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. The relative stability of exRNA within extracelluar vesicles (EVs) or in association with 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 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. Quality control prior to sequencing for samples derived from EV and exosome preparations is difficult due to lack of reliable 'housekeeping' markers, such as the ratio of 18S and 28S ribosomal RNAs. Variable presence of rRNA in mixtures of low- and high-density EVs; deterministic cleavage of structured smallRNA (tRNAs and piRNAs) and long 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 present at detectable levels in some biofluids, necessitating careful analysis to ensure these sequences are not in fact derived from endogenous RNA molecules. For these reasons, existing tools capable of aligning smallRNAseq datasets are not 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 main 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:

We developed the exceRpt pipeline to process and analyze RNA-seq data generated during exRNA profiling. It is composed of a series of discrete computational analyses designed to combat problems with sample quality and contamination by first aligning to sequence libraries expected to be present in a sample before aligning to libraries with a lower degree of expectation. The pipeline is highly modular by design, allowing the user to define which smallRNA libraries are used during read-mapping; for example, 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 preprocessing by automatically identifying and removing 3' adapters. 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. 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 then applies 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 the endogenous genome and transcriptome, and transcript abundances are calculated using custom software. 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. Strong correlations in miRNA counts produced by exceRpt and miRDeep2 were observed for both high and low coverage of samples (Figure S1). This concordance is likely due to the fact that miRNA sequences annotated in miRBase are typically well curated and are not prone to being confused for contamination. However, other transcript biotypes such as piRNAs and circularRNAs are not usually so reliable.

The custom endogenous transcriptome quantification engine was developed to support the random barcode quantification and support user-modifiable library prioritization. Based on the variety of RNA preparations available (totalRNA, smallRNA, miRNA), the libraries (miRBase, tRNAscan, piRNA, gencode, circRNA) can be set different priorities such that reads from a miRNA-seq prep can be assigned to miRBase miRNAs over Gencode miRNA annotations, facilitating interpretation. Likewise, reads from long or total RNA preparations can be assigned to Gencode transcripts before (or instead of) the other smallRNA libraries. This is particularly relevant to lower-confidence annotations; for example, piRNAs should be given lower priority than tRNAs to ensure correct read assignments.

Running the same sample through the pipeline with individual steps removed shows the effect of these 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.

Third, reads are aligned to curated libraries of annotated exogenous miRNAs in miRBase and exogenous rRNA sequences in the Ribosomal Database Project (RDP). Reads not aligned in the preceding stages are aligned to the full genomes of all sequenced bacteria, viruses, plants, fungi, protists, metazoa, and selected vertebrates. 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. By characterizing exogenous genomes alignments generated by exceRpt in terms of the NCBI taxonomy tree, users may obtain valuable information regarding the distribution of the flora in various exRNA samples and generate

phylogenetic trees for cross-sample comparison. This may be particularly interesting for samples from environments with robust microbiota such as saliva (Figure 2C).

To ensure that small exRNA-Seq datasets have high standards, the ERCC developed the following QC metrics: (1) individual RNA-Seq datasets are required to have a minimum of 100,000 reads that overlap with any annotated RNA transcript in the host genome, and (2) the fraction of reads that align to any annotated RNA over the reads that map to the host genome should be greater than 0.5. The first criterion ensures that a sufficient number of reads are generated to quantify the RNAs in the sample. The second criterion ensures that sample reads mostly align to RNA, as samples with low fractions likely have DNA contamination possibly from cellular sources. The metrics are uniformly applied by exceRpt on all exRNA samples. Applying these criteria to a set of 595 exceRpt-processed smallRNA-Seq data, we find that 557 (93.6%) meet both criteria (Figure 1C), with most datasets well above threshold.

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. 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 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 2B). This information is critical for understanding the composition of particular exRNA profiles and for interrogating their biological significance.

Discussion:

Due to the unique technical challenges of exRNA profiling, there was a need for a standardized processing platform. As the main smallRNA-seq pipeline of the ERCC, exceRpt has processed all of the datasets (over 2000 samples) in the 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 in a user-friendly, browser-based interface available at Genboree.org. As the field expands, exceRpt will play an increasingly important role in characterizing the functions of and developing biomarkers based on extracellular RNA.

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:

[FILL]

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

[METHOD DETAILS]

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 Docker image with all required dependencies may be used for installation on the user's own machine or cluster: https://hub.docker.com/r/rkitchen/excerpt/.
The exceRpt source code may be downloaded and installed manually for the most amount of flexibility: 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.
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.
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.
Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.

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.
Sai Lakshmi, S., and Agrawal, S. (2008). piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. Nucleic Acids Res *36*, D173-177.
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.
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.