**Introduction**

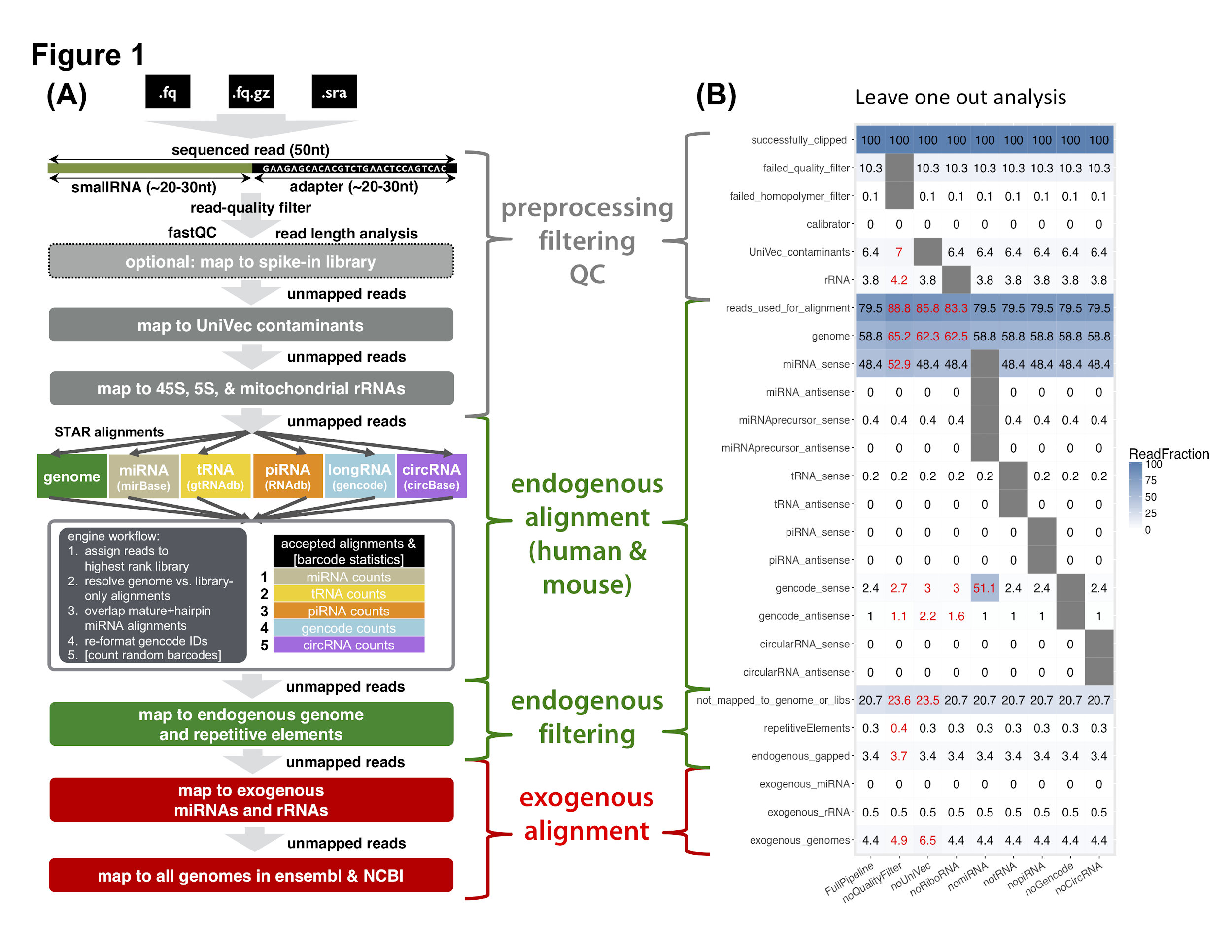
Recent research indicate that RNA can play significant roles as extracellular signaling molecules [1, 2], existing not only in exosomes or other vesicles, but also outside vesicles bound to carrier proteins, such as Argonaute (Ago) proteins [3, 4]. RNA has been associated with autocrine, paracrine, and endocrine signaling [1], and is therefore of great potential significance to many biological processes. Most body fluids contain microRNAs (miRNAs) and other non-coding RNAs (ncRNAs) [5], some of which have been identified as markers for various biological functional states, including pathologies, cancers, and toxicity effects [6-9]. Despite previous observations of widespread signaling via extracellular RNA (exRNA), little is known about the regulatory machinery that determines whether vesicle mediation occurs, the mechanisms guiding transport of specific RNAs from source to target cells, and target-cell uptake and the nature of the signal transmitted. It is also unclear when and why exosomal inclusion is needed, and how inclusion affects biological function.

Fully addressing these mechanisms requires exploring the biophysical properties of exRNA communication. Many exRNAs are transported in exosomes. Exosomes are membrane vesicles, between 30 and 100 nm in size, that are secreted by most cells. The molecular composition of exosomes depends on their size and cell of origin, and includes a mixture of membrane and cytoplasmic proteins, RNAs, and lipids [10, 11]. Previous studies have identified over 200 types of proteins [10] and a significant amount of RNA [10] [e.g., ribosomal RNA (rRNA), messenger RNA (mRNA), miRNA, and small RNA] within exosomes. Even though the general molecular structure of the vesicles is similar to that of the plasma membrane of the producing cell, the composition of membrane proteins, such as tetraspanins, might not be the same [12, 13]. These proteins and associated molecules might determine exosome target cell selection. Additionally, proteomic studies have identified subsets of cellular components that are targeted specifically to exosomes [14]. The variety of cellular material found in exosomes and their role in cell-cell communication is one of the primary foci of our exRNA project.

Furthermore, in addition to endogenous exRNAs (en-exRNAs), recent studies have reported the presence of exogenous exRNAs (ex-exRNAs) in circulation in the blood [3, 4]. Exogenous oligonucleotides can be taken up by cultivated cells, making their presence in plasma even more significant [15-17]. Using next-generation sequencing technology, we have observed that a significant fraction of the circulating RNA appears to originate from exogenous species [18, 19], including bacteria and fungi, as well as from plants (for example, miRNAs from plant-based food) [20]. Some of these RNAs can also be detected in intracellular complexes *in vitro* and thus may affect cell function. We have discovered that some of these molecules form a selective pathway from the gut to blood, suggesting a role as signaling molecules mediating the human-microbiome interaction [18, 19]. Ex-exRNAs represent an important component of the spectrum of exRNA; thus, we expect them to significantly contribute to the total exRNA-seq signal. Although studies have observed or proposed the above sources of ex-exRNA, to our knowledge no large-scale comparative study of multiple sources has been carried out.

**We have extensive experience in developing pipelines for analyzing exRNA-Seq data**

We built the extracellular RNA processing toolkit (exceRpt), a standardized small RNA-Seq analysis pipeline of the NIH Extracellular RNA Communication Consortium (ERCC) optimized for exRNA analysis, by adapting existing RNA-Seq workflows for mRNAs and miRNA (***Figure 1***). We built the exceRpt pipeline to address the need for a standardized bioinformatics processing platform in exRNA research. The toolkit is structured as a principled, biologically driven series of alignment, filtering, and quantification steps in which unmapped reads are used as inputs for the next step. The prioritization of steps is biased in favor of conservative estimates for RNA quantifications, with higher-confidence RNA reference sets (by degree of expectation or annotation quality) having higher priority. Our pipeline combines analysis methods and tools targeting RNA-Seq data, connects contig assembly and alternative splicing workflows with modifications of processing of secondary structure of small RNAs, and assigns reads to their species of origin, whether human, microbial, or plant. We have also addressed detection biases in the accurate measurement of RNAs smaller than sequencing read lengths. To date, exceRpt has uniformly processed and applied QC standards to all of the datasets in the ERCC exRNA Atlas (<http://exrna-atlas.org/>) [21]. The exceRpt pipeline enables a variety of user-specified customizations, including RNA reference prioritization, random barcoding, spike-in support, and detailed quantification reports. The pipeline is available as source (at github.gersteinlab.org/exceRpt) or wrapped in a user-friendly, browser-based interface available at genboree.org.



**Figure 1 exceRpt schema** (A): exceRpt schema: Samples in FASTA, FASTQ or SRA file formats are used as inputs to exceRpt. Adapter and random barcode sequences are removed, followed by a read-quality filter, optional spike-in quantification and removal, and UniVec contaminant removal. High quality filtered 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 miRNA, rRNA, and genomic libraries. (B): Leave-one-out analysis: Running the pipeline multiple times with individual steps removed shows the effect of those steps on subsequent alignments. The sample used for this analysis was SRR822433, a plasma exRNA plasma sample. 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. Removing the UniVec alignment step significantly increases the number of reads that, likely incorrectly, map to the exogenous genomes.

**We have extensive experience in transcriptome and exRNA analysis**

RNA-seq is essential for transcriptome studies. We have extensive experience in developing RNA-Seq processing pipelines [22] as part of the ENCODE and modENCODE consortia [23]25164755 [24]. These pipelines could work on RNA expression analysis, RNA variant-calling, and RNA fusion detection techniques. We have developed tools and methodologies for identifying non-coding transcripts and novel transcribed elements in the genome [24-28]. We have developed several tools and data formats to handle large quantities of data generated by RNA-Seq experiments [22, 24, 29]. We also developed statistical methods for data mining on the large-scale RNA-Seq datasets from ENCODE and modENCODE. In addition, privacy is receiving much attention with the unprecedented increase in the breadth and depth of biomedical datasets, particularly personal RNA-Seq datasets. To this end, we have developed methodologies to analyze sensitive information leakage from phenotype datasets [30, 31].We also have identified a set of disease-associated small exRNAs [32-34], and improved their functional annotations [35].

**References**

1. Dinger, M.E., T.R. Mercer, and J.S. Mattick, *RNAs as extracellular signaling molecules.* J Mol Endocrinol, 2008. **40**(4): p. 151-9.

2. Tsang, J.C. and Y.M. Lo, *Circulating nucleic acids in plasma/serum.* Pathology, 2007. **39**(2): p. 197-207.

3. Turchinovich, A., et al., *Characterization of extracellular circulating microRNA.* Nucleic Acids Res, 2011. **39**(16): p. 7223-33.

4. El-Hefnawy, T., et al., *Characterization of amplifiable, circulating RNA in plasma and its potential as a tool for cancer diagnostics.* Clin Chem, 2004. **50**(3): p. 564-73.

5. Weber, J.A., et al., *The MicroRNA Spectrum in 12 Body Fluids.* Clinical Chemistry, 2010. **56**(11): p. 1733-1741.

6. Garcia, J.M., et al., *Extracellular plasma RNA from colon cancer patients is confined in a vesicle-like structure and is mRNA-enriched.* Rna, 2008. **14**(7): p. 1424-1432.

7. Taylor, D.D. and C. Gercel-Taylor, *MicroRNA signatures of tumor-derived exosomes as diagnostic biomarkers of ovarian cancer.* Gynecologic Oncology, 2008. **110**(1): p. 13-21.

8. DeWitte-Orr, S.J., et al., *An Accessory to the 'Trinity': SR-As Are Essential Pathogen Sensors of Extracellular dsRNA, Mediating Entry and Leading to Subsequent Type I IFN Responses.* Plos Pathogens, 2010. **6**(3).

9. Turchinovich, A., et al., *Characterization of extracellular circulating microRNA.* Nucleic Acids Research, 2011. **39**(16): p. 7223-7233.

10. Valadi, H., et al., *Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells.* Nat Cell Biol, 2007. **9**(6): p. 654-9.

11. Taylor, D.D. and C. Gercel-Taylor, *MicroRNA signatures of tumor-derived exosomes as diagnostic biomarkers of ovarian cancer.* Gynecol Oncol, 2008. **110**(1): p. 13-21.

12. Nazarenko, I., et al., *Cell surface tetraspanin Tspan8 contributes to molecular pathways of exosome-induced endothelial cell activation.* Cancer Res, 2010. **70**(4): p. 1668-78.

13. Rana, S. and M. Zoller, *Exosome target cell selection and the importance of exosomal tetraspanins: a hypothesis.* Biochem Soc Trans, 2011. **39**(2): p. 559-62.

14. Thery, C., L. Zitvogel, and S. Amigorena, *Exosomes: composition, biogenesis and function.* Nat Rev Immunol, 2002. **2**(8): p. 569-79.

15. Kole, R. and P. Sazani, *Antisense effects in the cell nucleus: modification of splicing.* Curr Opin Mol Ther, 2001. **3**(3): p. 229-34.

16. Feinberg, E.H. and C.P. Hunter, *Transport of dsRNA into cells by the transmembrane protein SID-1.* Science, 2003. **301**(5639): p. 1545-7.

17. Valiunas, V., et al., *Connexin-specific cell-to-cell transfer of short interfering RNA by gap junctions.* J Physiol, 2005. **568**(Pt 2): p. 459-68.

18. Wang, K., et al., *Comparing the MicroRNA spectrum between serum and plasma.* PLoS One, 2012. **7**(7): p. e41561.

19. Wang, K., et al., *The complex exogenous RNA spectra in human plasma: an interface with human gut biota?* PLoS One, 2012. **7**(12): p. e51009.

20. Zhang, L., et al., *Exogenous plant MIR168a specifically targets mammalian LDLRAP1: evidence of cross-kingdom regulation by microRNA.* Cell Res, 2012. **22**(1): p. 107-26.

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

22. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.

23. Djebali, S., et al., *Landscape of transcription in human cells.* Nature, 2012. **489**(7414): p. 101-8.

24. Gerstein, M.B., et al., *Comparative analysis of the transcriptome across distant species.* Nature, 2014. **512**(7515): p. 445-8.

25. Bertone, P., et al., *Global identification of human transcribed sequences with genome tiling arrays.* Science, 2004. **306**(5705): p. 2242-6.

26. Clark, M.B., et al., *The reality of pervasive transcription.* PLoS Biol, 2011. **9**(7): p. e1000625; discussion e1001102.

27. Lu, Z.J., et al., *Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data.* Genome Res, 2011. **21**(2): p. 276-85.

28. Rozowsky, J.S., et al., *The DART classification of unannotated transcription within the ENCODE regions: associating transcription with known and novel loci.* Genome Res, 2007. **17**(6): p. 732-45.

29. Habegger, L., et al., *RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries.* Bioinformatics, 2011. **27**(2): p. 281-3.

30. Harmanci, A. and M. Gerstein, *Quantification of private information leakage from phenotype-genotype data: linking attacks.* Nat Methods, 2016. **13**(3): p. 251-6.

31. Greenbaum, D., et al., *Structuring supplemental materials in support of reproducibility.* Genome Biol, 2017. **18**(1): p. 64.

32. Shah, R., et al., *Discordant Expression of Circulating microRNA from Cellular and Extracellular Sources.* PLoS One, 2016. **11**(4): p. e0153691.

33. Freedman, J.E., et al., *Diverse human extracellular RNAs are widely detected in human plasma.* Nat Commun, 2016. **7**: p. 11106.

34. Mick, E., et al., *Stroke and Circulating Extracellular RNAs.* Stroke, 2017. **48**(4): p. 828-834.

35. Cheung, K.H., et al., *Extending gene ontology in the context of extracellular RNA and vesicle communication.* J Biomed Semantics, 2016. **7**: p. 19.