A Global Genome-Wide Scan with Optimal Cutoff Mining for Emerging Biomarkers in Head and Neck Squamous Cell Carcinoma

Li-Hsing Chi

The Ph.D. Program for Translational Medicine, College of Medical Science and Technology, Taipei Medical University and Academia Sinica, Taipei, Taiwan https://orcid.org/0000-0002-4476-2600

Alexander TH Wu

The Ph.D. Program for Translational Medicine, College of Medical Science and Technology, Taipei Medical University and Academia Sinica, Taipei, Taiwan

Michael Hsiao

Genomics Research Center, Academia Sinica, Taipei, Taiwan

Yu-Chuan (Jack) Li (jaak88@gmail.com)

Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan

Research

Keywords: HNSCC, TCGA, RNA-sequencing, Survival Analysis, Optimal Cutoff, Biomarker, Tumor Type-agnostic Therapy, Surgical Margin, Rstudio

Posted Date: September 28th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-80673/v1

RESEARCH

A Global Genome-wide Scan with Optimal Cutoff Mining for Emerging Biomarkers in Head and Neck Squamous Cell Carcinoma

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

[*]Correspondence:
jaak88@gmail.com
[1]The Ph.D. Program for Translational Medicine, College of Medical Science and Technology, Taipei Medical University and Academia Sinica, Taipei, Taiwan
Full list of author information is available at the end of the article
[†]Corresponding author

Abstract

Background: 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). In order to utilize a continuous gene expression for survival analysis, it is necessary to determine a cutoff point by the dichotomization of the patients. There is some optimization software for cutoff determination. However, those predetermined cutoffs by software usually set at the median, 1/4 quantile, or 3/4 quantile of RNA sequencing (RNA-Seq) value to find a significant P-value of the Kaplan-Meier curve. There are few clinicopathological features available on their pre-processed data sets.

Methods: We developed a comprehensive workflow by R script, running on the Rstudio platform. It includes data retrieving and pre-processing, feature selection, cutoff mining engine, Kaplan-Meier survival analysis, Cox proportional hazard modeling, and biomarker discovery.

Results: Using this workflow on the TCGA HNSCC cohort, we scanned human protein-coding genes (20,500) programmatically. After adjustment with other confounders, we found that the clinical tumor stage and the surgical margin involvement are independent risk factors in patient survival. According to the resulting tables with Bonferroni adjusted P-value under optimal cutoff as well as hazard ratio ($>= 1.5$), there were ten candidate biomarkers, named as DKK1, CAMK2N1, STC2, PGK1, SURF4, USP10, NDFIP1, FOXA2, STIP1, and DKC1, which are significantly associated with the poor prognosis of overall survival (OS). At the same time, the other ten genes were over-expressed in the better survival patients (with hazard ratio $<= 0.5$), named as ZNF557, ZNF266, IL19, MYO1H, FCGBP, LOC148709, EVPLL, PNMA5, IQCN (previous name as KIAA1683), and NPB. Further validations are warranted.

Conclusions: We suggested this analysis tool equipped with an optimal cutoff finder will help with biomarker discovery of protein-coding genes, in terms of tumor-agnostic therapy.

Keywords: HNSCC; TCGA; RNA-sequencing; Survival Analysis; Optimal Cutoff; Biomarker; Tumor Type-agnostic Therapy; Surgical Margin; Rstudio

Background

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, which guide the development of modern systemic therapy.

Accumulative knowledge showed that some biomarkers have prognostic significance in HNSCC. For example, node-negative HNSCC patients with p53 overexpression were found to have 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 over-expressed to promote invasion and metastasis of HNSCC. From 2000 to 2006, the 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 was approved by the US Food and Drug Administration (FDA) in 2008. In advance, the bio-RT could have proceeded with docetaxel, cisplatin, and 5-fluorouracil (Tax-PF) induction chemotherapy to overcome the radio-resistance of HNSCC[14].

However, Rampias and his colleagues suggested 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[15]. 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 TKI, 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. So far, cetuximab is still the only drug of choice with proven efficacy, which 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, which 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, pembrolizumab, and nivolumab, as a monotherapy for the platinum-treated patients of recurrent or metastatic HNSCC[21]. However, because of the complexity of immune-tumor interaction, ICI has not to guarantee the response to programmed death ligand 1 (PD-L1) expressed HNSCC[19]. 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].

In our previous proteomic study in 2017, thymosin beta-4 X-linked (TMSB4X) was reported to be related to tumor growth and metastasis of HNSCC[24]. It was also found by the subsequent investigations that TMSB4X 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 possible for tumor type-agnostic therapy[31] as a common biomarker crossing several types of cancer.

In summary, identifying predictive biomarkers for selecting standard-of-care or advanced systemic therapy[32] in HNSCC is crucial. However, there are three challenges of biomarker discovery from survival analysis, so far. Firstly, although TCGA genomics data were harmonized, there is unclean data, including null expressed genes, which over 50% of the cohort, should be manually investigated and cleaned. Second, we need to find a way to determine candidates from the expression level of 20,500 human protein-coding genes[33]. Usually, the investigators should get the rationale or revelation of the genes of interest on a specific cancer type. They should upload those genes manually onto bioinformatics tools, such as SurvExpress (http://bioinformatica.mty.itesm.mx:8080/Biomatec/SurvivaX.jsp, which has been lost since Oct/2019 and currently out of funds), analyze with TCGA cohort. 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. Third, we need to find an optimal cutpoint of those RNA expression data to maximize candidate mining coverage. The above mentioned online tools might set a cutpoint 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, such as Prognoscan[34], Cutoff Finder[35], Findcut[36], Human protein atlas[37], OptimalCutpoints[32], 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 combine the scanning of the protein-coding genes and cutoff optimization programmatically.

In our approach, this article described a comprehensive workflow implemented in the R script, which runs on the Rstudio server. Its functions include data retrieving and pre-processing, feature selection, cutoff mining engine, Kaplan-Meier survival analysis, Cox proportional hazard modeling, and biomarker selection. Using this workflow on the TCGA HNSCC cohort, the 20,500 human protein-coding genes were scanned. The analysis workflow is shown in Figure 1.

Materials and Methods

Patient Cohort

The Cancer Genome Atlas (TCGA) profiled 528 HNSCC clinical and genomic data, which has been standardized and available on a unified data portal, Genomic Data

Commons (GDC) of the the National Cancer Institute (NCI). GDC is available on 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)[38]. We utilized FirebrowseR with our analysis workflow (see Figure 1, step 1) to receive standardized data frames while avoiding data re-formatting, often causing some errors.

RNA Sequencing Data

The number of protein-coding genes was suggested as 20,500[33]. 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[39]. 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[40]. 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 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 50% of the cohort, to avoid the useless result.

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)[41]) 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 tcga_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)[42] along with he Union for International Cancer Control (UICC)[43] TNM system for clinical staging of HNSCC. We made data clean by removing duplicated rows and columns.

Cutoff Finder Core Engine

To evaluate the effect of gene expression on the patient's survival, we introduced the stratifying of 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 "HNSCC cohort"). 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 value (e.x. 0.027 for gene calcium/calmodulin dependent protein kinase II inhibitor 1, CAMK2N1). The optimal cutoff value and its correlated patient grouping size (e.x. low-expression in 262 persons vs. high-expression in 152 persons with gene CAMK2N1) 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[44][34]. In case there was no significant P-value, a median expression of this gene was set as its cutpoint as usual.

Statistical Consideration

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 is the widely accepted approach for modeling survival while accounting for confounding factors[45]. Univariate and multivariate Cox proportional regression model[46], using the "coxph" function in R package "survival", was applied to calculate hazard ratio, 95% confidence interval (95% CI) and its significance, and to estimate the independent contributions of each clinicopathological features to the OS. Results were considered statistically significant when a two-sided P-value < 0.05,

or a lower threshold if indicated. There were multiple correlated tests in the family of Kaplan-Meier survival hypotheses during our global scanning of 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.

Biomarker Selection

Those genes with prognostic impact, whose hazard ratio >= 1.5 or <= 0.5 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 final decision (see Figure 1, step 2).

Results

The 9416 Kaplan-Meier plots with associated Cox's univariate and multivariate tables were generated at workflow step 1 (see Figure 1) and justified by the ranking of hazard ratios. By uncorrected P-value below 0.05, we selected 967 genes in which hazard ratio (HR) is greater than 1.5 or less than 0.5 (see Figure 2A univariate, and Figure 2B multivariate plots). At the final step, a Bonferroni P-value correction was used to yield the twenty candidate genes, under the stringent criteria (see Figure 2C, D). The ten candidates, including DKK1, CAMK2N1, STC2, PGK1, SURF4, USP10, NDFIP1, FOXA2, STIP1, and DKC1, have significantly associated with the poor prognosis of OS (see Table 1), while the other ten genes were over-expressed in the better survival patients, named as ZNF557, ZNF266, IL19, MYO1H, FCGBP, LOC148709, EVPLL, PNMA5, IQCN (previous name as KIAA1683), and NPB (see Table 3, with their full name of genes). We made a volcano plot for 9416 genes by Kaplan-Meier P-value (less than 0.05, obtained during cutoff finding procedure) against the Cox hazard ratio (see Figure 3). The plot revealed that the most significant (Bonferroni-adjusted P-value < 0.05) candidate genes are located above the dotted line. At the same time, Cox's HR separated them on the two-side with prognostic impact.

Our top 1 candidate is dickkopf WNT signaling pathway inhibitor 1 (DKK1) (see Figure 4A). The Kaplan-Meier curve revealed 227 patients bearing the higher expression of DKK1 were suffered from only 40% of 5-year OS rate. In comparison, the other 187 patients with lower expression (the cutoff at -0.312(RSEM)) have been a better prognosis (adjusted P-value as 0.001). Figure 4B's cumulative P-value plot showed that the uncorrected 116 P-values (< 0.05) were estimated by a serial cut from 125 to 290 persons for grouping the cohort in our cutoff finding procedure (cutofFinder_func.R, see Figure 1, "HNSCC cohort"). The smallest P-value (8.9×10^{-8}), when cut on n=187 (45.2% of 414), was defined as optimal P-value. Conversely, the most associated gene with better survival is zinc finger protein 557 (ZNF557). In Figure 4C, a Kaplan-Meier curve revealed 264 patients bearing the higher expression of ZNF557 have 55% of 5-year OS survival rate (adjusted P-value = 0.001). The cutoff finding procedure (cutofFinder_func.R) generated cumulative P-value plots in Figure 4D. There were 166 uncorrected P-values estimated by a

serial cut from 125 to 290 for grouping the cohort. All was less than 0.05, and the smallest P-value (8.6×10^{-8}), when cut on n=150 (36.2% of total cohort 414), was defined as optimal P-value with a cutoff value -0.511(RSEM) of RNA-Seq.

Table 1 showed ten candidate genes over-expressed with poor prognosis in HNSCC, ranked by adjusted Kaplan-Meier P-value. We found their Cox's univariate and multivariate HR are all greater than 1.837. There were few published articles of surfeit 4 (SURF4) and NEDD4 family interacting protein 1 (NDFIP1) (NEDD4: neural precursor cell expressed, developmentally down-regulated 4), which were related to cancer research. In Table 2, after adjustment of confounders, it was considered the DKK1 over-expression is the independent prognostic factor (multivariate HR 2.135 [95% CI: 1.559-2.924, P-value < 0.001]), as well as clinical T stage (HR 1.978 [95% CI: 1.046-3.737, P-value $= 0.036$]) and surgical margins status (HR 1.601 [95% CI: 1.159-2.211, P-value $= 0.004$]). The age older than 65 year-old has negative influence on survival (HR 1.462 [95% CI: 1.078-1.983, P-value $= 0.015$]). The M stage could not be considered due to only 3 out of 414 patients which have distant metastasis.

There were also ten candidate genes over-expressed with a better prognosis of HNSCC, listed in Table 3. Cox's univariate and multivariate HR is just under 0.5. In Table 4, after adjustment of confounders, it revealed HR 1.961 [95% CI: 1.035-3.714, P-value $= 0.039$] in advance clinical T Status, HR 1.631 [95% CI: 1.18-2.254, P-value $= 0.003$] with positive surgical margin involvement, HR 1.453 [95% CI: 1.055-2.000, P-value $= 0.022$] with higher tobacco exposure, and a protective HR 0.499 [95% CI: 0.372-0.669, P-value < 0.001] in over-expressed ZNF557 gene.

In overall results, those 20 candidate biomarkers, clinical T stage, and surgical margin are independent prognosis factors in HNSCC.

Discussion

Besides ethnicity, age, gender, the TNM stage, radiation therapy, chemotherapy, and targeted therapy, the comprehensive adverse features of prognosis in HNSCC should 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)[47], and worst pattern of invasion score 5 (WPOI-5), which was defined as tumor dispersion $>= 1mm$ between tumor satellites or positive PNI/LVSI[42]. The features of DOI, LND, and tumor dispersion were not available on the TCGA dataset. The Brandwein-Gensler's risk model (lymphocytic host response, WPOI-5, and PNI)[48][49] was 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[50]. According to Table 2 and Table 4 in our study, the positive surgical margin yielded a hazard ratio greater than 1.6 to influence on patient's OS. It was suggested by authors [51][52][53][54][55][56][57][58][59][60] 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 is suggested as an independent surrogate for tumor dispersion, and important in the HNSCC study.

Thus, in the current study, we selected the common clinicopathological features, including gender, age, clinical T, clinical N, clinical M, surgical margin status, and tobacco exposure in the biomarker discovery for adjustment of confounders (details description at Materials and Methods section).

In our previous work, altered glucose metabolism (e.g., the Warburg effect[61]) 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[62]. In this study, LOC148709 (a long non-coding RNA (lncRNA)) was suggested as a biomarker of HNSCC (see Table 3). It was also found to has a contribution to the Warburg effect on esophagus cancer[63].

There is a trend toward cancer type-agnostic study. The success of pembrolizumab and nivolumab was based on a common biomarker (e.g., PD-1) crossing several types of cancer. It showed a precedent of tumor type-agnostic therapy[31]. Currently, there are several common biomarkers of ICI under evaluation, which include 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[32](ClinicalTrials.gov Identifier: NCT01968109) and II-IVA[64](ClinicalTrials.gov Identifier: NCT04080804) studies.

Although we combined the power of genome-wide scanning and an optimal cutoff finder for Kaplan-Meier 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) excited the cancer research community[65][66]. By the utility of the reverse-phase protein arrays (RPPAs) or reverse-phase protein lysate microarray (RPMA), a microarray of "Western blots" in the TCPA could help to test our hypotheses. However, there are only 237 antibodies available so far. We found our 20 candidates are not included in the TCPA database (v3.0[67]). Our strategy still has the strength to explore the more possible biomarkers from RNA-Seq datasets in cancer research. In line with tumor-agnostic research, we plan to explore more TCGA diseases to find common biomarkers. However, the GDC provided standardized data frames that could not directly fit our workflow's scope. Before the global genes scanning process, we needed to re-format, transpose and, merge the 528 patients' clinical datasets and correlated 20,500 expressions of bio-specimen. It should be carefully curated to confirm the data integrity with the correct definition. We also plan to upgrade our plain R script at the Rstudio platform to program in the C++ language and source it in R. The high performance of C++ could speed up the cutoff finding engine in this workflow involving heavy computations[68]. Thus, it is possible to introduce the Rstudio Shiny app (https://shiny.rstudio.com) as a web application of the "pvalueTex" packaged with our workflow in the future.

Conclusion

Our findings suggested 20 candidate biomarkers, DKK1, CAMK2N1, STC2, PGK1, SURF4, USP10, NDFIP1, FOXA2, STIP1, DKC1, as well as ZNF557, ZNF266, IL19, MYO1H, FCGBP, LOC148709, EVPLL, PNMA5, IQCN (previous name as

KIAA1683), and NPB, are all heavily associated with the prognosis of OS under optimal cutoff points with stringent Bonferroni P-values and proper confounders control. They also might be potential common biomarkers for subsequent study. We wish this bioinformatics tool will be available for the broad usage of tumor-agnostic research[69] to cross several TCGA diseases to make translational impacts.

List of abbreviations
95% CI 95% confidence interval

AJCC the American Joint Committee on Cancer
API Application Programmable Interface

CAMK2N1 calcium/calmodulin dependent protein kinase II inhibitor 1
CDE Common Data Element
CTLA-4 cytotoxic T lymphocyte antigen 4

DKK1 dickkopf WNT signaling pathway inhibitor 1
DOI depth of invasion

EGFR epidermal growth factor receptor
EMT epithelial-mesenchymal-transition
ENE extra-nodal extension
ERK extracellular signal-regulated kinases

FDA Food and Drug Administration
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
HPV human papillomavirus
HR hazard ratio
HRAS Harvey rat sarcoma viral oncoprotein

ICI immune-checkpoint inhibitor
ID identification
IFN-γ interferon gamma

LAG-3 lymphocyte activation gene 3
lncRNA long non-coding RNA
LND lymph node density
LVSI lymph-vascular space invasion

MMP matrix metalloproteinase

NCCN National Comprehensive Cancer Network
NCI the National Cancer Institute
NDFIP1 NEDD4 family interacting protein 1
NEDD4 neural precursor cell expressed, developmentally down-regulated 4

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
SURF4 surfeit 4

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

ZNF557 zinc finger protein 557

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials
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 in the GitHub repository, https://github.com/texchi2/pvalueTex.

Competing interests
The authors declare that they have no competing interests.

Funding
Not applicable.

Author's contributions
A.T.H. Wu, Y.C. Li, and M. Hsiao designed and supervised the study. L.H. Chi made the R coding and debugging, was a major contributor in writing the manuscript. Y.C. Li and M. Hsiao analyzed the results. L.H. Chi and A.T.H. Wu are involved in manuscript review and revision. All authors read and approved the final manuscript.

Acknowledgements
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://github.com/JDMCreator/LaTeXTableEditor. 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 TCGA affiliated biobanks.
The Mouse Tumor Biology Database. http://tumor.informatics.jax.org/mtbwi/index.do. Accessed 20 May 2013.

Author details

[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, Taipei Medical University Hospital, Taipei, Taiwan. [3]Genomics Research Center, Academia Sinica, Taipei, Taiwan. [4]Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan.

References

1. MOHW: 2017 Statistics of Causes of Death. Department of Statistics, Ministry of Health and Welfare, Taipei (2018). https://www.mohw.gov.tw/cp-3961-42866-2.html
2. MOHW: Statistics of General Health and Welfare 2018. Department of Statistics, Ministry of Health and Welfare (2018). https://www.mohw.gov.tw/lp-4614-2.html
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 **18**(7), 873–898 (2020). doi:10.6004/jnccn.2020.0031
4. HPA: Statistics of Health Promotion 2017. Health Promotion Administration, Ministry of Health and Welfare, Taiwan, Taipei (2019). https://www.hpa.gov.tw
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 **26**(1), 22–30 (2004). 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 **61**(7), 2911–2916 (2001)
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 **94**(3), 713–722 (2002). 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 **60**(4), 1121–1128 (2000)
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 **23**(24), 5560–5567 (2005). doi:10.1200/JCO.2005.06.411
10. Harrington, K.J.: Chemotherapy and Targeted Agents. In: Maxillofacial Surgery, pp. 339–354. Elsevier, ??? (2017). doi:10.1016/B978-0-7020-6056-4.00022-8. https://linkinghub.elsevier.com/retrieve/pii/B9780702060564000228
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 **354**(6), 567–578 (2006). 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 **359**(11), 1116–1127 (2008). 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 **9**(10), 1421–1428 (2009). 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 **31**(23), 2854–2860 (2013). 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 **20**(11), 2933–2946 (2014). 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 **104**, 1–8 (2018). 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 **32**(15_suppl), 6011 (2014)
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 **51**(1), 12–15 (2015). 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. Springer (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. Nature Publishing Group (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.M., 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.M., 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 **394**(10212), 1915–1928 (2019). 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 **84**(December 2019), 101977 (2020). doi:10.1016/j.ctrv.2020.101977
24. Chi, L.-H., Chang, W.-M., Chang, Y.-C., Chan, Y.-C., Tai, C.-C., Leung, K.-W., Chen, C.-L., Wu, A.T., Lai, T.-C., Li, Y.-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 **7**(1), 9031 (2017). 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 **7**(3), 419–423 (2008). 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 **131**(9), 2067–77 (2012). 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 **6**(41), 43869–43880 (2015). 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 **48**(9), 788–794 (2016). 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 **7**(December), 1–16 (2019). doi:10.3389/fcell.2019.00304
31. Yan, L., Zhang, W.: Precision medicine becomes reality-tumor type-agnostic therapy. Cancer communications (London, England) **38**(1), 6 (2018). doi:10.1186/s40880-018-0274-3
32. 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 **20**(21) (2019). doi:10.3390/ijms20215399
33. 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 **104**(49), 19428–19433 (2007). doi:10.1073/pnas.0709013104
34. Mizuno, H., Kitada, K., Nakai, K., Sarai, A.: PrognoScan: A new database for meta-analysis of the prognostic value of genes. BMC Medical Genomics **2**(1), 18 (2009). doi:10.1186/1755-8794-2-18
35. 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 **7**(12), 1–7 (2012). doi:10.1371/journal.pone.0051862
36. 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 **12**(4) (2017).

doi:10.1371/journal.pone.0176231
37. 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 **357**(6352), 2507 (2017). doi:10.1126/science.aan2507
38. Deng, M., Brägelmann, J., Kryukov, I., Saraiva-Agostinho, N., Perner, S.: FirebrowseR: An R client to the Broad Institute's Firehose Pipeline. Database **2017**(1) (2017). doi:10.1093/database/baw160
39. NCI Genomic Data Commons: mRNA Analysis Pipeline (2017). https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/
40. GDAC: Samples Report (2016). http://gdac.broadinstitute.org/runs/stddata__2016_01_28/samples_report/
41. NCI Genomic Data Commons: GDC Data Dictionary (2019). https://docs.gdc.cancer.gov/Data_Dictionary/
42. 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 **67**(2), 93–99 (2017). doi:10.3322/caac.21388
43. Brierley, J.D., Gospodarowicz, M.K., Wittekind, C.: TNM Classification of Malignant Tumours, 8th Edition, p. 272. Wiley-Blackwell, Hoboken (2016). https://www.wiley.com/en-us/
44. Halpern, J.: Maximally Selected Chi Square Statistics for Small Samples. Biometrics **38**(4), 1017 (1982). doi:10.2307/2529882
45. 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 **28**(4), 938–9486 (2019). doi:10.1016/j.celrep.2019.06.067
46. Andersen, P.K., Gill, R.D.: Cox's Regression Model for Counting Processes: A Large Sample Study. Annals of Statistics **10**(4), 1100–1120 (1982)
47. 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
48. 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 **34**(5), 676–688 (2010). doi:10.1097/PAS.0b013e3181d95c37
49. 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 **31**(5), 772–779 (2018). doi:10.1038/modpathol.2017.183
50. 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) **161**(2), 235–244 (2019). doi:10.1177/0194599819839006
51. 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 **152**(4), 354–360 (1986). doi:10.1016/0002-9610(86)90304-1
52. 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. Churchill Livingstone Inc. (2003). doi:10.1054/ijom.2002.0313
53. 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. http://doi.wiley.com/10.1002/hed.20036
54. 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 **39**(4), 370–7 (2010). doi:10.2310/7070.2010.090084
55. 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 **32**(11), 1444–1451 (2010). doi:10.1002/HED
56. Kuriakose, M.A., Trivedi, N.P.: Contemporary Oral Oncology vol. 2, 1st edn., pp. 147–187. Springer, Switzerland (2017). doi:10.1007/978-3-319-14917-2. https://www.springer.com/gp/book/9783319149165
57. Shapiro, M., Salama, A.: Margin Analysis: Squamous Cell Carcinoma of the Oral Cavity. Oral and Maxillofacial Surgery Clinics of North America **29**(3), 259–267 (2017). doi:10.1016/j.coms.2017.03.003
58. 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 **86**, 53–60 (2018). doi:10.1016/j.oraloncology.2018.09.005
59. 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 **24**(5), 603–609 (2019). doi:10.4317/medoral.23011
60. 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 **94**, 115–120 (2019). doi:10.1016/j.oraloncology.2019.05.020
61. Warburg, O.: On the Origin of Cancer Cells. Science **123**(3191), 309–314 (1956)
62. 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 and Oncology **10**(1), 1–12 (2017). doi:10.1186/s13045-016-0372-0
63. Liu, J., Liu, Z.-X., Wu, Q.-N., Lu, Y.-X., Wong, C.-W., Ju, H.-Q., Xu, R.-H.: Long non-coding RNA LOC148709 regulates PFKFB3-mediated glycolytic reprogramming in esophageal squamous cell carcinoma. Gut **68**(Suppl 1), 20–20 (2019). doi:10.1136/gutjnl-2019-IDDFAbstracts.37
64. 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
65. Li, J., Lu, Y., Akbani, R., Ju, Z., Roebuck, P.L., Liu, W., Yang, J.Y., Broom, B.M., Verhaak, R.G.W., Kane, D.W., Wakefield, C., Weinstein, J.N., Mills, G.B., Liang, H.: TCPA: A resource for cancer functional proteomics data. Nature Methods **10**(11), 1046–1047 (2013). doi:10.1038/nmeth.2650
66. 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 **77**(21), 51–54 (2017). doi:10.1158/0008-5472.CAN-17-0369
67. 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 **18**(8), 15–25 (2019). doi:10.1074/mcp.RA118.001260
68. Woodward, S.J.R., Beukes, P.C., Hanigan, M.D.: Molly reborn in C++ and R. animal **14**(S2), 250–256 (2020). doi:10.1017/S1751731120000270
69. Looney, A.M., Nawaz, K., Webster, R.M.: Tumour-agnostic therapies. Nature Reviews Drug Discovery **19**(6), 383–384 (2020). doi:10.1038/d41573-020-00015-1

Figures

Figure 1 A workflow of HNSCC biomarker discovery, step 1 (blue line: main procedure) and step 2 (orange line: analysis export). The "main procedure" includes data retrieving from TCGA GDC data portal, data process with merging and cleaning, then performing the survival analyses. The Cutoff engine (cutofFinder_func.HNSCC.R) might calculate all possible Kaplan-Meier P-value to find the optimal cutoff value of RNA-Seq for subsequent Cox modeling (a draft diagram shown on the upper right corner "HNSCC cohort", the serial cut for grouping patients with low [green] or high [red] expression of a specific gene, to yield a collection of P-values; please see Materials and Methods section for details). The step 2 "analysis export" performs dissecting and selection of candidate genes by Bonferroni adjusted P-value as well as a hazard ratio of Cox model, which was based on the results from the step 1. (HNSCC: head and neck squamous cell carcinoma; TCGA: the Cancer Genome Atlas; RNA-Seq: RNA sequencing; GDC: Genomic Data Commons.)

Figure 2 HNSCC Cox's hazard ratio and P-value plots. (a) Univariate HR versus uncorrected P-value; (b) Multivariate HR versus uncorrected P-value; (c) Univariate HR versus Bonferroni corrected P-value; and (d) Multivariate HR versus Bonferroni corrected P-value.

Figure 3 Volcano plot of genes under survival analyses. X axis: unadjusted P-value of Kaplan-Meier survival (-log10 transformed). Y axis: multivariate hazard ratio from Cox proportional regression model. Dotted line: significant Bonferroni corrected P-value. Red circles mark 10 candidate genes, which impact on poor prognosis ($HR >= 1.5$). Green circles mark 10 genes, which affect on better survival ($HR <= 0.5$).

Tables

Figure 4 Kaplan-Meier survival analyses, by cutoff finding. (a) Kaplan-Meier plot of DKK1 under optimal P-value, and (b) the cutoff was derived from cumulative P-value plots of DKK1. (c) Kaplan-Meier plot of ZNF557 under optimal P-value, and (d) the cutoff was derived from cumulative P-value plots of ZNF557.

Table 1 The 10 candidate genes over-expressed with poor prognosis in HNSCC (ranked by Bonferroni adjusted P-value)

Gene ID	Gene Description	Kaplan-Meier survival		Univariate		Multivariate		Remark
		P-value	Adjusted P-value	HR*	95% CI	HR*	95% CI	
DKK1	dickkopf WNT signaling pathway inhibitor 1	8.9×10^{-8}	0.001	2.266	1.666-3.082	2.135	1.559-2.924	**
CAMK2N1	calcium/calmodulin-dependent protein kinase II inhibitor 1	2.9×10^{-7}	0.002	2.101	1.572-2.809	2.007	1.490-2.704	**
STC2	stanniocalcin 2	6.5×10^{-7}	0.004	2.147	1.578-2.921	2.075	1.515-2.843	**
PGK1	phosphoglycerate kinase 1	9.1×10^{-7}	0.006	2.127	1.563-2.895	2.046	1.498-2.795	**
SURF4	surfeit 4	9.6×10^{-7}	0.006	2.055	1.531-2.757	2.089	1.543-2.829	3
USP10	ubiquitin specific peptidase 10	1.7×10^{-6}	0.012	2.083	1.532-2.834	2.119	1.551-2.895	**
NDFIP1	Nedd4 family interacting protein 1	2.6×10^{-6}	0.017	2.031	1.502-2.746	2.027	1.483-2.771	6
FOXA2	forkhead box A2	2.7×10^{-6}	0.018	1.976	1.479-2.640	1.914	1.426-2.569	**
STIP1	stress-induced-phosphoprotein 1	4.3×10^{-6}	0.029	1.958	1.463-2.621	1.957	1.451-2.640	**
DKC1	dyskeratosis congenita 1, dyskerin	6.3×10^{-6}	0.042	2.046	1.490-2.808	1.837	1.332-2.534	**
Selection criteria: Kaplan-Meier Bonferroni adjusted P-value < 0.05¬† Cox's univariate and multivariate$HR >= 1.5$								
* Cox's model: P-value < 0.001								
Remark: number of articles related with cancer research; ** as many								

Table 2 Univariate/multivariate Cox's proportional hazards regression analyses on OS time of DKK1 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	1.178	0.841-1.650	0.342
Age at diagnosis	$<= 65y$	1			1		
	$> 65y$	1.329	0.990-1.784	0.058	1.462	1.078-1.983	0.015
Clinical T Status	T1+T2	1			1		
	T3+T4	1.409	1.028-1.931	0.033	1.978	1.046-3.737	0.036
Clinical N Status	N0	1			1		
	N1-3	1.185	0.890-1.577	0.246	1.149	0.805-1.640	0.445
Clinical M Status	M0	1			1		
	M1	4.097	1.009-16.64	0.049	6.513	1.415-29.96	0.016
Clinical Stage	Stage I+II	1			1		
	Stage III+IV	1.245	0.882-1.759	0.213	0.597	0.277-1.287	0.188
Surgical Margin status	Negative	1			1		
	Positive	1.591	1.155-2.191	0.004	1.601	1.159-2.211	0.004
Tobacco Exposure	Low	1			1		
	High	1.364	1.008-1.844	0.044	1.302	0.943-1.797	0.109
RNA-Seq	Low	1			1		
	High	2.266	1.666-3.082	***	2.135	1.559-2.924	***
(P-value significant codes is denoted: $red < 0.05$; *** < 0.001)							

Table 3 The 10 candidate genes over-expressed with better prognosis in HNSCC (ranked by Bonferroni corrected P-value)

Gene ID	Gene Description	Kaplan-Meier survival		Univariate		Multivariate		Remark
		P-value	Adjusted P-value	HR*	95% CI	HR*	95% CI	
ZNF557	zinc finger protein 557	8.6×10^{-8}	0.001	0.465	0.348-0.619	0.499	0.372-0.669	0
ZNF266	zinc finger protein 266	2.2×10^{-7}	0.001	0.474	0.355-0.632	0.453	0.338-0.607	1
IL19	interleukin 19	3.7×10^{-7}	0.002	0.472	0.351-0.635	0.459	0.340-0.619	14
MYO1H	myosin 1H	3.8×10^{-7}	0.003	0.468	0.347-0.632	0.467	0.344-0.634	0
FCGBP	Fc fragment of IgG binding protein	1.2×10^{-6}	0.008	0.484	0.359-0.653	0.496	0.366-0.674	**
LOC148709	LncRNA LOC148709	1.5×10^{-6}	0.010	0.499	0.374-0.666	0.485	0.361-0.652	1
EVPLL	envoplakin-like protein	2.0×10^{-6}	0.013	0.490	0.363-0.661	0.494	0.364-0.672	0
PNMA5	paraneoplastic antigen like 5	2.6×10^{-6}	0.017	0.499	0.371-0.671	0.481	0.357-0.650	5
KIAA1683	new name as IQ Motif Containing N (IQCN)	3.1×10^{-6}	0.020	0.500	0.371-0.673	0.483	0.356-0.654	0
NPB	neuropeptide B	4.0×10^{-6}	0.027	0.460	0.328-0.646	0.457	0.324-0.646	4

Selection criteria:
Kaplan-Meier Bonferroni adjusted P-value < 0.05
Cox's univariate and multivariate HR $>= 1.5$
Cox's model: P-value < 0.001
Remark: number off articles related to cancer research; ** as many
lncRNA: Long non-coding RNA

Table 4 Univariate/multivariate Cox's proportional hazards regression analyses on OS time of ZNF557 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	1.163	0.833-1.625	0.375
Age at diagnosis	$<= 65y$	1			1		
	$> 65y$	1.329	0.990-1.784	0.058	1.328	0.976-1.808	0.071
Clinical T Status	T1+T2	1			1		
	T3+T4	1.409	1.028-1.931	0.033	1.961	1.035-3.714	0.039
Clinical N Status	N0	1			1		
	N1-3	1.185	0.890-1.577	0.246	1.179	0.824-1.686	0.367
Clinical M Status	M0	1			1		
	M1	4.097	1.009-16.64	0.049	8.478	1.847-38.92	0.006
Clinical Stage	Stage I+II	1			1		
	Stage III+IV	1.245	0.882-1.759	0.213	0.512	0.239-1.096	0.085
Surgical Margin status	Negative	1			1		
	Positive	1.591	1.155-2.191	0.004	1.631	1.180-2.254	0.003
Tobacco Exposure	Low	1			1		
	High	1.364	1.008-1.844	0.044	1.453	1.055-2.000	0.022
RNA-Seq	Low	1			1		
	High	0.465	0.348-0.619	***	0.499	0.372-0.669	***
(P-value significant codes is denoted: $red < 0.05$; *** < 0.001)							

Figures

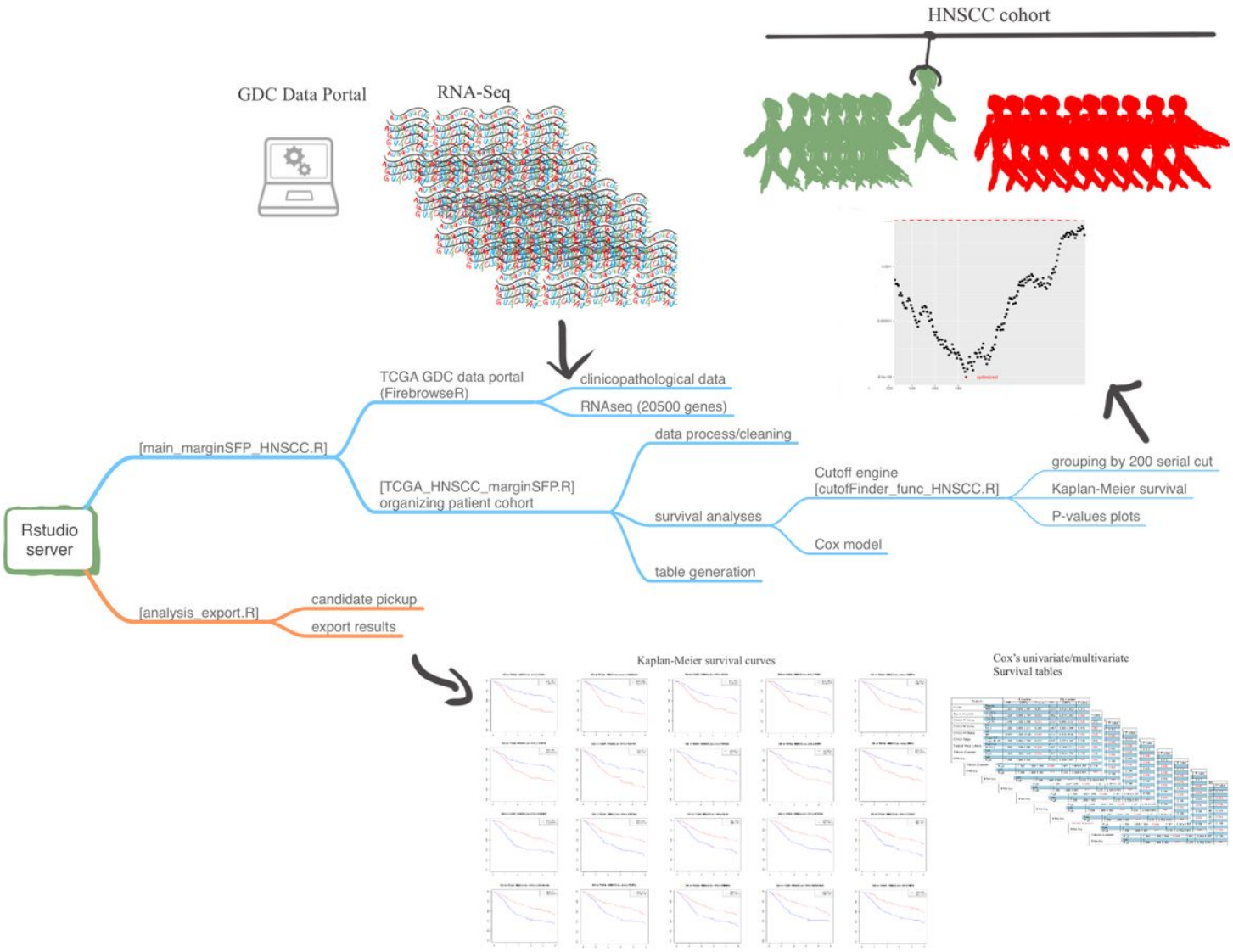

Figure 1

A workflow of HNSCC biomarker discovery, step 1 (blue line: main procedure) and step 2 (orange line: analysis export). The "main procedure" includes data retrieving from TCGA GDC data portal, data process with merging and cleaning, then performing the survival analyses. The Cutoff engine (cutofFinder func.HNSCC.R) might calculate all possible Kaplan-Meier P-value to find the optimal cutoff value of RNA-Seq for subsequent Cox modeling (a draft diagram shown on the upper right corner "HNSCC cohort", the serial cut for grouping patients with low [green] or high [red] expression of a specific gene, to yield a collection of P-values; please see Materials and Methods section for details). The step 2 "analysis export" performs dissecting and selection of candidate genes by Bonferroni adjusted P-value as well as a hazard ratio of Cox model, which was based on the results from the step 1. (HNSCC: head and neck squamous cell carcinoma; TCGA: the Cancer Genome Atlas; RNA-Seq: RNA sequencing; GDC: Genomic Data Commons.)

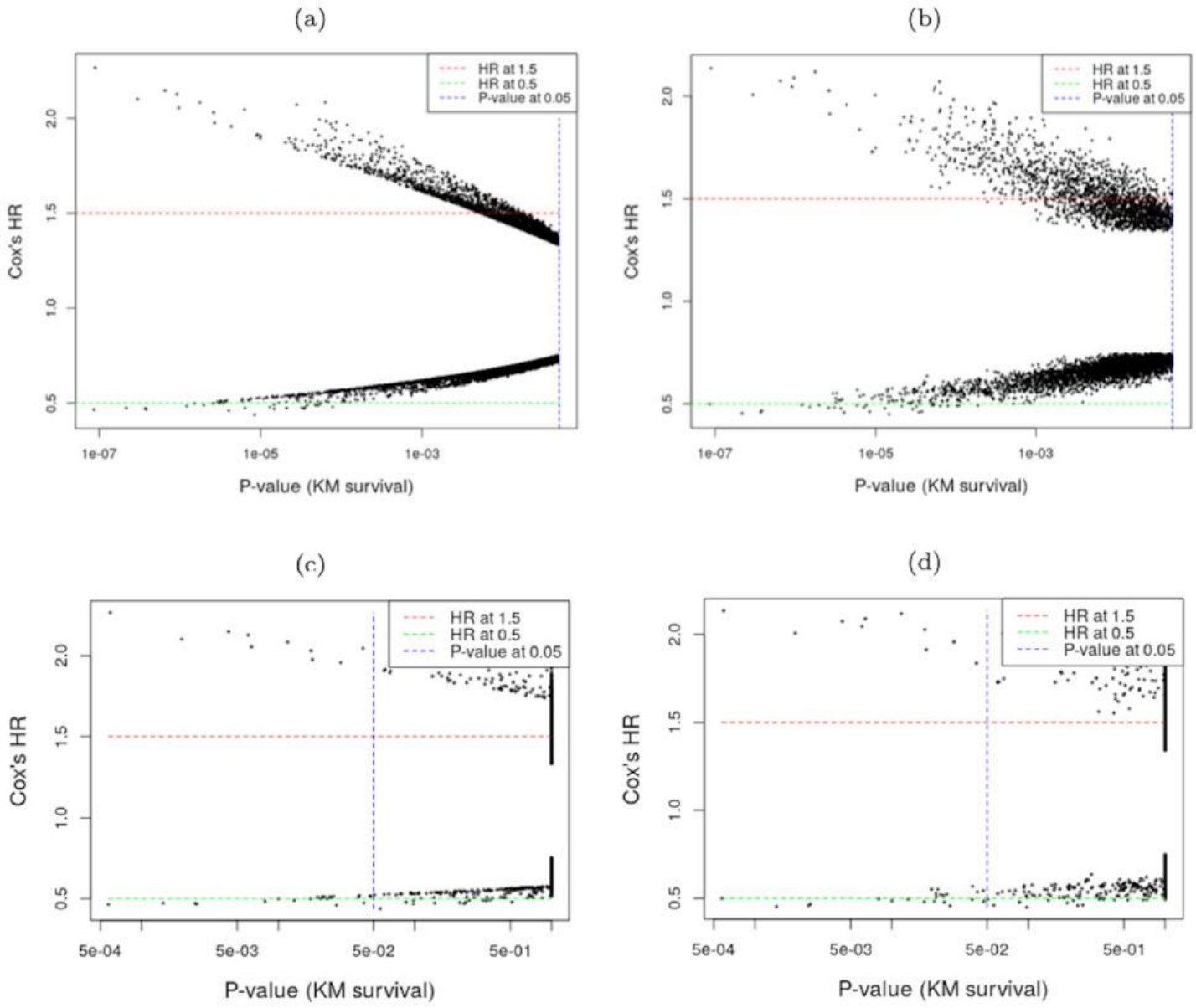

Figure 2

HNSCC Cox's hazard ratio and P-value plots. (a) Univariate HR versus uncorrected P-value; (b) Multivariate HR versus uncorrected P-value; (c) Univariate HR versus Bonferroni corrected P-value; and (d) Multivariate HR versus Bonferroni corrected P-value.

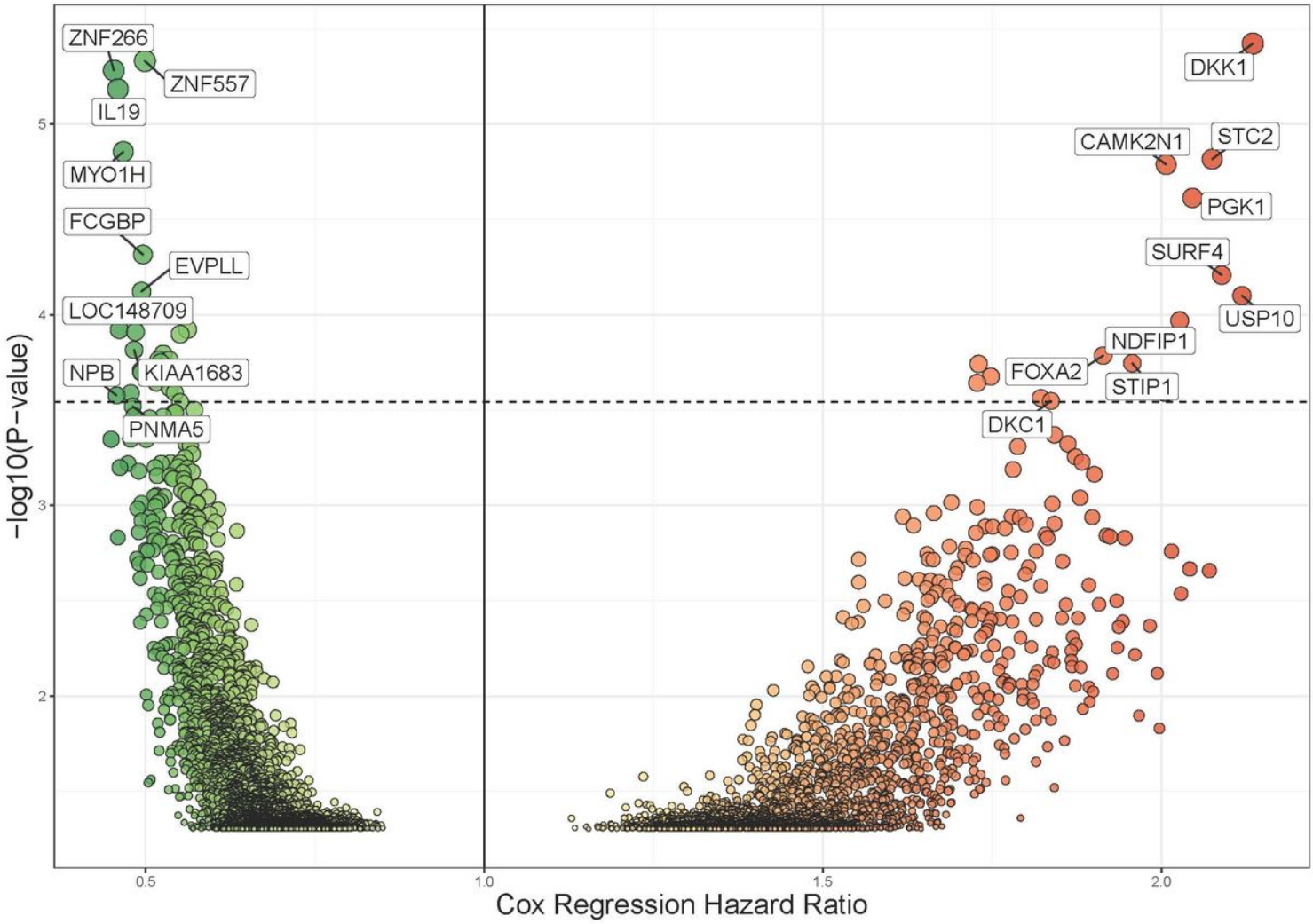

Figure 3

Volcano plot of genes under survival analyses. X axis: unadjusted P-value of Kaplan-Meier survival (-log10 transformed). Y axis: multivariate hazard ratio from Cox proportional regression model. Dotted line: significant Bonferroni corrected P-value. Red circles mark 10 candidate genes, which impact on poor prognosis (HR >= 1:5). Green circles mark 10 genes, which affect on better survival (HR <= 0:5).

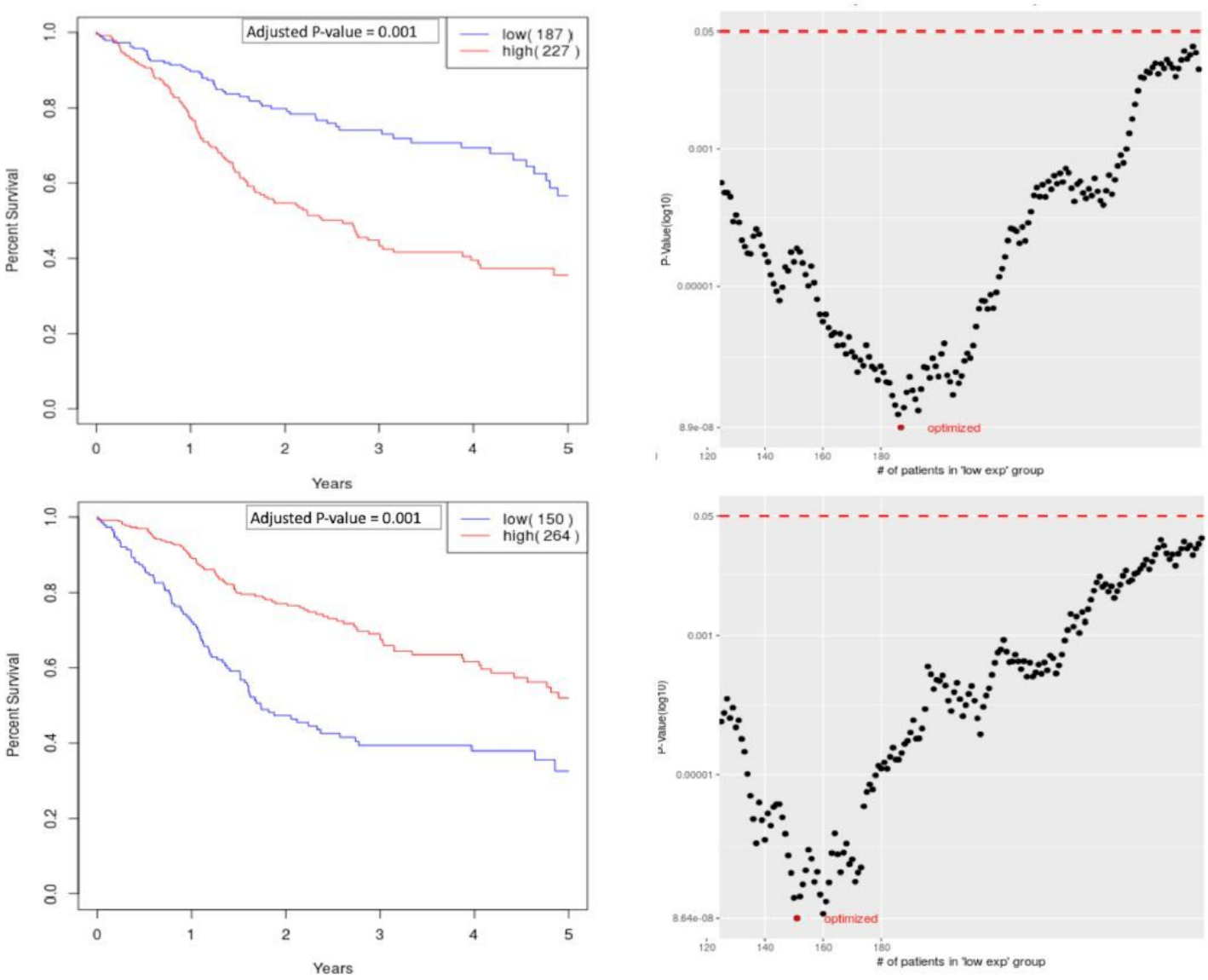

Figure 4

Kaplan-Meier survival analyses, by cutoff finding. (a) Kaplan-Meier plot of DKK1 under optimal P-value, and (b) the cutoff was derived from cumulative P-value plots of DKK1. (c) Kaplan-Meier plot of ZNF557 under optimal P-value, and (d) the cutoff was derived from cumulative P-value plots of ZNF557.