Title

Novel approaches for bioinformatic analysis of salivary RNA sequencing data for development.

Journal

Bioinformatics, 34(1)

ISSN

1367-4803

Authors

Kaczor-Urbanowicz, Karolina Elżbieta
Kim, Yong
Li, Feng
et al.

Publication Date

2018

DOI

10.1093/bioinformatics/btx504

Peer reviewed

Bioinformatics, 34(1), 2018, 1–8
doi: 10.1093/bioinformatics/btx504
Advance Access Publication Date: 14 August 2017
Original Paper

OXFORD

Genome analysis

Novel approaches for bioinformatic analysis of salivary RNA sequencing data for development

Karolina Elżbieta Kaczor-Urbanowicz[1], Yong Kim[1], Feng Li[1], Timur Galeev[2,3], Rob R. Kitchen[2,3], Mark Gerstein[2,3,4], Kikuye Koyano[5], Sung-Hee Jeong[6], Xiaoyan Wang[7], David Elashoff[7], So Young Kang[8], Su Mi Kim[9], Kyoung Kim[8], Sung Kim[9], David Chia[10], Xinshu Xiao[5], Joel Rozowsky[2,3] and David T. W. Wong[1,*]

[1]Center for Oral/Head & Neck Oncology Research, School of Dentistry, Division of Oral Biology & Medicine University of California at Los Angeles, Los Angeles, CA 90095, USA, [2]Department of Molecular Biophysics and Biochemistry, [3]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA, [4]The Department of Computer Science, Yale University, New Haven, CT 06520, [5]Department of Integrative Biology and Physiology, University of California at Los Angeles, Los Angeles, CA 90095-1570, USA, [6]Department of Oral Medicine, School of Dentistry, Pusan National University, Beomeo-ri, Mulgeum-eup, Yangsan-si, Gyeongsangnam-do 626-770, Korea, [7]Department of Biostatistics, University of California at Los Angeles, Los Angeles, CA 90024, USA, [8]Department of Pathology & Translational Genomics, [9]Department of Surgery, Samsung Medical Center, Sungkyunkwan University School of Medicine, Gangnam-gu, Seoul, Korea and [10]Department of Pathology & Laboratory Medicine, University of California at Los Angeles, Los Angeles, CA 90095, USA

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on May 24, 2017; revised on July 25, 2017; editorial decision on August 3, 2017; accepted on August 8, 2017

Abstract

Motivation: Analysis of RNA sequencing (RNA-Seq) data in human saliva is challenging. Lack of standardization and unification of the bioinformatic procedures undermines saliva's diagnostic potential. Thus, it motivated us to perform this study.

Results: We applied principal pipelines for bioinformatic analysis of small RNA-Seq data of saliva of 98 healthy Korean volunteers including either direct or indirect mapping of the reads to the human genome using Bowtie1. Analysis of alignments to exogenous genomes by another pipeline revealed that almost all of the reads map to bacterial genomes. Thus, salivary exRNA has fundamental properties that warrant the design of unique additional steps while performing the bioinformatic analysis. Our pipelines can serve as potential guidelines for processing of RNA-Seq data of human saliva.

Availability and implementation: Processing and analysis results of the experimental data generated by the exceRpt (v4.6.3) small RNA-seq pipeline (github.gersteinlab.org/exceRpt) are available from exRNA atlas (exrna-atlas.org). Alignment to exogenous genomes and their quantification results were used in this paper for the analyses of small RNAs of exogenous origin.

Contact: dtww@ucla.edu

1 Introduction

1.1 Past challenges in salivary RNA-Seq analysis and current directions for RNA-Seq optimization of saliva samples

RNA sequencing (RNA-Seq) is a newly developed approach to perform transcriptome profiling by the use of deep-sequencing technologies. Successful application of RNA-Seq is key to extracellular RNA (exRNA) biomarker discovery. Although various comprehensive RNA-Seq approaches have been applied to examine RNA expression in several body fluids (i.e. plasma, serum, urine, cerebrospinal fluid, etc.), the examination of RNA-Seq data in saliva encountered several unique challenges due to inadequate technique of RNA isolation, improper stabilization of RNA or RNA library construction, as well as the insufficient amount of extracted RNA that can be used for further analysis (Bahn *et al.*, 2015; Burgos *et al.*, 2013; Freedman *et al.*, 2016; Hu *et al.*, 2014; Spielmann *et al.*, 2012).

Our group has optimized RNA-Seq library prep methods for saliva samples such as saliva exRNA isolation, large and small RNA library construction, inclusion of spike-in standards and controls, RNA-Seq data storage and data analysis. Our efforts have led to the optimization of the procedural components such as enhanced yield of salivary exRNA, best kit for salivary RNA library construction (large and small), high concordance of abundance of spike-in RNA and exRNAs in biological replicas of the same donor (Spielmann *et al.*, 2012).

Our saliva Standard Operating Procedure (SOP) for these procedures has been widely used and shown to effectively remove all cells from saliva (St John *et al.*, 2004) and concurrently stabilize proteins and RNA by the inclusion of a protease inhibitor cocktail [aprotinin, phenylmethylsulfonyl fluoride (PMSF) and sodium orthovanadate] and RNase inhibitor (SUPERase•In; Ambion, Austin, TX). Any contamination by cellular elements will distort the discriminatory biomarker profile. Therefore, effective and complete cell removal from saliva is important as it contains $\sim 1 \times 10^6$ epithelial cells, 1.2×10^6 leukocytes and $\sim 0.5 \times 10^6$ erythrocytes (Aps *et al.*, 2002). Contaminant cellular removal can be monitored by exclusion of cellular genomic DNA and cell counting (St John *et al.*, 2004). While the QIAGEN MirVana method produces high quality RNA for microarray profiling applications, the RNA yield is insufficient for RNA-Seq. Therefore, for RNA-Seq, the use of the miRNeasy Micro Kit (Qiagen) is recommended as the most suitable that produces sufficient RNA (~500 ng per ml of CFS) after DNase I digestion with good purity (OD 260/280 ~1.8) (Fig. 1).

In addition, a number of RNA-Seq library construction methods have been developed in the literature (Van Dijk *et al.*, 2014). For each library, 500 ng of total RNA is recommended as input (i.e. Exiqon for small RNA-Seq and Life Technologies for long RNA-Seq) along with a predefined amount of synthetic spike-in RNAs added into each RNA sample equivalently. The spike-in RNA serves as an internal standard to evaluate library efficiency and reproducibility, and to normalize data across different samples in order to calculate absolute RNA abundance. The spike-in RNAs cover a wide spectrum of RNA abundance and sequence diversity, which are essential features of normalization standards. They enable a detailed evaluation of library quality and RNA quantification (Williams *et al.*, 2013).

Since it is known that RNA from cell-free saliva (CFS) are partially degraded between 20 and 200 nucleotides (Palanisamy *et al.*, 2010), the library generation methods should be modified to exclude polyA selection and include a size-selection step favoring RNAs below 200 nucleotides. For each CFS sample, two different RNA-Seq libraries should be created, small RNA-Seq and long RNA-Seq (Bahn *et al.*, 2015). In our studies, the NEB small RNA-Seq kit (for small RNAs) and the NEB directional RNA-Seq kit (for long RNAs) generated the most reproducible and sensitive profiling of respective types of exRNAs in CFS compared to other alternative commercially available kits targeting different types of RNA (i.e. Illumina, ClonTech, etc). We can typically obtain 500 ng of total RNA from 1 ml of CFS that is adequate to generate high quality profiles of small or long exRNAs (Bahn *et al.*, 2015).

1.2 The landscape of exRNAs in human saliva

Human saliva has been used successfully and efficiently for biomarker development to enable the most accessible and non-invasive detection of several human diseases (Wong, 2015, 2012). Since the discovery of salivary exRNA in 2004 (Li *et al.*, 2004; St John *et al.*, 2004), there has been a constant increase in number of publications related to exRNAs in saliva (Fig. 2). Many studies reported the usefulness of exRNAs as potential biomarkers for detection of oral cancer (Li *et al.*, 2004), Sjögren syndrome (Hu *et al.*, 2007), resectable pancreatic cancer (Zhang *et al.*, 2010a), lung cancer (Zhang *et al.*, 2012), ovarian cancer (Lee *et al.*, 2012) and breast cancer (Zhang *et al.*, 2010a).

By means of next generation sequencing, the landscape of human exRNAs present in saliva was revealed to be largely composed of messenger RNAs (mRNAs), microRNAs (miRNAs), piwi-interacting RNAs (piRNAs), small nucleolar RNAs (snoRNAs) and circular RNAs (circRNAs) (Bahn *et al.*, 2015; Ogawa *et al.*, 2013; Spielmann *et al.*, 2012; Tandon *et al.*, 2012). However, the entire profile of exRNA from saliva has not been fully discovered. Still,

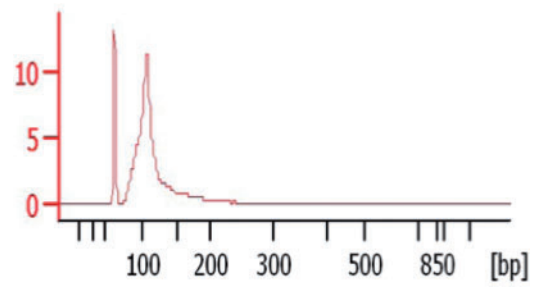

Fig. 1. A typical bioanalyzer trace of total RNA from CFS isolated by the miRNeasy Micro Kit (Qiagen)

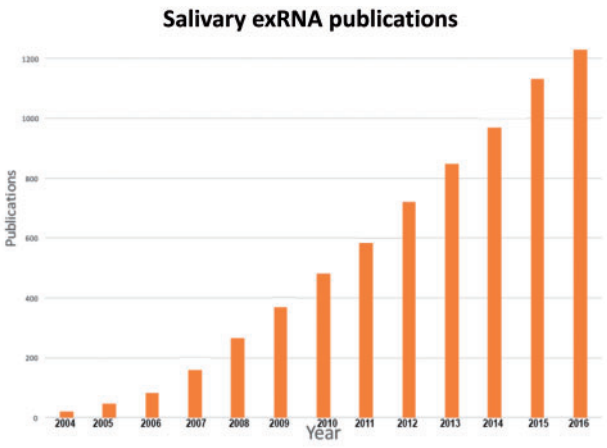

Fig. 2. Salivary exRNA publications

little is known about bacterial exRNAs present in human saliva, thus encouraging further comprehensive studies deciphering exogenous organisms in saliva. Therefore, we performed a genome-wide analysis of exogenous exRNAs in human cell-free saliva (CFS) by next generation sequencing.

1.3 Different approaches to bioinformatic analysis of RNA-Seq data in saliva

In the biomarker development process, there are currently various bioinformatic approaches to analyze RNA-Seq data from physiological fluids in humans. The pipeline schemes differ significantly depending on the final goal of the analysis (i.e. evaluation of human or microbial exRNAs) as well as the source (saliva, blood, cerebrospinal fluid, tissue, etc.).

Compared to other human biofluids, saliva is unique, as it contains abundant microbial species, while most of other physiological fluids (blood, cerebrospinal fluid, urine, etc.) are usually free of microorganisms (should be sterile) or contain just a few bacteria, which most probably constitute contamination from subject skin, collection process, isolation or kit preparations (Yeri *et al.*, 2017). Therefore, different RNA-Seq data analysis approaches for saliva are proposed, including those that provide additional mapping steps, which aim to remove microbial exRNAs in order to focus on human exRNAs.

After first quality control (QC) of raw reads (i.e. FastQC, RSeQC, etc.), the initial steps of analyzing salivary RNA-Seq data involve trimming adapters, filtering out contaminants, ribosomal RNAs (rRNAs) and calibration. The next step includes the alignment to the specific genomes of interest. We developed 2 major customized bioinformatic pipelines to identify different types of long exRNAs in saliva that may have originated either from human or microbial sources (Fig. 3). In case of human saliva:

1. If the fundamental goal of the bioinformatic analysis is to evaluate microbial exRNAs, the first step includes the alignment of the reads to the human genome (i.e. hg38) and subsequently aligning the unmapped reads to the bacterial genomes (i.e. 16S rRNA, HOMD, etc.) (Pipeline I, Fig. 3).
2. However, if the major goal of the analysis is to identify human exRNAs, the reads are first aligned to microbial genomes (i.e. 16S rRNA and HOMD) to eliminate those that possibly originated from microbial species, followed by the alignment of the unmapped reads to the human genome (i.e. hg38) (Pipeline II, Fig. 3).

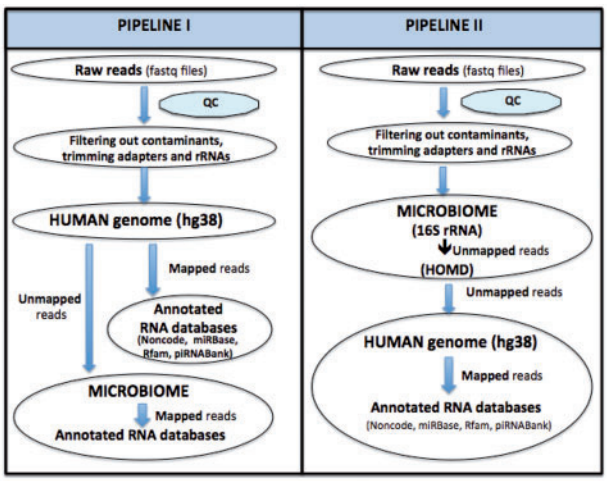

Fig. 3. Bioinformatic pipelines to identify different types of exRNAs in saliva originated from microbial or human sources

Thus, the above-mentioned two different pipelines are designated for different final goals.

However, when evaluating small exRNAs (i.e. miRNAs), the major problem is that the RNA sequences are very short, thus allowing them to easily map both to human and bacterial genomes. As a result, there are different proposed approaches for analyzing small RNA-Seq datasets:

1. Indirect mapping to bacterial genome(s) (including previous alignment to human genome and removing the mapped human RNA-Seq reads).
2. Direct mapping to bacterial genome(s) (without previous alignment to human genome and removing the human RNA-Seq reads).
3. Direct mapping to transcriptome (i.e. miRBase, piRNABank, etc.).

In order to set the standards for future studies, we compared major bioinformatics approaches for processing salivary RNA-Seq data, which have not been previously reported. In addition, we present the comprehensive report of existence of exogenous exRNAs (i.e. bacterial) in human saliva. Our findings open new venues for further functional, biological and biomarker discoveries related to microbial exRNAs in human saliva.

2 Materials and methods

2.1 Saliva collection, processing and RNA isolation

Unstimulated saliva samples were obtained prospectively from 98 healthy human volunteers of Korean origin (23 females and 75 males) with mean age of 54.84 ± 10.42 in accordance with a protocol approved by the University of California–Los Angeles (UCLA) Institutional Review Board. A written informed consent was obtained from each participant. Subjects were asked to refrain from eating, drinking, smoking or oral hygiene procedures for at least 1 h prior to saliva collection. Whole saliva samples were centrifuged at 2600*g* for 15 min at 4 °C. The supernatant (cell free saliva, CFS) was separated from the cellular pellet. RNase inhibitor (Superase-In, Ambion Inc., Austin, TX, USA) was then added to the CFS at a level of 20 U/ml. Total RNA was isolated using the miRNeasy Micro Kit (Qiagen) following manufacturer's instructions. Subsequently, RNA was treated with 50 U RNase-free DNase per μg RNA for 15 min at 37 °C and acid phenol/chloroform (pH 4.5) extraction (Roche) was performed to eliminate DNA. CFS RNA quality was measured by Bioanalyzer (Agilent).

2.2 Construction of small RNA-Seq libraries

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

2.3 CFS small RNA-Seq data analysis

RNA-Seq was performed using the Illumina sequencing platform. Adapter sequences and low quality reads were removed from small RNA-Seq reads. Subsequently, bioinformatic analysis of small

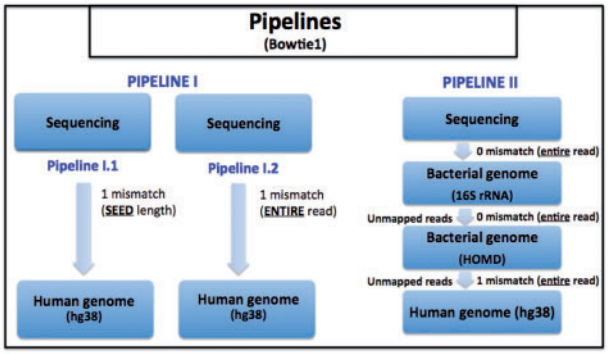

Fig. 4. Processing pipelines using Bowtie1 for small RNA- Seq data analysis in saliva

RNA-Seq data on saliva samples was performed using the two above-mentioned pipelines for processing of salivary small RNA-Seq data, that integrates read pre-processing, various alignment stages and exRNA detection (Fig. 3). However, our aim was to examine the influence of the applied alignment stringency as well as the differences in the sequence of the particular mapping steps on the total number of mapped reads. Thus, the following pipelines (Pipelines: I.1, I.2 and II) were applied using the same read aligner Bowtie1 (Langmead, 2010; Langmead *et al.*, 2009) with the aimed final goal of obtaining human reads (Fig. 4). Bowtie 1 was used as designed to be extremely fast for the alignment to a relatively short reference sequence (i.e. a bacterial genome).

- The first pipeline directly aligns to the human genome (hg38) allowing only 1 mismatch in the 19- nucleotide seed length region (Bowtie1 option: –n 1 –l 19 –best –strata hg38) (Fig. 4, Pipeline I.1).
- The second pipeline also directly aligns to the human genome (hg38) allowing 1 mismatch in the entire read (Bowtie1 option: –v 1 –best –strata hg38) (Fig. 4, Pipeline I.2).
- The last pipeline includes two major steps: initial filtering out of the bacterial reads by aligning to the 16S rRNA, followed by aligning the unmapped reads against the HOMD, permitting 0 mismatches (Bowtie1 option: -v 0). Then, remaining unmapped reads are aligned to the human genome (hg38) with up to 1 mismatch in the entire read (Bowtie1 option: -v 1 –best –strata -a hg38) (Fig. 4, Pipeline II).

In addition, we investigated the landscape of short exRNAs of exogenous origin in human CFS from individuals of this study. Quantifications of exogenous small RNA-Seq reads generated by the exceRpt toolkit (v4.6.3, implemented on the Genboree Workbench) were obtained from the exRNA Atlas (Subramanian et al., 2015), the data repository of the Extracellular RNA Communication Consortium (ERCC). The exceRpt pipeline processes each sample independently through a cascade of read-alignment steps designed to remove likely contaminants and endogenous sequences before aligning to exogenous sequences.

3 Results

3.1 Comparison of processing saliva samples using different pipelines and alignment options (stringency parameters)

The results of bioinformatics analyses including the mean number of mapped reads to the specific genomes of interest per sample using

Pipeline I: Direct alignment to the human genome (hg38)

Method	Starting mean number of reads per sample	Mean number of reads with at least one alignment/per sample (% starting reads)	Mean number of reads originated from intergenic regions (% reads aligned to hg38)
Pipeline I.1	20 484 385 (± 6 035 895) **(100%)**	3 482 345 (± 655 322) **(17%)**	905 409 (± 232 448) **(26%)**
Pipeline I.2	20 484 385 (± 6 035 895) **(100%)**	1 638 750 (± 388 443) **(8%)**	294 975 (± 76) **(18%)**

Pipeline II: Filtered (16s rRNA and HOMD) & mapping to human genome (hg38)

Method	Starting mean number of reads per sample	Mean number of reads mapping to 16S rRNA (% starting reads)	Mean number of reads mapping to HOMD (% starting reads)	Mean number of reads mapping to hg38 (% starting reads)	Mean number of reads originated from intergenic regions (% reads aligned to hg38)
Pipeline II	20 484 385 (± 6 035 895) **(100%)**	5 325 940 (± 788 012) **(26%)**	2 662 970 (± 455 275) **(13%)**	2 048 438 (± 348 2822) **(10%)**	491 625 (±127) **(24%)**

Fig. 5. Bioinformatics analysis of 98 small RNA-Seq datasets of saliva samples taken from healthy individuals using the pipelines from Figure 3

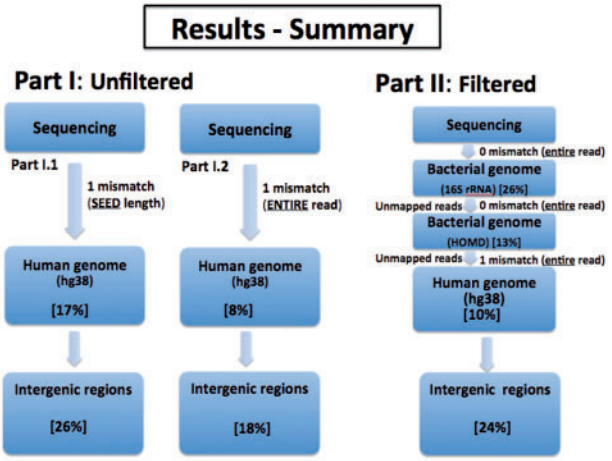

Fig. 6. Summary of the results of bioinformatics analysis

the 3 different pathways (Pipeline I.1; I.2 and II) are presented in Figure 5.

Direct alignment to the human genome (hg38) only allowing 1 mismatch in the seed length resulted in 17% read alignment and approximately 26% of those reads originated from intergenic regions (Pipeline I.1, Fig. 5). Allowance of 1 mismatch in the entire read contributed to fewer reads mapped to the human genome (hg38) (8%), with less reads originating from the intergenic regions (approximately 18%) (Pipeline I.2, Fig. 5). The second pipeline, which filters out the bacterial reads, revealed similar results to the pipeline I.2, amounting to 10% of aligned reads to the human genome (hg38) and 24% of them being intergenic reads (Pipeline II, Fig. 5).

In all alignment methods, approximately 8–17% of the RNA-Seq reads from CFS were aligned to the human genome (hg38), while 18–26% of the reads originated from the intergenic regions. However, the vast majority of the original reads mapped to the bacterial genomes (Fig. 6).

3.2 The landscape of exogenous miRNA in human saliva

Our findings show that circulating non-cellular small-RNAs, such as miRNAs, are consistently present in human saliva. The most

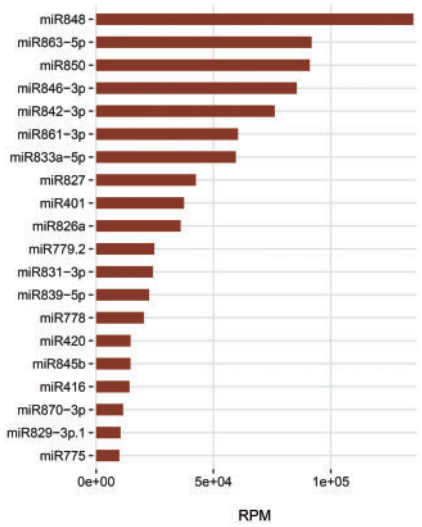

Fig. 7. Most abundant exogenous miRNAs identified in human CFS. Average RPM (reads mapped to exogenous miRNA per million of reads mapped to all exogenous miRNAs) values across 98 individuals

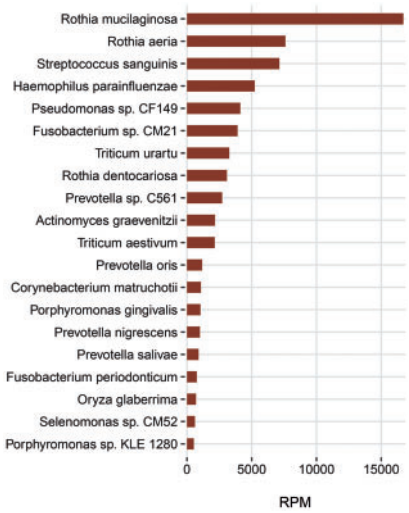

Fig. 8. Most abundant exogenous genomes identified in human CFS. Average RPM (number of species-specific reads per million of reads mapped to all exogenous genomes) values across all individuals

abundant exogneous miRNA were miR-848 (135183 RPM) and miR-863 (91832 RPM) (Fig. 7). These exogenous miRNAs might be from closely phylogenetically related plant species that are as yet unsequenced and can be a part of the subjects diet.

3.2.1 Analysis of exogenous genomes present in human saliva

Among the most abundant exogenous species present in human saliva were Rothia mucilanginosa (16711 RPM), Rotia aeria (7605 RPM) and Streptococcus sanguinis (7136 RPM) (Fig. 8).

Taxonomy-based quantification of exogenous reads revealed that among all the exogenous species that are present in human saliva, the vast majority constituted bacteria (90.4%), mostly of such phyla as Proteobacteria (34.8%) and Furmicutes (24%) (Fig. 9).

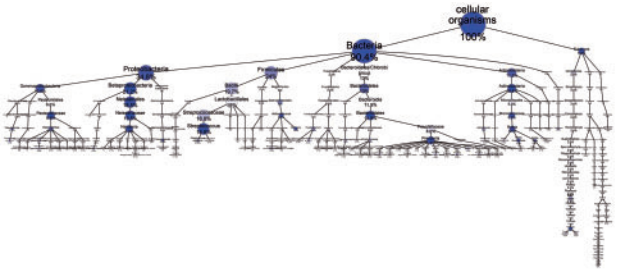

Fig. 9. Taxonomy-based quantification of exogenous reads. Percentage of reads that can be attributed to individual nodes out of all reads mapped to exogenous species, averaged across all individuals

4 Discussion

4.1 Current exRNA status in saliva

In 2002, the National Institute of Dental & Craniofacial Research (NIDCR) in the United States (U.S.) made a significant investment toward developing the science, translational and clinical utilities of saliva. Saliva has since become an emerging biofluid poised for translational and clinical applications. The U.S. National Institutes of Health (NIH) initiatives have resulted in a number of salient outcomes that have substantiated the scientific foundation for salivary research. These include sets of diagnostic toolboxes and point-of-care technologies. Five diagnostic alphabets are now known to be present in saliva including: proteome (Denny *et al.*, 2008; Yan *et al.*, 2009), transcriptome (Hu *et al.*, 2008; Li *et al.*, 2004, 2006), micro-RNA (Park *et al.*, 2009), metabolome (Sugimoto *et al.*, 2010) and microbiome (Wong, 2012). The presence of diagnostic constituents defined the clinical potentials of saliva and the translational utilities of salivary extracellular RNA (exRNA) biomarkers for detection of various systematic and oral diseases, as previously described.

4.2 Establishment of the Extracellular RNA Communication Consortium

The ability to widely detect and study exRNAs in different human physiological fluids was further enabled due to the initiative of the NIH Common Fund to establish the Extracellular RNA Communication Consortium (ERCC). The consortium of investigators aims to address the critical issues in the exRNA research field. Its major goal is to generate public resources for sharing principal scientific discoveries, protocols, innovative tools and technologies, including (Ainsztein *et al.*, 2015):

- creating a reference catalogue of exRNAs present in physiological fluids of healthy individuals that enables proper disease diagnosis and adequate therapies;
- description of the principal rules of exRNA biogenesis, distribution, uptake, function as well as a discovery of molecular tools, technologies and imaging modalities;
- development of definitive and highly discriminatory exRNA biomarkers for specific diseases;
- utilitizing exRNAs as therapeutic agents and development of modern technologies for these studies;
- establishing a public resource (**the exRNA Atlas**) to enable open source access for every individual to the scientific exRNA data, standardized exRNA protocols and other useful tools and technologies.

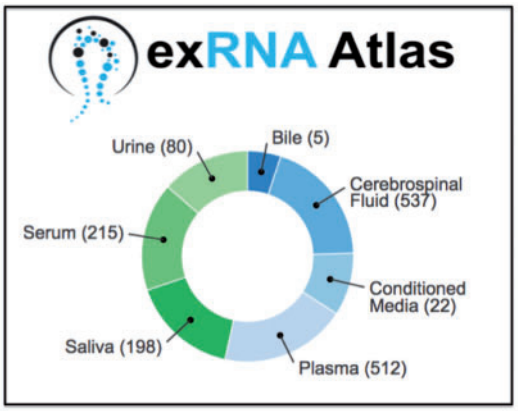

Fig. 10. Biofluids included in the exRNA Atlas established by the Extracellular RNA Communication Consortium (ERCC), National Institutes of Health (NIH), United States (http://genboree.org/exRNA-atlas/index.rhtml)

4.3 Saliva in exRNA Atlas

Currently, saliva constitutes about 13.3% of all human biofluids datasets in the exRNA Atlas (http://genboree.org/exRNA-atlas/index.rhtml) (Fig. 10).

Salivary exRNA has specific properties, including high content of bacterial components that warrant the design of unique additional steps in bioinformatic analysis of RNA-Seq data. RNA-seq enables the acquisition of nucleotide resolution of the human salivary transcriptome through alignment to multiple sequence databases. Salivary exRNA has a wide variety of applications, but no single analysis pipeline can be recommended in all cases. There is no gold standardized protocol for salivary RNA-Seq data analysis for experimental design, quality control, read alignment, quantification of gene and transcript levels, visualization, differential gene expression, alternative splicing, functional analysis, gene fusion detection, etc (Conesa *et al.*, 2016). Specific mapping steps and alignment options remains still largely uninvestigated. Therefore, our group compared principal comprehensive approaches for bioinformatic analysis of RNA-Seq data from saliva that integrates read pre-processing, alignment stages to the aimed genomes of interest, exRNA detection and quantification. Depending on the final goal of the performed analysis, filtering of reads mapped to the bacterial genomes (16S RNA and HOMD databases) before mapping to human genome (hg38) should or should not be performed. In either case, the stringency of the applied alignment criteria plays a key role.

4.4 The role of the applied stringency criteria for alignment quality and final results

Different aligner platforms (Bowtie1, Bowtie2, TopHat, Star, HISAT2, etc.) allow various ways of setting the stringency for the mapping parameters. For example, Bowtie 1 enables to limit the number of mismatches either in the entire read region (option -v) or only in the seed region (option –n), while HISAT2 uses different alignment option (i.e. –ignore-quals), when calculating a mismatch penalty. Bowtie2 rapidly narrows the number of possible mismatches in alignments by the possibilty of setting the adequate seed length (-L), the interval between the extracted seeds (-i) (in 'multiseed alignment') and the number of mismatches permitted per seed (-N). In the aligner STAR, discrete mismatches can be controlled by the use of alignment option: -outFilterMismatchNmax (where N = number of mismatches you wish to tolerate) or –outFilterMismatchNoverLmax (where alignment will be output only if its ratio of mismatches to mapped length is less than this value).

For reads that do not map to the human genome, alignment to the transcriptome (i.e. Ensembl genes) enables alignment of reads mapped to splice junctions. TopHat is a spliced read mapper which uses Bowtie1 to align reads and then identifies splice junction between junctions.

Increasing the mapping stringency noticeably decreases the number of the mapped reads but, simultaneously, reduces the error rate in the base alignment that may lead to misleading mapping results.

The comparison of the results originated from the various processing pathways revealed that, while using Bowtie1, the main difference in the results seems to be caused by limiting the number of mismatches in the seed length (-n option) as opposed to the entire read (-v option). Application of more strict criteria (0 & 1 mismatch in the entire read) results in obtaining less number of reads mapped to the human genome (hg38) as well as to the intergenic regions. Conversely, using less strict criteria (1 mismatch in the seed length, thus allowing more mismatches in the entire read) delivers larger alignment percentage of both human and intergenic reads.

ExRNA biomarkers that are specific and indicative of health or disease are greatly needed to serve as surrogates for clearly defined endpoints such as cancer as well as to monitor the health status, disease onset, treatment responsiveness and outcome (Phillips, 2006). Therefore, our group is currently working on developing highly discriminatory and definitively validated salivary exRNA biomarkers for gastric cancer detection using RNA-Seq and bioinformatic analyses.

4.5 Uniqueness of salivary RNA-seq analysis

Bioinformatic analysis of RNA-Seq data revealed unique property of saliva, compared to other physiological fluids (i.e. blood, urine, etc.), as it contains large amounts of bacterial RNA-Seq reads. Our approach enabled to reveal detailed delineation of exogenous species present in human saliva including bacteria, plants, etc. These findings are concordant with previous reports on a great abundance of exogenous species present in saliva, with much more reads aligned to bacterial species than to the human genome (Yeri *et al.*, 2017). In addition, our results show that circulating non-cellular small RNA and exogenous miRNAs are consistenty present in human saliva.

In summary, using stringent and sensitive criteria for alignment to a human genome results in fewer mapped human reads, but much more reads annotated to human RNA databases. Only a small percentage of the reads come from the intergenic regions. However, performing the bioinformatics analysis using less sensitive and more lenient alignment parameters contributes to more reads mapping to the human genome, but of lower quality (i.e. with more mismatches), so only small percentage of them can be annotated to human RNA databases and more reads are left without annotations (intergenic alignments). This should be alarming because it can impose misleading results such as DNA contamination.

Interestingly, strict alignment criteria results in similar number of reads mapped to the human genome (hg38) compared to alignment with filtering of the bacterial reads (16S rRNA and oral microbiome) versus alignment without prior filtering.

Optimizing the sequence of specific alignment steps and stringency parameters for processing RNA-Seq data has the potential to grossly improve the final data quality and enable greater insight into understanding how sequence information relates to the biology of

analyzed biofluid. In saliva, several additional steps are required for bioinformatics analysis such as eliminating bacterial reads before the alignment to human genome while applying lenient criteria or application of more strict criteria if bacterial reads are not filtered. Our bioinformatic analysis pipelines can serve as potential gudielines for processing of RNA-Seq data of human saliva, thus opening up a new avenue for conducting further investigation.

Finally, this study explored widespread presence of exogenous small RNAs in human saliva. The observation that bacterial exRNAs are present in human saliva in vast majority is noteworthy. Our findings are consistent with the literature, where bacterial species, such as *Neisseria flavescens*, *Rothia mucilaginosa* and *Streptococcus salivarius* accounted for about 67.3 ± 8.8% of the salivary bacterial population (Takeshita *et al.*, 2016). In addition, Firmicutes, Bacteroidete, Actinobacteria and Proteobacteria were also found to be the most abundant in saliva in other study, with eleven genera, Streptococcus, Prevotella, Veillonella, Neisseria, Haemophilus, Campylobacter, Fusobacterium, Rothia, Mycoplasma, Actinomyces and Aggregatibacter comprising 90% of the bacterial community in saliva (Hasan *et al.*, 2014). Apart from that, *Veillonella parvula, Prevotella melaninogenica, Fusobacterium periodonticum and Streptococcus mitis* were also predominant microorganisms identified in saliva of healthy individuals in previous studies (Diaz *et al.*, 2012, Pereira *et al.*, 2012).

Acknowledgement

We thank Tania Lombo of the NIH Common Fund for her assistance and guidance.

Funding

Public Health Service (PHS) grants from the National Institutes of Health/National Institute of Dental and Craniofacial Research (NIH/NIDCR) [UH3 TR000923 and R90 DE022734] as well as the 2017 Debbie's Dream Foundation – American Association for Cancer Research (AACR) Gastric Cancer Research Fellowship (Grant Number 17-40-41-KACZ). In addition, we highly acknowledge the donation made by Ronnie James Dio Stand Up and Shout Cancer Fund.

Conflict of Interest: D.T.W.W. is co-founder of RNAmeTRIX Inc., a molecular diagnostic company. He holds equity in RNAmeTRIX, and serves as a company Director and Scientific Advisor. The University of California also holds equity in RNAmeTRIX. Intellectual property that D.T.W.W. invented and which was patented by the University of California has been licensed to RNAmeTRIX. Additionally, he is a consultant to PeriRx, GlaxoSmithKlein, Wrigley and Colgate-Palmolive.

None of the other authors have a conflict of interest in relation to this study.

References

Abuín,J. *et al.* (2016) SparkBWA: speeding up the alignment of high-throughput DNA sequencing data. *PLoS One*, **11**, e0155461.

Ainsztein,A.M. *et al.* (2015) The NIH extracellular RNA communication consortium. *J. Extracell. Vesicles*, **4**, 27493.

Aps,J.K. *et al.* (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.

Bahn,J.H. *et al.* (2015) The landscape of microRNA, piwi-interacting RNA, and circular RNA in human saliva. *Clin. Chem.*, **61**, 221–230.

Brinkmann,O. *et al.* (2011) Oral squamous cell carcinoma detection by salivary biomarkers in a Serbian population. *Oral Oncol.*, **47**, 51–55.

Burgos,K.L. *et al.* (2013) Identification of extracellular miRNA in human cerebrospinal fluid by next-generation sequencing. *RNA*, **19**, 712–722.

Conesa,A. *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol.*, **17**, 13–32.

Denny,P. *et al.* (2008) The proteomes of human parotid and submandibular/sublingual gland salivas collected as the ductal secretions. *J. Proteome Res.*, **7**, 1994–2006.

Diaz,P.I. *et al.* (2012) Using high throughput sequencing to explore the biodiversity in oral bacterial communities. *Mol. Oral. Microbiol.*, **27**, 182–201.

Dobin,A. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Ellis,E.S. *et al.* (2013) RNA-Seq optimization with eQTL gold standards. *BMC Genomics*, **14**, 892–903.

Farrell,J.J. *et al.* (2012) Variations of oral microbiota are associated with pancreatic diseases including pancreatic cancer. *Gut*, **61**, 582–588.

Freedman,J.E. *et al.* (2016) Diverse human extracellular RNAs are widely detected in human plasma. *Nat. Commun.*, **26**, 11106.

Hasan,N.A. *et al.* (2014) Microbial community profiling of human saliva using shotgun metagenomic sequencing. *PLoS One*, **9**, e97699.

Hu,S. *et al.* (2010) Preclinical validation of salivary biomarkers for primary Sjogren's syndrome. *Arthritis Care Res. (Hoboken)*, **62**, 1633–1638.

Hu,S. *et al.* (2007) Salivary proteomic and genomic biomarkers for primary Sjogren's syndrome. *Arthritis Rheum.*, **56**, 3588–3600.

Hu,L. *et al.* (2014) Identification of microRNAs predominately derived from testis and epididymis in human seminal plasma. *Clin. Biochem.*, **47**, 967–972.

Hu,Z. *et al.* (2008) Exon-level expression profiling: a comprehensive transcriptome analysis of oral fluids. *Clin. Chem.*, **54**, 824–832.

Kim,D. *et al.* (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.

Langmead,B. (2010) Aligning short sequencing reads with Bowtie. *Curr. Protoc. Bioinf.*, **11**, 11.7.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Lee,Y.H. *et al.* (2012) Salivary transcriptomic biomarkers for detection of ovarian cancer: for serous papillary adenocarcinoma. *J. Mol. Med.*, **90**, 427–434.

Li,H. and Durbin,R. (2009) Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics*, **25**, 1754–1760.

Li,Y. *et al.* (2006) Serum circulating human mRNA profiling and its utility for oral cancer detection. *J. Clin. Oncol.*, **24**, 1754–1760.

Li,Y. *et al.* (2004) Salivary transcriptome diagnostics for oral cancer detection. *Clin. Cancer Res.*, **10**, 8442–8450.

Ogawa,Y. *et al.* (2013) Small RNA transcriptomes of two types of exosomes in human whole saliva determined by next generation sequencing. *Biol. Pharm. Bull.*, **36**, 66–75.

Palanisamy,V. *et al.* (2010) Nanostructural and transcriptomic analyses of human saliva derived exosomes. *PLoS One*, **5**, e8577.

Park,N.J. *et al.* (2009) Salivary microRNA: discovery, characterization, and clinical utility for oral cancer detection. *Clin. Cancer Res.*, **15**, 5473–5477.

Pereira,J.V. *et al.* (2012) Bacterial diversity in the saliva of patients with different oral hygiene indexes. *Braz. Dent. J.*, **23**, 409–416.

Phillips,C. (2006) Rinse and Spit: Saliva as a Cancer Biomarker Source. *NCI Cancer Bull.*, **3–5**.

Subramanian,S.L. *et al.* (2015) Integration of extracellular RNA profiling data using metadata, biomedical ontologies and Linked Data technologies. *J. Extracell Vesicles.*, **4**, 27497.

Spielmann,N. *et al.* (2012) The human salivary RNA transcriptome revealed by massively parallel sequencing. *Clin. Chem.*, **58**, 1314–1321.

St John,M.A. *et al*. (2004) Interleukin 6 and interleukin 8 as potential biomarkers for oral cavity and oropharyngeal squamous cell carcinoma. *Arch. Otolaryngol. Head Neck Surg*., **130**, 929–935.

Sugimoto,M. *et al*. (2010) Capillary electrophoresis mass spectrometry-based saliva metabolomics identified oral, breast and pancreatic cancer-specific profiles. *Metabolomics*, **6**, 78–95.

Takeshita,T. *et al*. (2016) Bacterial diversity in saliva and oral health-related conditions: the Hisayama Study. *Sci Rep*., **6**, 22164.

Tandon,M. *et al*. (2012) Deep sequencing of short RNAs reveals novel micrornas in minor salivary glands of patients with Sjogren's syndrome. *Oral Dis*., **18**, 127–131.

Van Dijk,E.L. *et al*. (2014) Library preparation methods for next-generation sequencing: tone down the bias. *Exp. Cell. Res*., **322**, 12–20.

Vijay,N. *et al*. (2013) Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Mol. Ecol*., **22**, 620–634.

Wei,F. *et al*. (2009) Electrochemical sensor for multiplex biomarkers detection. *Clin. Cancer Res*., **15**, 4446–4452.

Williams,Z. *et al*. (2013) Comprehensive profiling of circulating microRNA via small RNA sequencing of cDNA libraries reveals biomarker potential and limitations. *Proc. Natl. Acad. Sci. USA*, **110**, 4255–4260.

Wang,Z. *et al*. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet*., **10**, 57–63.

Wong,D.T. (2012) Salivaomics. *J. Am. Dent. Assoc*., **143**, 19S–24S.

Wong,D.T. (2015) Salivary extracellular non-coding RNA: emerging biomarkers for molecular diagnostics. *Clin. Ther*., **37**, 540–551.

Xiao,H. *et al*. (2012) Proteomic analysis of human saliva from lung cancer patients using two-dimensional difference gel electrophoresis and mass spectrometry. *Mol. Cell Proteomics*, **11**, M111.

Yan,W. *et al*. (2009) Systematic comparison of the human saliva and plasma proteomes. *Proteomics Clin. Appl*., **3**, 116–134.

Yeri,A. *et al*. (2017) Total extracellular small RNA profiles from plasma, saliva, and urine of healthy subjects. *Sci Rep*., **7**, 44061.

Zhang,L. *et al*. (2010a) Salivary transcriptomic biomarkers for detection of resectable pancreatic cancer. *Gastroenterology*, **138**, 949–957 e7.

Zhang,L. *et al*. (2010b) Discovery and preclinical validation of salivary transcriptomic and proteomic biomarkers for the non-invasive detection of breast cancer. *PLoS One*, **5**, e15573.

Zhang,L. *et al*. (2012) Development of transcriptomic biomarker signature in human saliva to detect lung cancer. *Cell. Mol. Life Sci*., **69**, 3341–3350.

Zhao,S. *et al*. (2015) A comprehensive evaluation of ensembl, RefSeq and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC Genomics*, **16**, 97–111.