A method for working with non-model organism RNA-seq data in R.

library(girafe)

Import bam file (This makes an AlignedGenomeIntervals object):

my_bamfile <- agiFromBam("file_path")

Import gff (This makes a Genome_intervals_stranded object):

my_gff <- readGFF3("file_path", isRightOpen=T)

> **Note: I had problems here because the gff file I was using had some formatting differences from what it was expecting. The girafe package contains sample data (mouse miRNA reads), so you can reformat your Gis_object to be the same. In my case I had to add two columns: my_gff$family (geneIDs alone) and my_gff$type (mRNA, CDS or whatever).

Align gff and read info:

my_bamfile.io <- interval_overlap(my_bamfile, my_gff)

Make counts table:

my_counts1 <- table(as.character(my_gff$family)[unlist(my_bamfile.io)])
my_count_df1 <- data.frame(my_counts1)
colnames(my_count_df1) <- c("gene_ID", "sample_name")

> ** repeat these steps for each sample. geneID column should be named the same in each data frame, but sample name should be different (sample_1, sample_2, etc.)

Merge samples and replace NAs with 0:

counts_merged <- merge(my_count_df1, my_count_df2, all=T)
counts_merged <- replace(counts_merged, is.na(counts_merged), 0)

> **Note: I usually do not overwrite like this just in case I make a mistake. You can assign to a temp object like "x", double check, and then assign to the real object name- longer but safer. Also, merging is best done pairwaise (if you have 3+ samples, merge 1 at a time)

Create row names from gene ID column (necessary formatting for edgeR)

gene_IDs <- counts_merged[,1]
rownames(counts_merged) <- gene_IDs
counts_merged <-counts_merged[,2:whatever last column is]

This final count table can be used as is in edgeR.

To get sample data:
exDir <- system.file("extdata", package="girafe")
load(file.path(exDir, "anno_mm_genint.RData"))