bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A curated human cellular microRNAome based on 196 primary cell types

Arun H. Patil[1], Andrea Baran[2], Zachary P. Brehm[2], Matthew N. McCall[2], Marc K. Halushka[1*]

1 - Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD, 21205 USA

2 - Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY 14642, USA

ahanuma2@jhmi.edu

andrea_baran@urmc.rochester.edu

zachary_brehm@urmc.rochester.edu

matthew_mccall@urmc.rochester.edu

mhalush1@jhmi.edu

* Corresponding author at:
Marc K. Halushka, M.D., Ph.D.
Johns Hopkins University School of Medicine
Ross Bldg. Rm 632B
720 Rutland Avenue
Baltimore, MD 21205
410-614-8138 (ph)

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Background: An incomplete picture of the expression distribution of microRNAs (miRNAs) across human cell types has long hindered our understanding of this important regulatory class of RNA. With the continued increase in available public small RNA sequencing datasets, there is an opportunity to more fully understand the general distribution of miRNAs at the cell level.

Results: From the NCBI Sequence Read Archive, we obtained 6,054 human primary cell datasets and processed 4,184 of them through the miRge3.0 small RNA-seq alignment software. This dataset was curated down, through shared miRNA expression patterns, to 2,077 samples from 196 unique cell types derived from 175 separate studies. Of 2,731 putative miRNAs listed in miRBase (v22.1), 2,452 (89.8%) were detected. Among reasonably expressed miRNAs, 108 were designated as cell specific/near specific, 59 as infrequent, 52 as frequent, 54 as near ubiquitous and 50 as ubiquitous. The complexity of cellular microRNA expression estimates recapitulates tissue expression patterns and informs on the miRNA composition of plasma.

Conclusions: This study represents the most complete reference, to date, of miRNA expression patterns by primary cell type. The data is available through the human cellular microRNAome track at the UCSC Genome Browser (https://genome.ucsc.edu/cgi-bin/hgHubConnect) and an R/Bioconductor package (https://bioconductor.org/packages/microRNAome/).

Keywords: microRNA; cell type; genomic; resource; MirGeneDB

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Background

microRNAs (miRNAs) are short, ~18-21 bp, critical regulatory elements that block protein translation [1, 2]. miRNAs are recognized as functional regulators of development and cellular biology. They also demonstrate altered expression levels in disease states that may have biomarker potential [3, 4]. Despite their importance, miRNAs, are a challenging biomolecule to characterize. A number of attributes of miRNAs have hampered progress in this area. One is the short, 6 bp, seed sequence for target recognition that has innumerable potential targets in the genome resulting in a vast over interpretation of miRNA regulatory roles [5]. A second is what short RNA sequences should legitimately be considered bona fide miRNAs versus some other form of non-miRNA that have yet to be accurately characterized [6]. A third is the unclear distribution of miRNA expression among cell types.

The confusion surrounding miRNAs expression by unique cell types is two-fold. There is general anonymity of miRNAs in which the numerical naming scheme (ex. miR-141, miR-142, miR-143, miR-144) hides marked differences in expression patterns and function [7]. The second is that early publications of general miRNA expression focused on tissues, which are comprised of numerous cell types and the localization of miRNAs, whether cell-specific or ubiquitous was not established [8, 9].

Recently, cell-specific miRNA atlases of greater and greater complexity have been published to understand the expression patterns of this important RNA class [10-12]. Previously, we described a cellular microRNAome based upon 46 primary cell types from 161 samples [12]. Separately, FANTOM5 reported data from 123 cell types from 304 samples [11]. With the continued output of small RNA sequencing datasets that have been placed in public repositories and the development of miRge3.0, a new, faster version of our small RNA sequencing analysis tool, we decided to readdress what was known about specific cellular expression patterns of miRNAs [13].

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Herein we describe a more complete microRNAome built upon 2,077 samples from 175 public datasets across 196 primary cell types. This deeply curated resource extends our knowledge of patterns of miRNA expression across diverse cell types.

Results

Generation of miRNA results across cell types

An initial search of cell-specific small RNA sequencing datasets identified 6,054 potential samples for study. An analysis of adaptors and other sequencing-specific factors of these downloaded FASTQ files identified 4,184 runs as appropriate for further analysis. miRNA annotation and quantification was performed using the miRge3.0 pipeline on these 4,184 run FASTQ files. Over 40.7 billion reads were processed. Of the initial 4,184 runs, 871 were removed due to a lack of clustering with other appropriate samples in UMAP based clustering (**Fig. 1A**). Outlier samples were removed for being tissue-contaminated, immortalized cells, treated with infectious agents, treated with drugs, technical error during processing, low read depth, and other discrepancies. Further, of 640 samples from the RNA-Atlas project [14], 608 (the non-immune cells) consistently clustered together irrespective of their class/cell type expression. These were also removed, resulting in 2,077 final samples from 196 cell types. For 173 of the 2,077 samples, we had 329 technical replicates, resulting in 2,406 total runs. The various cell types were broadly classified from their source of origin into “class” namely, epithelial cells, endothelial cells, brain cells, fat cells, red blood cells (RBCs), immune cells, fibroblasts, stem cells, and others (unclassified cells). Plasma, not a cell type, and platelets, fragments of cells, were also included in the dataset and represented two additional classes. The class distribution is shown in **Table 1**, while the details of each cell to corresponding class (**Supplementary Table S1)** and detailed metadata information/miRge3.0 summary information (**Supplementary Table S2)** are provided elsewhere. In total, ~9 billion of the ~20 billion trimmed reads of the 2,406 runs mapped to mature miRNAs covering 89.8%

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

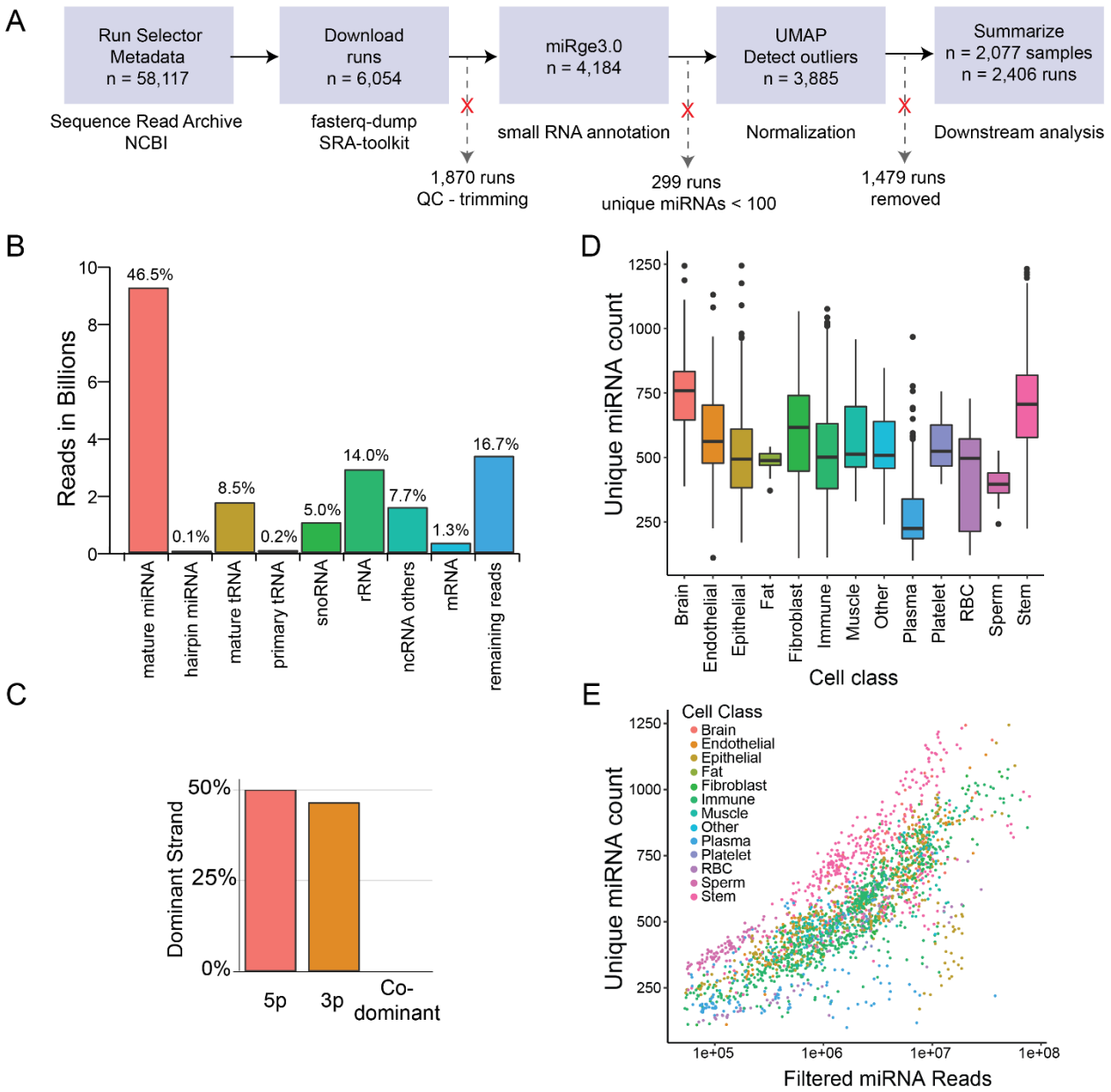

Figure 1. Overview of the human cellular microRNAome A. Workflow employed in obtaining, cleaning, and processing the human primary cell samples. **B.** The overall read distribution of the samples used across different small RNA libraries and the percent across each bar is the individual number of reads over total filtered reads (~20 billion) **C.** Strand dominance of the 5p and 3p arms among the 362 most abundant miRNAs. **D.** Distribution of unique miRNA counts for 2,406 samples across each cell class. **E.** Scatter plot of unique miRNA count abundance with the increase in sequencing depth of filtered miRNA Reads (log10).

of miRNAs in miRBase [15] and 99.6% of the mature miRNAs in MirGeneDB2.1 [16]. The read distribution across various small RNA types is provided in **Fig. 1B**, where the majority (~46%) of the

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

reads are mature miRNAs (~9 billion). Among the primary cells, 473 miRNAs had a max Reads Per Million (RPM) ≥ 1000 in at least one sample. The miRNA abundance from the 5p or the 3p arm suggests that there is no strand bias, as the dominant microRNA is equally found in both arms of the hairpin miRNA (**Fig. 1C**). An average of 546 unique miRNAs were identified across each of the 13 cell classes with the median range between 450 and 750 miRNAs (**Fig. 1D**). Plasma, which represents a collection of miRNAs from multiple cell sources, and sperm, which had low overall reads, had the fewest average number of unique miRNAs reported. As the number of miRNAs reads per sample increased, the identification of unique miRNAs increased (**Fig. 1E**). The complete read counts (**Supplementary Table S3**) and Reads Per Million mapped reads (RPM) (**Supplementary Table S4**) for all 2,406 runs mapped to miRBase annotations are available. A list of miRNAs with no reads are available in **Supplementary Table S5**.

DESeq2 VST provided superior normalization. Due to the large number of independent studies, technical causes of expression variation across shared cell types were a major concern. We employed a "leave one study out" cross-validation strategy to identify the normalization approach that resulted in the highest classification accuracy in correctly assigning cell types across 5 groups. The method assigned test samples to the cell type that minimized the Euclidean distance between the test sample and training cell type centroids. We specifically compared non-adjusted raw data to adjustments utilizing ComBat-Seq, DESeq2 VST, RUVr, RUVg and combinations of these approaches. The DESeq2 VST method, without additional batch correction, had the highest accuracy identifying cell types (96.8%, **Supplementary Table S6**) and was the normalization approach used for downstream analyses. The highest accuracy was for immune cells (99%), while the lowest accuracy was for neuron (93.6%), where ~6% matched fibroblasts, rather than neuron (**Table 2**). After appropriate normalization, a UMAP cluster of the entire dataset (**Fig. 2**) and cell-class specific clusters (**Supplementary Figs S1-11**) were generated. An HTML interactive UMAP with cell type information is available in the GitHub repository (https://github.com/mhalushka/miOme2/UMAP/Figures). These images demonstrate generalized

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

appropriate clustering of similar cell types, despite the range of studies they were pulled from. The read counts normalized with DESeq2 VST is available in **Supplementary Table S7.**

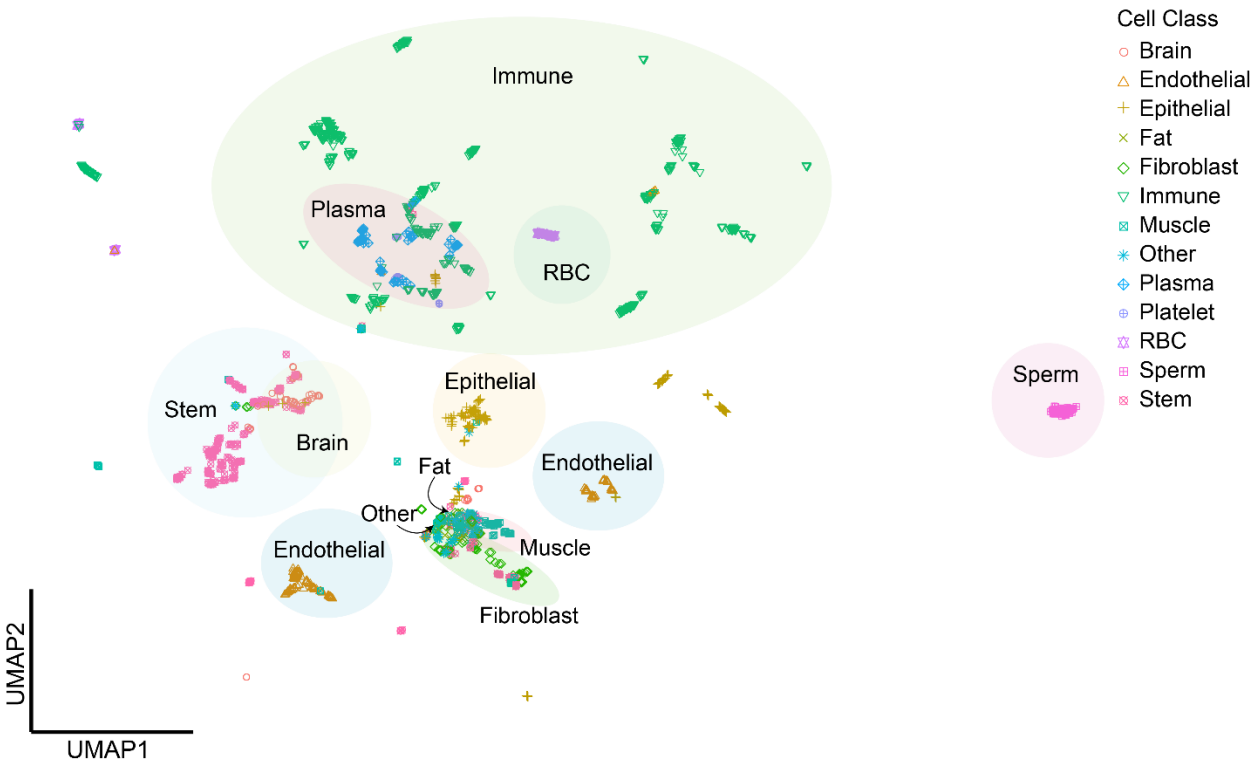

Figure 2. DESeq2 VST normalized miRNA counts are shown using UMAP clustering representing each class across all samples. There is general clustering by cell class.

Categorization of miRNAs by appearance in different cell types and class:

The cell specificity or ubiquitousness of individual miRNAs was determined across the 196 cell types. We focused only on the 323 leading strand miRNAs with a minimum RPM ≥100 of any cell type and presence in MirGeneDB. Of these, 108 were considered "cell specific/near specific" based on methods described below. This group included highly expressed miRNAs such as miR-7-5p found in beta cells and lowly expressed miRNAs such as miR-190b-5p found in sperm. Fifty-nine miRNAs were classified as "infrequent," 52 as "frequent" and 54 as "near ubiquitous." Fifty miRNAs were classified as "ubiquitous" including most of the well-known let-7 miRNAs and others such as miR-21-5p, miR-26a-5p and miR-30d-5p (**Supplementary Table S8**).

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We also evaluated miRNAs that demonstrated specificity among 7 cell classes (see methods), that are based on the similarities of the 196 cell types described above. (**Fig. 3A).** Plasma and platelets were grouped with immune cells into "blood." Many miRNAs are class-specific but can vary widely among specific cells of that class as observed in the epithelial class (**Fig. 3B**) and the blood class (**Fig. 3C**). For example, miR-122-5p is nearly exclusive to hepatocytes, while miR-205-5p is a more generic epithelial miRNA.

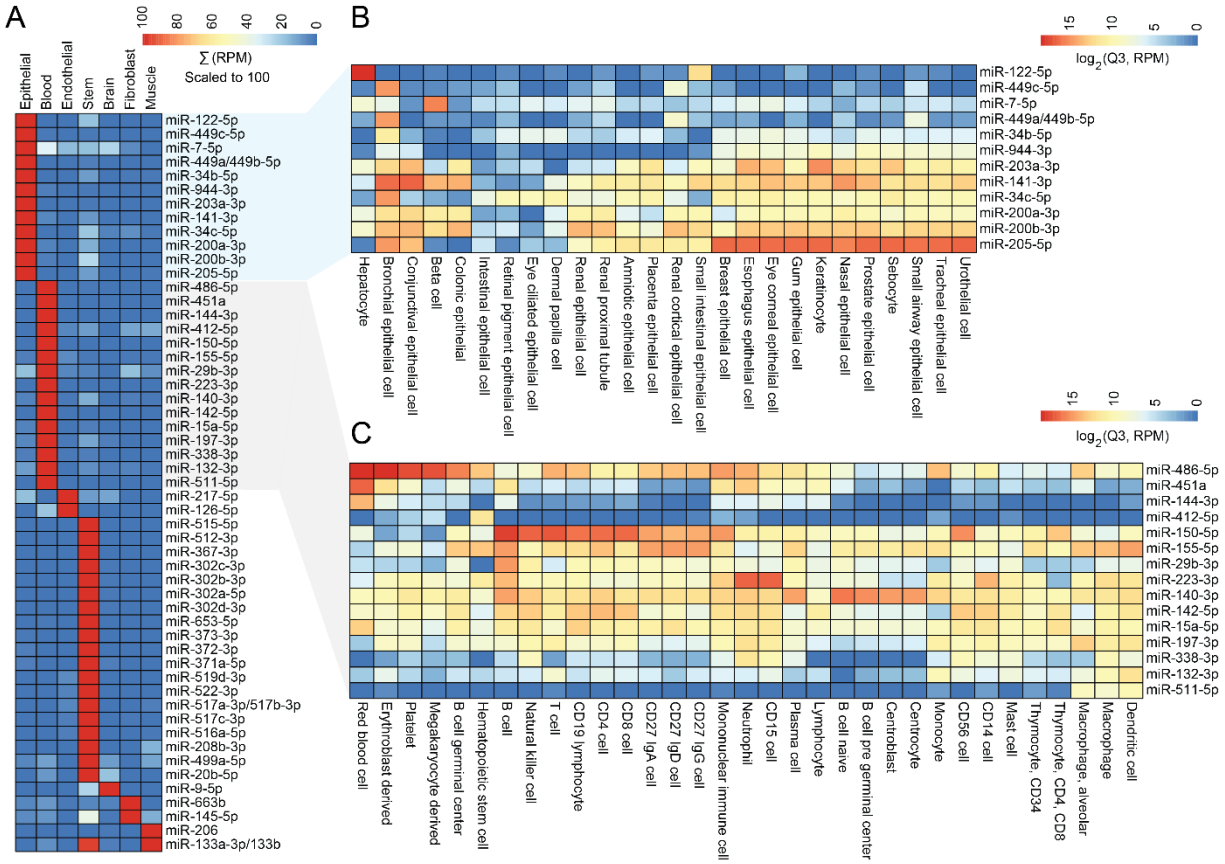

Figure 3. A. Heatmap showing enrichment of miRNA expression unique to cell class. All miRNA RPM values are summed and scaled to 100 across the 7 classes. **B.** Subset of epithelial cells, showing a wider range of expression among cells and **C.** miRNA expression enriched among a variety of blood cells, such as immune cells, RBCs, and platelets. For B. and C. individual miRNAs are log2 normalized and the 75th percentile (Q3) of the RPM value is shown.

Unique patterns of new cell types added to the cellular microRNAome

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In our current collection of 196 cell types, 30 were not part of previously published large cellular microRNAomes projects (McCall et al. [12] and de Rie et al. [11]). We identified several more specific patterns of enriched miRNAs expressed across these cells (**Fig. 4**). miR-184 is highly enriched in conjunctival epithelial cells (n=8), miR-199a-5p in fibroblast foreskin (n=14), and miR-4298 in hematopoietic stem cells (n= 6). The microRNAs miR-26b-5p, miR-29b-2-5p and miR-3150b-3p were enriched among CD27 cells. These miRNAs are based on the full miRBase collection.

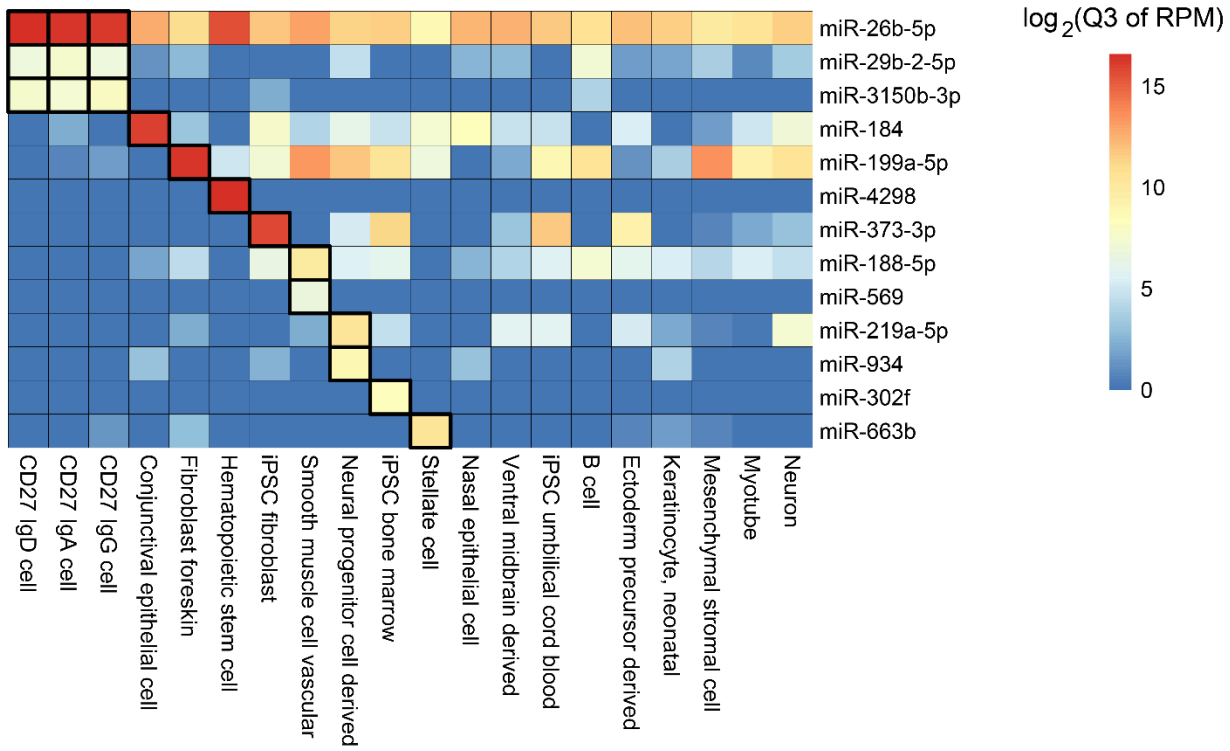

Figure 4. Heatmap showing miRNAs enriched for cells unique/new to this study, compared to prior cellular microRNAome studies.

Tissue microRNA expression is clarified by cellular expression patterns

Tissues are composed of numerous, diverse cell types. Thus, tissue miRNA expression estimates are a composite of the cell-specific expression patterns of the cell types they contain. To demonstrate this, we obtained miRNA expression estimates of 4 tissue samples (colon, liver, spleen, lymph node) for which the main cell types are present in our dataset. As seen in **Fig. 5**, the top 10 highest expressed

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

miRNAs in each sample, are both from specific cells and generally expressed across numerous cells. For

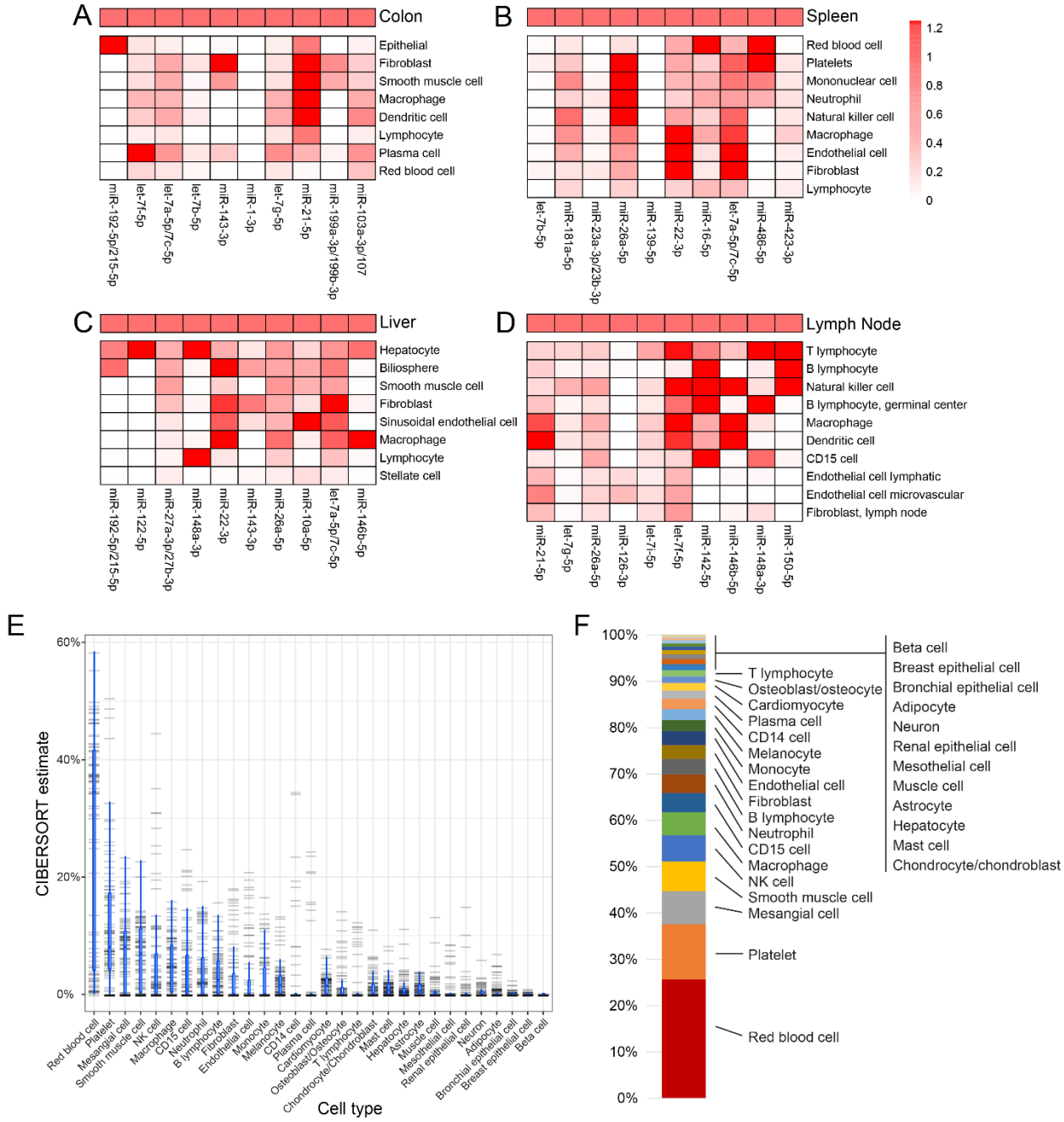

Figure 5. Contributions of individual cells toward tissue level miRNA expression Ratio of RPM expression between tissue miRNA and the individual cell miRNA RPM value for **A.** Colon, **B.** Spleen, **C.** Liver, and **D.** Lymph Node. Ratio between 0-1.25 (capped at 1.25 for illustration purposes). **E.** A boxplot of CIBERSORT estimates for each of 31 cell types with barcode strips overlaying each sample estimate for all 139 plasma samples. Cell types are in decreasing order of average composition estimate. **F.** Stacked box plot of the average composition of all plasma samples by the 31 cell types. Contribution ranges from <0.001% for beta cells to 27.4% for red blood cells.

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

example, in colon, miR-192-5p/215-5p is expressed exclusively in epithelial cells, while miR-103a-3p/107 is more ubiquitously expressed. Some tissue abundant miRNAs were not noted to be expressed in any of the common cell types including miR-1-3p in colon and miR-139-5p in spleen.

Plasma miRNAs are predominately derived from RBCs and platelets.

Blood plasma has been described as a full-body biopsy as the nucleic acid and protein material that it contains is derived from many cell types of the body. Based on the range of cell types in this microRNAome, we could evaluate the contributions of different cell types to plasma miRNA estimates. We deconvoluted 85 plasma samples from 30 representative cell types using CIBERSORT [17] and determined the major contributors to plasma miRNA are RBCs and platelets (38%) (**Fig. 5E and 5F**).

Accessing the human cellular microRNAome through R/Bioconductor and the UCSC Genome Browser.

To access the human cellular microRNAome, we have provided several useful tools. The first is an R/Bioconductor package "microRNAome" containing raw counts, RPM value and DESeq2 VST normalized values. The second is the "ABC of cellular microRNAome" barChart available under track hubs on the UCSC Genome Browser (https://genome.ucsc.edu/) (**Fig. 6**) [18]. This tool provides miRNA expression estimates for all 196 cell types (plus plasma and platelets) described in this project at a well-maintained website.

Discussion

This study represents the largest microRNAome of primary human cells that we are aware of to date. It consists of 2,077 samples from 2,406 runs representing 196 cell types, platelets and plasma, and covering 89.8% of miRNAs in the miRBase reference database. It generally replicated miRNA expression estimates from prior studies [11, 12, 19], while significantly adding to the number of samples upon which those estimates are based. Patterns of cell-enrichment and ubiquitousness are similar to those reported earlier with a few new associations based on new-to-this-study cell types. As most of these

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cells are grown in culture, this improved cellular microRNAome still suffers from ex vivo alterations in cell expression patterns. New methods to identify in vivo expression patterns may change absolute expression estimates substantially [20-22].

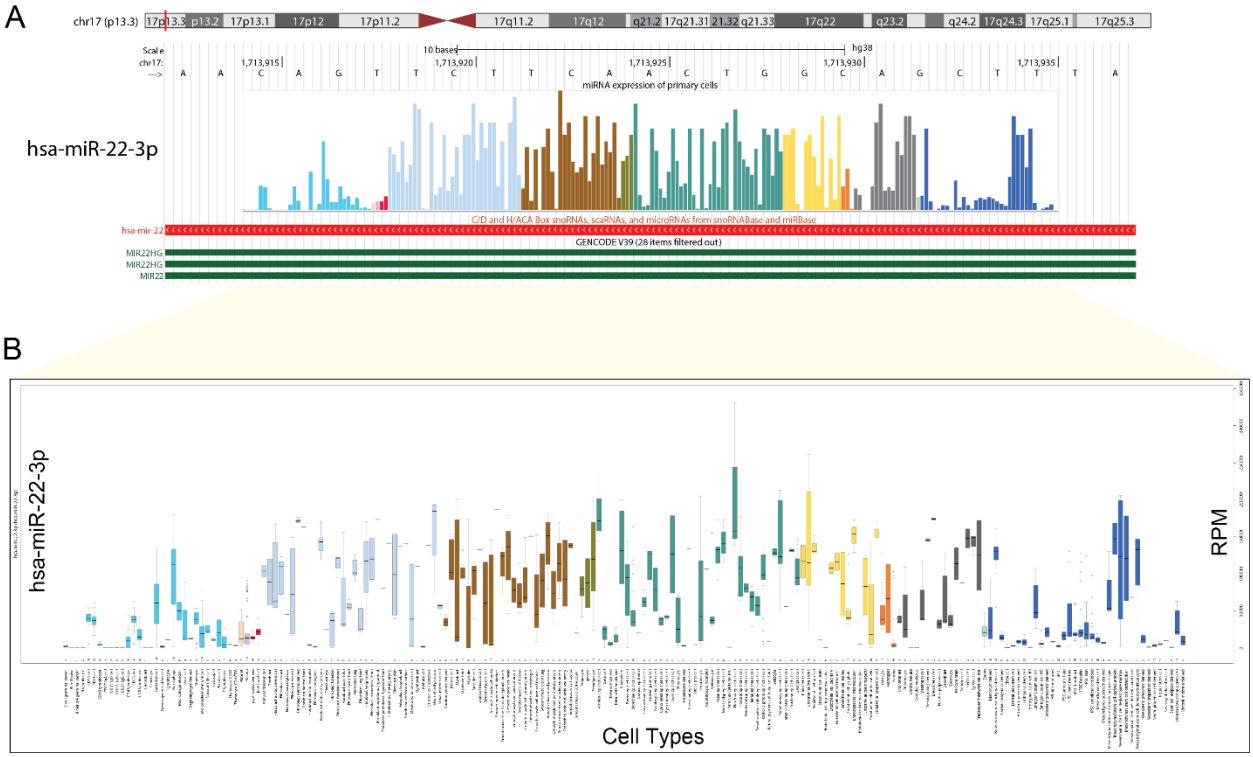

Figure 6. Screen capture of the UCSC Genome Browser ABC of Cellular microRNAome track hub. A. barchart of miR-22-3p expression. **B.** Box plot of miR-22-3p across 196 cell types, plasma and platelets.

Normalizing datasets from so many resources was a tremendous challenge. We chose to only include samples which had library preparations using Illumina TruSeq small RNA kits or which appeared similarly processed. We are aware of large expression estimate differences by library preparation methods due to ligation biases and other differences and felt excluding these other cases would improve batch correction [23, 24]. This limited the number of studies that appeared in the final analysis. Unlike our previous microRNAome effort, we initially included many non-control samples in this project, reasoning that some would have minor effects on miRNA expression, to increase the sample size. However, we removed those treated samples that did result in notable expression alterations. Ultimately, the DESeq2

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

VST normalization method proved to be a robust approach to normalize and stabilize the remaining samples, without adding a specific batch correction approach.

A significant loss to the number of samples and cells in our cellular microRNAome was the removal of 608 RNA Atlas runs due to their poor clustering relative to other cell types [14]. It was difficult to ascertain a pattern for why these cells were so different, but there were clear and consistent differences where some miRNAs were significantly higher or lower expressed in these 608 runs compared to matched runs from other studies. As the 32 hematologic cells from the RNA Atlas, clustered appropriately with other studies, we reasoned something related to the culturing method drove these changes, but what that was is unclear. We caution the use of this dataset relative to other microRNA resources [25]. Thus, our cellular microRNAome has several important biases. These relate to the library preparation method, inclusion of some treated cells, exclusion of most RNA Atlas samples, and cell culture passaging rather than in vivo isolation.

The selection of which miRNA library to align reads was a difficult decision. miRGeneDB has clearly established itself as the resource for bona fide miRNAs, while miRBase still includes scores of dubiously identified miRNAs [6, 16, 26-28]. The challenge is that our dataset appears as a UCSC Genome Browser Track Hub and this Genome Browser includes the full miRNA repertoire of miRBase. We chose to use the miRBase library to provide data for all “miRNAs” and demonstrate, unequivocally, how so many “miRNAs” are not expressed in many cell types. In fact, despite over 9 billion reads aligned, 280 “miRNAs” had no expression (**Supplementary Table S5**). This information, and the information on scores of other very lowly expressed miRNAs should be useful in the evaluation of miRNA reports which claim activity of miRNAs that are either not expressed or not expressed in the cell type of analysis [29, 30]. Another concern is that while not all reported miRNAs are bona fide, any short RNA could hypothetically represent a useful biomarker if expressed uniquely in a disease setting. Thus, even non-miRNAs could have value. Nonetheless, moving forward we strongly urge the use and reporting of miRNAs found in

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the MirGeneDB repository and only used miRNAs that were present in MirGeneDB for most of our deeper analyses.

We also chose to not search for novel miRNAs in these datasets. Too many "novel miRNAs" lack clear miRNAs features and have only served to further complicate the miRNA field [14, 31, 32]. We recently explored new chromosomal regions of the genome from the work of the T2T consortium and found no new novel miRNAs or paralagous miRNA loci [33]. With the exception of truly unique human cell types that have yet to be explored, we are confident that essentially all reasonably expressed bona fide human miRNAs have been identified.

The cell type specificity of any miRNA is dependent on the sample types to be compared. Thus, the comparison of our findings to other studies with a different collection of cell types, needs to be considered. With 196 cell types in this evaluation, we were reasonably confident we had good coverage of most human cell types. The majority (108) of evaluated miRNAs (323) were considered "cell specific/near specific," however, many of these were more lowly expressed (~100-500 RPM).

For many uses, having a cellular, rather than a tissue, microRNAome is preferred. As noted herein, a tissue signal is a composite of a number of different cell types, and it can easily, but wrongfully be assumed that one's miRNA of interest is present in a cell type of interest, if it's expression estimate is only obtained from tissues [29]. Conversely, some miRNAs are seen commonly in tissues that cannot be explained by cell data. For example miR-1, a known myomiR with skeletal and cardiac muscle expression, was noted in a colon sample here and was seen in a prostate sample previously [34, 35]. Skeletal and cardiac myocytes are not believed to be in these tissues and an absence of miR-1 in any reasonably expressed cell type from these tissues suggests an alternative cell-state or simple gap in our cellular coverage of tissues. In our tissue analysis, only a general idea of cellular contributions is conveyed as the exact composition of each tissue, with a deconvolution technique was not employed.

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Similar to a recent manuscript on cell free RNA in plasma [36], and consistent with other miRNA-based studies [37-39] we observed that most plasma miRNAs were RBC derived, followed by platelets, mesenchymal cells and immune cells. Of note, there was a very minor miRNA signal for brain-enriched miRNA, miR-9, and the contributions of neurons and astrocytes to the plasma miRNAs were estimated at 0.43% and 0.89% respectively. This calls into question plasma biomarker studies purported to show brain-specific changes in general miRNA expression estimates [40, 41].

In conclusion, we present the largest human cellular microRNAome project generated to date, which largely agrees with and expands upon prior knowledge of cell type miRNA expression patterns. It is easily accessible through the UCSC GenomeBrowser or through an R/Bioconductor experimental data package.

Methods

Sample Ascertainment:

Identification of suitable samples and their metadata were obtained from the NCBI Sequence Read Archive (SRA) with the query ((miRNA[All Fields] OR microRNA[All Fields] OR (small[All Fields] AND RNA[All Fields]) AND ("Homo sapiens"[Organism] OR ("Homo sapiens"[Organism] OR human[All Fields]))) AND "Homo sapiens"[Organism] AND (cluster_public[prop] AND "library layout single"[Properties] AND 1900[MDAT] : 2900[MDAT] NOT "strategy epigenomic"[Filter] NOT "strategy genome"[Filter] NOT "strategy exome"[Filter] AND "filetype fastq"[Properties])). This search was performed on January 22, 2021 and yielded 58,117 runs corresponding to 1,872 studies (**Fig. 1A**). Metadata from these samples was manually curated to obtain sample data exclusive to primary cell types. This curation excluded any sequencing runs that corresponded to paired-end sequencing, cancer cells, exosomes, and non-Illumina sequencing platforms. Further, category “Assay type” was filtered to only include “miRNA-Seq”, “ncRNA-Seq”, “RNA-Seq,” and the broad unknown category of “OTHER.” This

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

resulted in 6,054 runs that were downloaded using fasterq-dump/fastq-dump of the NCBI SRA Tookit (version 2.9.2) (https://github.com/ncbi/sra-tools) [42]. All runs were evaluated for adapter sequences and any samples with barcodes, unique molecular identifiers (UMIs), or adapter sequences on both ends were not processed (n = 1,870 runs were removed) due to the use of miREC in the processing [43]. Four tissue SRA runs, colon (SRR837824), spleen (SRR6853286), liver (SRR950887) and lymph noted (SRR14130226) were also obtained and processed.

Sample nomenclature:

The miRNA samples (n= 2,077) are derived from 175 different projects. We also included 329 duplicate runs of 173 samples, for a total of 2,406 runs processed. Due to the large number of uniquely named samples, cell types were clustered into batches for certain analyses. The classes for each cell type are: fibroblast, muscle, fat, epithelial, stem, endothelial, brain, immune, platelet, plasma, red blood cell (RBC), sperm and other (not easily classified cell types). Of note, plasma, the blood fluid, and platelets, megakaryocyte cell fragments, are not cells, but are listed as such for analyses, bringing the total "cells" to 198 in some analyses. Each project containing ≥2 samples was termed a batch (n=165). All singleton runs were collected into a single batch (batch 1). Groups (n=67) were defined as highly similar cell types (ex. all endothelial cells, regardless of tissue origin).

miRge3.0 Run Parameters:

The miRge3.0 pipeline was run in batches (an average of 11 samples, (-s <samples>)) on two computational clusters (BlueHive, University of Rochester and ARES, Johns Hopkins University) and locally on a PC (with 64-128Gb RAM and 12-40 CPUs) [13]. miRge3.0 default parameters were used along with parameters for miRNA error correction [43] and aligned to miRBase v22.1 [27, 28]. A typical run parameter is as follows:

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

miRge3.0 -s SRAS-file.fastq.gz -a <adapter_sequence> -gff -bam -trf -lib miRge3_Lib -on human -db miRBase -o OutputDir -mEC -ks 20 -ke 20

Dominant miRNA strand calculation:

The abundance of miRNA strands (5p/3p) was computed as described previously [12]. Briefly, the sum of RPM across all runs was generated and miRNAs with ≥ 1000 summed RPM were considered. The summed RPM ratio of 5p to 3p was calculated and any RPM value within 5% of a ratio of value of 1 was considered a co-dominant miRNA.

Multiple approaches to normalize for batch effect across datasets:

The raw read counts from all of the SRA runs were combined to form a single matrix with samples as columns and miRNAs as rows using the Pandas data frame in Python. Duplicate runs (technical replicates) were summed together for batch effect analysis. Four normalization methods and combinations of the methods were evaluated on this dataframe. These were variance stabilizing transformation (VST) in DESeq2 (v1.30.1) [44], Combat-Seq [45], RUVg and RUVr from RUVSeq package (v1.24.0) [46] or combinations of these approaches. The metadata information of all the samples was supplemented to these tools as matrix (CSV format) along with expression matrix (CSV format). All default parameters were used for each normalization method with the exception of the use of “group as design” in DESeq2; “batch and group” in CombatSeq; and “batch as design” in RUVr. The spike/control genes used in the RUVg method were "let-7a-5p/7c-5p", "let-7f-5p", "miR-103a-3p/107", "miR-125a-5p", "miR-181a-5p", "miR-186-5p", "miR-191-5p", "miR-22-3p", "miR-27a-3p/27b-3p", and "miR-30d-5p," based on ubiquitous expression pattern in SRA runs, described below. miRNAs which are also present in MirGeneDB [16, 26] and have an average RPM of ≥100 across all studies were used (n=670). The miRNA read counts were used for all normalization approaches and, to avoid errors pertaining to

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

divisible by zeros and/or infinity values, the value of one was summated across the matrix to replace zeros prior to applying normalization methods.

Solving ubiquitously expressed miRNAs for RUVg:

RUVg requires ubiquitous miRNAs from across the datasets to serve as spike-in controls. To identify these, we established an expression range using the Q1 and Q3 quartile values of let-7a-5p/7c-5p using the Excel function "QUARTILE.EXC". All miRNAs in the data matrix were queried and common miRNAs which could serve as RUVg control genes were found when true for this function: =IF(AND($QQ_1 \geq TQ_1$ - 1500, $QQ_3 \leq TQ_3$ + 1500),"T","F")

Where, TQ_1 = The threshold miRNA Q1 (lower quantile of let-7a), TQ_3 = The threshold miRNA Q3 (upper quantile of let-7a), and QQ_1 = Query miRNA Q1, QQ_3 = Query miRNA Q3. This resulted in the identification of 10 appropriate, ubiquitous control genes for RUVg.

Euclidean Distance Measurement:

To identify the best batch-correction optimization approach to our data, we investigated accurate cell type prediction based on different approaches. For this, four cell types (neuron, fibroblast, endothelial cell, lymphocyte) and plasma, containing 619 individual samples with a median of 5 samples per study (range 1-122), were used. A leave-one-study-out cross-validation strategy was used in which each study was used as the test set once with all other studies being used as a training set. In R, we generated cell type (and plasma) specific centroids by averaging gene counts over all training samples from a given cell type/plasma. The Euclidean distance was computed between each test sample and the cell type centroids, and we assigned each test sample to the cell type that minimized the Euclidean distance. Since there are 70 studies, this resulted in 70-fold cross-validation. Classification accuracy was assessed for datasets either using raw counts or after using ComBat-Seq, RUVSeq (RUVr and RUVg), DESeq2 VST and/or combinations of these approaches for corrections. As normalization occurred on all samples

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

prior to the leave one out approach, there was a common bias towards increasing homogeneity of samples in all of the ComBat-Seq, RUVr and RUVg approaches, likely inflating overall accuracy, but not affecting accuracy rank.

Uniform Manifold Approximation and Projection (UMAP) algorithm and outlier detection:

The Uniform Manifold Approximation and Projection (UMAP) algorithm was used for dimensionality reduction and plotting of the cell type clusters (v0.2.7.0) in R [47]. The UMAP clustering on the miRNA counts was performed to detect outliers for each class individually and included outclasses "fat" and "RBC" as controls for each cluster. The sum of miRNAs counts across all samples ≥ 5,000 were considered for UMAP clustering (n=1,111). Any samples that were outliers in the UMAP clusters for specific class (ex. epithelial) were individually evaluated for metadata-based or manuscript-based explanations of their unexpected differences to like samples. Some examples of elements that led to exclusion of a sample at this step were RNA source (nuclear only, exosome), protocols (drug stimulation, infectious agent use, siRNA use) and technical issues (low read depths, likely contamination due to isolation method). Such outlier samples were removed from the downstream analysis. R-based Plotly graphing library for ggplot, ggplotly (version 4.10.0) was used to create interactive HTML images of the UMAP clustering.

Determination of cell expression specificity of miRNAs:

The determination of cell expression specificity of miRNAs was performed for miRNAs that met the following conditions: present in the MirGeneDB database; leading strand (dominant); and reads per million (RPM) average value ≥ 100 for at least one of the 198 cell types. Expression patterns were classified into 5 groups. "Cell specific/near specific" indicated a miRNA in which expression was present in <5 dominant cells based on relative RPM peaks. "Infrequent" indicated a miRNA in which expression was present in ~5-10 dominant cells based on relative RPM peaks. "Frequent" indicated a miRNA

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

present in ~10-30 cell types based on dominant RPM peaks. "Near ubiquitous" was a miRNA with common expression in up to 178 cell types (<90%) at ≥100 RPM. "Ubiquitous" was a miRNA with common expression in >178 cell types (>90%) at ≥100 RPM with no dominant expression patterns. Not all miRNAs were easily placed in a category.

Analysis was performed at the cell type level (196) and at the cell class level for classes epithelial, endothelial, stem, brain, fibroblast and muscle as described above. Class "blood," used here, combines immune cells, red blood cells and platelets. A 75th percentile (Q3) of the RPM value was determined for individual miRNAs demonstrated to be cell class specific.

Cellular contributions to tissue miRNA expression:

Four representative tissues were obtained and processed through miRge3.0. The 10 most-highly expressed miRNAs were reported for each, as RPM. Expression levels of these 10 miRNAs were obtained from the 8-10 most common cell types of each tissue. For each miRNA, the tissue level RPM was divided by the average cell-type RPM level. Any miRNA expression level in a cell type greater than tissue was capped at a ratio of 1.25. A heat map of ratios (from 0-1.25) was generated for each tissue using Pheatmap in R.

CIBERSORT deconvolution of plasma miRNA expression:

A deconvolution of 85 plasma samples was performed from a reference dataset comprised of 30 cell types (1048 samples) using CIBERSORT [17]. The reference data was first batch corrected with the RUVSeq method [46]. The reference and mixture data were then normalized with the DESeq2 method [44], and the deconvolution was performed with CIBERSORT using q = 0.5 and a minimum of 50 and maximum of 200 signature genes per cell type. CIBERSORT was performed on each plasma sample individually and across a single averaged value of each miRNA for the 85 plasma cells.

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Generating bigBarChart for UCSC genome browser:

The RPM values of the miRNA expression across 196 primary cell types, platelets, and plasma were used to create a bigBarChart custom tracks for the UCSC genome browser [18]. A category file with two columns of named SRA runs and its corresponding cell type was created from the metadata. The genomic coordinates of miRNAs in the form of a BED file was obtained from miRBase (https://mirbase.org/). Two utility programs "expMatrixToBarchartBed" and "bedJoinTabOffset," obtained from the UCSC genome browser were used to transform the input expression matrix into a Browser Extensible Data (BED) bed6+5 file format (bed file). Another, utility "bedToBigBed" and chromosome sizes for Hg38 genome database "hg38.chrom.sizes" were downloaded from the UCSC genome browser. The "bedToBigBed" program was executed with default parameters except for parameter '-as=barChartBed.as' where definition of each field was slightly adjusted to represent miRNAs in the AutoSql format. The generated bigBed file along with all supporting information is provided in trackDb.txt and hub.txt files and are linked to UCSC genome browser via a GitHub repository [https://github.com/ mhalushka/miROme].

Data availability

Data from all 2,077 samples (2,406 runs) across 196 merged cell types, plasma and platelets are available through the track hubs feature at the UCSC genome browser (https://genome.ucsc.edu/). The track hub, "ABC of cellular microRNAome", allows one to query individual miRNAs. The expression patterns across different cell types can be visualized as a bar chart or a box plot. The raw counts, RPM values and DESeq2 VST normalized values are available for download as CSV files from the description page of the UCSC track hubs and Bioconductor repository

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(https://www.bioconductor.org/packages/devel/data/experiment/html/microRNAome.html). All custom scripts and code for this project are stored at GitHub: https://github.com/mhalushka/miROme.

Table 1: Table of primary cells included in the analysis, by general cell class. All cell classes, except sperm, had >1,000,000 average miRNA reads/sample. The average number of miRNAs detected across all 2,077 samples was 550.

Cell Class	Samples (Count)	Cell types (Count)	Total Input Reads (Average)	miRNA Reads (Average)	Unique miRBase miRNAs (Average)
Brain	77	3	11,836,198	5,011,006	748
Endothelial	147	14	10,160,565	4,446,952	588
Epithelial	216	36	8,730,464	5,036,673	524
Fat	19	3	2,114,216	1,201,275	485
Fibroblast	121	32	10,638,286	5,137,329	607
Immune	725	31	8,153,797	3,830,646	517
Muscle	124	24	9,043,430	4,228,642	579
Other	39	15	5,029,231	3,079,742	544
Plasma	85	1	5,469,997	1,535,698	295
Platelet	17	1	14,938,514	3,137,852	546
RBC	61	2	7,655,051	3,221,434	424
Sperm	89	1	3,421,680	166,735	401
Stem	357	35	9,429,860	4,690,630	714

Table 2: Comparison of cell type classification accuracy across normalization methods. Accuracy was defined as the number of predicted cell types that match the truth divided by the total number of predictions. DESeq2 VST had the highest accuracy. Only miRNAs present in MirGeneDB and ≥ 100 max(RPM) (n= 670) were included.

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Method	Accuracy	Neuron	Endothelial	Fibroblast	Immune	Plasma
Raw counts	56.5	47.6	38.1	43	91.6	30.6
Log_2(Raw counts)	81.1	100	90.5	64.4	76.3	85.9
DESeq2 VST	96.8	93.6	95.9	95.0	99.0	97.6
ComBat-Seq	79.3	100	89.1	62	73.4	85.9
ComBat-Seq + DESeq2 VST	95	95.2	93.9	90.9	98.5	94.1
RUVg	94	100	93.9	95.9	94.6	84.7
RUVg + DESeq2 VST	95.5	98.4	94.5	97.5	97.5	87.0
RUVr	86.3	92.0	90.5	96.7	79.8	75.3
RUVr + DESeq2 VST	84.5	87.3	88.4	96.7	78.3	72.9

Additional Files:

Supplementary Table S1: Cell type and cell class for all 2,077 samples and 2,406 runs used in this study.

Supplementary Table S2: Metadata for each sample from Sequence Read Archive (NCBI) and miRge3.0 run summary information for all 2,406 runs.

Supplementary Table S3: miRNA raw counts across all 2,406 runs.

Supplementary Table S4: miRNA RPM values across all 2,406 runs.

Supplementary Table S5: The unmapped miRNAs from miRBase

Supplementary Table S6: The Euclidean distance-based accuracy of cell type clustering among 5 classes of cells.

Supplementary Table S7: The DESeq2 VST normalized values of miRNA expression across all samples and used for further analysis in this project. All miRNAs are also present in MirGeneDB.

Supplementary Table S8: The cell type specificity of 323 miRNAs.

Supplementary Figure S1: UMAP clustering of cell class 'Brain' (n=77) corresponding to three distinct cell types and two outgroups fat (n=19) corresponding to three distinct cell types and Red blood cells (n=37).

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Figure S2: UMAP clustering of cell class 'Endothelial' (n=147) corresponding to 14 cell types and two outgroups fat (n=19) corresponding to three distinct cell types and red blood cells (n=37).

Supplementary Figure S3: UMAP clustering of cell class 'Epithelial' (n=216) corresponding to 36 cell types and two outgroups fat (n=19) corresponding to three distinct cell types and red blood cells (n=37).

Supplementary Figure S4: UMAP clustering of cell class 'Fibroblasts' (n=121) corresponding to 32 cell types and two outgroups fat (n=19) corresponding to three distinct cell types and red blood cells (n=37).

Supplementary Figure S5: UMAP clustering of cell class 'Immune' (n=725) corresponding to 31 cell types and two outgroups fat (n=19) corresponding to three distinct cell types and red blood cells (n=37).

Supplementary Figure S6: UMAP clustering of cell class 'Muscle' (n=124) corresponding to 24 cell types and two outgroups fat (n=19) corresponding to three distinct cell types and red blood cells (n=37).

Supplementary Figure S7: UMAP clustering of cell class 'Other' (n=39) corresponding to 15 cell types and two outgroups fat (n=19) corresponding to three distinct cell types and red blood cells (n=37).

Supplementary Figure S8: UMAP clustering of cell class 'Plasma' (n=85) and two outgroups fat (n=19) corresponding to three distinct cell types and red blood cells (n=37).

Supplementary Figure S9: UMAP clustering of cell class 'Platelet' (n=17) and two outgroups fat (n=19) corresponding to three distinct cell types and red blood cells (n=37).

Supplementary Figure S10: UMAP clustering of cell class 'Sperm' (n=89) and two outgroups fat (n=19) corresponding to three distinct cell types and red blood cells (n=37).

Supplementary Figure S11: UMAP clustering of cell class 'Stem' (n=357) corresponding to 35 cell types and two outgroups fat (n=19) corresponding to three distinct cell types and red blood cells (n=37).

Abbreviations

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

NCBI: National Center for Biotechnology Information; SRA: Sequence Read Archive; RNA: Ribonucleic Acid; UCSC: University of California, Santa Cruz; UMAP: Uniform Manifold Approximation and Projection; HTML: Hypertext Markup Language; RBCs: Red Blood Cells; RPM: Reads Per Million; VST: variance stabilizing transformation; RUVr: Remove Unwanted Variation Using Residuals; RUVg: Remove Unwanted Variation Using Control Genes; MDAT: Modification date; Prop: Properties; CSV: Comma-separated Values; Gg: Grammar of graphics (ggplot); Q3: 75th percentile; BED: Browser Extensible Data

Competing Interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Heart, Lung, and Blood Institute, R01HL137811, M Halushka; National Institute of General Medical Sciences, R01GM130564, M Halushka; National Institute of General Medical Sciences, R01GM139928, M N McCall; and the University of Rochester, CTSA, UL1TR002001, M N McCall.

Author Contributions

Design of Study: M.K.H., A.H.P.;

Funding Acquisition: M.K.H.; M.N.M.;

Data Analysis: A.H.P., A.B., Z.P.B., M.N.M.;

Draft Preparation: A.H.P., M.K.H.;

Review and editing: A.B., Z.P.B.

References:

1. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004;116 2:281-97.

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2. Mendell JT and Olson EN. MicroRNAs in stress signaling and human disease. Cell. 2012;148 6:1172-87. doi:10.1016/j.cell.2012.02.005.
3. Mitchell PS, Parkin RK, Kroh EM, Fritz BR, Wyman SK, Pogosova-Agadjanyan EL, et al. Circulating microRNAs as stable blood-based markers for cancer detection. Proc Natl Acad Sci U S A. 2008;105 30:10513-8. doi:10.1073/pnas.0804549105.
4. Halushka PV, Goodwin AJ and Halushka MK. Opportunities for microRNAs in the Crowded Field of Cardiovascular Biomarkers. Annual review of pathology. 2019;14:211-38. doi:10.1146/annurev-pathmechdis-012418-012827.
5. Pinzon N, Li B, Martinez L, Sergeeva A, Presumey J, Apparailly F, et al. microRNA target prediction programs predict many false positives. Genome Res. 2017;27 2:234-45. doi:10.1101/gr.205146.116.
6. Fromm B, Keller A, Yang X, Friedlander MR, Peterson KJ and Griffiths-Jones S. Quo vadis microRNAs? Trends in genetics : TIG. 2020;36 7:461-3. doi:10.1016/j.tig.2020.03.007.
7. Witwer KW and Halushka MK. Toward the promise of microRNAs - Enhancing reproducibility and rigor in microRNA research. RNA biology. 2016;13 11:1103-16. doi:10.1080/15476286.2016.1236172.
8. Lagos-Quintana M, Rauhut R, Yalcin A, Meyer J, Lendeckel W and Tuschl T. Identification of tissue-specific microRNAs from mouse. Current biology : CB. 2002;12 9:735-9.
9. Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, et al. A mammalian microRNA expression atlas based on small RNA library sequencing. Cell. 2007;129 7:1401-14. doi:10.1016/j.cell.2007.04.040.
10. Juzenas S, Venkatesh G, Hubenthal M, Hoeppner MP, Du ZG, Paulsen M, et al. A comprehensive, cell specific microRNA catalogue of human peripheral blood. Nucleic Acids Res. 2017;45 16:9290-301. doi:10.1093/nar/gkx706.
11. de Rie D, Abugessaisa I, Alam T, Arner E, Arner P, Ashoor H, et al. An integrated expression atlas of miRNAs and their promoters in human and mouse. Nature biotechnology. 2017;35 9:872-8. doi:10.1038/nbt.3947.
12. McCall MN, Kim MS, Adil M, Patil AH, Lu Y, Mitchell CJ, et al. Toward the human cellular microRNAome. Genome Res. 2017;27 10:1769-81. doi:10.1101/gr.222067.117.
13. Patil AH and Halushka MK. miRge3.0: a comprehensive microRNA and tRF sequencing analysis pipeline. NAR Genom Bioinform. 2021;3 3:lqab068. doi:10.1093/nargab/lqab068.
14. Lorenzi L, Chiu HS, Avila Cobos F, Gross S, Volders PJ, Cannoodt R, et al. The RNA Atlas expands the catalog of human non-coding RNAs. Nature biotechnology. 2021;39 11:1453-65. doi:10.1038/s41587-021-00936-1.
15. Kozomara A and Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. 2011;39 Database issue:D152-7. doi:10.1093/nar/gkq1027.
16. Fromm B, Hoye E, Domanska D, Zhong X, Aparicio-Puerta E, Ovchinnikov V, et al. MirGeneDB 2.1: toward a complete sampling of all major animal phyla. Nucleic Acids Res. 2021; doi:10.1093/nar/gkab1101.
17. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nature methods. 2015;12 5:453-7. doi:10.1038/nmeth.3337.
18. Navarro Gonzalez J, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney BJ, et al. The UCSC Genome Browser database: 2021 update. Nucleic Acids Res. 2021;49 D1:D1046-D57. doi:10.1093/nar/gkaa1070.
19. Halushka MK, Fromm B, Peterson KJ and McCall MN. Big Strides in Cellular MicroRNA Expression. Trends in genetics : TIG. 2018;34 3:165-7. doi:10.1016/j.tig.2017.12.015.

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

20. Rosenberg AZ, Wright C, Fox-Talbot K, Rajpurohit A, Williams C, Porter C, et al. xMD-miRNA-seq to generate near in vivo miRNA expression estimates in colon epithelial cells. Scientific reports. 2018;8 1:9783. doi:10.1038/s41598-018-28198-z.
21. Kuosmanen SM, Kansanen E, Sihvola V and Levonen AL. MicroRNA Profiling Reveals Distinct Profiles for Tissue-Derived and Cultured Endothelial Cells. Scientific reports. 2017;7 1:10943. doi:10.1038/s41598-017-11487-4.
22. Isakova A, Neff N and Quake SR. Single-cell quantification of a broad RNA spectrum reveals unique noncoding patterns associated with cell types and states. Proc Natl Acad Sci U S A. 2021;118 51 doi:10.1073/pnas.2113568118.
23. Wright C, Rajpurohit A, Burke EE, Williams C, Collado-Torres L, Kimos M, et al. Comprehensive assessment of multiple biases in small RNA sequencing reveals significant differences in the performance of widely used methods. BMC Genomics. 2019;20 1:513. doi:10.1186/s12864-019-5870-3.
24. Giraldez MD, Spengler RM, Etheridge A, Godoy PM, Barczak AJ, Srinivasan S, et al. Comprehensive multi-center assessment of small RNA-seq methods for quantitative miRNA profiling. Nature biotechnology. 2018;36 8:746-57. doi:10.1038/nbt.4183.
25. Fromm B, Zhong X, Tarbier M, Friedlander MR and Hackenberg M. The limits of human microRNA annotation have been met. RNA. 2022; doi:10.1261/rna.079098.122.
26. Fromm B, Domanska D, Hoye E, Ovchinnikov V, Kang W, Aparicio-Puerta E, et al. MirGeneDB 2.0: the metazoan microRNA complement. Nucleic Acids Res. 2020;48 D1:D132-D41. doi:10.1093/nar/gkz885.
27. Griffiths-Jones S. miRBase: the microRNA sequence database. Methods Mol Biol. 2006;342:129-38. doi:10.1385/1-59745-123-1:129.
28. Kozomara A, Birgaoanu M and Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res. 2019;47 D1:D155-D62. doi:10.1093/nar/gky1141.
29. Kent OA, McCall MN, Cornish TC and Halushka MK. Lessons from miR-143/145: the importance of cell-type localization of miRNAs. Nucleic Acids Res. 2014;42 12:7528-38. doi:10.1093/nar/gku461.
30. Halushka MK. MicroRNA-144 is unlikely to play a role in bronchiolitis obliterans syndrome. The Journal of heart and lung transplantation : the official publication of the International Society for Heart Transplantation. 2016;35 4:543. doi:10.1016/j.healun.2016.01.008.
31. Londin E, Loher P, Telonis AG, Quann K, Clark P, Jing Y, et al. Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs. Proc Natl Acad Sci U S A. 2015;112 10:E1106-15. doi:10.1073/pnas.1420955112.
32. Blanco-Dominguez R, Sanchez-Diaz R, de la Fuente H, Jimenez-Borreguero LJ, Matesanz-Marin A, Relano M, et al. A Novel Circulating MicroRNA for the Detection of Acute Myocarditis. The New England journal of medicine. 2021;384 21:2014-27. doi:10.1056/NEJMoa2003608.
33. Patil AH, Halushka MK and Fromm B. No evidence of paralogous loci or new bona fide microRNAs in telomere to telomere (T2T) genomic data. bioRxiv. 2021;https://doi.org/10.1101/2021.12.09.471935
34. Kumar B, Rosenberg AZ, Choi SM, Fox-Talbot K, De Marzo AM, Nonn L, et al. Cell-type specific expression of oncogenic and tumor suppressive microRNAs in the human prostate and prostate cancer. Scientific reports. 2018;8 1:7189. doi:10.1038/s41598-018-25320-z.
35. Sweetman D, Goljanek K, Rathjen T, Oustanina S, Braun T, Dalmay T, et al. Specific requirements of MRFs for the expression of muscle specific microRNAs, miR-1, miR-206 and miR-133. Developmental biology. 2008;321 2:491-9. doi:10.1016/j.ydbio.2008.06.019.
36. Vorperian SK, Moufarrej MN, Tabula Sapiens C and Quake SR. Cell types of origin of the cell-free transcriptome. Nature biotechnology. 2022; doi:10.1038/s41587-021-01188-9.

bioRxiv preprint doi: https://doi.org/10.1101/2022.05.16.492160; this version posted May 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

37. Kirschner MB, Kao SC, Edelman JJ, Armstrong NJ, Vallely MP, van Zandwijk N, et al. Haemolysis during sample preparation alters microRNA content of plasma. PLoS One. 2011;6 9:e24145. doi:10.1371/journal.pone.0024145.
38. Cheng HH, Yi HS, Kim Y, Kroh EM, Chien JW, Eaton KD, et al. Plasma processing conditions substantially influence circulating microRNA biomarker levels. PLoS One. 2013;8 6:e64795. doi:10.1371/journal.pone.0064795.
39. Pritchard CC, Kroh E, Wood B, Arroyo JD, Dougherty KJ, Miyaji MM, et al. Blood cell origin of circulating microRNAs: a cautionary note for cancer biomarker studies. Cancer Prev Res (Phila). 2012;5 3:492-7. doi:10.1158/1940-6207.CAPR-11-0370.
40. Coolen M, Katz S and Bally-Cuif L. miR-9: a versatile regulator of neurogenesis. Front Cell Neurosci. 2013;7:220. doi:10.3389/fncel.2013.00220.
41. Sheinerman KS, Tsivinsky VG, Crawford F, Mullan MJ, Abdullah L and Umansky SR. Plasma microRNA biomarkers for detection of mild cognitive impairment. Aging (Albany NY). 2012;4 9:590-605. doi:10.18632/aging.100486.
42. Kodama Y, Shumway M, Leinonen R and International Nucleotide Sequence Database C. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res. 2012;40 Database issue:D54-6. doi:10.1093/nar/gkr854.
43. Zhang X, Ping P, Hutvagner G, Blumenstein M and Li J. Aberration-corrected ultrafine analysis of miRNA reads at single-base resolution: a k-mer lattice approach. Nucleic Acids Res. 2021;49 18:e106. doi:10.1093/nar/gkab610.
44. Love MI, Huber W and Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15 12:550. doi:10.1186/s13059-014-0550-8.
45. Zhang Y, Parmigiani G and Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genom Bioinform. 2020;2 3:lqaa078. doi:10.1093/nargab/lqaa078.
46. Risso D, Ngai J, Speed TP and Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nature biotechnology. 2014;32 9:896-902. doi:10.1038/nbt.2931.
47. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nature biotechnology. 2018; doi:10.1038/nbt.4314.