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