

## A clinically validated human saliva metatranscriptomic test for global systems biology studies

Ryan Toma\*,<sup>1</sup>, Ying Cai<sup>1</sup>, Oyetunji Ogundijo<sup>1</sup>, Lan Hu<sup>1</sup>, Stephanie Gline<sup>1</sup>, Diana Demusaj<sup>1</sup>, Nathan Duval<sup>1</sup>, Pedro Torres<sup>1</sup>, Francine Camacho<sup>1</sup>, Guruduth Banavar<sup>1</sup> & Momchilo Vuyisich<sup>1</sup>

<sup>1</sup>Viome, Inc. Viome Life Sciences, Bothell/Bellevue WA 98011/98004. Viome Bioinformatics, New York, NY 10018, USA; \*Author for correspondence: ryan.toma@viome.com

BioTechniques 74: 00–00 (January 2023) 10.2144/btn-2022-0104

First draft submitted: 4 October 2022; Accepted for publication: 7 December 2022; Published online: 9 January 2023

### ABSTRACT

The authors report here the development of a high-throughput, automated, inexpensive and clinically validated saliva metatranscriptome test that requires less than 100  $\mu$ l of saliva. RNA is preserved at the time of sample collection, allowing for ambient-temperature transportation and storage for up to 28 days. Critically, the RNA preservative is also able to inactivate pathogenic microorganisms, rendering the samples noninfectious and allowing for safe and easy shipping. Given the unique set of convenience, low cost, safety and technical performance, this saliva metatranscriptomic test can be integrated into longitudinal, global-scale systems biology studies that will lead to an accelerated development of precision medicine, diagnostic and therapeutic tools.

### METHOD SUMMARY

This paper introduces a novel method for the preservation and metatranscriptomic analysis of low volumes of saliva. The method involves mixing saliva with a preservative at the time of sample collection, isolating RNA, depleting rRNAs from the sample and converting the RNA into directional, dual-barcoded libraries for sequencing. This method can be used for the analysis of microbial species, Kyoto Encyclopedia of Genes and Genomes orthologs and human genes present in saliva.

### KEYWORDS:

clinically validated • disease • metatranscriptomics • microbiome • population studies • saliva

Chronic diseases are a leading cause of morbidity and mortality globally, yet treatment and prevention options have shown limited success. Importantly, chronic diseases have been shown to be weakly associated with genetics; instead, the human microbiome and alterations to human and microbial gene expression patterns have been identified as the underlying driver of many chronic diseases [1,2]. To identify the etiology of chronic diseases and develop more effective preventative measures, comprehensive gene expression analysis of the human body and associated microbiomes is needed. Saliva is noninvasive, is easy to obtain and contains a diverse microbiome involved in many fundamental aspects of human physiology [3–6]. Saliva is also a suitable clinical specimen for the identification of human pathogens, such as SARS-CoV-2 [7]. Human transcripts in saliva can also produce an informative snapshot of human gene expression and its possible roles in human health and disease [8–10]. Saliva presents an opportunity to investigate the oral microbiome and oral human transcriptome and their roles in disease pathogenesis [6,11].

There is a critical need for noninvasive, at-home sample collection that can be used to predict, diagnose and inform therapeutic options for chronic diseases. Critically, methods for investigating multiple aspects of human physiology need to be pioneered to adequately shed light on the diversity of factors impacting chronic diseases. For example, systems biology transcriptomic methods have already been employed with great success in cancer therapies [12–14]. Methods to analyze the transcriptome are becoming invaluable in the understanding of chronic diseases with unknown etiologies, but large-scale adoption of metatranscriptomic analysis has not yet been possible due to the lack of low-cost and scalable methods.

The saliva microbiome has been associated with numerous chronic diseases, including Alzheimer's disease [15], Parkinson's disease [16], autism spectrum disorder [8], cancers [17], cardiovascular disease [18], diabetes [19], obesity [20] and autoimmune disorders [21,22]. As an example of the direct effect of salivary microorganisms on disease, *Fusobacterium nucleatum* originating in the oral cavity has been shown to contribute to colorectal cancer development [23,24]. The oral microbiome has also been shown to have a strong influence on the gut microbiome and the immune system, which are involved in a variety of chronic diseases [25]. In addition to the clear role of the saliva microbiome in human health and disease, there is substantial information to be gained from human gene

expression patterns in saliva. Alterations in salivary human gene expression and epigenetic markers have been observed in a range of disorders such as autism spectrum disorder [9], Parkinson's disease [26] and traumatic brain injury [27]. Saliva is clearly an informative sample type in human health and disease, with predictive biomarkers of several chronic diseases.

Despite the abundance of literature showing clear connections among the oral microbiome, human gene expression and chronic diseases, most of the evidence is based on 16S rRNA gene sequencing or metagenomic sequencing methods. Both 16S and metagenomics have intrinsic limitations that minimize the ability to discover biomarkers and to provide actionable therapeutic interventions. 16S is limited in its ability to reliably identify microorganisms at the species or strain level and, importantly, cannot provide insights into microbial or human gene expression [28–31]. Metagenomics is limited in its ability to detect RNA viruses and, again, critically cannot identify active microbial or human gene expression. The analysis of gene expression is an important component of understanding chronic diseases and has been shown to have a greater role than simply looking at the microbial composition [32,33]. These limitations of the standard methodologies result in an incomplete picture of the saliva microbiome ecosystem, which minimizes the ability to discover novel microbial or human-derived biomarkers [34]. Metatranscriptomic analysis of the saliva microbiome presents a more comprehensive and relevant snapshot of the microbial and human transcripts compared with traditional methodologies. Metatranscriptomic analyses address the limitations of both 16S rRNA gene sequencing and metagenomics by providing high-level taxonomic resolution (strain level) and by allowing for assessments of both human and microbial gene expression [35,36]. Metatranscriptomics is uniquely capable of providing much of the information needed for detailed biomarker discovery.

Saliva metatranscriptomic tests have already been used by the authors' laboratory to create comprehensive and accurate diagnostic indicators of oral cancer [37]. Importantly, these predictive models rely on microbial taxonomy, microbial gene expression and human gene expression signatures.

In addition to their utility in biomarker discovery, functional outputs are also critical in developing a comprehensive understanding of disease etiology and subsequent treatment options [38]. Gene expression analyses can also provide information about microbial pathogenicity, which is lost when simply looking at the microbiome composition. *Porphyromonas gingivalis*, for example, can be present in small amounts but has highly expressed functions with a strong negative impact on the microbial community and ultimately the host [32,39].

Scalable, affordable and at-home sample collection had limited metatranscriptomic clinical applicability in the past [40]. Effective RNA preservation has been challenging, traditionally requiring a cold chain that is expensive and complicated. In addition, the high abundance of noninformative RNA sequences (such as rRNA and tRNA) in clinical samples requires high sequencing depth and associated costs to overcome. By utilizing the selective depletion of noninformative RNAs, mRNA sequences can be enriched, which dramatically reduces sequencing costs while improving data resolution [35,41].

Here the authors present a comprehensive (sample collection-to-result) method for the quantitative metatranscriptomic analysis of saliva samples that can easily be applied to clinical studies and trials globally. The method is automated, high-throughput, inexpensive and clinically validated, and it includes a fully automated bioinformatic suite for strain-level taxonomic classification of all microorganisms and human genes and their quantitative gene expression levels.

## Methods

### Ethics statement

All procedures involving human subjects were performed in accordance with the ethical standards and approved by a federally accredited institutional review board committee. Informed consent was obtained from all participants, who were residents of the USA at the time.

### Sample collection

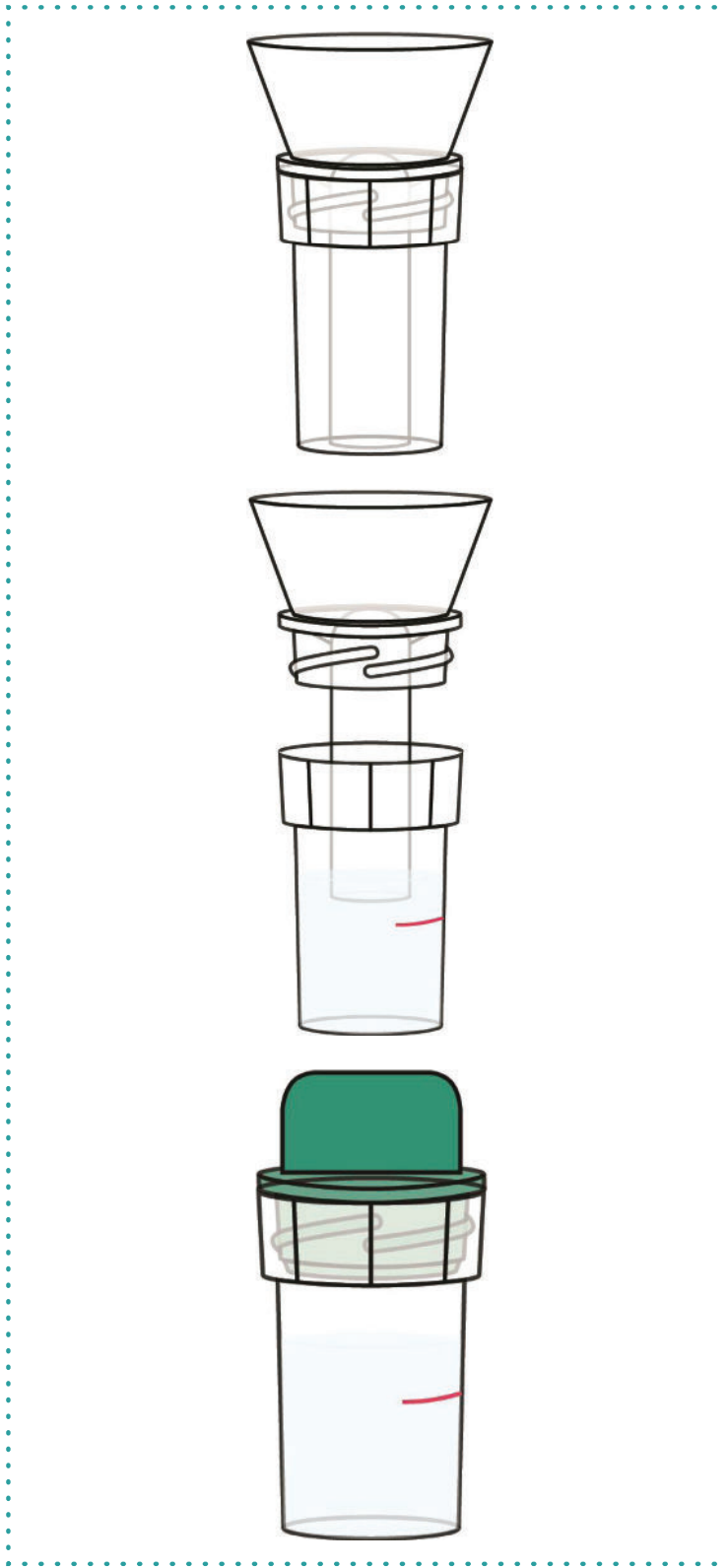
Viome has developed a patented and convenient sample collection device that allows anyone, even children, to easily collect sufficient volumes of saliva and preserve it for metatranscriptomic analysis (Figure 1). For future studies, saliva can be collected in any device that enables mixing of the Viome RNA preservative buffer (RPB) with the sample as soon as it is collected.

Participants were instructed to fast for 8 h prior to collection, to gently rinse their mouth with water for 10 s before collecting the sample, to collect 1.2 ml of saliva within 1 h of waking up and then to dispense the RPB and shake the collection tube vigorously for 15 s.

### Method validation

The performance of the saliva method was assessed by determining the method precision, sample stability and longitudinal changes in saliva. Precision was assessed in four technical replicates of saliva from ten donors. Sample stability was assessed by storing samples for 7 and 28 days at room temperature compared with the technical replicates analyzed on the day of collection (day 0). A subset of replicate samples was shipped and returned to Viome to assess the impact of shipping on sample stability. All sample stability analyses were performed with four technical replicates among three donors. Longitudinal changes and Hellinger distances in saliva were assessed by collecting saliva weekly for 5 weeks from eight donors.

To develop a comprehensive saliva cohort, saliva samples were collected from 1102 individuals. For the saliva cohort, the average age was 48 years (range: 12–88 years) with 436 males, 665 females and one participant identifying as another gender.



**Figure 1. At-home saliva collection device.** The insert contains the sample preservative that inactivates all types of pathogens and preserves both RNA and DNA for 28 days at ambient temperature. Once saliva is deposited inside the tube, the funnel and insert are removed, which releases the preservative and mixes it with the specimen.

## Laboratory analysis

The method for the analysis of saliva samples was similar to that which has been previously described by the authors' lab [42,43]. Briefly, saliva samples were chemically lysed using the RPB, followed by mechanical lysis using bead-beating beads. RNA extraction was performed with silica beads on 87.5  $\mu$ l of saliva. DNA was digested using DNase. rRNAs, both prokaryotic and eukaryotic, were removed via a subtractive hybridization method. Directional, dual-barcoded libraries were generated and analyzed with Qubit dsDNA (Thermo Fisher Scientific) and Fragment Analyzer (Advanced Analytical) methods. Library pools were sequenced on Illumina NovaSeq instruments using 300-cycle kits.

## Bioinformatic analysis

Viome's bioinformatic methods include quality control, strain-level taxonomic classification, microbial gene expression and human gene expression characterizations. The quality control includes per sample and per batch metrics, such as the level of barcode hopping, batch contamination, positive and negative process controls, DNase efficacy and number of reads obtained per sample. Following the quality control, the paired-end reads are aligned to a catalog containing rRNA, the human transcriptome and 53,660 genomes spanning archaea, bacteria, fungi, protozoa and viruses. Reads that map to rRNA are filtered out. Strain-level relative activities are computed from mapped reads via the expectation-maximization algorithm [44]. Relative activities at other levels of the taxonomic tree are then computed by aggregating according to taxonomic rank. Relative activities for the biological functions are computed by mapping paired-end reads to a catalog of 52,324,420 genes, quantifying gene-level relative activity with the expectation-maximization algorithm and then aggregating gene-level activity by Kyoto Encyclopedia of Genes and Genomes ortholog (KO) annotation [45]. The identified and quantified active microbial species and KOs for each sample are then used for downstream analysis.

For analysis of the human and microbial transcriptome, each sample was analyzed by two separate pipelines (for the microbiome and human transcriptome, respectively). It is possible to map reads to both pipelines.

To assess the ability of Viome's custom catalog to interrogate the oral microbiome, the authors compared the genera and species present in the Viome catalog with those present in the expanded Human Oral Microbiome Database V3, which contains 2123 genomes representing 539 species ([www.homd.org/genome/genome\\_table](http://www.homd.org/genome/genome_table)). Of the 539 species, 109 are uncultured oral taxa with undefined taxonomic classification. Of the remaining 430 species, 307 (71%) are present in Viome's catalog. The presence rate is higher at the genus level, with 158 out of 191 genera (83%) present in Viome's catalog. Viome's custom catalog is sufficiently comprehensive to provide insight into the oral microbiome.

For the presentation of metrics from the saliva cohort (Supplementary Table 6), the percentage of total reads that align to archaea, bacteria, eukaryotes, viruses and humans are reported. These percentages include only reads that align to their target and meet Viome's quality-control criteria. Reads that align to rRNAs, internal control sequences, rRNA depletion probes and unaligned reads are removed.

## Data analysis

Statistical parameters, including transformations and significance, are reported in the figures and figure legends. To compare pairs of samples, the authors report Spearman correlation coefficients (which are invariant to absolute expression levels of the genes and only consider the similarity of ranked expression), Pearson correlation coefficients (which compare the linear relationship between features and is sensitive to differences in expression levels) and Hellinger distance (an appropriate distance measure for compositional data). For correlations of species and KOs, the union of features between sample pairs was analyzed. For correlations of human genes, the overlap of features between sample pairs was analyzed. The multiplicative replacement method [46] was employed to deal with missing values, the data were transformed with centered log-ratio (CLR) [47] and Spearman correlation coefficients were computed. CLR transformation is commonly done to reduce false discoveries due to the compositional nature of sequencing data. CLR transformation breaks the dependence between features (i.e., transcripts) and makes data more normally distributed, reducing the impact of highly abundant transcripts that could be artificially driving high correlation values. Statistical analyses were performed in Python.

The relative abundance of transcripts was calculated by calculating the relative activity value of every molecular feature (gene, KO, taxa). The relative activity was calculated by taking the total number of mapped reads to that molecular feature, dividing it by effective sequencing depth (the total number of reads aligned to human transcripts, KOs or species that meet Viome's quality-control requirements) and then normalizing so that the sum across all features in a sample is equal to one. Multiplicative replacement was then performed to fill in the zero values.

To assess the relationship between expression bins (low, medium and high) within a sample, correlation coefficients were individually computed for each bin. The threshold between the low and medium bins was set to represent the 50th percentile, and the threshold between the medium and high bins was set to represent the 95th percentile of the CLR values of the samples. For the assessment of the lower limit of detection, the expression threshold that was needed to achieve an overall correlation of 0.6 was computed. All correlation values were above 0.6 for species and KOs regardless of thresholding, indicating that no specified lower limit of detection was needed for high-quality data. A detection threshold was needed for robust analysis of human genes, which is reported in the associated figures.

**Table 1. Total reads and effective sequencing depth (the number of reads aligned to human transcripts) of resequenced saliva samples for human gene expression precision analyses.**

Donor ID	Average total reads	Total reads SD	Average effective sequencing depth	Effective sequencing depth SD
PID-0026	73,653,434	38,925,007	1,155,273.803	592,131.0757
PID-0290	110,747,517	27,390,256	1,075,291.804	381,364.1083
PID-0291	84,602,795	45,127,258	1,674,916.386	861,131.1174
PID-0372	87,167,604	17,886,807	1,765,820.818	586,940.0628

## Results & discussion

### Saliva cohort metrics

For the cohort of 1102 saliva samples, the average number of sequencing reads per sample was 24,353,309. On average, there were 487 species, 2539 microbial KOs and 5174 human transcripts detected per sample. See Supplementary Tables 1–5 for a rank-ordered list of all detected genera, species, strains, microbial KOs and human transcripts in the saliva cohort. See Supplementary Table 6 for a breakdown of total reads, species/KO/human gene richness and the percentage of reads aligned to microorganisms and human genes per donor. The metatranscriptomic data demonstrate that the species and KO richness are less variable between people (%CV of 14.7% for species richness and 14.2% for KO richness) than the human gene richness (%CV of 59.8% for human gene richness). Additionally, the data show that a higher proportion of reads align to microorganisms (on average, 5.9%) than human genes (on average, 2.4%).

### Method precision

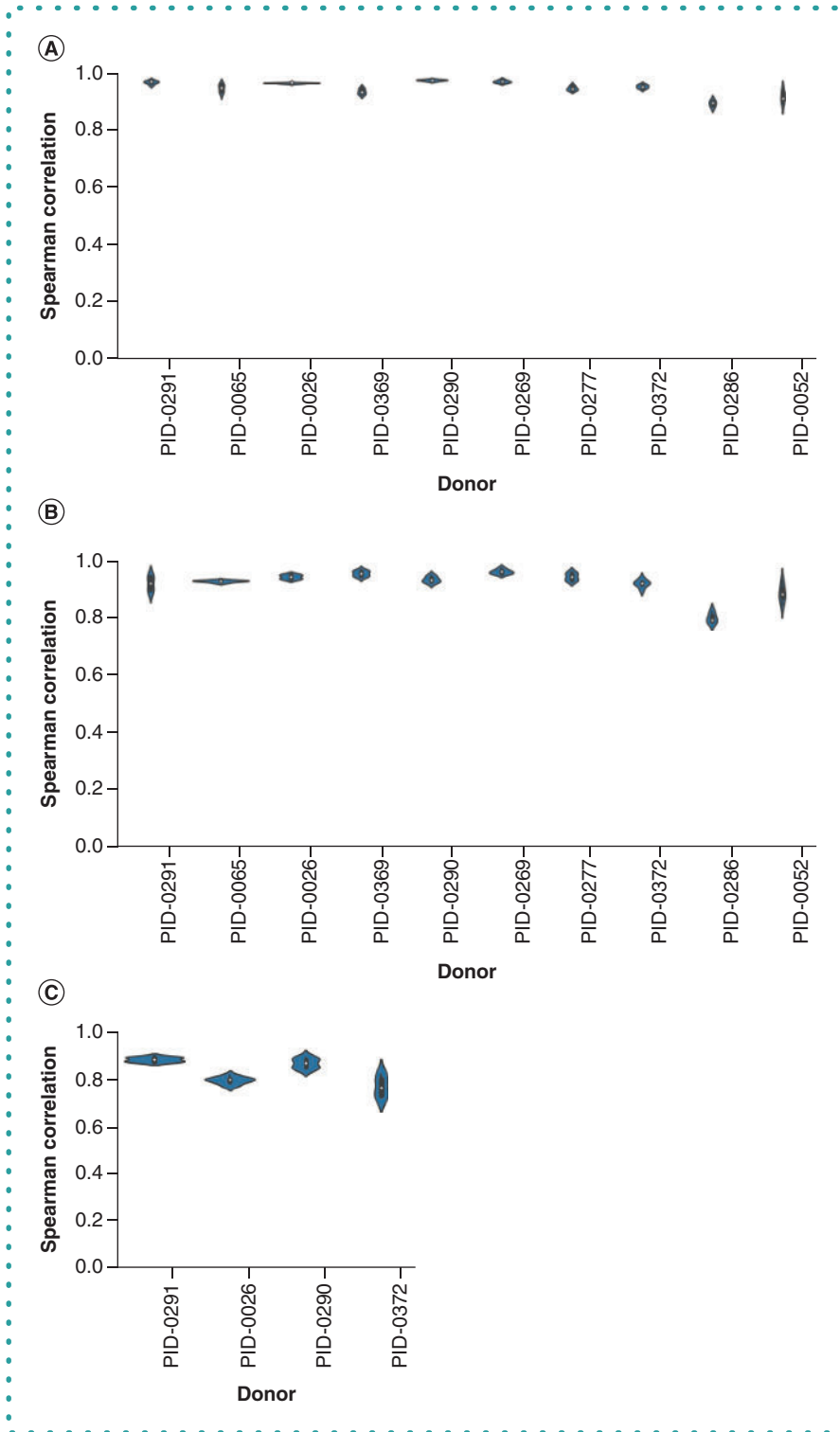
For longitudinal clinical studies in large populations that measure gene expression changes as a function of health and disease states, a method with high precision is extremely important, as it allows for the measurement of small changes in the levels of gene expression that are associated with or predictive of chronic disease. To determine the precision of the metatranscriptomic method, Spearman correlations were calculated for microbial taxonomy (species) and microbial functions (KOs) for four technical replicates from ten participants (Figure 2A & B). For one participant (PID-0286), one replicate was determined to be an outlier, with the log-transformed effective sequencing depth (the number of sequencing reads aligned to microbial species) being greater than three SDs away from the mean and was removed from the analysis, as it failed quality control. The method precision for microbial taxonomy and microbial functions is high, showing that the method is able to reliably reproduce the results (Figure 2A & B).

Human transcripts are present in low amounts in saliva and therefore require higher sequencing depth to analyze. Samples from four donors were randomly chosen for resequencing with higher sequencing depth to determine the method precision for human genes. The effective sequencing depth (the number of reads aligned to human transcripts) needs to be about 1 million reads to yield high-quality human gene expression data (Table 1 & Figure 2C). This demonstrates that the technology does allow for high-precision human gene expression analyses from saliva but requires higher sequencing depth.

In addition to the Spearman correlations, two random technical replicates were compared from each participant for microbial taxonomy, microbial KOs and human gene expression (Figure 3). To further categorize the correlations between technical replicates, correlations were computed for the low, medium and high expression bins (Supplementary Figure 1). The data indicated generally high correlations across expression bins. For a robust analysis of human genes, an average detection threshold of 10.25 p.p.m. was shown to be sufficient (Supplementary Figure 1C). These data show that Viome's metatranscriptomic saliva test is able to produce data with high precision.

### Sample stability

RNA is prone to rapid degradation, making proper sample preservation a critical component of any metatranscriptomic test [48]. As previously described in the authors' publications, their laboratory uses a proprietary RPB to prevent RNA degradation [42,43]. The ability of RPB to preserve RNA in saliva was validated by collecting saliva from three participants and storing it at ambient temperatures (72° F, 22°C) for 0, 7 and 28 days. Subsets of samples were also packaged and were shipped via the US Postal Service to a designated address, then shipped back to the laboratory using the same shipping method to emulate an at-home testing process. Four technical replicates per condition were analyzed using the test, and microbial taxonomy and microbial functions were compared among all samples using Spearman correlations (Figure 4) and Pearson correlations (Supplementary Figure 2). The correlation between microbial species in saliva samples in all conditions was very high, with Spearman correlation coefficients above 0.928 for all conditions tested (Figure 4A). The correlation between microbial functions in saliva samples in all conditions was also very high, with Spearman correlation coefficients above 0.901 for all conditions tested (Figure 4B). Venn diagrams showing the number of overlapping and unique features between each storage condition and their expression levels can be found in Supplementary Figure 3. The data demonstrate that the majority of features overlap even across storage conditions and that the expression levels of the overlapping features are significantly higher than the unique features ( $p < 0.05$ ). These data show that Viome's saliva metatranscriptomic analysis method can adequately preserve RNA for up to 28 days at ambient temperature, inclusive of sample shipping.



**Figure 2. Method precision is shown, comparing technical replicates from ten donors for microbial species taxonomy (A) and for microbial functions (B) and from four donors for human transcripts (C).**

### Longitudinal changes in the saliva transcriptome

It is important to understand the longitudinal stability of the salivary transcriptome so that large-scale studies can be reliably carried out. Toward this goal, the authors recruited eight subjects, who collected saliva samples weekly for 5 weeks. On average, 537.18 microbial species and 2693.54 microbial KOs were detected per sample. For all of the taxa and KOs that were detected, the correlation remained

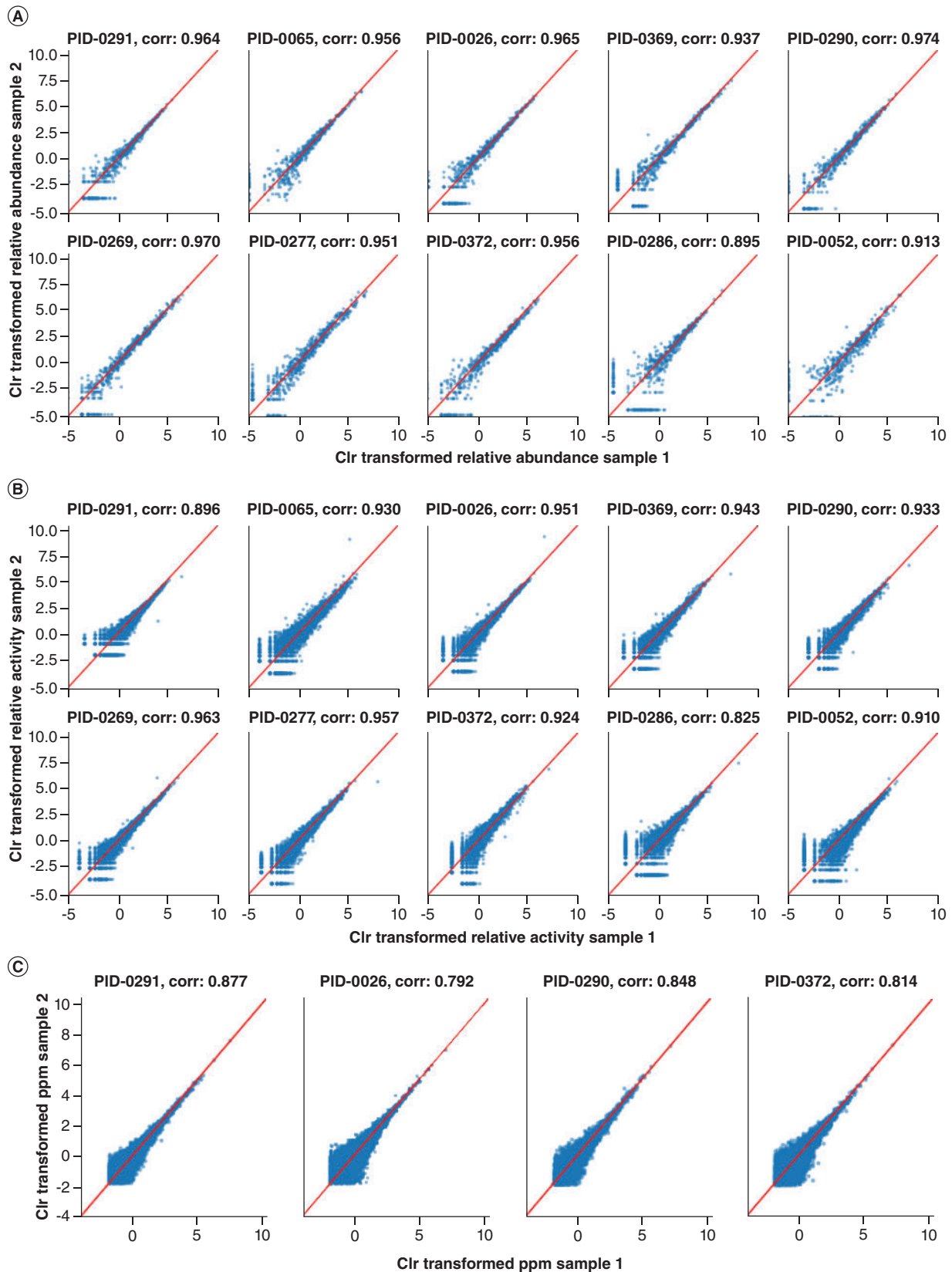
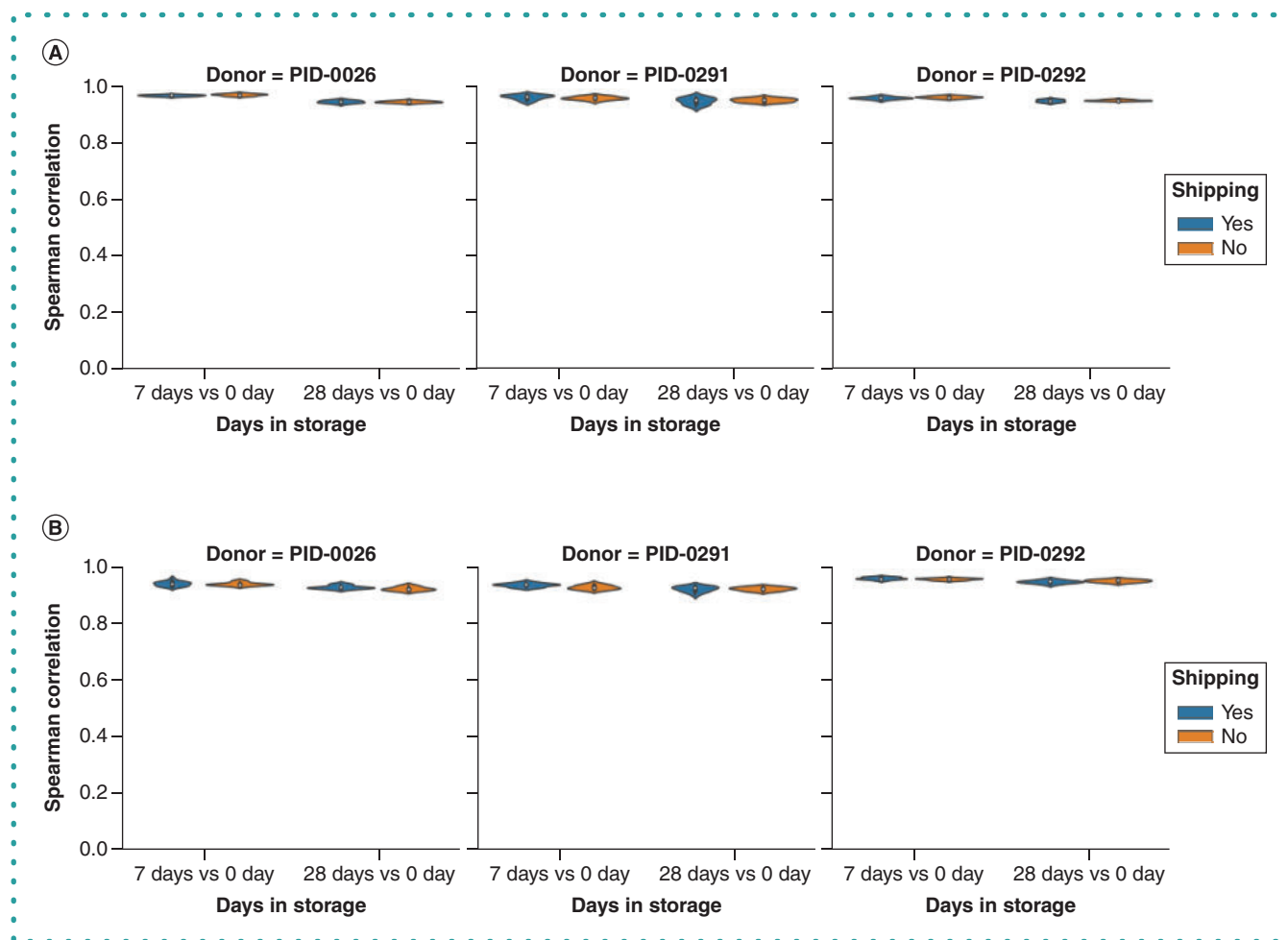


Figure 3. Relative abundance of species and transcripts are shown between two technical replicates from ten donors for microbial species taxonomy (A) and microbial functions (B) and from four donors for human genes (C).



**Figure 4. Spearman correlation coefficients of species taxonomy (A) and microbial functions (B) for sample stability of saliva samples stored at ambient temperature for 0, 7 and 28 days, with and without shipping conditions.**

very high across 5 weeks for each participant (Figure 5 & Supplementary Figure 1). Correlations for one participant across expression bins (low, medium and high) are presented in Supplementary Figure 4. The correlations for each expression bin were generally high, indicating robust method performance. The correlations observed between the 1-week and 5-week samples for taxonomy ranged from 0.87 to 0.94 (Figure 5A & Supplementary Figure 1A–H) and for KOs ranged from 0.88 to 0.94 (Figure 5B & Supplementary Figure 1I–P). These data show that the saliva microbiome is longitudinally stable in terms of both composition (taxonomy) and activity (KOs).

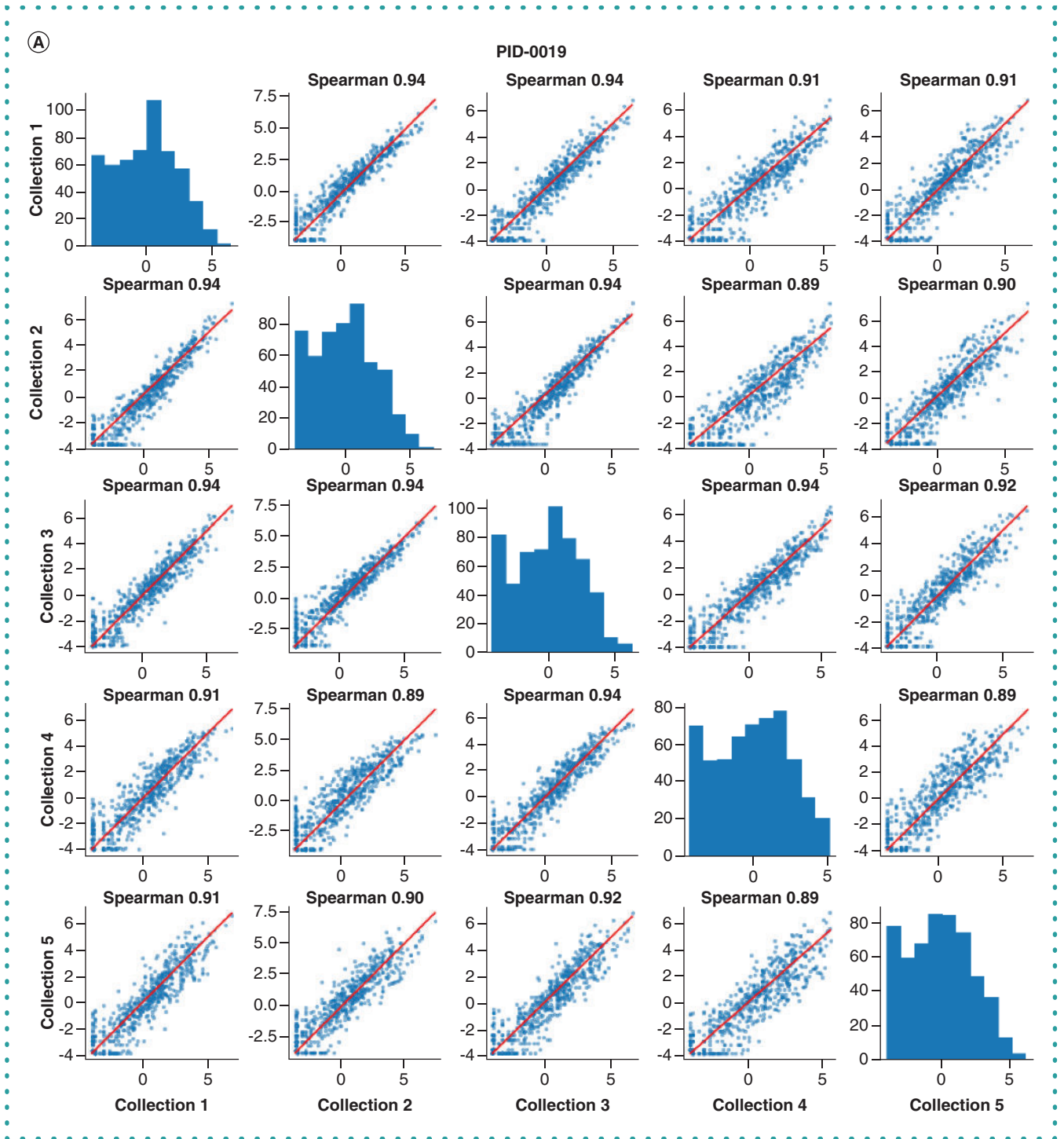
### Intrasample versus intersample precision

One of the more important parameters of the saliva test is to be sufficiently precise to distinguish very small changes among thousands of measured features. Such high precision would enable the identification of even minor transcriptome changes over time and changes related to health and disease. To assess the test precision in this context, the authors compared Hellinger distances among biological samples from the same person collected over time (intraperson distances in Figure 6) and biological samples from different people (interperson distances in Figure 6). Empirical cumulative distribution function plots of Hellinger distance show that when taking microbial species (Figure 6A), microbial functions (Figure 6B) and human genes (Figure 6C) into account, samples from one individual over 5 different weeks (intraperson distances) tend to be more similar than pairs of samples coming from two different participants (interperson distances).

### Conclusion

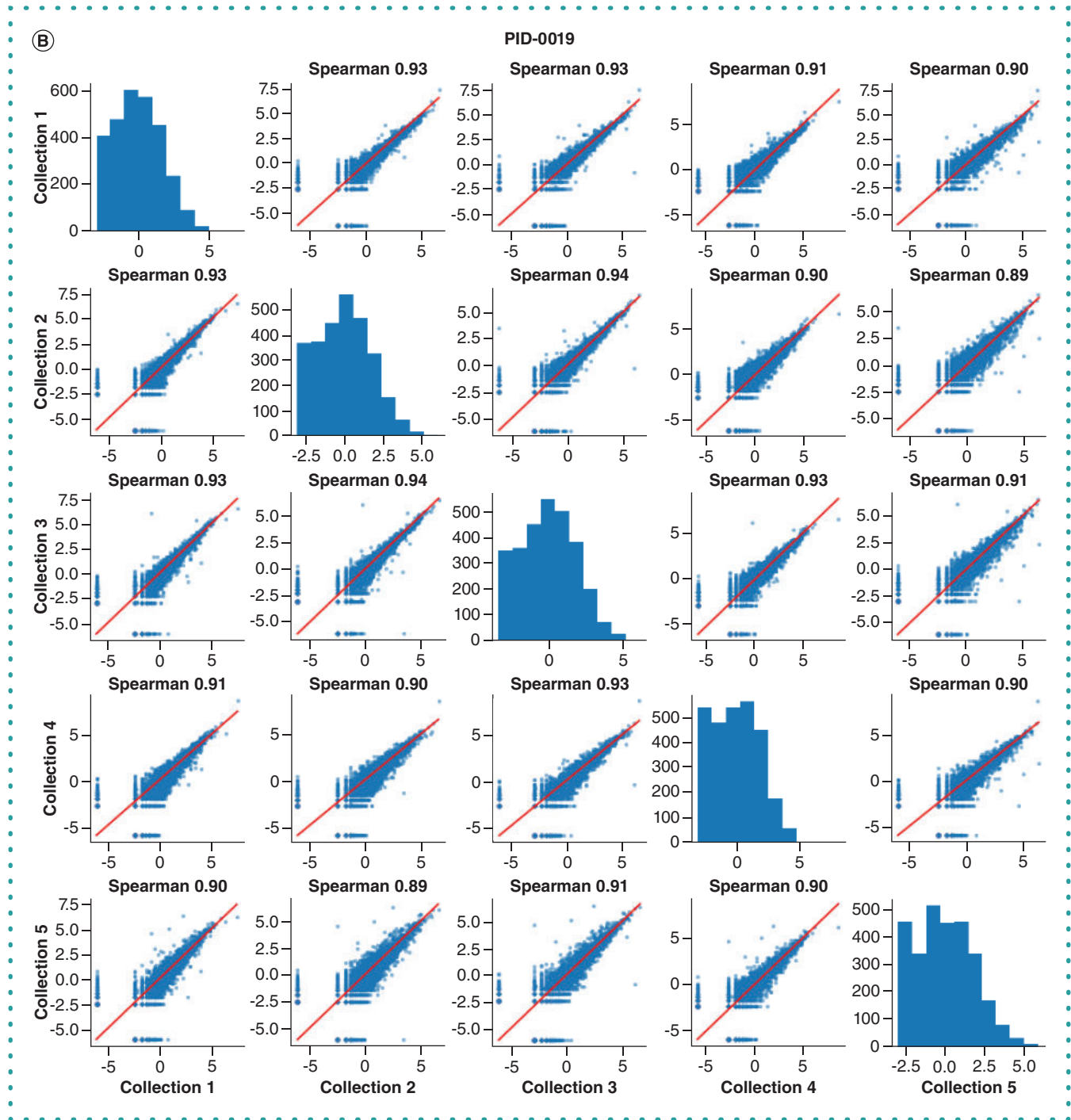
This paper outlines a novel saliva analysis method that can produce high-quality human and microbial gene expression data. This paper also includes, to the best of our knowledge, the largest list of population-scale data from saliva samples analyzed through a metatranscriptomic method (taxonomic classifications, microbial gene expression and human gene expression). The data demonstrate clinical utility with high levels of method precision, adequate sample stability for global shipping (including to and from traditionally underrep-





**Figure 5. Longitudinal stability of saliva transcriptome in eight study participants over a 5-week period (one participant is shown here; for all eight participants, see Supplementary Figure 1). (A) Scatter plots comparing the relative abundance values of each species at each collection time. (B) The same analysis was repeated for Kyoto Encyclopedia of Genes and Genomes orthologs comparing sum transcripts per million of each ortholog.**

resented communities in clinical research) and longitudinal stability suitable for population studies. This saliva method complements metatranscriptomic pipelines previously developed by our laboratory for the analysis of stool and blood [42,43]. With the integration of these tests, our lab has developed a robust and novel systems biology platform for the investigation of biomarkers critical to human health and disease.



**Figure 5. Longitudinal stability of saliva transcriptome in eight study participants over a 5-week period (one participant is shown here; for all eight participants, see Supplementary Figure 1) (cont.). (A) Scatter plots comparing the relative abundance values of each species at each collection time. (B) The same analysis was repeated for Kyoto Encyclopedia of Genes and Genomes orthologs comparing sum transcripts per million of each ortholog.**

Despite recent advancements in a variety of molecular biology techniques, limited insights have been generated into the cause of chronic and noncommunicable diseases. Systems biology approaches that take into account human and associated microbial features are rapidly becoming invaluable tools for the identification of biomarkers critical in human health and disease [37]. The saliva metatranscriptomic method described in this paper is easily deployable across the world, precise, cost effective, automated and clinically validated, and it overcomes many of the limitations previously encountered with saliva microbiome characterizations and metatranscrip-

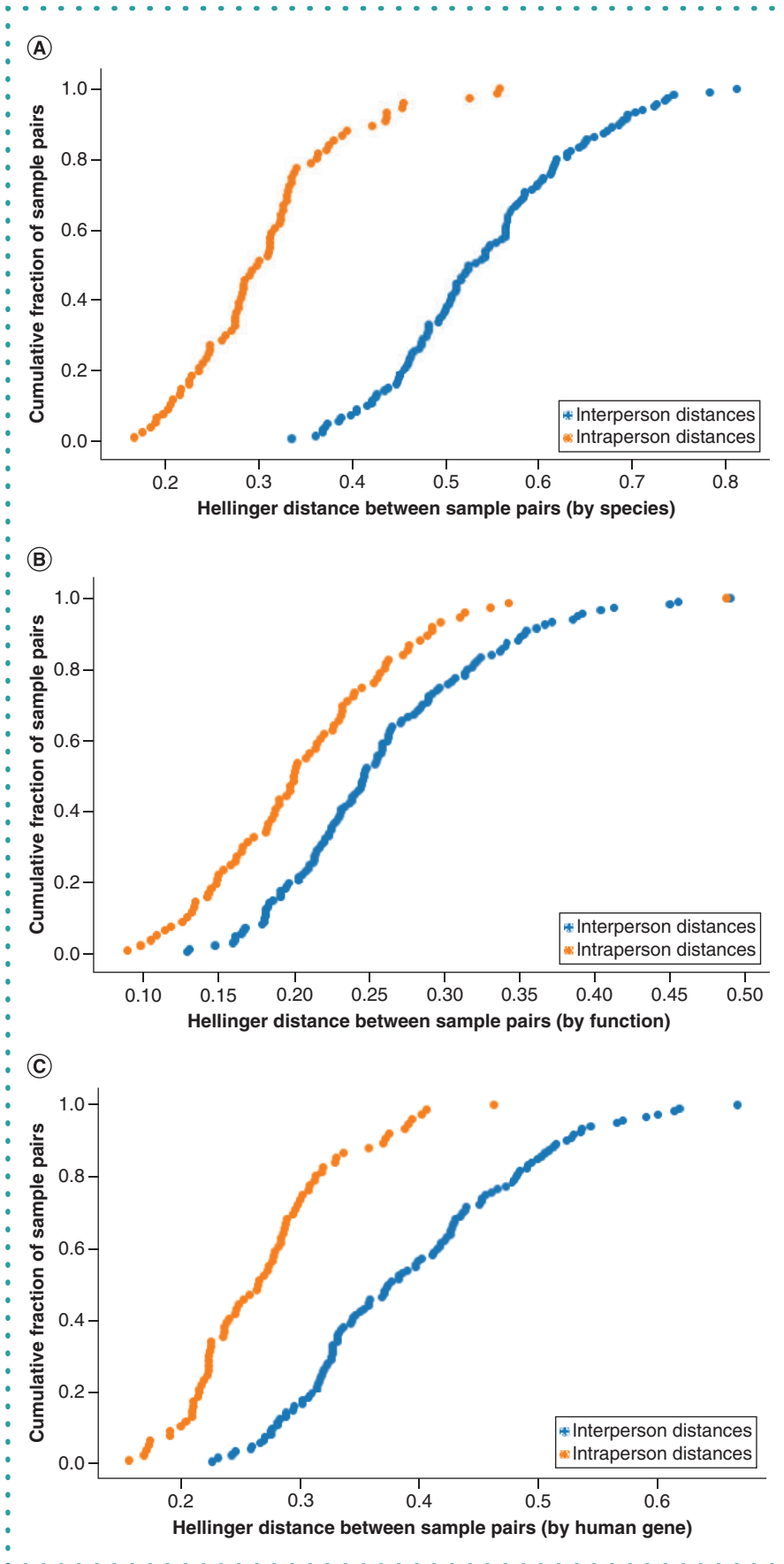


Figure 6. Empirical cumulative distribution function plots of Hellinger distance show that when taking microbial species (A) microbial functions (B) and human genes (C) into account, samples taken from one individual over five different weeks (intraperson distances) tended to be more similar than pairs of samples coming from two different participants (interperson distances).

tomic methods (e.g., sample stability [49]). This method is suitable for use in population-scale studies to elucidate the role of the saliva microbiome and associated human gene expression in human health and disease.

## Future perspective

The methods presented in this paper and other papers from our laboratory demonstrate the feasibility of cost-effective methods for interrogation of the holistic human system (microbial components and human components). We anticipate that over the course of the next 5–10 years these molecular biology techniques will be used in large-scale, population-based studies for the identification of chronic disease etiologies. This will pave the way for the development of accurate and effective diagnostic, therapeutic and preventative therapies to treat chronic diseases.

Since the methods presented by our laboratory allow for at-home sample collection, we also believe that these methods will allow for the democratization of healthcare. This will enable individuals to easily collect samples outside of primary care settings and to obtain critical health information on their own accord. We anticipate that this increased accessibility to health infrastructure will result in better and more widespread screening, which should lower the overall disease burden in the population. Additionally, we hope that shifting the healthcare setting from hospitals to the home will result in an overall reduction in healthcare costs and will enable the effective prioritization of healthcare resources.

Finally, we anticipate that the analysis of microbiomes and human gene expression will open the door for optimized personalized therapies. This personalized approach to healthcare will likely be more effective than traditional treatment strategies.

## Executive summary

### Introduction

- Chronic diseases are a leading cause of death, and the majority of their etiologies are not understood.
- Gene expression profiles are an important component of chronic diseases and have been shown to be more important than genetics alone.
- There is a need for methods that can investigate gene expression profiles of human systems.
- Saliva represents an important sample type for the investigation of the oral microbiome and human gene expression in health and disease.
- The authors' laboratory has developed a robust method for the metatranscriptomic analysis of saliva that can be deployed in population-scale studies.

### Methods

- This paper introduces a novel method for the preservation and metatranscriptomic analysis of low volumes of saliva.
- The method involves mixing saliva with a preservative at the time of sample collection, isolating RNA, depleting rRNAs from the sample and converting the RNA into directional, dual-barcoded libraries for sequencing.
- This method can be used for the analysis of microbial species, Kyoto Encyclopedia of Genes and Genomes orthologs (KOs) and human genes present in saliva.

### Results & discussion

- Human genes are present at lower abundances than microbial transcripts.
- The method is highly precise for species, KOs and human gene, with the results being reproducible between technical replicates.
- Species, KOs and human genes are stable in the authors' proprietary RNA preservative buffer, even after 28 days at room temperature.
- Species, KOs and human genes are stable within a person across 5 weeks.

### Conclusion

- The method presented herein allows for the generation of high-quality metatranscriptomic data from saliva samples.
- This method could be used in large-scale population studies for the identification of disease etiologies.

## Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: [www.future-science.com/doi/suppl/10.2144/btn-2022-0104](http://www.future-science.com/doi/suppl/10.2144/btn-2022-0104)

## Author contributions

M Vuyisich and R Toma conceived and designed the methods. R Toma, D Demusaj and N Duval performed data collection. P Torres and F Camacho developed the bioinformatic pipeline. L Hu, Y Cai, O Ogunjido, S Gline and G Banavar performed data analysis. All authors contributed to data interpretation. R Toma and N Duval contributed to the writing of the manuscript.

## Financial & competing interests disclosure

All authors are current or former employees of Viome, Inc. 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

The authors have obtained appropriate institutional review board approval or have followed the principles outlined in the Declaration of Helsinki for all human or animal experimental investigations. In addition, for investigations involving human subjects, informed consent has been obtained from the participants involved

## Open access

This work is licensed under the Attribution-NonCommercial-NoDerivatives 4.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>

## References

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

1. Rappaport SM. Genetic factors are not the major causes of chronic diseases. *PLOS ONE* 11(4), e0154387 (2016).
- Addresses the importance of investigating microbiomes and gene expression profiles for chronic diseases as opposed to only looking at genetic factors.
2. Durack J, Lynch SV. The gut microbiome: relationships with disease and opportunities for therapy. *J. Exp. Med.* 216(1), 20–40 (2019).
3. Nikitakis NG, Papaioannou W, Sakkas LI, Kousvelari E. The autoimmunity–oral microbiome connection. *Oral Dis.* 23(7), 828–839 (2017).
4. Solbiati J, Frias-Lopez J. Metatranscriptome of the oral microbiome in health and disease. *J. Dent. Res.* 97(5), 492–500 (2018).
5. Vanhatalo A, Blackwell JR, L'Heureux JE et al. Nitrate-responsive oral microbiome modulates nitric oxide homeostasis and blood pressure in humans. *Free Radic. Biol. Med.* 124, 21–30 (2018).
6. Willis JR, Gabaldón T. The human oral microbiome in health and disease: from sequences to ecosystems. *Microorganisms* 8(2), 308 (2020).
7. Wylie AL, Fournier J, Casanovas-Massana A et al. Saliva or nasopharyngeal swab specimens for detection of SARS-CoV-2. *N. Engl. J. Med.* 383(13), 1283–1286 (2020).
8. Hicks SD, Uhlig R, Afshari P et al. Oral microbiome activity in children with autism spectrum disorder. *Autism Res.* 11(9), 1286–1299 (2018).
9. Hicks SD, Carpenter RL, Wagner KE et al. Saliva microRNA differentiates children with autism from peers with typical and atypical development. *J. Am. Acad. Child Adolesc. Psychiatry* 59(2), 296–308 (2020).
10. Kinser HE, Pincus Z. MicroRNAs as modulators of longevity and the aging process. *Hum. Genet.* 139(3), 291–308 (2020).
11. Inchingolo F, Martelli FS, Gargiulo Isacco C et al. Chronic periodontitis and immunity, towards the implementation of a personalized medicine: a translational research on gene single nucleotide polymorphisms (SNPs) linked to chronic oral dysbiosis in 96 Caucasian patients. *Biomedicine* 8(5), 115 (2020).
12. Cabanero M, Tsao MS. Circulating tumour DNA in EGFR-mutant non-small-cell lung cancer. *Curr. Oncol. Tor. Ont.* 25(Suppl. 1), S38–S44 (2018).
13. Chauhan AK, Bhardwaj M, Chaturvedi PK. Molecular diagnostics in liver cancer. In: *Molecular Diagnostics in Cancer Patients* Shukla KK, Sharma P, Misra S (Eds). Springer, Singapore, 293–303 (2019).
14. Rosell R, Carcereny E, Gervais R et al. Erlotinib versus standard chemotherapy as first-line treatment for European patients with advanced EGFR mutation-positive non-small-cell lung cancer (EURTAC): a multicentre, open-label, randomised phase 3 trial. *Lancet Oncol.* 13(3), 239–246 (2012).
15. Bathini P, Foucras S, Dupanloup I et al. Classifying dementia progression using microbial profiling of saliva. *Alzheimers Dement. Diagn. Assess. Dis. Monit.* 12(1), e12000 (2020).
16. Mihaila D, Donegan J, Barns S et al. The oral microbiome of early stage Parkinson's disease and its relationship with functional measures of motor and non-motor function. *PLOS ONE* 14(6), e0218252 (2019).
17. Acharya A, Chan Y, Kheur S, Jin L, Watt R, Mattheos N. Salivary microbiome in non-oral disease: a summary of evidence and commentary. *Arch. Oral Biol.* 83, 169–173 (2017).
18. Pietiäinen M, Liljestrand JM, Kopra E, Pussinen PJ. Mediators between oral dysbiosis and cardiovascular diseases. *Eur. J. Oral Sci.* 126(Suppl. 1), S26–S36 (2018).
19. Long J, Cai Q, Steinwandl M et al. Association of oral microbiome with Type 2 diabetes risk. *J. Periodontol Res.* 52(3), 636–643 (2017).
20. Wu Y, Chi X, Zhang Q, Chen F, Deng X. Characterization of the salivary microbiome in people with obesity. *PeerJ* 6, e4458 (2018).
21. Sharma D, Sandhya P, Vellarikkal SK et al. Saliva microbiome in primary Sjögren's syndrome reveals distinct set of disease-associated microbes. *Oral Dis.* 26(2), 295–301 (2020).
22. Tong Y, Zheng L, Qing P et al. Oral microbiota perturbations are linked to high risk for rheumatoid arthritis. *Front. Cell. Infect. Microbiol.* 9 (2020). [www.frontiersin.org/articles/10.3389/fcimb.2019.00475/full?report=reader](http://www.frontiersin.org/articles/10.3389/fcimb.2019.00475/full?report=reader)
23. Brennan CA, Garrett WS. *Fusobacterium nucleatum* – symbiont, opportunist and oncobacterium. *Nat. Rev. Microbiol.* 17(3), 156–166 (2019).
24. Rubinstein MR, Baik JE, Lagana SM et al. *Fusobacterium nucleatum* promotes colorectal cancer by inducing Wnt/ $\beta$ -catenin modulator annexin A1. *EMBO Rep.* 20(4) (2019). [www.embopress.org/doi/abs/10.15252/embr.201847638](http://www.embopress.org/doi/abs/10.15252/embr.201847638)
25. Kitamoto S, Nagao-Kitamoto H, Jiao Y et al. The intermucosal connection between the mouth and gut in commensal pathobiont-driven colitis. *Cell* 182(2), 447–462.e14 (2020).
26. Cressatti M, Juwara L, Galindez JM et al. Salivary microR-153 and microR-223 levels as potential diagnostic biomarkers of idiopathic Parkinson's disease. *Mov. Disord.* 35(3), 468–477 (2020).
27. LaRocca D, Barns S, Hicks SD et al. Comparison of serum and saliva miRNAs for identification and characterization of mTBI in adult mixed martial arts fighters. *PLOS ONE* 14(1), e0207785 (2019).
28. Knight R, Urbanac A, Taylor BC et al. Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* 16(7), 410–422 (2018).
29. Poretsky R, Rodriguez-R LM, Luo C, Tsementzi D, Konstantinidis KT. Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. *PLOS ONE* 9(4), e93827 (2014).
30. Langille MGI, Zaneveld J, Caporaso JG et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* 31(9), 814–821 (2013).
31. Raymann K, Moeller AH, Goodman AL, Ochman H. Unexplored archaeal diversity in the great ape gut microbiome. *mSphere* 2(1) (2017). <https://msphere.asm.org/content/2/1/e00026-17>
32. Jølsen P, Turner KH, Gumus P, Nizam N, Buduneli N, Whiteley M. Metatranscriptomics of the human oral microbiome during health and disease. *mBio* 5(2) (2014). <https://mbio.asm.org/content/5/2/e01012-14>
- Outlines the importance of looking at microbial gene expression profiles in the context of human health and disease.
33. Koh A, Mannerås-Holm L, Yunn N-O et al. Microbial imidazole propionate affects responses to metformin through p38 $\gamma$ -dependent inhibitory AMPK phosphorylation. *Cell Metab.* 32(4), 643–653.e4 (2020).
34. Hidayat MFH, Milne T, Cullinan MP, Seymour GJ. Feasibility of the salivary transcriptome as a novel biomarker in determining disease susceptibility. *J. Periodontol Res.* 53(3), 369–377 (2018).
35. Bashardes S, Zilberman-Schapira G, Elina E. Use of metatranscriptomics in microbiome research. *Bioinform. Biol. Insights* 10, 19–25 (2016).
36. Gosalbes MJ, Durbán A, Pignatelli M et al. Metatranscriptomic approach to analyze the functional human gut microbiota. *PLOS ONE* 6(3), e17447 (2011).
37. Banavar G, Ogundijo O, Toma R et al. The salivary metatranscriptome as an accurate diagnostic indicator of oral cancer. *NPJ Genomics Med.* 6(1) (2021). [www.proquest.com/openview/a7a003c783b9dd7bb9a2451441c6d57d/1?pq-origsite=gscholar&cbl=2041923](http://www.proquest.com/openview/a7a003c783b9dd7bb9a2451441c6d57d/1?pq-origsite=gscholar&cbl=2041923)
- Demonstrates how both the method outlined in this publication and human/microbial gene expression profiles can be used for population-scale discovery of novel biomarkers related to oral cancer.
38. Kadosh E, Snir-Alkalay I, Venkatchalam A et al. The gut microbiome switches mutant p53 from tumour-suppressive to oncogenic. *Nature* 586(7827), 133–138 (2020).
39. Olsen I, Singhrao SK. *Porphyromonas gingivalis* infection may contribute to systemic and intracerebral amyloid-beta: implications for Alzheimer's disease onset. *Expert Rev. Anti. Infect. Ther.* 18(11), 1063–1066 (2020).
40. Knight R, Jansson J, Field D et al. Unlocking the potential of metagenomics through replicated experimental design. *Nat. Biotechnol.* 30(6), 513–520 (2012).
41. He S, Wurtzel O, Singh K et al. Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nat. Methods* 7(10), 807–812 (2010).

42. Hatch A, Horne J, Toma R *et al.* A robust metatranscriptomic technology for population-scale studies of diet, gut microbiome, and human health. *Int. J. Genomics* 2019, 1718741 (2019).
43. Toma R, Duval N, Pelle B *et al.* A clinically validated human capillary blood transcriptome test for global systems biology studies. *BioTechniques* 69(4), 289–301 (2020).
44. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* 39(1), 1–38 (1977).
45. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28(1), 27–30 (2000).
46. Martín-Fernández JA, Barceló-Vidal C, Pawlowsky-Glahn V. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Math. Geol.* 35(3), 253–278 (2003).
47. Aitchison J. The statistical analysis of compositional data. *J. R. Stat. Soc. Ser. B Methodol.* 44(2), 139–177 (1982).
48. de Souza MF, Kuasne H, Barros-Filho M de C *et al.* Circulating mRNAs and miRNAs as candidate markers for the diagnosis and prognosis of prostate cancer. *PLOS ONE* 12(9), e0184094 (2017).
49. Sullivan R, Heavey S, Graham DG *et al.* An optimised saliva collection method to produce high-yield, high-quality RNA for translational research. *PLOS ONE* 15(3), e0229791 (2020).