Salivary microRNAs identified by small RNA sequencing and machine learning as potential biomarkers of alcohol dependence

Andrew J Rosato¹, Xiaochun Chen¹, Yoshiaki Tanaka², Lindsay A Farrer^{3,4,5,6,7}, Henry R Kranzler⁸, Yaira Z Nunez¹, David C Henderson¹, Joel Gelernter^{2,9,10,11} & Huiping Zhang^{*,1,3}

¹Department of Psychiatry, Boston University School of Medicine, Boston, MA 02118, USA

²Department of Genetics, Yale University School of Medicine, New Haven, CT 06520, USA

³Department of Medicine (Biomedical Genetics), Boston University School of Medicine, Boston, MA 02118, USA

⁴Department of Neurology, Boston University School of Medicine, Boston, MA 02118, USA

⁵Department of Ophthalmology, Boston University School of Medicine, Boston, MA 02118, USA

⁶Department of Epidemiology & Boston University School of Public Health, Boston, MA 02118, USA

⁷Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA

⁸Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania & VISN4 MIRECC, Crescenzo VAMC, Philadelphia, PA 19104, USA

⁹Department of Psychiatry, Yale University School of Medicine, New Haven, CT 06511, USA

¹⁰Department of Neuroscience, Yale University School of Medicine, New Haven, CT 06510, USA

¹¹VA Connecticut Healthcare System, West Haven, CT 06516, USA

*Author for correspondence: Tel.: +1 617 358 3689; Fax: +1 617 414 1996; huipingz@bu.edu

Aim: Salivary miRNA can be easily accessible biomarkers of alcohol dependence (AD). **Materials & methods:** The miRNA transcriptome in the saliva of 56 African-Americans (AAs; 28 AD patients/28 controls) and 64 European-Americans (EAs; 32 AD patients/32 controls) was profiled using small RNA sequencing. Differentially expressed miRNAs were identified. Salivary miRNAs were used to predict the AD presence using machine learning with Random Forests. **Results:** Seven miRNAs were differentially expressed in AA AD patients, and five miRNAs were differentially expressed in EA AD patients. The AD prediction accuracy based on top five miRNAs (ranked by Gini index) was 79.1 and 72.2% in AAs and EAs, respectively. **Conclusion:** This study provided the first evidence that salivary miRNAs are AD biomarkers.

First draft submitted: 15 October 2018; Accepted for publication: 4 February 2019; Published online: 29 May 2019

Keywords: alcohol dependence • differential expression • machine learning • salivary microRNA • small RNA sequencing

Alcohol dependence (AD) is a common, complex and genetically influenced disorder. A reliable diagnostic tool for AD is needed to support efforts at prevention and treatment of the disorder. A current AD diagnosis depends primarily on self-reported symptoms, which are limited by inaccurate recall or reluctance of patients to give accurate information on their drinking behaviors or alcohol-related problems. Thus, there is considerable interest in the identification of biological measurements (or biomarkers) to assess a patient's current or past alcohol use.

Biochemical markers such as liver enzymes (e.g., γ -glutamyltransferase, aspartate aminotransferase and alanine aminotransferase) have been used to detect excessive ethanol consumption [1]. However, patients with liver diseases also have increased levels of these liver enzymes. Elevated erythrocyte macrocytic volume (MCV) is also common in AD patients [2], but the slow return of MCV to the reference value diminishes its potential as a relapse marker. Moreover, patients with macrocytic anemia also have an increased MCV [3]. Blood levels of alcohol and its byproducts (such as acetaldehyde, ethyl glucuronide and fatty acid ethyl ester) can reflect acute alcohol ingestion but not past drinking patterns or alcohol relapse [4]. Phosphatidylethanol (PEth), another byproduct of ethanol, can reflect drinking over the preceding weeks [5]. Most of the above biochemical markers (excluding PEth) are limited by their sensitivity and specificity in assessing alcohol abuse.

Genetic and epigenetic factors are potential biomarkers of AD. Genome-wide association studies have identified AD-associated genetic variants, particularly those located in alcohol-metabolizing enzyme genes [6,7]. Nevertheless, the effect size of these genetic variants on AD risk is small [8]. So far, no genetic markers have been used as practical biomarkers for diagnosis of AD. Epigenetic markers, particularly miRNAs, are potential AD biomarkers. miRNAs are a class of small noncoding RNA molecules (containing about 22 nucleotides) that regulate gene expression via either translational repression or mRNA degradation at the post-transcriptional level [9–11]. As each miRNA can target multiple mRNAs and each mRNA can be regulated by multiple miRNAs, miRNAs play diverse roles in many cellular processes [12]. miRNAs have been implicated in a number of diseases, particularly cancer [13]. Evidence from rat [14] and human *post mortem* brain [15,16] studies suggest that adaptations to alcohol may be due in part to altered expression of a group of miRNAs and their target genes.

Although information on miRNA expression in postmortem brains of patients is critical for unraveling the epigenetic mechanisms of neuropsychiatric disorders including AD, it is of little use clinically because brain tissues are not easily accessible. Besides the existence of miRNAs in tissues and cells, miRNAs are also present in extracellular or body fluids including saliva. Extracellular miRNAs are highly stable and also RNase resistant because they are either contained in membranous vesicles [17,18] or bound to Argonaute proteins [19] or HDL [20]. There is evidence that extracellular miRNAs (such as salivary miRNAs) can serve as informative biomarkers for assessing the severity or presence of diseases [21,22]. However, no study is known to have examined miRNA expression alterations in the saliva of AD subjects.

In the present study, we investigated whether salivary miRNAs are potential biomarkers for detection of AD. We first used miRNA sequencing (miRNA-seq) technology to profile miRNA transcriptomes in the saliva of AD patients and healthy control subjects from both African–American (AA) and European–American (EA) populations. We then identified salivary miRNAs that are differentially expressed in AD patients, and used a machine learning approach to explore the utility of salivary miRNAs as biomarkers for identifying AD.

Materials & methods

Participants

56 African–Americans (AAs; 28 AD patients and 28 control subjects) and 64 European–Americans (EAs; 32 AD patients and 32 control subjects) participated in the present study. Participants were recruited from the community using advertisements such as posted bulletin board flyers and online advertising as well as word of mouth referrals. Participants were screened via telephone and scheduled for appointments if they met the inclusion/exclusion criteria. They were interviewed at the Yale University School of Medicine (APT Foundation; CT, USA). All subjects gave written informed consent to participate in the study. They were assessed using the Semi-Structured Assessment for Drug Dependence and Alcoholism [23,24] to derive diagnoses for lifetime substance use disorders including AD. All subjects were not affected with major psychotic disorders (schizophrenia and bipolar disorder). Among the 60 AD patients, 88.3% of them had one or more co-morbid other substance use disorders (such as cocaine, opioid, nicotine, marijuana, sedative and stimulant dependence). Control subjects were not affected with these substance use disorders. The participants were not seeking treatment, and abstinence from substances was not an inclusion criterion for either group. Demographic characteristics of the sample are summarized in Table 1.

Saliva collection & total RNA extraction

Saliva samples were collected from the above 120 subjects (56 AAs and 64 EAs). Subjects were asked to refrain from eating, drinking, smoking or chewing gum for 30 min before giving saliva samples. Whole saliva (about 2 ml) was collected using the Oragene•RNA (RE-100) for Expression Analysis Self-Collection Kit (DNA Genotek, Ottawa, Canada). Total RNA was extracted from the cell-free supernatant using the TRIzol LS Reagent (Life Technologies, CA, USA). Extracted RNA samples were further purified using the Agencourt RNAClean XP Kit (Beckman Coulter, MA, USA). RNA was quantified using a NanoDrop 8000 Spectrophotometer (Thermo Fisher Scientific, MA, USA). The size distribution and quality of extracted RNAs were assessed on an Agilent 2100 Bioanalyzer using an Agilent RNA 6000 Nano Chip. The RNA yield was about 150 ng/ml of saliva supernatant, and the mean RNA integrity number was 6.7.

miRNA-seq library preparation & sequencing

miRNA-seq was conducted at the Yale Center for Genome Analysis. Total RNAs (250 ng) extracted from saliva were used to construct small RNA sequencing libraries using the NEBNext® Multiplex Small RNA Library Prep

Table 1. Demographic information of the sample.

Demographic categories	African-Americans		χ^2 /t-test	European-Americans		χ^2 /t-test
	AD cases (n = 28)	Controls (n = 28)		AD cases (n = 32)	Controls (n = 32)	
Sex, males	12 (42.9%)	12 (42.9%)	$\chi^2 = 1.00$, $p = 1.000$	16 (50.0%)	16 (50.0%)	$\chi^2 = 1.00$, $p = 1.000$
Age, year (mean $\pm$ SD)	39 $\pm$ 13	39 $\pm$ 12	$t = -0.11$, $p = 0.915$	40 $\pm$ 12	46 $\pm$ 16	$t = 0.99$, $p = 0.324$
RIN (mean $\pm$ SD)	6.7 $\pm$ 1.7	6.3 $\pm$ 1.8	$t = -0.86$, $p = 0.392$	7.2 $\pm$ 1.2	6.7 $\pm$ 1.6	$t = -1.34$, $p = 0.186$
Co-morbid substance use disorders:						
– Cocaine dependence	11 (39.3%)	0 (0%)	–	23 (71.9%)	0 (0%)	–
– Opioid dependence	4 (14.3%)	0 (0%)	–	17 (53.1%)	0 (0%)	–
– Nicotine dependence	12 (42.9%)	0 (0%)	–	21 (65.6%)	0 (0%)	–
– Marijuana dependence	20 (71.4%)	0 (0%)	–	11 (34.4%)	0 (0%)	–
– Sedative dependence	1 (3.6%)	0 (0%)	–	8 (25.0%)	0 (0%)	–
– Stimulant dependence	1 (3.6%)	0 (0%)	–	4 (12.5%)	0 (0%)	–

AD: Alcohol dependence; RIN: RNA integrity number; SD: Standard deviation.

Set for Illumina® (Set 1; New England Biolabs, MA, USA) following the manufacturer's instruction manual. First, the 3' SR adaptor was ligated to the 3' end of RNAs, the reverse transcription primer was hybridized to the excess of the 3' SR adaptor, and the 5' SR adaptor was ligated to the 5' end of the RNAs. Then, the ligation products were subjected to reverse transcription reactions to create single-stranded cDNAs. To enrich fragments with adapters on both ends selectively, cDNAs were amplified with 15 cycles of PCRs using a common primer and a primer containing an index tag (6 nt), which facilitated multiplexing and sequencing of different samples in a single lane of a flow cell. Size selection of miRNA-seq libraries (bands of 147 bp, corresponding to the size of adaptor-ligated miRNAs) was performed on a 6% polyacrylamide gel. The size, purity and concentration of miRNA-seq libraries were further assessed on an Agilent 2100 Bioanalyzer using a DNA 1000 chip. Finally, libraries generated from 12 salivary RNA samples were pooled and loaded in one lane of a flow cell for cluster formation. The colonized DNA served as the template for single-end 75-cycle sequencing using the HiSeq 2500 Sequencing System (Illumina, CA, USA).

miRNA-seq data processing

Raw sequence reads were processed by the Mapper module of miRDeep2 (v2.0.0.8) [25] to remove entries with noncanonical letters, clip the 3' adaptor sequence AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC, discard sequence reads shorter than 18 nt and collapse identical sequence reads. Mapping sequence reads to the human genome (hg19) was performed by miRDeep2 mapper.pl script with the '-e -q -r 100 -s -h -n -m -j -l 18 -k AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC' options. The output files with sequence reads mapped to the human genome (hg19) were further processed with the miRDeep2.pl script to identify human miRNAs that were annotated in the miRBase database (v20) [26]. The total number of sequence reads for each miRNA was normalized to counts per million (CPM) by the total number of mapped sequence reads per sample. Three samples (one AA control sample and two EA control samples) were failed in miRNA-seq.

miRNA differential expression analysis

We performed miRNA differential expression analysis with the Bioconductor package edgeR [27,28]. miRNA expression levels (or counts of miRNAs per sample) were imported to edgeR and converted to CPM. The CPM were then normalized using the method TMM (trimmed means of M-values), which removed the miRNAs that were extremely low or high in expression and those that differed extensively across samples. The negative binomial distribution was used to model the variance of miRNA expression levels. A generalized linear model framework was used to compare miRNA expression differences between cases and controls, which included covariates sex, age,

and RNA integrity number (RIN). To identify common salivary miRNAs as biomarkers of AD, only those with CPM ≥ 100 in at least half of the subjects were retained in the AD prediction analysis.

miRNA target gene prediction & functional annotation

Genes (or mRNAs) potentially targeted by differentially expressed miRNAs were predicted using miRWalk (v3.0), the online Database on Predicted and Validated miRNA Targets [29,30]. Target genes with a prediction score of more than 0.8 and validated by a third party database such as TargetScan [31], miRDB [32] or miRTarBase [33] were subjected to gene annotation enrichment analysis, which was performed using the online Database for Annotation, Visualization and Integrated Discovery (DAVID; v6.8) [34].

Machine learning & AD prediction by salivary miRNAs

A Random Forest (RF) machine learning approach was applied to identify influential miRNAs for AD prediction using the randomForest R package [35]. An RF model was run using the AD status and miRNA expression levels of subjects in the dataset. The RF algorithm treated the expression level of each miRNA as a different variable in each decision tree and calculated each variable's importance to the model. Included in the importance calculation was each variable's Gini index [36], a measure of the importance of miRNAs in the RF model in predicting the disease status of subjects. In this study, each miRNA was assigned a Gini index based on its contribution to differentiate samples of AD and control subjects. The top ten, five or three miRNAs ranked by the Gini index were applied in AD prediction. The dataset was divided into training and test sets using random sampling for multiple train/test ratios: 50, 60, 70, 80 and 90%. An RF neural network model was created using the training dataset containing the case and control phenotype information. AD prediction was then performed using the RF neural network model on the test dataset. A confusion matrix was generated by comparing the predicted AD status of subjects to their actual status in the test dataset, which yielded the accuracy, sensitivity and specificity of the prediction analysis. The analysis for each ratio was performed ten-times, each time using a seed generated from a different random number. The mean of each of the above three statistics (accuracy, sensitivity and specificity) in ten permutations served as the final result.

Results

Differentially expressed salivary miRNAs in AD subjects

The average number of miRNA-seq reads per AA subject was 2,313,896 ($\pm 1,731,731$), with 29.0% of the sequence reads mapped to the human genome and 0.2% mapped to human miRNA sequences. The average number of miRNA-seq reads per EA subject was 2,193,734 ($\pm 1,486,205$), with 27.6% of the sequence reads mapped to the human genome and 0.2% mapped to human miRNA sequences. Among 2,588 different miRNAs detected in the saliva, 399 were expressed at a level of no less than 100 CPM in at least half of the subjects. Expression differences of these 399 miRNAs between cases and controls were analyzed by edgeR, and the results were visualized by volcano plots (Figure 1A & B). A list of miRNAs with $p < 0.050$ and FC (fold-change) > 2 is provided in Table 2. Seven such miRNAs (miR-451a, miR-10a-5p, miR-100-5p, miR-3613-5p, miR-7704, miR-1290 and miR-4488) in AAs and five such miRNAs (miR-126-3p, miR-10a-5p, miR-1290, miR-4488 and miR-1273h-5p) in EAs were identified. Although the results did not withstand multiple testing correction (false discovery rate > 0.05), expression changes of three miRNAs (miR-10a-5p, miR-1290 and miR-4488) were cross-validated in both AAs and EAs. They showed similar fold changes and the same direction of expression changes in both AA and EA AD subjects.

Pathways enriched in genes potentially targeted by differentially expressed miRNAs

Genes potentially targeted by the above three cross-validated miRNAs (miR-10a-5p, miR-1290 and miR-4488) were predicted by miRWalk and validated by TargetScan, miRDB or miRTarBase. When the miRWalk prediction score was set at 0.8, 16 genes were predicted to be targets of miR-10a-5p, 104 genes were predicted to be targets of miR-1290, and 46 genes were predicted to be targets of miR-4488. Genes potentially targeted by these miRNAs were significantly overrepresented in gene ontology categories of DNA binding for miR-10a-5p ($P_{\text{Bonferroni}} = 0.019$), alternative splicing for miR-1290 ($P_{\text{Bonferroni}} = 0.021$), and calcium-dependent cell-cell adhesion for miR-4488 ($P_{\text{Bonferroni}} = 0.001$) (Supplementary Table 1).

Figure 1. Volcano plot of miRNA expression changes. The volcano plot shows miRNA expression changes in the saliva of AAs (A) and EAs (B). The y-axis corresponds to the $\log_{10}(\text{p-value})$, and the x-axis displays the $\log^2$ fold change value. The red dots represent differentially expressed miRNAs with $p < 0.01$; the blue dots represent differentially expressed miRNAs with $p < 0.05$; and black dots represent miRNAs whose expression levels did not reach statistical significance ($p > 0.05$).

AA: African-American; EA: European-American.

Table 2. Differentially expressed miRNAs in the saliva of subjects with alcohol dependence.

In AAs	logFC	logCPM	LR	p-value	FDR
hsa-miR-451a	1.93	10.75	7.49	0.006	>0.05
hsa-miR-10a-5p†	2.18	12.72	5.83	0.016	>0.05
hsa-miR-100-5p	1.43	12.49	4.79	0.029	>0.05
hsa-miR-3613-5p	1.77	10.71	4.22	0.040	>0.05
hsa-miR-7704	-1.07	11.93	4.21	0.040	>0.05
hsa-miR-1290†	-1.12	11.58	4.15	0.042	>0.05
hsa-miR-4488†	-1.06	12.64	4.15	0.042	>0.05
In EAs	logFC	logCPM	LR	p-value	FDR
hsa-miR-126-3p	2.00	12.90	6.77	0.009	>0.05
hsa-miR-10a-5p†	2.30	12.75	6.49	0.011	>0.05
hsa-miR-1290†	-1.35	11.40	6.17	0.013	>0.05
hsa-miR-4488†	-1.11	12.98	4.35	0.037	>0.05
hsa-miR-1273h-5p	-1.98	13.48	4.26	0.039	>0.05

† Consistent results in both AAs and EAs.
 AA: African-American; EA: European-American; FDR: False discovery rate; logCPM: $\log_2(\text{counts per million})$; logFC: $\log_2(\text{fold change})$; LR: Likelihood ratio.

AD prediction by machine learning

The machine learning RF algorithm was used to define a cluster of miRNAs that predict AD status. Based on the expression levels of 399 common salivary miRNAs and the AD status of each subject, the RF algorithm constructed a multitude of decision trees at the training time and then output the mode of classes. Each miRNA was assigned a Gini index based on its contribution to differentiate samples of AD and control subjects. The Gini index of the top ten miRNAs in AAs and EAs are shown in Figure 2A & 2B, respectively. When the top five miRNAs (ranked by Gini index or their importance to AD prediction) were included in RF prediction analyses using a train/test sample ratio of 80/20, the AD prediction accuracy was 79.1 and 72.2% in AAs and EAs, respectively (Table 3). Inclusion of the top ten miRNAs (ranked by Gini index or their importance to AD prediction) in the RF prediction analyses

Figure 2. Machine learning-generated Gini indexes of the top ten miRNAs. The x-axis corresponds to Gini indexes generated by the Random Forest machine learning algorithm. The y-axis corresponds to the top ten miRNAs in AAs (A) and EAs (B). AA: African-American; EA: European-American.

decreased the AD prediction accuracy to 73.6% in AAs but slightly increased it to 75.4% in EAs. If the top three miRNAs (ranked by GINI index or their importance to AD prediction) were included in RF prediction analyses, the AD prediction accuracy was decreased in AAs (76.4%) and to a greater extent in EAs (64.6%).

Discussion

The identification of disease-specific biomarkers in easily accessible body fluids such as saliva can result in the early diagnosis and treatment of diseases. Given the important role of miRNAs in post-transcriptional regulation of gene expression and their stability and detectability in saliva, salivary miRNAs have been proposed as specific and reliable

Table 3. Random Forest prediction of alcohol dependence using miRNAs with the highest values of Gini index.

African-Americans				European-Americans			
Train/test	Accuracy (%)	Sensitivity (%)	Specificity (%)	Train/test	Accuracy (%)	Sensitivity (%)	Specificity (%)
Predicted by top ten miRNAs							
50/50	70.4	76.2	65.0	50/50	65.8	65.1	67.3
60/40	68.6	72.8	64.8	60/40	66.8	69.7	64.5
70/30	71.8	75.0	66.5	70/30	72.1	74.2	70.5
80/20	73.6	79.2	67.6	80/20	75.4	77.1	73.6
90/10	68.3	66.0	71.7	90/10	74.3	89.0	56.7
Predicted by top five miRNAs							
50/50	69.3	75.5	63.9	50/50	64.8	66.3	63.9
60/40	70.5	74.8	66.1	60/40	67.2	67.3	67.8
70/30	75.9	79.9	70.1	70/30	72.6	72.0	72.5
80/20	79.1	84.9	71.5	80/20	72.2	71.0	73.9
90/10	78.3	87.7	72.5	90/10	65.7	75.0	54.2
Predicted by top three miRNAs							
50/50	67.1	76.4	57.8	50/50	62.6	65.4	60.9
60/40	68.8	74.6	58.2	60/40	66.0	70.9	60.4
70/30	75.9	79.9	70.0	70/30	66.3	68.8	63.4
80/20	76.4	87.4	61.1	80/20	64.6	66.7	60.0
90/10	75.0	89.7	56.7	90/10	68.6	76.5	51.7

Top ten miRNAs in African-Americans (ordered by Gini indexes from high to low): miR-1303, miR-4792, miR-499a-5p, miR-4488, miR-619-5p, miR-324-5p, miR-151b, miR-4763-5p, miR-151a-5p, miR-1307-5p.

Top ten miRNAs in European-Americans (ordered by Gini indexes from high to low): miR-1303, miR-6499-5p, miR-27a-5p, miR-589-5p, miR-4488, miR-499a-5p, miR-1307-5p, miR-324-5p, miR-378a-3p, and miR-4763-5p.

biomarkers for the noninvasive diagnosis of diseases, including AD. The present study provided initial evidence that salivary miRNAs are potential biomarkers for AD prediction.

First, we identified differentially expressed miRNAs in the saliva of AD subjects by miRNA-seq. Three AD-associated miRNAs (miR-10a-5p, miR-1290 and miR-4488) were cross-validated in both AAs and EAs. In addition, four miRNAs (miR-451a, miR-100-5p, miR-3613-5p and miR-7704) were specific for AA AD subjects and two miRNAs (miR-126-3p and miR-1273h-5p) were specific for EA AD subjects (Table 2). The identification of population-specific miRNAs for AD is not surprising. It is known that gene expression can be population-specific [37], and subjects from different ethnic groups can have different vulnerabilities to AD [38]. Population-specific miRNAs for AD may confer susceptibility of subjects from a specific ethnic group to AD. Because the same set of genes may participate in reward pathways and addiction in subjects from different populations, a set of miRNAs that regulate the expression of these genes may show differential expression in AD subjects from different populations. Our miRNA differential expression analysis considered only common miRNAs (i.e., miRNAs with CPM no less than 100 in at least half of the subjects), even though miRNA-seq can detect miRNAs expressed at a level of a single copy. Because miRNAs present at low levels cannot be accurately quantified and low-expression miRNAs may be indistinguishable from sampling noise [39], low-expression miRNAs are inadequate as biomarkers for disease prediction. Therefore, we excluded low-expression miRNAs from our analyses in order to maximize the sensitivity of detecting differentially expressed miRNAs.

Second, our study demonstrated that salivary miRNAs could be used as biomarkers for AD prediction. A cluster of five miRNAs could achieve a prediction accuracy of over 70% for AD (Table 3). Although certain liver enzymes and alcohol metabolites have been evaluated as AD predictors, their limited specificity precludes them as biomarkers for the diagnosis of AD [1,4]. Specific and reliable biomarkers need to be developed to supplement or replace biochemical measurements in order to predict AD with high confidence. To our knowledge, no other published studies have explored the potential use of salivary miRNAs as biomarkers of AD. miRNAs have been incorporated in the diagnosis of several other diseases particularly cancers [40,41]. There is evidence that miRNA expression profiles can more accurately cluster poorly differentiated tumors than mRNA profiles [42]. In contrast to the relatively small differences in mRNA expression levels between cancer and normal cells, the expression levels of miRNAs can exhibit fold changes of tens to hundreds [43]. In the present study, we found that a cluster of salivary

miRNAs predicted AD with an accuracy of 79.1% in AAs and 72.2% in EAs (Table 3). In comparison to the predictive accuracy of miRNAs in physical diseases such as cancers (97.6% for lung cancer, 97.8% for hepatocellular carcinoma, and 95.0% for bladder carcinoma in a study with 41 lung cancer samples, 47 hepatocellular carcinoma samples and 20 bladder cancer samples as well as adjacent or normal tissues as controls) [44], the miRNA prediction accuracy of AD was relatively low. This may be due to the heterogeneity of AD, which is a complex genetic disorder in which there are substantial gene–environment interactions. In contrast, cancer biomarkers often are derived from somatic mutations [45], which are often more readily assessed and show greater variation than germline mutations such as those contributing to complex diseases like AD.

Third, our findings showed the utility of a machine learning approach in the development of an effective prediction system for AD. The traditional multivariate logistic regression model is commonly used to predict the outcome of a categorical dependent variable (e.g., a disease phenotype) from a set of predictors or independent variables. However, the success of logistic regression model (either forward or backward stepwise) for disease prediction depends on (1) choosing the correct predictor variables, (2) avoiding inclusion of highly correlated variables which can reduce model efficiency, and (3) not including too many predictor variables which can lead to model over-fitting. The major difference between the traditional approach and the machine learning approach for disease prediction is the number of predictor variables that can be considered initially. Our application of the RF-based machine learning prediction algorithm considered 399 common salivary miRNAs from which it extracted a small number of miRNAs for inclusion in the highest performing predictive model. Of interest, the prediction accuracy was not improved when using the top ten versus the top five miRNAs based on the Gini scores, suggesting the number of potentially useful miRNAs in AD prediction is limited. However, the prediction accuracy was reduced when using too few miRNAs (e.g., only the top three miRNAs based on the Gini scores) (Table 3). Including too many predictor variables can dilute the true association and lead to a large standard error with a wide and imprecise confidence interval. Conversely, including too few predictor variables can lower the predictive power of the model.

Fourth, our machine learning prediction analysis results indicated that it was more favorable to use Gini score-ranked top miRNAs than differentially expressed miRNAs in AD prediction. When the train/test ratio was 80/20 and differentially expressed miRNAs ($p < 0.05$ and $\log_{2}FC > 1$) were included in the prediction analysis, the prediction accuracy was 63.6% in AAs and 56.2% in EAs (Supplementary Table 2). This suggests that a cluster of miRNAs generated by machine learning could be more powerful in predicting disease status than using differentially expressed miRNAs identified by statistical analysis, although the top miRNAs (e.g., miR-4488) generated by machine learning were also differentially expressed in AD subjects. We also noticed that the prediction accuracy was improved when the train/test ratio was increased (Table 3 & Supplementary Table 2). Previous studies have demonstrated that the training sample size is critical for training good classifiers [46]. In other words, more training data decrease the variance of the model, making it a more accurate general representation, and thus decreasing model overfitting.

Fifth, our study suggested that the integration of both machine learning and miRNA-seq approaches could yield a more precise prediction of AD. miRNA-seq is a type of RNA sequencing method (RNA-seq) for use in miRNA transcriptome profiling using a next-generation sequencing platform. It has higher detection sensitivity and specificity than gene expression microarray technology [47]. miRNA-seq can distinguish miRNAs with similar sequences and thus detect miRNA isoforms and novel miRNAs. When profiling miRNA transcriptomes in the saliva of human subjects by miRNA-seq, there is a concern that oral bacterial small RNAs can contaminate the sample. To avoid this, we aligned the sequences of 2588 miRNAs detected in the saliva of our AA and EA subjects against bacterial RNA sequences from the Human Oral Microbiome Database (www.homd.org). None of these salivary miRNA sequence reads were mapped to bacterial RNA sequences. Thus, miRNA-seq plus optimal sequence mapping algorithms facilitate accurate detection of human salivary miRNAs. The development of the disease prediction system using miRNA transcriptome profiling and machine learning is a novel approach with important potential clinical applications.

Conclusion & future perspective

Taken together, we described a method using miRNA-seq and RF-based machine learning to identify a cluster of salivary miRNAs as AD classifiers or predictors. We expect that this type of prediction system could be extended to other diseases based on miRNA expression changes that may also be present in the saliva of patients affected with them. To increase disease prediction accuracy, one could expand the machine learning component to include

additional types of biomarkers such as genetic variants and other epigenetic markers (DNA methylation and histone protein modifications), which regulate gene expression at the transcriptional level. This comprehensive approach could improve the diagnosis and treatment of a variety of diseases.

Summary points

- The current diagnosis of alcohol dependence (AD) depends primarily on self-reported symptoms. These may be augmented by alcohol-related biomarkers that reflect alcohol drinking patterns. The present study aimed to investigate whether salivary miRNAs are easily accessible AD biomarkers.
- We performed the first exploratory study on differential expression of salivary miRNAs in patients with AD in two populations (African-Americans [AAs] and European-Americans [EAs]) and using salivary miRNAs as easily accessible biomarkers to predict the presence of AD by the Random Forest machine learning method.
- Seven miRNAs were differentially expressed ($p < 0.05$ and $> \text{two-fold change}$) in AA AD patients, and five miRNAs were differentially expressed ($p < 0.05$ and $> \text{two-fold change}$) in EA AD patients. Three of the above salivary miRNAs (miR-10a-5p, miR-1290 and miR-4488) showed the same direction of expression changes in both AA and EA AD subjects. These three miRNAs potentially target genes involved in DNA binding, alternative splicing or calcium-dependent cell–cell adhesion.
- When the train/test sample ratio of 80/20 and the top five mRNAs (ranked by Gini index or their importance to AD prediction) were integrated in Random Forest prediction analyses, the AD prediction accuracy was 79.1 and 72.2% in AAs and EAs, respectively.
- These findings reflect miRNA expression changes in the saliva of AD patients, providing evidence that salivary miRNAs are potential biomarkers that identify the presence of AD. The findings need to be validated in a larger sample.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.futuremedicine.com/doi/full/10.2217/epi-2018-0177

Author contributions

H Zhang and J Gelernter were responsible for the study concept and design. H Zhang and YZ Nunez performed the experiments. AJ Rosato, X Chen and H Zhang analyzed the data and wrote the manuscript. LA Farrer, HR Kranzler, YZ Nunez, DC Henderso and J Gelernter provided helpful comments on the manuscript. All authors critically reviewed content and approved the final version for publication.

Acknowledgments

We thank all subjects for participating in this study. We are grateful to M Cusumano, L Frederick and R Gordon at the APT foundation, CT, USA for collecting saliva samples and managing the clinical data for this study.

Financial & competing interests disclosure

This work was supported by grants (R21AA023068 and R01AA025080) from the National Institute on Alcohol Abuse and Alcoholism (NIAAA). The findings achieved herein are solely the responsibility of the authors. HR Kranzler has been an advisory board member, consultant or continuing medical education speaker for Alkermes, Indivior and Lundbeck. He is a member of the American Society of Clinical Psychopharmacology's Alcohol Clinical Trials Initiative (ACTIVE), which was supported in the last 3 years by AbbVie, Alkermes, Amygdala Neurosciences, Arbor Pharmaceuticals, Ethypharm, Indivior, Lilly, Lundbeck, Otsuka and Pfizer. HR Kranzler and J Gelernter are named as inventors on PCT patent application #15/878,640 entitled: "Genotype-guided dosing of opioid agonists," filed 24 January 2018. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Ethical conduct of research

All participants provided written informed consent approved by the Yale University Human Investigation Committee (HIC) (Protocol #: 0102012183).

References

Papers of special note have been highlighted as: • of interest; •• of considerable interest

1. Hietala J, Koivisto H, Anttila P, Niemela O. Comparison of the combined marker GGT-CDT and the conventional laboratory markers of alcohol abuse in heavy drinkers, moderate drinkers and abstainers. *Alcohol Alcohol.* 41(5), 528–533 (2006).
2. Seppa K, Laippala P, Saarni M. Macrocytosis as a consequence of alcohol abuse among patients in general practice. *Alcohol. Clin. Exp. Res.* 15(5), 871–876 (1991).
3. Davenport J. Macrocytic anemia. *Am. Fam. Physician* 53(1), 155–162 (1996).
4. Peterson K. Biomarkers for alcohol use and abuse – a summary. *Alcohol Res. Health* 28(1), 30–37 (2004).
5. Justice AC, McGinnis KA, Tate JP *et al.* Validating harmful alcohol use as a phenotype for genetic discovery using phosphatidylethanol and a polymorphism in *ADH1B*. *Alcohol. Clin. Exp. Res.* 41(5), 998–1003 (2017).
6. Quillen EE, Chen XD, Almasy L *et al.* *ALDH2* is associated to alcohol dependence and is the major genetic determinant of “daily maximum drinks” in a GWAS study of an isolated rural Chinese sample. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 165B(2), 103–110 (2014).
7. Gelernter J, Kranzler HR, Sherva R *et al.* Genome-wide association study of alcohol dependence: significant findings in African- and European-Americans including novel risk loci. *Mol. Psychiatry* 19(1), 41–49 (2014).
8. Edenberg HJ. Common and rare variants in alcohol dependence. *Biol. Psychiatry* 70(6), 498–499 (2011).
9. Filipowicz W, Bhattacharyya SN, Sonenberg N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat. Rev. Genet.* 9(2), 102–114 (2008).
10. Liu X, Fortin K, Mourelatos Z. MicroRNAs: biogenesis and molecular functions. *Brain Pathol.* 18(1), 113–121 (2008).
11. Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell* 136(2), 215–233 (2009).
- **Describes the important role of miRNAs in regulating mRNA expression.**
12. Krol J, Loedige I, Filipowicz W. The widespread regulation of microRNA biogenesis, function and decay. *Nat. Rev. Genet.* 11(9), 597–610 (2010).
13. Nebbioso A, Tambaro FP, Dell'Aversana C, Altucci L. Cancer epigenetics: moving forward. *PLoS Genet.* 14(6), e1007362 (2018).
14. Tapocik JD, Solomon M, Flanigan M *et al.* Coordinated dysregulation of mRNAs and microRNAs in the rat medial prefrontal cortex following a history of alcohol dependence. *Pharmacogenomics J.* 13(3), 286–296 (2013).
15. Lewohl JM, Nunez YO, Dodd PR, Tiwari GR, Harris RA, Mayfield RD. Up-regulation of microRNAs in brain of human alcoholics. *Alcohol. Clin. Exp. Res.* 35(11), 1928–1937 (2011).
- **Provides evidence that miRNAs are differentially expressed in alcoholic subjects.**
16. Wang F, Gelernter J, Zhang H. Differential expression of miR-130a in postmortem prefrontal cortex of subjects with alcohol use disorders. *J. Addict. Res. Ther.* 4(155), pii 18179 (2013).
17. Hunter MP, Ismail N, Zhang X *et al.* Detection of microRNA expression in human peripheral blood microvesicles. *PLoS ONE* 3(11), e3694 (2008).
18. Zernecke A, Bidzhekov K, Noels H *et al.* Delivery of microRNA-126 by apoptotic bodies induces CXCL12-dependent vascular protection. *Sci. Signal.* 2(100), ra81 (2009).
19. Arroyo JD, Chevillet JR, Kroh EM *et al.* Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma. *Proc. Natl Acad. Sci. USA* 108(12), 5003–5008 (2011).
- **An important study regarding the stability of circulating miRNAs.**
20. Vickers KC, Palmisano BT, Shoucri BM, Shamburek RD, Remaley AT. MicroRNAs are transported in plasma and delivered to recipient cells by high-density lipoproteins. *Nat. Cell Biol.* 13(4), 423–433 (2011).
21. Bahn JH, Zhang Q, Li F *et al.* The landscape of microRNA, Piwi-interacting RNA, and circular RNA in human saliva. *Clin. Chem.* 61(1), 221–230 (2015).
22. Li F, Yoshizawa JM, Kim KM *et al.* Discovery and validation of salivary extracellular RNA biomarkers for noninvasive detection of gastric cancer. *Clin. Chem.* 64(10), 1513–1521 (2018).
23. Pierucci-Lagha A, Gelernter J, Feinn R *et al.* Diagnostic reliability of the semi-structured assessment for drug dependence and alcoholism (SSADDA). *Drug Alcohol Depend.* 80(3), 303–312 (2005).
24. Pierucci-Lagha A, Gelernter J, Chan G *et al.* Reliability of DSM-IV diagnostic criteria using the semi-structured assessment for drug dependence and alcoholism (SSADDA). *Drug Alcohol Depend.* 91(1), 85–90 (2007).
25. Friedlander MR, Chen W, Adamidi C *et al.* Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotech.* 26(4), 407–415 (2008).
- **Describes an important biostatistical program for miRNA sequencing data analysis.**
26. Ambros V, Bartel B, Bartel DP *et al.* A uniform system for microRNA annotation. *RNA* 9(3), 277–279 (2003).

27. Robinson BG, Mealy K, Wilmore DW, Majzoub JA. The effect of insulin-induced hypoglycemia on gene expression in the hypothalamic-pituitary-adrenal axis of the rat. *Endocrinology* 130(2), 920–925 (1992).
28. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* 40(10), 4288–4297 (2012).
29. Dweep H, Sticht C, Pandey P, Gretz N. miRWalk–database: prediction of possible miRNA binding sites by “walking” the genes of three genomes. *J. Biomed. Inform.* 44(5), 839–847 (2011).
- **An important miRNA database.**
30. Dweep H, Gretz N. miRWalk2.0: a comprehensive atlas of microRNA-target interactions. *Nat. Methods* 12(8), 697 (2015).
31. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120(1), 15–20 (2005).
32. Wong N, Wang X. miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic Acids Res.* 43(Database issue), D146–D152 (2015).
33. Chou CH, Shrestha S, Yang CD *et al.* miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.* 46(D1), D296–D302 (2018).
34. Huang Da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4(1), 44–57 (2009).
35. Breiman L. Random forests. *Machine Learning* 45, 5–32 (2001).
- **An important review of the Random Forest algorithm.**
36. Menze BH, Kelm BM, Masuch R *et al.* A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinform.* 10(1), 213 (2009).
- **The concept of Gini index.**
37. Wang L, Rishishwar L, Marino-Ramirez L, Jordan IK. Human population-specific gene expression and transcriptional network modification with polymorphic transposable elements. *Nucleic Acids Res.* 45(5), 2318–2328 (2017).
38. Wickramasinghe SN, Corridan B, Izaguirre J, Hasan R, Marjot DH. Ethnic differences in the biological consequences of alcohol abuse: a comparison between south Asian and European males. *Alcohol Alcohol.* 30(5), 675–680 (1995).
39. Bourgon R, Gentleman R, Huber W. Independent filtering increases detection power for high-throughput experiments. *Proc. Natl Acad. Sci. USA* 107(21), 9546–9551 (2010).
40. Park NJ, Zhou H, Elashoff D *et al.* Salivary microRNA: discovery, characterization, and clinical utility for oral cancer detection. *Clin. Cancer Res.* 15(17), 5473–5477 (2009).
41. Matse JH, Yoshizawa J, Wang X *et al.* Discovery and prevalidation of salivary extracellular microRNA biomarkers panel for the noninvasive detection of benign and malignant parotid gland tumors. *Clin. Cancer Res.* 19(11), 3032–3038 (2013).
42. Lu J, Getz G, Miska EA *et al.* MicroRNA expression profiles classify human cancers. *Nature* 435(7043), 834–838 (2005).
43. Jiang J, Lee EJ, Gusev Y, Schmittgen TD. Real-time expression profiling of microRNA precursors in human cancer cell lines. *Nucleic Acids Res.* 33(17), 5394–5403 (2005).
44. Salim A, Amjesh R, Chandra SS. An approach to forecast human cancer by profiling microRNA expressions from NGS data. *BMC Cancer* 17(1), 77 (2017).
45. Kranzler HR, Smith RV, Schnoll R, Moustafa A, Greenstreet-Akman E. Precision medicine and pharmacogenetics: what does oncology have that addiction medicine does not? *Addiction* 112(12), 2086–2094 (2017).
46. Beleites C, Neugebauer U, Bocklitz T, Krafft C, Popp J. Sample size planning for classification models. *Anal. Chim. Acta* 760, 25–33 (2013).
47. Tam S, De Borja R, Tsao MS, McPherson JD. Robust global microRNA expression profiling using next-generation sequencing technologies. *Lab. Invest.* 94(3), 350–358 (2014).
- **Describes the next generation sequencing technology for profiling genome-wide miRNA expression.**