Article

Transcriptomic Analysis for Prognostic Value in Head and Neck Squamous Cell Carcinoma

Li-Hsing Chi [1,2,3], Alexander TH Wu [1], Michael Hsiao [4]* and Yu-Chuan (Jack) Li [1,5]*

[1] The Ph.D. Program for Translational Medicine, College of Medical Science and Technology, Taipei Medical University and Academia Sinica, Taipei, Taiwan
[2] Division of Oral and Maxillofacial Surgery, Department of Dentistry, Wan Fang Hospital, Taipei Medical University
[3] Division of Oral and Maxillofacial Surgery, Department of Dentistry, Taipei Medical University Hospital, Taipei Medical University
[4] Genomics Research Center, Academia Sinica, Taipei, Taiwan
[5] Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, No.172-1, Sec. 2, Keelung Rd., Taipei 106, Taiwan
* Correspondence: Hsiao: mhsiao@gate.sinica.edu.tw; Li: jaak88@gmail.com

Abstract: The survival analysis of the Cancer Genome Atlas (TCGA) dataset is a well-known method to discover the gene expression-based prognostic biomarkers of head and neck squamous cell carcinoma (HNSCC). A cutoff point is usually used in survival analysis for the patients' dichotomization in the continuous gene expression. There is some optimization software for cutoff determination. However, the software's predetermined cutoffs are usually set at the median or quantiles of gene expression value to perform the analyses. There are also few clinicopathological features available on their pre-processed data sets. We applied an in-house workflow, including data retrieving and pre-processing, feature selection, sliding-window cutoff selection, Kaplan-Meier survival analysis, and Cox proportional hazard modeling for biomarker discovery. In our approach for the TCGA HNSCC cohort, we scanned human protein-coding genes to find optimal cutoff values. After adjustment with confounders, the clinical tumor stage and the surgical margin involvement are independent risk factors for patients' prognosis. According to the resulting tables with Bonferroni-adjusted *P* value under the optimal cutoff and the hazard ratio, three biomarker candidates, CAMK2N1, CALML5, and FCGBP, are significantly associated with the patients' overall survival. We validated this discovery by using the other independent HNSCC dataset (GSE65858). Thus, we suggest the transcriptomic analysis could help for biomarker discovery.

Keywords: Head and Neck Squamous Cell Carcinoma (HNSCC); the Cancer Genome Atlas (TCGA); Transcriptomic Analysis; Survival Analysis; Optimal Cutoff; Effect Size; calcium/calmodulin dependent protein kinase II inhibitor 1 (CAMK2N1); calmodulin like 5 (CALML5); Fc fragment of IgG binding protein (FCGBP); Mindfulness Meditation

1. Introduction

Head and neck squamous cell carcinoma (HNSCC), including oral, oropharyngeal, and hypopharyngeal origin, is the fourth leading cancer causes of death for males in Taiwan[1]. The age-standardized incidence rate of HNSCC in males is 42.43 per 100,000 persons[2]. The treatment strategies of HNSCC are surgery alone, systemic therapy with concurrent radiation therapy (systemic therapy/RT), or surgery with adjuvant systemic therapy/RT (according to National Comprehensive Cancer Network, NCCN Clinical Practice Guidelines in HNSCC, Version 2.2020)[3]. Despite the improvement in those interventions, the survival of HNSCC has improved only marginally over the past decade worldwide[4]. The critical advancement of targeted therapy and immuno-oncology should benefit from emerging prognostic biomarkers guiding modern systemic therapy.

Accumulative knowledge shows that some biomarkers have prognostic significance in HNSCC. For example, node-negative HNSCC patients with p53 overexpression were found to hold lower survival[5]. Overexpression of hypoxia-inducible factor (HIF)-1 alpha[6] or Ki-67[7] was found to be correlated with poor response to radiotherapy of HNSCC. The epidermal growth factor receptor (EGFR)[8][9] and matrix metalloproteinase (MMP)[10] were found to be overexpressed to promote invasion and metastasis of HNSCC. From 2000 to 2006, the first anti-EGFR antibody-drug (cetuximab) has been developed and combined with radiotherapy, known as bio-RT, to increase survival of unresectable locoregionally advanced disease[11]. The systemic therapy of cetuximab plus platinum-fluorouracil chemotherapy (EXTREME regimen) improves overall survival when given as first-line treatment in patients with recurrent or metastatic HNSCC[12][13]. It has been approved by the US Food and Drug Administration (FDA) since 2008. In advance, the bio-RT could be proceeded with docetaxel, cisplatin, and 5-fluorouracil (Tax-PF) induction chemotherapy to overcome radio-resistance of HNSCC[14].

However, Rampias and his colleagues[15] suggested that Harvey rat sarcoma viral oncoprotein (HRAS) mutations could mediate cetuximab resistance in systemic therapy of HNSCC via the EGFR/rat sarcoma (RAS)/extracellular signal-regulated kinases (ERK) signaling pathway. After that, the EGFR tyrosine kinase inhibitor (TKI) was introduced to help cetuximab in 2018. The anti-tumor activity was observed in a phase 1 trial for HNSCC patients using cetuximab and afatinib, a TKI of EGFR, human epidermal growth factor receptor (HER)2, and HER4[16]. Other EGFR TKIs, such as gefitinib, erlotinib, osimertinib, were also developed to treat advanced HNSCC. Although 90% of HNSCC has overexpression of EGFR, cetuximab has only 10% to 20% response rate on those patients. Before the year 2019, cetuximab is still the only drug of choice with proven efficacy, which has targeted the selected HNSCC patients[17].

Until the immuno-oncology era, immune-checkpoint inhibitor (ICI) was introduced since 2014 for treating HNSCC[18][19]. The ICI works on immune checkpoint molecules, including programmed death 1 (PD-1), cytotoxic T lymphocyte antigen 4 (CTLA-4), T-cell immunoglobulin mucin protein 3 (TIM-3), lymphocyte activation gene 3 (LAG-3), T cell immunoglobin and immunoreceptor tyrosine-based inhibitory motif (TIGIT), glucocorticoid-induced tumor necrosis factor receptor (GITR), and V-domain Ig suppressor of T-cell activation (VISTA)[20]. The US FDA has approved the anti-PD-1 agents (e.g. pembrolizumab and nivolumab) as a monotherapy for the platinum-treated patients with recurrent or metastatic HNSCC[21]. According to the phase 3 KEYNOTE-048 study, PD-L1 is a validated biomarker used in clinical guidance for candidate selection of pembrolizumab[22][23]. However, due to the complexity of immune-tumor interaction, ICI has 20% response rate to programmed death ligand 1 (PD-L1) expressed patients (over 50% in immunohistochemistry, IHC staining of HNSCC)[19][23].

According to our previous proteomic study from 2010 to 2017, thymosin beta-4 X-linked (TMSB4X) was reported to be related to tumor growth and metastasis of HNSCC[24]. It was also reported by the subsequent investigations that TMSB4X has engaged in tumor aggressiveness through epithelial-mesenchymal-transition (EMT) on pancreatic[25], gastric[26], colorectal[27], lung[28], ovarian[29], and melanoma[30] cancers. Thus, it might be suggested that TMSB4X is a candidate for tumor type-agnostic therapy[31] as a common biomarker crossing several types of cancer.

The Cancer Genome Atlas (TCGA) profiled HNSCC (528 participants) clinical and genomic data, which were standardized and are available at a unified data portal, Genomic Data Commons (GDC) of the the National Cancer Institute (NCI). The highlight of the advantages of applying the TCGA data for cancer biomarker identification includes:

- To the best of our knowledge, the TCGA database is the largest collection (in both cancer types and cohort size, especially in HNSCC) of comprehensive genomics with survival data available in the field of cancer research so far. The whole-genome sequencing data were harmonized across all Genome Data Analysis Center

(GDAC)s. Many databases adopt the essential demographic data from TCGA since it has comprehensive physical and social features of patients, such as exposure to alcohol, asbestos, radioactive radon, tobacco smoking, or cigarettes.
- TCGA has a remarkable advantage for computational and life scientists who study cancer since useful web-based tools and Application Programmable Interface (API) are ready to analyze and visualize TCGA data. It might be getting help quickly from the research community for trouble-shooting.
- Many achievements of diagnoses, treatments, and prevention from the TCGA data were already published and keep growing[32].

Usually, researchers developed an in-house workflow of gene expression analysis of TCGA data to find HNSCC biomarkers. It would be helpful to show that alteration in gene expression correlates with phenotypes of HNSCC. Some researchers[33][34][35][36][37][38][39] tried to find differentially expressed genes (DEGs) of the HNSCC samples at both genotypic and phenotypic levels (without survival information) for biomarker discovery. Gene expression data were downloaded from the TCGA or Gene Expression Omnibus (GEO) databases (e.g., GSE117973[39], HIPO-HNC cohort has n = 87). They used the Database for Annotation, Visualization, and Integrated Discovery—DAVID (available at https://david.ncifcrf.gov/) to obtain information for Gene Ontology (GO), including biological processes, the cellular component, and molecular function. The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis was also used to annotate the potential functions of their biomarker candidates. The pathway enrichment analysis of DEGs was also performed by DAVID, STRING (available at https://string-db.org), or Cytoscape software[37][38]. Li[36] and his colleagues made an R package (GDCRNATool) for the implementation of those workflows for gene expression analyses of the TCGA. Xu and his colleagues[40] also identified their biomarkers by DEGs analysis. The significant impact of genes on patients' overall survival was evaluated by Kaplan–Meier survival curves with a log-rank test (P value < 0.01), and univariate Cox regression. They validated the candidate genes by using the web-based tools of Gene Expression Profiling Interactive Analysis tool—GEPIA and Human protein atlas (HPA) databases. GEPIA was developed, using TCGA datasets, by Zefang Tang and his colleagues (version 1[41], version 2[42], and GEPIA2021[43], available at http://gepia2021.cancer-pku.cn/). HPA[44] applied immunohistochemistry (IHC) for the TCGA database (please see details in the Discussions section "Validation by Web-based Tools"). Finally, their biomarkers were verified by using the gene expression profile from the GEO, HNSCC cell lines and tissues.

The other investigators should get the rationale or revelation of the genes of interest on a specific cancer type. They should upload those genes manually onto web-based tools, such as SurvExpress[45] (available at http://bioinformatica.mty.itesm.mx:8080/Biomatec/SurvivaX.jsp), then analyze the cohort of interest (e.g., TCGA). After downloading the survival results, they could curate plots and tables carefully. It is not possible to scan the whole human protein-coding genome in this way. The web-based tools might set a cutoff at the median, 1/4 quantile, or 3/4 quantile for subsequent analyses. There are several visualization software or R packages which deal with cutoff determination[46], such as Prognoscan[47], Cutoff Finder[48], Findcut[49], OptimalCutpoints[50], cutpointr (available at https://github.com/thie1e/cutpointr), and cutoffR (available at https://cran.r-project.org/web/packages/cutoffR). However, non of them could perform jointly survival analysis with cutoff selection and whole-genome scanning.

In summary, identifying predictive biomarkers for selecting standard-of-care or advanced systemic therapy[50] in HNSCC is crucial. Our approach describes an in-house workflow implemented in the R script, which runs on the Rstudio server. Its function includes data retrieving and pre-processing, feature selection, sliding-window cutoff selection, Kaplan-Meier survival analysis, Cox proportional hazard modeling, and biomarker discovery. The independent HNSCC dataset (GSE65858)[51] was used to validate this strategy. The workflow, shown in Figure 1, has scanned 20,500 human protein-coding genes of the TCGA HNSCC cohort to yield a model with biomarker estimate by gene-expression based survival analysis.

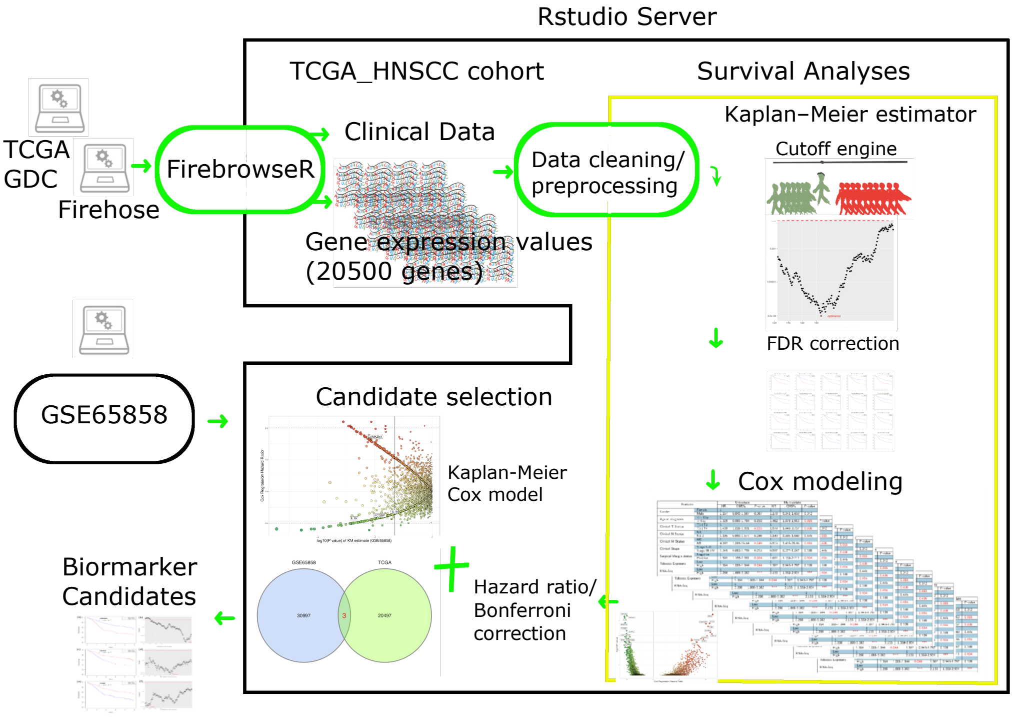

Figure 1. A workflow of HNSCC biomarker discovery The workflow includes data retrieval from TCGA GDC data portal, data process with merging and cleaning, then performing the survival analyses (within yellow square). The Cutoff engine (in R script: cutofFinder_func.HNSCC.R, a serial cut for grouping patients with low or high expression of a specific gene, to yield a collection of *P* values; please see Materials and Methods section for details) might calculate all possible Kaplan-Meier *P* values (corrected by false discovery rate, FDR, method) to find the optimal cutoff value of gene expression for subsequent Cox modeling. The candidate selection performs (1) dissecting and selection of candidate genes by further Bonferroni adjusted *P* values as well as a hazard ratio of Cox model, based on the results from the survival analyses; (2) survival analyses of the other HNSCC dataset (GSE65858) using Kaplan-Meier estimate (with FDR correction) and Cox modeling.

The biomarker candidates were a consensus result of TCGA and GSE65858. (HNSCC: head and neck squamous cell carcinoma; TCGA: the Cancer Genome Atlas; RNA-Seq: RNA sequencing; GDC: Genomic Data Commons.)

2. Results

TCGA HNSCC cohort was applied for exploration of the biomarker candidate. A total of 9416 Kaplan-Meier plots (under sliding-window cutoff selection) with associated Cox's univariate and multivariate tables were generated by Cox modeling (see Figure 1) and justified by the ranking of hazard ratios. The 967 out of 9416 genes were kept by criteria of FDR-adjusted Kaplan-Meier *P* value (< 0.05) and hazard ratio (HR) derived from Cox's model (please see Figure 2(a), (b), initial trial). In the next step, a stringent Bonferroni P-value correction was used to yield 20 genes (please see Figure 2(c), (d)).

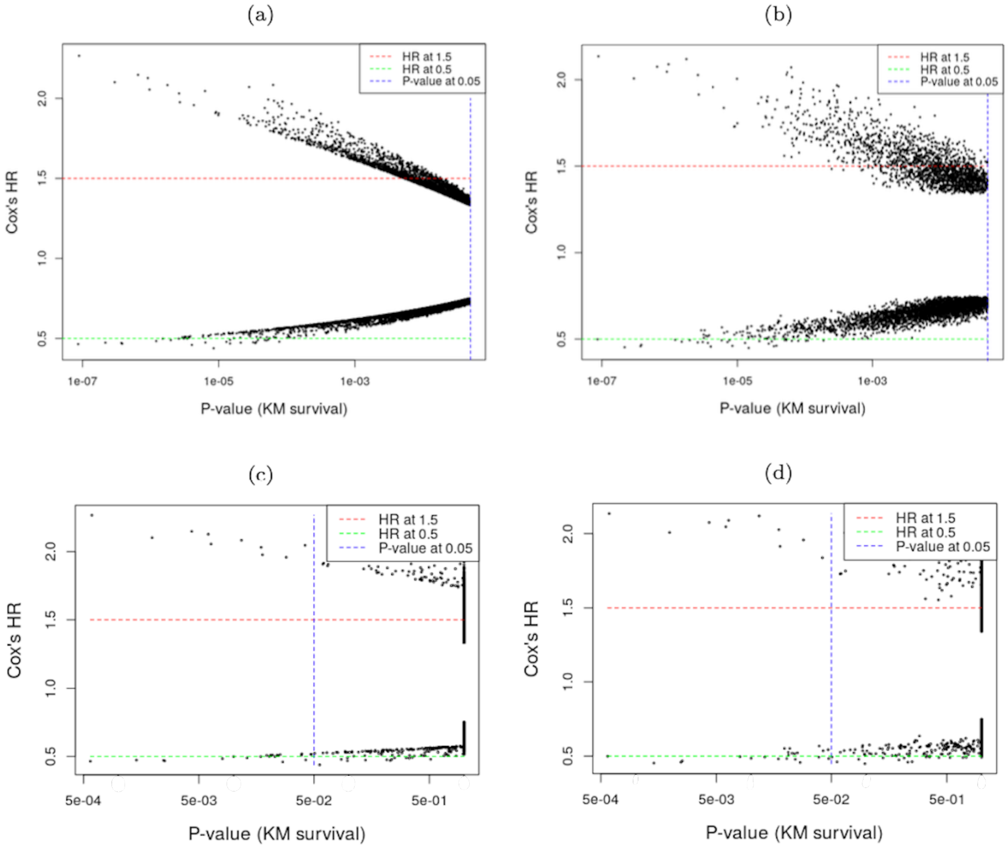

Figure 2. The initial progress of candidate selection from TCGA HNSCC cohort. The *P* values of Kaplan-Meier survival is one of the selection criteria. The effect size is estimated by Cox's hazard ratio. Initial trial step: (a) univariate HR versus *P* value; (b) multivariate HR versus *P* value. After stringent criteria by Bonferroni-adjusted *P* value, and the Cox's HR, few top-ranked genes are shown in (c) univariate HR versus Bonferroni-adjusted *P* value; and (d) multivariate HR versus Bonferroni-adjusted *P* value.
(TCGA: the Cancer Genome Atlas; HR: hazard ratio)

These CAMK2N1, CALML5, FCGBP and the other 17 genes (DKK1, STC2, PGK1, SURF4, USP10, NDFIP1, FOXA2, STIP1, DKC1, ZNF557, ZNF266, IL19, MYO1H, EVPLL, PNMA5, IQCN, and NPB) have significant FDR-adjusted *P* values (< 0.0003) in Kaplan-Meier estimate, and greater hazard ratio (HR) (> 1.8 or < 0.6) in Cox's model (please see Figure 3; $log_{10}(0.0003) = -3.5$). The plot reveals that the top 20 genes (Bonferroni-adjusted $P < 0.05$) are located on the peaks. At the same time, Cox's HR separates them on the two-side with significant prognostic impact.

In the study of GSE65858 cohort with median cutoffs, CAMK2N1, CALML5, and FCGBP (3 out of those 20 genes discovered in the TCGA cohort), keep ahead of the curve by their FDR-corrected *P* value (< 0.05), and Cox's HR (> 1.8 or < 0.6) (please see Table 1). However, the significance of the other 17 genes are replaced by genes such as DUSP6,

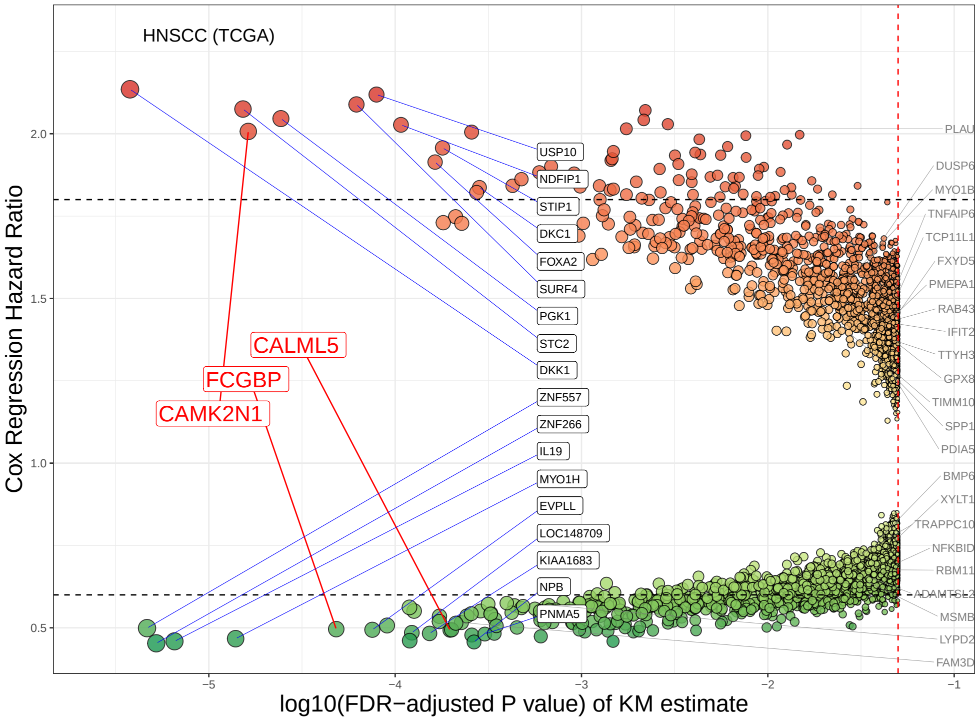

Figure 3. Volcano plot of genes under survival analyses of TCGA HNSCC. This cohort was applied for exploration of the candidate biomarkers. A total of 9416 genes have unadjusted P value less than 0.05. CAMK2N1, CALML5, FCGBP and the 17 genes (marked in black square) have hazard ratio (HR) $>$ 1.8 or $<$ 0.6. The 22 genes, listed on the side, have hazard ratio between 0.6 and 1.5.
(Red spots: $HR > 1.0$, Green spots: $HR < 1.0$);
(X-axis: Kaplan-Meier survival estimates, with FDR-adjusted P value (log10 transformed));
(Y-axis: HR of Cox proportional hazard regression model.)

MSMB, and RBM11 (please see Figure 4). Conversely, there are 22 genes, which have greater hazard ratio ($>$ 1.8 or $<$ 0.6) in GSE65858 cohort (Figure 4), drop their hazard ratio between 0.6 and 1.5 in the study of TCGA HNSCC (Figure 3). Thus, there is a consensus between the TCGA and GSE65858 cohorts that CAMK2N1, CALML5, and FCGBP are significant candidates for the HNSCC biomarker.

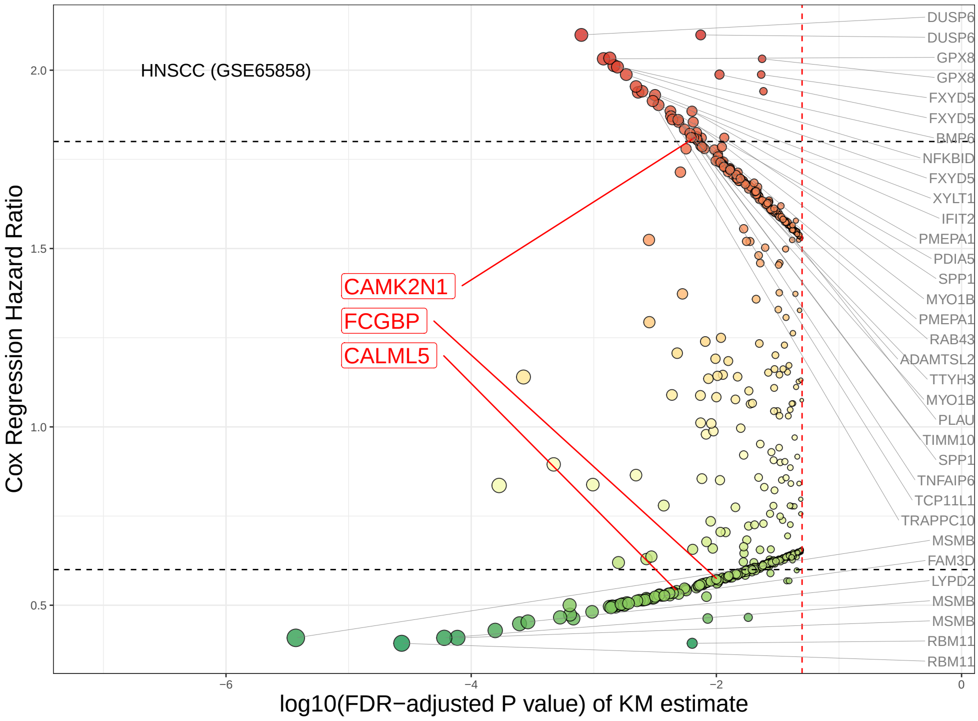

Figure 4. Volcano plot of genes under survival analyses of GSE65858 cohort This HNSCC cohort has been applied for filtering of our candidate genes: CAMK2N1, CALML5, and FCGBP. Total 534 genes has FDR-adjusted P value less than 0.05 (Red spots: hazard ratio is greater than 1.0); (Green spots: hazard ratio is under than 1.0).
The 22 genes, listed on the side, have hazard ratio > 1.8 or < 0.6.
(X-axis: Kaplan-Meier survival estimates, with FDR-adjusted P value with log10 transformed); (Y-axis: the hazard ratio (HR) under Cox proportional hazard regression model)

Table 1: The consensus between the TCGA and GSE65858 cohorts in Kaplan-Meier survival and Cox's model

Gene Symbol	KM P value		FDR-adjusted P value		Cox's univariate HR	
	TCGA	GSE65858	TCGA	GSE65858	TCGA	GSE65858
CAMK2N1	2.97×10^{-7}	6.87×10^{-3}	1.63×10^{-5}	0.038	2.101	1.814
CALML5	5.87×10^{-6}	4.75×10^{-3}	1.97×10^{-4}	0.035	0.493	0.541
FCGBP	1.21×10^{-6}	0.01	4.83×10^{-5}	0.039	0.484	0.573

(FDR: false discovery rate; HR: hazard ratio)

Our top 1 candidate is calcium/calmodulin dependent protein kinase II inhibitor 1 (CAMK2N1). The Kaplan-Meier curve reveals that 152 patients bearing the higher expression of CAMK2N1 were suffered from only 35% of 5-year OS rate. In comparison, the other 262 patients with lower expression had a better prognosis (Bonferroni-adjusted $P = 0.002$) (see Figure 5(a)). Figure 5(b)'s cumulative P value plot shows that the uncorrected 147 P values (< 0.05) have been estimated by a serial cut from 144 to 290 persons for grouping the cohort in our cutoff finding procedure (cutofFinder_func.R, see Figure 1, cutoff engine). The smallest P value (2.97×10^{-7}), when cut at n = 262 (63.3% of total cohort 414, with the cutoff value 0.027 in RNA-Seq by Expectation-Maximization, RSEM), has been defined as an optimal P value. The plot in Figure 5(b) shows a "backlash" curve with the half of values below 1.0×10^{-3}. Conversely, the most associated gene with better survival is calmodulin like 5 (CALML5). In Figure 5(c), a Kaplan-Meier curve reveals 200 patients bearing the higher expression of CALML5 had 60% of 5-year OS survival rate (Bonferroni-adjusted $P = 0.039$). The sliding-window cutoff selection generated cumulative P value plot in Figure 5(d). This plot reveals a

"V" curve with the minimum at middle portion. The 166 uncorrected P values were estimated by a serial cut from 125 to 290 for grouping the cohort. The smallest P value (5.87×10^{-6}), when cut at n = 214 (51.7% of total cohort 414), has been defined as an optimal P value with a cutoff value -0.359 RSEM of RNA-Seq. The third candidate is Fc fragment of IgG binding protein (FCGBP). It relates with also better survival in both the TCGA and GSE65858 cohorts. In Figure 5(e), a Kaplan-Meier curve reveals 282 patients bearing the higher expression of FCGBP had 60% of 5-year OS survival rate (Bonferroni-adjusted $P = 0.008$). The sliding-window cutoff selection generated cumulative P value plot in Figure 5(f). This plot has a "W-shaped" curve with the most majority of values far below 1.0×10^{-3}. The 166 uncorrected P values were estimated by a serial cut from 125 to 290 for grouping the cohort. The smallest P value (1.21×10^{-6}), when cut at n = 132 (31.9% of total cohort 414), has been defined as an optimal P value with a cutoff value -0.472 RSEM of RNA-Seq.

Table 2: The top 3 genes with prognostic impact in HNSCC

Gene ID	Gene Description	Kaplan-Meier survival		Cox Univariate		Cox Multivariate	
		FDR *P* value	Bonferroni *P* value	HR*	95% CI	HR*	95% CI
CAMK2N1	calcium/calmodulin-dependent protein kinase II inhibitor 1	1.63×10^{-5}	0.002	2.101	1.572-2.809	2.007	1.490-2.704
CALML5	calmodulin like 5	1.97×10^{-4}	0.039	0.51	0.379-0.686	0.493	0.364-0.667
FCGBP	Fc fragment of IgG binding protein	4.83×10^{-5}	0.008	0.484	0.359-0.653	0.496	0.366-0.674

Selection criteria (fit all):
(1) Kaplan-Meier Bonferroni-adjusted P <0.05;
(2) Cox's univariate and multivariate HR >= 1.8 or <= 0.6 in TCGA cohort;
(3) Cox's univariate and multivariate HR >= 1.8 or <= 0.6 in GSE65858 cohort
* Cox's model: P <0.001

In Table 3, after adjustment of confounders, CAMK2N1 overexpression is the independent prognostic factor (multivariate HR 2.007 [95% CI: 1.490-2.704, $P < 0.001$]), as well as clinical T stage (HR 1.982 [95% CI: 1.048-3.745, P = 0.035]) and surgical margins status (HR 1.631 [95% CI: 1.182-2.250, P = 0.003]). Older age (more than 65) also worse the survival (HR 1.391 [95% CI: 1.025-1.888, P = 0.034]). The M stage could be ignored in this cohort due to only 3 out of 414 patients with distant metastasis.

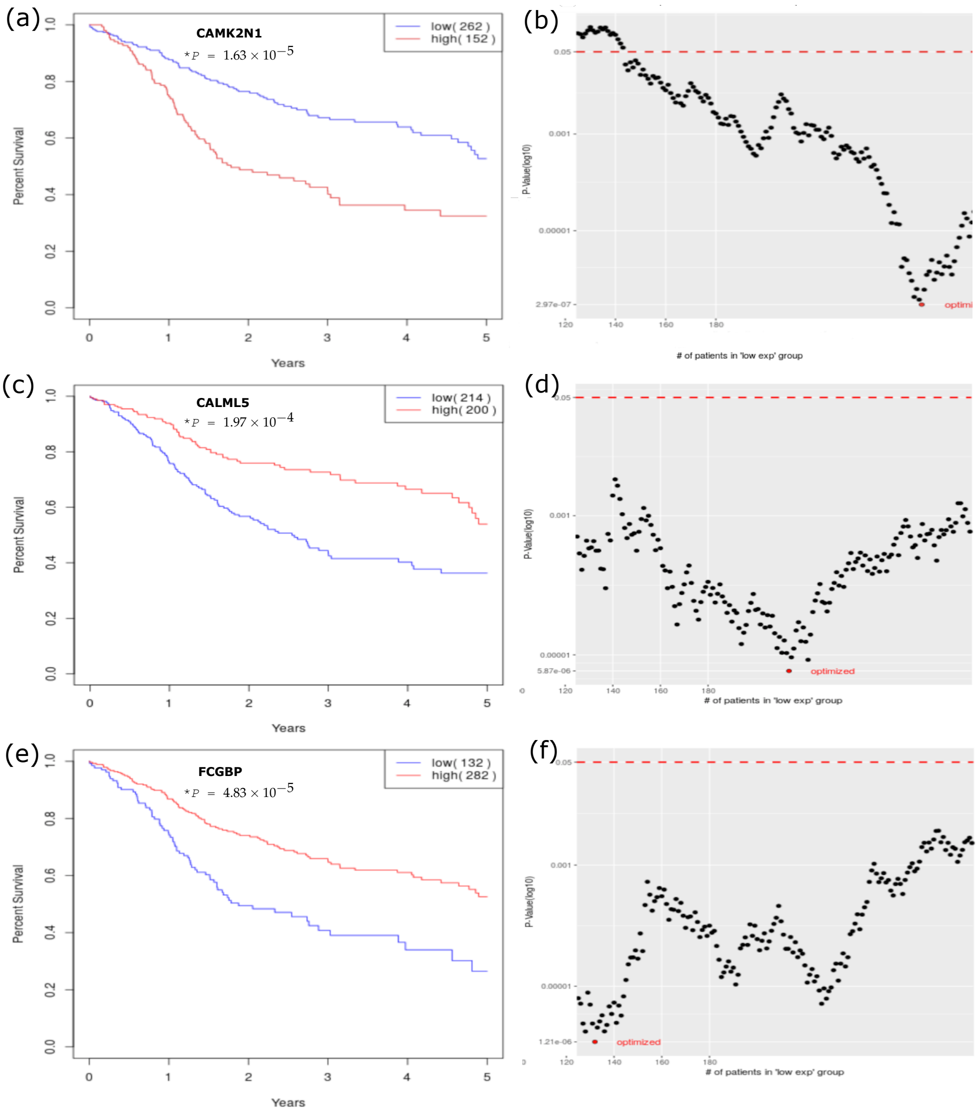

Figure 5. Kaplan-Meier survival analyses, by cutoff finding. The Kaplan-Meier curves of (a) CAMK2N1, (c) CALML5, and (e) FCGBP under optimal *P* value. The cutoffs, in the cumulative *P* value plots of (b) CAMK2N1, (d) CALML5, and (f) FCGBP, respectively, shows that over 50% of those unadjusted *P* values are below 0.001 derived by sliding-window cutoff finding procedure. (* *P*: *P* value adjusted by false discovery rate, FDR)

In summary, those three biomarker candidates , clinical T stage, and surgical margin are independent prognosis factors in HNSCC. We also found those candidates have

Table 3: Univariate/multivariate Cox's proportional hazards regression analyses on OS time of CAMK2N1 gene expression in HNSCC

Features		Univariate			Multivariate		
		HR	CI95%	*P* value	HR	CI95%	*P* value
Gender	Female	1			1		
	Male	1.157	0.843-1.587	0.367	1.076	0.767-1.510	0.671
Age at diagnosis	$<= 65y$	1			1		
	$> 65y$	1.329	0.990-1.784	0.058	1.391	1.025-1.888	0.034
Clinical T Status	T1+T2	1			1		
	T3+T4	1.409	1.028-1.931	0.033	1.982	1.048-3.745	0.035
Clinical N Status	N0	1			1		
	N1-3	1.185	0.890-1.577	0.246	1.145	0.801-1.636	0.457
Clinical M Status	M0	1			1		
	M1	4.097	1.009-16.644	0.049	7.314	1.590-33.631	0.011
Clinical Stage	Stage I+II	1			1		
	Stage III+IV	1.245	0.882-1.759	0.213	0.621	0.287-1.343	0.226
Surgical Margin status	Negative	1			1		
	Positive	1.591	1.155–2.191	0.004	1.631	1.182-2.250	0.003
Tobacco Exposure	Low	1			1		
	High	1.364	1.008-1.844	0.044	1.363	0.990-1.875	0.058
Gene Expression	Low	1			1		
	High	2.101	1.572-2.809	***	2.007	1.490-2.704	***
(OS: overall survival ; *P* value significant codes is denoted: red < 0.05; *** < 0.001)							

proper effect size—the Cox's HR either > 1.8 or < 0.6. Thus, the prognosis model with coefficients is established from TCGA and GSE65858 HNSCC cohorts.

3. Discussion

3.1. Feature Selection for Survival Modeling

Besides ethnicity, age, gender, TNM stage, radiation therapy, chemotherapy, and targeted therapy, the comprehensive adversely prognostic features in HNSCC should also include tobacco exposure, EGFR amplification, human papillomavirus (HPV) status, positive/close surgical margin ($< 5mm$), extra-nodal extension (ENE), lymph-vascular space invasion (LVSI), perineural invasion (PNI), depth of invasion (DOI) ($> 5mm$), as well as metastatic lymph node density (LND)[52], and worst pattern of invasion score 5 (WPOI-5), which is defined as tumor dispersion (1 mm apart between tumor satellites) or positive PNI/LVSI[53]. The features of DOI, LND, and tumor dispersion are not available on the TCGA dataset. The Brandwein-Gensler's risk model (lymphocytic host response, WPOI-5, and PNI)[54][55] has been suggested to be routinely performed on pathological examination. In previous reports of HNSCC, the loco-regional failure will be high when the initial frozen section has a positive/close surgical margin, and even the final margin revision revealed negative[56]. According to Table 3 in our study, the positive surgical margin yields a hazard ratio greater than 1.6 to influence on patient's OS. It is suggested by authors [57][58][59][60][61][62][63][64][65][66] that the reason of positive/close surgical margin is possibly due to tumor aggressiveness or dispersion (WPOI-5) instead of iatrogenic reason of surgery. The surgical margin status has also been suggested as an independent surrogate for tumor dispersion in the HNSCC study. Thus, we selected common clinicopathological features in the current biomarker discovery, including gender, age, clinical T, clinical N, clinical M, surgical margin status, and tobacco exposure, to adjust confounders (details description at Materials and Methods section).

3.2. The Purpose of a Sliding-window Cutoff Selection

Trying to find an optimal cutpoint of that RNA expression data to maximize candidate mining coverage, this strategy could identify more but sometimes weak "biomarkers". Thus, we should try our best to handle the effect size from Cox's modeling. And validation of those candidates is required by using other independent dataset.

Statistical significance (P value) is affected by both sample size, error, effect size (substantive significance)[67][68], and cutoff. The effect size is the magnitude of the difference (e.g., hazard ratio) among comparing groups. The effect size is independent of the sample size[67].

In a study with large sample size, the difference could be noticed easily (i.e., P value < 0.05) due to a decreased standard error[67]. However, the small effect size (non-zero) is often meaningless or substantive insignificance (e.g., hazard ratio between 0.8 and 1.2). Conversely, the effect size can be large but fail to get statistical significance if the sample size is small. The following error could also impact the P value:

- A random error, defined as the variability in data, is not considered a bias but rather occurs randomly across the entire study population and can distort the measurement process (e.g., RNA-Seq experiments). The larger sample size could reduce the random error.
- A systematic error is a bias, which includes selection bias, information bias, and confounding. It could deleteriously impact the statistical significance. The larger sample size could not affect the systematic error.

While statistical significance can inform the researcher whether an effect exists, the P value will not directly tell the effect size. Thus, if there is no error in two study groups, and the sample size is the same (not small), the group, which has a larger effect size, will has a small P value[68]. If a skewed cutoff that split, for example, 425 versus 75 between the two groups, it will also get statistical significance by increasing effect size artificially.

There is a benefit for using sliding-window cutoff method (between $30\% - 70\%$ quantile) in Kaplan-Meier analysis of TCGA HNSCC cohort at the beginning. We compared the results of cutoff at optimal P value by sliding window or just at the median of gene expression. It shows that the number of genes (with unadjusted P value < 0.05) is 6284 versus 3118, respectively. After FDR correction, it becomes 967 versus 209, respectively. The sliding-window cutoff method could catch more potential candidates, which have P values far less than 0.001, for subsequent Cox's modeling. That is because of the properly selected cutoff improving the statistical significance. To find a smaller P value might predict large effect size (HR) associated biomarkers. Then, these preliminary candidates will have an opportunity to be carefully selected by using FDR then stringent Bonferroni correction, their effect size (Cox's HR), and the other independent cohort (GSE65858) to prevent the false discovery.

We explain aforementioned situation by examples. Reviewing the special case of genes, such as NDFIP1, DKC1, PNMA5, and NPB, we noticed that NDFIP1, with a P value of 0.05 at 50% quantile (median) cutoff, could achieve a P value of 2.62×10^{-6} at 70% quantile cutoff. NDFIP1 has the FDR-adjusted P value as 1.07×10^{-4} (please see supplementary Figure S1, a "S"- or "W"-shaped acute-bending on the P value plot). However, it was excluded as a candidate by small effect size (HR = 1.33 in GSE65858) less than 1.8.

The other example is IL19. It has a P value plot with acute S-curve bending at the median zone, that lets the FDR-adjusted P value has large difference between 50% quantile cutoff (FDR 0.115) and optimal 48% quantile cutoff (FDR 6.54×10^{-6}). This optimal cutoff method could boost it's statistic significance to pass the correction by both FDR and Bonferroni methods. Even IL19 has competed as a candidate by its effect size (HR = 0.472 in TCGA cohort), but failed by the validation with GSE65858 cohort (small effect size as HR = 0.630, and FDR-KM P = 0.031).

3.3. Technical Consideration

There are two essential points of biomarker discovery from survival analysis of the TCGA HNSCC dataset.

First, since TCGA genomic data were harmonized, the pre-processing of TCGA RNA-Seq in our workflow has been done as follows:

- HNSCC samples without complete clinical information have been removed;

- Null expressed genes in more than 30% of the HNSCC samples have been excluded;
- Updated number of protein-coding genes in the TCGA HNSCC is 20500.

After investigation of mRNA expression dataset obtained through NCI's Firehose API, we found that expression value of two genes (gene ID: 9906 and gene ID: 728661) was saved together under the entity of gene symbol "SLC35E2". The expression file of SLC35E2 has almost double in size than those of SLC35E1 or SLC35E3. According to the Human Gene Database (available at https://www.genecards.org/Search/Keyword?queryString=SLC35E2), SLC35E2A (Gene ID: 9906) and SLC35E2B (Gene ID: 728661) should be the correct entities for the TCGA HNSCC dataset. SLC35E2 is the previous symbol of SLC35E2A (reference at https://www.genenames.org/data/gene-symbol-report/#!/hgnc_id/HGNC:20863). Thus, we made reassignment of the expression value of SLC35E2A and SLC35E2B and updated the number of protein-coding genes in this TCGA HNSCC dataset from 20499 to 20500.

Second, we analyzed the error log during the cutoff finding and Cox modeling, the result shows that program running could be halted under several technical situation. These include:

- 32.2% event has "one group" issue in confusion matrix of Chi-square test in Cox regression (coxph), due to zero in M (distant metastasis) patient subgroup;
- 21.05% error occurs by "one group" issue in log-rank test (survdiff or survdiff.fit function in R package "survival") in Kaplan-Meier estimate;
- 0.78% has unknown reason (so those 159 genes has been excluded in our workflow).

These technical problems could not be detected prior to program running. It might be due to skewed distribution of the expression value or even random error derived during the RNA sequencing procedure.

3.4. Validation by Web-based Tools

The Human Protein Atlas project (HPA) has proteomics analysis based on 26,941 antibodies targeting 17,165 unique proteins. The HPA's Pathology Atlas analyzes each protein in patients, using IHC analysis based on tissue microarrays (TMAs) adopted from TCGA. Kaplan-Meier survival analyses are based on RNA-Seq expression levels of human genes in HNSCC tissue and the clinical outcome. All transcriptomic data were retrieved from the TCGA, and all proteomics were generated in-house using the same antibodies. The our candidate, CAMK2N1, is also on the list of unfavorable prognostic genes for HNSCC from HPA (available at https://www.proteinatlas.org/humanproteome/pathology/head+and+neck+cancer, Version: 20.0 updated: 2020-11-19). CALML5 and FCGBP are on the list of favorable prognostic genes as well.

3.5. Future Direction in Translational Medicine

3.5.1. Proteomics Validation

Although we combine the power of genome-wide scanning and an optimal cut-off finder for survival analysis, the study has some limitations. We are aware of the importance of direct assessment of protein products comprising the basic functional units in cancer cells' biological processes. The announcement of the Cancer Proteome Atlas (TCPA, http://tcpaportal.org) excites the cancer research community[69][70]. By the utility of the reverse-phase protein arrays (RPPAs) or reverse-phase protein lysate microarray (RPMA), a microarray kind of "Western blots" in the TCPA could help to test our hypotheses from RNA-Seq study. However, in the TCPA database (v3.0[71]), there are only 237 antibodies available, not covering our candidates so far.

3.5.2. Laboratory Validation

It is encouraged for multidisciplinary studies that use complementary computational and experimental approaches to address challenging cancer research. These in vitro and in vivo validation experiments will be undertaken in our laboratory. We plan to analyze mRNA (e.g., qRT-PCR) and protein (e.g., Western blot) of HNSCC cell lysate to confirm the candidate genes' expression. The effect of overexpression and knockdown of the gene by lentiviral clones should be observed on cell function assay (e.g., proliferation, migration, and invasion) and mouse xenograft model (e.g., tumor growth).

Moreover, this bioinformatics paper provides targets and supports the community's rationale for looking into these HNSCC candidates by in vitro and in vivo validation. We aim to promote a reproducible bioinformatics[72][73] method allowing successful repetition and extension of analyses based on the TCGA or other in-house HNSCC datasets. A good research reproducibility practice is necessary to allow the reuse of code and results for new projects. It may turn out to be a time-saver in the longer run. When multiple scientists can reproduce a result, it will also validate our initial results and readiness to progress to the next research phase. Once our laboratory or the community confirms those candidates as the targets, the compound screening[74][75][76] could facilitate more personalized therapy for HNSCC patients.

3.5.3. Cancer Type-agnostic Study

Our strategy still has the strength to explore the more possible biomarkers from RNA-Seq datasets in cancer research. In our previous work, altered glucose metabolism—the Warburg effect[77] promotes the progression of HNSCC, which is partially attributed to the solute carrier family 2 member A4 (SLC2A4, or glucose transporters 4, GLUT4) and tripartite motif-containing 24 (TRIM24) pathway[78][79]. Lactic acidosis induced GLUT4 overexpression was also found in lung cancer cells[80]. Currently, Pembrolizumab and nivolumab's success has been based on a common biomarker (e.g., PD-1) crossing several types of cancer. It shows a model of tumor type-agnostic therapy[31]. There are several common biomarkers of immune-checkpoint inhibitor (ICI) under evaluation, including tumor-infiltrating lymphocytes (TIL), interferon gamma (IFN-γ), and tumor mutational burden (TMB)[23]. The other ICI, anti-LAG-3 (pelatlimab), is currently evaluated under the phase I/IIA[50] (ClinicalTrials.gov Identifier: NCT01968109) and II-IVA[81] (NCT04080804) studies.

In line with tumor-agnostic research, we plan to explore common biomarkers crossing TCGA diseases. However, the GDC provided standardized data frames that could not directly fit our workflow's scope. Before the global genes scanning process, it is necessary to re-format, transpose and, merge the 528 patients' clinical datasets and correlated 20,500 expressions of bio-specimen. This process should be carefully curated to confirm the data integrity within the correct definition[82]. We also plan to upgrade our R script of the cutoff engine to the C++ language and source it in the Rstudio server. The high performance of C++ could speed up the critical steps in this workflow involving heavy computation of matrix data[83]. Moreover, it is possible to introduce the Rstudio Shiny app (https://shiny.rstudio.com) as a web-based tools (named "pvalueTex") packaged with our workflow in the future.

3.5.4. Holistic Cancer Care

There are 81 physical, pathological and social conditions derived from participants available for survival modeling in the TCGA, such as age, gender, residual tumor, vital status, days-to-last-followup, cancer stage, smoking duration, exposure to alcohol, asbestos, radioactive radon. However, the TCGA did not collect other features related to holistic care. Going for holistic cancer care[84][85], spiritual and emotional condition is equally essential comparing with physical and social status (see Figure 6). Psychosocial stress is associated with cancer incidence[86][87][85], metastasis[86][88][89][90], and poor survival[91]. These impacts might be mediated through the hypothalamic-pituitary-adrenal (HPA) axis[92]. Holistic healthcare providers will engage his/her patients with eye-to-eye contact in terms of mind-to-mind connection. Their empathy, sympathy, and compassion are induced by the suffering of patients from those diseases. They will try to treat patients by prescribing medicine (and "themselves") or performing surgery. Thus, the healing resilience of patients will be induced by unconditional positive regard. The patients will trust those who take care of them and have the confidence to increase the capacity to recover from diseases through a mind-brain-body connection manner (see Figure 6).

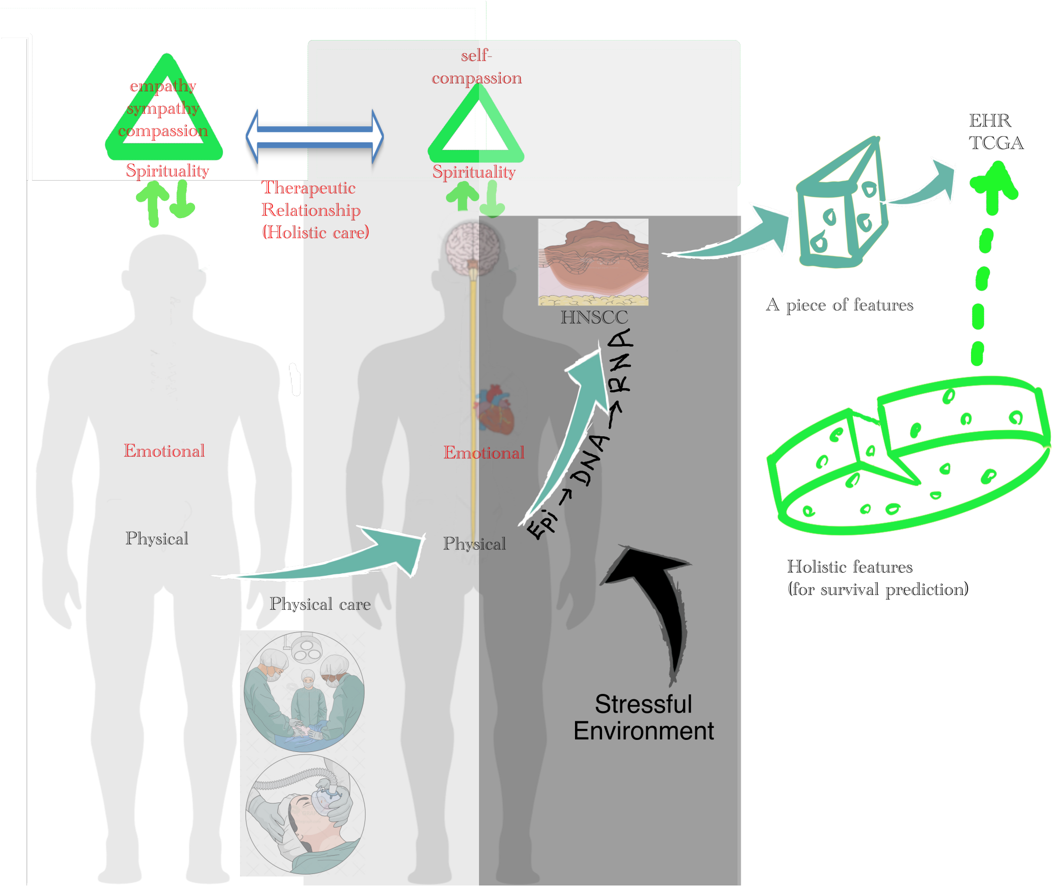

Figure 6. The concept of holistic care for HNSCC patients.
Beyond carcinogenesis - Under the mind-brain-body axis, a stressful environment (giant black arrow) will trigger emotional reception. The subconscious mind (brain) will release stress hormones and inflammation signals in response to the emotion. The physical body's internal environment (cells) alters epigenetic control in gene regulation and mRNA expression. In a long incubation time, the tissue/cells will be transformed into dysplasia then malignancy (e.g., HNSCC) with helping from known carcinogens.
Cancer care - Holistic care should take care of cancer patients' spiritual, emotional, physical, and socioeconomic needs. Physical care will be carried out by medication therapy or surgery. After establishing a therapeutic relationship (TR), the physicians' spiritual properties (empathy, sympathy, and compassion) will engage their patients and recover their self-compassion to gain resilience against the disease through their mind-brain-body axis. Thus, we suggest that electric healthcare records (EHR) should include physical, pathological, psychological data, and even more spiritual information. The TCGA might collect those "holistic features" (green dashed line) for further study of personalized medicine.

4. Materials and Methods

4.1. Patient Cohort

A large-scale cancer database, aggregating many independent features, is necessary to facilitate the biomarker discovery. The Cancer Genome Atlas (TCGA) project[93] has been developed since 2005 and supervised by the National Cancer Institute's (NCI) Center for Cancer Genomics and the National Human Genome Research Institute (NHGRI), funded by the US government. TCGA represents comprehensive genomics and clinic data from 84,392 patients among 33 major cancer types (data release 27.0 - Oct 29, 2020, available at https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/studied-cancers). TCGA and other collaborated Genome Data Analysis Center (GDAC) generated and analyzed DNA (mutation, copy number variation, methylation, simple nucleotide polymorphism, SNP), RNA (microarray, RNA-Seq, microRNA), and protein (reverse protein phased array) data derived from biospecimen. Sample types available at TCGA are primary solid tumors, recurrent solid tumors, blood-derived normal and tumor, metastatic, and solid tissue normal.

The NCI's Genomic Data Commons (GDC, available at https://portal.gdc.cancer.gov) receives, processes, and distributes genomic, clinical, and biospecimen data from TCGA database and other cancer research programs. The clinical features has been defined by TCGA GDC data dictionary: Common Data Element (CDE)[94]. The RNA-Seq expression data has been harmonized and re-aligned against an official reference genome build (Genome Reference Consortium Homo sapiens genome assembly 38, GRCh38). TCGA, GDC and some research communities offer several computational tools to the public for facilitating cancer research. GDC Data Portal is the official web-based TCGA data analysis tools. Other available web-based tools have been reviewed by Zhang et al.[95] and Matthieu Foll (availalbe at https://github.com/IARCbioinfo/awesome-TCGA). One of GDACs, the Broad TCGA Data and Analyses (Broad GDAC), provides Firehose, a repository of the TCGA public-accessible Level 3 data and Level 4 analyses. Broad GDAC Firehose is an analytical infrastructure that analyses algorithms not performed by the GDC (e.g., GISTIC, MutSig2CV, correlation with clinical variables, mRNA clustering). A web-based version of Broad GDAC Firehose is Firebrowse (available at firebrowse.org, Version: 1.1.40, 2019-10-13). Broad GDAC Firebrowse provides graphical tools like viewGene to explore expression levels and iCoMut to explore a mutation analysis of each TCGA disease.

GDC's Application Programmable Interface (API) is designed under the Representational State Transfer (REST) architecture and provides accessibility to external users for programmatic access to the same functionality found through GDC Portals. Those functions include searching, viewing, submitting, and downloading subsets of data files, metadata, and annotations based on specific parameters. Moreover, if restricted data is requested, the user must provide a token along with the API call. This token can be downloaded directly from the GDC Portals. (paragraph is provided here courtesy of the National Cancer Institute) Broad GDAC Firebrowse RESTful API could be accessed using an R package, FirebrowseR (available at https://github.com/mariodeng/FirebrowseR)[96].

GDC is available at https://portal.gdc.cancer.gov/projects/TCGA-HNSC. TCGA offers several computational tools to the public for facilitating cancer research. Broad Genome Data Analysis Center (GDAC) Firebrowse (firebrowse.org, version 1.1.35, 2016_09_27) is one of those tools to provide data access to each TCGA disease through a Representational State Transfer (REST) Application Programmable Interface (API). The 528 TCGA HNSCC patients' clinical information and RNA-Seq data were obtained from the Firebrowse RESTful API with an R package, FirebrowseR (available at https://github.com/mariodeng/FirebrowseR)[96]. We utilized FirebrowseR with our analysis workflow (see Figure 1, green square) to receive standardized data frames while avoiding data re-formatting, often causing some errors.

4.1.1. RNA Sequencing Data

The number of protein-coding genes was suggested as 20,500[97]. The GDC Data Portal provided TCGA data has been harmonized and re-aligned RNA sequencing data against an official reference genome build (Genome Reference Consortium Homo sapiens genome assembly 38, GRCh38). RNA-Seq expression level read counts produced by Illumina HiSeq have been normalized using the Fragments per kilobase per million reads mapped (FPKM) method, as described in reference[98]. The RNA-Seq preprocessor of Broad GDAC picked the RNA-Seq by Expectation-Maximization (RSEM) value from Illumina HiSeq/GA2 messenger RNA-Seq level_3 (v2) dataset of NCI GDC. It made the messenger RNA-Seq matrix with log2 transformed for the downstream analysis, as described in their reference[99]. We utilized FirebrowseR's function call, Samples.mRNASeq(cohort = "HNSC", gene=GeneName, format="csv"), to download each RNA-Seq data of all HNSCC patients and to save as 20,499 data frame files, named as "HNSCC.mRNA.Exp.[GeneName].Fire.Rda". After careful investigation of the genomics dataset, the RNA-Seq values of "solute carrier family 35 member E2A (SLC35E2A)" and "solute carrier family 35 member E2B (SLC35E2B)" should be considered as two distinct expression entities. We concluded that the number of protein-coding genes in the TCGA dataset is 20,500. We removed null expressed genes, which over 30% of the cohort, to avoid the useless result.

4.1.2. Clinical Data

We utilized FirebrowseR's function call, Samples.Clinical(cohort = "HNSC", format="csv"), to get all 81 clinical features (including pathological data, defined by TCGA GDC data dictionary: Common Data Element (CDE)[94]) of all 528 HNSCC patients, which saved as one data frame file: "HNSCC.clinical.Fire.Rda" (accessed November 2019).

One "HNSCC.clinical.Fire.Rda" tables and 20,500 "HNSCC.mRNA.Exp.[GeneName].Fire.Rda" tables were transposed and merged by their _participant_barcode (unique patient identification, ID) to yield a data frame with 528 rows (participants) against 20,581 columns (81 clinical features as well as 20,500 protein-coding RNA-Seq of cancer specimen). The clinicopathological features selected for our workflow included gender, age, clinical tumor size, clinical cervical lymph node metastases, clinical distant metastasis assessment, pathological surgical margin, and tobacco exposure with their corresponding survival data. The tumor size (T), cervical lymph node metastases (N), and distal metastasis status (M) were classified according to the American Joint Committee on Cancer (AJCC)[53] along with he Union for International Cancer Control (UICC)[100] TNM system for clinical staging of HNSCC. We made data clean by removing duplicated rows and columns.

4.2. Cutoff Finder Core Engine

To evaluate the effect of gene expression on the patient's survival, we introduce a sliding window cutoff selection by stratifying patients with Kaplan-Meier survival analysis according to each gene's low/high expression. Our cutofFinder_func subroutine employs the minimum *P* value approach to recognizing cutoff points in continuous gene expression measurement for patients sub-population. First, patients were ordered by RNA-Seq value (RSEM) of a given gene. Next, patients were stratified at a serial cut (counted by person ranked in 30% to 70% percentile of the cohort; please see Figure 1 cutoff engine). The survival risk differences of the two groups were estimated by log-rank test to yield around 165 Kaplan-Meier *P* values for each gene. Then, the optimal cutoff of RNA-Seq, giving the minimum *P* value, was selected by the cutofFinder_func subroutine. This iteration method could calculate all possible cutoff of each gene expression in this cohort. At each run of cutofFinder_func function call for an individual gene, it returned an optimal cutoff at specific patient grouping (e.g., low-expression in 262 persons versus high-expression in 152 persons with gene calcium/calmodulin dependent protein kinase

II inhibitor 1, CAMK2N1, please see Figure 5). The cutoff would be returned to the main program to allow downstream Cox survival analysis. The percentile range we applied as 30% to 70% was used to avoid a small grouping effect[101][47]. In case there was no significant P value, a median expression of this gene was set as its cutpoint as usual. The false discovery rate (FDR) (< 0.05) correction[102] shows which genes should be retained for subsequent univariate and multivariate analysis. It ensures the control of type I error of multiple test P values during our cutoff finding procedure. Then Bonferroni adjustment of that P values has been used for candidate selection.

4.3. Statistical Consideration for Survival Analysis

Our workflow has loops to call function survival_marginSFP(GeneName) with given GeneName to process the survival analysis gene by gene. We dichotomized the clinicopathological features, which includes gender (male/ female), age at diagnosed (below/beyond 65 year-old), clinical tumor size (T1-2/T3-4), clinical nodal status (negative/positive), clinical distant metastasis (negative/positive), TNM staging (early/late), surgical margin status (negative/positive) and tobacco exposure (low/high). The patients were grouped by an RNA-Seq value of each gene, cut at low- or high-expression on an optimal P value obtained from the cutofFinder_func subroutine (see the section of "Cutoff Finder Core Engine"). Pearson's chi-square test was used for these binary variables. Kaplan-Meier survival was analyzed using the log-rank test for two groups OS comparison.

The Cox proportional-hazards regression model[103][104] is commonly used for modeling survival data. It allows analyzing survival for one or more variables and to provides the effect size (coefficients, i.e., hazard ratios) for them[105]. The Cox model is also accounting for confounding factors[106]. It is expressed by the hazard function denoted by h(t). The hazard function represents the risk of a specific event (e.g., death) at time t. It can be estimated as follow:

$$h(t) = h_0(t) \times exp(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_n X_n)$$

where,

- t represents the survival time;
- $h(t)$ is the hazard function determined by a set of n covariates ($X_1 ... X_n$, e.g., clinicopathological features, including age, gender, gene expression, cancer stage (tumor size, nodal metastasis, distant metastasis), surgical margin, smoking, and alcohol; unfortunately, spiritual, emotional, and social status are not available in TCGA database;
- The coefficients ($\beta_1 ... \beta_n$) measure the impact—the effect size of covariates;
- The term h_0 is called the baseline hazard. It corresponds to the hazard value if all the X_i are equal to zero. The "t" in $h(t)$ indicates the hazard may vary over time.

Thus, the biomarker discovery strategy is survival modeling through a collection of $X_1 ... X_n$ features from cancer datasets.

Univariate Cox proportional regression model, using the "coxph" function in R package "survival", has been applied to calculate hazard ratio, 95% confidence interval (95% CI) and its significance, and to estimate the independent contributions of each clinicopathological features and gene expression to the overall survival individually.

In a multivariate test, those covariates used include the clinicopathological features (gender, age at diagnosed separated by 65 year-old, clinical tumor size [T1-2/T3-4], clinical nodal status [N0/N+], clinical distant metastasis [M0/M1], TNM staging [stage 1-2/stage 3-4], surgical margin status [negative/positive] and tobacco exposure [low/high]), and gene expression level [low/high] defined by the cutoff. All covariates have been pooled in the hazard function h(t) to estimate their impact on the overall survival.

Results were considered statistically significant when a two-sided $P < 0.05$, or a lower threshold if indicated. The false discovery rate (FDR) (< 0.05) could be used to pick up the optimal P value to ensure the control for type I error of the Kaplan-Meier survival test during the cutoff finding procedure. There were also multiple correlated tests of null hypotheses during our global scanning of 205,00 protein-coding genes. The stringent Bonferroni correction could result in an adjusted P value to ensure the control for type I error.

The resulting data, including Kaplan-Meier curves, cumulative P value plots, and Cox regression tables, were exported to ".xlsx" and ".Rda" files (by R package "r2excel") for subsequent biomarker selection.

4.4. Biomarker Selection and Validation

Those genes with prognostic impact, whose hazard ratio $>= 1.8$ or $<= 0.6$ in both Cox's univariate/multivariate model, were assigned as provisional candidates. Bonferroni adjusted (Kaplan-Meier) P value was used to make a ranking of candidates for the decision (see Figure 1, candidate selection).

GSE65858[51] is a dataset for helping of candidate selection in our workflow. The Gene Expression Omnibus (GEO) database[107], founded by National Center for Biotechnology Information (NCBI), is a public repository supporting MIAME-compliant data, including microarray and sequence-based experiments. R package of GEOquery[108] was used to download RNA-Seq dataset (in SOFT or MINiML format) of a HNSCC cohort, GSE65858, from the GEO database (available at https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE65858). GSE65858 has OS event, RFS event, and survival time. There are 270 HNSCC participants involved in this cohort. The expression data was generated using the platform GPL10558 (Illumina HumanHT-12 v4.0 Expression BeadChip), which targets on more than 30,330 annotated genes (47,000 probes, derived from the NCBI Reference Sequence, release 38 on 7 November, 2009). We have performed Kaplan-Meier (with FDR-correction of P value) and Cox survival analyses under the cutoff of each gene expression at their median value. The biomarker candidates were a consensus result of TCGA and GSE65858 analyses.

5. Conclusions

Our findings suggested 3 biomarker candidates—CAMK2N1, CALML5, and FCGBP, are all heavily associated with the prognosis of OS under optimal cutoff with stringent Bonferroni P values and proper effect size (HR).

The microenvironment of HNSCC, influenced by the mind-brain-body axis, requires further exploration and understanding using holistic multi-parametric approaches. Since mindfulness meditation will be helpful in cancer healthcare, we continually educate our cancer patients that they should repentantly confess for not taking care of body and spirit in the past, and sincere thanks for their physical body's hard working.

Supplementary Materials: The following are available online at https://www.mdpi.com/2075-4426/1/1/0/Supplementary_figures.pdf
Figure S1: P value plot of the gene NDFIP1
Table S1: 10 genes have unfavorable prognosis
Table S2: 10 genes have favorable prognosis

Author Contributions: Conceptualization, Yu-Chuan (Jack) Li; Data curation, Alexander TH Wu; Formal analysis, Li-Hsing Chi; Investigation, Michael Hsiao; Methodology, Michael Hsiao; Resources, Alexander TH Wu; Software, Li-Hsing Chi and Yu-Chuan (Jack) Li; Supervision, Michael Hsiao and Yu-Chuan (Jack) Li; Writing – original draft, Li-Hsing Chi; Writing – review and editing, Alexander TH Wu. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded by Taipei Medical University

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: All data process and analyses were performed with R programming language (https://www.r-project.org/, version 4.0.2 2020-06-22) and R packages "firebrowseR", "survival", "reshape", "data.table", "ggplot2", "R.utils", "xlsx", "r2excel", "rJava" and "rms" at Rstudio server (version 1.2.5001) based on Google cloud platform under operation system Linux (Ubuntu LTS, release v18.04.3). The R script codes and datasets generated during the current study are available at the GitHub repository, https://github.com/texchi2/pvalueTex.

Acknowledgments: The results shown here are in whole based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. The tables were generated by latex code with the help from https://www.tablesgenerator.com and https://github.com/JDMCreator /LaTeXTableEditor (accessed at https://www.latex-tables.com/v1/index.html). We like to thank Dr. Wen-Chang Wang, College of Medical Science and Technology, Taipei Medical University, for bioinformatics consultation and helps. We like to thank Sylvain Hallé, Department of Computer Science and Mathematics, Univerité du Québec à Chicoutimi, Canada, for latex technique help with his TeXtidote. We sincerely want to express our thanks to the HNSCC patients who donated their data to University Hospital Leipzig (GSE65858) and TCGA affiliated biobanks.

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

Abbreviations

The following abbreviations are used in this manuscript:

95% CI, 95% confidence interval; AJCC, the American Joint Committee on Cancer; API, Application Programmable Interface; CALML5, calmodulin like 5; CAMK2N1, calcium/calmodulin dependent protein kinase II inhibitor 1; CDE, Common Data Element; CTLA-4, cytotoxic T lymphocyte antigen 4; DEGs, differentially expressed genes; DOI, depth of invasion; EGFR, epidermal growth factor receptor; EMT, epithelial-mesenchymal-transition; ENE, extra-nodal extension; ERK, extracellular signal-regulated kinases; FCGBP, Fc fragment of IgG binding protein; FDA, Food and Drug Administration; FDR, false discovery rate; FPKM, Fragments per kilobase per million reads mapped; GDAC, Genome Data Analysis Center; GDC, Genomic Data Commons; GITR, glucocorticoid-induced tumor necrosis factor receptor; GLUT4, glucose transporters 4; GRCh38, Genome Reference Consortium Homo sapiens genome assembly 38; HER, human epidermal growth factor receptor; HIF, hypoxia-inducible factor; HNSCC, head and neck squamous cell carcinoma; HPA, the Human Protein Atlas; HPV, human papillomavirus; HR, hazard ratio; HRAS, Harvey rat sarcoma viral oncoprotein; ICI, immune-checkpoint inhibitor; ID, identification; IFN-γ, interferon gamma; IHC, immunohistochemistry; LAG-3, lymphocyte activation gene 3; LND, lymph node density; LVSI, lymph-vascular space invasion; MMP, matrix metalloproteinase; NCBI, National Center for Biotechnology Information; NCCN, National Comprehensive Cancer Network; NCI, the National Cancer Institute; OS, overall survival; PD-1, programmed death 1; PD-L1, programmed death ligand 1; PNI, perineural invasion; RAS, rat sarcoma; REST, Representational State Transfer; RNA, ribonucleic acid; RNA-Seq, RNA sequencing; RPMA, reverse-phase protein lysate microarray; RPPAs, reverse-phase protein arrays; RSEM, RNA-Seq by Expectation-Maximization; RT, radiation therapy; SLC2A4, solute carrier family 2 member A4; SLC35E2A, solute carrier family 35 member E2A; SLC35E2B, solute carrier family 35 member E2B; Tax-PF, docetaxel, cisplatin, and 5-fluorouracil; TCGA, the Cancer Genome Atlas; TCPA, the Cancer Proteome Atlas; TIGIT, T cell immunoglobin and immunoreceptor tyrosine-based inhibitory motif; TIL, tumor-infiltrating lymphocytes; TIM-3, T-cell immunoglobulin mucin protein 3; TKI, tyrosine kinase inhibitor; TMB, tumor mutational burden; TMSB4X, thymosin beta-4 X-linked; TNM, the tumor size (T), cervical lymph node metastases (N), and distal metastasis status (M); TRIM24, tripartite motif-containing 24; UICC, he Union for International Cancer Control; US, United States; VISTA, V-domain Ig suppressor of T-cell activation; WPOI-5, worst pattern of invasion score 5

References

1. MOHW. 2017 Statistics of Causes of Death, 2018.
2. MOHW. 2018 Statistics of General Health and Welfare, 2018.
3. Pfister, D.G.; Spencer, S.; Adelstein, D.; Adkins, D.; Anzai, Y.; Brizel, D.M.; Bruce, J.Y.; Busse, P.M.; Caudell, J.J.; Cmelak, A.J.; Colevas, A.D.; Eisele, D.W.; Fenton, M.; Foote, R.L.; Galloway, T.; Gillison, M.L.; Haddad, R.I.; Hicks, W.L.; Hitchcock, Y.J.; Jimeno, A.; Leizman, D.; Maghami, E.; Mell, L.K.; Mittal, B.B.; Pinto, H.A.; Ridge, J.A.; Rocco, J.W.; Rodriguez, C.P.; Shah, J.P.; Weber, R.S.; Weinstein, G.; Witek, M.; Worden, F.; Yom, S.S.; Zhen, W.; Burns, J.L.; Darlow, S.D. Head and Neck Cancers, Version 2.2020, NCCN Clinical Practice Guidelines in Oncology. *Journal of the National Comprehensive Cancer Network J Natl Compr Canc Netw* **2020**, *18*, 873–898. doi:10.6004/jnccn.2020.0031.
4. HPA. Statistics of Health Promotion 2017, 2019.
5. De Vicente, J.C.; Gutiérrez, L.M.J.; Zapatero, A.H.; Forcelledo, M.F.F.; Hernández-Vallejo, G.; López Arranz, J.S. Prognostic significance of p53 expression in oral squamous cell carcinoma without neck node metastases. *Head and Neck* **2004**, *26*, 22–30. doi:10.1002/hed.10339.
6. Aebersold, D.M.; Burri, P.; Beer, K.T.; Laissue, J.; Djonov, V.; Greiner, R.H.; Semenza, G.L. Expression of hypoxia-inducible factor-1α: A novel predictive and prognostic parameter in the radiotherapy of oropharyngeal cancer. *Cancer Research* **2001**, *61*, 2911–2916.
7. Couture, C.; Raybaud-Diogène, H.; Têtu, B.; Bairati, I.; Murry, D.; Allard, J.; Fortin, A. p53 and Ki-67 as markers of radioresistance in head and neck carcinoma. *Cancer* **2002**, *94*, 713–722. doi:10.1002/cncr.10232.
8. O-charoenrat, P.; Modjtahedi, H.; Rhys-Evans, P.; Court, W.J.; Box, G.M.; Eccles, S.A. Epidermal growth factor-like ligands differentially up-regulate matrix metalloproteinase 9 in head and neck squamous carcinoma cells. *Cancer Research* **2000**, *60*, 1121–1128.
9. Bentzen, S.M.; Atasoy, B.M.; Daley, F.M.; Dische, S.; Richman, P.I.; Saunders, M.I.; Trott, K.R.; Wilson, G.D. Epidermal growth factor receptor expression in pretreatment biopsies from head and neck squamous cell carcinoma as a predictive factor for a benefit from accelerated radiation therapy in a randomized controlled trial. *Journal of Clinical Oncology* **2005**, *23*, 5560–5567. doi:10.1200/JCO.2005.06.411.
10. Harrington, K.J. Chemotherapy and Targeted Agents. In *Maxillofacial Surgery*; Elsevier, 2017; pp. 339–354. doi:10.1016/B978-0-7020-6056-4.00022-8.
11. Bonner, J.A.; Harari, P.M.; Giralt, J.; Azarnia, N.; Shin, D.M.; Cohen, R.B.; Jones, C.U.; Sur, R.; Raben, D.; Jassem, J.; Ove, R.; Kies, M.S.; Baselga, J.; Youssoufian, H.; Amellal, N.; Rowinsky, E.K.; Ang, K.K. Radiotherapy plus cetuximab for squamous-cell carcinoma of the head and neck. *New England Journal of Medicine* **2006**, *354*, 567–578. doi:10.1056/NEJMoa053422.
12. Vermorken, J.B.; Mesia, R.; Rivera, F.; Remenar, E.; Kawecki, A.; Rottey, S.; Erfan, J.; Zabolotnyy, D.; Kienzer, H.R.; Cupissol, D.; Peyrade, F.; Benasso, M.; Vynnychenko, I.; De Raucourt, D.; Bokemeyer, C.; Schueler, A.; Amellal, N.; Hitt, R. Platinum-based chemotherapy plus cetuximab in head and neck cancer. *New England Journal of Medicine* **2008**, *359*, 1116–1127. doi:10.1056/NEJMoa0802656.
13. Rivera, F.; García-Castaño, A.; Vega, N.; Vega-Villegas, M.E.; Gutiérrez-Sanz, L. Cetuximab in metastatic or recurrent head and neck cancer: The EXTREME trial. *Expert Review of Anticancer Therapy* **2009**, *9*, 1421–1428. doi:10.1586/ERA.09.113.
14. Blanchard, P.; Bourhis, J.; Lacas, B.; Posner, M.R.; Vermorken, J.B.; Hernandez, J.J.C.; Bourredjem, A.; Calais, G.; Paccagnella, A.; Hitt, R.; Pignon, J.P. Taxane-Cisplatin-Fluorouracil As Induction Chemotherapy in Locally Advanced Head and Neck Cancers: An Individual Patient Data Meta-Analysis of the Meta-Analysis of Chemotherapy in Head and Neck Cancer Group. *Journal of Clinical Oncology* **2013**, *31*, 2854–2860. doi:10.1200/JCO.2012.47.7802.
15. Rampias, T.; Giagini, A.; Siolos, S.; Matsuzaki, H.; Sasaki, C.; Scorilas, A.; Psyrri, A. RAS/PI3K crosstalk and cetuximab resistance in head and neck squamous cell carcinoma. *Clinical Cancer Research* **2014**, *20*, 2933–2946. doi:10.1158/1078-0432.CCR-13-2721.
16. Gazzah, A.; Boni, V.; Soria, J.C.; Calles, A.; Even, C.; Doger, B.; Mahjoubi, L.; Bahleda, R.; Ould-Kaci, M.; Esler, A.; Nazabadioko, S.; Calvo, E. A phase 1b study of afatinib in combination with standard-dose cetuximab in patients with advanced solid tumours. *European Journal of Cancer* **2018**, *104*, 1–8. doi:10.1016/j.ejca.2018.07.011.
17. Taberna, M.; Oliva, M.; Mesía, R. Cetuximab-containing combinations in locally advanced and recurrent or metastatic head and neck squamous cell carcinoma, 2019. doi:10.3389/fonc.2019.00383.
18. Seiwert, T.Y.; Burtness, B.; Weiss, J.; Gluck, I.; Eder, J.P.; Pai, S.I.; Dolled-Filhart, M.; Emancipator, K.; Pathiraja, K.; Gause, C.; Iannone, R.; Brown, H.; Houp, J.; Cheng, J.D.; Chow, L.Q.M. A phase Ib study of MK-3475 in patients with human papillomavirus (HPV)-associated and non-HPV-associated head and neck (H/N) cancer. *Journal of Clinical Oncology* **2014**, *32*, 6011.
19. Swanson, M.S.; Sinha, U.K. Rationale for combined blockade of PD-1 and CTLA-4 in advanced head and neck squamous cell cancer - Review of current data. *Oral Oncology* **2015**, *51*, 12–15. doi:10.1016/j.oraloncology.2014.10.010.
20. Mei, Z.; Huang, J.; Qiao, B.; yin Lam, A.K. Immune checkpoint pathways in immunotherapy for head and neck squamous cell carcinoma, 2020. doi:10.1038/s41368-020-0084-8.
21. Cramer, J.D.; Burtness, B.; Le, Q.T.; Ferris, R.L. The changing therapeutic landscape of head and neck cancer, 2019. doi:10.1038/s41571-019-0227-z.
22. Burtness, B.; Harrington, K.J.; Greil, R.; Soulières, D.; Tahara, M.; de Castro, G.; Psyrri, A.; Basté, N.; Neupane, P.; Bratland, Å.; Fuereder, T.; Hughes, B.G.; Mesía, R.; Ngamphaiboon, N.; Rordorf, T.; Wan Ishak, W.Z.; Hong, R.L.; González Mendoza, R.; Roy, A.; Zhang, Y.; Gumuscu, B.; Cheng, J.D.; Jin, F.; Rischin, D.; Lerzo, G.; Tatangelo, M.; Varela, M.; Zarba, J.J.; Boyer, M.; Gan, H.;

Gao, B.; Hughes, B.G.; Mallesara, G.; Taylor, A.; Burian, M.; Barrios, C.H.; de Castro Junior, D.O.; Castro, G.; Franke, F.A.; Girotto, G.; Lima, I.P.F.; Nicolau, U.R.; Pinto, G.D.J.; Santos, L.; Victorino, A.P.; Chua, N.; Couture, F.; Gregg, R.; Hansen, A.; Hilton, J.; McCarthy, J.; Soulieres, D.; Ascui, R.; Gonzalez, P.; Villanueva, L.; Torregroza, M.; Zambrano, A.; Holeckova, P.; Kral, Z.; Melichar, B.; Prausova, J.; Vosmik, M.; Andersen, M.; Gyldenkerne, N.; Jurgens, H.; Putnik, K.; Reinikainen, P.; Gruenwald, V.; Laban, S.; Aravantinos, G.; Boukovinas, I.; Georgoulias, V.; Kwong, D.; Al-Farhat, Y.; Csoszi, T.; Erfan, J.; Horvai, G.; Landherr, L.; Remenar, E.; Ruzsa, A.; Szota, J.; Billan, S.; Gluck, I.; Gutfeld, O.; Popovtzer, A.; Benasso, M.; Bui, S.; Ferrari, V.; Licitra, L.; Nole, F.; Fujii, T.; Fujimoto, Y.; Hanai, N.; Hara, H.; Matsumoto, K.; Mitsugi, K.; Monden, N.; Nakayama, M.; Okami, K.; Oridate, N.; Shiga, K.; Shimizu, Y.; Sugasawa, M.; Takahashi, M.; Takahashi, S.; Tanaka, K.; Ueda, T.; Yamaguchi, H.; Yamazaki, T.; Yasumatsu, R.; Yokota, T.; Yoshizaki, T.; Kudaba, I.; Stara, Z.; Cheah, S.K.; Aguilar Ponce, J.; Gonzalez Mendoza, R.; Hernandez Hernandez, C.; Medina Soto, F.; Buter, J.; Hoeben, A.; Oosting, S.; Suijkerbuijk, K.; Bratland, A.; Brydoey, M.; Alvarez, R.; Mas, L.; Caguioa, P.; Querol, J.; Regala, E.E.; Tamayo, M.B.; Villegas, E.M.; Kawecki, A.; Karpenko, A.; Klochikhin, A.; Smolin, A.; Zarubenkov, O.; Goh, B.C.; Cohen, G.; du Toit, J.; Jordaan, C.; Landers, G.; Ruff, P.; Szpak, W.; Tabane, N.; Brana, I.; Iglesias Docampo, L.; Lavernia, J.; Mesia, R.; Abel, E.; Muratidu, V.; Nielsen, N.; Cristina, V.; Rothschild, S.; Wang, H.M.; Yang, M.H.; Yeh, S.P.; Yen, C.J.; Soparattanapaisarn, N.; Sriuranpong, V.; Aksoy, S.; Cicin, I.; Ekenel, M.; Harputluoglu, H.; Ozyilkan, O.; Harrington, K.J.; Agarwala, S.; Ali, H.; Alter, R.; Anderson, D.; Bruce, J.; Campbell, N.; Conde, M.; Deeken, J.; Edenfield, W.; Feldman, L.; Gaughan, E.; Goueli, B.; Halmos, B.; Hegde, U.; Hunis, B.; Jotte, R.; Karnad, A.; Khan, S.; Laudi, N.; Laux, D.; Martincic, D.; McCune, S.; McGaughey, D.; Misiukiewicz, K.; Mulford, D.; Nadler, E.; Nunnink, J.; Ohr, J.; O'Malley, M.; Patson, B.; Paul, D.; Popa, E.; Powell, S.; Redman, R.; Rella, V.; Rocha Lima, C.; Sivapiragasam, A.; Su, Y.; Sukari, A.; Wong, S.; Yilmaz, E.; Yorio, J. Pembrolizumab alone or with chemotherapy versus cetuximab with chemotherapy for recurrent or metastatic squamous cell carcinoma of the head and neck (KEYNOTE-048): a randomised, open-label, phase 3 study. *The Lancet* **2019**, *394*, 1915–1928. doi:10.1016/S0140-6736(19)32591-7.

23. Gavrielatou, N.; Doumas, S.; Economopoulou, P.; Foukas, P.G.; Psyrri, A. Biomarkers for immunotherapy response in head and neck cancer. *Cancer Treatment Reviews* **2020**, *84*, 101977. doi:10.1016/j.ctrv.2020.101977.
24. Chi, L.H.H.; Chang, W.M.M.; Chang, Y.C.C.; Chan, Y.C.C.; Tai, C.C.C.; Leung, K.W.W.; Chen, C.L.L.; Wu, A.T.; Lai, T.C.C.; Li, Y.C.C.; Hsiao, M. Global Proteomics-based Identification and Validation of Thymosin Beta-4 X-Linked as a Prognostic Marker for Head and Neck Squamous Cell Carcinoma. *Scientific Reports* **2017**, *7*, 9031. doi:10.1038/s41598-017-09539-w.
25. Zhang, Y.; Feurino, L.W.; Zhai, Q.; Wang, H.; Fisher, W.E.; Chen, C.; Yao, Q.; Li, M. Thymosin beta 4 is overexpressed in human pancreatic cancer cells and stimulates proinflammatory cytokine secretion and JNK activation. *Cancer Biology and Therapy* **2008**, *7*, 419–423. doi:10.4161/cbt.7.3.5415.
26. Ryu, Y.K.; Lee, Y.S.; Lee, G.H.; Song, K.S.; Kim, Y.S.; Moon, E.Y. Regulation of glycogen synthase kinase-3 by thymosin beta-4 is associated with gastric cancer cell migration. *International journal of cancer. Journal international du cancer* **2012**, *131*, 2067–77. doi:10.1002/ijc.27490.
27. Gemoll, T.; Strohkamp, S.; Schillo, K.; Thorns, C.; Jens, K. MALDI-imaging reveals thymosin beta-4 as an independent prognostic marker for colorectal cancer. *Oncotarget* **2015**, *6*, 43869–43880. doi:10.18632/oncotarget.6103.
28. Huang, D.; Wang, S.; Wang, A.; Chen, X.; Zhang, H. Thymosin beta 4 silencing suppresses proliferation and invasion of non-small cell lung cancer cells by repressing Notch1 activation. *Acta Biochimica et Biophysica Sinica* **2016**, *48*, 788–794. doi:10.1093/abbs/gmw070.
29. Chu, Y.; You, M.; Zhang, J.; Gao, G.; Han, R.; Luo, W.; Liu, T.; Zuo, J.; Wang, F. Adipose-Derived Mesenchymal Stem Cells Enhance Ovarian Cancer Growth and Metastasis by Increasing Thymosin Beta 4X-Linked Expression. *Stem Cells International* **2019**, *2019*. doi:10.1155/2019/9037197.
30. Makowiecka, A.; Malek, N.; Mazurkiewicz, E.; Mrówczyńska, E.; Nowak, D.; Mazur, A.J. Thymosin β4 Regulates Focal Adhesion Formation in Human Melanoma Cells and Affects Their Migration and Invasion. *Frontiers in Cell and Developmental Biology* **2019**, *7*, 1–16. doi:10.3389/fcell.2019.00304.
31. Yan, L.; Zhang, W. Precision medicine becomes reality-tumor type-agnostic therapy. *Cancer communications (London, England)* **2018**, *38*, 6. doi:10.1186/s40880-018-0274-3.
32. Tomczak, K.; Czerwińska, P.; Wiznerowicz, M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge, 2015. doi:10.5114/wo.2014.47136.
33. Loraine, A.E.; Blakley, I.C.; Jagadeesan, S.; Harper, J.; Miller, G.; Firon, N. Analysis and visualization of RNA-Seq expression data using RStudio, Bioconductor, and Integrated Genome Browser. *Methods in molecular biology (Clifton, N.J.)* **2015**, *1284*, 481–501. doi:10.1007/978-1-4939-2444-8_24.
34. Tonella, L.; Giannoccaro, M.; Alfieri, S.; Canevari, S.; De Cecco, L. Gene Expression Signatures for Head and Neck Cancer Patient Stratification: Are Results Ready for Clinical Application? *Current Treatment Options in Oncology* **2017**, *18*. doi:10.1007/s11864-017-0472-2.
35. Zhao, X.; Sun, S.; Zeng, X.; Cui, L. Expression profiles analysis identifies a novel three-mRNA signature to predict overall survival in oral squamous cell carcinoma. *American journal of cancer research* **2018**, *8*, 450–461.
36. Li, R.; Qu, H.; Wang, S.; Wei, J.; Zhang, L.; Ma, R.; Lu, J.; Zhu, J.; Zhong, W.D.D.; Jia, Z.; We, J.; Zhang, L.; Ma, R.; Lu, J.; Zhu, J.; Zhong, W.D.D.; Jia, Z. GDCRNATools: An R/Bioconductor package for integrative analysis of lncRNA, miRNA and mRNA data in GDC. *Bioinformatics* **2018**, *34*, 2515–2517. doi:10.1093/bioinformatics/bty124.
37. Huang, G.z.; Wu, Q.q.; Zheng, Z.n.; Shao, T.r.; Lv, X.Z. Identification of Candidate Biomarkers and Analysis of Prognostic Values in Oral Squamous Cell Carcinoma, 2019.

38. Shen, Y.; Liu, J.; Zhang, L.; Dong, S.; Zhang, J.; Liu, Y.; Zhou, H.; Dong, W. Identification of Potential Biomarkers and Survival Analysis for Head and Neck Squamous Cell Carcinoma Using Bioinformatics Strategy: A Study Based on TCGA and GEO Datasets. *BioMed Research International* **2019**, *2019*, 7376034. doi:10.1155/2019/7376034.
39. Schmitt, K.; Molfenter, B.; Laureano, N.K.; Tawk, B.; Bieg, M.; Hostench, X.P.; Weichenhan, D.; Ullrich, N.D.; Shang, V.; Richter, D.; Stögbauer, F.; Schroeder, L.; de Bem Prunes, B.; Visioli, F.; Rados, P.V.; Jou, A.; Plath, M.; Federspil, P.A.; Thierauf, J.; Döscher, J.; Weissinger, S.E.; Hoffmann, T.K.; Wagner, S.; Wittekindt, C.; Ishaque, N.; Eils, R.; Klussmann, J.P.; Holzinger, D.; Plass, C.; Abdollahi, A.; Freier, K.; Weichert, W.; Zaoui, K.; Hess, J. Somatic mutations and promotor methylation of the ryanodine receptor 2 is a common event in the pathogenesis of head and neck cancer. *International Journal of Cancer* **2019**, *145*, 3299–3310. doi:10.1002/ijc.32481.
40. Xu, C.; Zhang, Y.; Shen, Y.; Shi, Y.; Zhang, M.; Zhou, L. Integrated Analysis Reveals ENDOU as a Biomarker in Head and Neck Squamous Cell Carcinoma Progression, 2021.
41. Tang, Z.; Li, C.; Kang, B.; Gao, G.; Li, C.; Zhang, Z. GEPIA: A web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Research* **2017**, *45*, W98–W102. doi:10.1093/nar/gkx247.
42. Tang, Z.; Kang, B.; Li, C.; Chen, T.; Zhang, Z. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Research* **2019**, *47*, W556–W560. doi:10.1093/nar/gkz430.
43. Li, C.; Tang, Z.; Zhang, W.; Ye, Z.; Liu, F. GEPIA2021: integrating multiple deconvolution-based analysis into GEPIA. *Nucleic Acids Research* **2021**. doi:10.1093/nar/gkab418.
44. Uhlen, M.; Zhang, C.; Lee, S.; Sjöstedt, E.; Fagerberg, L.; Bidkhori, G.; Benfeitas, R.; Arif, M.; Liu, Z.; Edfors, F.; Sanli, K.; Von Feilitzen, K.; Oksvold, P.; Lundberg, E.; Hober, S.; Nilsson, P.; Mattsson, J.; Schwenk, J.M.; Brunnström, H.; Glimelius, B.; Sjöblom, T.; Edqvist, P.H.; Djureinovic, D.; Micke, P.; Lindskog, C.; Mardinoglu, A.; Ponten, F. A pathology atlas of the human cancer transcriptome. *Science* **2017**, *357*, eaan2507. doi:10.1126/science.aan2507.
45. Aguirre-Gamboa, R.; Gomez-Rueda, H.; Martínez-Ledesma, E.; Martínez-Torteya, A.; Chacolla-Huaringa, R.; Rodriguez-Barrientos, A.; Tamez-Peña, J.G.; Treviño, V. SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. *PloS one* **2013**, *8*, e74250. doi:10.1371/journal.pone.0074250.
46. Abel, U.; Berger, J.; Wiebelt, H. CRITLEVEL: An Exploratory Procedure for the Evaluation of Quantitative Prognostic Factors. *Methods of Information in Medicine* **1984**, *23*, 154–156. doi:10.1055/s-0038-1635335.
47. Mizuno, H.; Kitada, K.; Nakai, K.; Sarai, A. PrognoScan: A new database for meta-analysis of the prognostic value of genes. *BMC Medical Genomics* **2009**, *2*, 18. doi:10.1186/1755-8794-2-18.
48. Budczies, J.; Klauschen, F.; Sinn, B.V.; Gyorffy, B.; Schmitt, W.D.; Darb-Esfahani, S.; Denkert, C. Cutoff Finder: A Comprehensive and Straightforward Web Application Enabling Rapid Biomarker Cutoff Optimization. *PLOS ONE* **2012**, *7*, 1–7. doi:10.1371/journal.pone.0051862.
49. Chang, C.; Hsieh, M.K.; Chang, W.Y.; Chiang, A.J.; Chen, J. Determining the optimal number and location of cutoff points with application to data of cervical cancer. *PLoS ONE* **2017**, *12*. doi:10.1371/journal.pone.0176231.
50. Cristina, V.; Herrera-Gómez, R.G.; Szturz, P.; Espeli, V.; Siano, M. Immunotherapies and future combination strategies for head and neck squamous cell carcinoma. *International Journal of Molecular Sciences* **2019**, *20*. doi:10.3390/ijms20215399.
51. Wichmann, G.; Rosolowski, M.; Krohn, K.; Kreuz, M.; Boehm, A.; Reiche, A.; Scharrer, U.; Halama, D.; Bertolini, J.; Bauer, U.; Holzinger, D.; Pawlita, M.; Hess, J.; Engel, C.; Hasenclever, D.; Scholz, M.; Ahnert, P.; Kirsten, H.; Hemprich, A.; Wittekind, C.; Herbarth, O.; Horn, F.; Dietz, A.; Loeffler, M. The role of HPV RNA transcription, immune response-related gene expression and disruptive TP53 mutations in diagnostic and prognostic profiling of head and neck cancer. *International Journal of Cancer* **2015**, *137*, 2846–2857. doi:10.1002/ijc.29649.
52. Cheraghlou, S.; Otremba, M.; Kuo Yu, P.; Agogo, G.O.; Hersey, D.; Judson, B.L. Prognostic Value of Lymph Node Yield and Density in Head and Neck Malignancies, 2018. doi:10.1177/0194599818756830.
53. Amin, M.B.; Greene, F.L.; Edge, S.B.; Compton, C.C.; Gershenwald, J.E.; Brookland, R.K.; Meyer, L.; Gress, D.M.; Byrd, D.R.; Winchester, D.P. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA: A Cancer Journal for Clinicians* **2017**, *67*, 93–99. doi:10.3322/caac.21388.
54. Brandwein-Gensler, M.; Smith, R.V.; Wang, B.; Penner, C.; Theilken, A.; Broughel, D.; Schiff, B.; Owen, R.P.; Smith, J.; Sarta, C.; Hebert, T.; Nason, R.; Ramer, M.; De Lacure, M.; Hirsch, D.; Myssiorek, D.; Heller, K.; Prystowsky, M.; Schlecht, N.F.; Negassa, A. Validation of the histologic risk model in a new cohort of patients with head and neck squamous cell carcinoma. *American Journal of Surgical Pathology* **2010**, *34*, 676–688. doi:10.1097/PAS.0b013e3181d95c37.
55. Sinha, N.; Rigby, M.H.; McNeil, M.L.; Taylor, S.M.; Trites, J.R.; Hart, R.D.; Bullock, M.J. The histologic risk model is a useful and inexpensive tool to assess risk of recurrence and death in stage i or II squamous cell carcinoma of tongue and floor of mouth. *Modern Pathology* **2018**, *31*, 772–779. doi:10.1038/modpathol.2017.183.
56. Bulbul, M.G.; Tarabichi, O.; Sethi, R.K.; Parikh, A.S.; Varvares, M.A. Does Clearance of Positive Margins Improve Local Control in Oral Cavity Cancer? A Meta-analysis. *Otolaryngology - Head and Neck Surgery (United States)* **2019**, *161*, 235–244. doi:10.1177/0194599819839006.
57. Scholl, P.; Byers, R.M.; Batsakis, J.G.; Wolf, P.; Santini, H. Microscopic cut-through of cancer in the surgical treatment of squamous carcinoma of the tongue. Prognostic and therapeutic implications. *The American Journal of Surgery* **1986**, *152*, 354–360. doi:10.1016/0002-9610(86)90304-1.

58. Sutton, D.N.; Brown, J.S.; Rogers, S.N.; Vaughan, E.D.; Woolgar, J.A. The prognostic implications of the surgical margin in oral squamous cell carcinoma, 2003. doi:10.1054/ijom.2002.0313.
59. Shaw, R.J.; Brown, J.S.; Woolgar, J.A.; Lowe, D.; Rogers, S.N.; Vaughan, E.D. The influence of the pattern of mandibular invasion on recurrence and survival in oral squamous cell carcinoma, 2004. doi:10.1002/hed.20036.
60. Guillemaud, J.P.; Patel, R.S.; Goldstein, D.P.; Higgins, K.M.; Enepekides, D.J. Prognostic impact of intraoperative microscopic cut-through on frozen section in oral cavity squamous cell carcinoma. *Journal of Otolaryngology - Head and Neck Surgery* **2010**, *39*, 370–7. doi:10.2310/7070.2010.090084.
61. Patel, R.S.; Goldstein, D.P.; Guillemaud, J.; Bruch, G.A.; Brown, D.; Gilbert, R.W.; Gullane, P.J.; Higgins, K.M.; Irish, J.; Enepekides, D.J.; Leoncini, E.; Ricciardi, W.; Cadoni, G.; Arzani, D.; Petrelli, L.; Paludetti, G.; Brennan, P.; Luce, D.; Stucker, I.; Matsuo, K.; Talamini, R.; La Vecchia, C.; Olshan, A.F.; Winn, D.M.; Herrero, R.; Franceschi, S.; Castellsague, X.; Muscat, J.; Morgenstern, H.; Zhang, Z.F.; Levi, F.; Dal Maso, L.; Kelsey, K.; McClean, M.; Vaughan, T.L.; Lazarus, P.; Purdue, M.P.; Hayes, R.B.; Chen, C.; Schwartz, S.M.; Shangina, O.; Koifman, S.; Ahrens, W.; Matos, E.; Lagiou, P.; Lissowska, J.; Szeszenia-Dabrowska, N.; Fernandez, L.; Menezes, A.; Agudo, A.; Daudt, A.W.; Richiardi, L.; Kjaerheim, K.; Mates, D.; Betka, J.; Yu, G.P.; Schantz, S.; Simonato, L.; Brenner, H.; Conway, D.I.; Macfarlane, T.V.; Thomson, P.; Fabianova, E.; Znaor, A.; Rudnai, P.; Healy, C.; Boffetta, P.; Chuang, S.C.; Lee, Y.C.; Hashibe, M.; Boccia, S. Impact of positive frozen section microscopic tumor cut-through revised to negative on oral carcinoma control and survival rates. *Head & Neck* **2010**, *32*, 1444–1451. doi:10.1002/HED.
62. Kuriakose, M.A.; Trivedi, N.P. *Contemporary Oral Oncology*, 1 ed.; Vol. 2, Springer International Publishing: Switzerland, 2017; pp. 147–187. doi:10.1007/978-3-319-14917-2.
63. Shapiro, M.; Salama, A. Margin Analysis: Squamous Cell Carcinoma of the Oral Cavity. *Oral and Maxillofacial Surgery Clinics of North America* **2017**, *29*, 259–267. doi:10.1016/j.coms.2017.03.003.
64. Saidak, Z.; Clatot, F.; Chatelain, D.; Galmiche, A. A gene expression profile associated with perineural invasion identifies a subset of HNSCC at risk of post-surgical recurrence. *Oral Oncology* **2018**, *86*, 53–60. doi:10.1016/j.oraloncology.2018.09.005.
65. Migueláñez-Medrán, B.D.C.; Pozo-Kreilinger, J.J.; Cebrián-Carretero, J.L.; Martínez-García, M.Á.; López-Sánchez, A.F. Oral squamous cell carcinoma of tongue: Histological risk assessment. A pilot study. *Medicina Oral Patologia Oral y Cirugia Bucal* **2019**, *24*, e603–e609. doi:10.4317/medoral.23011.
66. Saidak, Z.; Pascual, C.; Bouaoud, J.; Galmiche, L.; Clatot, F.; Dakpé, S.; Page, C.; Galmiche, A. A three-gene expression signature associated with positive surgical margins in tongue squamous cell carcinomas: Predicting surgical resectability from tumour biology? *Oral Oncology* **2019**, *94*, 115–120. doi:10.1016/j.oraloncology.2019.05.020.
67. Sullivan, G.M.; Feinn, R. Using Effect Size—or Why the P Value Is Not Enough . *Journal of Graduate Medical Education* **2012**, *4*, 279–282. doi:10.4300/jgme-d-12-00156.1.
68. Thiese, M.S.; Ronna, B.; Ott, U. P value interpretations and considerations. *Journal of Thoracic Disease* **2016**, *8*, E928–E931. doi:10.21037/jtd.2016.08.16.
69. Li, J.; Lu, Y.; Akbani, R.; Ju, Z.; Roebuck, P.L.; Liu, W.; Yang, J.Y.; Broom, B.M.; Verhaak, R.G.; Kane, D.W.; Wakefield, C.; Weinstein, J.N.; Mills, G.B.; Liang, H. TCPA: A resource for cancer functional proteomics data. *Nature Methods* **2013**, *10*, 1046–1047. doi:10.1038/nmeth.2650.
70. Li, J.; Akbani, R.; Zhao, W.; Lu, Y.; Weinstein, J.N.; Mills, G.B.; Liang, H. Explore, Visualize, and Analyze Functional Cancer Proteomic Data Using the Cancer Proteome Atlas. *Cancer research* **2017**, *77*, e51–e54. doi:10.1158/0008-5472.CAN-17-0369.
71. Chen, M.J.M.; Li, J.; Wang, Y.; Akbani, R.; Lu, Y.; Mills, G.B.; Liang, H. TCPA v3.0: An integrative platform to explore the pan-cancer analysis of functional proteomic data. *Molecular and Cellular Proteomics* **2019**, *18*, S15–S25. doi:10.1074/mcp.RA118.001260.
72. Stodden, V. *Implementing Reproducible Research*; Chapman and Hall/CRC, 2014; chapter 1, pp. 185–218. doi:10.1201/b16868.
73. Kulkarni, N.; Alessandrì, L.; Panero, R.; Arigoni, M.; Olivero, M.; Ferrero, G.; Cordero, F.; Beccuti, M.; Calogero, R.A. Reproducible bioinformatics project: A community for reproducible bioinformatics analysis pipelines. *BMC Bioinformatics* **2018**, *19*, 349. doi:10.1186/s12859-018-2296-x.
74. Yang, J.M.; Chen, C.C. GEMDOCK: A Generic Evolutionary Method for Molecular Docking. *Proteins: Structure, Function and Genetics* **2004**, *55*, 288–304. doi:10.1002/prot.20035.
75. Hsu, K.C.; Chen, Y.F.; Lin, S.R.; Yang, J.M. Igemdock: A graphical environment of enhancing gemdock using pharmacological interactions and post-screening analysis. *BMC Bioinformatics* **2011**, *12*, S33. doi:10.1186/1471-2105-12-S1-S33.
76. Pathak, N.; Chen, Y.T.; Hsu, Y.C.; Hsu, N.Y.; Kuo, C.J.; Tsai, H.P.; Kang, J.J.; Huang, C.H.; Chang, S.Y.; Chang, Y.H.; Liang, P.H.; Yang, J.M. Uncovering flexible active site conformations of SARS-COV-2 3Cl proteases through protease pharmacophore clusters and covid-19 drug repurposing. *ACS Nano* **2021**, *15*, 857–872. doi:10.1021/acsnano.0c07383.
77. Warburg, O. On the Origin of Cancer Cells. *Science* **1956**, *123*, 309–314.
78. Chang, Y.C.; Chi, L.H.; Chang, W.M.; Su, C.Y.; Lin, Y.F.; Chen, C.L.; Chen, M.H.; Chang, P.M.H.; Wu, A.T.H.; Hsiao, M. Glucose transporter 4 promotes head and neck squamous cell carcinoma metastasis through the TRIM24-DDX58 axis. *Journal of Hematology & Oncology* **2017**, *10*, 11. doi:10.1186/s13045-016-0372-0.
79. Mani, S.; Swargiary, G.; Singh, K.K. Natural agents targeting mitochondria in cancer, 2020. doi:10.3390/ijms21196992.
80. Prado-Garcia, H.; Campa-Higareda, A.; Romero-Garcia, S. Lactic Acidosis in the Presence of Glucose Diminishes Warburg Effect in Lung Adenocarcinoma Cells. *Frontiers in Oncology* **2020**, *10*, 807. doi:10.3389/fonc.2020.00807.
81. Neal, M.E.H.; Haring, C.T.; Mann, J.E.; Brenner, J.C.; Spector, M.E.; Swiecicki, P.L. Novel immunotherapeutic approaches in head and neck cancer. *Journal of Cancer Metastasis and Treatment* **2019**, *2019*. doi:10.20517/2394-4722.2019.32.

82. Rendleman, M.C.; Buatti, J.M.; Braun, T.A.; Smith, B.J.; Nwakama, C.; Beichel, R.R.; Brown, B.; Casavant, T.L. Machine learning with the TCGA-HNSC dataset: Improving usability by addressing inconsistency, sparsity, and high-dimensionality. *BMC Bioinformatics* **2019**, *20*, 339. doi:10.1186/s12859-019-2929-8.
83. Woodward, S.J.R.; Beukes, P.C.; Hanigan, M.D. Molly reborn in C++ and R. *animal* **2020**, *14*, s250–s256. doi:DOI: 10.1017/S1751731120000270.
84. Mehta, R.; Sharma, K.; Potters, L.; Wernicke, A.G.; Parashar, B. Evidence for the Role of Mindfulness in Cancer: Benefits and Techniques. *Cureus* **2019**, *11*, e4629. doi:10.7759/cureus.4629.
85. Iftikhar, A.; Islam, M.; Shepherd, S.; Jones, S.; Ellis, I. Cancer and stress: Does it make a difference to the patient when these two challenges collide? *Cancers* **2021**, *13*, 1–29. doi:10.3390/cancers13020163.
86. Lutgendorf, S.K.; Sood, A.K.; Antoni, M.H. Host factors and cancer progression: Biobehavioral signaling pathways and interventions, 2010. doi:10.1200/JCO.2009.26.9357.
87. Powell, N.D.; Tarr, A.J.; Sheridan, J.F. Psychosocial stress and inflammation in cancer, 2013. doi:10.1016/j.bbi.2012.06.015.
88. Moreno-Smith, M.; Lutgendorf, S.K.; Sood, A.K. Impact of stress on cancer metastasis. *Future Oncology* **2010**, *6*, 1863–1881. doi:10.2217/fon.10.142.
89. Du, P.; Zeng, H.; Xiao, Y.; Zhao, Y.; Zheng, B.; Deng, Y.; Liu, J.; Huang, B.; Zhang, X.; Yang, K.; Jiang, Y.; Ma, X. Chronic stress promotes EMT-mediated metastasis through activation of STAT3 signaling pathway by miR-337-3p in breast cancer. *Cell Death and Disease* **2020**, *11*, 761. doi:10.1038/s41419-020-02981-1.
90. Xu, X.R.; Xiao, Q.; Hong, Y.C.; Liu, Y.H.; Liu, Y.; Tu, J. Activation of dopaminergic VTA inputs to the mPFC ameliorates chronic stress-induced breast tumor progression. *CNS Neuroscience and Therapeutics* **2021**, *27*, 206–219. doi:10.1111/cns.13465.
91. Chida, Y.; Hamer, M.; Wardle, J.; Steptoe, A. Do stress-related psychosocial factors contribute to cancer incidence and survival? *Nature clinical practice. Oncology* **2008**, *5*, 466–475. doi:10.1038/ncponc1134.
92. Hsiao, F.H.; Jow, G.M.; Kuo, W.H.; Chang, K.J.; Liu, Y.F.; Ho, R.T.H.; Ng, S.M.; Chan, C.L.W.; Lai, Y.M.; Chen, Y.T. The Effects of Psychotherapy on Psychological Well-Being and Diurnal Cortisol Patterns in Breast Cancer Survivors. *Psychotherapy and Psychosomatics* **2012**, *81*, 173–182. doi:10.1159/000329178.
93. Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.; Ozenberger, B.A.; Ellrott, K.; Sander, C.; Stuart, J.M.; Chang, K.; Creighton, C.J.; Davis, C.; Donehower, L.; Drummond, J.; Wheeler, D.; Ally, A.; Balasundaram, M.; Birol, I.; Butterfield, Y.S.; Chu, A.; Chuah, E.; Chun, H.J.E.; Dhalla, N.; Guin, R.; Hirst, M.; Hirst, C.; Holt, R.A.; Jones, S.J.; Lee, D.; Li, H.I.; Marra, M.A.; Mayo, M.; Moore, R.A.; Mungall, A.J.; Robertson, A.G.; Schein, J.E.; Sipahimalani, P.; Tam, A.; Thiessen, N.; Varhol, R.J.; Beroukhim, R.; Bhatt, A.S.; Brooks, A.N.; Cherniack, A.D.; Freeman, S.S.; Gabriel, S.B.; Helman, E.; Jung, J.; Meyerson, M.; Ojesina, A.I.; Pedamallu, C.S.; Saksena, G.; Schumacher, S.E.; Tabak, B.; Zack, T.; Lander, E.S.; Bristow, C.A.; Hadjipanayis, A.; Haseley, P.; Kucherlapati, R.; Lee, S.; Lee, E.; Luquette, L.J.; Mahadeshwar, H.S.; Pantazi, A.; Parfenov, M.; Park, P.J.; Protopopov, A.; Ren, X.; Santoso, N.; Seidman, J.; Seth, S.; Song, X.; Tang, J.; Xi, R.; Xu, A.W.; Yang, L.; Zeng, D.; Auman, J.T.; Balu, S.; Buda, E.; Fan, C.; Hoadley, K.A.; Jones, C.D.; Meng, S.; Mieczkowski, P.A.; Parker, J.S.; Perou, C.M.; Roach, J.; Shi, Y.; Silva, G.O.; Tan, D.; Veluvolu, U.; Waring, S.; Wilkerson, M.D.; Wu, J.; Zhao, W.; Bodenheimer, T.; Hayes, D.N.; Hoyle, A.P.; Jeffreys, S.R.; Mose, L.E.; Simons, J.V.; Soloway, M.G.; Baylin, S.B.; Berman, B.P.; Bootwalla, M.S.; Danilova, L.; Herman, J.G.; Hinoue, T.; Laird, P.W.; Rhie, S.K.; Shen, H.; Triche, T.; Weisenberger, D.J.; Carter, S.L.; Cibulskis, K.; Chin, L.; Zhang, J.; Sougnez, C.; Wang, M.; Getz, G.; Dinh, H.; Doddapaneni, H.V.; Gibbs, R.; Gunaratne, P.; Han, Y.; Kalra, D.; Kovar, C.; Lewis, L.; Morgan, M.; Morton, D.; Muzny, D.; Reid, J.; Xi, L.; Cho, J.; Dicara, D.; Frazer, S.; Gehlenborg, N.; Heiman, D.I.; Kim, J.; Lawrence, M.S.; Lin, P.; Liu, Y.; Noble, M.S.; Stojanov, P.; Voet, D.; Zhang, H.; Zou, L.; Stewart, C.; Bernard, B.; Bressler, R.; Eakin, A.; Iype, L.; Knijnenburg, T.; Kramer, R.; Kreisberg, R.; Leinonen, K.; Lin, J.; Liu, Y.; Miller, M.; Reynolds, S.M.; Rovira, H.; Shmulevich, I.; Thorsson, V.; Yang, D.; Zhang, W.; Amin, S.; Wu, C.J.; Wu, C.C.; Akbani, R.; Aldape, K.; Baggerly, K.A.; Broom, B.; Casasent, T.D.; Cleland, J.; Dodda, D.; Edgerton, M.; Han, L.; Herbrich, S.M.; Ju, Z.; Kim, H.; Lerner, S.; Li, J.; Liang, H.; Liu, W.; Lorenzi, P.L.; Lu, Y.; Melott, J.; Nguyen, L.; Su, X.; Verhaak, R.; Wang, W.; Wong, A.; Yang, Y.; Yao, J.; Yao, R.; Yoshihara, K.; Yuan, Y.; Yung, A.K.; Zhang, N.; Zheng, S.; Ryan, M.; Kane, D.W.; Aksoy, B.A.; Ciriello, G.; Dresdner, G.; Gao, J.; Gross, B.; Jacobsen, A.; Kahles, A.; Ladanyi, M.; Lee, W.; Lehmann, K.V.; Miller, M.L.; Ramirez, R.; Rätsch, G.; Reva, B.; Schultz, N.; Senbabaoglu, Y.; Shen, R.; Sinha, R.; Sumer, S.O.; Sun, Y.; Taylor, B.S.; Weinhold, N.; Fei, S.; Spellman, P.; Benz, C.; Carlin, D.; Cline, M.; Craft, B.; Goldman, M.; Haussler, D.; Ma, S.; Ng, S.; Paull, E.; Radenbaugh, A.; Salama, S.; Sokolov, A.; Swatloski, T.; Uzunangelov, V.; Waltman, P.; Yau, C.; Zhu, J.; Hamilton, S.R.; Abbott, S.; Abbott, R.; Dees, N.D.; Delehaunty, K.; Ding, L.; Dooling, D.J.; Eldred, J.M.; Fronick, C.C.; Fulton, R.; Fulton, L.L.; Kalicki-Veizer, J.; Kanchi, K.L.; Kandoth, C.; Koboldt, D.C.; Larson, D.E.; Ley, T.J.; Lin, L.; Lu, C.; Magrini, V.J.; Mardis, E.R.; McLellan, M.D.; McMichael, J.F.; Miller, C.A.; O'Laughlin, M.; Pohl, C.; Schmidt, H.; Smith, S.M.; Walker, J.; Wallis, J.W.; Wendl, M.C.; Wilson, R.K.; Wylie, T.; Zhang, Q.; Burton, R.; Jensen, M.A.; Kahn, A.; Pihl, T.; Pot, D.; Wan, Y.; Levine, D.A.; Black, A.D.; Bowen, J.; Frick, J.; Gastier-Foster, J.M.; Harper, H.A.; Helsel, C.; Leraas, K.M.; Lichtenberg, T.M.; McAllister, C.; Ramirez, N.C.; Sharpe, S.; Wise, L.; Zmuda, E.; Chanock, S.J.; Davidsen, T.; Demchok, J.A.; Eley, G.; Felau, I.; Sheth, M.; Sofia, H.; Staudt, L.; Tarnuzzer, R.; Wang, Z.; Yang, L.; Zhang, J.; Omberg, L.; Margolin, A.; Raphael, B.J.; Vandin, F.; Wu, H.T.; Leiserson, M.D.; Benz, S.C.; Vaske, C.J.; Noushmehr, H.; Wolf, D.; Veer, L.V.T.; Anastassiou, D.; Yang, T.H.O.; Lopez-Bigas, N.; Gonzalez-Perez, A.; Tamborero, D.; Xia, Z.; Li, W.; Cho, D.Y.; Przytycka, T.; Hamilton, M.; McGuire, S.; Nelander, S.; Johansson, P.; Jörnsten, R.; Kling, T. The cancer genome atlas pan-cancer analysis project, 2013. doi:10.1038/ng.2764.
94. NCI Genomic Data Commons. GDC Data Dictionary, 2019.
95. Zhang, Z.; Li, H.; Jiang, S.; Li, R.; Li, W.; Chen, H.; Bo, X. A survey and evaluation of Web-based tools/databases for variant analysis of TCGA data. *Briefings in Bioinformatics* **2018**, *20*, 1524–1541. doi:10.1093/bib/bby023.

96. Deng, M.; Brägelmann, J.; Kryukov, I.; Saraiva-Agostinho, N.; Perner, S. FirebrowseR: An R client to the Broad Institute's Firehose Pipeline. *Database* **2017**, *2017*. doi:10.1093/database/baw160.
97. Clamp, M.; Fry, B.; Kamal, M.; Xie, X.; Cuff, J.; Lin, M.F.; Kellis, M.; Lindblad-Toh, K.; Lander, E.S. Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **2007**, *104*, 19428–19433. doi:10.1073/pnas.0709013104.
98. NCI Genomic Data Commons. mRNA Analysis Pipeline, 2017.
99. GDAC. Samples Report, 2016.
100. Brierley, J.D.; Gospodarowicz, M.K.; Wittekind, C. *TNM Classification of Malignant Tumours, 8th Edition*; Wiley-Blackwell: Hoboken, 2016; p. 272.
101. Halpern, J. Maximally Selected Chi Square Statistics for Small Samples. *Biometrics* **1982**, *38*, 1017. doi:10.2307/2529882.
102. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **1995**, *57*, 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.
103. Cox, D.R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **1972**, *34*, 187–202. doi:https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.
104. Andersen, P.K.; Gill, R.D. Cox's Regression Model for Counting Processes: A Large Sample Study. *Annals of Statistics* **1982**, *10*, 1100–1120.
105. Bradburn, M.J.; Clark, T.G.; Love, S.B.; Altman, D.G. Survival analysis part II: multivariate data analysis–an introduction to concepts and methods. *British journal of cancer* **2003**, *89*, 431–6. doi:10.1038/sj.bjc.6601119.
106. Magen, A.; Das Sahu, A.; Lee, J.S.; Sharmin, M.; Lugo, A.; Gutkind, J.S.; Schäffer, A.A.; Ruppin, E.; Hannenhalli, S. Beyond Synthetic Lethality: Charting the Landscape of Pairwise Gene Expression States Associated with Survival in Cancer. *Cell Reports* **2019**, *28*, 938–948.e6. doi:10.1016/j.celrep.2019.06.067.
107. Chung, C.H.; Parker, J.S.; Ely, K.; Carter, J.; Yi, Y.; Murphy, B.A.; Ang, K.K.; El-Naggar, A.K.; Zanation, A.M.; Cmelak, A.J.; Levy, S.; Slebos, R.J.; Yarbrough, W.G. Gene expression profiles identify epithelial-to-mesenchymal transition and activation of nuclear factor-kappaB signaling as characteristics of a high-risk head and neck squamous cell carcinoma. *Cancer research* **2006**, *66*, 8210–8218. doi:10.1158/0008-5472.CAN-06-1213.
108. Sean, D.; Meltzer, P.S. GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **2007**, *23*, 1846–1847. doi:10.1093/bioinformatics/btm254.