Inadequate surgical removal of the primary tumour is one important factor associated with patient survival [24]. The presence of residual disease or epithelial dysplasia in the surgical resection margins is recognised as a

contributor to disease recurrence, and molecular analysis is one approach that has been explored to address this challenge [6,25].

In this study, we investigated a multivariate statistical method for its ability to identify classifying gene signatures for OSCC relative to normal tissue and for identifying biological features that distinguish OSCC margins. We also sought to examine the application of this methodology in the detection of abnormality in surgical margins. We were interested in exploring the application of sPLS-DA, since this relatively recently described statistical approach has demonstrated advantages in classification from high-dimensional data but has not as yet been widely applied to tumour classification and biomarker derivation [12]. Our strategy used paired tissue samples from the same patient group, since that molecular context should be more informative for the investigation of abnormality in surgical margins, rather than normal tissue from unrelated patients. The methodological approaches to the identification of classifying gene signatures in OSCC to date have predominantly employed differential gene expression and focussed upon genes that are overexpressed in tumour tissue using univariate statistical methods.

Since classifying biomarkers derived from discovery-based analyses are frequently significant contributors to disease processes [8], we initially applied sPLS-DA to identify features that discriminated between the three tissue groups. Interestingly, it was the downregulation of a panel of genes expressed in N and M zones that was most effective at classifying T samples (Figure 1C), suggesting that the dysregulation of particular tumour suppressors is a more common feature for classifying tumour tissue. Both *MAMDC2* and *SYNPO2* have been characterised as tumour suppressors in solid cancers, although not previously described in OSCC [26,27], and *ABCA8* has been reported downregulated in OSCC [28]. In the tumour zone, we found the downregulation of *ROR1*, a molecule that has been associated with tumour progression and proposed as a therapeutic target in a number of cancers other than HNSCC [29]. Our findings of lower expression in OSCC tumour tissue are consistent with another recent study [30] and suggest a different role in this cancer. Furthermore, there is evidence showing *ROR1* is expressed in a number of normal tissues, including oesophageal epithelium [31], indicating that targeted therapies may result in toxicity in normal tissues, including oral epithelia, according to our expression data. These classifiers showed excellent discrimination of the T samples but were less effective at discriminating N and M. We speculated that the significant molecular differences between T and the other zones influenced the derivation of the biomarkers, and hence, we performed an analysis of the N and M samples only. In this way, we sought to explore the molecular differences that might differentiate marginal tissue from more distant normal sites to gain insights into genes' driving biology in peritumoral tissue. We were able to generate a panel of genes that discriminated these zones. It should be noted that such markers and molecular differences can reflect both the potential for residual tumour tissue within the margin and also the influence of soluble tumour-derived factors upon the marginal cells and their compositions. Many of the classifying genes for the classification of normal and marginal tissues have not previously been described with reference to OSCC but are dysregulated in other cancers. *ARMH4* (C14orf37, UT2) has been implicated as a tumour suppressor in myeloma and was found here downregulated in margins relative to the normal tissue [32]. *FGFBP2* is an effector of cytotoxic cell activity and potentially antitumour immunity [33], and the downregulation seen in the marginal tissue may reflect the suppression of antitumour immunity.

For discrimination of the normal and tumour sampling zones, we were able to derive an effective gene-based classifying model for molecular abnormality based on a panel of eight genes (*MMP1, MMP12, MYO1B, TNFRSF12A, WDR66, LAMC2, SLC16A1* and *PLAU*). Since we ultimately aimed for biomarkers that could be used to detect abnormality in marginal tissues, we focussed upon genes upregulated in the tumour tissue. Many of the genes identified in the resulting classifying biomarkers have been previously identified as dysregulated in cancer, including *MMP1* and *MMP12*. Matrix metalloproteinases like *MMP1* and *MMP12* are key enzymes involved in extracellular matrix degradation during

tissue remodelling and are frequently reported as upregulated in cancer, associated with invasion and metastases and potential therapeutic targets [34]. In OSCC, *MMP1* has previously been reported as overexpressed in tumours relative to normal tissue [28,35] and proposed as a biomarker of disease recurrence [20,36]. *MMP12* has been reported as upregulated in tongue SCC [28]. The upregulation of *MYO1B* has been associated with lymph node metastases and cellular invasion in HNSCC [37].

The immune-related gene *TNFRSF12A* has been significantly associated with overall survival (OS) in HNSCC, with the high-protein expression of TNFRSF12A correlated with poor OS [38]. *TNFRSF12A* was also significantly correlated with activated CD8+ T cells, CD4+ T cells, follicular helper T cells and regulatory T cells, implying the possible involvement of this gene in immunoregulation and the development of HNSCC.

PLAU (plasminogen activator urokinase) mRNA has also been shown to be significantly elevated in HNSCC tumour samples over normal specimens, with evidence for *PLAU* as an independent indicator for HNSCC prognosis [39] associated with poorer clinical outcomes [40]. Furthermore, *PLAU* expression has been positively correlated with the presence of M1-type macrophages and negatively associated with CD4+ T cells, Treg cells and follicular helper T cells. Intriguingly, this finding may be explained by evidence that *PLAU* can regulate the expression of *TNFRSF12A* [40], previously implicated in immune regulation in HNSCC [38]. Collectively, these results suggest that *PLAU* may be an independent biomarker for predicting the outcomes of HNSCC patients and potentially contribute to tumour immunosuppression either in concert with or via *TNFRSF12A* [39].

The laminin C gene family is associated with the progression of diverse cancers through epithelial-mesenchymal transition (EMT). *LAMC2*, which encodes the laminin subtype LN-332 γ2 chain, is associated with poor prognosis, infiltrating immune cells and higher levels of LAMC2 protein expression, particularly in tumour-budding areas in HNSCC contributing to proliferation, migration, invasion and metastasis [41]. *LAMC2*, a target gene of miR-134, activates the EGFR/MAPK pathway but is also downregulated by miR-134, thereby inhibiting the migration and invasion of OSCC tumour stem cells by suppressing the PI3K-Akt signalling pathway [42]. Not only is *LAMC2* a good prognostic biomarker candidate, but it also shows utility as an early diagnostic marker for OSCC [43]. Previously, we examined miRNA-mRNA interactions in this data and reported 100 putative miRNA–mRNA interactions between 40 miRNA and 96 mRNA [14]. We did not find the dysregulation of miR-134 in our data, suggesting that some other mechanism is responsible for the overexpression of *LAMC2* that we see here.

SLC16A1 encodes the monocarboxylate transporter MCT1, a member of a family involved in transporting lactate into and out of tumour cells, serving as a key mediator in maintaining tumour microenvironment homeostasis [44]. Recently, *SLC16A1* has been shown to be upregulated in HNSCC tumour tissues compared to adjacent normal marginal tissues, emphasising its potential oncogenic role in cancer progression likely through the NF-κB pathway [45]. Additionally, *SLC16A1* expression was closely associated with smoking and drinking behaviours in HNSCC patients. Interestingly, *MMP1*, *LAMC2* and *SLC16A1* have been shown to be significantly upregulated in HNSCC but downregulated by S100A8/A9 expression, contributing to increased proliferation, malignant transformation and disease progression in HNSCC [46].

There has been limited investigation of the role of *WDR66* (WD-repeat-containing protein 66) in HNSCC. WD-repeat proteins are frequently upregulated in various cancers, impacting upon transcription regulation, cell cycle control and apoptosis. *WDR66* was reported as upregulated in oesophageal SCC, associated with poor overall patient survival and a driver of EMT, identifying *WDR66* as a novel marker for risk stratification [47]. In salivary adenoid cystic carcinoma, the knockdown of *WDR66* decreases cellular proliferation, migration and invasion [48].

The key test of such a model is its applicability across independent populations, and we selected datasets from international populations to test the generalisability of this model. While there are a number of studies that have reported gene expression data for tumour and

matched normal tissues, the sampling sites for normal tissues are not always well-defined, and studies that have reported tumour, peritumoral margins and more distant normal tissues are limited. We were unable to find a comparable study to our own, which used NBI to define the sampling zones; however, we used an independent cohort of Canadian OSCC patients [20] with three similar sampling zones to benchmark our classifying models. The authors of this study focussed their investigation upon gene expression in marginal tissues associated with recurrence. Although there is some ambiguity in the definition of sample zone locations, we found that our three-zone classifying model was highly accurate in the prediction of T samples but classified all N and M samples as marginal. This may reflect differences in the sampling of N tissues between the studies given the use of NBI in our study or the relatively inferior prediction accuracy for N and M samples that we found in our own dataset using this model. We speculated that the classification of marginal samples using our binary classifier of (T vs. N) might have application in the prediction of recurrence. In our own primary dataset, such an analysis was not possible, because recurrence was extremely low due to the application of NBI for margin determination [15]. In the external dataset for which recurrence data was available, we found that the margin classifying tumour signature was not suitable for the prediction of recurrence in this cohort. It seems possible that the identification of tumour signatures in marginal tissue is insufficiently discriminating for the prediction of recurrence, and there may be distinct marginal gene expression signatures predictive of this. Moving forward, refinement of these biomarker panels will require testing in additional cohorts with the well-defined sampling of marginal tissue and populations with different genetic backgrounds. Furthermore, in some Asian populations, there are alternative risk factors such as betel nut chewing that have different carcinogenic mechanisms and that may impact the relevance of some biomarkers. As much as possible, these populations need to be included early in follow-up studies to ensure the broad applicability of biomarker panels in clinical practice.

In conclusion, we investigated the performance of multivariate statistical methods in the discovery of diagnostic biomarker panels for OSCC from transcriptomic data. The novel classifying gene expression signatures derived here will contribute to the ongoing development of diagnostic and prognostic biomarker panels for OSCC. In addition, many of the classifying genes have not been previously reported for OSCC and represent promising putative targets for further molecular analysis.

Author Contributions: C.S.F. conceived and designed the study. C.S.F., S.A.F. and M.V. performed the data acquisition and analysis. S.A.F. and M.V. performed the bioinformatics. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: The study conformed to the principles outlined in the Declaration of Helsinki (2008), with ethical approval from the Royal Brisbane and Women's Hospital Human Research Ethics Committee (HREC/08/QRBW20 and HREC/10/QRBW336).

Informed Consent Statement: All subjects provided written informed consent to participate in the study.

Data Availability Statement: Microarray data raw files and normalised expression data are available on ArrayExpress EMBL-EBI under database accession numbers E-MTAB-5933 and E-MTAB-6470.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gupta, B.; Johnson, N.W.; Kumar, N. Global Epidemiology of Head and Neck Cancers: A Continuing Challenge. *Oncology* **2016**, *91*, 13–23. [CrossRef] [PubMed]
2. Cronin, K.A.; Lake, A.J.; Scott, S.; Sherman, R.L.; Noone, A.M.; Howlader, N.; Henley, S.J.; Anderson, R.N.; Firth, A.U.; Ma, J.; et al. Annual Report to the Nation on the Status of Cancer, Part I: National Cancer Statistics. *Cancer* **2018**, *124*, 2785–2800. [CrossRef] [PubMed]

3.   Gatta, G.; Botta, L.; Sánchez, M.J.; Anderson, L.A.; Pierannunzio, D.; Licitra, L.; on behalf of the EUROCARE Working Group. Prognoses and Improvement for Head and Neck Cancers Diagnosed in Europe in Early 2000s: The EUROCARE-5 Population-Based Study. *Eur. J. Cancer* **2015**, *51*, 2130–2143. [CrossRef]

4.   Binahmed, A.; Nason, R.W.; Abdoh, A.A. The Clinical Significance of the Positive Surgical Margin in Oral Cancer. *Oral Oncol.* **2007**, *43*, 780–784. [CrossRef] [PubMed]

5.   González-García, R.; Naval-Gías, L.; Román-Romero, L.; Sastre-Pérez, J.; Rodríguez-Campo, F.J. Local Recurrences and Second Primary Tumors from Squamous Cell Carcinoma of the Oral Cavity: A Retrospective Analytic Study of 500 Patients. *Head Neck* **2009**, *31*, 1168–1180. [CrossRef]

6.   Braakhuis, B.J.M.; Bloemena, E.; Leemans, C.R.; Brakenhoff, R.H. Molecular Analysis of Surgical Margins in Head and Neck Cancer: More than a Marginal Issue. *Oral Oncol.* **2010**, *46*, 485–491. [CrossRef]

7.   Wood, H.M.; Daly, C.; Chalkley, R.; Senguven, B.; Ross, L.; Egan, P.; Chengot, P.; Graham, J.; Sethi, N.; Ong, T.K.; et al. The Genomic Road to Invasion-Examining the Similarities and Differences in the Genomes of Associated Oral Pre-Cancer and Cancer Samples. *Genome Med.* **2017**, *9*, 53. [CrossRef]

8.   McDermott, J.E.; Wang, J.; Mitchell, H.; Webb-Robertson, B.J.; Hafen, R.; Ramey, J.; Rodland, K.D. Challenges in Biomarker Discovery: Combining Expert Insights with Statistical Analysis of Complex Omics Data. *Expert Opin. Med. Diagn.* **2013**, *7*, 37–51. [CrossRef]

9.   Szymańska, E.; Saccenti, E.; Smilde, A.K.; Westerhuis, J.A. Double-Check: Validation of Diagnostic Statistics for PLS-DA Models in Metabolomics Studies. *Metab. Off. J. Metab. Soc.* **2012**, *8*, 3–16. [CrossRef]

10. Christin, C.; Hoefsloot, H.C.J.; Smilde, A.K.; Hoekman, B.; Suits, F.; Bischoff, R.; Horvatovich, P. A Critical Assessment of Feature Selection Methods for Biomarker Discovery in Clinical Proteomics. *Mol. Cell. Proteom.* **2013**, *12*, 263–276. [CrossRef]

11. Boulesteix, A.L.; Strimmer, K. Partial Least Squares: A Versatile Tool for the Analysis of High-Dimensional Genomic Data. *Brief. Bioinform.* **2007**, *8*, 32–44. [CrossRef] [PubMed]

12. Lê Cao, K.A.; Boitard, S.; Besse, P. Sparse PLS Discriminant Analysis: Biologically Relevant Feature Selection and Graphical Displays for Multiclass Problems. *BMC Bioinform.* **2011**, *12*, 253. [CrossRef] [PubMed]

13. Farah, C.S.; Dalley, A.J.; Nguyen, P.; Batstone, M.; Kordbacheh, F.; Perry-Keene, J.; Fielding, D. Improved Surgical Margin Definition by Narrow Band Imaging for Resection of Oral Squamous Cell Carcinoma: A Prospective Gene Expression Profiling Study. *Head Neck* **2016**, *38*, 832–839. [CrossRef] [PubMed]

14. Farah, C.S.; Fox, S.A.; Dalley, A.J. Integrated MiRNA-MRNA Spatial Signature for Oral Squamous Cell Carcinoma: A Prospective Profiling Study of Narrow Band Imaging Guided Resection. *Sci. Rep.* **2018**, *8*, 823. [CrossRef] [PubMed]

15. Farah, C.S. Narrow Band Imaging-Guided Resection of Oral Cavity Cancer Decreases Local Recurrence and Increases Survival. *Oral Dis.* **2018**, *24*, 89–97. [CrossRef] [PubMed]

16. Wu, Z.; Irizarry, R.A.; Gentleman, R.; Martinez-Murillo, F.; Spencer, F. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *J. Am. Stat. Assoc.* **2004**, *99*, 909–917. [CrossRef]

17. Huang, D.W.; Sherman, B.T.; Lempicki, R.A. Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nat. Protoc. Lond.* **2009**, *4*, 44–57. [CrossRef]

18. Rouillard, A.D.; Gundersen, G.W.; Fernandez, N.F.; Wang, Z.; Monteiro, C.D.; McDermott, M.G.; Ma'ayan, A. The Harmonizome: A Collection of Processed Datasets Gathered to Serve and Mine Knowledge about Genes and Proteins. *Database* **2016**, *2016*. [CrossRef]

19. Rohart, F.; Gautier, B.; Singh, A.; Cao, K.A.L. MixOmics: An R Package for 'omics Feature Selection and Multiple Data Integration. *PLoS Comput. Biol.* **2017**, *13*, e1005752. [CrossRef]

20. Reis, P.P.; Waldron, L.; Perez-Ordonez, B.; Pintilie, M.; Galloni, N.N.; Xuan, Y.; Cervigne, N.K.; Warner, G.C.; Makitie, A.A.; Simpson, C.; et al. A Gene Signature in Histologically Normal Surgical Margins Is Predictive of Oral Carcinoma Recurrence. *BMC Cancer* **2011**, *11*, 437. [CrossRef]

21. Zou, H.; Hastie, T. Regularization and Variable Selection via the Elastic Net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2005**, *67*, 301–320. [CrossRef]

22. Lohavanichbutr, P.; Méndez, E.; Holsinger, F.C.; Rue, T.C.; Zhang, Y.; Houck, J.; Upton, M.P.; Futran, N.; Schwartz, S.M.; Wang, P.; et al. A 13-Gene Signature Prognostic of HPV-Negative OSCC: Discovery and External Validation. *Clin. Cancer Res.* **2013**, *19*, 1197–1203. [CrossRef] [PubMed]

23. Ziober, A.F.; Patel, K.R.; Alawi, F.; Gimotty, P.; Weber, R.S.; Feldman, M.M.; Chalian, A.A.; Weinstein, G.S.; Hunt, J.; Ziober, B.L. Identification of a Gene Signature for Rapid Screening of Oral Squamous Cell Carcinoma. *Clin. Cancer Res.* **2006**, *12*, 5960–5971. [CrossRef] [PubMed]

24. Anderson, C.R.; Sisson, K.; Moncrieff, M. A Meta-Analysis of Margin Size and Local Recurrence in Oral Squamous Cell Carcinoma. *Oral Oncol.* **2015**, *51*, 464–469. [CrossRef]

25. Pierik, A.S.; Leemans, C.R.; Brakenhoff, R.H. Resection Margins in Head and Neck Cancer Surgery: An Update of Residual Disease and Field Cancerization. *Cancers* **2021**, *13*, 2635. [CrossRef] [PubMed]

26. Lee, H.; Park, B.C.; Soon Kang, J.; Cheon, Y.; Lee, S.; Jae Maeng, P. MAM Domain Containing 2 Is a Potential Breast Cancer Biomarker That Exhibits Tumour-Suppressive Activity. *Cell Prolif.* **2020**, *53*, e12883. [CrossRef] [PubMed]

27. Hossain, M.S.; Quadery Tonmoy, M.I.; Islam, M.N.; Islam, M.S.; Afif, I.K.; Singha Roy, A.; Fariha, A.; Al Reza, H.; Bahadur, N.M.; Rahaman, M.M. MicroRNAs Expression Analysis Shows Key Affirmation of Synaptopodin-2 as a Novel Prognostic and Therapeutic Biomarker for Colorectal and Cervical Cancers. *Heliyon* **2021**, *7*, e07347. [CrossRef]

28. Ye, H.; Yu, T.; Temam, S.; Ziober, B.L.; Wang, J.; Schwartz, J.L.; Mao, L.; Wong, D.T.; Zhou, X. Transcriptomic Dissection of Tongue Squamous Cell Carcinoma. *BMC Genom.* **2008**, *9*, 69. [CrossRef]

29. Menck, K.; Heinrichs, S.; Baden, C.; Bleckmann, A. The WNT/ROR Pathway in Cancer: From Signaling to Therapeutic Intervention. *Cells* **2021**, *10*, 142. [CrossRef] [PubMed]

30. Nema, S.; Kallianpur, S.; Kumar, A.; Nema, R.; Vishwakarma, S.; Nema, S.K. Do Patients with Oral Squamous Cell Carcinoma Express Receptor Tyrosine Kinase-like Orphan Receptor 1? Results of an Observational Study. *J. Oral Maxillofac. Pathol.* **2021**, *25*, 105–109. [CrossRef] [PubMed]

31. Balakrishnan, A.; Goodpaster, T.; Randolph-Habecker, J.; Hoffstrom, B.G.; Jalikis, F.G.; Koch, L.K.; Berger, C.; Kosasih, P.L.; Rajan, A.; Sommermeyer, D.; et al. Analysis of ROR1 Protein Expression in Human Cancer and Normal Tissues. *Clin. Cancer Res.* **2017**, *23*, 3061–3071. [CrossRef] [PubMed]

32. Lee, D.; Wang, Y.H.; Kalaitzidis, D.; Ramachandran, J.; Eda, H.; Sykes, D.B.; Raje, N.; Scadden, D.T. Endogenous Transmembrane Protein UT2 Inhibits PSTAT3 and Suppresses Hematological Malignancy. *J. Clin. Investig.* **2016**, *126*, 1300–1310. [CrossRef] [PubMed]

33. Alonso, S.; Mayol, X.; Nonell, L.; Salvans, S.; Pascual, M.; Pera, M.; on behalf of the Colorectal Cancer Research Group (Hospital del Mar Medical Research Institute). Peripheral Blood Leucocytes Show Differential Expression of Tumour Progression-Related Genes in Colorectal Cancer Patients Who Have a Postoperative Intra-Abdominal Infection: A Prospective Matched Cohort Study. *Colorectal Dis.* **2017**, *19*, O115–O125. [CrossRef] [PubMed]

34. Galliera, E.; Tacchini, L.; Corsi Romanelli, M.M. Matrix Metalloproteinases as Biomarkers of Disease: Updates and New Insights. *Clin. Chem. Lab. Med.* **2015**, *53*, 349–355. [CrossRef] [PubMed]

35. Kuriakose, M.A.; Chen, W.T.; He, Z.M.; Sikora, A.G.; Zhang, P.; Zhang, Z.Y.; Qiu, W.L.; Hsu, D.F.; McMunn-Coffran, C.; Brown, S.M.; et al. Selection and Validation of Differentially Expressed Genes in Head and Neck Cancer. *Cell. Mol. Life Sci.* **2004**, *61*, 1372–1383. [CrossRef] [PubMed]

36. Ginos, M.A.; Page, G.P.; Michalowicz, B.S.; Patel, K.J.; Volker, S.E.; Pambuccian, S.E.; Ondrey, F.G.; Adams, G.L.; Gaffney, P.M. Identification of a Gene Expression Signature Associated with Recurrent Disease in Squamous Cell Carcinoma of the Head and Neck. *Cancer Res.* **2004**, *64*, 55–63. [CrossRef]

37. Ohmura, G.; Tsujikawa, T.; Yaguchi, T.; Kawamura, N.; Mikami, S.; Sugiyama, J.; Nakamura, K.; Kobayashi, A.; Iwata, T.; Nakano, H.; et al. Aberrant Myosin 1b Expression Promotes Cell Migration and Lymph Node Metastasis of HNSCC. *Mol. Cancer Res.* **2015**, *13*, 721–731. [CrossRef]

38. Yao, Y.; Yan, Z.; Lian, S.; Wei, L.; Zhou, C.; Feng, D.; Zhang, Y.; Yang, J.; Li, M.; Chen, Y. Prognostic Value of Novel Immune-Related Genomic Biomarkers Identified in Head and Neck Squamous Cell Carcinoma. *J. Immunother. Cancer* **2020**, *8*, e000444. [CrossRef]

39. Li, Z.; Chen, C.; Wang, J.; Wei, M.; Liu, G.; Qin, Y.; She, L.; Liu, Y.; Huang, D.; Tian, Y.; et al. Overexpressed PLAU and Its Potential Prognostic Value in Head and Neck Squamous Cell Carcinoma. *PeerJ* **2021**, *9*, e10746. [CrossRef]

40. Chen, G.; Sun, J.; Xie, M.; Yu, S.; Tang, Q.; Chen, L. PLAU Promotes Cell Proliferation and Epithelial-Mesenchymal Transition in Head and Neck Squamous Cell Carcinoma. *Front. Genet.* **2021**, *12*, 651882. [CrossRef] [PubMed]

41. Jiang, P.; He, S.; Li, Y.; Xu, Z. Identification of Therapeutic and Prognostic Biomarkers of Lamin C (LAMC) Family Members in Head and Neck Squamous Cell Carcinoma. *Med. Sci. Monit. Int. Med. J. Exp. Clin. Res.* **2020**, *26*, e925735. [CrossRef] [PubMed]

42. Zhou, Y.M.; Yao, Y.L.; Liu, W.; Shen, X.M.; Shi, L.J.; Wu, L. MicroRNA-134 Inhibits Tumor Stem Cell Migration and Invasion in Oral Squamous Cell Carcinomas via Downregulation of PI3K-Akt Signaling Pathway by Inhibiting LAMC2 Expression. *Cancer Biomark.* **2020**, *29*, 51–67. [CrossRef]

43. Meireles Da Costa, N.; Mendes, F.A.; Pontes, B.; Nasciutti, L.E.; Ribeiro Pinto, L.F.; Palumbo Júnior, A. Potential Therapeutic Significance of Laminin in Head and Neck Squamous Carcinomas. *Cancers* **2021**, *13*, 1890. [CrossRef] [PubMed]

44. Sun, X.; Wang, M.; Wang, M.; Yao, L.; Li, X.; Dong, H.; Li, M.; Sun, T.; Liu, X.; Liu, Y.; et al. Role of Proton-Coupled Monocarboxylate Transporters in Cancer: From Metabolic Crosstalk to Therapeutic Potential. *Front. Cell Dev. Biol.* **2020**, *8*, 651. [CrossRef] [PubMed]

45. Jin, Y.; Qin, X. Comprehensive Analysis of Transcriptome Data for Identifying Biomarkers and Therapeutic Targets in Head and Neck Squamous Cell Carcinoma. *Ann. Transl. Med.* **2020**, *8*, 282. [CrossRef]

46. Khammanivong, A.; Sorenson, B.S.; Ross, K.F.; Dickerson, E.B.; Hasina, R.; Lingen, M.W.; Herzberg, M.C. Involvement of Calprotectin (S100A8/A9) in Molecular Pathways Associated with HNSCC. *Oncotarget* **2016**, *7*, 14029–14047. [CrossRef] [PubMed]

47. Wang, Q.; Ma, C.; Kemmner, W. Wdr66 Is a Novel Marker for Risk Stratification and Involved in Epithelial-Mesenchymal Transition of Esophageal Squamous Cell Carcinoma. *BMC Cancer* **2013**, *13*, 137. [CrossRef] [PubMed]

48. Cao, Y.; Liu, H.; Xia, S.L.; Zhang, X.; Bai, H.; Yang, Q.; Li, J.; Gao, L.; Jin, F.; Wei, M.J.; et al. PTEN Downregulates WD Repeat-containing Protein 66 in Salivary Adenoid Cystic Carcinoma. *Oncol. Rep.* **2019**, *41*, 1827–1836. [CrossRef] [PubMed]