Supplemental Table of Contents

Figures

Tables

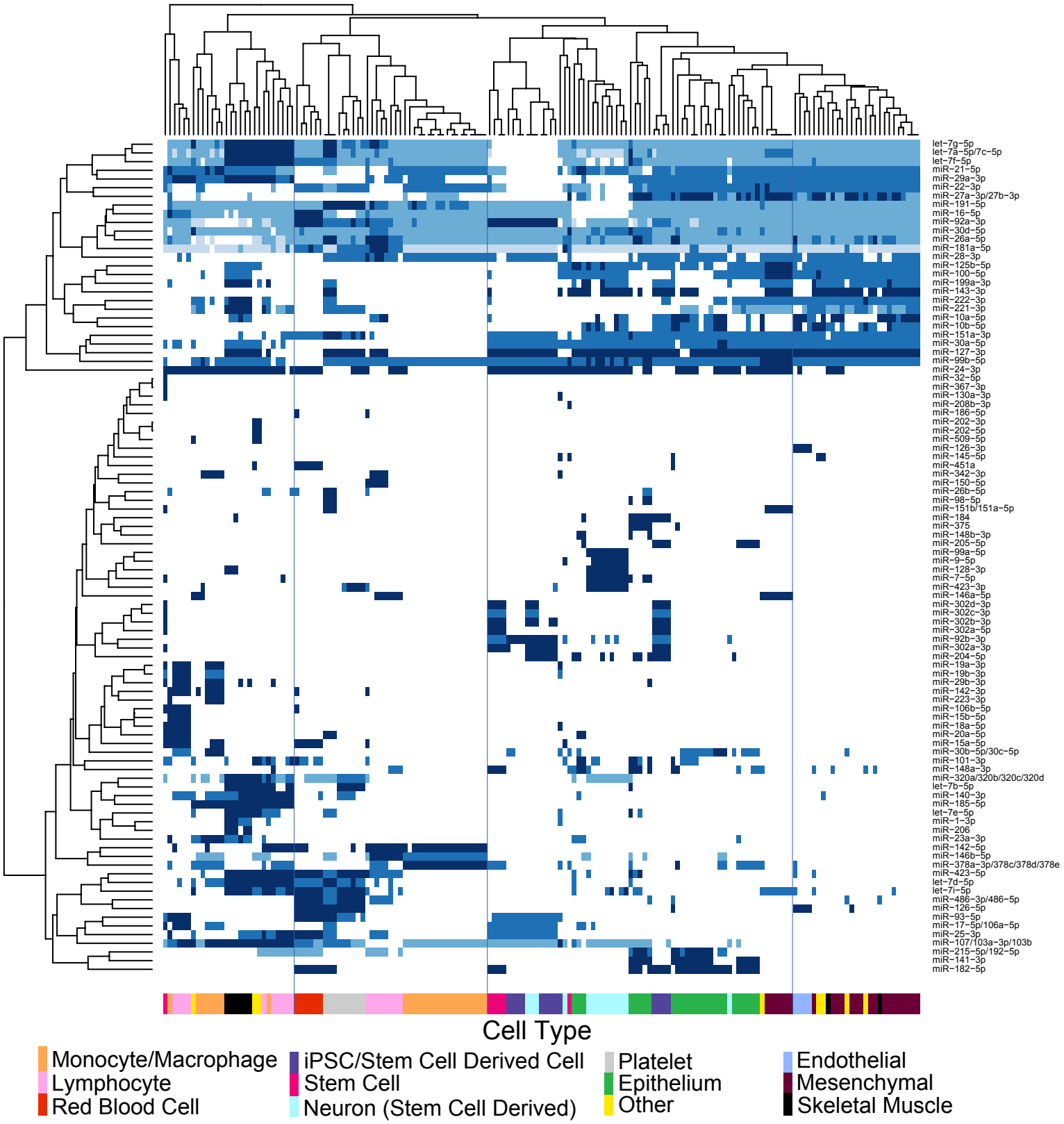
11. Supplemental Table S11 - Primary cell dominant isomiR for each microRNA
12. Supplemental Table S12 – IsomiRs with altered 5p starting positions
13. Supplemental Table S13 - Cancer cell dominant isomiR for each microRNA
14. Supplemental Table S14 - Lonza Cell lines
15. Supplemental Table S15 - Ago2 Clip RNAseq samples used to detect novel microRNAs
16. Supplemental Table S16 - PCR primers and conditions for novel microRNA detection

Scripts

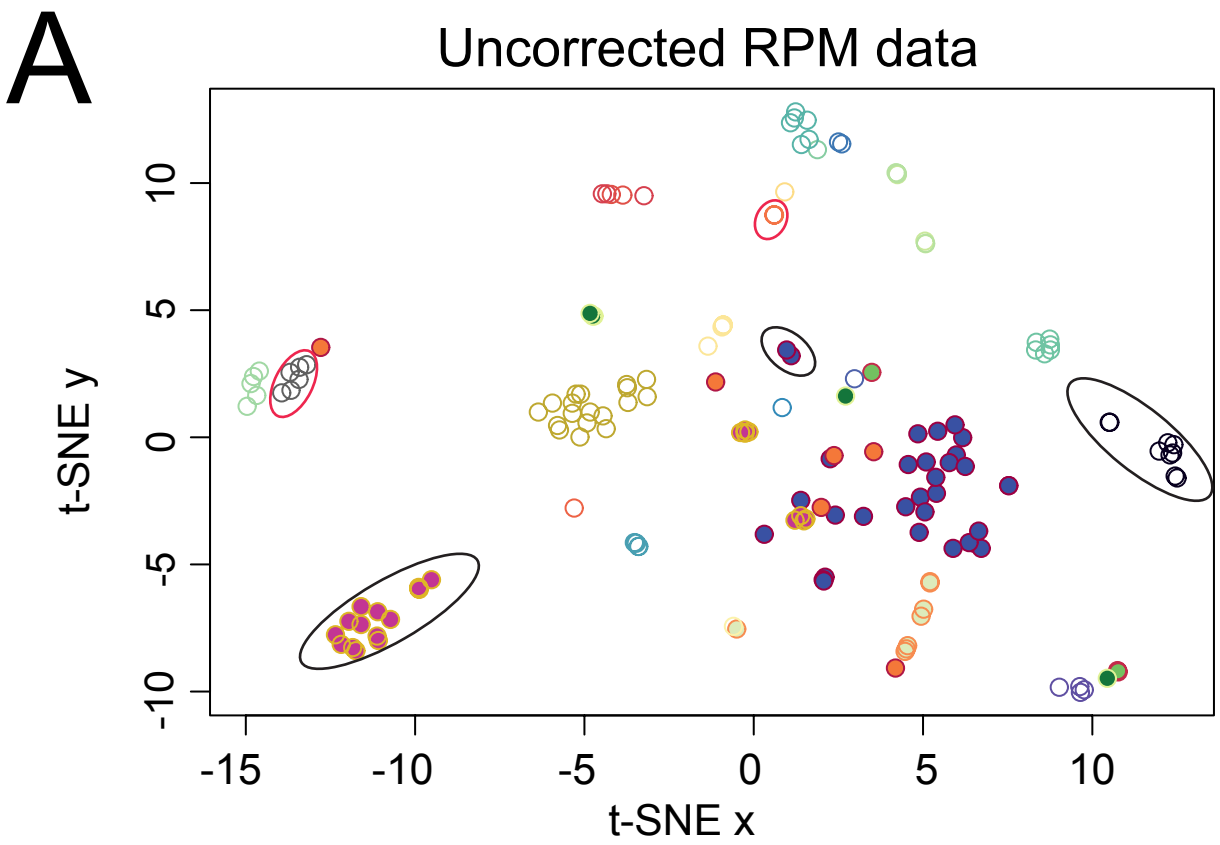
1. Supplemental Scripts (.zip)
    a. isomir_analysis_v12.pl
    b. Phylop_calculate_JHU_conservation.py
    c. Phylop_calculate_TJH_miRBase_conservation.py
    d. Phylop_preprocess_wigfix_file.py
    e. Unmapped_miRNA2fasta.py

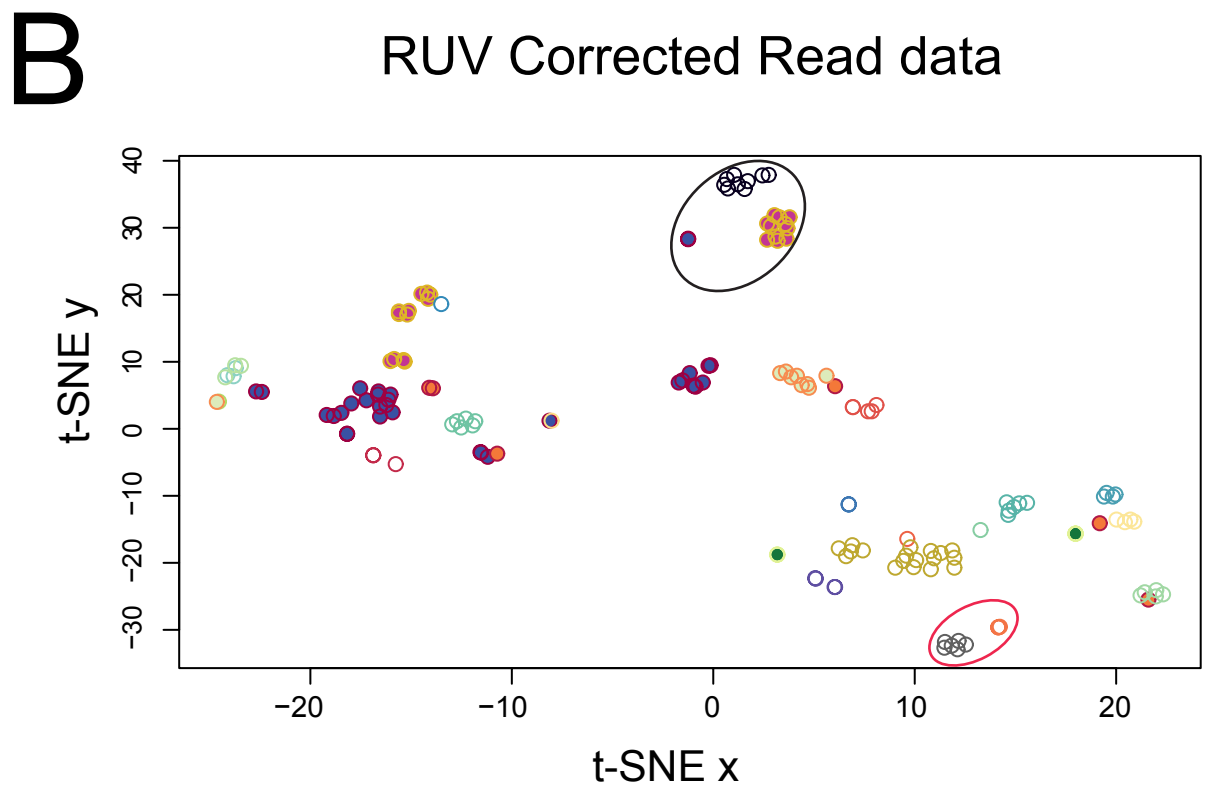
Packages

1. Supplemental Bioconductor Package (.tar.gz)

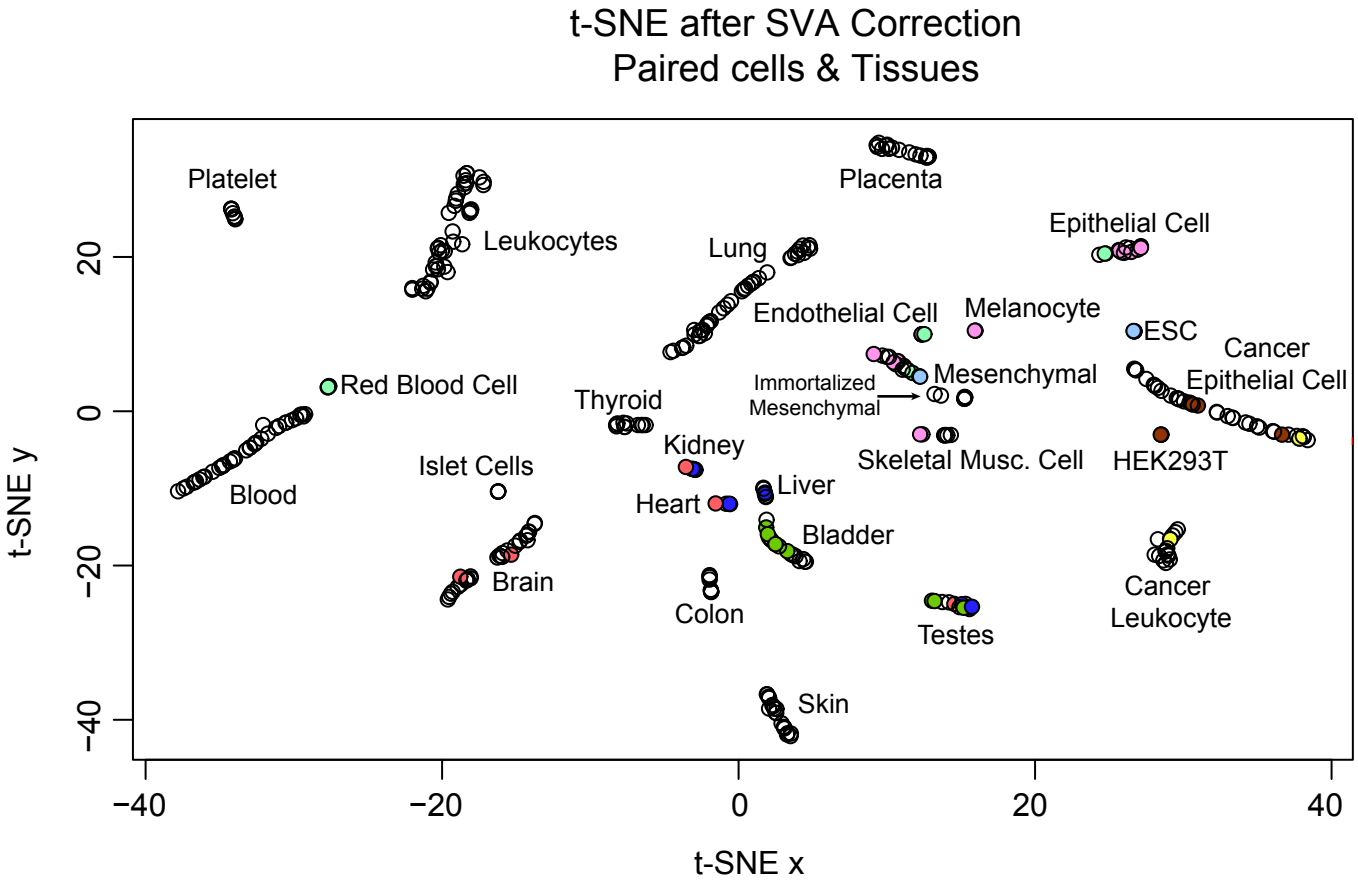


Supplemental Figure S1: DEXUS analysis of primary cells. Hierarchical clustering of cell-types based on discrete levels of microRNA expression identifies generalized cell-type specific patterns of microRNA expression that are inherently robust to batch effects. A five component negative binomial mixture model was fit to estimate discrete levels of microRNA expression across cell-types. We selected the 96 microRNAs that were deemed multimodal (INI score > 0.1) and highly expressed in at least one mixture component (mean greater than 50,000). The estimated mixture distributions were used to assign up to five levels of expression to each microRNA across 161 samples. All mixture components with a mean less than 2,500 were considered unexpressed. The "other" category includes melanocyte, corona radiata, cumulus oophorous, osteoblast, mesangial, astrocyte and chondrocyte cells.

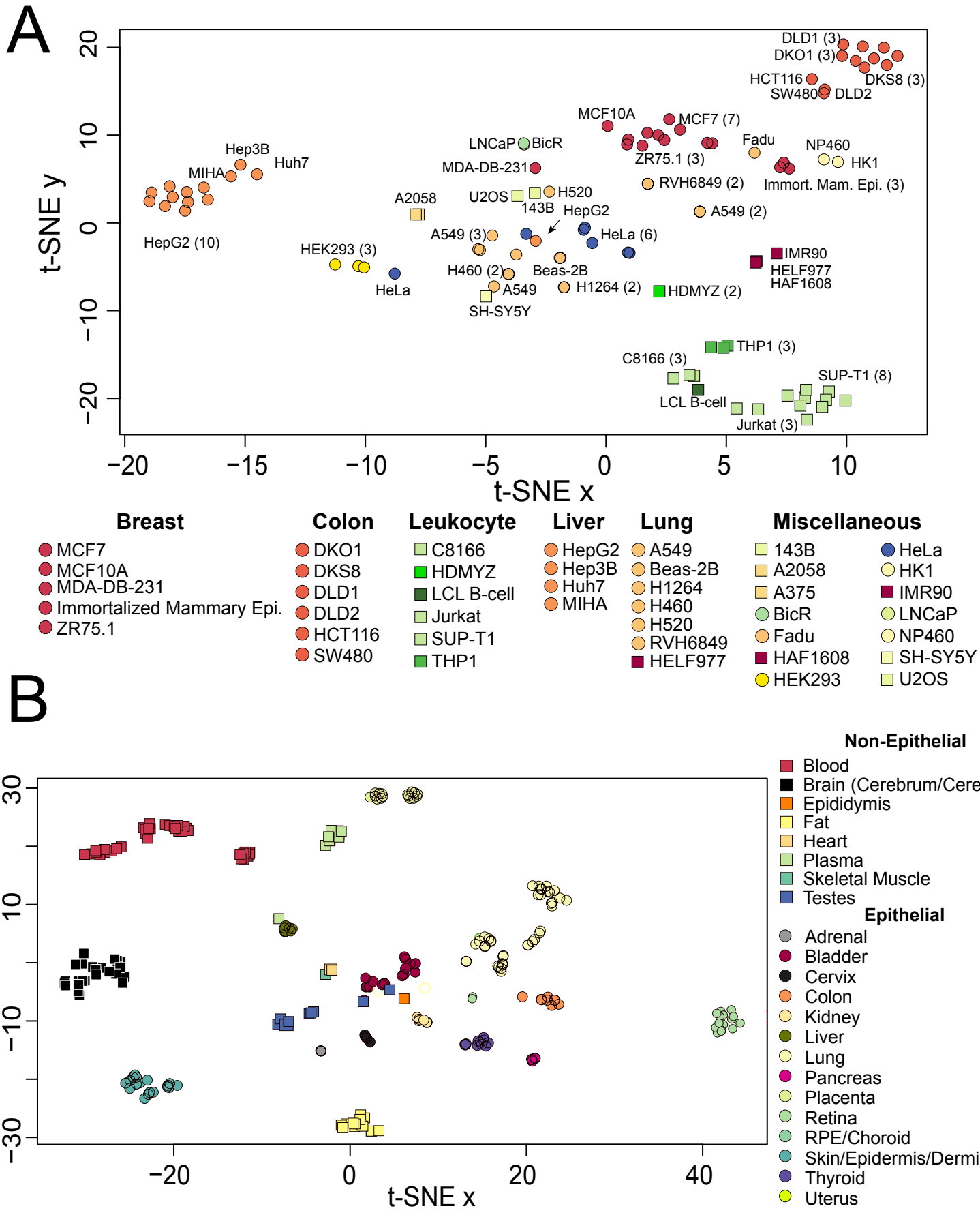




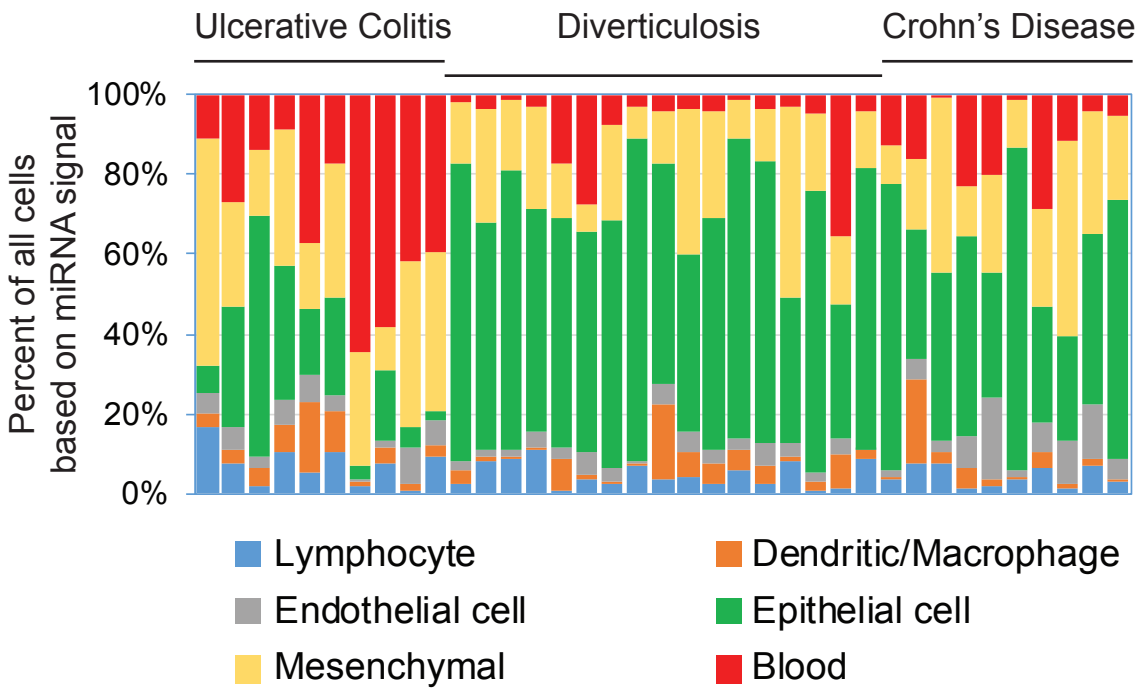
Supplemental Figure S2: Adjusted and unadjusted t-Stochastic Neighber Embedding (t-SNE) on primary cells. (A) Unadjusted RPM values were used to identify clustering patterns. Each circle is a sample and the circle coloring is based on 27 technical batches. Closed colored circles are samples that had more than one cell type present in the technial batch. The closed orange circle batch contained a smooth muscle cell, lymphocyte, red blood cell, endothelial cell, epithelial cell and fibroblast. The closed blue/red circle batch contained epithelial cells, mesenchymal cells, endothelial cells, skeletal muscle cells and others. The red ovals are two separate platelet samples and the black ovals are a batch of neural stem cells and a batch of H9 ESC differentiated brain samples.Without adjustment, technical batches tend to cluster by themselves unless multiple cell types are present. (B) After adjustment by RUV on read data, the sample clustering is improved. Here the platelet samples and neural samples (red and black ovals) cluster together. A further description of the various cell types present is seen in Figure 2b.



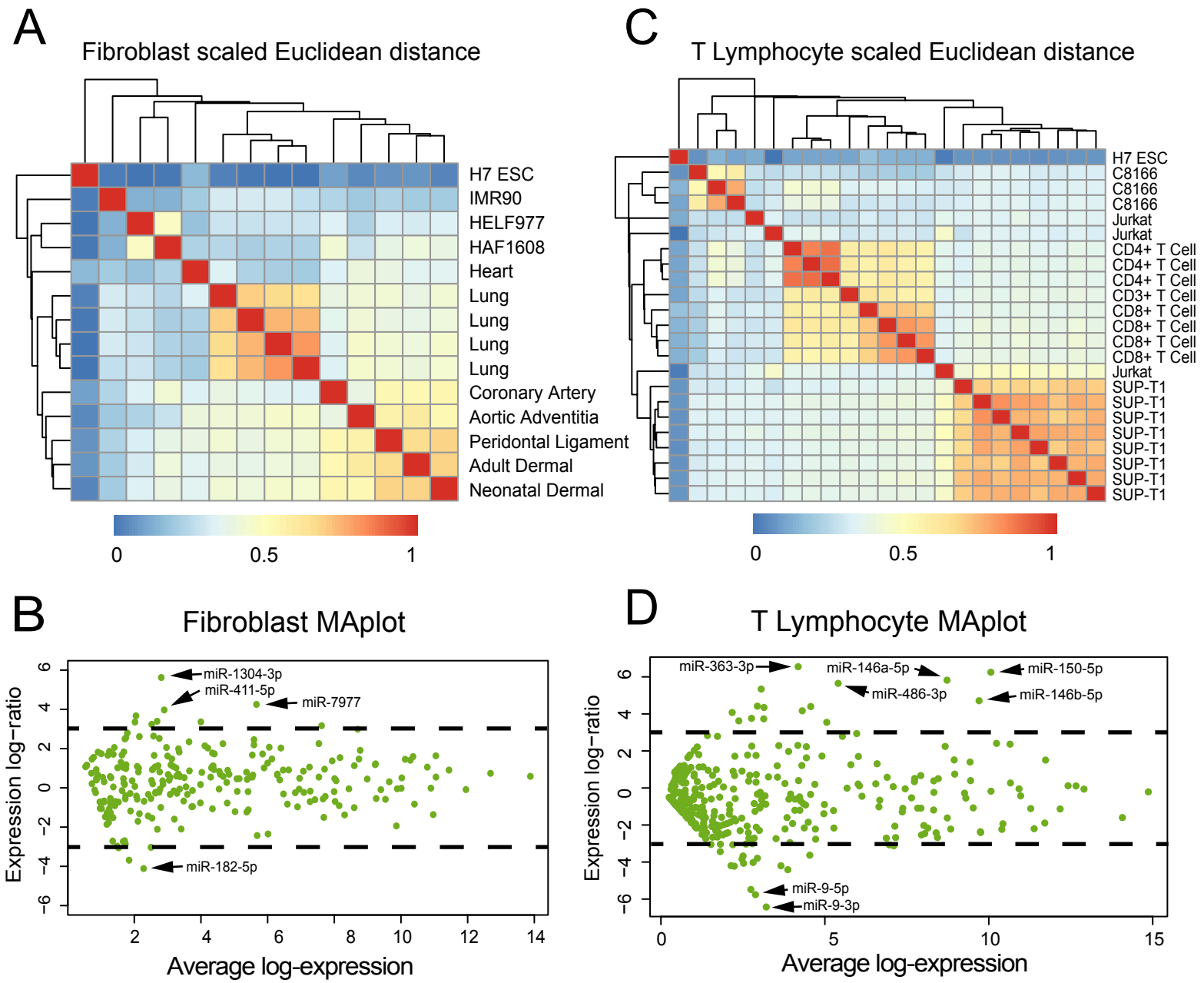
Supplemental Figure S3: Biological clustering of samples using t-Stochastic Neighbor Embedding (t-SNE). 387 samples were identified across primary cells, cancer/immortalized cells and tissues that had RNA-seq data from more than one experiment. Surrogate variable analysis (SVA) was performed using 26 loosely representative "biologic clusters" of tissues and cell types as the primary variable. SVA corrected for 32 surrogate variables. The t-SNE plot strongly clusters these biologic groupings. Of note, some similar cell/tissue types such as blood and red blood cells cluster nearby, as do mesenchymal and immortalized mesenchymal cells. The colored circles represent samples that were from the same study, but were found in more than one cell type or tissue cluster (ex. dark blue samples were from SRA bioproject PRJEB2604). Eight separate bioprojects are represented by colored circles.



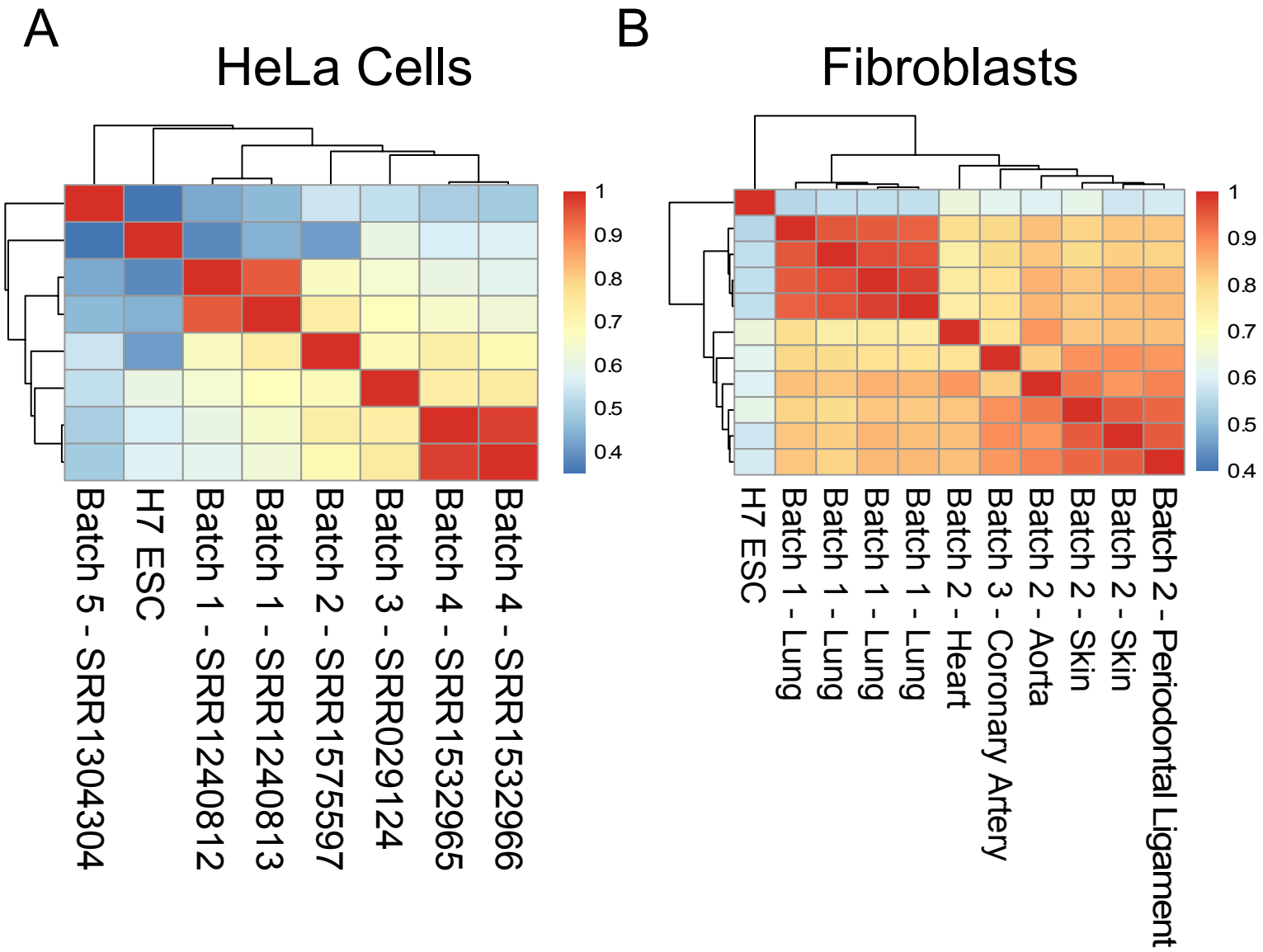
Supplemental Figure S4: T-Distributed Stochastic Neighbor Embedding (t-SNE) image of cancer/immortalized cells and tissues. (A) One hundred samples are evaluated. There is strong clustering by general cell types (leukocyte, liver, breast, colon and lung adenocarcinoma). Round symbols represent epithelial cells. Occasional outlier cells (HeLa, HepG2) may represent batch effects, improperly identified cell cultures or differences in stocks of these cells due to different passages. (B) Two hundred and sixty-nine samples from 26 tissues show variable patterns of clustering. Homogeneous and compositionally different tissues (blood/plasma/brain) cluster away from the epithelial organs which are generally found in the middle of the figure. Round symbols represent epithelial organs.



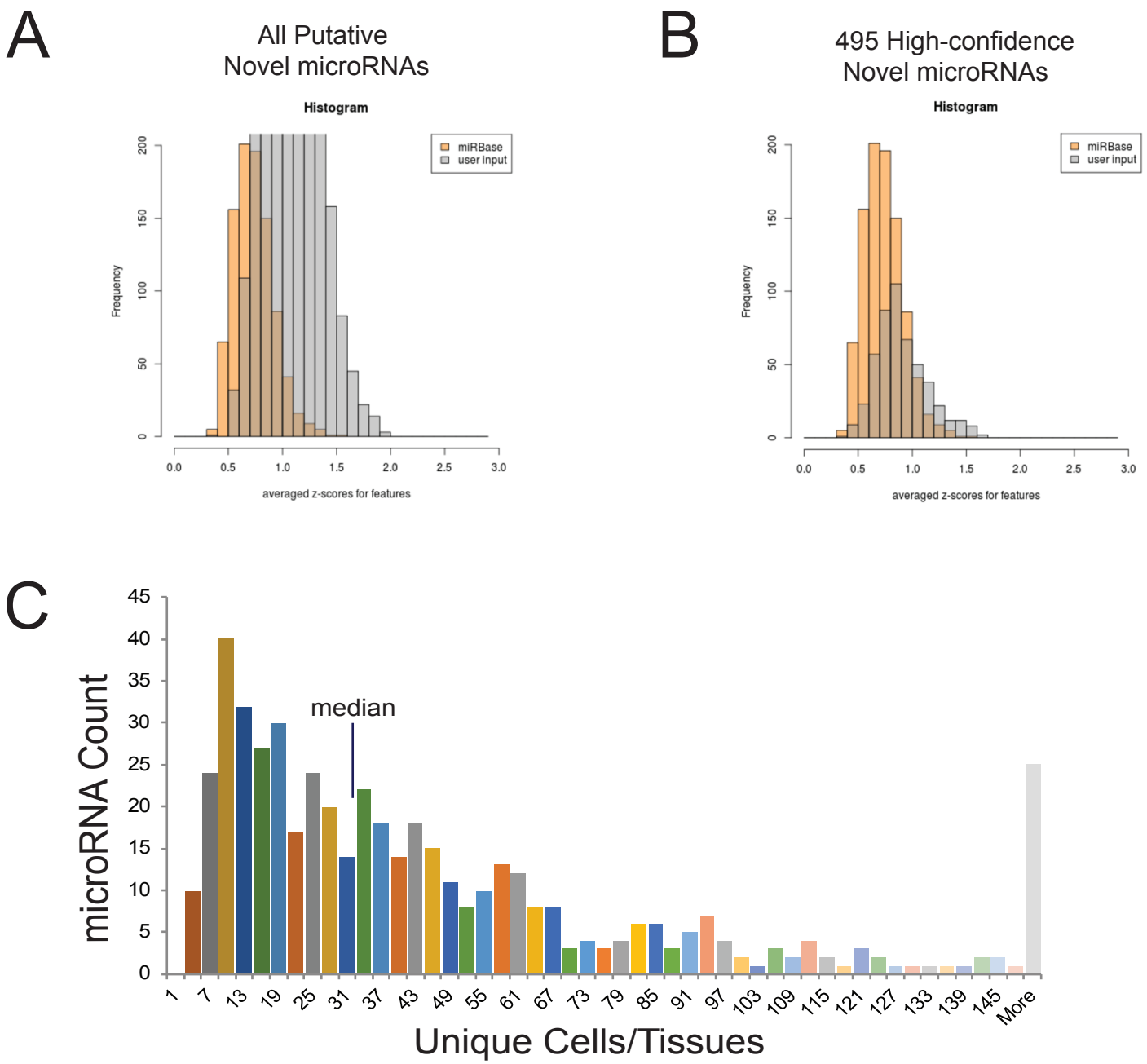
Supplemental Figure S5: CIBERSORT analysis of 37 colon samples. Colon samples had been obtained from a single study (Lin et al. 2016). We used the microRNAome data with the CIBERSORT deconvolution algorithm to show variable levels of specific cell types driving the global expression pattern of these tissues. In general, there was a higher percentage of blood in the ulcerative colitis samples (31.8% vs 7.9% and 12.4% for diverticulosis and Crohn's disease respectively), although this was variable between samples. The percent of epithelial cells was lower in the ulcerative colitis sample (19.9% vs 61.6% and 44.4%), while the percent of lymphocytes (avg. 7.2% vs. 4.7% and 4.3%) was increased. These large changes in the composition of the tissues will, in the absence of correction, impact on any cell-restricted microRNA differences that would be reported between these disease states.



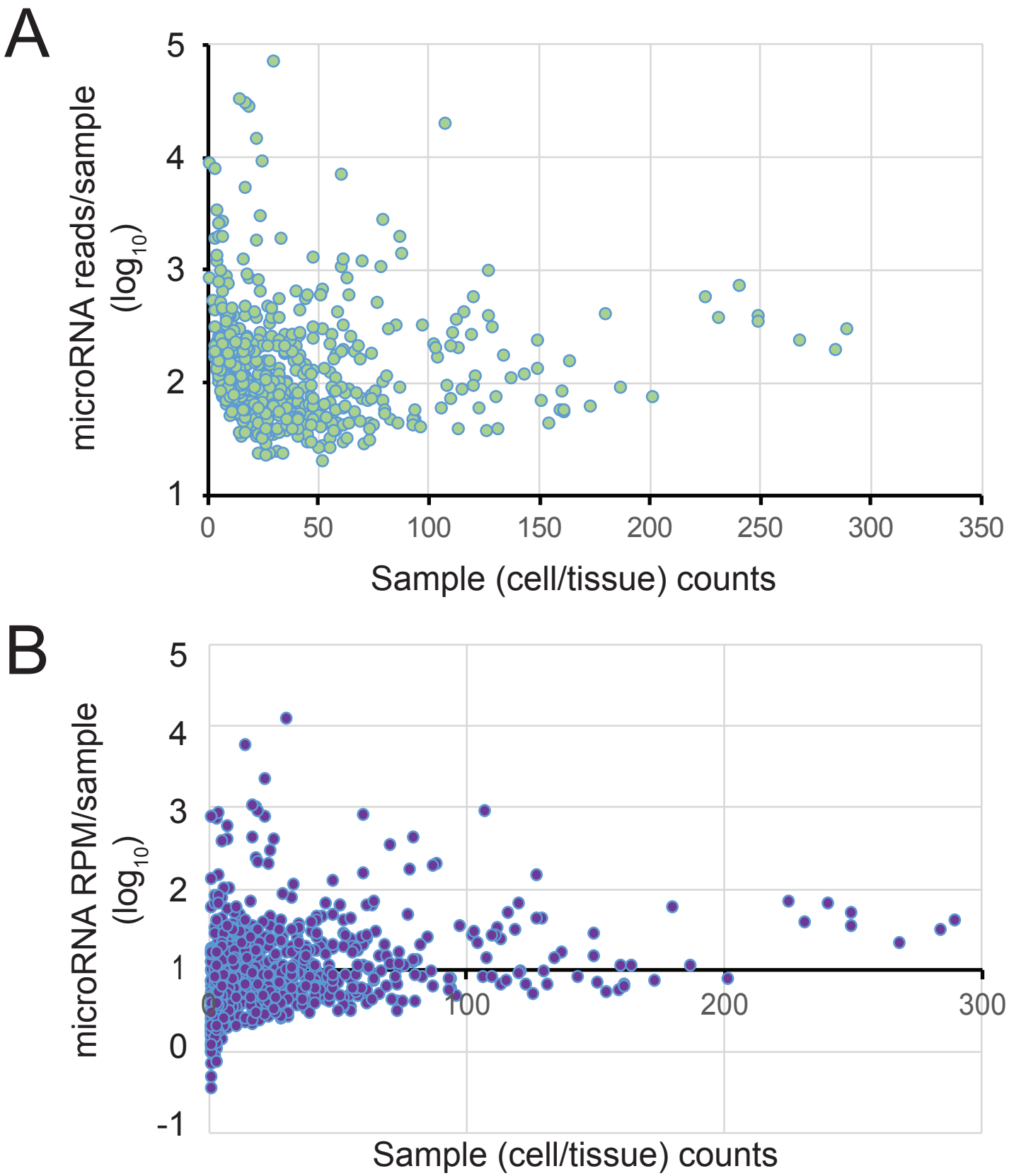
Supplemental Figure S6: Primary and immortalized fibroblasts and T lymphocytes generally have similar microRNA expression patterns. (A) A heat map of scaled Euclidean distances for 13 fibroblast lines demonstrate general clustering by primary and immortalized cells. HELF977, HAF1608 and IMR90 were immortalized from embryonic lung, adult skin and embryonic lung respectively. A H7 embryonic stem cell (ESC) line serves as an outgroup. (B) A MAplot of microRNA expression between primary and immortalized fibroblasts, based on the 250 most abundant microRNAs reveals that most microRNAs have similar expression. An arbitrary dotted line denotes a 3-fold change in expression. Only miR-1304-3p was markedly higher in primary fibroblasts. (C) Scaled Euclidean distances of 22 T lymphocyte primary and cancer-derived (Jurkat, SUP-T1 and C8166) samples show general clustering by primary or malignant cell type. H7 ESC was again used as an outgroup. (D) A MAplot of microRNA expression between primary and cancer T lymphocytes, based on the 350 most abundant microRNAs reveals that most microRNAs have similar expression levels. However >7 microRNAs have markedly different expression levels including the lymphocyte specific microRNA miR-150-5p which is 6 $\log_2$ fold higher in primary T lymphocytes. The primary biological function of these microRNAs may not be easily identified in these transformed cell lines.



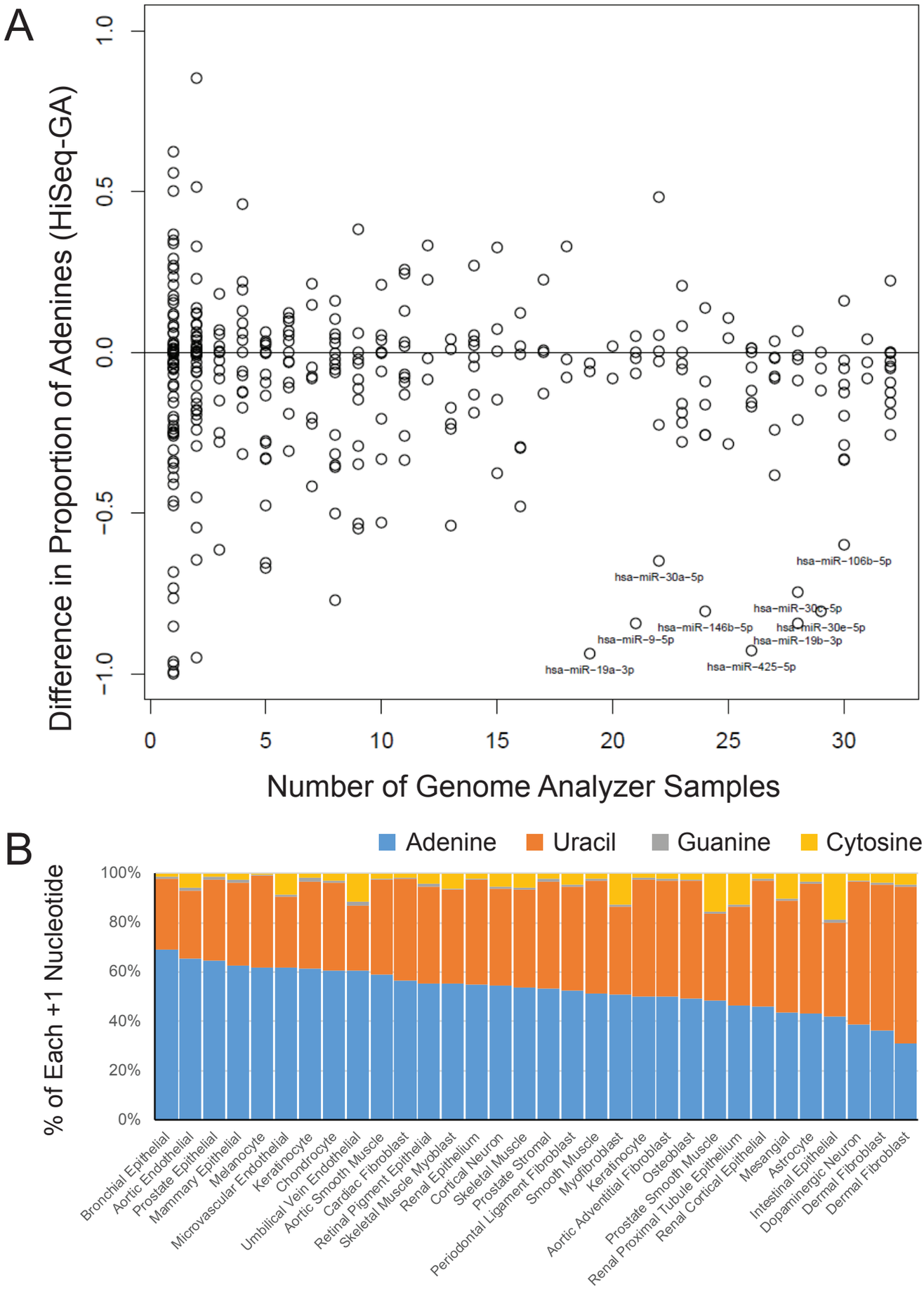
Supplemental Figure S7: HeLa cells are more variable between batches than fibroblasts. (A) After RUV correction, 7 HeLa cell samples and 10 primary fibroblast cell lines from 5 and 3 batches respectively were compared. A H7 ESC cell was used as an out group. The pairwise correlation for the HeLa Cell batches (SRA data from separate submissions) was between 0.45-0.75. One HeLa cell line, SRR1304304, clustered less closely to other HeLa cell samples than the H7 ESC cell line data. (B) For the fibroblasts, the correlation between batches was between 0.75-0.9, despite multiple organ sources (coronary artery, heart, lung, skin, & periodontal ligament). Fibroblasts demonstrated poorer correlation with the H7 ESC cell line.



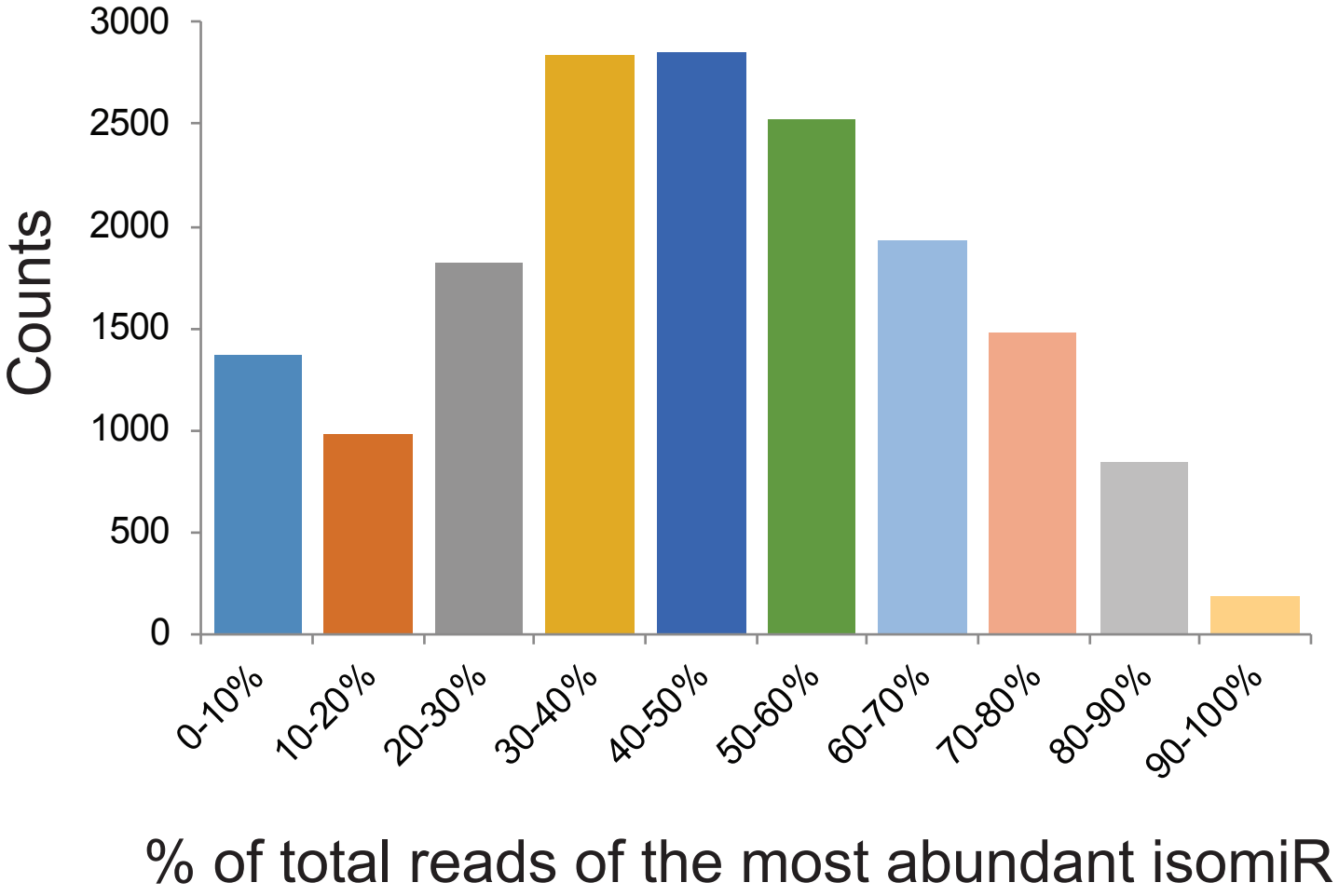
Supplemental Figure S8: Context of the novel microRNAs. (A) Histogram of z-scores for 3,333 representative novel microRNAs (grey) from the original 21,338 loci as computed by NovoMiRank and compared to the z-scores of all miRBase versions (orange). The average z-score was 1.07. A higher z-score indicates a novel microRNA is less similar to known microRNAs based on 24 features. (B) Histogram of z-scores for the 495 highest-confidence putative novel microRNAs. The average z-score was 0.90 which was much more similar to the miRBase collection. (C) Of 495 highest-confidence putative novel microRNAs, over half were found in 10-40 samples. The median number of samples per microRNA was 33. Twenty-five microRNAs were identified in more than 150 samples (max = 293) and this is represented by the last data column compressing the long tail in the data.



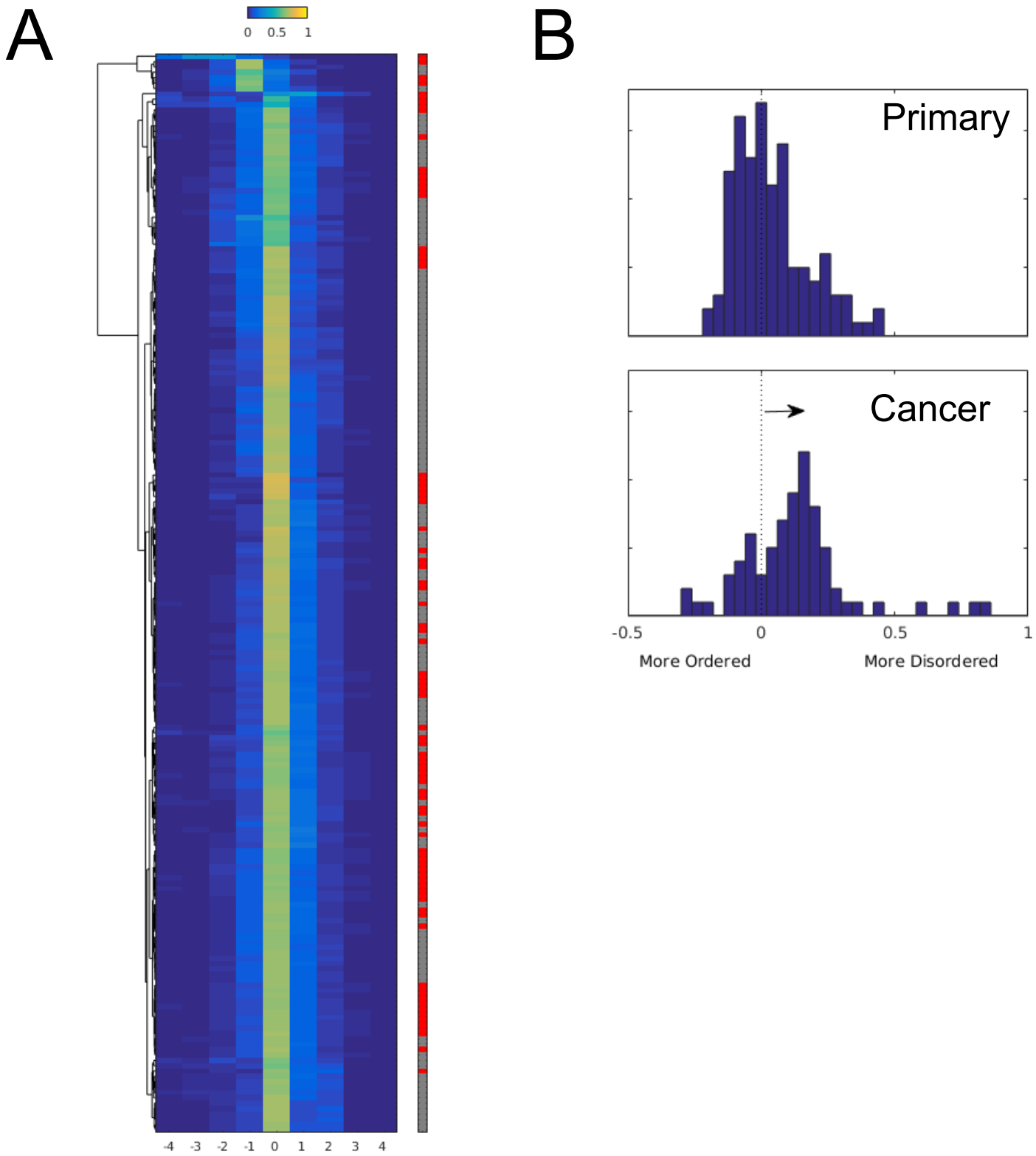
Supplemental Figure S9: No correlation between total reads and samples for 495 novel microRNAs. (A) The average number of microRNA reads/sample or (B), RPM/sample (normalized by miRNA read counts in Supplemental Tables 2-4) were compared to the number of samples each microRNA was detected in. There was no overall correlation suggesting that some abundant microRNAs were present specifically in certain samples and other microRNAs were present in many samples at lower levels.



Supplemental Figure S10: The distribution of +1 non-templated nucleotide additions. A) The proportion of nontemplated adenine additions were compared between samples run on a Genome Analyzer (GA) or HiSeq Illumina system for all primary and cancer cell line samples. Nine microRNAs were outliers driving a difference between systems. Analysis of these microRNAs revealed specific library preparation kits driving this difference in specific cell types. **B)** Thirty-two primary cell lines were grown in culture and processed for RNA-seq in two batches at the same sequencing center. At the +1 position from the most abundant canonical sequence, the non-templated nucleotide varied considerably between samples. A non-templated adenine was the most common substitution for most cell types, however uracils were more common amongst 6 samples including two dermal fibroblast lines.



Supplemental Figure S11: There is a wide distribution of most abundant isomiR sequences relative to total sequences for all microRNAs. All microRNAs with 1000+ total reads were evaluated and the most abundant templated sequence (-4 to +4 of the genomic sequence) or most abundant +1 non-templated length isomiR was identified for each. The total read counts of that sequence were divided by the total number of reads in that isomiR family and plotted above. The most abundant sequence for any given microRNA averaged 45% of all reads. For only 6% of microRNAs, the most abundant microRNA sequence represented 80+% of all reads.



Supplemental Figure S12: Cancer cell lines have more disorder in their canonical sequence distribution than primary cells. (A) The most abundant canonical length sequence was determined for all evaluated primary (grey) and cancer (red) cell lines. These were plotted on a heat map showing the abundance of each length from -4 to +4 relative to the canonical sequence in miRBase.org. Most samples behaved in a similar fashion with the miRBase canonical sequence being the most abundant length overall. Eight samples had their overall most abundant canonical sequences at a non 0 position and were removed from the analysis in (B). (B) A rank sum test of order in the samples identified increased disorder ($p<0.005$) among the cancer cell lines as a result of a greater frequency of both increased and decreased length variation in the canonical sequence. This finding was more pronounced with the 8 removed samples in the analysis.