International Journal of Legal Medicine (2024) 138:401–412
https://doi.org/10.1007/s00414-023-03100-3

METHOD PAPER

# Spitting in the wind?—The challenges of RNA sequencing for biomarker discovery from saliva

Annica Gosch[1] · Regine Banemann[2] · Guro Dørum[3] · Cordula Haas[3] · Thorsten Hadrys[4] · Nadescha Haenggi[3] · Galina Kulstein[2] · Jacqueline Neubauer[3] · Cornelius Courts[1]

Received: 31 July 2023 / Accepted: 25 September 2023 / Published online: 17 October 2023

## Abstract

Forensic trace contextualization, i.e., assessing information beyond who deposited a biological stain, has become an issue of great and steadily growing importance in forensic genetic casework and research. The human transcriptome encodes a wide variety of information and thus has received increasing interest for the identification of biomarkers for different aspects of forensic trace contextualization over the past years. Massively parallel sequencing of reverse-transcribed RNA ("RNA sequencing") has emerged as the gold standard technology to characterize the transcriptome in its entirety and identify RNA markers showing significant expression differences not only between different forensically relevant body fluids but also within a single body fluid between forensically relevant conditions of interest. Here, we analyze the quality and composition of four RNA sequencing datasets (whole transcriptome as well as miRNA sequencing) from two different research projects (the RNAgE project and the TrACES project), aiming at identifying contextualizing forensic biomarker from the forensically relevant body fluid saliva. We describe and characterize challenges of RNA sequencing of saliva samples arising from the presence of oral bacteria, the heterogeneity of sample composition, and the confounding factor of degradation. Based on these observations, we formulate recommendations that might help to improve RNA biomarker discovery from the challenging but forensically relevant body fluid saliva.

**Keywords** Forensic RNA analysis · Saliva · Massive parallel sequencing

## Introduction

The main aim of forensic molecular biological analysis is the individualization of a trace, i.e., unequivocally linking a biological trace to its donor, which is commonly performed via DNA-based STR profiling. Apart from and complementary to that, the contextualization of traces has become an issue of great and steadily growing importance. If the donor of a trace is not contested in a criminal court case, it can be crucial in the reconstruction of the course of events to contextualize the trace, meaning to explain, based on physical evidence, by which activity, how long ago, at which time of day, as part of which body fluid or organ tissue, etc., and the trace in question has been deposited.

Because of the high and complex information content of the transcriptome [1] represented by its differential and dynamically changing composition, the analysis of RNA readily lends itself to the assessment of several forensic contextual aspects. Among these, the identification of body fluids via gene expression analysis [2] is routinely applied in forensic casework in different laboratories [3]. Besides, several research projects assess the potential of transcriptomic analysis for assessing further aspects of forensic relevance, such as time since trace deposition [4], post-mortem interval estimation [5], wound age [6], the biological age of the donor [7], and time of day of deposition [8]. Since 2009, also microRNA (miRNA), a small, non-coding regulatory type of RNA, about 18–25 nt in length, is being investigated in the context of its forensic potential [9] and ongoing research into

✉ Cornelius Courts
cornelius.courts@uk-koeln.de

[1] Institute of Legal Medicine, University Hospital of Cologne, Cologne, Germany

[2] Federal Criminal Police Office, Forensic Science Institute, Wiesbaden, Germany

[3] Zurich Institute of Forensic Medicine, University of Zurich, Zurich, Switzerland

[4] State Criminal Police Office, Forensic Science Institute, Munich, Germany

a wide array of forensic applications of miRNA analysis has covered a lot of ground up to this point [10, 11].

Frequently, the selection of mRNA and miRNA candidates whose differential expression and/or degradation state [12] informs on a particular aspect of forensic interest will be performed by perusing (not necessarily forensic) literature and testing previously identified markers for their informative value in the setup of interest (for example performed in [8, 13]). However, previously published markers may not always be ideal or even available to answer the question of forensic interest at hand, and thus, traditional statistical and machine learning algorithms based upon raw data generated by massively parallel sequencing (MPS) of whole transcriptomes [14] and miRNomes [15] are increasingly being used for unsupervised marker identification.

Saliva is a body fluid regularly encountered in several criminal contexts with cases of sexual assault (e.g., licked, bitten, or spat on body parts, oral rape) being of particular impact. Therefore, the reliable and sensitive detection, identification, and contextualization of saliva in (mixed) trace material is very desirable and can be crucial in the assessment of the trace's weight of evidence. Consequently, analysis of salivary RNA from forensic samples is well represented in the literature [16–20], with a particular focus on body fluid identification. Additionally, saliva is often included in marker identification studies for other trace contextualization aspects [4, 21].

Within this article, the authors—a European consortium of forensic RNA researchers working on different aspects of RNA-based trace contextualization—present RNA sequencing data from different sets of salivary samples and experimental setups and discuss the challenges associated with whole transcriptome sequencing of saliva samples.

## Material and methods

Evaluations are based on four different RNA sequencing datasets from saliva samples obtained in two different forensic genetic research projects: RNA-based Age Estimation (RNAgE) and Transcriptomic Analysis for the Contextualization of Evidential Stains (TrACES). Both research projects have the aim of identifying RNA markers for a forensically relevant aspect of trace contextualization. The first project aimed at correlating the transcriptome with the sample donor's age (hereafter referred to as "RNAgE" project), while the second project explored potential correlations of the transcriptome with the time of day of sample collection (hereafter referred to as "TrACES" project). In both research projects, donors provided samples of blood as well as saliva. Results for the transcriptomic analysis of blood samples are reported separately for the "RNAgE" (manuscript under preparation) and the "Traces" [22] projects.

In both projects, salivary transcriptomes were analyzed using both whole transcriptome (WT) and microRNA (miRNA) sequencing, resulting in a total of four datasets (Table 1).

Details on sample collection and sample processing are provided in Bioinformatic processing of datasets was performed as described in Table 2.

Table 1. All protocols were performed according to the manufacturers' instructions.

Bioinformatic processing of datasets was performed as described in Table 2.

## Results

In two different research projects, RNA sequencing was performed to identify biomarkers (human RNA transcripts or human miRNA) showing differential expression either between individuals of different ages ("RNAgE" project) or between samples deposited at different times of the day ("TrACES" project). An evaluation of the distribution of sequencing reads in the four different datasets from human saliva samples showed that only a low percentage of reads mapped to the actual RNA species of interest (Fig. 1). The percentage of human on-target reads was lowest in the "RNAgE-WT" dataset (average: 0.1%, range: 0.007–1.1%) and highest in the "TrACES-miRNA" dataset (average: 38.4%, range: 3.8–86.2%). For a better understanding of the reasons for these low on-target read count percentages, we performed an in-depth analysis of the read distribution of the four sequencing datasets (Fig. 1, Supplementary Tables 1 and 2).

### Non-human RNA

Taxonomic classification of reads revealed that the largest proportion of reads in both WT datasets is of bacterial origin (Fig. 1). Sequences from a diverse set of microbial species were detected (Supp. File 1 and 2), most of which are commonly encountered in the human oral cavity [39]. The percentage of bacterial reads was consistently > 75% in the "RNAgE-WT" dataset and thus accounted for the majority of reads in every sample of this dataset. The "TrACES-WT" dataset contained lower percentages of bacterial reads (average: 58%, range: 3–97%) with large variability between individuals as well as between samples from the same individual taken at different time points of the day (Supp. Table 1).

In both miRNA datasets, the percentage of microbial reads was lower compared to the whole transcriptome sequencing dataset from the corresponding research project (average < 30% in both datasets, Supp. Table 2).

### Human RNA

In the "RNAgE-WT" dataset, the majority of human reads were attributed to ribosomal RNA (rRNA) (average:

**Table 1** Sample collection and processing information for the four different saliva sample sets

| Sample set abbreviation | RNAgE—WT | TrACES—WT | RNAgE—miRNA | TrACES—miRNA |
|---|---|---|---|---|
| Aspect of contextualization | Donor age | Time of day | Donor age | Time of day |
| Targeted RNA type | Total RNA (excluding rRNA) | Total RNA (excluding rRNA) | miRNA | miRNA |
| Number of Samples and Individuals | 67 individuals (39 females, 28 males; 8–77 years); 1 sample per individual | 3 individuals (2 males, 1 female, 25–31 years), 8 samples per individual (8 time points: 8 AM, 11 AM, 2 PM, 5 PM, 8 PM, 11 PM, 2 AM, 5 AM); 24 samples in total | 85 individuals (56 females, 29 males; 0–96 years); 1 sample per individual | 10 individuals (5 males, 5 females; 19–31 years), 8 samples per individual (8 time points: 8 AM, 11 AM, 2 PM, 5 PM, 8 PM, 11 PM, 2 AM, 5 AM); 80 samples in total** |
| Sample collection | Cotton swab (Forensix sample collection system, ThermoFisher Scientific) soaked in liquid saliva* (after spitting) | Buccal mucosa collected on a cotton swab (Sarstedt)* | Cotton swab (Forensix sample collection system, ThermoFisher Scientific) soaked in liquid saliva* (after spitting) | Buccal mucosa collected on a cotton swab (Sarstedt)* |
| Sample stabilization/storage | Dried and stored at RT for up to 11 days | Stabilized in DNA/RNA shield stabilization solution (Zymo Research) and stored at – 80 °C | Dried and stored at RT for up to 11 days | Stabilized in DNA/RNA shield stabilization solution (Zymo Research) and stored at – 80 °C |
| RNA extraction | ReliaPrep RNA Miniprep System, (Promega) | mirVana miRNA extraction kit (total RNA protocol), (Thermo Fisher Scientific) | miRNeasy mini Kit (QIAGEN) | mirVana miRNA extraction kit (total RNA protocol), (Thermo Fisher Scientific) |
| DNAse treatment | TURBO DNase free Kit, Thermo Fisher Scientific) | TURBO DNase free Kit, Thermo Fisher Scientific) | TURBO DNase free Kit, Thermo Fisher Scientific) | TURBO DNase free Kit, Thermo Fisher Scientific) |
| RNA quantification | QuantiFluor RNA System, HS assay, (Promega) | QuantiFluor RNA System, HS assay, (Promega) | QuantiFluor RNA System, HS assay, (Promega) | QuantiFluor RNA System, HS assay, (Promega) |
| Quality control | Not assessed | RNA Pico Chip Kit on a 2100 Bioanalyzer Instrument (Agilent Technologies) | Not assessed | RNA Pico Chip Kit on a 2100 Bioanalyzer Instrument (Agilent Technologies) |
| Library preparation | TRIO RNA Seq Library Preparation Kit, no rRNA depletion (Tecan) | Stranded total RNA with Ribo-Zero Plus protocol, Ribo-Zero Plus rRNA depletion kit (Illumina), performed at the Competence Centre for Genomic Analyses in Kiel, Germany | NEBNext multiplex small RNA library Prep Set for Illumina (NEB) | NextFlex Small RNA library preparation kit (Bioo Scientific) performed at the Competence Centre for Genomic Analyses in Kiel, Germany |
| Sequencing | NovaSeq 6000 platform, Illumina, S1 flowcell (100 nt, single end, 15 Mio. Reads/sample), performed at the Functional Genomics Center Zurich (FGCZ), Switzerland | NovaSeq 6000 platform, Illumina, S4 flowcell, (2×50 bp, paired-end, 30 Mio. Reads/sample), performed at the Competence Centre for Genomic Analyses in Kiel, Germany | NextSeq 500 platform, Illumina, High output flowcell, (75 nt, single end, 13 Mio. Reads/sample), performed at the Functional Genomics Center Zurich (FGCZ), Switzerland | NovaSeq 6000 platform, Illumina, S1 flowcell, (2×50 bp, paired-end, 16 Mio. Reads/sample), performed at the Competence Centre for Genomic Analyses in Kiel, Germany |

*Donors refrained from eating and drinking (except water) for ½ h prior to sampling

**In the TrACES – miRNA dataset, library preparation failed for two samples, thus only 78 out of 80 samples were submitted to sequencing

**Table 2** Bioinformatic processing information of RNA sequencing data for the four different saliva sample sets. All processing steps for the WT datasets were performed on the free public European Galaxy server UseGalaxy.eu [23]

| Sample set abbreviation | RNAgE—WT | TrACES—WT | RNAgE—miRNA | TrACES—miRNA |
|---|---|---|---|---|
| Preprocessing | Trimmomatic (Adapter trimming, Removal of 5 bases from the start of the read, keeping reads of a minimal length of 20 nt and a minimum average quality of 10) [24] | Trim Galore! (Adapter trimming, Removal of 1 base from the start of the read, removal of low-quality ends from reads, keeping reads of a minimal length of 20 nt) [25] | sRNAbench (default settings for the NEBnext protocol) [26, 27] | sRNAbench (default settings for the Bioo Scientific Nextflex (v2,v3) protocol), processing of forward strand reads (R1) only [26, 27] |
| Mapping | STAR mapping algorithm, single-end reads, using the GenCode primary assembly reference genome and primary assembly comprehensive gene annotation (GRCh38. p13, Release 43) [28, 29] | STAR mapping algorithm, paired-end reads, using the GenCode primary assembly reference genome and primary assembly comprehensive gene annotation (GRCh38. p13, Release 43) [28, 29] | sRNAbench (default settings for the NEBnext protocol, annotation reference database: miRBase release 22.1, species: *Homo sapiens*) [26, 27, 30] | sRNAbench (default settings for the Bioo Scientific Nextflex (v2,v3) protocol, annotation reference database: miRBase release 22.1, species: *Homo sapiens*), processing of R1 reads only [26, 27, 30] |
| Assessment of ribosomal RNA (rRNA) content | SortMeRNA on trimmed reads, rRNA databases:<br>2.1b-rfam-5 s-database-id98<br>2.1b-silva-arc-23 s-id98<br>2.1b-silva-euk-28 s-id98<br>2.1b-silva-bac-23 s-id98<br>2.1b-silva-euk-18 s-id95<br>2.1b-silva-bac-16 s-id90<br>2.1b-rfam-5.8 s-database-id98<br>2.1b-silva-arc-16 s-id9 [31–33]* | SortMeRNA on trimmed reads, rRNA databases:<br>2.1b-rfam-5 s-database-id98<br>2.1b-silva-arc-23 s-id98<br>2.1b-silva-euk-28 s-id98<br>2.1b-silva-bac-23 s-id98<br>2.1b-silva-euk-18 s-id95<br>2.1b-silva-bac-16 s-id90<br>2.1b-rfam-5.8 s-database-id98<br>2.1b-silva-arc-16 s-id9<br>[31–33]* | sRNAbench (default settings for the NEBnext protocol)<br>[26, 27] | sRNAbench (default settings for the Bioo Scientific Nextflex (v2,v3) protocol), processing of R1 reads only<br>[26, 27] |
| Assessment of bacterial RNA content | Kraken2 on trimmed reads, Bracken, Krakentools, Krona [34–38] | Kraken2 on trimmed reads, Bracken, Krakentools, Krona [34–38] | sRNAblast on unassigned reads from sRNAbench analysis [26, 27] | sRNAblast on unassigned reads from sRNAbench analysis [26, 27] |
| Assessment of on-target read counts | "reads per gene" output from the STAR mapping algorithm [28] | "reads per gene" output from the STAR mapping algorithm [28] | sRNAbench (default settings for the NEBnext protocol), SA read counts provided in the mature_sense_SA output file [26, 27] | sRNAbench (default settings for the Bioo Scientific Nextflex (v2,v3) protocol), SA read counts provided in the mature_sense_SA output file [26, 27] |

*Reads mapping to human mitochondrial rRNA (ENSG00000210082.2 and ENSG00000211459.2) were additionally counted as rRNA reads and excluded from the on-target read count table

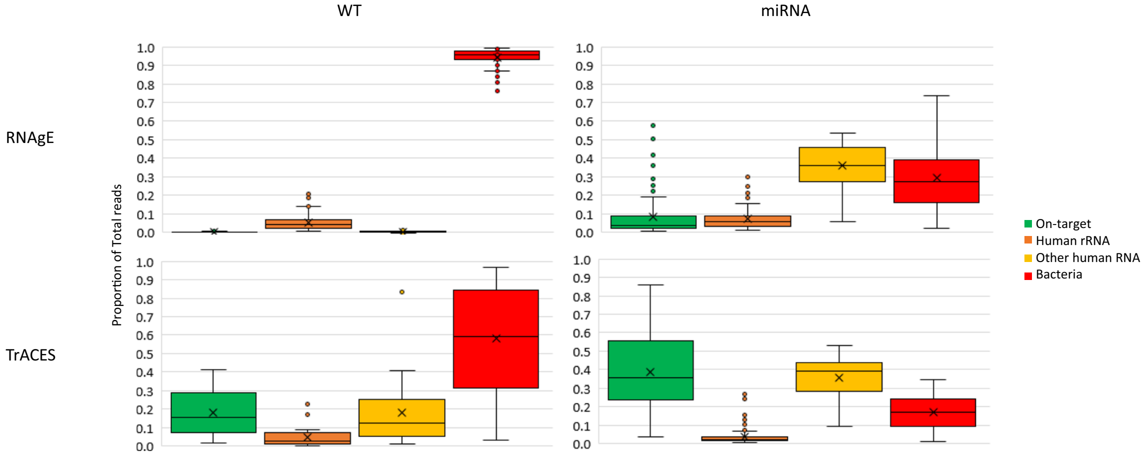

**Fig. 1** Distribution of sequencing reads in four RNA sequencing datasets from saliva samples. Boxplots *indicate the distribution of the proportion of sequencing reads over n = 67, 24, 85, 78 (RNAgE-WT, TrACES-WT, RNAgE-miRNA, and* TrACES-miRNA datasets respectively) samples. The boxplots indicate the median and interquartile ranges (IQR), whiskers indicate the minimum and maximum values (within 1.5*IQR)*, and outliers (> 1.5*IQR) are indicated by individual dots. Additionally, the average value is marked by a cross. WT*, whole transcriptome. "On-target" reads are defined as reads mapping to exonic regions of genes in the WT datasets (cf. Supp. Table 1) and reads mapping to miRbase (sense) in the miRNA datasets (cf. Supp. Table 2). "Other human RNA" reads are assigned to the human transcriptome but did not map to the target RNA type or human rRNA (i.e., ambiguously mapped reads, multi-mapping reads, intronic and intergenic reads for the WT datasets), and reads assigned to RNA types other than miRNA and rRNA (e.g., small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), mRNA, long-non coding RNA (lncRNA) in miRNA datasets)

94.1%, Fig. 1, Supp. Table 1). Thus, among the already low proportion of reads mapping to the human transcriptome, the proportion of reads mapping to regions of interest (i.e., genes) was also low (average: 2.5%, Supp. Table 1). In comparison, the percentage of reads mapping to rRNA was considerably reduced in the "TrACES" dataset (average: 11%, Supp. Table 1). In the "RNAgE" dataset, most of the ribosomal reads were from mitochondrial rRNA (average mitochondrial to nuclear rRNA ratio: 12.7, range: 1.5–83.0), whereas in the "TrACES" dataset, a higher proportion of nuclear rRNA reads was observed (average ratio: 0.1, range: 0.02–0.3, Supp. Table 1). In the "TrACES" dataset, the percentage of human reads mapping to regions of interest was on average 48.4% (Supp. Table 1). Besides the reads mapping to genes, a relevant proportion (average: 27%) also mapped to intronic or intergenic regions of the transcriptome (assigned as "no Feature" by the STAR mapping algorithm, Supp. Table 1).

In the miRNA sequencing datasets, rRNAs made up a relatively small proportion of the total reads of human origin (average of 13.6% and 4.7% in the "RNAgE" and "TrACES" datasets respectively, Supp. Table 2). However, significant proportions of human reads were attributed to other (small) RNA species, resulting in average "on target"-miRNA reads of 13.7 and 46.9% of total human reads in the "RNAgE" and "TrACES" datasets, respectively (Supp. Table 2).

## "Useful reads" for biomarker discovery

To apply algorithms for the identification of biomarkers using differential gene expression analysis, potential RNA markers must be reliability detected and quantified. Thus, markers with very low read counts are commonly excluded from datasets prior to the application of differential gene expression algorithms [40].

The number of RNA markers detected above a certain read count in each of the four sequencing datasets is plotted in Fig. 2. (Note: Data is shown as raw read counts in Fig. 2. For the purpose of performing differential gene expression analysis within datasets, these read counts would have to be normalized. However, as in this step we were interested in the number of markers whose expression can reliably be quantified (i.e., can be differentiated from noise), raw read counts were considered here.)

It is evident that in each of the four datasets, the number of RNA markers that can be reliably quantified (and thus may be eligible for biostatistical analyses for marker discovery) is very limited. In the "RNAgE-WT" dataset, an average of 29 markers (range: 0–264) exceeded an absolute read count threshold of 100. In the "TrACES –WT" dataset, a higher number of markers reached this threshold (average: 1731, range: 130–4020); however, this number remains low compared to a whole transcriptome

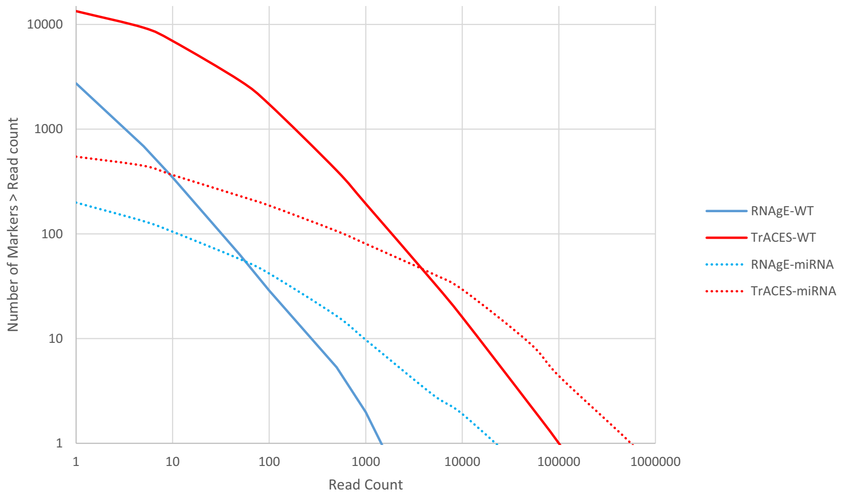

**Fig. 2** Number of markers above a read count in each of the four RNA sequencing datasets. The plot indicates average values over $n=67$, 24, 85, s78 (RNAgE-WT, TrACES-WT, RNAgE-miRNA, and TrACES-miRNA datasets respectively) samples

sequencing dataset of human whole blood that had been processed in a similar way in the “TrACES” project (average: 10,669, range: 8602–16,483 in a total of 80 samples from ten individuals, data not shown).

The number of miRNA markers reaching an absolute read count threshold of 100 was 42 (4–156) and 186 (49–273) in the “RNAgE” and “TrACES” datasets, respectively.

## Discussion

We analyzed four different RNA sequencing datasets from saliva samples originating from two different research projects. In each of these datasets, data analysis proved to be challenging due to a large heterogeneity between samples and consistently low percentages of read counts aligning with the RNA targets of interest. Nevertheless, differences were observed between datasets that were analyzed under different conditions, indicating that some of the challenges associated with RNA sequencing of saliva samples may be addressed by adjusting the analysis procedures. It needs to be emphasized that this study was not designed to systematically assess how individual aspects within each of the different workflows impacted the total outcome. Nonetheless, the observations reported herein allow for some conclusions that merit consideration for future biomarker discovery studies for forensic (and non-forensic) purposes.

Based on the in-depth analysis of four RNA sequencing datasets (whole transcriptome and miRNome) from two different forensic research projects, we identified three factors contributing to the challenges of RNA sequencing from saliva samples.

### Non-human RNA content

Whole transcriptome analyses showed that a large proportion of the RNA present in the salivary samples was of bacterial origin, which is consistent with previous studies [41–43]. For example, Ostheim et al. describe that the bacterial RNA content in human saliva is on average 1145 times higher than the human RNA content (based on 18S/16S rRNA ratio measurements) [42].

In comparison to the “RNAgE-WT” dataset, the bacterial RNA content was reduced in the “TrACES” dataset. The stabilization of buccal mucosa samples right after collection (possibly preventing bacterial growth), as well as the use of the Ribo-Zero Plus rRNA depletion kit (depleting not only human but also gram-negative (*Escherichia coli*) and gram-positive (*Bacillus subtilis*) bacterial 5S, 16S, and 23S rRNA sequences), might have contributed to the reduction of bacterial reads in this dataset.

As bacteria are naturally present in human saliva, it is hardly possible to select and sequence only the human component of the salivary transcriptome. Biomedical studies often analyze cell-free saliva (obtained by centrifugation) rather than whole saliva, as it has been shown to contain a lower proportion of bacterial RNA than whole saliva [43]. However, this approach would not be applicable to forensic saliva stains as these are usually dried (e.g., on surfaces at a crime scene) hindering a clear separation into cellular and cell-free fractions.

In another previous study, the authors observed higher percentages of reads mapping to the human transcriptome when enriching for poly-A-tailed RNAs (rather than depleting rRNAs), as polyadenylation of mRNAs is unique to eukaryotic cells [44]. However, the authors also remark that this approach restricts biomarker discovery to polyadenylated mRNA, whereas non-polyadenylated transcripts (such as non-coding RNAs) will not be detectable [44]. Additionally, it has been experimentally proven that more comprehensive and reliable results can be obtained from low-quality/degraded samples with rRNA-depletion-based library preparation methods as compared to oligo(dT)-enrichment-based methods [45], which is why this approach is usually recommended for samples expected to show degradation to some extent (cf. "Heterogeneity of sample composition").

Thus, the presence of oral bacteria in saliva has to be accepted and needs to be accounted for when processing salivary samples: Based on our observations, we recommend adjusting the sequencing depth to account for the high percentage of reads expected to map to bacterial rather than human RNA. Additionally, an adjustment of bioinformatics processing workflows was recommended by Kaczor-Urbanowicz et al. Reads mapping to bacterial genomes should be filtered out in a first bioinformatics processing step, prior to analyzing the reads mapping to the human transcriptome [46].

Alternatively, the oral microbiome could represent a target for the discovery of (forensic) biomarkers: In biomedical studies, the composition of the oral microbiome has been associated with a number of oral as well as non-oral diseases (including periodontitis, cardiovascular disease, and pneumonia, e.g., summarized in [47, 48]). In a forensic context, microbial signatures have very early been suggested for analysis in addition to human transcripts for the differentiation of forensically relevant body fluids [49–51]. More recently, it has also been shown that changes in the composition of microbial transcripts could be used as a biomarker to analyze time since deposition of forensically relevant body fluids (including saliva) [41]. Hence, the salivary microbiome could be eligible for biomarker discovery for forensic trace contextualization by considering not only its composition on a species-level but also individual differentially expressed bacterial transcripts.

A second alternative solution would be to target the miRNome rather than the whole transcriptome. miRNAs are small, regulatory RNAs present only in eukaryotes and thus naturally absent from bacterial transcriptomes [52]. Indeed, in both research projects, the percentage of reads assigned to the human transcriptome was higher in the miRNA sequencing datasets as compared to the WT sequencing datasets.

Nonetheless, bacterial reads were still present in the majority of the samples. In both library preparation procedures performed in this study, small RNAs are enriched by selecting RNA molecules of a defined fragment length (corresponding to the combined length of small RNAs of interest and adjoined adapters). Thus, fragmented bacterial RNA of the same length will also be included in the resulting small RNA sequencing libraries.

As described for whole transcriptome sequencing of saliva samples, the issue of bacterial reads in miRNA sequencing can be addressed and might (at least partially) be resolved by the choice of sample preparation protocol, adjustment of sequencing depth to account for off-target reads as well as modified bioinformatic processing protocols [46].

## Complexity of the human transcriptome

In the "RNAgE" WT dataset, the majority of reads mapping to the human transcriptome was determined to be rRNA. rRNAs are known to make up a considerable proportion (≥ 80%) of the human transcriptome [45, 53].

Therefore, it is common to perform some sort of enrichment of target RNA species (either enrichment of poly-A-tailed (mostly) mRNA ("poly-A-enrichment") or selective depletion of rRNA ("rRNA depletion") [54]. When deciding on an RNA sequencing strategy for the "RNAgE" project, it was taken into consideration that for the TRIO RNA Seq Library Preparation kit, the rRNA depletion step had previously been observed to have a negative influence on the sequencing quality and on downstream analyses [55]. Additionally, studies suggest that rRNAs may carry age-relevant information [56, 57]. Thus, the rRNA depletion step was omitted in the "RNAgE" project. Therefore, it may plausibly be assumed that the relevant reduction of rRNA reads in the "TrACES" dataset as compared to the "RNAgE" dataset can be explained by the inclusion of the rRNA depletion step during library preparation of samples in this study.

Notably, we observed a large proportion of rRNA reads mapping to mitochondrial RNA in the "RNAgE" dataset. As mitochondria have evolutionary originated from incorporated prokaryotes [58], it may be hypothesized that a proportion of these reads was of true bacterial origin and incorrectly mapped to the human mitochondrial rRNA (mt-rRNA). However, as the absolute reads mapped to mt-rRNA reads did not relevantly decrease when STAR mapping was performed on reads not aligned to rRNA in the SortMeRNA step (data not shown), a true human mitochondrial origin is more likely. In single-cell-RNA sequencing, high proportions of mitochondrial reads are considered indicative of damaged cells (and corresponding cells are usually excluded from analysis) [59, 60]. Our observations might thus suggest a high proportion of damaged cells in human saliva (which is to be expected). However, to the best of our knowledge, no other studies report similarly high proportions of mt-rRNA in whole transcriptome sequencing datasets from saliva (available studies either performed rRNA depletion or did not specifically report rRNA read proportions [43, 46, 61, 62]). The association between high mt-rRNA reads and cell damage in whole transcriptome sequencing datasets thus remains speculative and would have to be experimentally assessed in future studies.

Besides exonic reads, relevant percentages of reads mapping to intronic and intergenic regions were observed in both WT datasets. It has previously been reported that high percentages of intronic and intergenic reads (e.g., representing non-coding transcripts or nascent mRNAs) are commonly seen in rRNA-depleted whole transcriptome sequencing libraries (as opposed to poly-A-enriched sequencing libraries in which all transcripts lacking poly-A-tails would be excluded) [63, 64]. In comparison to poly-A-enriched libraries, rRNA depleted libraries capture information encoded in the entirety of the transcriptome and therefore require a higher sequencing depth to achieve a similar exonic coverage [63].

Compared to the WT datasets, miRNA sequencing resulted in higher percentages of on-target read counts. The higher proportion of human reads mapping to miRNAs in the "TrACES" dataset compared to the "RNAgE" dataset might be attributed to the differences in sampling procedures (liquid saliva dried on cotton swabs vs. stabilized buccal swabs). Sullivan et al. observed significant differences in miRNome compositions between whole saliva samples stored with and without RNA stabilizer (with significantly higher miRNA read counts for samples stored in stabilizer). Moreover, they found differences attributable to the collection method (with significantly higher miRNA read counts for samples collected by swabbing compared to samples collected by expectoration), and highlight the relevance of consistent sample collection and storage procedures within studies [65].

Besides, the higher proportion of miRNAs reads in the "TrACES" dataset might also be caused by the use of a different library preparation procedure. Comparative evaluations of small RNA library preparation procedures have repeatedly reported higher proportions of miRNA reads and larger numbers of miRNAs detected when using the Bioo Scientific NextFlex Small RNA-Seq library preparation kit as compared to the NEBNext Small RNA Library Prep Kit [66, 67].

miRNA sequencing libraries also showed a large complexity comprising a variety of different RNA types including mRNAs, rRNAs, lncRNAs, snRNAs, snoRNAs, and other small RNAs (not specifically targeted by the miRNA mapping algorithm applied in this study [26, 27]). Types of small RNA other than miRNA, e.g., PIWI-interacting RNA (piRNA), snRNA, and snoRNA have already been reported as biomarkers in previous forensic studies [68, 69] and could be possibly included in future marker identification studies as well.

## Heterogeneity of sample composition

The transcriptomic composition in the datasets from two different studies showed a large heterogeneity, both between as well as within studies. Differences between sample sets partially arise from different technologies (as discussed above) but can also be attributable to the different sample types (liquid saliva samples in the "RNAgE" and buccal swab samples in the "TrACES" dataset):

In forensic as well as biomedical studies, both liquid saliva and swabbed samples of the buccal mucosa ("buccal swabs") are used to study the body fluid "saliva." However, the two sample types have been shown to possess a markedly different biological composition [70]. Liquid saliva is a complex mixture of fluids secreted from various glands in the oral cavity. It is composed of > 99% water with a pH between 6 and 7 under normal conditions, and contains a large variety of electrolytes as well as macromolecules, such as mucins, enzymes, and immunoglobulins [71]. The cellular content of liquid saliva is low and mainly consists of leukocytes, erythrocytes, and epithelial cells shed from the oral mucosa [72]. A study analyzing microscopy slides with saliva observed that 47.3% (± 6.2) of the cells found in liquid saliva from adult study participants were of epithelial origin, compared to 83.4% (± 6.8) in the same participants' buccal swabs [70]. As cellular heterogeneity impacts heterogeneity in the transcriptome, it is not recommended to mix liquid saliva and buccal swab samples in biomarker discovery studies.

In forensic casework, salivary samples may be recovered in a variety of different contexts, deriving from kissing, licking, spitting, drooling, chewing, biting, or speaking. Both liquid saliva and buccal swabs may represent only a subset of these forensically relevant salivary sample types [73] and it thus remains open for discussion which sample type is best suited for forensic biomarker discovery studies.

By microscopic evaluation, both liquid saliva and buccal swab samples showed large inter-individual differences in cell type composition (with liquid saliva showing a higher variability than buccal swabs) [70]. This is consistent with larger inter- as well as intra-individual variability observed in the datasets presented in this study.

Apart from the previously described cellular RNA, cell-free RNA (cfRNA) has been reported to contribute to the transcriptome of human saliva [74, 75]. cfRNAs may originate from dead or damaged cells [69, 71], but have also been shown to reside within exosomes secreted by cells in human saliva [76]. For miRNAs, it has even been suggested that exosomes are the main source of this RNA species in human saliva [75].

The heterogeneity of the salivary transcriptome's composition is further increased by degradation. Previous studies have reported both full-length and partially degraded RNA molecules in human saliva [76, 77]. Studies analyzing the time-wise stability of the transcriptome in different body fluids observed that the salivary transcriptome showed low integrity even right after sample deposition, and salivary transcripts degraded more rapidly compared to other body fluids (blood, semen, and vaginal secretions) [4, 12, 78]. Conditions in the oral cavity (warmth, moisture, presence of ribonucleases) are assumed to promote salivary RNA degradation, but it has also been shown that exogenously introduced mRNA degrades more rapidly under these conditions than endogenous salivary RNA, suggesting that the salivary RNA might be (partially) protected [77].

Previous studies have measured the extent of overall as well as transcript-wise degradation based on the assessment of the sequencing coverage along the transcript, with degraded transcripts showing increased coverage at their 3' ends [79]. However, this approach enables quantification of the extent of degradation only for sequencing libraries that have been prepared using the poly-A-enrichment strategy, as it selects the 3'-fragments of transcripts, whereas a 3' bias is not expected to be observed for degraded samples after selective depletion of rRNAs [80].

While we are thus unable to exactly quantify and compare the amount of degradation in the samples analyzed in this study, it may reasonably be assumed based on observations from previous studies that RNA from human saliva is degraded to a certain extent, and that the stochastic phenomenon of degradation increases heterogeneity within sample sets.

Due to their short length, miRNAs are assumed to be less prone to degradation, and previous studies have indeed observed these small RNAs to be more stable than mRNA transcripts [81–83]. Hence, despite our lack of knowledge of the true extent of miRNA degradation in our datasets, there is sufficient ground to assume that the increased quality and on-target read counts in our miRNA compared to the WT datasets may partly be attributable to the lower impact of degradation on miRNAs as compared to longer transcripts.

## Conclusion

In summary, our analysis of four different RNA sequencing datasets from two different forensic research projects indicates that biomarker discovery from saliva samples through RNA sequencing is challenging. This can be attributed to a multitude of factors that decrease the proportion of sequencing reads suitable for biomarker discovery, and at the same time increase between-sample heterogeneity. This includes the presence of oral bacteria, the heterogeneity of cellular and cell-free RNA, and the confounding factor of degradation.

It is important to note that our study was not explicitly designed to assess the impact of individual impact of factors such as sample collection procedures, stabilization reagent, or library preparation methods. As a result specific recommendations for an optimal sample processing protocol cannot be given based on our observations. However, our results in combination with discussed outcomes of previous studies may be helpful to inform decisions and designs for future biomarker discovery studies.

To address the issue of low read counts for the RNAs of interest, we recommend adjusting the sequencing depth or undertaking measures to enrich for the RNA type of interest. This may include the control for microbial growth by sample stabilization after collection, selective depletion of (bacterial and) human rRNA, or selective enrichment of poly-A-tailed RNA.

Besides human mRNA, other biomarkers may be more suitable for saliva samples: The high bacterial load might be exploited in metagenomic/metatranscriptomic analyses. Alternatively, small RNAs, which are less prone to degradation and whose sequencing results are less impacted by the presence of bacteria, might also be promising targets for biomarker discovery studies.

In conclusion, careful experimental planning that should account for the challenges associated with the important, but difficult-to-tackle body fluid “saliva,” and adjusting for individual research aims, will be necessary to successfully parse the salivary transcriptome for biomarkers that can potentially contextualize forensically relevant saliva stains.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s00414-023-03100-3.

**Acknowledgements** We would like to thank all anonymous sample donors for taking part in this study!

**Funding** Open Access funding enabled and organized by Projekt DEAL. We thank the Deutsche Forschungsgemeinschaft (DFG) for funding part of this project (CO-992/10–1, “TrACES” study). This part of the study work was additionally supported by the DFG Research Infrastructure NGS_CC (project 407495230) as part of the Next Generation Sequencing Competence Network (project 423957469).

We also thank the European Union Internal Security Fund for supporting part of this project (Grant Agreement Number IZ25-5793–2019-44, “RNAgE” study).

### Declarations

In the “RNAgE” project, all sample donors provided written informed consent. The study protocols were reviewed and approved by the CEBES review board at the University of Zurich (case number 2021-11e).

In the “TrACES” project, all sample donors provided written informed consent. The study protocols were reviewed and approved by the ethics committee of the University Hospital Schleswig–Holstein.

**Competing interests** The authors declare no competing interests.

**Role of the funding source** The funders of the study had no role in the study design, data collection, analysis, interpretation, or writing of the manuscript.

## References

1. Frith MC, Pheasant M, Mattick JS (2005) The amazing complexity of the human transcriptome. Eur J Hum Genet 13:894–897. https://doi.org/10.1038/sj.ejhg.5201459
2. Lynch C, Fleming R (2020) RNA -based approaches for body fluid identification in forensic science. WIREs Forensic Sci. https://doi.org/10.1002/wfs2.1407
3. Salzmann AP, Bamberg M, Courts C, Dørum G, Gosch A, Hadrys T, Hadzic G, Neis M, Schneider PM, Sijen T, den van Berge M, Wiegand P, Haas C (2021) mRNA profiling of mock casework samples: results of a FoRNAP collaborative exercise, Forensic Sci Int Genet 50. https://doi.org/10.1016/j.fsigen.2020.102409
4. Salzmann AP, Russo G, Kreutzer S, Haas C (2021) Degradation of human mRNA transcripts over time as an indicator of the time since deposition (TsD) in biological crime scene traces. Forensic Sci Int Genet 53:102524. https://doi.org/10.1016/j.fsigen.2021.102524
5. Scrivano S, Sanavio M, Tozzo P, Caenazzo L (2019) Analysis of RNA in the estimation of post-mortem interval: a review of current evidence. Int J Legal Med 133:1629–1640. https://doi.org/10.1007/s00414-019-02125-x
6. Hassan Gaballah M, Fukuta M, Maeno Y, Seko-Nakamura Y, Monma-Ohtaki J, Shibata Y, Kato H, Aoki Y, Takamiya M (2016) Simultaneous time course analysis of multiple markers based on DNA microarray in incised wound in skeletal muscle for wound aging. Forensic Sci Int 266:357–368. https://doi.org/10.1016/j.forsciint.2016.06.027
7. Zubakov D, Liu F, Kokmeijer I, Choi Y, van Meurs JBJ, van IJcken WFJ, Uitterlinden AG, Hofman A, Broer L, van Duijn CM, Lewin J, Kayser M (2016) Human age estimation from blood using mRNA, DNA methylation, DNA rearrangement, and telomere length, Forensic Sci. Int. Genet. 24:33–43. https://doi.org/10.1016/j.fsigen.2016.05.014
8. Lech K, Liu F, Ackermann K, Revell VL, Lao O, Skene DJ, Kayser M (2016) Evaluation of mRNA markers for estimating blood deposition time: towards alibi testing from human forensic stains with rhythmic biomarkers. Forensic Sci Int Genet 21:119–125. https://doi.org/10.1016/j.fsigen.2015.12.008
9. Hanson EK, Lubenow H, Ballantyne J (2009) Identification of forensically relevant body fluids using a panel of differentially expressed microRNAs. Anal Biochem 387:303–314. https://doi.org/10.1016/j.ab.2009.01.037
10. Glynn CL (2020) Potential applications of microRNA profiling to forensic investigations. RNA 26(1):1–9. https://doi.org/10.1261/rna.072173.119
11. Rocchi A, Chiti E, Maiese A, Turillazzi E, Spinetti I (2020) MicroRNAs: an update of applications in forensic science. Diagnostics (Basel) 11. https://doi.org/10.3390/diagnostics11010032
12. Lin MH, Jones DF, Fleming R (2015) Transcriptomic analysis of degraded forensic body fluids. Forensic Sci Int Genet 17:35–42
13. Alshehhi S, Haddrill PR (2019) Estimating time since deposition using quantification of RNA degradation in body fluid-specific markers. Forensic Sci Int 298:58–63. https://doi.org/10.1016/j.forsciint.2019.02.046
14. Haas C, Neubauer J, Salzmann AP, Hanson E, Ballantyne J (2021) Forensic transcriptome analysis using massively parallel sequencing. Forensic Sci Int Genet 52:102486. https://doi.org/10.1016/j.fsigen.2021.102486
15. Dørum G, Ingold S, Hanson E, Ballantyne J, Russo G, Aluri S, Snipen L, Haas C (2019) Predicting the origin of stains from whole miRNome massively parallel sequencing data. Forensic Sci Int Genet 40:131–139. https://doi.org/10.1016/j.fsigen.2019.02.015
16. Dørum G, Ingold S, Hanson E, Ballantyne J, Snipen L, Haas C (2018) Predicting the origin of stains from next generation sequencing mRNA data. Forensic Sci Int Genet 34:37–48. https://doi.org/10.1016/j.fsigen.2018.01.001
17. Haas C, Hanson E, Anjos MJ, Banemann R, Berti A, Borges E, Carracedo A, Carvalho M, Courts C, de Cock G, Dötsch M, Flynn S, Gomes I, Hollard C, Hjort B, Hoff-Olsen P, Hríbiková K, Lindenbergh A, Ludes B, Maroñas O, McCallum N, Moore D, Morling N, Niederstätter H, Noel F, Parson W, Popielarz C, Rapone C, Roeder AD, Ruiz Y, Sauer E, Schneider PM, Sijen T, Court DS, Sviežená B, Turanská M, Vidaki A, Zatkalíková L, Ballantyne J (2013) RNA/DNA co-analysis from human saliva and semen stains–results of a third collaborative EDNAP exercise. Forensic Sci Int Genet 7:230–239. https://doi.org/10.1016/j.fsigen.2012.10.011
18. Lindenbergh A, de Pagter M, Ramdayal G, Visser M, Zubakov D, Kayser M, Sijen T (2012) A multiplex (m)RNA-profiling system for the forensic identification of body fluids and contact traces. Forensic Sci Int Genet 6:565–577. https://doi.org/10.1016/j.fsigen.2012.01.009
19. Sauer E, Reinke A-K, Courts C (2016) Differentiation of five body fluids from forensic samples by expression analysis of four microRNAs using quantitative PCR. Forensic Sci Int Genet 22:89–99. https://doi.org/10.1016/j.fsigen.2016.01.018
20. Sakurada K, Watanabe K, Akutsu T (2020) Current methods for body fluid identification related to sexual crime: focusing on saliva, semen, and vaginal fluid, diagnostics (Basel) 10. https://doi.org/10.3390/diagnostics10090693
21. DíezLópez C, Kayser M, Vidaki A (2021) Estimating the time since deposition of saliva stains with a targeted bacterial DNA approach: a proof-of-principle study. Front Microbiol 12:647933. https://doi.org/10.3389/fmicb.2021.647933
22. Gosch A, Bhardwaj A, Courts C (2023) TrACES of time: transcriptomic analyses for the contextualization of evidential stains – identification of RNA markers for estimating time-of-day of bloodstain deposition. Forensic Sci Int: Genet 102915. https://doi.org/10.1016/j.fsigen.2023.102915
23. The Galaxy Community (2022) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. Nucleic Acids Res. 50:W345-51. https://doi.org/10.1093/nar/gkac247
24. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170
25. Krüger F (n.d.) Trim Galore. https://github.com/FelixKrueger/TrimGalore. Accessed 8 Sept 2023
26. Aparicio-Puerta E, Lebrón R, Rueda A, Gómez-Martín C, Giannoukakos S, Jaspez D, Medina JM, Zubkovic A, Jurak I, Fromm B, Marchal JA, Oliver J, Hackenberg M (2019) sRNAbench and sRNAtoolbox 2019: intuitive fast small RNA profiling and differential expression. Nucleic Acids Res. 47:W530–W535. https://doi.org/10.1093/nar/gkz415
27. Aparicio-Puerta E, Gómez-Martín C, Giannoukakos S, Medina JM, Scheepbouwer C, García-Moreno A, Carmona-Saez P, Fromm B, Pegtel M, Keller A, Marchal JA, Hackenberg M (2022) sRNAbench and sRNAtoolbox 2022 update: accurate miRNA and sncRNA profiling for model and non-model organisms. Nucleic Acids Res. 50(2022):W710–W717. https://doi.org/10.1093/nar/gkac363

28. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21. https://doi.org/10.1093/bioinformatics/bts635
29. Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, Sisu C, Wright JC, Armstrong J, Barnes I, Berry A, Bignell A, Boix C, Carbonell Sala S, Cunningham F, Di Domenico T, Donaldson S, Fiddes IT, GarcíaGirón C, Gonzalez JM, Grego T, Hardy M, Hourlier T, Howe KL, Hunt T, Izuogu OG, Johnson R, Martin FJ, Martínez L, Mohanan S, Muir P, Navarro FCP, Parker A, Pei B, Pozo F, Riera FC, Ruffier M, Schmitt BM, Stapleton E, Suner M-M, Sycheva I, Uszczynska-Ratajczak B, Wolf MY, Xu J, Yang YT, Yates A, Zerbino D, Zhang Y, Choudhary JS, Gerstein M, Guigó R, Hubbard TJP, Kellis M, Paten B, Tress ML, Flicek P (2021) GENCODE 2021. Nucleic Acids Res. 49:D916–D923. https://doi.org/10.1093/nar/gkaa1087
30. Kozomara A, Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 39:D152–D157. https://doi.org/10.1093/nar/gkq1027
31. Kopylova E, Noé L, Touzet H (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28:3211–3217. https://doi.org/10.1093/bioinformatics/bts611
32. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590–D596. https://doi.org/10.1093/nar/gks1219
33 Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, Eddy SR, Gardner PP, Bateman A (2013) Rfam 11.0: 10 years of RNA families. Nucleic Acids Res 41:D226-32. https://doi.org/10.1093/nar/gks1005
34. Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15. https://doi.org/10.1186/gb-2014-15-3-r46
35. Lu J, Breitwieser FP, Thielen P, Salzberg SL (2017) Bracken: estimating species abundance in metagenomics data. PeerJ Computer Science 3:e104. https://doi.org/10.7717/peerj-cs.104
36. Ondov BD, Bergman NH, Phillippy AM (2011) Interactive metagenomic visualization in a Web browser. BMC Bioinformatics 12. https://doi.org/10.1186/1471-2105-12-385
37. Cuccuru G, Orsini M, Pinna A, Sbardellati A, Soranzo N, Travaglione A, Uva P, Zanetti G, Fotia G (2014) Orione, a web-based framework for NGS analysis in microbiology. Bioinformatics 30:1928–1929. https://doi.org/10.1093/bioinformatics/btu135
38. Lu J, Rincon N, Wood DE, Breitwieser FP, Pockrandt C, Langmead B, Salzberg SL, Steinegger M (2022) Metagenome analysis using the Kraken software suite. Nat Protoc 17:2815–2839. https://doi.org/10.1038/s41596-022-00738-y
39. Dewhirst FE, Chen T, Izard J, Paster BJ, Tanner ACR, Yu W-H, Lakshmanan A, Wade WG (2010) The human oral microbiome. J Bacteriol 192:5002–5017. https://doi.org/10.1128/JB.00542-10
40. Koch CM, Chiu SF, Akbarpour M, Bharat A, Ridge KM, Bartom ET, Winter DR (2018) A beginner's guide to analysis of RNA sequencing data. Am J Respir Cell Mol Biol 59:145–157. https://doi.org/10.1165/rcmb.2017-0430TR
41. Salzmann AP, Arora N, Russo G, Kreutzer S, Snipen L, Haas C (2021) Assessing time dependent changes in microbial composition of biological crime scene traces using microbial RNA markers. Forensic Sci Int Genet 53:102537. https://doi.org/10.1016/j.fsigen.2021.102537
42. Ostheim P, Tichý A, Sirak I, Davidkova M, Stastna MM, Kultova G, Paunesku T, Woloschak G, Majewski M, Port M, Abend M (2020) Overcoming challenges in human saliva gene expression measurements. Sci Rep 10:11147. https://doi.org/10.1038/s41598-020-67825-6
43. Spielmann N, Ilsley D, Gu J, Lea K, Brockman J, Heater S, Setterquist R, Wong DTW (2012) The human salivary RNA transcriptome revealed by massively parallel sequencing. Clin Chem 58:1314–1321. https://doi.org/10.1373/clinchem.2011.176941
44. Yen E, Kaneko-Tarui T, Maron JL (2020) Technical considerations and protocol optimization for neonatal salivary biomarker discovery and analysis. Front Pediatr 8:618553. https://doi.org/10.3389/fped.2020.618553
45. Adiconis X, Borges-Rivera D, Satija R, DeLuca DS, Busby MA, Berlin AM, Sivachenko A, Thompson DA, Wysoker A, Fennell T, Gnirke A, Pochet N, Regev A, Levin JZ (2013) Comparative analysis of RNA sequencing methods for degraded or low-input samples. Nat Methods 10:623–629. https://doi.org/10.1038/nmeth.2483
46. Kaczor-Urbanowicz KE, Kim Y, Li F, Galeev T, Kitchen RR, Gerstein M, Koyano K, Jeong S-H, Wang X, Elashoff D, Kang SY, Kim SM, Kim K, Kim S, Chia D, Xiao X, Rozowsky J, Wong DTW (2018) Novel approaches for bioinformatic analysis of salivary RNA sequencing data for development. Bioinformatics 34:1–8. https://doi.org/10.1093/bioinformatics/btx504
47. Krishnan K, Chen T, Paster BJ (2017) A practical guide to the oral microbiome and its relation to health and disease. Oral Dis 23:276–286. https://doi.org/10.1111/odi.12509
48. Verma D, Garg PK, Dubey AK (2018) Insights into the human oral microbiome. Arch Microbiol 200:525–540. https://doi.org/10.1007/s00203-018-1505-3
49. Fleming RI, Harbison S (2010) The use of bacteria for the identification of vaginal secretions. Forensic Sci Int Genet 4:311–315. https://doi.org/10.1016/j.fsigen.2009.11.008
50. Nakanishi H, Kido A, Ohmori T, Takada A, Hara M, Adachi N, Saito K (2009) A novel method for the identification of saliva by detecting oral streptococci using PCR. Forensic Sci Int 183:20–23. https://doi.org/10.1016/j.forsciint.2008.10.003
51. Donaldson AE, Taylor MC, Cordiner SJ, Lamont IL (2010) Using oral microbial DNA analysis to identify expirated bloodspatter. Int J Legal Med 124:569–576. https://doi.org/10.1007/s00414-010-0426-8
52. Carthew RW, Sontheimer EJ (2009) Origins and mechanisms of miRNAs and siRNAs. Cell 136:642–655. https://doi.org/10.1016/j.cell.2009.01.035
53. Westermann AJ, Gorski SA, Vogel J (2012) Dual RNA-seq of pathogen and host. Nat Rev Microbiol 10:618–630. https://doi.org/10.1038/nrmicro2852
54. van Dijk EL, Jaszczyszyn Y, Thermes C (2014) Library preparation methods for next-generation sequencing: tone down the bias. Exp Cell Res 322:12–20. https://doi.org/10.1016/j.yexcr.2014.01.008
55. Salzmann AP, Russo G, Aluri S, Haas C (2019) Transcription and microbial profiling of body fluids using a massively parallel sequencing approach. Forensic Sci Int Genet 43:102149. https://doi.org/10.1016/j.fsigen.2019.102149
56. Wang M, Lemos B (2019) Ribosomal DNA harbors an evolutionarily conserved clock of biological aging. Genome Res 29:325–333. https://doi.org/10.1101/gr.241745.118
57. Ganley ARD, Kobayashi T (2014) Ribosomal DNA and cellular senescence: new evidence supporting the connection between rDNA and aging. FEMS Yeast Res 14:49–59. https://doi.org/10.1111/1567-1364.12133
58. Taanman J-W (n.d.) The mitochondrial genome: structure, transcription, translation and replication, Biochem. Pharmacol
59. Galow A-M, Kussauer S, Wolfien M, Brunner RM, Goldammer T, David R, Hoeflich A (2021) Quality control in scRNA-Seq can discriminate pacemaker cells: the mtRNA bias. Cell Mol Life Sci 78:6585–6592. https://doi.org/10.1007/s00018-021-03916-5

60. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, Teichmann SA (2016) Classification of low quality cells from single-cell RNA-seq data. Genome Biol 17:29. https://doi.org/10.1186/s13059-016-0888-1
61. Mias GI, Singh VV, Rogers LRK, Xue S, Zheng M, Domanskyi S, Kanada M, Piermarocchi C, He J (2021) Longitudinal saliva omics responses to immune perturbation: a case study. Sci Rep 11:710. https://doi.org/10.1038/s41598-020-80605-6
62. Li F, Kaczor-Urbanowicz KE, Sun J, Majem B, Lo H-C, Kim Y, Koyano K, Rao SL, Kang SY, Kim SM, Kim K-M, Kim S, Chia D, Elashoff D, Grogan TR, Xiao X, Wong DTW (2018) Characterization of human salivary extracellular RNA by next-generation sequencing. Clin Chem 64:1085–1095. https://doi.org/10.1373/clinchem.2017.285072
63. Zhao S, Zhang Y, Gamini R, Zhang B, von Schack D (2018) Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: PolyA+ selection versus rRNA depletion. Sci Rep 8:4781. https://doi.org/10.1038/s41598-018-23226-4
64. Morlan JD, Qu K, Sinicropi DV (2012) Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue. PLoS ONE 7:e42882. https://doi.org/10.1371/journal.pone.0042882
65. Sullivan R, Montgomery A, Scipioni A, Jhaveri P, Schmidt AT, Hicks SD (2022) Confounding factors impacting microRNA expression in human saliva: methodological and biological considerations. Genes (Basel) 13. https://doi.org/10.3390/genes13101874
66. Coenen-Stass AML, Magen I, Brooks T, Ben-Dov IZ, Greensmith L, Hornstein E, Fratta P (2018) Evaluation of methodologies for microRNA biomarker detection by next generation sequencing. RNA Biol 15:1133–1145. https://doi.org/10.1080/15476286.2018.1514236
67. Yeri A, Courtright A, Reiman R, Carlson E, Beecroft T, Janss A, Siniard A, Richholt R, Balak C, Rozowsky J, Kitchen R, Hutchins E, Winarta J, McCoy R, Anastasi M, Kim S, Huentelman M, van Keuren-Jensen K (2017) Total extracellular small RNA profiles from plasma, saliva, and urine of healthy subjects. Sci Rep 7:44061. https://doi.org/10.1038/srep44061
68. Wang S, Wang Z, Tao R, Wang M, Liu J, He G, Yang Y, Xie M, Zou X, Hou Y (2019) Expression profile analysis of piwi-interacting RNA in forensically relevant biological fluids. Forensic Sci Int Genet 42:171–180. https://doi.org/10.1016/j.fsigen.2019.07.015
69. Liu Z, Wang Q, Wang N, Zang Y, Wu R, Sun H (2022) A comprehensive characterization of small RNA profiles by massively parallel sequencing in six forensic body fluids/tissue. Genes (Basel) 13. https://doi.org/10.3390/genes13091530
70. Theda C, Hwang SH, Czajko A, Loke YJ, Leong P, Craig JM (2018) Quantitation of the cellular content of saliva and buccal swab samples. Sci Rep 8:6944. https://doi.org/10.1038/s41598-018-25311-0
71. Humphrey SP, Williamson RT (2001) A review of saliva: normal composition, flow, and function 85:162–169
72. Aps JK, van den Maagdenberg K, Delanghe JR, Martens LC (2002) Flow cytometry as a new method to quantify the cellular content of human saliva and its relation to gingivitis. Clin Chim Acta 321:35–41. https://doi.org/10.1016/S0009-8981(02)00062-1
73. Ambroa-Conde A, Girón-Santamaría L, Mosquera-Miguel A, Phillips C, Casares de Cal MA, Gómez-Tato A, Álvarez-Dios J, de La Puente M, Ruiz-Ramírez J, Lareu MV, Freire-Aradas A (2022) Epigenetic age estimation in saliva and in buccal cells. Forensic Sci Int Genet 61:102770. https://doi.org/10.1016/j.fsigen.2022.102770
74. Fábryová H, Celec P (2014) On the origin and diagnostic use of salivary RNA. Oral Dis 20:146–152. https://doi.org/10.1111/odi.12098
75. Gallo A, Tandon M, Alevizos I, Illei GG (2012) The majority of microRNAs detectable in serum and saliva is concentrated in exosomes. PLoS ONE 7:e30679. https://doi.org/10.1371/journal.pone.0030679
76. Palanisamy V, Sharma S, Deshpande A, Zhou H, Gimzewski J, Wong DT (2010) Nanostructural and transcriptomic analyses of human saliva derived exosomes. PLoS ONE 5:e8577. https://doi.org/10.1371/journal.pone.0008577
77. Park NJ, Li Y, Yu T, Brinkman BMN, Wong DT (2006) Characterization of RNA in saliva. Clin Chem 52:988–994. https://doi.org/10.1373/clinchem.2005.063206
78. Weinbrecht KD, Fu J, Payton M, Allen R (2017) Time-dependent loss of mRNA transcripts from forensic stains. RRFMS 7:1–12. https://doi.org/10.2147/RRFMS.S125782
79. Wang L, Nie J, Sicotte H, Li Y, Eckel-Passow JE, Dasari S, Vedell PT, Barman P, Wang L, Weinshiboum R, Jen J, Huang H, Kohli M, Kocher J-PA (2016) Measure transcript integrity using RNA-seq data. BMC Bioinformatics 17:58. https://doi.org/10.1186/s12859-016-0922-z
80. Sigurgeirsson B, Emanuelsson O, Lundsberg J (2014) Sequencing degraded RNA addressed by 3′ tag counting. PLoS ONE 9. https://doi.org/10.1371/journal.pone.0091851.g001
81. Hall JS, Taylor J, Valentine HR, Irlam JJ, Eustace A, Hoskin PJ, Miller CJ, West CML (2012) Enhanced stability of microRNA expression facilitates classification of FFPE tumour samples exhibiting near total mRNA degradation. Br J Cancer 107:684–694. https://doi.org/10.1038/bjc.2012.294
82. Bamberg M, Bruder M, Dierig L, Kunz SN, Schmidt M, Wiegand P (2022) Best of both: a simultaneous analysis of mRNA and miRNA markers for body fluid identification. Forensic Sci Int: Genet 102707. https://doi.org/10.1016/j.fsigen.2022.102707
83. Mayes C, Houston R, Seashols-Williams S, LaRue B, Hughes-Stamm S (2019) The stability and persistence of blood and semen mRNA and miRNA targets for body fluid identification in environmentally challenged and laundered samples. Leg Med (Tokyo) 38:45–50. https://doi.org/10.1016/j.legalmed.2019.03.007