Developing Quality Control Metrics for Extracellular RNA Communications Consortium (ERCC) exRNA-Seq Data

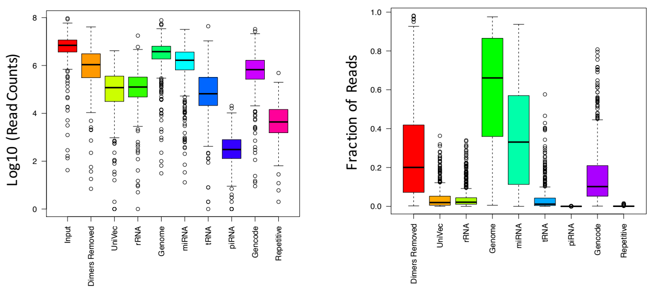
Rozowsky J[1,2], Kitchen R[1,2,3], Diao J[2], Subramanian S[4], Roth M[4],
Milosavljevic A[4], Jensen K[5], Gerstein M[1,2,6]

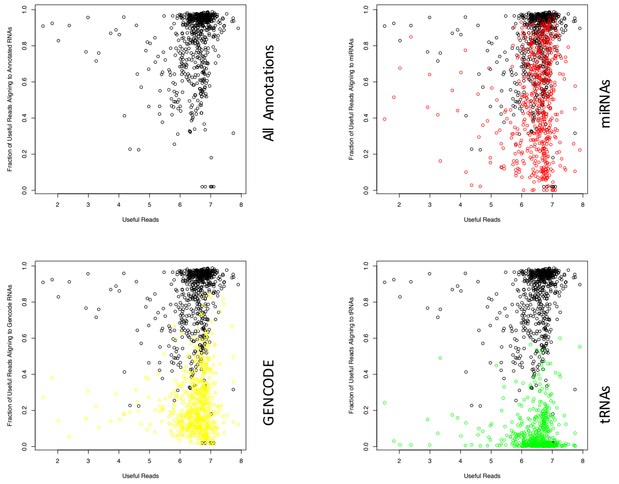
1. Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT. 2. Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT. 3. Department of Psychiatry, Yale University, New Haven, CT. 4. Bioinformatics Research Laboratory, Molecular and Human Genetics Department, Baylor College of Medicine, Houston, TX. 5. Neurogenomics Division, Translational Genomics Research Institute, Phoenix, AZ. 6. Department of Computer Science, Yale University, New Haven, CT

We present the current progress toward developing quality control (QC) standards for data being generated by the Extracellular RNA Communications Consortium (ERCC). In particular, we will focus on the QC metrics for small exRNA-Seq data which forms the majority of the data that is currently being produced by members of the consortium. Metrics such as for example mapped read counts and fraction of reads mapping to categories such as annotated miRNAs or potential laboratory contaminants. The metrics presented here were developed using data from seven different ERCC labs that were shared with the DMRR for the purposes of this analysis.

Many of the sample-level QC metrics discussed here are already built in to the exceRpt small-RNA pipeline, which automatically reports detailed QC information for each sample. However there is a clear desire amongst the international exRNA community to understand and interpret the quality of their RNA-seq samples. We therefore intend to include automatic QC evaluation of any sample submitted by any user based on the current set of 'gold-standard' exRNA samples provided by ERCC members. We discuss the various QC metrics, and acceptable tolerances, for what we define as 'high quality' RNA-seq data from an extracellular preparation.

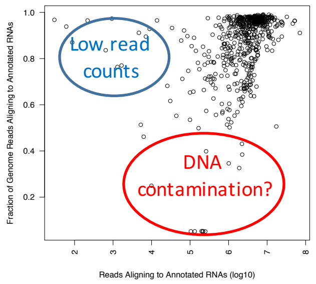
Exploration of 595 small exRNA-Seq datasets

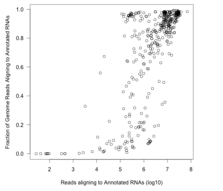




Useful Reads = Input Reads – Reads Mapping to Contaminants, rRNAs or Adaptor Dimers

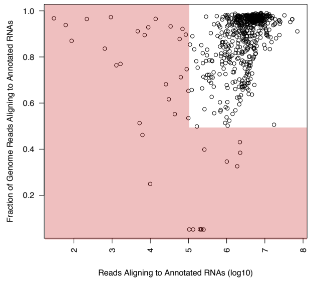
Map of 595 small exRNA-Seq Datasets





400 cellular small RNA-Seq datasets from SRA

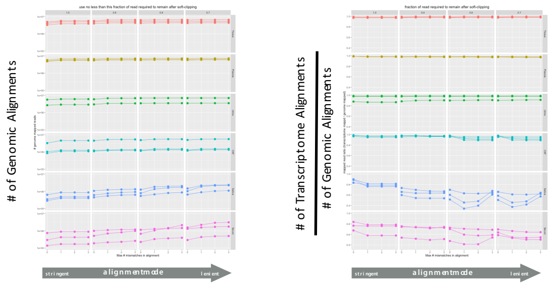
QC Standards for exRNA-Seq Datasets_



581/595 (94.1%) datasets meet both criteria

1. $\textit{Reads mapping to annotated RNA} > 10^5$
2. $\dfrac{\textit{Reads mapping to annotated RNAs}}{\textit{Reads mapping to Genome}} > 0.5$

Parameters for Genomic Alignments



Acknowledgements

Funding from the NIH Common Fund Extracellular RNA Communication Consortium.

Members of the ERCC consortium for the data presented (data available from exRNA.org).