AWARD NUMBER: W81XWH-12-1-0330

TITLE: Salivary Proteomic and microRNA Biomarkers Development for Lung Cancer Detection

PRINCIPAL INVESTIGATOR: David Wong, DMD, DMSc

CONTRACTING ORGANIZATION: The Regents of the University of California
Los Angeles, CA 90095

REPORT DATE: August 2015

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.
**REPORT DOCUMENTATION PAGE**

**Form Approved**

**OMB No. 0704-0188**

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. **REPORT DATE**
   August 2015

2. **REPORT TYPE**
   Annual

3. **DATES COVERED**
   1 Aug 2014 – 31 Jul 2015

4. **TITLE AND SUBTITLE**
   Salivary Proteomic and microRNA Biomarkers Development for Lung Cancer Detection.

5a. **CONTRACT NUMBER**
   Detection.

5b. **GRANT NUMBER**
   W81XWH-12-1-0330

5c. **PROGRAM ELEMENT NUMBER**

5d. **PROJECT NUMBER**

5e. **TASK NUMBER**

5f. **WORK UNIT NUMBER**

6. **AUTHOR(S)**
   David Wong, DMD, DMSc

   E-Mail: dtww@ucla.edu

7. **PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
   University of California, Los Angeles
   11000 Kinross Ave., Suite 211
   Los Angeles, CA 90095-2000

8. **PERFORMING ORGANIZATION REPORT NUMBER**

9. **SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
   U.S. Army Medical Research and Materiel Command
   Fort Detrick, Maryland 21702-5012

10. **SPONSOR/MONITOR’S ACRONYM(S)**

11. **SPONSOR/MONITOR’S REPORT NUMBER(S)**

12. **DISTRIBUTION / AVAILABILITY STATEMENT**
    Approved for Public Release; Distribution Unlimited

13. **SUPPLEMENTARY NOTES**

14. **ABSTRACT**
    This is a lung cancer biomarker development project to test the hypothesis that there are discriminatory miRNA and proteomic biomarkers in saliva that can detect lung cancer with the aim to reduce the number unnecessary diagnostic workups (bronchoscopy) in patients with suspicious chest symptoms. Preliminary data is in place to support that our salivary biomarker technologies can discover and validate lung cancer biomarkers in saliva. The major goal is to perform a properly powered biomarker discovery and definitive validation of salivary proteomic and miRNA biomarkers for detection of lung cancer based on PRoBE design principles (prospective-specimen-collection and retrospective-blinded-evaluation). The outcome of this three-year proposal will be a panel of definitively validated non-invasive saliva-based proteomic and micro-RNA biomarkers for detection of lung cancer.

15. **SUBJECT TERMS**
    Lung cancer, Early detection, Saliva, Biomarkers

16. **SECURITY CLASSIFICATION OF:**
    a. **REPORT**
       Unclassified
    b. **ABSTRACT**
       Unclassified
    c. **THIS PAGE**
       Unclassified

17. **LIMITATION OF ABSTRACT**
    Unclassified

18. **NUMBER OF PAGES**
    20

19a. **NAME OF RESPONSIBLE PERSON**
    USAMRMC

19b. **TELEPHONE NUMBER (include area code)**

**Standard Form 298 (Rev. 8-98)**
Prescribed by ANSI Std. Z39.18
Table of Contents

1. Introduction ................................................................. 2
2. Keywords .................................................................. 3
3. Overall Project Summary .............................................. 3
4. Key Research Accomplishments ................................. 6
5. Conclusion .................................................................. 6
7. Inventions, Patents and Licenses ..................................... 6
8. Reportable Outcomes ................................................... 6
9. Other Achievements ..................................................... 7
10. References ................................................................. 7
11. Appendices ............................................................... 7
1. Introduction

This is a lung cancer biomarker development project to test the hypothesis that there are discriminatory biomarkers in saliva that can detect lung cancer with the aim to reduce the number unnecessary diagnostic workups (bronchoscopy) in patients with suspicious chest symptoms. The major goal is to perform a properly powered biomarker discovery and validation of salivary miRNA and proteomic biomarkers for detection of lung cancer based on PRoBE design principles (prospective-specimen-collection and retrospective-blinded-evaluation). The outcome of this project will be a panel of validated non-invasive saliva-based biomarkers for detection of lung cancer.
2. Keywords: Lung cancer, Early detection, Saliva, Biomarkers

3. Overall project summary

This is the third year of this DoD CDMRP Lung Cancer Research Program (LCRP) Investigator-initiated Translational Research Award project titled “Salivary Proteomic and microRNA Biomarkers Development for Lung Cancer Detection”.

The first year of this lung cancer biomarker development project was spent in the obtainment of regulatory (IRB) approvals from the two performance sites of the project, University of California Los Angeles and the Greater Los Angeles VA (GLA-VA), as well as with the Human Research Protection Office (HRPO) at the US Army Medical Research and Materiel Command (USAMRMC). These lengthy regulatory procedures caused a year of setbacks delaying the initiation our translational research study to develop salivary biomarkers for lung cancer detection. We have since obtained full approval of the informed consent changes and the HRPO of USAMRMC have approved the use of human subjects of this lung cancer biomarker development study. On November 15, 2013 we obtained approval from Dr. Sheilah Rowe, the Scientific Officer, that our project was delayed for one year and consider the need of an extension of the performance period.

The second year of the project was focused on Aim 1 to continue accrual of lung cancer and control subjects. We also began the salivary biomarker discovery phase, Aim 2 of the project. As reported in the last progress report, significant efforts were made to integrate the emerging technology of RNA-Seq in saliva to discovery miRNA in saliva of lung cancer patients. The efforts to optimize the RNA-Seq technology to saliva for extracellular RNA discovery will allow us to obtain un-parallel detailed information of known and novel miRNA in saliva that can be developed for lung cancer non-invasive biomarkers. 30 lung cancer and 30 non-lung cancer control saliva RNA-libraries were constructed for biomarker discovery. A high impact paper was published based on these novel and impactful efforts to decipher the salivary extracellular transcriptome and the article was featured in a special issue of Clinical Chemistry titled “Molecular Diagnostics: A Revolution in Progress” (1). Support by the CDMRP/DoD was cited.

This progress report contains the research accomplishment of the Specific Aims 1 and 2 as contained in the original Statement of Work.

**Aim 1: Accrual of Lung Cancer and Control Subjects- Based on PROBE Design**

**Milestone 1:** Accrual of 1560 saliva samples from patients with suspicious chest symptoms. Based on current practice, we anticipate 624 lung cancers and 936 are cancer free patients at the Greater Los Angeles VA hospital (GLA-VA) procured based on the PROBE-study design.

As of August 28, 2015 we have screened 2486 patients with chest symptoms at the GLA-VA (159% of the targeted enrollment of 1560). Of these 215 subjects were endoscoped and 93 were confirmed with diagnosed of lung cancer. Our original study design anticipated 624 lung cancer cases by the end of year 01 with nodular sizes on CT > 1cm. The lung cancer yield turned out to be 32 cases with nodular size >1 cm. This is lower than anticipated and necessitated us to modify the study design for the biomarker discovery Aim.

We used 30 lung cancer and 30 non-cancer saliva samples for the biomarker discovery. In addition to the cancer status of these lesions, we have also correlated the tumor size of these lesion based on their CT data. This inclusion is of clinical relevance and impact since the ability to develop salivary biomarkers that can predict cancer from non-cancer patients will be clinically impactful. By examining the plot of sample size against the proportion of genes exceeding the power threshold, we estimated that the sample size of 30 per group (cases and control) will
prove statistical power of at least 99% for 98% of the genes whose true effects exceed a fold-change of 2. Saliva from 30 lung cancer patients and 30 matched controls were used for the discovery studies. Controls were matched for gender, age, smoking history and ethnicity. This matching ensured a distributional match on potential confounders.

**Aim 2:** Salivary extracellular RNA (exRNA) Biomarkers Discovery, Statistical and Systems Approaches to Candidate Biomarkers Selection

**Milestone 2:** Optimized salivary biomarker discovery technologies and a systems approach will be used to identify candidate exRNA salivary biomarkers for lung cancer detection in a discovery cohort of 30 cases and 30 controls randomly selected from Aim 1. Salivary biomarker optimized data mining approach will be used to identify to candidate markers.

RNA-Seq is emerging technology to obtain the most detailed information of RNA in a biological sample. While we originally proposed to use the Taqman MicroRNA Array Card for saliva miRNA discovery, the significant advantages to use RNA-Seq for saliva miRNA discovery for known and novel miRNA is compelling. We published the first RNA-Seq study on salivary RNA using the SOLID™ system (2). In this project, we will use Illumina sequencing systems. We have generated data that support the quality, reproducibility and feasibility of our approach. We have compared multiple library generation methods, constructed different types of libraries to capture the whole spectrum of exRNA in saliva, evaluated the reproducibility of our methods, and obtained a preliminary landscape of relative and absolute concentration of various types of exRNAs in saliva (1).

A number of RNA-Seq library construction methods have been developed in the literature. We have evaluated the performance of alternative methods, we used multiple commercially available kits (NEB, Illumina, Clonetech and NuGen) targeting different types of RNA. A typical bioanalyzer profile of saliva exRNA is shown in Fig. 1. For each library, 500ng of total RNA was used as input. Importantly, predefined amount of synthetic spike-in RNAs were added into each RNA sample equivalently, which will serve as internal standards to evaluate library efficiency, reproducibility, to normalize data across different samples, and to calculate absolute RNA abundance. The synthetic RNAs were purchased from Exiqon and Life Technologies for small RNA-Seq and regular RNA-Seq, respectively. The synthetic RNA pool consists of many distinct RNA species (>40 for small RNA) to ensure abundance and sequence diversity. Since it is known that RNA from saliva is partially degraded with size between 20 and 200nt, we modified the library generation methods to exclude polyA selection and include a size-selection step favoring RNAs below 200nt. Depletion of ribosomal RNA was not carried out since it is known that saliva has relatively less rRNA compared to cellular RNA. Note that although the regular RNA-Seq spike-in RNAs were polyadenylated, the random priming method used in regular RNA-Seq still allows their usage as reference standards. Using these optimized steps for salivary exRNA, we constructed RNA-libraries of the 60 saliva samples (30 lung cancer, 30 controls). All samples were randomized to minimize batched effect for RNA-Sequencing.

All libraries were sequenced using Illumina HiSeq 2500 sequencers at the UCLA core facility. A total of 30-50 million single-end (50nt) reads were obtained for each library. We have developed a customized bioinformatic pipelines to identify different types of non-coding RNAs (ncRNA) present in these data sets originated from human, microbiome, plants or any other species (Fig. 2). Small RNA-Seq data were analyzed for miRNAs and other ncRNAs. Although
the small RNA libraries capitalize on the fact that canonical miRNAs have a 5'-phosphate and 3'-OH, other ncRNAs may be identified if their processing steps also lead to such footprints. Since the RNA-Seq libraries used random priming to generate cDNAs, they can theoretically capture all different types of RNAs in the selected size range (20-200nt) with adequate abundance. We have analyzed whether the RNA-Seq data sets may capture all long, small and circular ncRNAs. Mapping uniqueness was required for reads mapped to spike-in RNAs, known genes, IncRNAs and circular RNAs, but not for reads mapped to microbiome or 16S. Small RNA reads were not required to be unique either since small RNAs (miRNAs, piRNAs, etc.) may have multiple copies or similar family members in the human genome. All libraries yielded high quality reads, with an average of ~50% reads mapped to 16S and microbiome. To evaluate potential contamination by cellular RNA in our samples, we examined a number of genes (e.g., ESRP1/2, OVOL1/2, HBA1, APOC1 etc.) that are known to be highly specific to epithelial cells or leukocytes, the major types of cells in saliva. Most of these genes are not expressed based on the RNA-Seq data, supporting the effectiveness of our saliva SOP in removing cells.

Table 1: Salivary miRNA candidates for lung cancer detection

<table>
<thead>
<tr>
<th>miRNA</th>
<th>Control % Read Count &gt;10 Control median</th>
<th>Cancer % Read Count &gt;10 Cancer median</th>
<th>P.Wilcoxon R</th>
<th>AdjP.DESeq</th>
<th>P.DESeq</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR_4488</td>
<td>36</td>
<td>45</td>
<td>1</td>
<td>2.5</td>
<td>26.75</td>
</tr>
<tr>
<td>miR_3656</td>
<td>12</td>
<td>18</td>
<td>1</td>
<td>0.214503</td>
<td>0.611561</td>
</tr>
<tr>
<td>miR_10b_5p</td>
<td>16</td>
<td>45</td>
<td>2</td>
<td>6.5</td>
<td>14.49004</td>
</tr>
<tr>
<td>miR_181_3p</td>
<td>44</td>
<td>50</td>
<td>7</td>
<td>4.25</td>
<td>17</td>
</tr>
<tr>
<td>miR_210</td>
<td>92</td>
<td>77</td>
<td>95</td>
<td>80.2</td>
<td>287</td>
</tr>
<tr>
<td>miR_664a_5p</td>
<td>48</td>
<td>68</td>
<td>9</td>
<td>18</td>
<td>8.25</td>
</tr>
<tr>
<td>miR_1307_3p</td>
<td>96</td>
<td>95</td>
<td>126</td>
<td>66</td>
<td>0.023440</td>
</tr>
<tr>
<td>miR_342_5p</td>
<td>64</td>
<td>68</td>
<td>14</td>
<td>8</td>
<td>7.25</td>
</tr>
<tr>
<td>miR_177_3p</td>
<td>96</td>
<td>100</td>
<td>183.5</td>
<td>50</td>
<td>49.943</td>
</tr>
<tr>
<td>miR_451a</td>
<td>88</td>
<td>82</td>
<td>41</td>
<td>23</td>
<td>225.5</td>
</tr>
<tr>
<td>miR_125b_2_3p</td>
<td>36</td>
<td>59</td>
<td>6</td>
<td>13</td>
<td>4.5</td>
</tr>
<tr>
<td>miR_222_5p</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.375</td>
</tr>
<tr>
<td>miR_660_5p</td>
<td>76</td>
<td>86</td>
<td>53</td>
<td>18</td>
<td>110.75</td>
</tr>
<tr>
<td>miR_148a_5p</td>
<td>100</td>
<td>100</td>
<td>25175</td>
<td>8172</td>
<td>51507</td>
</tr>
<tr>
<td>miR_183_5p</td>
<td>88</td>
<td>91</td>
<td>201</td>
<td>118</td>
<td>276</td>
</tr>
<tr>
<td>miR_100_5p</td>
<td>96</td>
<td>91</td>
<td>111</td>
<td>73</td>
<td>235</td>
</tr>
<tr>
<td>miR_211_5p</td>
<td>44</td>
<td>50</td>
<td>6</td>
<td>45</td>
<td>10</td>
</tr>
<tr>
<td>miR_184</td>
<td>88</td>
<td>95</td>
<td>62</td>
<td>28</td>
<td>120</td>
</tr>
</tbody>
</table>

Sequencing data was preprocessed to generate counts for small exRNA species. A few different normalization methods (raw counts, RPM counts, Quantile normalization and DESeq normalization) were used to generate analytic datasets. We filtered the small exRNA species by requiring counts ≥ 5 in at least 20% of samples for inclusion in the statistical analysis. The Wilcoxon rank sum test and the Negative Binomial DE-Seq test were used to compare cancer vs. non-cancer for each species. Initial small exRNA candidates were selected based statistical significance and additional criteria of biological relevance. The first set of candidates was assayed with ddPCR to evaluate optimal data analysis strategies for the candidate selection process. We have found that Quantile normalization of the sequencing data did not correlate well with ddPCR concentrations. Normalizing the ddPCR targets to reference gene also negatively affected the concordance between techniques. We observed correlations ~(0.70-
0.90) between untransformed ddPCR concentrations and number of RNA-seq reads (either raw or RPM).

Based on these findings we revised the rationale for miRNA candidates' selection for verification. Table 1 showed the identification of 18 salivary (miRNA 9, miR 4488, miR 3656, miR 10b 5p, miR 381 3p, miR 210, miR 664a 5p, miR 1307 3p, miR 342 5p, miR 27b 3p, miR 451a, miR 125b 2 3p, miR 122 5p, miR 660 5p, miR 148a 3p, miR 183 5p, miR 100 5p, miR 2115 5p, miR 184) that are significantly different between the lung cancer and non-lung cancer controls. These 18 miRNA will be subjected for validation in Aim 3.

4. Key Research Accomplishments

• Accrual of 2486 patients with chest symptoms (159% of the targeted enrollment of 1560)
• Biomarker discovery cohort of 30 lung cancer patients with lung nodules on CT> 1cm and 30 non-lung cancer matched controls fully adhering to prospective-specimen-collection and retrospective-blinded-evaluation (PRoBE) design
• RNA library construction of the 30 lung cancer and 30 non lung cancer controls
• RNA-Sequencing of 30 lung cancer and 30 non lung cancer controls
• Data analysis and select 18 salivary miRNAs that are significantly altered in lung cancer from non lung cancer controls

5. Conclusion

During the third year of the project, scientific progress has been sound. Targeted enrollment has been attained despite the lung cancer cases fulfilling the inclusion criteria was less than expected. We have successfully performed, using optimized salivary RNA library construction and RNA-sequencing technologies for the biomarker discovery cohort of 30/30. Data analysis of the RNA-seq data revealed 18 salivary miRNAs are significantly altered in lung cancer from non lung cancer subjects.

So What Section: The frontier technology of RNA-Seq for salivary miRNA development have led to the discovery of salivary biomarkers that have discriminatory power to detect lung cancer in patients with symptomatic chest symptoms and nodules of > 1cm.

6. Publications, abstracts, and presentations


7. Invention, Patents, and Licenses:

None to report.

8. Reportable Outcomes

Not applicable.
9. Other achievements

Not applicable.

10. References


11. Appendices

The Landscape of MicroRNA, Piwi-Interacting RNA, and Circular RNA in Human Saliva

Jae Hoon Bahn,1,2† Qing Zhang,1,2† Feng Li,3 Tak-Ming Chan,1,2 Xianzhi Lin,1,2 Yong Kim,3,4,5 David T.W. Wong,2,3,4,6,7* and Xinshu Xiao1,2,4,6*

BACKGROUND: Extracellular RNAs (exRNAs) in human body fluids are emerging as effective biomarkers for detection of diseases. Saliva, as the most accessible and noninvasive body fluid, has been shown to harbor exRNA biomarkers for several human diseases. However, the entire spectrum of exRNA from saliva has not been fully characterized.

METHODS: Using high-throughput RNA sequencing (RNA-Seq), we conducted an in-depth bioinformatic analysis of noncoding RNAs (ncRNAs) in human cell-free saliva (CFS) from healthy individuals, with a focus on microRNAs (miRNAs), piwi-interacting RNAs (piRNAs), and circular RNAs (circRNAs).

RESULTS: Our data demonstrated robust reproducibility of miRNA and piRNA profiles across individuals. Furthermore, individual variability of these salivary RNA species was highly similar to those in other body fluids or cellular samples, despite the direct exposure of saliva to environmental impacts. By comparative analysis of >90 RNA-Seq datasets of different origins, we observed that piRNAs were surprisingly abundant in CFS compared with other body fluid or intracellular samples, with expression levels in CFS comparable to those found in embryonic stem cells and skin cells. Conversely, miRNA expression profiles in CFS were highly similar to those in serum and cerebrospinal fluid. Using a customized bioinformatics method, we identified >400 circRNAs in CFS. These data represent the first global characterization and experimental validation of circRNAs in any type of extracellular body fluid.

CONCLUSIONS: Our study provides a comprehensive landscape of ncRNA species in human saliva that will facilitate further biomarker discoveries and lay a foundation for future studies related to ncRNAs in human saliva.

© 2014 American Association for Clinical Chemistry
been shown to act as miRNA sponges, thus playing a role in mediating miRNA targeting (20). It is expected that additional functions of circRNAs may be described soon (19). The stable nature of circRNAs makes these moieties intriguing candidates as functional molecules in circulating body fluid.

We performed a comprehensive analysis of extracellular noncoding RNAs (ncRNAs) in cell-free saliva (CFS) by next-generation sequencing. In addition, we carried out a genome-wide analysis of possible existence of circular RNAs in CFS with RNA-Seq. To our best knowledge, this is the first report and validation of the existence of circRNAs in any body fluid. Our findings credential the presence of salivary ncRNA and, importantly, pave ways for further functional, biological, and biomarker discoveries related to ncRNAs in human saliva.

Materials and Methods

SALIVA COLLECTION AND PROCESSING AND RNA ISOLATION

Unstimulated saliva samples were obtained from healthy volunteers in accordance with a protocol approved by the University of California–Los Angeles (UCLA) Institutional Review Board as described previously (21). More details are included in Supplemental Methods, which accompanies the online version of this article at http://www.clinchem.org/content/vol61/issue1.

CONSTRUCTION OF SMALL RNA-SEQ LIBRARIES

We isolated total RNA directly from CFS as described above. Spike-in RNAs (Exiqon) were added to the total RNA samples (1 reaction volume per microgram RNA) before library construction as internal controls. With the spiked total RNA samples, we prepared small RNA-Seq libraries with the NEBNext Small RNA library Prep kit (NEB). The final libraries were purified with 6% PAGE gel.

CONSTRUCTION OF circRNA-SEQ LIBRARIES

To obtain enriched circular RNAs from the total RNA samples, we used 3 U/μg RNase R (Epigence) to treat the total RNA (from CFS directly) for 20 min at 37 °C. Subsequently, RNA was extracted with acid phenol/chloroform (pH 4.5). We prepared sequencing libraries with the NEBNext Ultra directional RNA library Prep kit (NEB) followed by AMPure XP Beads size selection (Beckman Coulter).

CFS SMALL RNA-SEQ DATA ANALYSIS

Small RNA-Seq reads were first processed to remove adapter sequences and low quality reads. The reads were then aligned to the human genome with Bowtie (22) allowing at most 1 mismatch. We parsed the mapping results to identify reads mapped to miRNAs (miRBase, release 19), piwi-interacting RNAs (piRNAs) (piRNABank, November 2013), and other known small RNAs (RFam, version 11.0). We also mapped reads to the Human Oral Microbiome Databases (23) to eliminate those that possibly originated from microbial species. For human miRNAs, only uniquely mapped reads were retained. Uniqueness was not required for reads mapped to piRNAs or other small RNAs owing to their repetitive nature and/or presence of multiple copies in the genome. In parallel, reads were also aligned to the spike-in controls, allowing no mismatches. The number of reads mapped to each miRNA was normalized with the spike-in controls and total number of mapped reads in each library.

Detailed bioinformatic methods, exosome isolation, and experimental validation procedures are described in online Supplemental Methods.

Results and Discussion

SMALL RNA SEQUENCING OF CFS

We used our widely adopted protocol to isolate CFS from fresh saliva samples (see online Supplemental Methods) (9). The protocol previously was shown to effectively remove cells as evaluated by exclusion of cellular genomic DNA and cell counting (9). In addition, we examined several genes [e.g., ESRP1/ESRP2 (epithelial splicing regulatory protein 1 and 2)],9 OVOL1/2 (ovo-like zinc finger 1 and 2), HBA1 (hemoglobin, α1), APOC1 (apolipoprotein C-1) that are known to be highly specific to epithelial cells or leukocytes, the major types of cells in saliva. Most of these genes were not expressed (on the basis of RNA-Seq data collected for a separate study, data not shown), supporting the effectiveness of our CFS protocol in removing cells. We obtained a total of 165 million reads from 8 CFS small RNA sequencing libraries (see online Supplemental Table 1). Total RNA was isolated directly from the CFS, and synthetic small RNAs were added into the total RNA samples before library construction to serve as spike-in controls. As shown in online Supplemental Fig. 1, expression levels of spike-in controls were highly correlated between samples, supporting the technical consistency of our data. After adapter removal, the reads showed a length distribution that peaked at 22 nt (Fig. 1A), consistent with the expected length of miRNAs. Interestingly, we also observed a second peak at 29 nt that may correspond to piRNAs. The most
Fig. 1. miRNA expression in CFS.

(A), Example length distribution of a small RNA sequencing library from CFS. Library adapters have been trimmed. The read lengths of the major peak (22 nt) and minor peak (29 nt) are illustrated, which correspond to the known lengths of miRNAs and piRNAs, respectively. (B), Scatter plot of miRNA expression (log2 RPM) across 2 individuals. Pearson correlation coefficient is shown. (C), Experimental validation of miRNA expression in exosome fraction (E) and exosome-free fraction (NE) by use of ddPCR. For each miRNA, the ddPCR fluorescence intensity is shown for E and NE samples in each individual. Negative control (no template) was run together with the actual samples in the same batch of experiments, and fluorescent signal was barely detected. (D), Histogram of ISI for all expressed miRNAs calculated with the 6 independent saliva samples (biological replicates not included). The distributions of ISI values of other public data sets are also shown for comparison.
abundant types of small RNAs in our data included human miRNAs (6.0% of reads on average), piRNAs (7.5% of reads), and snoRNAs (0.02% of reads) (see online Supplemental Table 1). In addition, 58.8% of reads corresponded to microbial RNA sequences, reflecting the enriched presence of microorganisms in saliva (8). Our results suggest that the small RNA sequencing experiment can capture a wide spectrum of noncoding exRNAs in human saliva.

miRNAs ARE ABUNDANT IN HUMAN CFS
In each saliva sample, a total of 127–418 miRNAs were detected, with an expression level of \( \geq 1 \) reads per million mapped reads (RPM) (see online Supplemental Table 2). The most abundant miRNA was miR-223–3p that had an average expression level of 19442 RPM across all samples. miRNA expression levels were highly correlated across biological replicates \( (r \geq 0.977) \) (see online Supplemental Fig. 2), again supporting the reproducibility of our data. Importantly, different human subjects also demonstrated highly correlated miRNA expression levels (Fig. 1B; online Supplemental Fig. 3). For example, among the top 10 highly expressed miRNAs in each sample, 8 were shared by at least 4 of the 6 nonreplicated samples.

Previous reports suggested that the majority of miRNAs in saliva were concentrated in exosomes (24). We tested the presence of miRNAs in exosome and nonexosome fractions of saliva in a subset of samples (Fig. 1C; online Supplemental Fig. 4). Two miRNAs (miR-223–3p and miR-148a-3p) detected in RNA-Seq data of multiple individuals were chosen for this experiment. Plasma/serum miR-223 together with other miRNAs have been shown to be closely associated with the tumorigenesis and metastasis of gastric carcinoma (25), hepatocellular carcinoma or chronic hepatitis (26), sepsis (27), and lung cancer (28). Dysregulated miR-148a has been reported in ovarian cancer (29), liver injury (30), and gastric cancer (31).

To measure expression levels of the 2 miRNAs, exosomes were isolated from fresh human saliva with the conventional differential centrifugation method (see online Supplemental Methods), the effectiveness of which has been well established in our laboratory (32–34). We used the droplet digital PCR (ddPCR) method because it can measure absolute concentration of each miRNA and does not need internal controls. As shown in Fig. 1C, both miRNAs can be detected in all 3 subjects, and they are predominantly present in the exosomal RNA fraction in 2 of 3 subjects. For 1 subject (S1), both miRNAs were detected in both exosomal and nonexosomal fractions. This observation is not likely to be due to failed separation of the exosomal fractions, because piRNAs in this subject showed predominant localization in the exosomal fraction (see below). It should be noted that these data only serve as a qualitative evaluation rather than a quantitative validation of the RNA-Seq data, because the RNA was obtained in very different ways in the 2 types of experiments. In addition, the exosome isolation step introduced relatively large technical variation across samples, which may explain the large interindividual variation observed in miRNA expression level. Nevertheless, our result is consistent with previous findings that miRNAs are mainly localized in exosomes in most individuals. However, as shown in Fig. 1C, it is likely that there also exist vesicle-free ncRNAs in saliva (e.g., for subject S1), which should be further investigated.

VARIATION OF miRNA EXPRESSION ACROSS INDIVIDUALS
As shown in Fig. 1B, miRNA expression values are generally significantly correlated across individuals, as measured by RNA-Seq of CFS total RNA. However, there does exist noticeable variation across different subjects. We asked whether salivary miRNAs are particularly subject to individual variability, given that saliva is readily exposed to and communicative with the external environment that can be highly individual specific. To examine this question, we defined an individual specificity index (ISI) for each miRNA (see online Supplemental Methods). This index has a value between 0 and 1, with larger values representing higher interperson variability. For comparison, we also analyzed several other data sets derived from primary brain tissues (35), skin (36), B cells (37), cerebral spinal fluid (CSF), and serum (15). As shown in Fig. 1D, miRNAs generally demonstrated a wide range of individual variability in all types of samples. Interestingly, the salivary ISI distribution was relatively similar to those of other body fluids (CSF and serum) and intracellular RNA (B cells). Compared with other samples, brain and skin miRNAs showed higher individual variability, possibly reflecting the heterogeneous cell type composition in these samples. Overall, our results suggest that the individual variability observed in salivary miRNAs is at a level similar to those observed in other extracellular and intracellular data sets. Given the remarkable diversity of salivary environment across individuals, this observation supports the effectiveness of our cell removal method in enriching for physiological extracellular RNA rather than environmentally related RNA. Thus, extracellular RNA in CFS can serve as stable biomarkers with individual variability similar to that of other body fluids, with the advantage of being highly accessible and noninvasive.

miRNA EXPRESSION PROFILES OF CFS CLUSTER WITH THOSE OF OTHER BODY FLUIDS
We next conducted a comprehensive comparison of miRNA expression profiles in different body fluids and
A total of 95 small RNA sequencing data sets (including our 8 data sets) were analyzed for miRNA expression with the same method (Fig. 2; online Supplemental Fig. 3; online Supplemental Methods). Raw sequencing reads were used, except for a few data sets in which read counts of miRNAs were directly taken from the original publication owing to lack of raw data (as noted in online Supplemental Fig. 3). Batch effects or technical variations across laboratories may be a significant confounding factor in this type of analysis. We used a data normalization method similar to that adopted in DESeq (38) to alleviate batch effects. In addition, the correlation across data sets was evaluated with the Kendall \(\tau\) method, a rank-based nonparametric correlation analysis (see more details in online Supplemental Methods). In this manner, data comparison is less sensitive to the quantitative values of miRNA expression levels, which may fluctuate due to technical variation.

As shown in online Supplemental Fig. 3, data sets generated from the same tissue or cell type (brain, B cells, or ES cells) by different laboratories clustered together. This observation suggests that batch effects have been adequately reduced, although it may not have been possible to reach a complete elimination. Strikingly, all the data sets derived from extracellular body fluids [CFS, serum, CSF, and plasma (exosome-associated RNA)] clustered together, with relatively smaller distances compared with their distances to intracellular RNA samples. This observation could not have been due to batch effects, since different laboratories generated the data sets of various body fluids. Thus, our analysis suggests that extracellular miRNAs in different body fluids share similar profiles, indicating existence of commonality in the biogenesis of these miRNAs.

Nevertheless, miRNAs in CFS may also have distinct expression patterns that reflect the local cellular environment of the salivary glands and oral mucosa. Epithelial and other cells may release cellular miRNAs into the extracellular space and contribute to the miRNA profile in CFS. Interestingly, among cellular
data sets that are relatively close to the CFS samples in the clustered heat map (Fig. 2; online Supplemental Fig. 3), several may contain cell types similar or related to the cellular environment of saliva, such as mammary epithelial cells, skin cells, endometrium (with epithelial cells as a major layer), fibroblasts, etc.

piRNAs ARE RELATIVELY ABUNDANT IN HUMAN CFS
As mentioned above, piRNAs constitute another group of small RNAs in CFS with an appreciable peak in the length distribution of sequencing reads (Fig. 1A). A total of 32–109 piRNAs were detected in each saliva sample with at least 1 RPM. Compared with the total number of piRNAs in public databases (23,439 in piRNABank (39)), the number of piRNAs detected in our study was small. Nevertheless, some of the piRNAs were expressed at relatively high expression levels in CFS. For example, the most abundant piRNA, piR-018570, had an average expression level of 32,296 RPM across all samples (see online Supplemental Table 3).

The expression levels of piRNAs across different CFS samples were highly correlated (Fig. 3A; online Supplemental Fig. 5), although not as strongly as that for miRNAs, possibly because of the relatively small number of piRNAs at high expression levels.

The ISI values of piRNAs in CFS were overall similar to those in B cells (Fig. 3B) but lower than in brain or skin tissues. This observation is similar to that for miRNAs (Fig. 1D). However, the ISI values of piRNAs in CFS were generally larger than those of CFS miRNAs, indicating a higher degree of individual variability. This result may be explained by the possible diversity of cellular origins of CFS piRNAs, which is discussed below.

To test the presence of piRNAs in exosomes, 2 highly expressed piRNAs (piR-001184 and piR-014923) were chosen for ddPCR analysis (Fig. 3C). Both piRNAs were identified in at least 2 subjects, which demonstrated predominant localization in exosomal RNA fractions. Interestingly, in contrast to the

Fig. 3. piRNA expression in CFS.
(A), Scatter plot of piR expression (log2 RPM) across 2 individuals. Pearson correlation coefficient is shown. (B), Histogram of ISI for all expressed piRNAs calculated with the 6 independent saliva samples (biological replicates were not included). The distributions of ISI values of other public data sets are also shown for comparison. (C), Experimental validation of piRNA expression in exosome fraction (E) and exosome-free fraction (NE), similar to Fig. 1C.
ddPCR results of miRNAs in subject S1, both piRNAs were mainly detected in exosomes of this subject. Overall, our data suggest that the 2 piRNAs in this experiment had a predominant localization in exosomes in all subjects with detectable signals.

COMPARISON OF piRNA EXPRESSION PROFILES OF CFS AND OTHER SAMPLES

As for miRNAs, we compared the piRNA expression profiles across a large number of samples (Fig. 4; online Supplemental Fig. 5). Raw sequencing data were analyzed, and across-data set normalization was conducted in the same way as for miRNAs. Importantly, since the number of piRNAs is relatively small, we carried out the normalization procedures by combining all data related to miRNAs and piRNAs to avoid potential bias due to the small number of variables. We observed that many piRNAs were highly specific to only a subset of data sets. For example, only 23% of piRNAs were detected with >1 read in ≥50% of the samples included in online Supplemental Fig. 5, whereas 40% of miRNAs were found in a similar analysis. Due to this type of scarcity, we did not conduct a rank-based correlation analysis for piRNAs. Instead, the heat maps in Fig. 4 and online Supplemental Fig. 5 directly visualize piRNA expression levels with hierarchical clustering.

As shown in Fig. 4 and online Supplemental Fig. 5, CFS, embryonic stem (ES) cells, and skin cells were among those having the highest expression levels of piRNAs. The same observation can also be appreciated in online Supplemental Fig. 6, in which the expression levels of piRNAs are directly shown as empirical cumulative distributions. In addition, the data were examined for the fraction of piRNAs with at least 10 reads (after data normalization) in each sample (see online Supplemental Fig. 7). Again, CFS, ES cells, and skin cells were among those with the highest fraction of moderately or highly expressed piRNAs. In contrast, the piRNA expression in plasma (exosome-bound) was relatively low (see online Supplemental Figs. 5–7). These data suggest that that most salivary piRNAs may not have originated from circulating RNAs in blood. The similarity of piRNA profiles in CFS to those in ES cells and skin cells indicates that cells producing piRNAs in CFS (possibly salivary glands, oral mucosa, etc.) may have stem-like properties or regenerative capacity, which should be an area for further investigation.
IDENTIFICATION OF circRNAs IN HUMAN CFS

circRNAs constitute an emerging type of RNA recently highlighted in several studies involving different cell types and tissues (20, 40–44). It is not yet known whether circRNAs exist extracellularly in body fluid. To examine this question, we constructed strand-specific sequencing libraries with a circRNA enrichment step using RNase R (42). We developed a customized pipeline to identify unique back-spliced circular junctions within the sequencing reads (see online Supplemental Methods). It should be noted that this method searches for de novo circular junctions and does not depend on annotations of known exons and genes. Nevertheless, we observed that many of the predicted circRNAs were generated from known exons with canonical splice site signals (see online Supplemental Table 4). Interestingly, most such canonical circRNAs were also predicted as circRNAs in previous studies of intracellular RNA samples. A total of 95 putative canonical circRNAs were identified with at least 2 distinct circular junction reads among the 4 samples in this study or with 1 read but also reported as intracellular circRNA (http://circbase.org).

In addition to the canonical ones, many predicted circRNAs were not associated with canonical splice site signals (see online Supplemental Table 4). Because previous studies of intracellular RNA data did not consider such noncanonical circRNAs, we imposed a slightly more stringent criterion in calling such circRNAs by requiring at least 3 distinct reads overlapping the putative circular junction. A total of 327 noncanonical circRNAs were identified in this way. Thus, together, our study predicted 422 putative circRNAs in human CFS. Among these predictions, 28, 6, and 1 were common to at least 2, 3, and 4 individuals, respectively (see online Supplemental Table 4). The low degree of overlap may indicate that circRNAs are highly individual-specific. Alternatively, a much larger number of circRNAs may exist in each sample than detected in our study. Considering the large number of predicted intracellular circRNAs (>9000 in http://circbase.org), the latter possibility is likely true.

To date, the functional relevance of most intracellular circRNAs remains largely unknown. To gain some functional insights, we carried out a gene ontology analysis of the genes overlapping putative circRNAs in human CFS. Interestingly, a number of closely related categories were highly enriched, such as chemotaxis, inflammatory response, establishment of T cell polarity, cellular movement, actin cytoskeleton organization, and integrin-mediated signaling pathway (see online Supplemental Table 5). Overall, this result indicates that salivary circRNAs may be involved in intercellular signaling and inflammatory response. This observation is in line with the fact that inflammation is manifested via periodontal diseases, the most common disease in the oral cavity, and that exRNAs are now known to mediate cellular communications.

EXPERIMENTAL VALIDATION OF circRNAs IN HUMAN CFS

To validate the presence of circRNAs, primers capable of amplifying the predicted circular junctions (outward-facing primers, Fig. 5A) were used in a RT-PCR experiment (see online Supplemental Methods; online Supplemental Table 6). PCR product was visualized on a PAGE gel to confirm the expected size of the circular junction. Subsequently, DNA was purified and subject to Sanger sequencing. We randomly picked 6 putative circRNAs with the corresponding circles formed by 1 or 2 exons using canonical splice site signals. As shown in Fig. 5A, all 6 circular junctions were confirmed on the PAGE gel and, more importantly, with their exact sequences confirmed by Sanger sequencing. Thus, our data provide the first evidence that circRNAs exist in an extracellular body fluid.

Because a large number of our predicted circRNAs do not have canonical splice site signals, we conducted the above validation experiment on 3 additional noncanonical circRNAs. However, instead of direct Sanger sequencing, we conducted clonal sequencing of the TOPO-cloned PCR products because of the small size of these predicted circRNAs. The candidates were picked to represent 2 main categories of observed genomic locations that give rise to noncanonical circRNAs in our data: introns and long noncoding RNA (lncRNA) transcripts. Two of the 3 candidates were confirmed in this experiment (Fig. 5B). The circular junction (chr13:23270854_23270908) generated by a lncRNA was not confirmed, possibly owing to its low expression level and that only a limited number of clones were included for clonal sequencing. For the EIF3E (eukaryotic translation initiation factor 3, subunit E) intron–derived circRNA, the RNA-Seq reads led to discovery of multiple alternative circular junctions that differed by a small number of nucleotides (see online Supplemental Table 4). Our validation confirmed one of the most abundant forms. Interestingly, the ENCODE RNA-Seq data (polyA-fraction) derived from GM12878 cells (available at the UCSC Genome Browser, http://genome.ucsc.edu) included reads that correspond to this predicted circRNA, serving as an independent evidence to support its existence. The EIF3E intron harboring this circRNA is relatively long, and it is unlikely that the circular RNA was generated from the complete intron lariat after splicing. In contrast, the other confirmed circRNA (chr1:20979041_20979113) was derived from an intron of DDOST [dolichyl-diphosphooligosaccharide–protein glycosyltransferase subunit (non-catalytic)], the length of
which is consistent with its generation from the intron lariat. Therefore, diverse biogenesis pathways may exist for these circRNAs, which need to be further investigated in the future.

In summary, we conducted a comprehensive study of the extracellular ncRNA profile of human saliva. To our best knowledge, this is the first report and validation of circRNAs in an extracellular body fluid. In addition, our study, for the first time, revealed the distinction and similarities between the landscapes of miRNAs and piRNAs in CFS and those of other body fluids and intracellular RNA samples. The validated presence of ncRNA in saliva provides novel insights into the biology and regulatory roles that saliva constituents can exert locally in the oral environment as well as systemically as saliva is being swallowed into the gastrointestinal tract. The insights generated in our work lay a foundation for future functional, mechanistic,
and translational discoveries related to ncRNAs in human saliva.

Author Contributions: All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.

Authors’ Disclosures or Potential Conflicts of Interest: Upon manuscript submission, all authors completed the author disclosure forms. Disclosures and/or potential conflicts of interest:

Employment or Leadership: D.T. Wong, RNAmemRIX Inc.
Consultant or Advisory Role: D.T. Wong, PeriRxRNAmemRIX Inc.

References


