Stratification of extracellular miRNA samples by biofluid of origin

James Diao[1]

Abstract
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 bodily fluids and how they differ from their better-studied intracellular cousins. Using public RNA-Seq data from the Sequence Read Archive and the exRNA Atlas, we demonstrate that samples most strongly cluster by tissue and biofluid of origin. These patterns are demonstrable as visual clusters on dimensionality-reduced plots, and can be predicted using supervised machine learning techniques. Moreover, the miRNA signatures of related tissues and biofluids are demonstrated to be loosely correlated, with related signatures and meaningful cross-predictions. The data echoes a growing need to develop standardized data collection and processing tools to account for stratification by sample origin.

[1] *Department of Molecular Biophysics and Biochemistry, Yale University*

Contents

1. Introduction

RNA molecules were once thought to exist only as stable molecules inside cells. However, microarray and RNA sequencing experiments have discovered many classes of RNA in extracellular fluids, including messenger (mRNA), transfer (tRNA), micro (miRNA), small interfering (siRNA), piwi-interacting (piRNA), and long non-coding (lncRNA).[1, 2] These molecules are stably present in plasma, cerebrospinal fluid (CSF), urine, bile, and saliva despite the presence of RNase in these fluids.[3] Since their discovery, extracellular RNAs have been implicated in cell-to-cell communication and associated with a wide variety of diseases.[4] Because of their well-understood cellular export processes, miRNAs have been the focus of most work on exRNA. miRNAs are short (22 nucleotides) single-stranded non-coding RNAs that base pair with complementary mRNA to regulate gene expression for apoptosis, proliferation, differentiation, and other functions.[1] Because miRNAs have tissue-specific expression patterns, they are particularly promising as a class of molecular biomarkers for the early detection of disease.[3, 5, 6] However, associating signatures with phenotypes will necessarily require a deeper understanding of the biases and convolutions that underlie RNA sequencing data.

2. Methods

2.1 Data Description

exRNA Atlas Small RNA and RT-qPCR exRNA sequencing profiles from human biofluids were collected from the exRNA Atlas, the central data repository of the Extracellular RNA Communication Consortium (ERCC). The total dataset is 3,290 mapped small RNA reads vs 751 samples. Individual samples are labeled with disease condition and biofluid source.

Sequence Read Archive Small RNA sequencing profiles from human tissues were collected from the Sequence Read Archive (SRA), the NIH's primary archive for high-throughput sequencing data. The total dataset is 1,804 mapped small RNA reads vs 201 samples. Individual samples are labeled with healthy/perturbed condition and tissue of origin.

2.2 Processing

Normalization The RNA-Seq read counts were normalized to reads-per-million (RPM).

Sparsity Both data matrices had non-zero values in fewer than 25% of all data entries. Missing values were imputed by iteratively redefining each sample as the average of its nearest neighbors. This reduces the sparsity of the matrix and was shown to have little effect on the general cluster structure.

Dimensionality Reduction We reduced the very high dimensionality of the data from more than 2,000 to only the top 30 principle components (PCs). These PCs together explain 58% and 66% of the variance in the extracellular and cellular data, respectively.

Outlier Filtering The 2nd principle component almost perfectly separates out plasma samples from the TPATE healthy vs. cancer dataset. Since TPATE derives from diverse patient phenotypes but does not cluster with other plasma-derived data, this is likely due to a significant batch effect. The other 10 datasets were indistinguishable under any pairwise top PCA components. We chose to exclude this anomalous dataset from the exRNA analysis (192/1267 samples). After removal, the PCA plots show no separation of the data, suggesting that the structure of the data is strongly non-linear.

2.3 Unsupervised Analysis

tSNE Visualization The availability of sample labels makes color-coded visualization a reliable heuristic for evaluating cluster structure. Besides the principle components analysis used to perform dimensionality reduction and data filtering, we also visualized the data along the top 2 dimensions given by a t-distributed stochastic neighbor embedding (tSNE). Additionally, Gaussian mixture models fit to the data, and resulting cluster labels were used to color the plots as well.

2.4 Supervised Classification

Techniques Supervised machine learning techniques were applied for classification of exRNA samples by biofluid. Classification was performed on dimensionality-reduced features (top 30 PCA axes). Because the data has been shown to be distinctly nonlinear, we have avoided using linear classifiers. Instead, we use random forests and support vector machines to identify key decision features.

Cross-Classification The classifier could be applied to cellular samples as well, to ascertain the relationship between specific tissue types and biofluids. Expected matches include saliva/lung, brain/CSF, and liver/bile.

Biological Relevance These non-linear methods will be compared to the non-negative matrix factorization (NMF) method typically used for mRNA gene expression data, and compared with a decision tree for biological relevance.

3. Results

3.1 Unsupervised Analysis

Visualization The plot in figure 1 reveals strong separation by biofluids. But because datasets are often homogeneous in biofluid origins, we should be wary of batch effects. This may be observed by comparing figure 1 with figure 2 (colored by dataset membership). The separation between KJENS and LLAUR serum samples suggest such an effect. On the other hand, the internal clustering of CSF, plasma, and serum within the KJENS data seems to provide strong evidence for robust biofluid stratification.

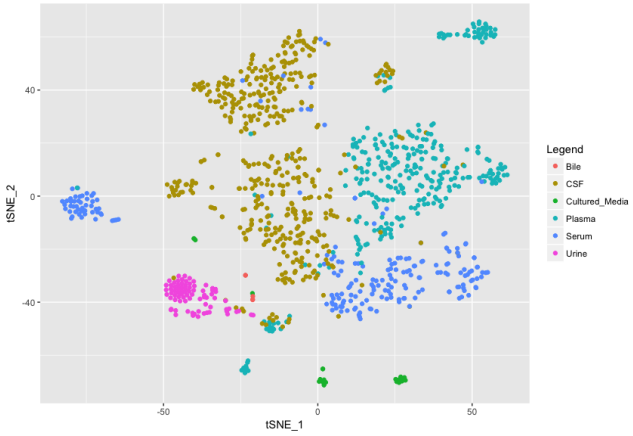

Figure 1. tSNE: exRNA Clustering by Biofluid

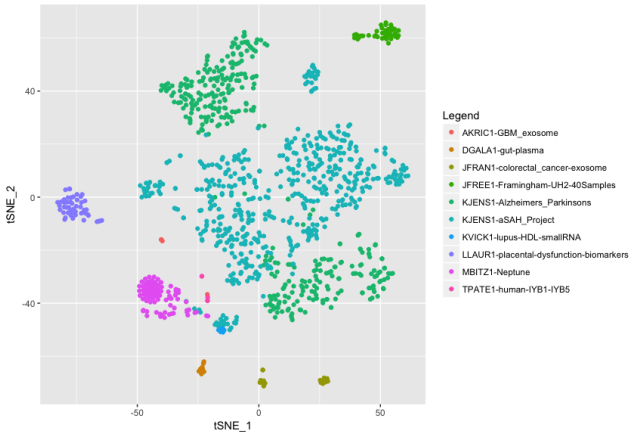

Figure 2. tSNE: exRNA Clustering by Dataset Membership

Gaussian mixture models (GMM) The data was modeled as a weighted sum of Gaussian distributions. This model cleanly identifies some of the tighter clusters, and if an accurate representation, suggests subdivision within groups.

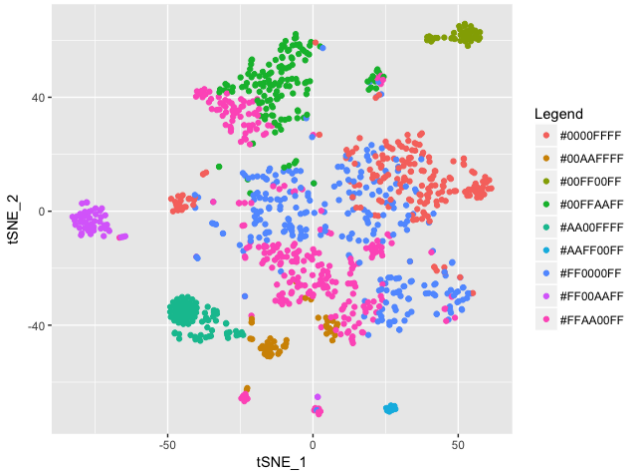

Figure 3. tSNE: exRNA Clustering by GMM

3.2 Supervised Classification

Accuracy Various models were trained on 80% of the extracellular and cellular miRNA samples by 10-fold cross-validation. These included random forests (RF), support vector machine (SVM), decision tree (DT), and non-negative matrix factorization (NMF). The error from the validation set was used to estimate the accuracy of the classifier, and these numbers were corroborated by model accuracy in the test set (20% of samples).

Table 1. Accuracy of Biofluid Classifiers

Method	Cellular	Extracellular
RF	82.6%	93.4%
SVM	80.0%	87.1%
DT	74.2%	80.9%
NMF	74.2%	79.6%

This suggests that non-linear methods such as random forests and support vector machines are the ideal methods. We hypothesize that NMF performs less well because it assumes that each sample comes from a linear combination of categories, while the other two methods do not make such an assumption.

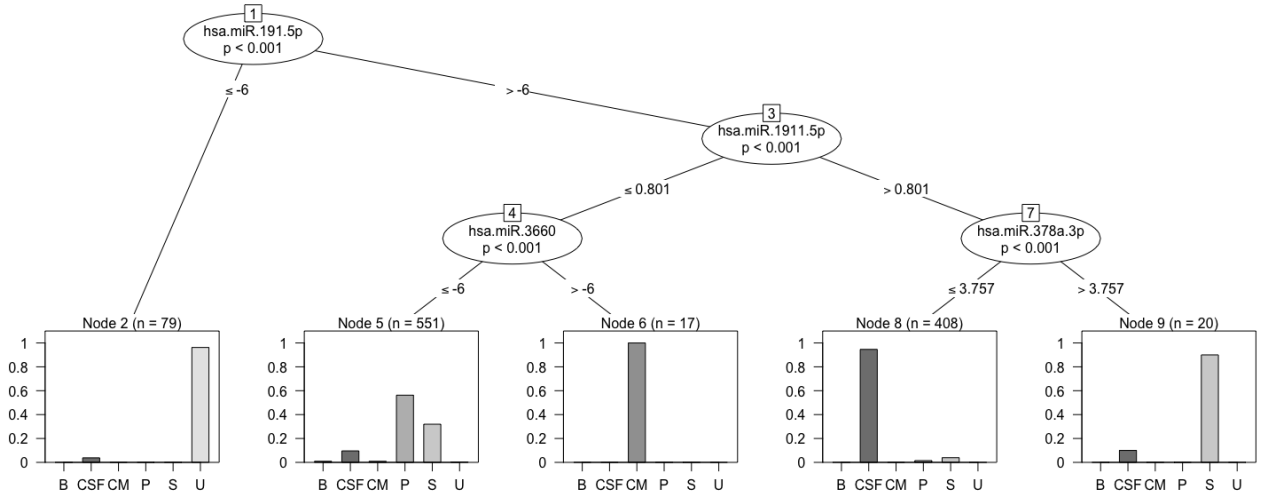

Figure 4. Decision Tree Identifies Signature miRNAs

3.3 Interpretation

Signatures Each tissue is enriched in different miRNAs, but very few (20-30) are required to make reasonable predictions. The median miRNA levels of the 6 most common tissue categories are shown below.

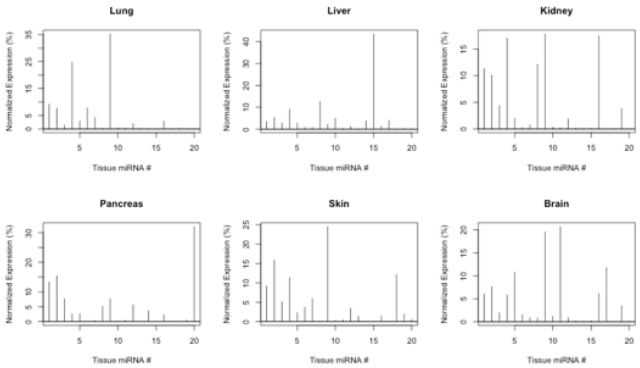

Figure 5. Tissue Signatures by miRNA

RF Importance The ranked importance scores given by the random forest support the significance of the miRNAs partitions chosen by the decision tree algorithm.

Table 2. Importance Scores of Partitioned miRNAs

miRNA	Importance	Rank
miR-378a-3p	14.4	2 / 2411
miR-1911-5p	8.9	6 / 2411
miR-486-5p	8.8	8 / 2411
miR-3660-5p	5.7	19 / 2411

Decision Tree This distinctiveness is similarly seen in the decision tree diagram [Figure 4]. For example, urinary samples are characterized by depletion in miR-191, and conditioned-media is characterized by the presence of miR-3660. Other relationships can similarly be found in this classification.

Cross-Classification Although the exRNA classifier is trained to recognize sample biofluids, it is possible to apply this to cellular samples to examine correlations between biofluids and tissue types. We chose to use the three-layer decision tree for its simplicity. Most of the cellular samples were classified into cultured media, despite the low prior coming from the biofluid samples (2%). This is reasonable because a large majority of cellular samples came from cultured tissues, such as HEK-293 liver cells. We also report that all 4 samples classified into CSF were from brain tissue. The reverse classification (using a tissue classifier on exRNA data) was unable to differentiate between samples. We hypothesize that this is because exRNA reads are far more homogeneous than cellular RNA reads.

4. Discussion

As the field of extracellular RNA matures, it will become increasingly important to establish standards and quality controls for the diverse data being generated and processed in various laboratories.

The data suggests that the miRNAs are strongly stratified by biofluid, even after accounting for batch effects. Distinctive miRNAs were identified by decision tree, and corroborated by random forest importance scores and cross-classification of cellular RNAs. We present evidence that the distribution of exRNAs by biofluid cannot be modeled as a linear combination. This suggests that matrix factorization techniques common for mRNA analysis (e.g. DeconRNASeq) would be less effective for exRNA. Additionally, we show that classification models can effectively isolate the miRNAs that are distinctive to certain biofluids.

Further work to develop a stratification workflow would be useful for three reasons. First, a controlling for stratification will allow association studies of greater power, if they do not have to limit their samples to a single biofluid. Second, labeled samples can be classified in a quality control step to detect sample mixing. Third, the distinct miRNAs in each biofluid may hint at different biomarkers that may be found in each fluid type. Still, more work is needed to interpret these findings and connect them to biological relevance. Further efforts may broaden the analysis to other small RNAs (piRNA, snRNA, tRNA, etc) and aim at a literature search on the backgrounds of signature miRNAs and their connections to tissue and disease states.

5. Acknowledgments

Thank you to everyone in the Gerstein Lab who guided my exploration of this data, and to Prof. Krishnaswamy and Nripesh, who taught me most of the analysis used in this project.

References

[1] James G Patton, Jeffrey L Franklin, Alissa M Weaver, Kasey Vickers, Bing Zhang, Robert J Coffey, K Mark Ansel, Robert Blelloch, Andrei Goga, Bo Huang, Noelle L'Etoille, Robert L Raffai, Charles P Lai, Anna M Krichevsky, Bogdan Mateescu, Vanille J Greiner, Craig Hunter, Olivier Voinnet, and Michael T McManus. Biogenesis, delivery, and function of extracellular RNA. *Journal of extracellular vesicles*, 4:27494, 2015.

[2] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63, 2009.

[3] David P. Bartel. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function, 2004.

[4] Esther E. Creemers, Anke J. Tijsen, and Yigal M. Pinto. Circulating MicroRNAs: Novel biomarkers and extracellular communicators in cardiovascular disease?, 2012.

[5] Jasmina S. Redzic, Leonora Balaj, Kristan E. van der Vos, and Xandra O. Breakefield. Extracellular RNA mediates and marks cancer progression, 2014.

[6] Zev Williams, Iddo Z Ben-Dov, Rony Elias, Aleksandra Mihailovic, Miguel Brown, Zev Rosenwaks, and Thomas Tuschl. Comprehensive profiling of circulating microRNA via small RNA sequencing of cDNA libraries reveals biomarker potential and limitations. *Proceedings of the National Academy of Sciences of the United States of America*, 110(11):4255–60, 2013.