Available online at www.sciencedirect.com

ScienceDirect

journal homepage: http://www.elsevier.com/locate/rpor

Original research article

Characteristic miRNA expression signature and random forest survival analysis identify potential cancer-driving miRNAs in a broad range of head and neck squamous cell carcinoma subtypes

Yury O. Nunez Lopez[a,*], *Berta Victoria*[b], *Pawel Golusinski*[c], *Wojciech Golusinski*[c], *Michal M. Masternak*[b,c]

[a] *Translational Research Institute for Metabolism & Diabetes, Florida Hospital, 301 East Princeton St., Orlando, FL 32804, USA*

[b] *Burnett School of Biomedical Sciences, College of Medicine, University of Central Florida, 6900 Lake Nona Blvd., Orlando, FL 32827, USA*

[c] *Department of Head and Neck Surgery, The Greater Poland Cancer Centre, 15 Garbary St., 61-866 Poznan, Poland*

ARTICLE INFO

Article history:
Received 21 March 2017
Received in revised form
27 August 2017
Accepted 22 October 2017
Available online 20 November 2017

Keywords:
Cancer
HNSCC
MicroRNA
Random forest
Survival
Classifier

ABSTRACT

Aim: To characterize the miRNA expression profile in head and neck squamous cell carcinoma (HNSSC) accounting for a broad range of cancer subtypes and consequently identify an optimal miRNA signature with prognostic value.

Background: HNSCC is consistently among the most common cancers worldwide. Its mortality rate is about 50% because of the characteristic aggressive behavior of these cancers and the prevalent late diagnosis. The heterogeneity of the disease has hampered the development of robust prognostic tools with broad clinical utility.

Materials and methods: The Cancer Genome Atlas HNSC dataset was used to analyze level 3 miRNA-Seq data from 497 HNSCC patients. Differential expression (DE) analysis was implemented using the *limma* package and multivariate linear model that adjusted for the confounding effects of age at diagnosis, gender, race, alcohol history, anatomic neoplasm subdivision, pathologic stage, T and N stages, and vital status. Random forest (RF) for survival analysis was implemented using the *randomForestSRC* package.

Results: A characteristic DE miRNA signature of HNSCC, comprised of 11 upregulated (i.e., miR-196b-5p, miR-1269a, miR-196a-5p, miR-4652-3p, miR-210-3p, miR-1293, miR-615-3p, miR-503-5p, miR-455-3p, miR-205-5p, and miR-21-5p) and 9 downregulated (miR-376c-3p, miR-378c, miR-29c-3p, miR-101-3p, miR-195-5p, miR-299-5p, miR-139-5p, miR-6510-3p, miR-375) miRNAs was identified. An optimal RF survival model was built from seven variables including age at diagnosis, miR-378c, miR-6510-3p, stage N, pathologic stage, gender, and race (listed in order of variable importance).

* *Corresponding author.*
E-mail address: Yury.Nunez-Lopez@flhosp.org (Y.O. Nunez Lopez).

https://doi.org/10.1016/j.rpor.2017.10.003

Conclusions: The joint differential miRNA expression and survival analysis controlling for multiple confounding covariates implemented in this study allowed for the identification of a previously undetected prognostic miRNA signature characteristic of a broad range of HNSCC.

1. Background

Head and neck squamous cell carcinomas (HNSCC) are malignancies of epithelial origin that are consistently among the most common cancers worldwide, with one of the highest mortality rates. The peak incidence occurs at the ages of 50–70 years, however, the proportion of elderly patients with HNSCC is increasing.[1] Worldwide, an estimated 650,000 new cases occur per year[2] and approximately half of all patients with the disease die after exhausting multiple treatment efforts including surgery, radiation, and chemotherapy.[3,4] Among the main reasons for this high mortality are the characteristic aggressive behavior of these cancers and the prevalent late diagnosis of the disease.[5] Regarding the disease etiology, it is currently known that human papillomavirus (HPV) infection and smoking are two important factors implicated (possibly synergistically) in the development of HNSCC.[6] However, because of the heterogeneity of this type of malignancy, there is currently a lack of specific molecular biomarkers with relevant clinical utility.

Recent initiatives implementing large scale genomics, such as The Cancer Genome Atlas (TCGA), have generated important insights for classification of HNSCC and identification of potential prognostic biomarkers including microRNAs.[3,5,7–9] MicroRNAs (miRNAs) are a type of small noncoding RNA that play major roles in maintaining cellular and tissue homeostasis often via post-transcriptional and translational regulatory processes. In addition, accumulating evidence suggest that miRNAs function in cell-to-cell communication both in normal physiology and pathological processes.[10–12] Other studies on HNSCC suggest that miRNAs present in altered concentrations in the patients' circulation may be useful biomarkers of cancer development and tumor stage.[13,14] However, most of these studies have circumvented the need for adjusting for confounding effects by limiting their analyses to subsets of cancer subtypes based, for example, on the original tumor location or alternative characteristics of interest such as human papillomavirus infection, among others. Such approaches have produced valuable information on a variety of subtype comparisons, yet our knowledge of a characteristic miRNA signature covering the broad range of HNSCC subtypes is lacking. Such a characteristic signature would identify miRNAs that potentially represent drivers of this type of cancer, provided that a growth advantage for the tumor cells and positive selection of the particular miRNA alteration in the tumor microenvironment can be additionally demonstrated.[15] We reason that miRNAs found to be both prognostic of survival and robustly differentially expressed between normal and tumor samples from a broad range of anatomical subdivisions and pathological stages, after accounting for potential confounding effects of additional covariates, are suggested to play important roles in disease development and progression. Therefore, characteristic miRNAs that track with survival could represent important potential targets for novel therapeutic strategies.

2. Aim

The aim of this study was to identify a characteristic and prognostic miRNA signature of HNSCC by implementing a comprehensive differential miRNA expression analysis (DEA) that accounted for cancer-specific and dataset-specific confounders and by modeling survival using an iteratively-evolved semi-exhaustive Random Forest approach, using the complete TCGA-HNSC miRNA-Seq dataset.

3. Materials and methods

3.1. *Study design*

Level-3 miRNA-Seq and clinical data from 524 HNSCC cases from the TCGA project was downloaded (on December 7, 2016) using the *TCGAbiolinks* R package. Samples missing tumor stage T or stage N information (extracted from the variable containing the TNM category information) were removed. In total, tumor samples from 497 patients, 44 of which had normal adjacent tissue samples/data were kept for the final analyses. Differential expression analysis comparing tumor and normal tissue miRNA expression and survival analysis using a machine learning approach were implemented as described below.

3.2. *Differential expression analysis*

DEA of miRNA profiles was implemented using level 3 miRNA-Seq data from the TCGA HNSC Project. Confounding effects of several known covariates for which data was existent (i.e., age at diagnosis, gender, race, alcohol history, anatomic neoplasm subdivision, pathologic stage, T and N stages, and vital status) were accounted for using multivariate linear modeling. Only miRNAs with expression levels greater than 5 count per million reads (CPM) in more than 50% of the samples were included in the analysis. Filtered CPM data was subjected to voom normalization[16] using the *limma* R package[17] followed by multivariate linear modeling of the expression of each detected miRNA as a function of the sample group and the above-mentioned covariates. A fold change (FC) filter was implemented to remove miRNAs showing relatively small changes in expression levels ($FC < 2$) from further analyses. Expression differences were considered

significant at $P \leq 0.01$. P values were corrected for multiple testing using the Benjamini-Hochberg correction.[18]

3.3. *Iterative semi-exhaustive random forest (iseRF) for survival analysis*

To identify which miRNAs from the characteristic DE miRNA signature may have prognostic value, we implemented a machine learning strategy that uses the random forest (RF) approach for survival analysis[19] in an iteratively evolving fashion. The RF approach is based on the construction of many decision (classification) trees that are used to classify a given input data vector. To build each decision tree, the RF algorithm uses different bootstrapped and randomly split samples of the original dataset. Approximately 33% of the samples are left out of the bootstrapped training set and used as a corresponding cross-validation set (test set). The ensemble of bootstrapped test sets (different for every tree construction) is called the out-of-bag (OOB) sample set and is used for an unbiased estimation of the prediction error of the model generated in the corresponding ensemble of training sets ("in-bag" sample set). A variable importance (VIMP) measure can be calculated based on the impact that each variable effect on the OOB prediction error as the respective variable values are randomly permuted.[20–22]

Our specific strategy was comprised of three main stages (described below). At each tree construction step of the first 2 stages, the number of trees in the forest (ntree value) was set to 5000, whereas in the last (longer) stage ntree was set to 2000. The mtry number of variables to be randomly sampled at each split was set equal to the default mtry = $\sqrt{\text{total number of variables}}$ in all stages.

Step 1. Iterative backward elimination step: Using the *randomForestSRC* package in the R.3.3.2 environment, an RF survival model was built on the complete study cohort (primary tumor data) starting with the normalized logCPM values for all 20 characteristic and differentially expressed (DE) miRNAs and the values for 31 clinical variables available for the TCGA-HNSC dataset (accounting for a total of 51 variables). An iterative backward elimination process was implemented[23] by successively eliminating the less informative (lower VIMP) variable, one at a time, after each of 51 iterations of the RF modeling approach until a single variable remain. The OOB prediction error rates calculated for each RF iteration are compared and the instance of the RF classifier generating the lower error rate selected as the optimal RF model from this step.

Step 2. Iterative forward incremental step: Starting with the variables selected in step 2, each single variable was independently used to generate single variable RF models. The single variable generating the RF instance with the lowest OOB prediction error rate was then selected and tested in combination with every other single non-selected variable. The pair generating a new instance of the RF model with the lowest OOB error rate for the specific iteration was then selected and tested in combination with every single non-selected variable. This iterative incremental process is repeated until a last model testing the combination of all variables from step 2 are tested. The combination of variables contributing the RF model with the lowest OOB prediction error rate is then selected as the optimal combination of variables from this step and used in step 3 to identify the final optimal model.

Step 3. Exhaustive combinatorial step: Starting with the best combination of variables selected in step 2, an exhaustive search will build RF models with every possible combination of those variables. The combination of variables producing the RF model with the lower OOB prediction error rate will then be selected as the optimal iseRF model.

3.4. *Advanced Gene Set Enrichment Analysis (GSEA) of lists of validated miRNA-target genes*

Validated target genes of specific differentially expressed miRNAs were identified using the multiMiR package.[24] Functional enrichment analysis for the list of validated miRNA-targeted genes was then performed using the SetRank package,[25] a novel advanced GSEA algorithm. Enriched categories are defined by the SetRank score (a value that reflects the prominence of a gene set in a gene set network and is calculated using the PageRank algorithm).[26] SetRank setPCutoff was set to 0.05 and fdrCutoff to 0.2.

4. Results

4.1. *Characteristics of study cohort and dataset*

Demographic and clinical characteristics of the study cohort are presented in Table 1 . The age of the cancer patients ranged from 19 years of age (yoa) for the youngest to 90 for the oldest, with an average of 61 (95% CI: 53 to 69) yoa. The majority of patients were male ($N = 360$ out of 497, 72%), white ($N = 427$, 86%), not of hispanic origins (438, 88%), and with documented alcohol drinking ($N = 330$, 66%). The most common primary tumors in this cohort (accounting for 74% of all) were located in the tongue ($N = 126$, 25%), the larynx ($N = 112$, 23%), the oral cavity ($N = 70$, 14%), and the floor of the mouth ($N = 59$, 12%). Over a quarter of the samples contributing data to this study were sourced from MD Anderson Cancer Center in Texas, USA [tissue source site (TSS) code: CV, $N = 134$, 27%]. Other two main sourcing sites were University of Pittsburgh in Pennsylvania, USA (TSS code: CN, $N = 72$, 14%) and Vanderbilt University in Tennessee, USA (TSS code: CR, $N = 53$, 11%).

4.2. *Profiling of miRNA expression identify a characteristic HNSCC signature*

DEA using mirnome-wide data (level 3 miRNA-Seq data) collected from the TCGA HNSC project was used to identify a characteristic miRNA signature of HNSCC. For this, we implemented the limma-voom normalization approach on the count-per-million reads (CPM) data followed by multivariate linear modeling of the expression of each detected miRNA as a function of the sample group (Tumor or Normal) and select covariates (see Section 3.2 for details). Out of 1870 miRNAs detected by miRNA-Seq in the TCGA HNSC dataset, 334 were detected with greater than 5 CPM in more than 50% of the samples. Of these filtered subset, 20 miRNAs were found to be differentially expressed between the tumor and normal tissue with a fold change greater than 2, $P < 0.01$, and FDR ≤ 0.17

Table 1 – Clinical characteristics of the study cohort.

	Normal (N = 44)	Tumor (N = 497)	P value
Vital status: Dead	70% (31)	33% (165)	**<0.001**
Age at diagnosis	64 (54 to 68)	61 (53 to 69)	0.377
Gender: Male	68% (30)	72% (360)	0.547
Race: Black or African American (AA)	5% (2)	9% (45)	0.866
American Indian or Alaskan Native (AI)	0% (0)	0% (2)	
Asian (As)	2% (1)	2% (10)	
White (W)	91% (40)	86% (427)	
Not Reported (NR)	2% (1)	3% (13)	
Ethnicity: Hispanic or Latino (HLA)	7% (3)	5% (25)	0.876
Not Hispanic or Latino (NHLA)	86% (38)	88% (438)	
Not Reported (NR)	7% (3)	7% (34)	
Alcohol history documented: NO	43% (19)	32% (159)	**0.097**
YES	52% (23)	66% (330)	
Not Reported (NR)	5% (2)	2% (8)	
Tobacco smoking history: 1	26% (11)	24% (118)	0.656
2	26% (11)	34% (165)	
3	21% (9)	14% (67)	
4	28% (12)	27% (132)	
5	0% (0)	0% (2)	
Anatomic Neoplasm Subdivision: Alveolar_Ridge	0% (0)	4% (18)	**0.057**
Base_of_tongue	5% (2)	5% (23)	
Buccal_Mucosa	0% (0)	4% (20)	
Floor_of_mouth	7% (3)	12% (59)	
Hard_Palate	0% (0)	1% (6)	
Hypopharynx	0% (0)	2% (9)	
Larynx	27% (12)	23% (112)	
Lip	0% (0)	1% (3)	
Oral_Cavity	32% (14)	14% (70)	
Oral_Tongue	30% (13)	25% (126)	
Oropharynx	0% (0)	2% (9)	
Tonsil	0% (0)	8% (42)	
Stage T: 0	100% (44)	0% (0)	**<0.001**
1	0% (0)	7% (35)	
2	0% (0)	29% (144)	
3	0% (0)	28% (138)	
4	0% (0)	36% (180)	
Stage N: 0	100% (44)	48% (241)	**<0.001**
1	0% (0)	17% (84)	
2	0% (0)	33% (164)	
3	0% (0)	2% (8)	
Stage M	0% (0)	1% (6)	0.459
Stage event pathologic stage: NR (Not Reported)	0% (0)	15% (73)	N/A
Stage_I	0% (0)	5% (24)	
Stage_II	0% (0)	14% (69)	
Stage_III	0% (0)	14% (72)	
Stage_IVA	0% (0)	49% (246)	
Stage_IVB	0% (0)	2% (12)	
Stage_IVC	0% (0)	0% (1)	
Tissue source site: 4P	0% (0)	0% (1)	**<0.001**
BA	0% (0)	7% (34)	
BB	0% (0)	4% (19)	
C9	0% (0)	0% (2)	
CN	0% (0)	14% (72)	
CQ	0% (0)	7% (34)	
CR	0% (0)	11% (53)	
CV	86% (38)	27% (134)	
CX	0% (0)	1% (4)	
D6	0% (0)	4% (20)	
DQ	0% (0)	3% (14)	
F7	0% (0)	1% (6)	
H7	5% (2)	1% (6)	
HD	7% (3)	2% (12)	
HL	0% (0)	0% (1)	
IQ	0% (0)	2% (12)	
KU	0% (0)	1% (4)	
MT	0% (0)	1% (6)	
MZ	0% (0)	1% (3)	
P3	0% (0)	3% (14)	
QK	0% (0)	3% (17)	
RS	0% (0)	0% (2)	
T2	0% (0)	1% (4)	
T3	0% (0)	0% (2)	
TN	0% (0)	0% (1)	
UF	0% (0)	4% (18)	
UP	0% (0)	0% (1)	
WA	2% (1)	0% (1)	

Bold fonts highlight P values lower than 0.1, which were considered indicative of baseline imbalance between the normal and tumor samples.

(Fig. 1A). Considering the direction of the change, 9 miRNAs were downregulated and 11 miRNAs were upregulated.

4.3. *Random forest survival analysis identifies miR-378c and miR-6510-3p among the most important variables comprising a prognostic survival model*

To efficiently identify characteristic DE miRNAs with prognostic value, we implemented an iteratively-evolving and semi-exhaustive random forest-based approach for survival analysis (herein dubbed iseRF) using consecutive cycles of backward elimination,[19] forward incremental modeling, and exhaustive combinatorial modeling on a multimodal combination of demographic, clinical, and molecular variables. Seven variables including, in order of importance, age at diagnosis, the tumor tissue levels of characteristic DE miR-378c and miR-6510-3p, stage N, pathologic stage, gender and race comprised the iseRF model producing the lowest OOB error rate (OOB = 0.28, Fig. 1B–D). The corresponding Harrell's concordance C index was equal to 0.72 and the Somers' Dxy rank correlation for the censored response variable equal to 0.43 ($P < 0.0001$, as empirically determined after 10,000 iterations of a permutation test), as calculated using the *rcorr.cens* function from the *Hmisc* package, Fig. 1E. This instance of the iseRF model is denominated hereafter as the "optimal" iseRF model. Of note, the first five variables are shown to be very informative (with the two characteristic DE miRNAs highlighted as the second and third most informative variables, only surpassed by the age at diagnosis), whereas gender and race appear less informative.

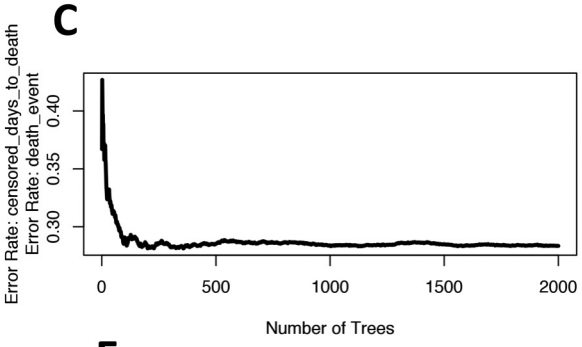
A

	Average Expression (CPM)	Average logCPM (as in plots)	Fold Change (FC)	logFC	*P* Value	FDR
hsa-miR-196b-5p	99.8	6.6	11.3	3.5	3.54E-13	1.18E-10
hsa-miR-1269a	8.3	3.0	9.0	3.2	3.34E-03	0.08
hsa-miR-196a-5p	31.8	5.0	8.3	3.0	1.93E-05	7.16E-04
hsa-miR-4652-3p	2.0	1.0	6.6	2.7	1.88E-06	1.05E-04
hsa-miR-210-3p	751.3	9.6	5.9	2.6	6.47E-06	3.09E-04
hsa-miR-1293	7.6	2.9	4.8	2.3	4.42E-03	0.10
hsa-miR-615-3p	1.5	0.6	4.8	2.3	3.07E-05	1.03E-03
hsa-miR-503-5p	6.7	2.7	4.6	2.2	1.58E-07	2.63E-05
hsa-miR-455-3p	346.5	8.4	4.5	2.2	1.26E-06	8.39E-05
hsa-miR-205-5p	12022.8	13.6	3.0	1.6	8.68E-03	0.15
hsa-miR-21-5p	255876.1	18.0	3.0	1.6	1.76E-05	7.16E-04
hsa-miR-376c-3p	5.5	2.5	-2.9	-1.5	0.010	0.17
hsa-miR-378c	8.4	3.1	-3.1	-1.7	7.07E-03	0.13
hsa-miR-29c-3p	1465.2	10.5	-3.2	-1.7	7.25E-03	0.13
hsa-miR-101-3p	36.6	5.2	-3.3	-1.7	1.80E-04	5.47E-03
hsa-miR-195-5p	23.5	4.6	-3.3	-1.7	4.07E-04	0.01
hsa-miR-299-5p	5.4	2.4	-3.6	-1.8	8.09E-04	0.02
hsa-miR-139-5p	50.1	5.6	-4.0	-2.0	4.41E-07	3.68E-05
hsa-miR-6510-3p	6.5	2.7	-4.6	-2.2	5.41E-03	0.11
hsa-miR-375	389.5	8.6	-13.4	-3.7	2.55E-07	2.84E-05

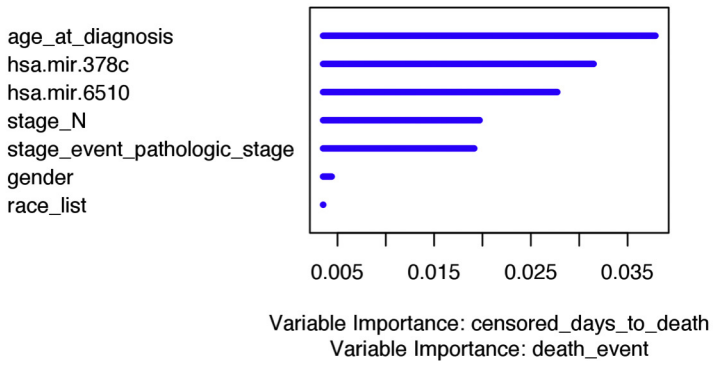
B

```
               Sample size: 497
        Number of deaths: 165
        Was data imputed: yes
         Number of trees: 2000
Minimum terminal node size: 3
Average no. of terminal nodes: 133.8
No. of variables tried at each split: 3
        Total no. of variables: 7
                Analysis: RSF
                  Family: surv
        Splitting rule: logrank *random*
Number of random split points: 10
              Error rate: 28.34%
```

C



D



E

Harrell's Concordance C Index:	0.72
Somer's Dxy Rank Corr.:	0.43
Corr. St. Dev.:	0.04
Non-missing n:	497
Missing n:	0
Uncensored n:	165
Relevant Pairs:	78220
Concordant Pairs:	56054
Randomization *P* value:	<0.0001

Fig. 1 – Characteristic and prognostic characteristic HNSCC miRNA signature identified by differential expression (DE) analysis and iterative semi-exhaustive random forest (iseRF) survival analysis. (A) Characteristic DE miRNA signature: miRNAs are sorted by logFC value. Red color indicates upregulation and blue indicates downregulation; green bars represent the average logCPM values for each miRNA in the complete dataset, including tumor and control data. (B) Summary report

4.4. *The optimal iseRF model performs well by a variety of external validation measures*

Performance measures based on external validation (OOB-based) measures are presented in Fig. 2. In the OOB mortality plot shown in Fig. 2A we can observed that the optimal iseRF model performed suitably by assigning higher mortality rates to patients that die (blue dots) earlier. Note the higher density of blue dots at higher mortality probabilities and earlier observed times and the negatively sloped fitting curve (discontinuous blue line), which contrast with the flat fitting line for the group with censored observations (black dots). Similarly, the survival probabilities predicted by the optimal iseRF model on the OOB validation set (presented for every patient in Fig. 2B and for each of the risk quartile subsets in Fig. 2E), show that the patients experiencing a death event (blue lines in Fig. 2B) were predicted to have, in general, a lower survival probability than those not experiencing death (black lines) while on trial or follow up. The related plot in Fig. 2E shows that the four distinct mortality risk quartiles have significantly different survival profiles over time, as indicated by the stratified Kaplan–Meier estimates and corresponding log-rank test P value ≤ 0.0001. In addition, the plot of the time-dependent OOB Brier score (shown in Fig. 2C, the s score measures the average discrepancies between the true vital status and the estimated predictive values over time – the higher the score, the lower the model accuracy, with a score of 0.25 equivalent to random guess) demonstrates that the iseRF model perform particularly well for predicting survival of the low risk quartile sample set (0–25% line) in the first 5 years, for predicting survival of the 25–50% risk quartile in the first 3 years, and for predicting survival of the two higher risk quartiles (50–75% and 75–100%) in first 2 years after diagnosis. Survival predictions for the highest risk (75–100%) quartile are also very accurate after 3 years from diagnosis. Furthermore, the time-dependent area under the receiver operating characteristic curves (AUC) in Fig. 2D, demonstrated good overall performance of the optimal iseRF model to predict survival in the cross-validation OOB ensemble set during the first 5 years (0–1824 days, AUCs > 0.5). The overall performance was particularly sensitive and specific during the first two years after diagnosis, with AUCs ranging from over 0.60 to approximately 0.80.

4.5. *Partial dependence plots underscore the impact of age at diagnosis, miR-378c, and miR-6510-3p on overall mortality and time-dependent survival probabilities*

To assess the predicted risk-adjusted responses we evaluated the partial dependence plots for each of the variables comprising the optimal iseRF survival model. These plots are generated by "integrating out" the effects of variables beside the covariate of interest.[27] The partial dependence of the standardize mortality plots are shown in Fig. 3, and the partial dependence of survival at 6-month, 1-year, and 5-year are shown in Supplementary Figs. SF1, SF2, and SF3. On one hand, note the strong non-linear contribution of the select variables. On the other, note how the first three most important variables, namely age at diagnosis, miR-378c, and miR-6510-3p, display the largest risk-adjusted impact on overall mortality and survival over a time range, as compared to the other less informative variables. By comparing the risk adjusted estimates of mortality shown in Fig. 3 with the corresponding non-adjusted plots shown in Fig. 4, we can observe that the effect of stage N, pathologic stage, gender, and race on mortality become flatten when the risk-adjusted partial dependence estimates are considered. Furthermore, by comparing the shape of the partial dependence survival curves at the three different times (6-month, 1-year, and 5-year, Supplementary Figs. SF1, SF2, and SF3), we can observe that the hazard responses are generally not proportional as the time progresses.

4.6. *miR-378c and miR-6510-3p levels correlate with the cancer T stage*

To gain additional insight into the pathophysiological roles of miR-378c and miR-6510-3p downregulation, we conducted correlation analysis between the miRNA expression levels and available clinicopathological measures. As shown in Fig. 5A and B, both miRNAs demonstrated significant (Spearman's $P \ll 0.001$) association with the cancer T stage, one of the cancer descriptors comprising the TNM classification system of malignant tumors.[28,29] The T stage describes the size of the primary tumor and the presence (or absence) of invasion into surrounding tissue. Note that the mean expression of these miRNAs is the highest in the normal tissue (assigned to a "dummy" stage 0 for the purpose of the correlation analysis to indicate the non-cancerous state of the tissue) and progressively decreased as the cancer T stage increased. The effect was stronger for miR-378c ($r = 0.31$, $P = 7.5 \times 10^{-4}$). Importantly, miR-378c and miR-6510-3p displayed a significant co-expression pattern between themselves (Pearson's $r = 0.34$, $P = 5.9 \times 10^{-16}$, Fig. 5C). However, differences in specific correlations with other characteristic DE miRNAs (e.g., miR-205, miR-299, and miR-376c, among others) were also detected (Fig. 5D and Supplementary Figs. SF4 and SF5).

4.7. *Novel HNSCC-related miR-6510-3p and miR-4652-3p target cancer-relevant pathways*

To predict functional roles of novel HNSCC-related miR-6510-3p and miR-4652-3p, we identified validated target genes of these two miRNAs using the multiMiR package (which represents a comprehensive compilation of predicted and validated miRNA-target interactions from 14 different databases) and conducted GSEA using the R package SetRank. SetRank implements an advanced GSEA algorithm that eliminate many false positive hits, a common problem to functional enrichment tools based on gene sets

Fig. 1 – *(continued)* for the iseRF survival model. (C) Plot of the out of bag (OOB) prediction error rate for each tree constructed in the iseRF model. (D) Plot of variable importance in the iseRF model (blue color is irrelevant). (E) Summary report of Harrell's concordance and Somer's rank correlation.

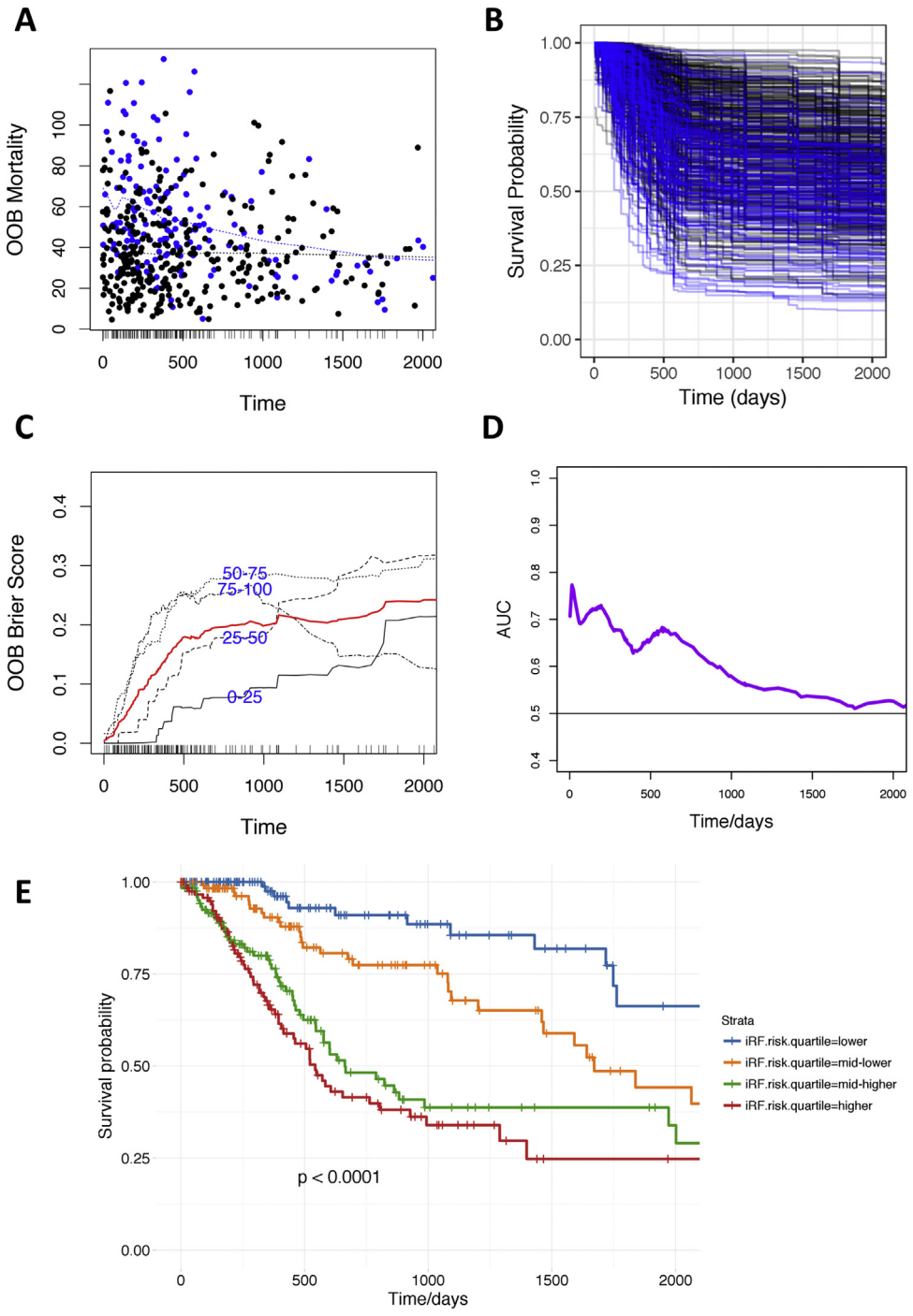


Fig. 2 – Performance measures for the optimal iseRF model based on external (OOB-based) validation measures. (A) Plot of OOB mortality of each individual versus the observed event time. Points in blue correspond to death events, black points are censored observations. (B) Plot of survival probabilities estimated for each patient in the OOB ensemble (color code as in A). (C) OOB time-dependent Brier Score (0 = perfect, 1 = poor, and 0.25 = guessing). The score is shown stratified by ensemble mortality into four groups corresponding to the 0–25, 25–50, 50–75 and 75–100 percentile values of mortality. The red line is the overall (non-stratified) time-dependent Brier score. (D) Plot of OOB time-dependent area under the ROC curve (AUC). (E) Plot of Kaplan–Meier estimates for HNSCC patients stratified by risk quartiles as predicted by the optimal iseRF model in the OOB ensemble.

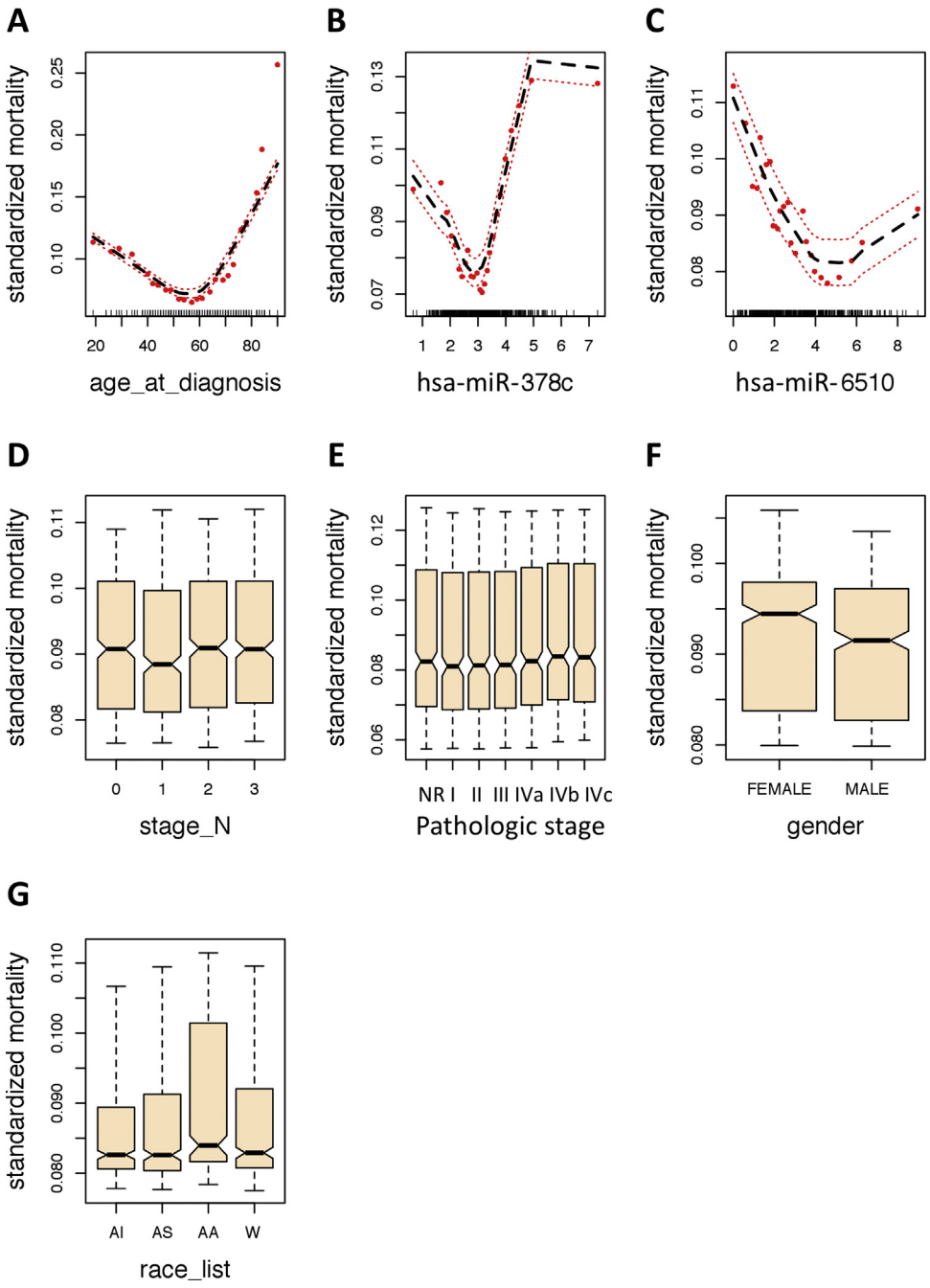


Fig. 3 – Partial dependence OOB mortality curves for each variable comprising the optimal iseRF model. These plots are the risk-adjusted alternative to the variable dependence plots shown in Fig. 4. The non-linearity of the effect of each variable is evidenced, most prominently in the three most important variables (A–C). Red dots are the partial values for the predicted response and the discontinued lines represent the trend line and the 95% confident interval. *Abbreviations*: AI: American Indian or Alaskan Native, AS: Asian, AA: Black or African American, W: white race, NR: not reported.

compiled from gene and pathway annotation databases. multiMiR identified 35 and 93 validated target genes for miR-6510-3p and miR-4652-3p, respectively (Supplementary Tables ST1–ST2). The GSEA identified significantly enriched ontologies and pathways relevant to cancer development and anti-cancer drug resistance (Table 2). Relevant gene memberships are presented in Table 2 and Supplementary Table ST3.

5. Discussion

Differential expression analysis using mirnome-wide data (miRNA-Seq) collected from TCGA (HNSC Project, level 3 data) and random forest survival analysis were used to identify a characteristic and prognostic miRNA signature in HNSCC that may contribute to the development and progression of

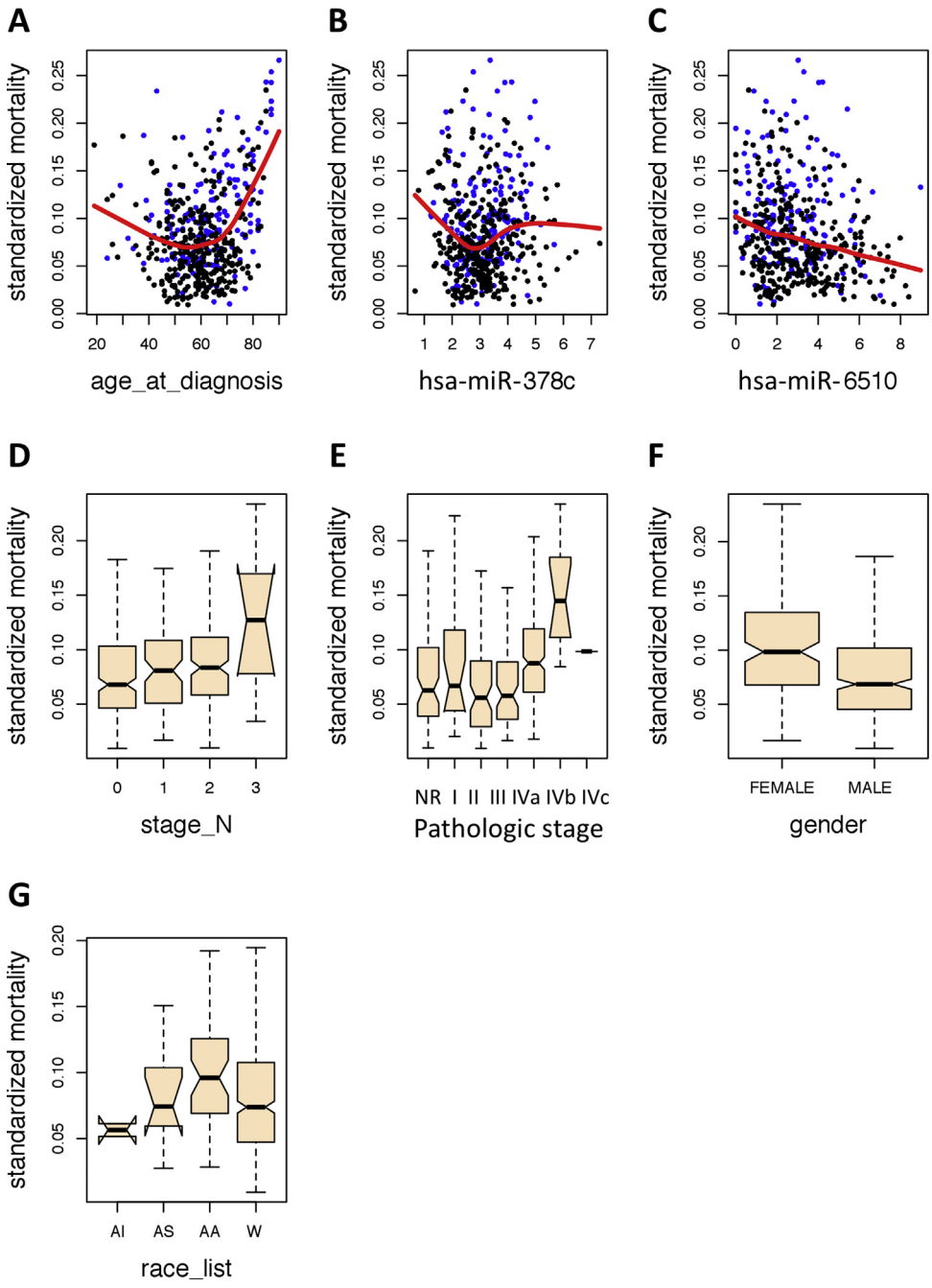


Fig. 4 – Variable dependence OOB mortality curves for each variable comprising the optimal iseRF model. Each point in the plot represent the predicted iseRF response relative to the OOB set observation for the specific covariate of interest (variables comprising the optimal iseRF model). The predicted response points are dependent on the full combination of covariates, not only on the covariate displayed in the dependence plot. The red line indicates the smooth loess trend of the prediction over the change in the specific variable. Censored cases are shown as solid black circles and death event as solid red circles. Partial dependence plots shown in Fig. 3 display more specific, risk-adjusted partial effects for each covariate of interest. *Abbreviations*: AI: American Indian or Alaskan Native, AS: Asian, AA: Black or African American, W: white race, NR: not reported.

the disease. We reasoned that the most biologically relevant miRNAs specifically involved with HNSCC development and progression are likely those with a common altered expression profile along the different cancer subtypes and tumor stages, which in addition associate with a prospective measure of survival. Such a signature could provide insight into the specific cancer-driving miRNA biology.

The first part of our strategy included the identification of a characteristic (common) differentially expressed miRNA signature relevant to a broad range of HNSCC, as

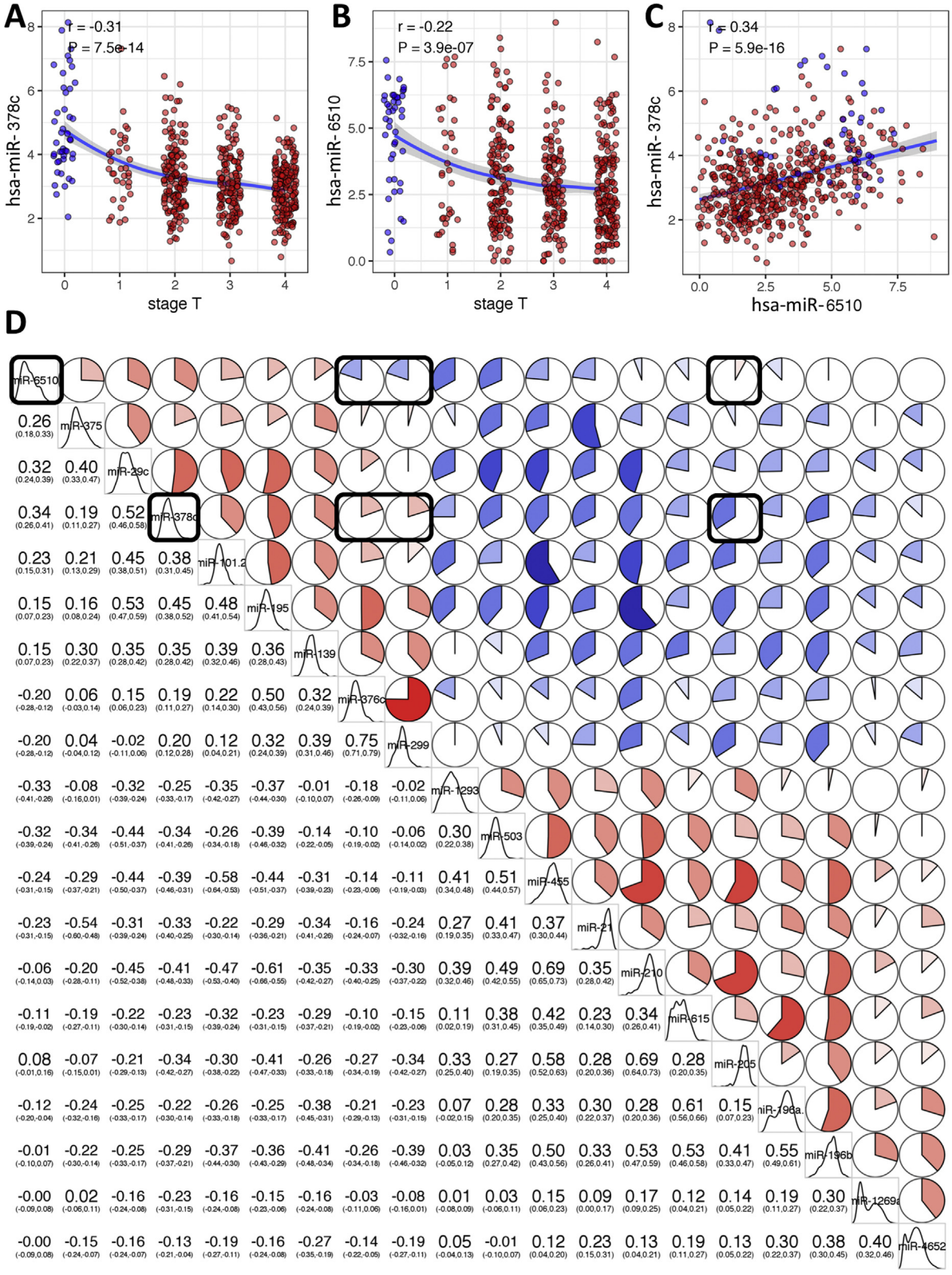


Fig. 5 – Correlations between prognostic miRNAs and clinicopathological measures. (A and B) Significant Spearman's correlation between stage T and miR-378c and miR-6510-3p, respectively. (C) Pearson's correlation between miR-378c and miR-6510-3p. Blue dots represent the observations for normal samples and red dots, for the primary tumor samples. Blue lines represent the smooth loess (A and B) or the linear (C) fit with 95% confidence interval. Correlation and corresponding *P* value are displayed in the top left corner of the graph. (D) Correlogram of all pairwise Pearson's correlations between

compared to normal tissue. Our hypothesis with this approach was that, given the diversity of tissue types and cancer locations involved, such a characteristic miRNA expression signature must be enriched for cancer-driving miRNAs, not only with biomarker potential but with real functional roles in cancer development and progression. Remarkably, sixteen (80%) of the identified DE miRNAs, including miR-21-5p,[30–33] miR-101-3p,[34–36] miR-139-5p,[37–39] miR-195-5p,[40–42] miR-196a/b,[3,43–47] miR-205-5p,[48,49] miR-210-3p,[50–52] miR-299-5p,[53,54] miR-375,[55–58] miR-376c-3p,[59–61] miR-378c,[62–66] miR-455-5p,[67–69] miR-503-3p,[70,71] and miR-1269a,[72,73] have been indeed reported to be functionally involved in the regulation of important cancer-driving pathways. Two other miRNAs, namely miR-615-3p and miR-1293, have also recently started to accumulate some evidence for a functional role in cancer.[74–76] Importantly, our approach was capable of identifying two novel miRNAs (i.e., miR-4652-3p and miR-6510-3p), for which no single report is yet found in the scientific literature. One of these miRNAs (i.e., miR-6510-3p) additionally demonstrated high importance in our random forest survival analysis and was one of the only two miRNAs that comprised the optimal iseRF survival model. These results underscore the efficiency of the data analysis approach presented in this work to suitably identify true cancer-driving miRNA-related events.

The second part of our strategy comprised the implementation of an iteratively evolved semi-exhaustive random forest strategy, dubbed iseRF, to generate an unbiased survival model with validated performance. We chose an approach based on randomized tree ensembles because this methodology has been proven highly effective to evaluate right-censored survival data, because it does not rely on restrictive assumptions (e.g., such as the proportional hazards assumption in classical Cox regression) and model the associations between covariates and hazards coherently and automatically.[19,21,77] The application of such an approach to the TCGA-HNSC miRNA-Seq data, as in this particular study, proved importantly productive and clinically relevant by identifying an efficient signature with only 7 variables that can produce the lowest and more reliable (2000-fold cross-validated) prediction error estimates reported to date (as to the best of our knowledge) in a heterogeneous cohort affected by a broad range of HNSCC subtypes. Of note, our approach accounted for a variety of confounding effects, broad range of tumor subtypes and stages, and elicited insight into the common and potentially cancer-driving HNSCC biology.

Performance measures based on external validation including the calculation of the OOB prediction error rate and time-dependent Brier score, the OOB mortality plot, the time-dependent AUC plots, and the risk-stratified survival plots, demonstrated the quality of the optimal iseRF survival model. Altogether, these results underscore the clinical value of the optimal iseRF classifier to predict the risk of death in a broad range of HNSCC patients. The high importance of three model variables, namely age at diagnosis, and tumor tissue levels of miR-378c and miR-6510-3p is demonstrated. Notably, the expression levels of both these miRNAs appeared to be significantly coregulated and to track with the cancer T stage. However, the occurrence of significant correlative differences between these miRNAs and other characteristic DE miRNAs suggests subtle and complex relationships in the HNSCC-relevant miRNA regulatory network. Based on these results and evidence from published literature, a mechanistic and potentially cancer-driving role for these characteristic and prognostic miRNAs is suggested. Evidence support a role for the downregulation of miR-378c as an important prognostic marker and potential early cancer-driving event.[62,65,66] In particular, downregulated levels of miR-378c together with changes in four other miRNAs identified in this work (i.e., miR-21, miR-196a, miR-205, and miR-210) showed altered progressive expression that reflected progression of Barrett's esophagus to esophageal adenocarcinoma.[62] No literature reports were available for miR-6510-3p, neither in cancer nor in an alternative medical research field before submission of this manuscript. Remarkably, while our article was being peer-reviewed, an original research by Chen and collaborators[78] just reported that miR-6510-3p is significantly downregulated (two-fold) in the oral mucosa of patients with oral lichen planus (OLP), as compared to healthy controls. OLP is a chronic mucocutaneous inflammatory disorder of stratified squamous epithelium, of suspected autoimmune etiology, which affects 0.1–4% of the general population.[79–81] Additional studies report that areas of erosive OLP or lichenoid lesions have increased risk of squamous cell carcinoma (SCC).[80,82–84] This evidence suggest that downregulation of miR-6510-3p, early in the process of premalignant transformation, may represent a bona fide event related to SCC development. Our own group has recently gathered additional confirmatory evidence for downregulation of miR-6510-3p in independent cohort of HNSCC samples (data not shown). We reason that the effect of miR-6510-3p downregulation in cancer (particularly in HNSCC) may have eluded confident detection due to the modest effect size and the relatively low expression levels in normal tissue. Therefore, our finding represents an important first reference suggesting a potential tumor suppressor role for human miR-6510-3p. Our GSEA implementing a highly specific (low rate of false positive) algorithm[25] revealed that genes ADH5 and AKR7A2, both involved in the metabolism of cellular aldehyde and xenobiotics, may represent critical targets of miR-6510-3p with a potential role in HNSCC development and/or the development of resistance to anti-cancer drugs. Supporting our findings, multiple ADH genes have previously been implicated in upper aerodigestive cancer etiology.[85,86] In addition, elevated ADH5 was reported to increase the risk of death in patients with pancreatic adenocarcinoma independent of a history of alcohol.[87] Similarly, aldehyde reductase AKR7A2 was reported to be highly expressed in pancreatic cancer but not in normal adjacent tissue and benign tumors.[88]

Fig. 5 – *(continued)* differentially expressed miRNAs. Top right panel displays the correlations as a pie chart, whereas the bottom left panel displays the correlation value and the 95% confidence interval. The diagonal displays the miRNA ID and the respective density distribution of the observed miRNA levels. Boxed pie charts highlight most striking correlative differences between miR-378c and miR-6510-3p.

Table 2 – Gene Set Enrichment Analysis results for miR-6510-3p and miR-4652-3p validated targets, using the SetRank package.

A. miR-6510-3p (downregulated in HNSCC)

Name	Description	Database	Size	SetRank score	*P* value	Adjusted *P* value	Gene membership		
							ADH5	AKR7A2	ABCF3
GO:0006081	Cellular aldehyde metabolic process	GOBP	85	0.48	0.0178	0.0357	X	X	.
hsa00980	Metabolism of xenobiotics by cytochrome P450	KEGG	74	0.26	0.0178	0.0357	X	X	.
GO:0009158	Ribonucleoside monophosphate catabolic process	GOBP	388	0.26	0.0294	0.0357	.	.	X

B. miR-4652-3p (upregulated in HNSCC)

Name	Description	Database	Size	SetRank score	*P* value	Adjusted *P* value	Gene membership
GO:0070898	RNA polymerase III transcriptional preinitiation complex assembly	GOBP	2	0.13	0.0225	0.1798	Supplementary Table ST3
WP2446	Retinoblastoma (RB) in Cancer	WikiPathways	88	0.10	0.0112	0.1236	Supplementary Table ST3
hsa05216	Thyroid cancer	KEGG	29	0.07	0.0112	0.1236	Supplementary Table ST3
GO:0000398	mRNA splicing, via spliceosome	GOBP	288	0.07	0.0268	0.1803	Supplementary Table ST3
GO:0016579	Protein deubiquitination	GOBP	290	0.07	0.0306	0.1803	Supplementary Table ST3
GO:0000126	Transcription factor TFIIIB complex	GOCC	2	0.07	0.0225	0.1798	Supplementary Table ST3
GO:0001156	TFIIIC-class transcription factor binding	GOMF	3	0.05	0.0225	0.1798	Supplementary Table ST3
WP2377	Integrated Pancreatic Cancer Pathway	WikiPathways	197	0.04	0.0084	0.1006	Supplementary Table ST3
GO:0005200	Structural constituent of cytoskeleton	GOMF	110	0.04	0.0112	0.1236	Supplementary Table ST3
WP2857	Mesodermal Commitment Pathway	WikiPathways	157	0.04	0.0112	0.1236	Supplementary Table ST3
GO:0002072	Optic cup morphogenesis involved in camera-type eye development	GOBP	8	0.04	0.0199	0.1793	Supplementary Table ST3
R-HSA-76071	RNA Polymerase III Transcription Initiation From Type 3 Promoter	REACTOME	29	0.04	0.0225	0.1798	Supplementary Table ST3
GO:0017124	SH3 domain binding	GOMF	124	0.04	0.0258	0.1803	Supplementary Table ST3
GO:0071013	Catalytic step 2 spliceosome	GOCC	100	0.04	0.0268	0.1803	Supplementary Table ST3
R-HSA-5688426	Deubiquitination	REACTOME	301	0.04	0.0306	0.1803	Supplementary Table ST3
WP3624	Lung fibrosis	WikiPathways	63	0.04	0.0337	0.1803	Supplementary Table ST3
GO:0006412	Translation	GOBP	433	0.04	0.0358	0.1803	Supplementary Table ST3
GO:0042472	Inner ear morphogenesis	GOBP	100	0.04	0.0392	0.1803	Supplementary Table ST3
GO:0005216	Ion channel activity	GOMF	432	0.04	0.0444	0.1803	Supplementary Table ST3

Rows sorted by SetRank Score (all *P* value < 0.05, Adj. *P* value < 0.2).

6. Conclusions

A characteristic HNSCC miRNA signature comprised of 20 DE miRNAs is reported in this study. This signature is highly enriched for miRNAs demonstrated to be involved in the regulation of important cancer-driving pathways and processes common to multiple types of cancers. These miRNAs act as tumor suppressors or oncomirs and are suggested to contribute to the development and/or progression of the disease. By association and highly specific GSEA, two novel HNSCC characteristic miRNAs, namely miR-4652-3p and miR-6510-3p, are suggested to function as oncomir and tumor suppressor, respectively.

MicroRNAs miR-378c and miR-6510-3p were specifically demonstrated to be among the three most important (most informative) variables comprising an optimal iseRF survival model with externally validated prognostic accuracy and sensitivity. These miRNAs likely act as a tumor suppressors and their downregulation in tissues might be involved in the carcinogenesis process leading to HNSCC.

Conflict of interest

No conflicts of interest are reported by the authors. Institutional organizations were involved neither in the study design, nor in the data analysis/manuscript writing, nor in the decision to submit the article for publication. YONL is the guarantor of this work and, as such, has full access to all the data in the study and takes responsibility for the integrity of the data and the data analysis.

Financial disclosure

The authors thank University of Central Florida, and The Greater Poland Cancer Centre for providing logistic support to conduct this study.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.rpor.2017.10.003.

REFERENCES

1. Gugic J, Strojan P. Squamous cell carcinoma of the head and neck in the elderly. *Rep Pract Oncol Radiother* 2012;**18**(1):16–25.
2. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer* 2010;**127**(12):2893–917.
3. (TCGA) TCGAN. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* 2015;**517**(7536):576–82.
4. González Ferreira JA, Jaén Olasolo J, Azinovic I, Jeremic B. Effect of radiotherapy delay in overall treatment time on local control and survival in head and neck cancer: review of the literature. *Rep Pract Oncol Radiother* 2015;**20**(5):328–39.
5. Gross AM, Orosco RK, Shen JP, et al. Multi-tiered genomic analysis of head and neck cancer ties TP53 mutation to 3p loss. *Nat Genet* 2014;**46**(9):939–43.
6. Sinha P, Logan HL, Mendenhall WM. Human papillomavirus, smoking, and head and neck cancer. *Am J Otolaryngol* 2012;**33**(1):130–6.
7. Zou AE, Zheng H, Saad MA, et al. The non-coding landscape of head and neck squamous cell carcinoma. *Oncotarget* 2016.
8. Wong N, Khwaja SS, Baker CM, et al. Prognostic microRNA signatures derived from The Cancer Genome Atlas for head and neck squamous cell carcinomas. *Cancer Med* 2016;**5**(7):1619–28.
9. Salyakina D, Tsinoremas NF. Non-coding RNAs profiling in head and neck cancers. *NPJ Genomic Med* 2016;**1**:15004.
10. Cortez MA, Bueso-Ramos C, Ferdin J, Lopez-Berestein G, Sood AK, Calin GA. MicroRNAs in body fluids – the mix of hormones and biomarkers. *Nat Rev Clin Oncol* 2011;**8**(8):467–77.
11. Kosaka N, Yoshioka Y, Hagiwara K, Tominaga N, Katsuda T, Ochiya T. Trash or treasure: extracellular microRNAs and cell-to-cell communication. *Front Genet* 2013;**4**:173.
12. Victoria B, Nunez Lopez YO, Masternak MM. MicroRNAs and the metabolic hallmarks of aging. *Mol Cell Endocrinol* 2017.
13. Victoria Martinez B, Dhahbi JM, Nunez Lopez YO, et al. Circulating small non-coding RNA signature in head and neck squamous cell carcinoma. *Oncotarget* 2015;**6**(22):19246–63.
14. Koshizuka K, Hanazawa T, Fukumoto I, Kikkawa N, Okamoto Y, Seki N. The microRNA signatures: aberrantly expressed microRNAs in head and neck squamous cell carcinoma. *J Hum Genet* 2017;**62**(1):3–13.
15. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature* 2009;**458**(7239):719–24.
16. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 2014;**15**(2):R29.
17. Smyth GK, Ritchie M, Thorne N, Wettenhall J, Shi W, Hu Y. *limma: linear models for microarray and RNA-Seq data user's guide*; 2016. p. 144 [Rev. 01.03.16].
18. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B (Methodol)* 1995;**57**(1):289–300.
19. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat* 2008:841–60.
20. Ishwaran H, Kogalur UB. Consistency of random survival forests. *Stat Probab Lett* 2010;**80**(13–14):1056–64.
21. Breiman L. Random forests. *Mach Learn* 2001;**45**(1):5–32.
22. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. *Random survival forests*; 2008. p. 841–60.
23. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. *BMC Bioinform* 2006;**7**(1):3.
24. Ru Y, Kechris KJ, Tabakoff B, et al. The multiMiR R package and database: integration of microRNA–target interactions along with their disease and drug associations. *Nucleic Acids Res* 2014;**42**(17):e133.
25. Simillion C, Liechti R, Lischer HE, Ioannidis V, Bruggmann R. Avoiding the pitfalls of gene set enrichment analysis with SetRank. *BMC Bioinform* 2017;**18**(1):151.
26. Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. In: *Proceedings of the seventh international conference on World Wide Web 7*. 1998.
27. Ehrlinger J. *ggRandomForests: visually exploring a random forest for regression. arXiv*; 2015.
28. Takes RP, Rinaldo A, Silver CE, et al. Future of the TNM classification and staging system in head and neck cancer. *Head Neck* 2010;**32**(12):1693–711.
29. Brierley JD, Gospodarowicz MK, Wittekind C. *TNM classification of malignant tumours*. John Wiley & Sons; 2016.

30. Melnik BC. MiR-21: an environmental driver of malignant melanoma? *J Transl Med* 2015;**13**:202.
31. Hirata Y, Murai N, Yanaihara N, et al. MicroRNA-21 is a candidate driver gene for 17q23-25 amplification in ovarian clear cell carcinoma. *BMC Cancer* 2014;**14**:799.
32. Cao J, Liu J, Xu R, Zhu X, Liu L, Zhao X. MicroRNA-21 stimulates epithelial-to-mesenchymal transition and tumorigenesis in clear cell renal cells. *Mol Med Rep* 2016;**13**(1):75–82.
33. Bera A, Das F, Ghosh-Choudhury N, Kasinath BS, Abboud HE, Choudhury GG. microRNA-21-induced dissociation of PDCD4 from rictor contributes to Akt-IKKbeta-mTORC1 axis to regulate renal cancer cell invasion. *Exp Cell Res* 2014;**328**(1):99–117.
34. Liu P, Ye F, Xie X, et al. mir-101-3p is a key regulator of tumor metabolism in triple negative breast cancer targeting AMPK. *Oncotarget* 2016;**7**(23):35188–98.
35. Gong J, Chu Y, Xu M, Huo J, Lv L. Esophageal squamous cell carcinoma cell proliferation induced by exposure to low concentration of cigarette smoke extract is mediated via targeting miR-101-3p/COX-2 pathway. *Oncol Rep* 2016;**35**(1):463–71.
36. Cui Y, Zhang F, Zhu C, Geng L, Tian T, Liu H. Upregulated lncRNA SNHG1 contributes to progression of non-small cell lung cancer through inhibition of miR-101-3p and activation of Wnt/beta-catenin signaling pathway. *Oncotarget* 2017.
37. Maoa R, Zou F, Yang L, et al. The loss of MiR-139-5p promotes colitis-associated tumorigenesis by mediating PI3K/AKT/Wnt signaling. *Int J Biochem Cell Biol* 2015;**69**:153–61.
38. Krowiorz K, Ruschmann J, Lai C, et al. MiR-139-5p is a potent tumor suppressor in adult acute myeloid leukemia. *Blood Cancer J* 2016;**6**(12):e508.
39. Luo HN, Wang ZH, Sheng Y, et al. MiR-139 targets CXCR4 and inhibits the proliferation and metastasis of laryngeal squamous carcinoma cells. *Med Oncol (Northwood, London, England)* 2014;**31**(1):789.
40. Xu H, Hu YW, Zhao JY, et al. MicroRNA-195-5p acts as an anti-oncogene by targeting PHF19 in hepatocellular carcinoma. *Oncol Rep* 2015;**34**(1):175–82.
41. Qu Q, Chu X, Wang P. MicroRNA-195-5p suppresses osteosarcoma cell proliferation and invasion by suppressing naked cuticle homolog 1. *Cell Biol Int* 2017;**41**(3):287–95.
42. Luo Q, Zhang Z, Dai Z, et al. Tumor-suppressive microRNA-195-5p regulates cell growth and inhibits cell cycle by targeting cyclin dependent kinase 8 in colon cancer. *Am J Transl Res* 2016;**8**(5):2088–96.
43. Shen S, Pan J, Lu X, Chi P. Role of miR-196 and its target gene HoxB8 in the development and proliferation of human colorectal cancer and the impact of neoadjuvant chemotherapy with FOLFOX4 on their expression. *Oncol Lett* 2016;**12**(5):4041–7.
44. Schotte D, Lange-Turenhout EA, Stumpel DJ, et al. Expression of miR-196b is not exclusively MLL-driven but is especially linked to activation of HOXA genes in pediatric acute lymphoblastic leukemia. *Haematologica* 2010;**95**(10):1675–82.
45. Lu YC, Chang JT, Liao CT, et al. OncomiR-196 promotes an invasive phenotype in oral cancer through the NME4-JNK-TIMP1-MMP signaling pathway. *Mol Cancer* 2014;**13**:218.
46. Liu CJ, Tsai MM, Tu HF, Lui MT, Cheng HW, Lin SC. miR-196a overexpression and miR-196a2 gene polymorphism are prognostic predictors of oral carcinomas. *Ann Surg Oncol* 2013;**20**(Suppl. 3):S406–14.
47. Chen C, Zhang Y, Zhang L, Weakley SM, Yao Q. MicroRNA-196: critical roles and clinical applications in development and cancer. *J Cell Mol Med* 2011;**15**(1):14–23.
48. Xin W, Liu X, Ding J, et al. Long non-coding RNA derived miR-205-5p modulates human endometrial cancer by targeting PTEN. *Am J Transl Res* 2015;**7**(11):2433–41.
49. De Cola A, Volpe S, Budani MC, et al. miR-205-5p-mediated downregulation of ErbB/HER receptors in breast cancer stem cells results in targeted therapy resistance. *Cell Death Dis* 2015;**6**:e1823.
50. Yoshino H, Yonemori M, Miyamoto K, et al. microRNA-210-3p depletion by CRISPR/Cas9 promoted tumorigenesis through revival of TWIST1 in renal cell carcinoma. *Oncotarget* 2017.
51. Yang Y, Zhang J, Xia T, et al. MicroRNA-210 promotes cancer angiogenesis by targeting fibroblast growth factor receptor-like 1 in hepatocellular carcinoma. *Oncol Rep* 2016;**36**(5):2553–62.
52. Ullmann P, Qureshi-Baig K, Rodriguez F, et al. Hypoxia-responsive miR-210 promotes self-renewal capacity of colon tumor-initiating cells by repressing ISCU and by inducing lactate production. *Oncotarget* 2016;**7**(40):65454–70.
53. Shevde LA, Metge BJ, Mitra A, et al. Spheroid-forming subpopulation of breast cancer cells demonstrates vasculogenic mimicry via hsa-miR-299-5p regulated de novo expression of osteopontin. *J Cell Mol Med* 2010;**14**(6B):1693–706.
54. Formosa A, Markert EK, Lena AM, et al. MicroRNAs, miR-154, miR-299-5p, miR-376a, miR-376c, miR-377, miR-381, miR-487b, miR-485-3p, miR-495 and miR-654-3p, mapped to the 14q32.31 locus, regulate proliferation, apoptosis, migration and invasion in metastatic prostate cancer cells. *Oncogene* 2014;**33**(44):5173–82.
55. Liu AM, Poon RT, Luk JM. MicroRNA-375 targets Hippo-signaling effector YAP in liver cancer and inhibits tumor properties. *Biochem Biophys Res Commun* 2010;**394**(3):623–7.
56. de Souza Rocha Simonini P, Breiling A, Gupta N, et al. Epigenetically deregulated microRNA-375 is involved in a positive feedback loop with estrogen receptor alpha in breast cancer cells. *Cancer Res* 2010;**70**(22):9175–84.
57. Bierkens M, Krijgsman O, Wilting SM, et al. Focal aberrations indicate EYA2 and hsa-miR-375 as oncogene and tumor suppressor in cervical carcinogenesis. *Genes Chromosom Cancer* 2013;**52**(1):56–68.
58. Kinoshita T, Hanazawa T, Nohata N, Okamoto Y, Seki N. The functional significance of microRNA-375 in human squamous cell carcinoma: aberrant expression and effects on cancer pathways. *J Hum Genet* 2012;**57**(9):556–63.
59. Zhang YH, Fu J, Zhang ZJ, Ge CC, Yi Y. LncRNA-LINC00152 down-regulated by miR-376c-3p restricts viability and promotes apoptosis of colorectal cancer cells. *Am J Transl Res* 2016;**8**(12):5286–97.
60. Tu L, Zhao E, Zhao W, et al. hsa-miR-376c-3p regulates gastric tumor growth both in vitro and in vivo. *BioMed Res Int* 2016;**2016**:9604257.
61. Chang WM, Lin YF, Su CY, et al. Dysregulation of RUNX2/Activin-A axis upon miR-376c downregulation promotes lymph node metastasis in head and neck squamous cell carcinoma. *Cancer Res* 2016;**76**(24):7140–50.
62. Slaby O, Srovnal J, Radova L, et al. Dynamic changes in microRNA expression profiles reflect progression of Barrett's esophagus to esophageal adenocarcinoma. *Carcinogenesis* 2015;**36**(5):521–7.
63. Hummel R, Sie C, Watson DI, et al. MicroRNA signatures in chemotherapy resistant esophageal cancer cell lines. *World J Gastroenterol* 2014;**20**(40):14904–12.
64. Yu BL, Peng XH, Zhao FP, et al. MicroRNA-378 functions as an onco-miR in nasopharyngeal carcinoma by repressing TOB2 expression. *Int J Oncol* 2014;**44**(4):1215–22.

65. Scapoli L, Palmieri A, Lo Muzio L, et al. MicroRNA expression profiling of oral carcinoma identifies new markers of tumor progression. *Int J Immunopathol Pharmacol* 2010;**23**(4):1229–34.
66. Liu X, Luo HN, Tian WD, et al. Diagnostic and prognostic value of plasma microRNA deregulation in nasopharyngeal carcinoma. *Cancer Biol Ther* 2013;**14**(12):1133–42.
67. Shoshan E, Mobley AK, Braeuer RR, et al. Reduced adenosine-to-inosine miR-455-5p editing promotes melanoma growth and metastasis. *Nat Cell Biol* 2015;**17**(3):311–21.
68. Liu J, Zhang J, Li Y, Wang L, Sui B, Dai D. MiR-455-5p acts as a novel tumor suppressor in gastric cancer by down-regulating RAB18. *Gene* 2016;**592**(2):308–15.
69. Cheng CM, Shiah SG, Huang CC, Hsiao JR, Chang JY. Up-regulation of miR-455-5p by the TGF-beta-SMAD signalling axis promotes the proliferation of oral squamous cancer cells by targeting UBE2B. *J Pathol* 2016;**240**(1):38–49.
70. Zhao Z, Fan X, Jiang L, et al. miR-503-3p promotes epithelial-mesenchymal transition in breast cancer by directly targeting SMAD2 and E-cadherin. *J Genet Genomics = Yi chuan xue bao* 2017;**44**(2):75–84.
71. Noguchi T, Toiyama Y, Kitajima T, et al. miRNA-503 promotes tumor progression and is associated with early recurrence and poor prognosis in human colorectal cancer. *Oncology* 2016;**90**(4):221–31.
72. Min P, Li W, Zeng D, et al. A single nucleotide variant in microRNA-1269a promotes the occurrence and process of hepatocellular carcinoma by targeting to oncogenes SPATS2L and LRP6. *Bulletin du cancer* 2017.
73. Bu P, Wang L, Chen KY, et al. miR-1269 promotes metastasis and forms a positive feedback loop with TGF-beta. *Nat Commun* 2015;**6**:6879.
74. Mukai R, Tomimaru Y, Nagano H, et al. miR-615-3p expression level in bone marrow is associated with tumor recurrence in hepatocellular carcinoma. *Mol Clin Oncol* 2015;**3**(3):487–94.
75. Guled M, Lahti L, Lindholm PM, et al. CDKN2A, NF2, and JUN are dysregulated among other genes by miRNAs in malignant mesothelioma – a miRNA microarray analysis. *Genes Chromosom Cancer* 2009;**48**(7):615–23.
76. De Sarkar N, Roy R, Mitra JK, et al. A quest for miRNA bio-marker: a track back approach from gingivo buccal cancer to two different types of precancers. *PLOS ONE* 2014;**9**(8):e104839.
77. Datema FR, Moya A, Krause P, et al. Novel head and neck cancer survival analysis approach: random survival forests versus Cox proportional hazards regression. *Head Neck* 2012;**34**(1):50–8.
78. Chen J, Du G, Wang Y, Shi L, Mi J, Tang G. Integrative analysis of mRNA and miRNA expression profiles in oral lichen planus: preliminary results. *Oral Surg Oral Med Oral Pathol Oral Radiol* 2017.
79. Sugerman PB, Savage NW, Walsh LJ, et al. The pathogenesis of oral lichen planus. *Crit Rev Oral Biol Med* 2002;**13**(4):350–65.
80. Edwards PC, Kelsch R. Oral lichen planus: clinical presentation and management. *J Can Dent Assoc* 2002;**68**(8):494–9.
81. Gupta S, Jawanda MK. Oral lichen planus: an update on etiology, pathogenesis, clinical presentation, diagnosis and management. *Indian J Dermatol* 2015;**60**(3):222–9.
82. Fitzpatrick SG, Hirsch SA, Gordon SC. The malignant transformation of oral lichen planus and oral lichenoid lesions: a systematic review. *J Am Dent Assoc (1939)* 2014;**145**(1):45–56.
83. Peng Q, Zhang J, Ye X, Zhou G. Tumor-like microenvironment in oral lichen planus: evidence of malignant transformation? *Expert Rev Clin Immunol* 2017;**13**(6):635–43.
84. Aghbari SMH, Abushouk AI, Attia A, et al. Malignant transformation of oral lichen planus and oral lichenoid lesions: a meta-analysis of 20095 patient data. *Oral Oncol* 2017;**68**:92–102.
85. Oze I, Matsuo K, Suzuki T, et al. Impact of multiple alcohol dehydrogenase gene polymorphisms on risk of upper aerodigestive tract cancers in a Japanese population. *Cancer Epidemiol Biomark Prev* 2009;**18**(11):3097–102.
86. Hashibe M, McKay JD, Curado MP, et al. Multiple ADH genes are associated with upper aerodigestive cancers. *Nat Genet* 2008;**40**(6):707–9.
87. Liao X, Huang R, Liu X, et al. Distinct prognostic values of alcohol dehydrogenase mRNA expression in pancreatic adenocarcinoma. *OncoTargets Ther* 2017;**10**:3719–32.
88. Cui Y, Tian M, Zong M, et al. Proteomic analysis of pancreatic ductal adenocarcinoma compared with normal adjacent pancreatic tissue and pancreatic benign cystadenoma. *Pancreatology* 2009;**9**(1–2):89–98.