Data Integration and Analysis Component (DIAC) of the DMRR

Continued Development and Support for Extracellular RNA-Seq Analysis Pipeline

We have developed a custom analysis pipeline, the extracellular RNA processing toolkit (exceRpt), for uniformly processing extracellular RNA-Seq data from participants in the ERCC. We have currently submitted a paper describing the current version of this paper as part of the ERCC paper package to Cell Systems. We would like to continue the development and support for this analysis pipeline in response to the needs of the reference profiling grants which will still be generating of RNA-Seq data after the end of the current funding period of the DMRR. In particular we would like to future develop the features of the RNA-Seq analysis pipeline such as improved support for and analysis of random barcodes. We will also continue support for the analysis pipeline as implemented as the DCC fixing any bugs that are encountered during the next period. We will also develop a more comprehensive documentation of the pipeline for ERCC users as well as external users. We will also continue the evaluation and potential modification of the current quality control standards that have been developed by the DIAC in consultation with the members of the consortium and the reference profiling grants. The quality control standards as applied to new datasets submitted to the DCC requires routine monitoring.

Collaboration with U01 Reference Profiling Grants

In addition to integrative analysis we will continue our pairwise interactions with members of the reference profiling grants and support for their individual analysis needs for the period where the references profiling grants are funded beyond the current DMRR funding period. In particular we have been collaborating with Prof. Jane Freedman (UMass), Prof. David Galas (PNDRI), Prof. Ionita Ghiran (Harvard) and Prof. Louise Laurent (UCSD).

Our current ongoing collaborations with these reference profiling grants are as follows: We are collaborating with Prof. Freedman about the analysis the extracellular RNA-Seq data in order to identify eQTLs for extracellular RNAs using genotype data for the sequenced samples from members of the Framingham Heart Study (FHS) and Multi-Ethnic Study of Atherosclerosis (MESA) cohorts. We are collaborating with Prof. Galas about the utility and analysis of random 4N barcodes used by the reference profiling grants in order to both compensate for adaptor ligations biases as well as account for PCR jackpotting which is a potential issue for small RNA quantification which can be mitigated using the barcodes. We are collaborating with Prof. Ghiran in order analyse his reference profiling data for extracellular RNAs for individuals during the daily circadian cycle in order to identify small RNAs that are correlated with these cycles. We are also collaborating with Prof. Laurent in order to help the development of spike-in sequences for use with extracellular RNA-Seq experiments.

We have helped analysis potentially spike-in sequences for a variety of size ranges in order to exclude sequences that have potential matches (within a number of mismatches) to both the

host genomes (human) as well as all the full exogenous genomes analyzed by the exceRpt analysis pipeline.

Results of Integrative Analysis of Data in exRNA Atlas

Once the reference profiling grants begin depositing their data to the DCC the number of RNA-Seq datasets will increase by a factor of 5-10 which will significantly improve the coverage of datasets from samples across increased number of biofluids and human physiological phenotypes (such as sex, age, etc...). This increased data will enable a more comprehensive analysis of the exRNA Atlas. We will identify a set of informative miRNAs (and other small RNA biotypes) that best enable the classification of samples from available biofluids and disease states.

We will explore developing methods enabling the normalization of the quantifications of small RNAs from extracellular RNA-Seq datasets from a variety of different experimental assays and protocols which introduce systematic differences. We will potentially use available reference datasets (using common shared RNA samples from plasma) generated by the difference U01 reference profiling centers in order to facilitate normalization. We would like to generate a huge normalized matrix served from the Atlas website with the rows of all annotated small RNAs and columns all the samples in the Atlas - this matrix will be of maximum utility to the broader scientific community in order enable even further computational integrative analysis using the data in the exRNA Atlas.

The discovery of stable RNAs in circulating fluids has focused interest towards the information-content of extracellular RNAs, especially with respect to tissue states and pathologies. However, little is known about the dominant features of RNA sequences derived from extracellular fluids and how they differ from their much better-studied cellular cousins. We would like to further study the identification of the tissue of origin for exRNAs from different biofluids using decomposition algorithms; this analysis has been somewhat limited by the sparseness of the data in earlier releases of the Atlas.

In collaboration with the DCC we would like to publish a paper describing the results of this integrative of the final set of data in the exRNA Atlas as well as making the results of this analysis including the outputs of the different analyses available via the exRNA Atlas portal for members of the wider scientific community to use.

Tools for the Analysis of Exogenous RNAs

As the last step in the exceRpt pipeline once we have identified reads from extracellular RNA-Seq datasets that could potentially come from the host genome we take the remaining reads and map them against potential exogenous sources. We first align these reads against databases of exogenous ribosomal RNAs and miRNAs and then we align the remaining reads

against the full genome sequences of all available bacteria, fungi, protist, viruses and selected eukaryotes. We then independently construct the phylogenetic trees of reads abundance mapping to the the exogenous genomes and exogenous rRNAs. Reads that multimap to different exogenous genomes are assigned to the most parsimonious node in the phylogenetic tree consistent with the mappings.

In collaboration with Prof. David Wong (UCLA) we have applied this pipeline to cell-free extracellular RNA-Seq datasets from saliva which contains a large fraction of RNA-Seq reads from bacterial genomes mostly corresponding the oral microbiome (Kaczor-Urbanowicz *et al.* 2018). While saliva seems somewhat special in that it contains a significantly large fraction of exogenous RNA content we have observed this phenomena to a lesser extent in many other human biofluids. We would like to further develop the analysis of exogenous RNAs present in human extracellular biofluids. In particular we would like to further improve the algorithms for aligning potential exogenous reads to the non-human genome sequences which is very computationally intensive. We would like to update and expand the number of exogenous genomes used using newer sequenced genomes in NCBI that have been added in the last year. We would also like to improve the algorithms for constructing the phylogenetic read count trees for exogenous genome and rRNA reads and develop novel methods for comparing the significance of the concordance of these trees which is intuitively evident by eye.

Development of Tools for Visualization of Data in exRNA Atlas

Using public data from the Sequence Read Archive (SRA) and the exRNA Atlas, we would like to explore how RNA-Seq samples cluster tightly by tissue as well as biofluid of origin. These patterns are demonstrable as visual patterns on dimensionality-reduced plots (such as PCA), and are even better classifiable using nonlinear supervised techniques (such as t-SNE, see Fig. XX). Furthermore, we have shown using preliminary data that the miRNA signatures of related tissues and biofluids share similar signatures and allow for meaningful cross-predictions and classifications. Our work echoes a growing need to develop standardized data collection and processing tools to account for stratification by sample origin. We have developed a prototype tool for visualizing the data in the exRNA Atlas using a variety of dimensionality reduction techniques (https://exrna-atlas.org/exat/precomputedJobs). We would like to further develop this visualization tool implementing and developing novel methodologies for exploring the extracellular RNA data.

We would like to identify a subset of small RNAs that are most informative for classifying samples from different biofluids and disease states. While we have made initial attempt to perform this analysis this will be much better enabled using the dense data in the Atlas once the data from reference profiling grants are available. We would also like to compare how the classification of samples compares using different small RNA biotypes (such as miRNAs as compared to tRNAs or piRNAs).

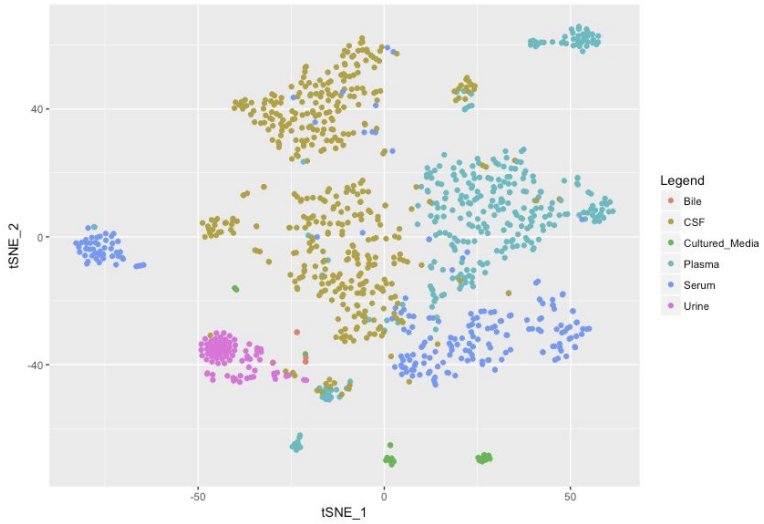

Figure XX: t-SNE decomposition of exRNA-Seq data in Atlas. The resulting plot reveals strong separation by biofluids.

Notes from MBG
5 to 10
pages falling in w/in spec aims
tasks
3-4 pages
tasks -
integrativa analysis
publish a final paper

profIles
final bolis
==
incremental improvement or extensions

file - data resource of standaridized expression levels
annotation of ncrna
by biofluid
level of expression
carrier

==
collabs w mapping centers
focus on u01 - ref.

==
exogenous exrna - theme - look at that Q
couldn't have been addressed by indiv inv
david wong - saliva
no meetings
u13 for a conf.