Title: Principled, Comprehensive Analytic Platform for Extracellular RNA Analysis

Summary:

Highlights and eTOC Blurb:

- exceRpt processes and analyzes exRNA profiling data
- Generates quality control metrics, RNA biotype abundance estimates, and processing reports
- User-friendly, browser-based graphical interface available
- Processes all RNA-seq datasets in the exRNA Atlas

Introduction:

Recent discoveries of extracellular RNA (exRNA) in the blood and other body fluids have added a new dimension to the paradigm of intercellular signaling (Patton et al., 2015; Skog et al., 2008). The relative stability of exRNA within extracelluar vesicles (EVs) or bound to proteins or lipids (Yanez-Mo et al., 2015), coupled to the availability of sensitive and specific tools such as RNA-seq, underpins the emergence of exRNA profiling as an approach for biomarker discovery. exRNA-based liquid biopsy is particularly attractive as a non-invasive mode for monitoring disease due to the significantly increased accessibility of biofluids over tissues, thereby allowing more frequent and longitudinal sampling (Byron et al., 2016). With better characterization of the differences between profiles secreted by diseased and healthy tissues, the diagnostic and prognostic utility of exRNA-based profiling is increasingly becoming a reality (Akat et al., 2014; Yuan et al., 2016).

However, exRNA profiling faces unique challenges. Biochemical methods for extraction, purification, and sequencing of exRNAs are much more vulnerable to contamination and artifacts than cellular RNA preparations, in large part due to relative low abundance (Danielson et al., 2017). Quality control prior to sequencing for samples derived from EV or exosome preparations is difficult due to the lack of reliable 'housekeeping' markers, such as the ratio of 18S and 28S ribosomal RNAs (Tataruch-Weinert et al., 2016). The variable presence of rRNA in mixtures of low- and high-density EVs (Lasser et al., 2017); deterministic cleavage of structured smallRNA (tRNAs and piRNAs) and longer RNA molecules; and imperfect annotation of miRNAs, piRNAs, and tRNAs all pose challenges for quantification and functional interpretation. Furthermore, it has been suggested that exogenous exRNAs may be also present at detectable levels in some biofluids (Freedman et al., 2016; Yeri et al., 2017), necessitating careful analysis to ensure that these sequences are not in fact derived from endogenous RNA molecules. For these reasons, existing computational tools capable of analyzing smallRNA-seq data are not as well suited to the new field of exRNA analysis.

To address these analytical challenges, we present here the extracellular RNA processing tool (exceRpt). exceRpt is the primary smallRNA analysis pipeline of the NIH Extracellular RNA Communication Consortium (ERCC). By providing an optimized and standardized bioinformatics platform, exceRpt reduces technical bias and allows for cross-study analyses to potentiate meaningful insights into exRNA biology.

Results:

The exceRpt pipeline is composed of a cascade of computational steps (Fig. 1A) where the input reads at a given step are aligned against a set of annotations and the unmapped reads are the input to the next step, with the prioritization of the steps based on our level of confidence in the annotations. For example, in order to combat potential contamination in a library, mapping to known contaminants occurs before the host genome, since if the steps were reversed contaminant sequences could be incorrectly quantified as endogenous RNAs which is less preferable than having false positive

contaminant reads. The pipeline is also highly modular, allowing the user to define the order of which smallRNA annotations are used during read-mapping; it includes support for random-barcoded libraries and spike-in sequences for calibration or titration. The general workflow comprises steps for preprocessing, endogenous alignment, and exogenous alignment (Figure 1A).

First, exceRpt begins the preprocessing step by automatically identifying and removing 3' adapter sequences. Randomly barcoded 5' and/or 3' adapter sequences are increasingly being used in smallRNA sequencing in an attempt to identify and compensate for ligation and/or amplification artifacts that have the potential to affect downstream quantification (Fu et al., 2014). exceRpt is capable of removing and quantifying these biases at both the insert level, which reveals ligation/amplification bias, and the transcript level, which provides an opportunity to compensate for the bias by counting unique N-mer barcodes rather than counting the number of inserts. The pipeline than aligns against an input library of known spike-ins sequences if used in the library construction, followed by a filter to remove low-quality reads and reads with large homopolymer repeats. As the final preprocessing step, exceRpt aligns reads to likely sequences in the UniVec database and to endogenous ribosomal RNAs, both of which are highly variable in abundance in EV preparations. This is designed for filtration of common laboratory contaminants.

Second, reads are aligned to either the human or mouse endogenous genome and transcriptome, and transcript abundances are calculated (RNAs are quantified using both raw read counts and normalized reads per million (RPM)). Based on the variety of RNA preparations available (totalRNA, smallRNA, miRNA), the user can prioritize the order that the annotations (miRBase, tRNAscan, piRNA, gencode, circRNA) are used for quantification based on our confidence in the presence of a given annotation in a given sample. For example, reads from a miRNA-seq prep can be assigned to miRBase miRNA annotations before piRNA annotations. Likewise, reads from long or total RNA preparations can be assigned to longer GENCODE transcripts before (or instead of) the other smallRNA libraries. This feature is particularly relevant for lower-confidence annotations; piRNAs, for example, are generally given lower priority than tRNAs to ensure correct read assignments.

Third, we designed exceRpt from the beginning to enable confident assessment of non-human sequences in biofluids after careful, explicit removal of as many known or likely contaminants as possible. Before we analyze the remaining reads for potential exogenous sequences, we perform a second pass alignment against the host genome using a more relaxed mapping criteria and data based on known repetitive sequences. This serves to remove sequences that could potentially be from the host genome and to be fairly conservative in the identification of exogenous sequences. Reads are then aligned to curated libraries of annotated exogenous miRNAs in miRBase and exogenous rRNA sequences in the Ribosomal Database Project (RDP), followed by alignment to the full genomes of all sequenced bacteria, viruses, plants, fungi, protists, metazoa, and selected vertebrates that are potentially part of the host diet. By characterizing exogenous genome alignments generated by exceRpt in terms of the NCBI taxonomy tree and assigning reads to the most specific node in the phylogenetic tree (many reads can only be assigned to nodes higher up in the phylogenetic tree due to not uniquely mapping to a specific genome of a sub-species), users may obtain valuable information regarding the contribution of the flora to various exRNA samples and generate phylogenies for cross-sample comparison.

The pipeline generates bulk statistics for differential abundance of the various RNA biotypes in addition to sample-level quality control (QC) metrics and processing reports. Descriptions of the post-processing output files and diagnostic plots generated

by exceRpt are listed in Table 1. As a performance evaluation, we found that the endogenous miRNA abundance estimates produced by exceRpt are in close agreement with existing tools. Comparing exceRpt-filtered read counts for miRBase miRNAs, we obtain an average Pearson correlation of 99.99% to the counts produced by miRDeep2 (Figure S1). As another performance evaluation, running the same sample through the pipeline with individual steps excluded shows the effect of the filters and alignments on downstream quantifications (Figure 1B). Most obvious from this analysis is that the pre-filtering of low quality and low-complexity reads and reads that align to UniVec or rRNA sequences account for a sizeable fraction of the total number sequenced and, without explicit removal, do align to the human genome leading to potential confounding and added quantification variability. UniVec has the largest effect on the fraction of reads aligning to exogenous genomes, and leaving it out substantially increases the number of reads that appear to be, but are not, exogenous in origin.

These bulk statistics can be used to differentiate biofluids (or tissues, if exceRpt is run on cellular samples) on the basis of their RNA distribution. For example, results from samples selected from the exRNA Atlas (Figure 2A) show that, relative to other biofluids, saliva samples tend to have more reads that are unmapped or that map to exogenous genomes, which is consistent with saliva's high potential for bacterial contamination and exposure to the external environment. Moreover, abundance quantifications for specific RNA biotypes can show which miRNAs (or other RNA biotype) are most highly represented in a particular sample (Figure S2). This information is critical for understanding the composition of particular exRNA profiles and for interrogating their biological significance.

In Figure 2C and 2D we present the phylogenetic trees of the reads that we assign to bacterial ribosomal and genome sequences for a specific saliva sample. Saliva biofluids are distinguished from other biofluids by their exposure to a robust and complex bacterial community in the oral cavity (Hasan et al., 2014), which causes a greater contribution of reads of bacterial origin (and not human genome) to the sample. In both the phylogenetic trees constructed using bacterial ribosomal and genome mapped reads, we find an abundance of reads assigned to the node corresponding to the genus *Streptococcus*. We have a high degree of confidence in these results given that the reads used for constructing these two trees are disjoint.

Discussion:

The exceRpt pipeline was built to address the need for a standardized bioinformatics processing platform in extracellular RNA research, and is structured as a series of filtering and quantification steps where unmapped reads are used as inputs to the next step. The prioritization of steps is biased towards conservative estimates for RNA quantifications, with higher confidence libraries (by degree of expectation or annotation quality) having higher priority. The exceRpt has uniformly processed all of the datasets in the ERCC exRNA Atlas (http://exrna-atlas.org/) in a principled, comprehensive manner. The pipeline applies ERCC-defined QC standards, allows for user-specification for library prioritization, offers barcoding and spike-in support, and generates detailed quantification reports, all of which can be done with the source code available in a Github repository or in a user-friendly, browser-based interface available at Genboree.org.

Tables:

File Name	Description of File
QC Data	
exceRpt_DiagnosticPlots.pdf	All diagnostic plots automatically generated by the tool
exceRpt_readMappingSummary.txt	Read-alignment summary including total counts for each library
exceRpt_ReadLengths.txt	Read-lengths (after 3' adapters/barcodes are removed)
Raw Transcriptome Quantifications	
exceRpt_miRNA_ReadCounts.txt	miRNA read-counts quantifications
exceRpt_tRNA_ReadCounts.txt	tRNA read-counts quantifications
exceRpt_piRNA_ReadCounts.txt	piRNA read-counts quantifications
exceRpt_gencode_ReadCounts.txt	gencode read-counts quantifications
exceRpt_circularRNA_ReadCounts.txt	circularRNA read-count quantifications
Normalized Transcriptome Quantifications	
exceRpt_miRNA_ReadsPerMillion.txt	miRNA RPM quantifications
exceRpt_tRNA_ReadsPerMillion.txt	tRNA RPM quantifications
exceRpt_piRNA_ReadsPerMillion.txt	piRNA RPM quantifications
exceRpt_gencode_ReadsPerMillion.txt	gencode RPM quantifications
exceRpt_circularRNA_ReadsPerMillion.txt	circularRNA RPM quantifications
R Objects	
exceRpt_smallRNAQuants_ReadCounts.RData	All raw data (binary R object)
exceRpt_smallRNAQuants_ReadsPerMillion.RData	All normalized data (binary R object)

Figure Legends:

Figure 1(A): exceRpt Schema
Sample inputs in FASTA or SRA file formats for the input to excerpt. Adapter and random barcode sequences are removed, followed by a read-quality filter, optional spike-in library removal, and contaminant library removal. Unmapped reads then enter the endogenous quantification engine, with RNA library prioritization defined by the user. After a second-pass endogenous genome and repetitive elements filter, reads are mapped to the exogenous libraries.

Figure 1(B): Leave-one-out Analysis
Removing the UniVec alignment step significantly increases the number of reads that map to the exogenous genomes.

Figure 2(A): Read Distributions
exceRpt outputs endogenous alignment quantifications which can be used to see RNA type distributions in exRNA samples. Here, saliva has a higher proportion of exogenous sequences than other samples, and urine has a higher proportion of tRNA sequences. Quantifications can also be performed for cellular datasets, such as ENCODE samples.

Figure 2(C): Exogenous Alignment Phylogeny
Exogenous sequence quantifications can be used to construct phylogenetic trees using rRNA reads and exogenous genome reads.

Figure S3: Quality Control Metrics
ERCC QC metrics are based on number of transcriptome reads and ratio of RNA annotated reads to the genome reads. The horizontal and vertical lines define QC cutoffs, and most exRNA Atlas samples meet the standards in the upper right quadrant.

STAR Methods:

KEY RESOURCES TABLE:

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
exRNA Atlas	ERCC	https://exrna-atlas.org/
Human reference genome build GRCh38 (UCSC hg38)	Genome Reference Consortium	https://www.ncbi.nlm.nih.gov/grc/human
Human reference genome build GRCh37 (UCSC hg19)	Genome Reference Consortium	https://www.ncbi.nlm.nih.gov/grc/human
Mouse reference genome build GRCm38 (UCSC mm10)	Genome Reference Consortium	http://www.ncbi.nlm.nih.gov/grc/mouse
miRBase version 21	(Griffiths-Jones, 2004)	http://www.mirbase.org/
GtRNAdb	(Chan and Lowe, 2009)	http://gtrnadb.ucsc.edu/
piRNABank	(Sai Lakshmi and Agrawal, 2008)	http://pirnabank.ibab.ac.in/
Gencode version 24 (hg38)	(Harrow et al., 2012)	http://www.gencodegenes.org/
Gencode version 18 (hg19)	(Harrow et al., 2012)	http://www.gencodegenes.org/
Gencode version M9 (mm10)	(Mudge and Harrow, 2015)	http://www.gencodegenes.org/
circBase	(Glazar et al., 2014)	http://www.circbase.org/
UniVec	NCBI	ftp://ftp.ncbi.nlm.nih.gov/pub/UniVec/
Ribosomal Database Project	(Cole et al., 2014)	http://rdp.cme.msu.edu/
Software and Algorithms		
exceRpt version 4.6.2	This paper	http://genboree.org/theCommons/projects/exrna-tools-may2014/wiki/Small%20RNA-seq%20Pipeline
Java	Oracle Corporation	https://www.java.com/
R version 3.2	The R Project	https://www.r-project.org/
FASTX version 0.0.14	Hannon Lab	http://hannonlab.cshl.edu/fastx_toolkit/
STAR version 2.4.2a	(Dobin et al., 2013)	https://github.com/alexdobin/STAR/releases
Bowtie 2 version 2.2.6	(Langmead and Salzberg, 2012)	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Samtools version 1.3.1	(Li et al., 2009)	http://www.htslib.org/
FastQC v0.11.2	Babraham Bioinformatics	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
SRA-Toolkit version 2.3	NCBI	https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software

Quality Control

To evaluate the samples themselves and identify outliers, the ERCC developed QC data standards which exceRpt evaluates uniformly on all input samples: (1) datasets are required to have at least 100,000 reads that overlap with any annotated RNA transcript in the host genome, and (2) over 50% of the reads that map to host genome also align to any RNA annotation. The first criterion ensures that enough reads are generated for quantification (the minimal read depth required for the minimal normalized expression of an annotated RNA to be greater than 1 RPM) and the second ensures that the reads mostly align to RNA, as opposed to DNA contamination from cellular sources. We find that 95% of the ~2500 exRNA-Seq datasets that have been uniformly processed in the exRNA Atlas with exceRpt meet both criteria (Figure S3), with most datasets well above both thresholds.

QUANTIFICATION AND STATISTICAL ANALYSIS:

All statistical analyses were performed in R.

DATA AND SOFTWARE AVAILABILITY:
The graphical, browser-based, user-friendly interface for uploading and processing exRNA-seq datasets with exceRpt is available at the Genboree Workbench: http://genboree.org/theCommons/projects/exrna-tools-may2014/wiki/Small%20RNA-seq%20Pipeline.
The exceRpt source code may be downloaded and installed manually for the most amount of flexibility. Moreover, the exceRpt Docker image with all required dependencies may be used for installation on the user's own machine or cluster: https://github.com/gersteinlab/exceRpt/.

ADDITIONAL RESOURCES:
The ERCC exRNA Atlas can be found here: https://exrna-atlas.org/
The ERCC quality control standards can be found here: https://exrna.org/resources/data/data-quality-control-standards/

Supplemental Information:
[FILL]

References:

Akat, K.M., Moore-McGriff, D., Morozov, P., Brown, M., Gogakos, T., Correa Da Rosa, J., Mihailovic, A., Sauer, M., Ji, R., Ramarathnam, A., *et al.* (2014). Comparative RNA-sequencing analysis of myocardial and circulating small RNAs in human heart failure and their utility as biomarkers. Proc Natl Acad Sci U S A *111*, 11151-11156.
Byron, S.A., Van Keuren-Jensen, K.R., Engelthaler, D.M., Carpten, J.D., and Craig, D.W. (2016). Translating RNA sequencing into clinical diagnostics: opportunities and challenges. Nat Rev Genet *17*, 257-271.
Chan, P.P., and Lowe, T.M. (2009). GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Res *37*, D93-97.
Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., and Tiedje, J.M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res *42*, D633-642.
Danielson, K.M., Rubio, R., Abderazzaq, F., Das, S., and Wang, Y.E. (2017). High Throughput Sequencing of Extracellular RNA from Human Plasma. PLoS One *12*, e0164644.
Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.
Freedman, J.E., Gerstein, M., Mick, E., Rozowsky, J., Levy, D., Kitchen, R., Das, S., Shah, R., Danielson, K., Beaulieu, L., *et al.* (2016). Diverse human extracellular RNAs are widely detected in human plasma. Nat Commun *7*, 11106.
Fu, G.K., Xu, W., Wilhelmy, J., Mindrinos, M.N., Davis, R.W., Xiao, W., and Fodor, S.P. (2014). Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. Proc Natl Acad Sci U S A *111*, 1891-1896.

Glazar, P., Papavasileiou, P., and Rajewsky, N. (2014). circBase: a database for circular RNAs. RNA *20*, 1666-1670.
Griffiths-Jones, S. (2004). The microRNA Registry. Nucleic Acids Res *32*, D109-111.
Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S.*, et al.* (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res *22*, 1760-1774.
Hasan, N.A., Young, B.A., Minard-Smith, A.T., Saeed, K., Li, H., Heizer, E.M., McMillan, N.J., Isom, R., Abdullah, A.S., Bornman, D.M.*, et al.* (2014). Microbial community profiling of human saliva using shotgun metagenomic sequencing. PLoS One *9*, e97699.
Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.
Lasser, C., Shelke, G.V., Yeri, A., Kim, D.K., Crescitelli, R., Raimondo, S., Sjostrand, M., Gho, Y.S., Van Keuren Jensen, K., and Lotvall, J. (2017). Two distinct extracellular RNA signatures released by a single cell type identified by microarray and next-generation sequencing. RNA Biol *14*, 58-72.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.
Mudge, J.M., and Harrow, J. (2015). Creating reference gene annotation for the mouse C57BL6/J genome assembly. Mamm Genome *26*, 366-378.
Patton, J.G., Franklin, J.L., Weaver, A.M., Vickers, K., Zhang, B., Coffey, R.J., Ansel, K.M., Blelloch, R., Goga, A., Huang, B.*, et al.* (2015). Biogenesis, delivery, and function of extracellular RNA. J Extracell Vesicles *4*, 27494.
Sai Lakshmi, S., and Agrawal, S. (2008). piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. Nucleic Acids Res *36*, D173-177.
Skog, J., Wurdinger, T., van Rijn, S., Meijer, D.H., Gainche, L., Sena-Esteves, M., Curry, W.T., Jr., Carter, B.S., Krichevsky, A.M., and Breakefield, X.O. (2008). Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers. Nat Cell Biol *10*, 1470-1476.
Tataruch-Weinert, D., Musante, L., Kretz, O., and Holthofer, H. (2016). Urinary extracellular vesicles for RNA extraction: optimization of a protocol devoid of prokaryote contamination. J Extracell Vesicles *5*, 30281.
Yanez-Mo, M., Siljander, P.R., Andreu, Z., Zavec, A.B., Borras, F.E., Buzas, E.I., Buzas, K., Casal, E., Cappello, F., Carvalho, J.*, et al.* (2015). Biological properties of extracellular vesicles and their physiological functions. J Extracell Vesicles *4*, 27066.
Yeri, A., Courtright, A., Reiman, R., Carlson, E., Beecroft, T., Janss, A., Siniard, A., Richholt, R., Balak, C., Rozowsky, J.*, et al.* (2017). Total Extracellular Small RNA Profiles from Plasma, Saliva, and Urine of Healthy Subjects. Sci Rep *7*, 44061.
Yuan, T., Huang, X., Woodcock, M., Du, M., Dittmar, R., Wang, Y., Tsai, S., Kohli, M., Boardman, L., Patel, T.*, et al.* (2016). Plasma extracellular RNA profiles in healthy and cancer patients. Sci Rep *6*, 19413.