‘Cellular abundance shapes function in piRNA-guided genome defense’

"Pavol Genzor, Parthena Konstantinidou*, Daniel Stoyko* *equal contributors*
Amirhossein Manzourolajdad, Celine Marlin Andrews, Alexandra R. Elchert, Constantinos Stathopoulos, Astrid D. Haase

This vignette describes the computational materials & methods associated with this manuscript. Please visit **HaaseLab/piRNA_Diversity github repository** to download functions used in the various scripts and analyses. Please refer to the GEO data set **GSE156058** associated with this study for adapter sequences and raw data. The analysis in this vignette was not performed with full data sets, but only subset of the data to demonstrate the materials and methods.

About small RNA Libraries.
To be able to account for the PCR duplication when quantifying individueal piRNA sequences, cloning procedure utilized adapter sequences containing multiple random nucleotides. Each ligated read contains 10 N’s or **U**nique **M**olecular **I**dentifiers **(UMIs)**. There are 8N at the 5-prime and 2N at the 3-prime of the small RNA that allows for 4^10 or 1,048,576 possible combinations.

Library structure: 5-FivePrimeAdapter–*NNNNNNNN*–**smallRNA**–*NN*–ThreePrimeAdapter-3

Pre-requisites
* Acquire the raw sequencing data from your facility or GEO at NCBI * Ensure that you have the appropriate 5-prime and 3-prime adapter sequences * Ensure you have access to computing cluster and a computer with R environment * Ensure that you have all the necessary software installed and running * NOTE: some images may not have rendered perfectly

Vignette Content

DATA PREPARATION

1. Pre-process fastq files.

First, the raw sequencing data was cleaned up with ***cutadapt*** by removing the adapter sequences and keeping the UMIs. Refer to ***cutadapt*** manual for option description. The following commands were executed in the *bash* environment in the terminal. Remember to placeholder variables for your own sequences, directories, and files on you machine.

```
bash$ # initiate appropriate vertion of cutadapt
bash$ module load cutadapt/2.3
bash$
bash$ # split the sequencing lane without trimming
bash$ cutadapt --no-trim -m 19 \
               -g FivePrimeAdapter \
               -a file\:InputFiles/ThreePrimeAdapterSequences.fa \
               -o OutputDirectory/{name}.SNT.fastq.gz \
               --untrimmed-o OutputDirectory/untrimmed.SNT.fastq.gz \
               InputDirectory/SequencingLane_R1_001.fastq.gz
bash$
bash$ # remove the constant adapter sequences
bash$ cutadapt -j 6 --trimmed-only -m 19 \
               -g FivePrimeAdapter \
               -a ThreePrimeAdapter \
               -o OutputDirectory/SampleName.10N.fastq.gz \
               InputDirectory/SampleName.SNT.fastq.gz
bash$
bash$

# SNT = split not trimmed, 10N = 10 UMIs
```

2. Generate unique sequence fasta files

Next, the small RNAs with UMIs can be cleared of PCR duplication. Full script of all functions can be found at ***piRNA__Diversity github***. In the next few chunks we will:

- initiate R session, and load custom function
- remove PCR duplicates using UMIs
- remove UMIs
- export the cleaned reads into **.fasta**

```
# NOTE: If you are working on small files or have a lot of memory,
#       this can be run on laptop

bash$ # initiate an R session
bash$ module load R/4.0.5
```

```
bash$
bash$ ## enter R environment
bash$ R
```

```
# NOTE: Below script sets up the directories and loads the functions

# IMPORTANT: directory on your machine with all functions in this vignette
#            downloaded from github
FUN.DIR <- "/Users/YourUsername/YourDiversityScriptsFolder/"

# File directories
FASTQ.FILE = "SampleName_conditionA_replicate1.SNT.fastq.gz"
OUTPUT.DIR = "/YourComputer/WorkingDirectory/fastaFiles/"

## Load function
source(paste0(FUN.DIR,"prepareFastq.R"))

## Run the analysis on single file
prepareFastq(FASTQ.FILE = FASTQ.FILE,
             OUTPUT.DIR = OUTPUT.DIR,
             REMOVE.PCR.DUPLICATES = TRUE,
             REMOVE.UMI.N = TRUE,
             FIVE.PRIME.N.NUMBER = 8,
             THREE.PRIME.N.NUMBER = 2,
             FILTER.BY.SIZE = TRUE,
             SIZE.RANGE = c(18,32))
```

- The prepareFastq() function will generate three folders:
  - **libraryStats**
    - contains statistics with information about processing (*.stats.txt)
  - **totalReads**
    - contains all the piRNA reads including duplicates (*.ALLREADS.fa)
  - **uniqueSequences**
    - contains only unique piRNA sequences only (*.UNIQSEQS.fa)

3. Remove structural contaminants and align to the genome

Next, we proceed with previously generated ****.UNIQSEQS.fa*** and use ***STAR*** aligner to align data:

- First, align sequences to structural genome (tRNA, rRNA, snoRNA, …)
  - Structural RNAs can be downloaded from USCS
  - Follow STAR manual for instruction on how to prepare genome index
- Second, align un-mapped sequences from previous step to specie-specific genome
  - There are various sources where species genome files can be downloaded
  - Follow STAR manual for instruction on how to prepare genome index

```
# NOTE: Aligning to structural genome

bash$ # load appropriate version of aligner
bash$ module load STAR/2.5.2b
bash$
bash$ # align to structural genome
bash$ STAR  --runThreadN 6 \
            --runMode alignReads \
            --genomeDir /LocationOfYourStructuralGenome/ \
            --outSAMtype BAM SortedByCoordinate \
            --limitBAMsortRAM 10000000000 \
            --outFilterMultimapNmax 100 \
            --outFilterMismatchNmax 1 \
            --outReadsUnmapped Fastx \
            --outSAMattributes NH HI NM MD AS nM \
            --readFilesIn /InputDirectory/SampleName.UNIQSEQS.fa \
            --outFileNamePrefix /OutputDirectory/mappedToStructural/SampleName.
```

```
# NOTE: Aligning to specie-specific genome
# NOTE:   - use *.Unmapped.out.mate1 file generated after
          previous code has been executed

bash$ # align to specie genome
bash$ STAR  --runThreadN 6 \
            --runMode alignReads \
            --genomeDir /LocationOfYourSpeciesGenome/ \
            --sjdbGTFfile /LocationOfYourSpeciesFiles/gene.gtf \
            --outSAMtype BAM SortedByCoordinate \
            --limitBAMsortRAM 10000000000 \
            --outFilterMultimapNmax 100 \
            --winAnchorMultimapNmax 100 \
            --sjdbOverhang 100 \
            --outSAMattributes NH HI NM MD AS nM \
            --readFilesIn /InputDirectory/mappedToStructural/SampleName.Unmapped.out.mate1 \
            --outFileNamePrefix /OutputDirectory/mappedToGenome/SampleName.
```

4. Load and pre-process files in the R.

Once the replicate ***bam** files are generated in the previous chunks, we load the data in R environment and prepare it for analysis:

- Load data into R
- Remove potential miRNA contaminants
- Combine replicate data sets
- Calculate sequence overlap between samples
- Combine overlapped data to get Total, Common, and Rare groups
- Full script of all functions can be found at ***piRNA_Diversity github***.

```
# NOTE: LOAD DATA INTO R
```

```r
# load function
source(paste0(FUN.DIR,"filterBam.R"))

## named vector of paths
BAM.PATH.L <- c("rep1"="/YourSTARoutputFolder/FH-Piwi-IP_1_1.Aligned.sortedByCoord.out.bam",
                "rep2"="/YourSTARoutputFolder/FH-Piwi-IP_1_2.Aligned.sortedByCoord.out.bam",
                "rep3"="/YourSTARoutputFolder/FH-Piwi-IP_1_3.Aligned.sortedByCoord.out.bam")

## load multiple files using lapply loop
BAM.L <- lapply(names(BAM.PATH.L), function(s){
  message(paste0("Processing: ",s))
  filterBam(BAMFILE = BAM.PATH.L[[s]],
            BSSPECIES = "Dmelanogaster",
            EXTENTION = ".Aligned.sortedByCoord.out.bam",
            SIMPLECIGAR = TRUE,
            INCLUDE.SECONDARY.ALIGNEMNT=FALSE,
            STANDARD.CONTIGS.ONLY = TRUE,
            GET.ORIGINAL.SEQUENCE = FALSE,
            PERFECT.MATCH.ONLY = TRUE,
            FILTER.BY.FLAG = TRUE,
            SELECTFLAG = c(0,16),
            USE.SIZE.FILTER=TRUE,
            READ.SIZE.RANGE = c(18,32),
            TAGS = c("NH","NM","MD"),
            WHAT = c("flag"),
            SPLIT.NAME.BY = "-") })
names(BAM.L) <- names(BAM.PATH.L)
```

```r
# NOTE: REMOVE miRNA CONTAMINATS

# libraries
library("GenomicRanges")

# load the function
source(paste0(FUN.DIR,"miRbase2BED.R"))

# path to miRbase annotation
miRBASE.PATH = "/Users/genzorp/Documents/DATA/Annotation/Dmelanogaster/
                dm6_miRBase_dme.gff3"

# make Genomic Range
mi.GR <- makeGRangesFromDataFrame(df = miRbase2BED(miRBASEFILE = miRBASE.PATH),
                                  keep.extra.columns = TRUE)

# Filter data
BAM.L <- lapply(BAM.L, function(s){subsetByOverlaps(x = s, ranges = mi.GR,
                                                    type = "any", invert = TRUE)})

## save ALL image to work with later
#save.image(file = "/Users/genzorp/Documents/GITHUB/piDiversity/sessions/revisions/
#three_filtered_bams_all.RData")
```

```
# NOTE: for simplicity only fraction of data was used here
#       - first 500 thousand sequences

SUBSET_500K=500000
BAM.L <- lapply(BAM.L,function(s){s[1:SUBSET_500K]})

## save loaded 500K to object for faster loading later
save.image(file = "/Users/genzorp/Documents/GITHUB/piDiversity/sessions/revisions/
           three_filtered_bams_500K.RData")
```

```
# NOTE: COMBINE REPLICATES

# load data
load(file = "/Users/genzorp/Documents/GITHUB/LIVE/piRNA_Diversity/exampleRData/three_filtered_bams_500K

# Load the function
source(paste0(FUN.DIR,"combineThreeGRS.R"))

# combine three technical replicates
BIO.REP <- suppressMessages(
  combineThreeGRS(GRL = BAM.L,
                  REPLICATE.NAMES = c("rep1","rep2","rep3"),
                  MC.CORES = 3))

# show results
BIO.REP
```

```
## GRanges object with 926340 ranges and 3 metadata columns:
##              seqnames            ranges strand |         N        NH      MULT
##                 <Rle>         <IRanges>  <Rle> | <integer> <integer> <integer>
##         [1]    chr2L         4646-4663      - |         3        54         4
##         [2]    chr2L         4714-4739      - |         2        55         3
##         [3]    chr2L         4718-4744      - |         1        55         1
##         [4]    chr2L         4721-4747      - |         2        55         2
##         [5]    chr2L         4722-4742      - |         2        55         2
##         ...      ...               ...    ... .       ...       ...       ...
##    [926336]    chr2L 19536850-19536869      + |         1         1         1
##    [926337]    chr2L 19536850-19536873      + |         1         1         1
##    [926338]    chr2L 19536850-19536875      + |         1         1         7
##    [926339]    chr2L 19536850-19536876      + |         1         1         1
##    [926340]    chr2L 19536851-19536875      + |         1         1         1
##   -------
##   seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

```
# all unique sequences
summary(duplicated(BIO.REP))
```

```
##    Mode   FALSE
## logical  926340
```

- While combining replicate data sets we:
  - Counted in how many replicates is each sequence present (N)
  - Recorded the highest sequence multi-mapping score (NH-tag)
  - Calculated the sequence cumulative abundance for identical genome positions (MULT)

```
# NOTE: CALCULATE SEQUENCE OVERLAPS

## Load the function
source(paste0(FUN.DIR,"threeSampleSequenceOverlaps.R"))

## Generate overlaps by sequence
BIOREP.OV.L <- suppressMessages(
  threeSampleSequenceOverlaps(GRL = BAM.L,
                              BSSPECIES = "Dmelanogaster",
                              MC.CORES = 3))
```

```
##
##  FALSE
## 500000
##
##  FALSE
## 500000
##
##  FALSE
## 500000
##    tempName realName Input_N Common_N Exclusive_N Pairwise Pairwise_N
## 1:       FA     rep1  500000   151910      166999       AB     243917
## 2:       FB     rep2  500000   151910      167334       BC     240659
## 3:       FC     rep3  500000   151910      170257       CA     240994
```

```
## show list content
names(BIOREP.OV.L)
```

```
## [1] "Common"     "Exclusive"  "Pairwise"   "SampleInfo"
```

- The objects generated during sequence overlap process are lists:
  - **Common** GR - found in each biological replicate
  - **Rare (Exclusive)** GR - found only in individual biological replicates
  - **Pairwise** GR - shared between pairs of biological replicates
  - **SampleInfo** table summarizing overlaps that can be use to create a Venn diagram

```
# NOTE: COMBINTE TO GET TOTAL, COMMON, AND RARE
# BEWARE: combining large data sets can be computationally very expensive

# TOTAL
TOTAL.GR <- suppressMessages(
  combineThreeGRS(GRL = BAM.L,
                  REPLICATE.NAMES = names(BAM.L),
                  MC.CORES = 3))

# ensure there are no duplicates
```

```
#summary(duplicated(TOTAL.GR))

## EXCLUSIVE
RARE.GR <- c(BIOREP.OV.L[["Exclusive"]][[1]],
             BIOREP.OV.L[["Exclusive"]][[2]],
             BIOREP.OV.L[["Exclusive"]][[3]])

# ensure there are no duplicates
#summary(duplicated(RARE.GR))

## COMMON
COMMON.GR <- suppressMessages(
  combineThreeGRS(GRL = BIOREP.OV.L[["Common"]],
                  REPLICATE.NAMES = names(BIOREP.OV.L[["Common"]]),
                  MC.CORES = 3))

# ensure each is found in all samples
#unique(mcols(COMMON.GR)[["N"]])
```

- After performing all the above operation, it is convenient to identify object that are intended for analysis, and save them into a *RData object for faster loading. Otherwise, depending on the size of the samples, the data preparation will take long time.

FIGURE-RELATED SCRIPTS

- **Figure 1**
- ***"The sequence diversity of piRNAs exceeds the capacity of an individual cell and generates cell-to-cell variability"***

```
# NOTE: Figure 1A

# load saved data
load("/Users/genzorp/Documents/GITHUB/piDiversity/sessions/revisions/three_filtered_bams_all_brep_ov_co

# load libraries
suppressPackageStartupMessages({library("data.table"); library("dplyr");
  library("ggplot2"); library("ggpubr"); library("GenomicRanges")})

# calculate Sequence versus Read ratio (SoR)
SoR <- lapply(names(BAM.L), function(i){
  GR <- BAM.L[[i]]
  DT <- suppressWarnings(as.data.table(GR, keep.rownames = TRUE))
  DT <- DT[,c("rn","MULT")]
  NDT <- data.table(Sample = i,
                    Sequences = nrow(DT),
                    Reads = sum(DT[["MULT"]]),
                    SeqPerRead = nrow(DT)/sum(DT[["MULT"]]))
```

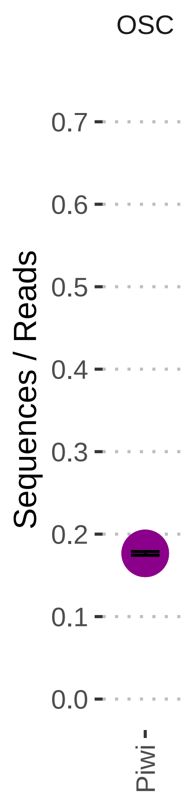
```r
  return(NDT) })

# Combine and view results
SoR <- setDT(bind_rows(SoR))
SoR
```

```
##    Sample Sequences    Reads SeqPerRead
## 1:   rep1   4499298 25621559  0.1756059
## 2:   rep2   4478531 25566855  0.1751694
## 3:   rep3   4423222 24681183  0.1792143
```

```r
# add information and summarize
SoR[["Sample"]] <- c(paste0("OSC",1:3))
SoR[["Species"]] <- c(rep("OSC",3))
SoR[["Protein"]] <- c(rep("Piwi",3))
SoR[["Mean"]] <- mean(SoR[["SeqPerRead"]])
SoR[["SD"]] <- sd(SoR[["SeqPerRead"]])
SoR
```

```
##    Sample Sequences    Reads SeqPerRead Species Protein      Mean          SD
## 1:   OSC1   4499298 25621559  0.1756059     OSC    Piwi 0.1766632 0.002220082
## 2:   OSC2   4478531 25566855  0.1751694     OSC    Piwi 0.1766632 0.002220082
## 3:   OSC3   4423222 24681183  0.1792143     OSC    Piwi 0.1766632 0.002220082
```

```r
# plot
ggplot() + theme_pubclean() +
  geom_point(data = SoR, aes(x=factor(Protein), y = Mean, colour = Protein),
             shape = 16, size=8) +
  geom_errorbar(data = SoR, aes(x = Protein, ymin = Mean-SD, ymax = Mean+SD),
                colour = "black", width = 0.4, size = 0.5) +
  scale_colour_manual(values =c("darkmagenta")) +
  facet_wrap(~Species, ncol = 3, scales = "free_x") +
  scale_y_continuous(limits = c(0,0.75), breaks = seq(0,1,0.1)) +
  labs(x="", y="Sequences / Reads", fill="") +
  theme(aspect.ratio = 8, legend.position = "none",
        strip.background = element_blank(),
        axis.text.x = element_text(angle = 90, hjust = 0, vjust = 0.5),
        axis.text = element_text(size = 10))
```



```
# NOTE: Figure 1B

# Accumulation Curve#
# If you want to use the rational function approximation that asymptotically approaches
# a linear curve, you have to download the old version 1.2.1 at
# https://cran.r-project.org/src/contrib/Archive/preseqR/ and use the old function
# preseqR.rfa.species.accum.curve (set the option asym.linear=1 inside the function).
# before proceeding:
# install.packages('preseqR') after locally downloading the above version.
# install.packages("local destination",repos=NULL,type="source")

# libraries
library('stringr');library('gplots');library('ggplot2')
library('data.table');library('plyr');library('dplyr')
library('RColorBrewer');library('dendextend');library('cluster')
library('reshape2');library('rdist');library('rdist')
library('factoextra');library('treeClust');library('clv')
library('preseqR');library('scales')

# ELBOW POINT:
# Function to calculate the elbow point. Point with maximum curvature is defined as
# elbow point of saturation.This function elbow takes input arrays (x and y).
# It will find the point with maximum curvature at index'ind' and returns values x(ind) and y(ind).

elbow <-function(inputx,inputy,scalex,scaley){
  ninputx=inputx/(scalex)
  ninputy=inputy/(scaley)
  ninputx=smooth(ninputx);ninputy=smooth(ninputy)
  curvature=rep(0,length(inputy))
  ypvector=rep(0,length(inputy))
  for (i in 2:(length(inputy)-1)){
    yp=(ninputy[i+1]-ninputy[i-1])/(ninputx[i+1]-ninputx[i-1])
    ypp=(ninputy[i+1]-2*ninputy[i]+ninputy[i-1])/(((ninputx[i+1]-ninputx[i-1])/2)^2)
    curvature[i]= abs(ypp)/((1+(yp^2))^(3/2))
    ypvector[i]=yp }
  curvature[1]=curvature[2]
  ypvector[1]=ypvector[2]
  curvature[i+1]=curvature[i]
  ypvector[i+1]=ypvector[i]
  curvature=smooth(curvature)
  inputx=smooth(inputx);inputy=smooth(inputy)
  dx=(inputx[which.max(curvature)+1]-inputx[which.max(curvature)-1])/2
  dy=(inputy[which.max(curvature)+1]-inputy[which.max(curvature)-1])/2
  ret=c(inputx[which.max(curvature)],inputy[which.max(curvature)],dx,dy)
  return(ret) }

#======= load data ================
Exp1path="data/03162020/MILI_PS_1_Multiplicity.csv"
Exp2path="data/03162020/MILI_PS_2_Multiplicity.csv"
Exp3path="data/03162020/MIWI_PS_1_Multiplicity.csv"
Exp4path="data/03162020/MIWI_PS_2_Multiplicity.csv"
Exp5path="data/03122020/FH-Piwi-IP_1_Multiplicity.csv"
Exp6path="data/03122020/FH-Piwi-IP_2_Multiplicity.csv"
Exp7path="data/03122020/FH-Piwi-IP_3_Multiplicity.csv"

Exp1=read.csv(Exp1path,sep=",")
Exp2=read.csv(Exp2path,sep=",")
Exp3=read.csv(Exp3path,sep=",")
Exp4=read.csv(Exp4path,sep=",")
Exp5=read.csv(Exp5path,sep=",")
Exp6=read.csv(Exp6path,sep=",")
Exp7=read.csv(Exp7path,sep=",")

exp1=data.matrix(Exp1,rownames.force = NA)
exp2=data.matrix(Exp2,rownames.force = NA)
exp3=data.matrix(Exp3,rownames.force = NA)
exp4=data.matrix(Exp4,rownames.force = NA)
exp5=data.matrix(Exp5,rownames.force = NA)
exp6=data.matrix(Exp6,rownames.force = NA)
exp7=data.matrix(Exp7,rownames.force = NA)

#==================================
#Real Data:
total=c(sum(exp1[,1]*exp1[,2]),sum(exp2[,1]*exp2[,2]),
        sum(exp3[,1]*exp3[,2]),sum(exp4[,1]*exp4[,2]),
        sum(exp5[,1]*exp5[,2]),sum(exp6[,1]*exp6[,2]),
        sum(exp7[,1]*exp7[,2]))

unique=c(sum(exp1[,2]),sum(exp2[,2]),
         sum(exp3[,2]),sum(exp4[,2]),
         sum(exp5[,2]),sum(exp6[,2]),
         sum(exp7[,2]))

maxtotal = max(total)

#Estimation:
estimatorExp1 <- preseqR.rSAC.bootstrap(exp1, r=1)
estimatorExp2 <- preseqR.rSAC.bootstrap(exp2, r=1)
estimatorExp3 <- preseqR.rSAC.bootstrap(exp3, r=1)
estimatorExp4 <- preseqR.rSAC.bootstrap(exp4, r=1)
estimatorExp5 <- preseqR.rSAC.bootstrap(exp5, r=1)
estimatorExp6 <- preseqR.rSAC.bootstrap(exp6, r=1)
estimatorExp7 <- preseqR.rSAC.bootstrap(exp7, r=1)

#==================================
predict=20
predict1=predict/(total[1]/maxtotal)
predict2=predict/(total[2]/maxtotal)
predict3=predict/(total[3]/maxtotal)
predict4=predict/(total[4]/maxtotal)
predict5=predict/(total[5]/maxtotal)
predict6=predict/(total[6]/maxtotal)
predict7=predict/(total[7]/maxtotal)
t1=c(seq(0,0.99,by=0.01),seq(1,predict1,by=0.01))
t2=c(seq(0,0.99,by=0.01),seq(1,predict2,by=0.01))
t3=c(seq(0,0.99,by=0.01),seq(1,predict3,by=0.01))
t4=c(seq(0,0.99,by=0.01),seq(1,predict4,by=0.01))
t5=c(seq(0,0.99,by=0.01),seq(1,predict5,by=0.01))
t6=c(seq(0,0.99,by=0.01),seq(1,predict6,by=0.01))
t7=c(seq(0,0.99,by=0.01),seq(1,predict7,by=0.01))
uniqueExp1=estimatorExp1$f(t1)
uniqueExp2=estimatorExp2$f(t2)
uniqueExp3=estimatorExp3$f(t3)
uniqueExp4=estimatorExp4$f(t4)
uniqueExp5=estimatorExp5$f(t5)
uniqueExp6=estimatorExp6$f(t6)
uniqueExp7=estimatorExp7$f(t7)

xExp1=total[1]*t1
xExp2=total[2]*t2
xExp3=total[3]*t3
xExp4=total[4]*t4
xExp5=total[5]*t5
xExp6=total[6]*t6
xExp7=total[7]*t7

setnames <-c(Exp1path,Exp2path,Exp3path,Exp4path,Exp5path,Exp6path,Exp7path)
csvfilenames <- substring(setnames, regexpr("\\/[^\\/]*$", setnames) +1 )
multnames <- as.character(c(str_remove(csvfilenames, ".csv"),"real"))

x.lim = min(max(xExp1),max(xExp2),max(xExp3),max(xExp4),max(xExp5),max(xExp6),max(xExp7))
y.lim = max(uniqueExp1,uniqueExp2,uniqueExp3,uniqueExp4,uniqueExp5,uniqueExp6,uniqueExp7)
xcut=x.lim*1.05
xExp1max=length(xExp1[which(xExp1 <= xcut)])
```

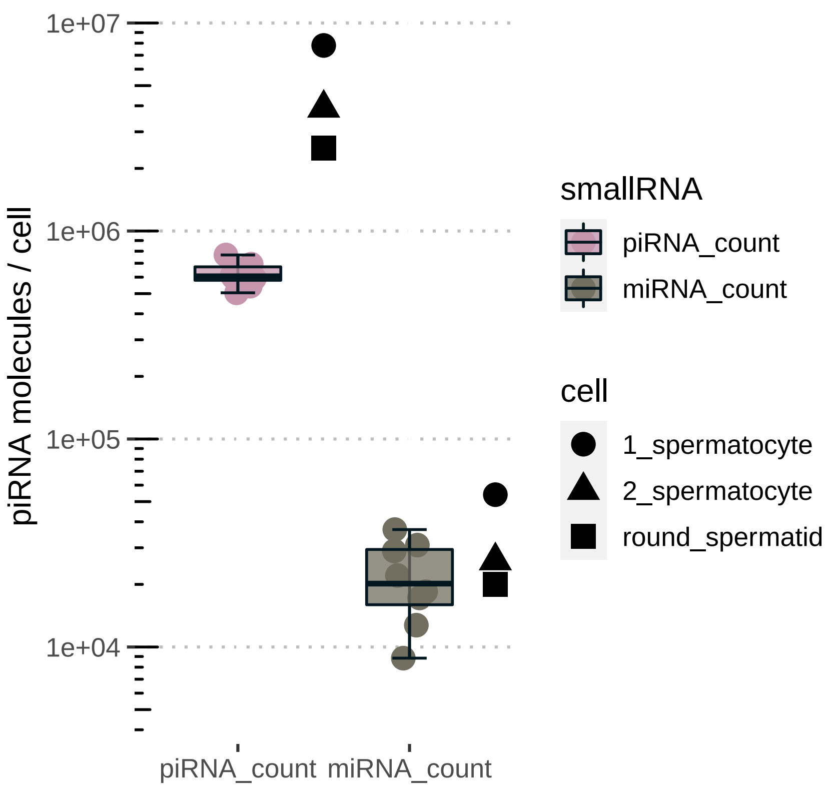
```r
xExp2max=length(xExp2[which(xExp2 <= xcut)])
xExp3max=length(xExp3[which(xExp3 <= xcut)])
xExp4max=length(xExp4[which(xExp4 <= xcut)])
xExp5max=length(xExp5[which(xExp5 <= xcut)])
xExp6max=length(xExp6[which(xExp6 <= xcut)])
xExp7max=length(xExp7[which(xExp7 <= xcut)])

df1 <- data.frame("abundance" = xExp1[1:xExp1max], "unique" = uniqueExp1[1:xExp1max],
                  set = multnames[1], stringsAsFactors = FALSE)
df12 <- data.frame("abundance" = xExp2[1:xExp2max], "unique" = uniqueExp2[1:xExp2max],
                   set = multnames[2], stringsAsFactors = FALSE)
df123 <- data.frame("abundance" = xExp3[1:xExp3max], "unique" = uniqueExp3[1:xExp3max],
                    set = multnames[3], stringsAsFactors = FALSE)
df1234 <- data.frame("abundance" = xExp4[1:xExp4max], "unique" = uniqueExp4[1:xExp4max],
                     set = multnames[4], stringsAsFactors = FALSE)
df12345 <- data.frame("abundance" = xExp5[1:xExp5max], "unique" = uniqueExp5[1:xExp5max],
                      set = multnames[5], stringsAsFactors = FALSE)
df123456 <- data.frame("abundance" = xExp6[1:xExp6max], "unique" = uniqueExp6[1:xExp6max],
                       set = multnames[6], stringsAsFactors = FALSE)
df1234567 <- data.frame("abundance" = xExp7[1:xExp7max], "unique" = uniqueExp7[1:xExp7max],
                        set = multnames[7], stringsAsFactors = FALSE)

# real data
dfreal <- data.frame("abundance" = total, "unique" = unique, set = "real")

# combine tables
df <- setDT(rbind(df1,df12,df123,df1234,df12345,df123456,df1234567,dfreal))
df[["set"]] <- as.character(df[["set"]])
legendTitle = "set"
df[["set"]] <- factor(df[["set"]], levels = multnames)
plot_colors <- c("dark green","light green","dark blue","light blue",
                 "purple","red","pink","black")

# Plot
ggplot(df, aes(x = abundance, y = unique, colour = set)) +
  scale_color_manual(values = plot_colors, name = paste(legendTitle,sep = ""),
                     labels = paste(multnames, sep = "")) +
  geom_vline(aes(xintercept=df[set=="real",]$abundance[1]),colour=plot_colors[1],
             linetype="dashed",size = 0.3) +
  geom_vline(aes(xintercept=df[set=="real",]$abundance[2]),colour=plot_colors[2],
             linetype="dashed",size = 0.3) +
  geom_vline(aes(xintercept=df[set=="real",]$abundance[3]),colour=plot_colors[3],
             linetype="dashed",size = 0.3) +
  geom_vline(aes(xintercept=df[set=="real",]$abundance[4]),colour=plot_colors[4],
             linetype="dashed",size = 0.3) +
  geom_vline(aes(xintercept=df[set=="real",]$abundance[5]),colour=plot_colors[5],
             linetype="dashed",size = 0.3) +
  geom_vline(aes(xintercept=df[set=="real",]$abundance[6]),colour=plot_colors[6],
             linetype="dashed",size = 0.3) +
  geom_vline(aes(xintercept=df[set=="real",]$abundance[7]),colour=plot_colors[7],
             linetype="dashed",size = 0.3) +
  geom_point(data = subset(df, set == "real"), colour = plot_colors[8]) +
  geom_line(data = subset(df, set != "real")) +
  scale_x_continuous(limits = c(0,4e8), breaks = seq(0,4e8,5e7)) +
  scale_y_continuous(limits = c(0,13e6), breaks = seq(0,13e6,1e6)) +
  ggtitle(outputpath) +
  theme_bw() +
  theme(plot.title = element_text(size = 8),
        panel.grid = element_blank(),
        aspect.ratio = 1,
        legend.position = "bottom")
```

```
# NOTE: Figure 1C

# load quantification data
QUANT.CSV="/Users/genzorp/Documents/GITHUB/piDiversity/data/QUANT2/Thenia/piRNA_quantification_data.csv
QUANT.DTM <- melt.data.table(data = fread(QUANT.CSV),
                             id.vars = c("Sample"),
                             measure.vars = c("piRNA_count","miRNA_count"),
                             variable.name = "smallRNA",
                             value.name = "count" )
# take a peak
head(QUANT.DTM, n=3)
```

```
##    Sample    smallRNA    count
## 1:      A piRNA_count 592158.9
## 2:      B piRNA_count 767060.5
## 3:      C piRNA_count 692749.4
```

```
# published data of piRNA and miRNA quantification: PMID:30193099
ZAM.PIMI.DT <- data.table("cell"=c("1_spermatocyte","2_spermatocyte","round_spermatid"),
                          "piRNA_count"=c(7800000,3900000,2500000),
                          "miRNA_count"=c(54000,26000,20000))

ZAM.PIMI.DTM <- melt.data.table(data = ZAM.PIMI.DT, id.vars = "cell",
                                variable.name = "smallRNA",
                                value.name = "count")

# plot settings
two_cols <- c("#C695AE","#726E60")

# plot the piRNA abundances
ggplot() + theme_pubclean() +
  geom_point(data = QUANT.DTM, aes(x = smallRNA, y = count, colour = smallRNA),
             position = position_jitterdodge(),
             size = 4, shape = 16) +
  stat_boxplot(data = QUANT.DTM, aes(x = smallRNA, y = count, fill = smallRNA),
               geom = "errorbar", width = 0.2, colour = "#01161E") +
  geom_boxplot(data = QUANT.DTM, aes(x = smallRNA, y = count, fill = smallRNA),
               alpha = 0.75, width = 0.5, colour = "#01161E") +
  geom_point(data = ZAM.PIMI.DTM, aes(x = smallRNA, y = count, shape = cell),
             size = 4, position = position_nudge(x = 0.5)) +
  scale_fill_manual(values = two_cols) +
  scale_colour_manual(values = two_cols) +
  scale_y_continuous(trans = "log10", limits = c(5e3,1e7)) +
  annotation_logticks(sides = "l") +
  ylab("piRNA molecules / cell") + xlab("") +
  theme(aspect.ratio = 2, legend.position = "right")
```

```
## Warning: Ignoring unknown aesthetics: fill
```

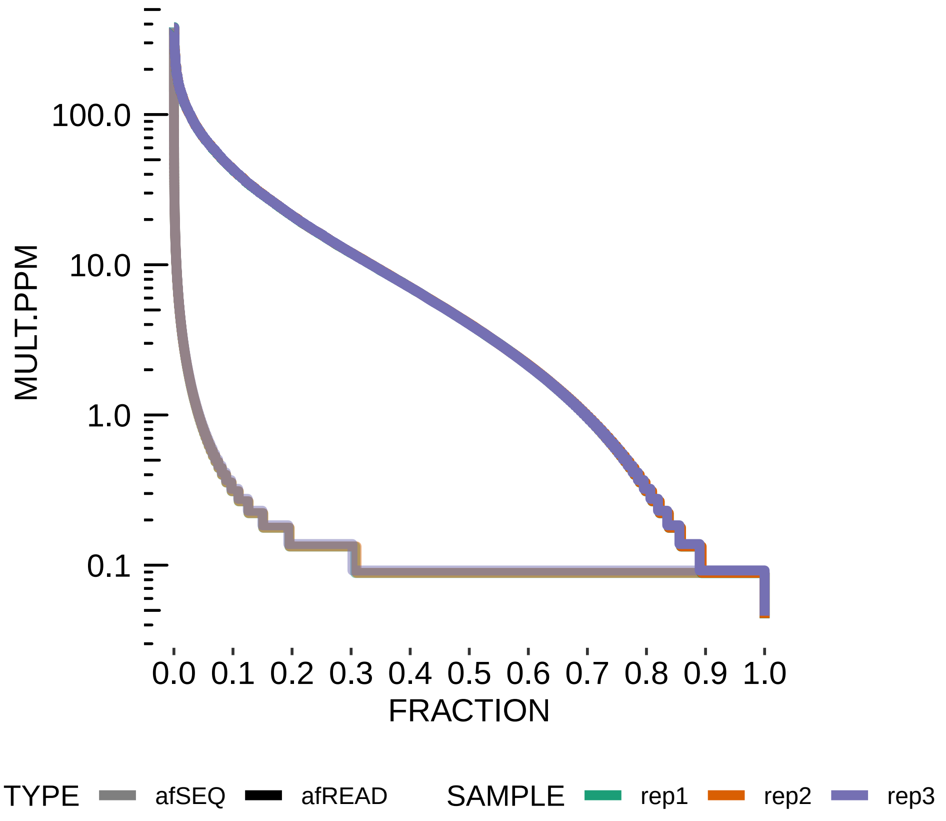


- **Figure 2**
- ***"A skewed distribution of sequence abundance results in a few common and many rare piRNAs."***

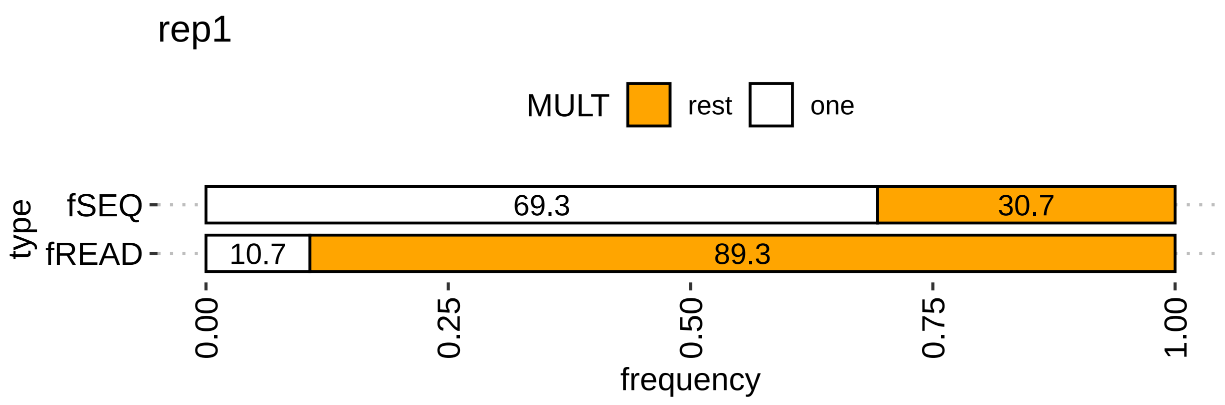
```
# NOTE: Figure 2A

# Load the functions
source(paste0(FUN.DIR,"simpleGRFilter.R"))
source(paste0(FUN.DIR,"simpleStepsPlot.R"))
source(paste0(FUN.DIR,"simpleMultBarPlot.R"))
source(paste0(FUN.DIR,"simpleSRViolin.R"))

# select only the sequences within desired size range
OSC.S249.BIOREP.L <- lapply(BAM.L, function(s){
  suppressMessages(
    simpleGRFilter(GR = s,
                   SIZE.RANGE = c(24,29))) })

## Fig 2A-C
```

```r
# Plot Steps
suppressMessages(
  simpleStepsPlot(GRL = OSC.S249.BIOREP.L,
                  Y.PPM = TRUE))
```

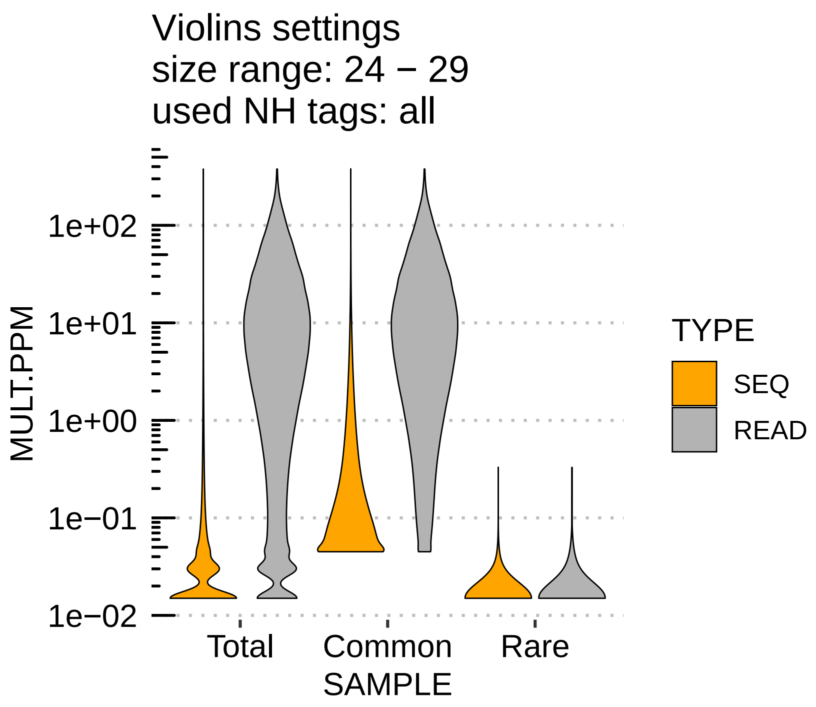
```
## Warning: Using alpha for a discrete variable is not advised.
```

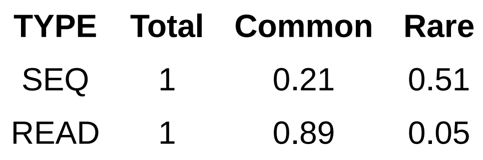


```r
# Plot fraction
suppressMessages(
  simpleMultBarPlot(GR = OSC.S249.BIOREP.L[["rep1"]],
                    RANGE.NAME = "rep1"))
```



```r
## Fig 2D-E

# Combine ranges into list
GRL <- list("Common"=COMMON.GR, "Rare"=RARE.GR, "Total"=TOTAL.GR)

# Plot violins
suppressMessages(
  simpleSRViolin(GRL = GRL,
                 SOURCE.DIR = FUN.DIR,
                 NH.TAG = NULL,
                 SIZE.RANGE = c(24,29),
                 SAMPLE.ORDER = c("Total","Common","Rare"),
                 Y.PPM = TRUE))
```



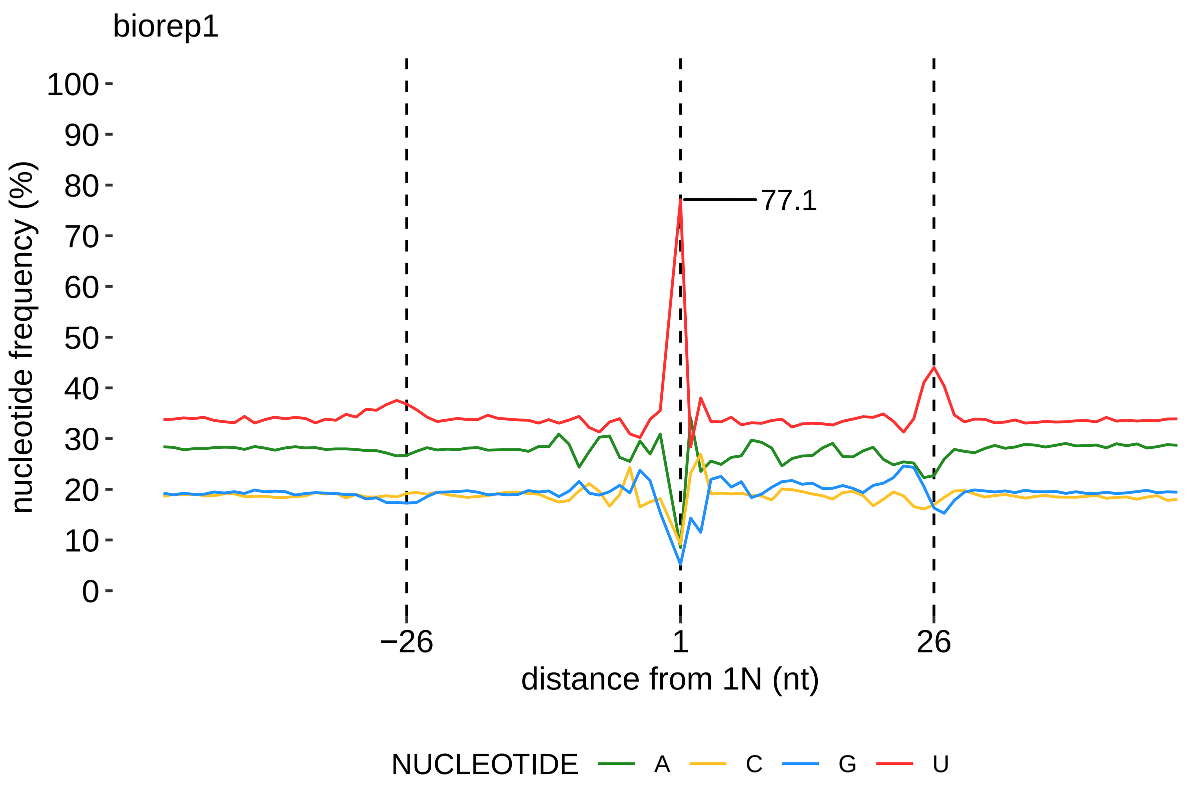


TYPE	Total	Common	Rare
SEQ	1	0.21	0.51
READ	1	0.89	0.05

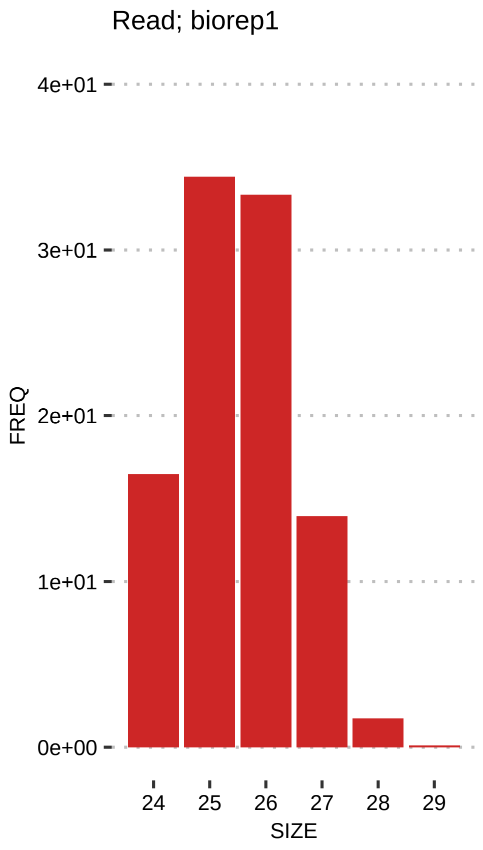
```
## TableGrob (2 x 1) "arrange": 2 grobs
##   z     cells    name              grob
## 1 1 (1-1,1-1) arrange     gtable[layout]
## 2 2 (2-2,1-1) arrange gtable[colhead-fg]
```

- **Figure S2 and Figure 3B**
- ***Metagene analysis***

```r
# NOTE: Metagene analysis for Figure S2 and Figure 3B

# load functions
source(paste0(FUN.DIR,"simpleGRFilter.R"))
source(paste0(FUN.DIR,"simpleMetageneTab.R"))
source(paste0(FUN.DIR,"simpleMetageneRegularPlotGG.R"))
source(paste0(FUN.DIR,"simpleSDfromGR.R"))

# Filter the data
BIO.REP.U.S249 <- suppressMessages(
  simpleGRFilter(GR = BIO.REP,
                 RANGE.NAME = "biorep1",
                 NH.TAG = 1,
                 SIZE.RANGE = c(24,29)))

# Generate metagene data tables
#   - For 5-prime, for 3-prime change the ALIGN.END
MG.TAB <- suppressMessages(
  simpleMetageneTab(GR = BIO.REP.U.S249,
                    RANGE.NAME = "biorep1",
                    BSSPECIES = "Dmelanogaster",
                    USE.READS = TRUE,
                    ALIGN.END = 5,
                    EXPAND.BY = 50) )

# plot 5-prime metagene
suppressMessages(
  simpleMetageneRegularPlotGG(METAGENE.DT = MG.TAB[["frequency"]],
                              SAMPLE.NAME = "biorep1",
                              Y.LIMITS = c(0,100),
                              PIRNA.SIZE = 26,
                              ASPECT.RATIO = 0.5) )
```



```
# plot size distribution of the sequences or reads
suppressMessages(
  simpleSDfromGR(GR = BIO.REP.U.S249,
                 SAMPLE.NAME = "biorep1",
                 USE.READS = TRUE,
                 PLOT.FREQ = TRUE,
                 YLIMS = c(0,40),
                 BAR.FILL = "firebrick3",
                 BAR.LINE = NA,
                 ASPECT.RATIO = 2) )
```



- **Figure 3 & 4**
- ***"Precursor and processing preferences determine piRNA sequence abundance."***
  - Before we can plot the results we have to re-process the data by:
    - Generate piRNA precursor index for **Rsubread** aligner
    - Export the piRNAs into new fasta file
    - Align the piRNA sequences to piRNA precursors
    - Load the new bam files back into R
    - Filter the data
    - Count and calculate the summaries
    - Select data for plotting
    - Plot the results for various analyses

```
# NOTE: Generate piRNA precursor index

## load library
library("Rsubread");library("GenomicRanges");
library("data.table"); library("ShortRead")

## Load precursors into GR
OSC.PRECURSORS.GR <- makeGRangesFromDataFrame(df = fread("/Users/genzorp/Documents/GITHUB/piDiversity/da
                            col.names = c("chr","start","end","name","rank","strand")),
                            keep.extra.columns = TRUE)
```

```r
## Extract genome sequence
OSC.PRECURSORS.SEQ <- getSeq(Dmelanogaster, OSC.PRECURSORS.GR)
names(OSC.PRECURSORS.SEQ) <- mcols(OSC.PRECURSORS.GR)[["name"]]

## Export fasta
GENOME_FA_DIR <- "/Users/genzorp/Documents/GITHUB/piDiversity/data/temp/genome_fa/"
writeFasta(object = OSC.PRECURSORS.SEQ, file = paste0(GENOME_FA_DIR,"osc_precursors.fa"))


## INDEX GENOME
GENOME_IDX_DIR <- "/Users/genzorp/Documents/GITHUB/piDiversity/data/temp/genome_rsub/"
OSC.PRECURSORS.IDX <- suppressMessages(
  buildindex(basename = paste0(GENOME_IDX_DIR,"osc_precursors"),
             reference = paste0(GENOME_FA_DIR,"osc_precursors.fa")) )
```

```r
# NOTE: Export the piRNAs into new fasta file

# load function
source(paste0(FUN.DIR,"makeFastaFromGR.R"))

# make fasta from the Total, Common, and Rare Genomic Ranges file
FA_DIR = "/Users/genzorp/Documents/GITHUB/piDiversity/data/temp/granges_fa/"
suppressMessages(
  GR.SEQ.L <- makeFastaFromGR(GRL = GRL,
                  BSSPECIES = "Dmelanogaster",
                  SAVE.TO.FILE = TRUE,
                  FA.DIR = FA_DIR,
                  MC.CORES = 3 ) )
```

```r
# NOTE: Align the piRNA sequences to piRNA precursors

# Provide path to newly made fasta files
GR.FASTA.L <- paste0(FA_DIR,grep("*all.fa$",list.files(FA_DIR), value = TRUE))
names(GR.FASTA.L) <- c(nth(tstrsplit(list.files(FA_DIR),split="_"),1))

# Set oath for alignment output
SUBREAD_OUTPUT <- "/Users/genzorp/Documents/GITHUB/piDiversity/data/temp/output_rsub/"

# Align files to genome with Rsubread
OSC.ALIGN.PRECURSORS <- lapply(names(GR.FASTA.L),function(s){
  suppressMessages(
    aRES <- suppressMessages(
      align(index = paste0(GENOME_IDX_DIR,"osc_precursors"),
            readfile1 = GR.FASTA.L[[s]],
            output_file = paste0(SUBREAD_OUTPUT,s,".bam"),
            sortReadsByCoordinates = TRUE,
            maxMismatches = 1,
            nthreads = 3,
            unique = FALSE,
            nBestLocations = 100) ) )
  return(aRES)})
```

```r
# NOTE: Load the new bam files back into R

# load function
source(paste0(FUN.DIR,"filterBam.R"))

# create named vector of paths
MAPPED.GR.PATHS <- paste0(SUBREAD_OUTPUT, grep("bam$",list.files(SUBREAD_OUTPUT), value = TRUE))
names(MAPPED.GR.PATHS) <- nth(tstrsplit(grep("bam$",list.files(SUBREAD_OUTPUT), value = TRUE), split="\

# load bam files into new object
MAPPED.BAM.L <- lapply(names(MAPPED.GR.PATHS), function(s){
  message(paste0("Processing: ",s))
  suppressMessages(
    filterBam(BAMFILE = MAPPED.GR.PATHS[[s]],
              BSSPECIES = "Dmelanogaster",
              EXTENTION = ".bam",
              STANDARD.CONTIGS.ONLY = FALSE,
              TAGS = c("NH","NM"),
              SPLIT.NAME.BY = "_")) })
```

```
## Processing: Common

##     INPUT SIZE_FILTER FLAG_FILTER MISMATCH_FILTER  FINAL
## 1: 896609         100         100        96.08748 861529

## Processing: Rare

##      INPUT SIZE_FILTER FLAG_FILTER MISMATCH_FILTER   FINAL
## 1: 1097066         100         100        97.21457 1066508

## Processing: Total

##      INPUT SIZE_FILTER FLAG_FILTER MISMATCH_FILTER   FINAL
## 1: 2720283         100         100        96.85165 2634639
```

```r
names(MAPPED.BAM.L) <- names(MAPPED.GR.PATHS)
```

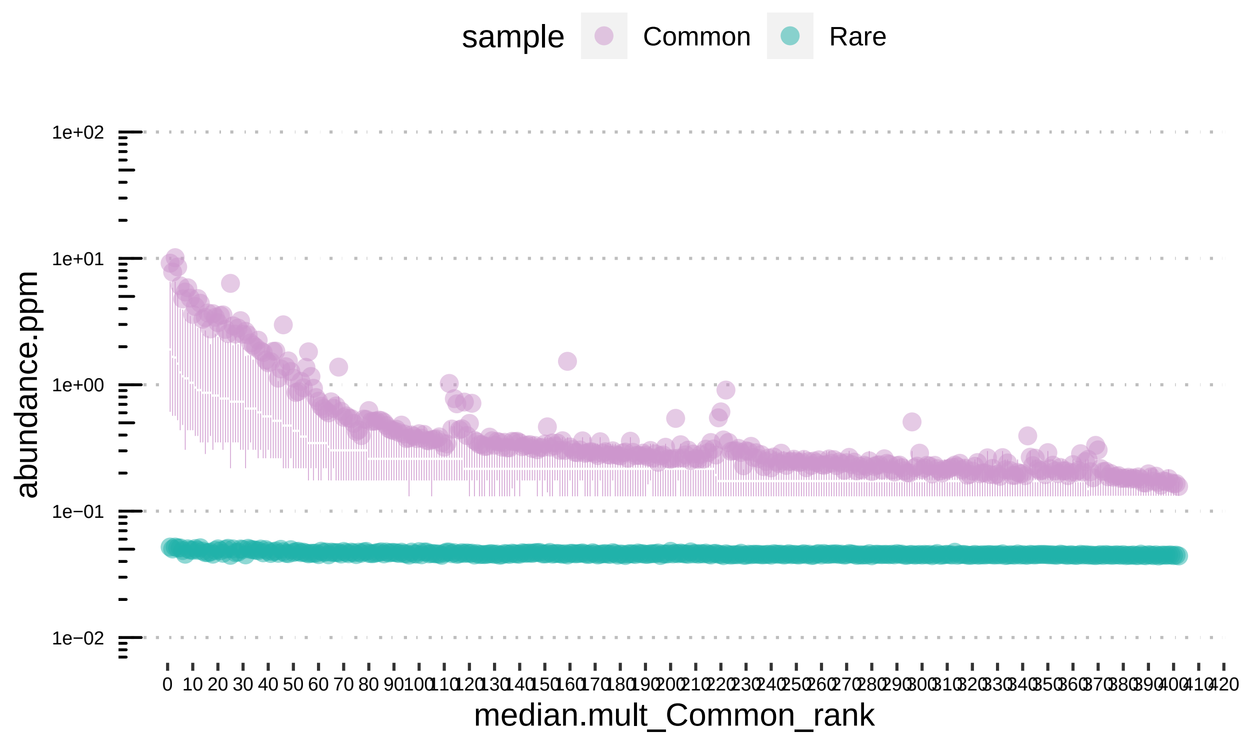
```r
# NOTE: Filter the data

# load function
source(paste0(FUN.DIR,"simpleGRFilter.R"))

# filter conditions
#  - only single mappers (NH=1)
#  - appropriate size range (24-29-nt long)
#  - originate from piRNA precursor (STRAND=YES)

MAPPED.FILTERED.BAM.L <- lapply(names(MAPPED.BAM.L),function(r){
  filtered.GR <- suppressMessages(
    simpleGRFilter(GR = MAPPED.BAM.L[[r]],
                   RANGE.NAME = r,
                   NH.TAG = 1,
                   SIZE.RANGE = c(24,29),
                   STRAND = "YES",
                   SEQNAMES.SPLIT = ":") )
  seqlevels(filtered.GR) <- seqlevelsInUse(filtered.GR)
  return(filtered.GR) })
names(MAPPED.FILTERED.BAM.L) <- names(MAPPED.BAM.L)
```

```
# NOTE: Count and calculate summaries

# load function
source(paste0(FUN.DIR,"simpleClusterSRCountsPerGR.R"))
source(paste0(FUN.DIR,"simpleClusterStatsPerGR.R"))

# Count sequences and reads
SR.COUNTS.DTL <- lapply(names(MAPPED.FILTERED.BAM.L), function(r){
  suppressMessages(
    simpleClusterSRCountsPerGR(GR = MAPPED.FILTERED.BAM.L[[r]],
                               RANGE.NAME = r,
                               CLUSTER.NAME.COLUMN = "seqnames")) })
names(SR.COUNTS.DTL) <- names(MAPPED.FILTERED.BAM.L)

# Calculate mean and median
R.STAT.DTL <- lapply(names(MAPPED.FILTERED.BAM.L), function(r){
  suppressMessages(
    simpleClusterStatsPerGR(GR = MAPPED.FILTERED.BAM.L[[r]],
                            RANGE.NAME = r,
                            CLUSTER.NAME.COLUMN = "seqnames")) })
names(R.STAT.DTL) <- names(MAPPED.FILTERED.BAM.L)

# Create piRNA precursor table
PRECURSOR.DT <- setDT(join_all(dfs = list(join_all(dfs = SR.COUNTS.DTL,
                                                   by = "CLUSTER",
                                                   type = "full"),
                                          join_all(dfs = R.STAT.DTL,
                                                   by = "CLUSTER",
                                                   type = "full")),
                               by = "CLUSTER", type = "full"))

# Filter: only regions with at least 100 seqeunces
SEQ.CUTOFF = 100
SEQ.CUTOFF.COLUMN <- grep("_Common",grep("seq.count",colnames(PRECURSOR.DT),
                                         value = TRUE), value = TRUE)
PRECURSOR.FILTERERD.DT <- PRECURSOR.DT[PRECURSOR.DT[[SEQ.CUTOFF.COLUMN]] >= SEQ.CUTOFF]

# Rank: based on the median common piRNA abundance
RANK.COLUMN <- grep("_Common",grep("median",colnames(PRECURSOR.DT),
                                   value = TRUE),value = TRUE)
setorderv(x = PRECURSOR.FILTERERD.DT, cols = RANK.COLUMN, order = -1)
PRECURSOR.FILTERERD.DT[,paste0(RANK.COLUMN,"_rank") := seq(1,nrow(PRECURSOR.FILTERERD.DT))]
PRECURSOR.FILTERERD.DT[["CLUSTER"]] <- droplevels(PRECURSOR.FILTERERD.DT[["CLUSTER"]])
```

```r
# NOTE: Select data for plotting

# Prepare data table needed for box plot
BOX.TABLE.L <- lapply(names(MAPPED.FILTERED.BAM.L), function(r){
  # a range & table
  a.GR <- MAPPED.FILTERED.BAM.L[[r]]
  a.DT <- as.data.table(a.GR)

  # subset
  a.DT <- a.DT[,c("seqnames", "MULT")]
  colnames(a.DT) <- c("CLUSTER","abundance")
  a.DT[["sample"]] <- r

  # return
  return(a.DT)})

# Calculations
PPM = 1000000
BP.DT <- rbindlist(l = BOX.TABLE.L)
BP.DT[,"total.reads" := sum(mcols(MAPPED.FILTERED.BAM.L[["Total"]])[["MULT"]])]
BP.DT[,"abundance.ppm" := (abundance / total.reads) * PPM]
BP.DT[,"median.abundance.ppm" := lapply(.SD, median),
      by = c("sample","CLUSTER"), .SDcols = c("abundance.ppm")]
BP.DT[,"mean.abundance.ppm" := lapply(.SD, mean),
      by = c("sample","CLUSTER"), .SDcols = c("abundance.ppm")]

# Extract ranked information only for piRNA precursors with sufficient data
BP <- BP.DT[CLUSTER %in% PRECURSOR.FILTERERD.DT[["CLUSTER"]]]
BPM <- setDT(join_all(
  dfs = list(BP,PRECURSOR.FILTERERD.DT[,.SD,
                                       .SDcols= c("CLUSTER",paste0(RANK.COLUMN,"_rank"))]),
                        by = "CLUSTER", type = "full"))
BPM[["CLUSTER"]] <- factor(BPM[["CLUSTER"]], levels = PRECURSOR.FILTERERD.DT[["CLUSTER"]])


# Subset for graphing
BPM.common <- BPM[sample %in% "Common"]
BPM.rare <- BPM[sample %in% "Rare"]
BP.mean <- BPM[, lapply(.SD, mean),
               by = c("CLUSTER",paste0(RANK.COLUMN,"_rank"),"sample"),
               .SDcols = "abundance.ppm"]
BP.mean <- BP.mean[sample %in% c("Common","Rare")]
```

```r
# NOTE: Plot the results - Figure 3A & Figure 4A

# plot settings
FAM = "Helvetica"; XYT = 12; TCOL = "black"

# plot
ggplot() + theme_pubclean() +
  ## common
  geom_boxplot(data = BPM.common, aes_string(x = paste0(RANK.COLUMN,"_rank"),
                                             y = "abundance.ppm",
                                     group = "CLUSTER", colour = "sample",
                                     fill = "sample"),
               outlier.shape = NA, coef = 0, lwd = 0.2,
               colour = "white", fill = "plum3") +
  ## rare
  geom_boxplot(data = BPM.rare, aes_string(x = paste0(RANK.COLUMN,"_rank"),
                                         y = "abundance.ppm",
                                   group = "CLUSTER", colour = "sample",
                                   fill = "sample"),
               outlier.shape = NA, coef = 0, lwd = 0.2,
               colour = "white", fill = "lightseagreen") +

  ## points
  geom_point(data = BP.mean,
             aes_string(x = paste0(RANK.COLUMN,"_rank"), y = "abundance.ppm",
                        colour = "sample"),
             shape = 16, size = 3, alpha = 0.5) +
  ## scales
  scale_y_continuous(trans = "log10", limits = c(0.01,100)) +
  scale_x_continuous(breaks = seq(0,500,10)) +
  scale_colour_manual(values = c("plum3", "lightseagreen")) +
  scale_fill_manual(values = c("plum3", "lightseagreen")) +

  ## theme
  annotation_logticks(sides = "l") +
  theme(aspect.ratio = 0.5,
        axis.ticks.y = element_blank(),
        axis.text = element_text(family = FAM, size = XYT-5, color = TCOL),
        axis.title = element_text(family = FAM, size = XYT, colour = TCOL))
```

```
## Warning: Removed 718 rows containing non-finite values (stat_boxplot).
```



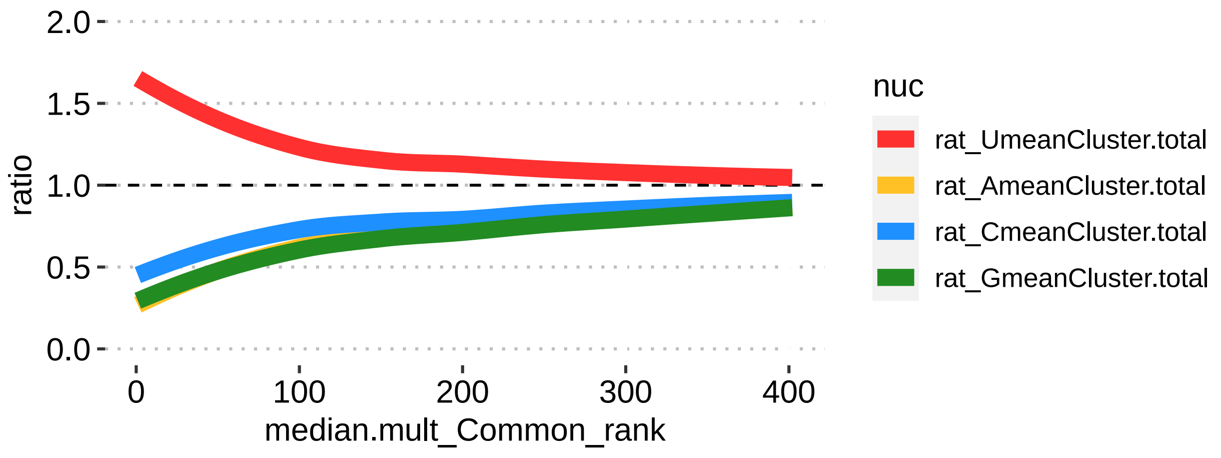
```r
# NOTE: Calculate first-nucleotide metrics

## load function
source(paste0(FUN.DIR,"simpleClusterFirstNucStatsPerGR.R"))

## extract first nucleotide count and calculate statistics
N1PC.DTL <- lapply(names(MAPPED.FILTERED.BAM.L),function(s){
  suppressMessages(
    simpleClusterFirstNucStatsPerGR(GR = MAPPED.FILTERED.BAM.L[[s]],
                                    RANGE.NAME = s,
                                    BSSPECIES = "Dmelanogaster",
                                    SOURCE.DIR = FUN.DIR)) })
names(N1PC.DTL) <- names(MAPPED.FILTERED.BAM.L)

## Combine all
N1DT <- setDT(join_all(dfs = N1PC.DTL, by = "CLUSTER", type = "full"))

## Add normalization to parts per million
PPM = 1000000
N1DT[,"mean.T.total.ppm" := (mean.T_Total/total.reads_Total) * PPM]
N1DT[,"mean.A.total.ppm" := (mean.A_Total/total.reads_Total) * PPM]
N1DT[,"mean.G.total.ppm" := (mean.G_Total/total.reads_Total) * PPM]
N1DT[,"mean.C.total.ppm" := (mean.C_Total/total.reads_Total) * PPM]
N1DT[,"mean.AGC.total.ppm" := (mean.AGC_Total/total.reads_Total) * PPM]
N1DT[,"median.T.total.ppm" := (median.T_Total/total.reads_Total) * PPM]
N1DT[,"median.A.total.ppm" := (median.A_Total/total.reads_Total) * PPM]
N1DT[,"median.G.total.ppm" := (median.G_Total/total.reads_Total) * PPM]
N1DT[,"median.C.total.ppm" := (median.C_Total/total.reads_Total) * PPM]
N1DT[,"median.AGC.total.ppm" := (median.AGC_Total/total.reads_Total) * PPM]
```

```r
## calculate ratios
N1DT[,"rat.UnonU.mean.total.raw" := mean.T_Total / mean.AGC_Total]
N1DT[,"rat_UmeanCluster.total" := mean.T_Total / cluster.mean_Total]
N1DT[,"rat_AmeanCluster.total" := mean.A_Total / cluster.mean_Total]
N1DT[,"rat_CmeanCluster.total" := mean.C_Total / cluster.mean_Total]
N1DT[,"rat_GmeanCluster.total" := mean.G_Total / cluster.mean_Total]
N1DT[,"rat.UnonU.median.total.raw" := median.T_Total / median.AGC_Total]
N1DT[,"rat_UmedianCluster.total" := median.T_Total / cluster.median_Total]
N1DT[,"rat_AmedianCluster.total" := median.A_Total / cluster.median_Total]
N1DT[,"rat_CmedianCluster.total" := median.C_Total / cluster.median_Total]
N1DT[,"rat_GmedianCluster.total" := median.G_Total / cluster.median_Total]

# Combine with ranked order from previous figure and setup for plotting
N1.TAB <- setDT(join_all(dfs = list(N1DT,
                                    BP.mean[sample %in% "Common",
                                            .SD,
                                            .SDcols=c("CLUSTER",paste0(RANK.COLUMN,"_rank") )]),
                         by = "CLUSTER", type = "full"))
N1.TAB <- na.omit(N1.TAB)
setorderv(N1.TAB, cols = c(paste0(RANK.COLUMN,"_rank")), order = 1)
DT.F3D <- N1.TAB[,.SD,
                 .SDcols=c("CLUSTER",
                           paste0(RANK.COLUMN,"_rank"),
                           grep("meanCluster",colnames(N1.TAB),value=T))]

DT.F3Dm <- melt.data.table(data = DT.F3D,
                           id.vars = c("CLUSTER",
                                       paste0(RANK.COLUMN,"_rank")),
                           variable.name = "nuc",
                           value.name = "ratio")
```

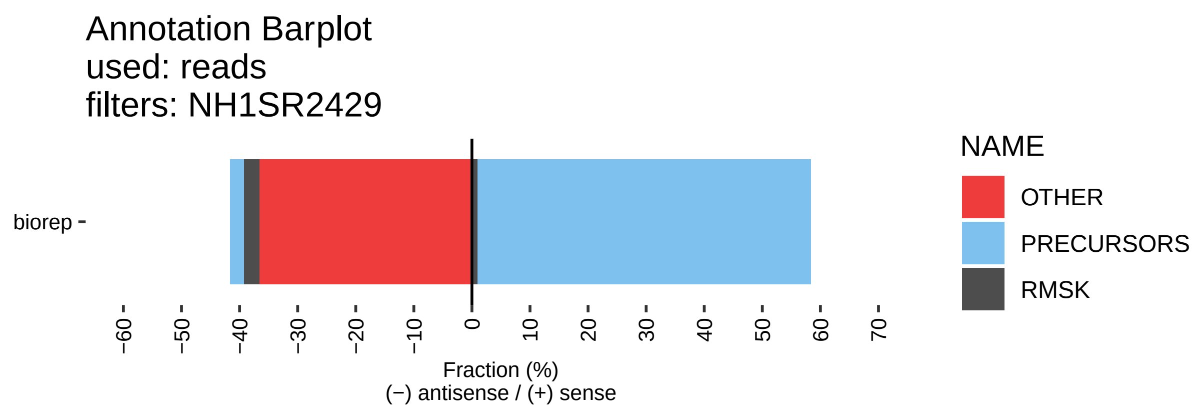
```r
# NOTE: Figure 3D & 4C plots

# Plot
ggplot() + theme_pubclean() +
  geom_hline(yintercept = 1, linetype = "dashed") +

  geom_smooth(data = DT.F3Dm, aes_string(x = paste0(RANK.COLUMN,"_rank"),
                                         y = "ratio",
                                         colour = "nuc"),
              se = FALSE, size = 3, alpha = 0.5) +
  scale_y_continuous(limits = c(0,2)) +
  scale_colour_manual(values = c("firebrick1","goldenrod1","dodgerblue1","forestgreen")) +
  theme(aspect.ratio = 0.5, legend.position = "right",
        axis.text = element_text(family = FAM, size = XYT, color = TCOL),
        axis.title = element_text(family = FAM, size = XYT, colour = TCOL))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

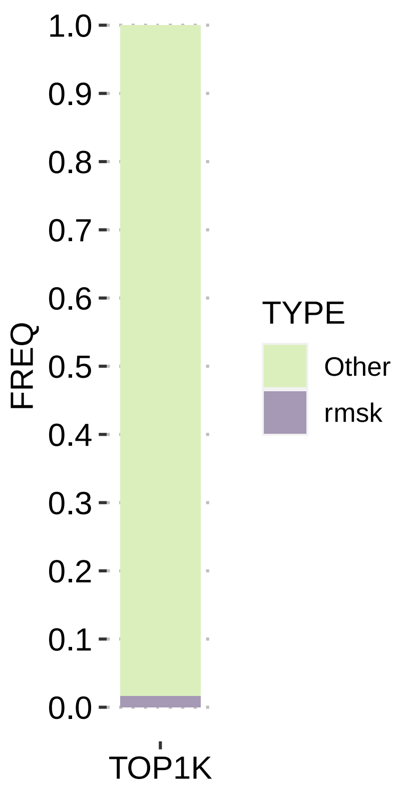


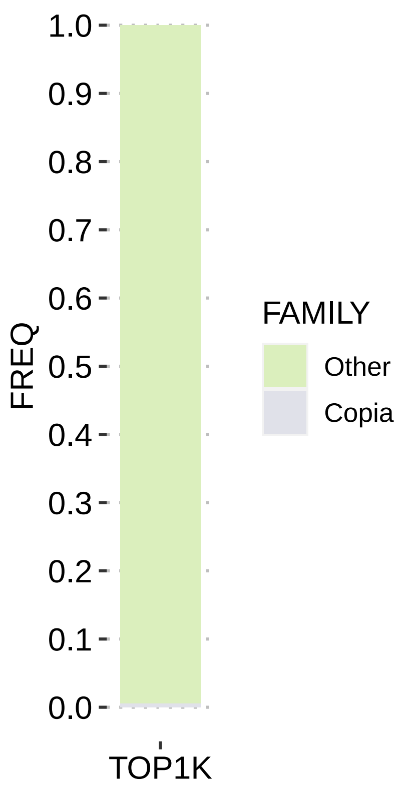
```
# NOTE: Annotating piRNAs by origin and target

# load function
source(paste0(FUN.DIR,"rmskGTF2BED.R"))
source(paste0(FUN.DIR,"annotateGRanked.R"))
source(paste0(FUN.DIR,"annotateRankedBP.R"))

# load rmsk annotations
RMSK.PATH= "/Users/genzorp/Documents/DATA/Annotation/Dmelanogaster/dm6_rmsk_TE.gtf"
RMSK.GR <- suppressMessages(
  rmskGTF2BED(RMSK.GTF = RMSK.PATH, RETURN.GR = TRUE))

# compare RMSK with piRNA precursors
#OSC.PRECURSORS.GR

# annotation in order CAT1 > CAT2 ... > CAT5
ANOT.DT.L <- suppressMessages(
  annotateGRanked(GR = BIO.REP.U.S249,
                  SAMPLE.NAME = "biorep",
                  NH.TAG = 1,
                  SIZE.RANGE = c(24,29),
                  CATEGORY.NAMES = c("PRECURSORS","RMSK"),
                  CATEGORY.1.GR = OSC.PRECURSORS.GR,
                  CATEGORY.2.GR = RMSK.GR,
                  SOURCE.DIR = FUN.DIR) )

# plot the distribution of annotation categories
annotateRankedBP(ANN.TAB.L = ANOT.DT.L,
                 COORD.FLIP = TRUE,
                 ASPECT.RATIO = 0.2,
                 Y.LIMS = c(-60,70))
```



```r
# NOTE: Figure 3A, Annotating transposable elements

suppressPackageStartupMessages({
  library("ggplot2"); library("ggbio"); library("ggpubr")})

# load function
source(paste0(FUN.DIR,"simpleGRfilter.R"))
source(paste0(FUN.DIR,"simpleAnnotateTE.R"))

# Use previously generated RMSK GR
# RMSK.GR

# Filter the data
BIO.REP.U.S249 <- suppressMessages(
  simpleGRFilter(GR = BIO.REP,
                 RANGE.NAME = "biorep1",
                 NH.TAG = 1,
                 SIZE.RANGE = c(24,29)))

## Order by abundance, calculate ppm and add order
BSSPECIES = "Dmelanogaster"
BIO.REP.U.S249.RANKED <- BIO.REP.U.S249[order(mcols(BIO.REP.U.S249)[["MULT"]],
                                              decreasing = TRUE)]
mcols(BIO.REP.U.S249.RANKED)[["PPM"]] <- (mcols(BIO.REP.U.S249.RANKED)[["MULT"]] /
                                            sum(mcols(BIO.REP.U.S249.RANKED)[["MULT"]]) )*1000000
mcols(BIO.REP.U.S249.RANKED)[["ORDER"]] <- seq(1,length(BIO.REP.U.S249.RANKED))


## TOP 1000 IN RMSK
TOP.IN.RMSK <- suppressMessages(
  simpleAnnotateTE(GR = BIO.REP.U.S249.RANKED[1:1000],
                   RMSK.GR = RMSK.GR,
                   OVERLAP.TYPE = "within") )

TOP.IN.RMSK.FAM <- TOP.IN.RMSK[["FAMILY"]]
TOP.IN.RMSK.FAM[["GROUP"]] = "TOP1K"
TOP.IN.RMSK.FAM[["TYPE"]] = ifelse(TOP.IN.RMSK.FAM[["FAMILY"]] %in% "OTHER","Other","rmsk")

# settings
TCOL="black"; FAM="Helvetica"; XYT=12
```

```
# plot
plot.TE <- ggplot() + theme_pubclean() +
  geom_bar(data = TOP.IN.RMSK.FAM,
           aes(x = GROUP, y = FREQ, fill = TYPE), stat = "identity")+
  scale_y_continuous(breaks = seq(0,1,0.1)) + xlab("") +
  scale_fill_manual(values = c("#DBEFBD","#A599B5")) +
  theme(aspect.ratio = 7, legend.position = "right",
        axis.text.x = element_text(family = FAM, size = XYT, color = TCOL),
        axis.text.y = element_text(family = FAM, size = XYT, color = TCOL))


## TOP 1000 in TE families
TOP.IN.RMSK.FAM.AS <- TOP.IN.RMSK.FAM[FREQ > 0 & STRAND %in% "AS"]
TOP.Other <- data.table(FAMILY="Other",STRAND="NA", COUNT=1000-sum(TOP.IN.RMSK.FAM.AS[["COUNT"]]),
                        FREQ=1-sum(TOP.IN.RMSK.FAM.AS[["FREQ"]]),GROUP="TOP1K",TYPE="rmsk")
TOP.IN.RMSK.FAM.AS <- rbindlist(l = list(TOP.IN.RMSK.FAM.AS,TOP.Other))
TOP.IN.RMSK.FAM.AS[["FAMILY"]] <- factor(TOP.IN.RMSK.FAM.AS[["FAMILY"]],
                                         levels = rev(TOP.IN.RMSK.FAM.AS[["FAMILY"]]))

## plot
plotTEFAM <- ggplot() + theme_pubclean() +
  geom_bar(data = TOP.IN.RMSK.FAM.AS,
           aes(x = GROUP, y = FREQ, fill = FAMILY), stat = "identity")+
  scale_y_continuous(breaks = seq(0,1,0.1)) +xlab("") +
  scale_fill_manual(values = c("#DBEFBD","#E0E1E9","#E7CEE3","#E4B4C2","#FF99BD")) +
  theme(aspect.ratio = 7, legend.position = "right",
        axis.text.x = element_text(family = FAM, size = XYT, color = TCOL),
        axis.text.y = element_text(family = FAM, size = XYT, color = TCOL))

## ARRANGE AND PRINT
arranged3Fplots <- ggarrange(plotlist = list(plot.TE,plotTEFAM),
                             ncol = 2)
arranged3Fplots
```





```
## NOTE: Figure 3F, flamenco specific mapping

# libraries
library("BSgenome.Dmelanogaster.UCSC.dm6"); library("GenomicAlignments")
library("GenomeInfoDb"); library("parallel"); library("dplyr");
library("BSgenome"); library("data.table")

# load functions
source(paste0(FUN.DIR,"simpleAnnotateTE.R"))
source(paste0(FUN.DIR,"rmskGTF2BED.R"))

# Start with raw bam files
# named vector of paths
BAM.PATH.L <- c("rep1"="/Users/genzorp/Documents/GITHUB/piDiversity/data/bamFiles/FHIP/STAR/FH-Piwi-IP_
                "rep2"="/Users/genzorp/Documents/GITHUB/piDiversity/data/bamFiles/FHIP/STAR/FH-Piwi-IP_
                "rep3"="/Users/genzorp/Documents/GITHUB/piDiversity/data/bamFiles/FHIP/STAR/FH-Piwi-IP_

# quick load settings
TAGS = c("NH","NM","MD"); WHAT = c("flag"); SIMPLE.CIGAR = TRUE; BSSPECIES = "Dmelanogaster"
PARAM = Rsamtools::ScanBamParam(flag = Rsamtools::scanBamFlag(isUnmappedQuery = FALSE),
                                tag = TAGS, simpleCigar = SIMPLE.CIGAR, what = WHAT)
ORIGINAL.FLAM <- GRanges(seqnames = "chrX", ranges=IRanges(start= 21631891, end = 21790731), strand = "

# Load and filter the data
# - only perfectly mapping to flam, 24-29-nt long
OFLAM.ALL.GRL <- mclapply(names(BAM.PATH.L),mc.cores = 3, function(s){
  message("Loading GA for ...")
  message(paste0("\t",s))
  GA <- GenomicAlignments::readGAlignments(file = BAM.PATH.L[[s]],
                                              use.names = TRUE, param = PARAM)
  message("\tfiltering")
  GA.FLAM <- subsetByOverlaps(x = GA, ranges = ORIGINAL.FLAM, type = "any")
  seqlevels(x = GA.FLAM, pruning.mode="coarse") <- c("chrX")
  GA.FLAM <- GA.FLAM[mcols(GA.FLAM)[["NM"]] %in% 0]
  GA.FLAM <- GA.FLAM[width(GA.FLAM) %in% seq(24,29)]

  message("\tformatting")
  FLAM.GA.GR <- GenomicRanges::granges(GA.FLAM, use.names = TRUE, use.mcols = TRUE)
  mcols(FLAM.GA.GR)[["MULT"]] <- dplyr::nth(tstrsplit(
    dplyr::nth(tstrsplit(names(FLAM.GA.GR), split = "-"), -1), split = "M"),-1)
  mcols(FLAM.GA.GR)[["SEQ"]] <- BSgenome::getSeq(eval(parse(text = BSSPECIES)), FLAM.GA.GR)

  message("\tcleaning")
  mcols(FLAM.GA.GR)[["flag"]] <- NULL
  mcols(FLAM.GA.GR)[["MD"]] <- NULL
  mcols(FLAM.GA.GR)[["NM"]] <- NULL
  return(FLAM.GA.GR) })

## combine into a single GR
OFLAM.ALL.GR <- do.call("c",OFLAM.ALL.GRL)
rm(OFLAM.ALL.GRL)

# save image
save.image(file = "/Users/genzorp/Documents/GITHUB/piDiversity/sessions/revisions/flamOnly_biorep_filte
```

```
## NOTE Figure 3F plot A

# libraries
library("ggplot2"); library("ggbio"); library("ggpubr");
library("BSgenome"); library("BSgenome.Dmelanogaster.UCSC.dm6")

## load the data
load("/Users/genzorp/Documents/GITHUB/piDiversity/sessions/revisions/flamOnly_biorep_filtered.RData")

# Sort and prepare biological replicate
BSSPECIES="Dmelanogaster"
GR.B <- BIO.REP[order(mcols(BIO.REP)[["MULT"]], decreasing = TRUE)]
mcols(GR.B)[["PPM"]] <- (mcols(GR.B)[["MULT"]] / sum(mcols(GR.B)[["MULT"]]) )*1000000
mcols(GR.B)[["ORDER"]] <- seq(1,length(GR.B))
mcols(GR.B)[["SEQ"]] <- BSgenome::getSeq(eval(parse(text = BSSPECIES)), GR.B)

# Convert GR to table and add info
OFLAM.ALL.DT <- as.data.table(OFLAM.ALL.GR)
mcols(OFLAM.ALL.GR)[["MAPPING"]] <- ifelse(mcols(OFLAM.ALL.GR)[["NH"]] %in% 1, "1","2")
mcols(OFLAM.ALL.GR)[["FIRST"]] <- ifelse(substr(mcols(OFLAM.ALL.GR)[["SEQ"]],
                                                start = 1, stop = 1) %in% "T", "1","2")
OFLAM.ALL.DT[["readNames"]] <- names(OFLAM.ALL.GR)

## add ppm to the sequences
B1000.INFLAM.GR <- OFLAM.ALL.GR[mcols(OFLAM.ALL.GR)[["SEQ"]] %in% mcols(GR.B[1:1000])[["SEQ"]]]
B1000.INFLAM.DT <- as.data.table(B1000.INFLAM.GR)
B1000.DT <- as.data.table(GR.B[1:1000])
```

```r
B1000.INFLAM.DT <-setDT(join_all(dfs = list(B1000.INFLAM.DT, B1000.DT[,c("SEQ","ORDER","PPM")]),
                                 by = "SEQ", type = "left"))
B1000.INFLAM.GR <- makeGRangesFromDataFrame(df = B1000.INFLAM.DT, keep.extra.columns = TRUE)

## extract only unique sequences
B1000.INFLAM.U.GR <- unique(B1000.INFLAM.GR)
B1000.INFLAM.U.DT <- as.data.table(B1000.INFLAM.U.GR)
B1000.INFLAM.U.DT[["FIRST"]] <- factor(B1000.INFLAM.U.DT[["FIRST"]], levels = c(2,1))

# flamenco start and end
FST = 21630000; FEN = 21800000

## plot
ggplot() + theme_pubclean() +
  geom_point(data = B1000.INFLAM.U.DT[FIRST %in% 1], aes(x = start, y = PPM),
             shape = 16, size = 5, alpha = 0.5, color = "#CD1F13") +
  geom_point(data = B1000.INFLAM.U.DT[FIRST %in% 2], aes(x = start, y = PPM),
             shape = 16, size = 5, alpha = 0.5, color = "black") +
  scale_y_continuous(trans = "log10", limits = c(10,1000)) +
  scale_x_continuous(limits = c(FST, FEN), breaks = seq(FST,FEN,50000)) +
  ggtitle(paste0("flamenco region\n","chrX:",FST,"-",FEN)) +
  annotation_logticks(sides = "l") +
  xlab("genome position") +
  ylab("Seq. abundance (ppm)\nTop 1000 most abundant piRNAs") +
  annotate(geom = "text", x = FST, y = 12, hjust = 0,
           label="RED= 1U piRNAs, BLACK= non-1U piRNAs") +
  theme(aspect.ratio = 0.4,
        axis.ticks.y = element_blank())
```

```r
# NOTE: Figure 3F: Mappability

# select dataset
AllMap <- GR.B
UnqMap <- AllMap[AllMap$NH == 1]

# select cluster, (original flam)
TheCluster <- GRanges(seqnames = "chrX", ranges=IRanges(start= 21631891, end = 21790731), strand = "+")
mcols(TheCluster)[["WIDTH"]] <- end(TheCluster)-start(TheCluster)

#make shortcut objects
ReadsInClust <- IRanges::subsetByOverlaps(UnqMap, TheCluster, type = "within")
mcols(TheCluster)[["DENSITY"]] <- sum(as.data.table(ReadsInClust)$MULT)/mcols(TheCluster)[["WIDTH"]]
ReadsInClust <- IRanges::subsetByOverlaps(UnqMap, TheCluster, type = "within")
RangeVect <- c(as.data.table(TheCluster)$start:as.data.table(TheCluster)$end)
TheChromo <- as.vector(as.data.table(TheCluster)$seqnames)
TheStrand <- as.vector(as.data.table(TheCluster)$strand)
RefGen <- BSgenome::getBSgenome("BSgenome.Dmelanogaster.UCSC.dm6")

#size distribution
ReadDistrib <- as.data.table(GR.B)[,sum(MULT), by=width]
ReadDistrib[,V1 := V1/sum(ReadDistrib$V1)]

# Create dm6 genome index using Rsubread
# - refer to manual for instructions

# create all possible piRNAs
NtoT <- lapply(c(18:32), function(a){
  vect <- split(RangeVect, ceiling(seq_along(RangeVect)/a))
  vect <- Filter(function(z){length(z) == a},vect)
  vectDT <- data.table(start = unlist(lapply(vect, min)), end = unlist(lapply(vect, max)))
  manyvect <- lapply(c(0:(a-1)), function(b){
    return(data.table(start = (vectDT$start+b), end = (vectDT$end+b)))
  })
  vectF <- bind_rows(manyvect)
  vectF <- vectF[end <= max(RangeVect)]
  GR1 <- GRanges(seqnames = TheChromo, ranges=IRanges(start=vectF$start, end = vectF$end),
                 strand = TheStrand)
  return(as.data.table(GR1)) })

NtoT <- bind_rows(NtoT)
NtoT <- GRanges(seqnames = NtoT$seqnames, ranges=IRanges(start=NtoT$start, end = NtoT$end),
                strand = NtoT$strand)
names(NtoT) <- paste0("Num_",c(1:length(NtoT)))

# extract sequences for each
NtoT_seq <- BSgenome::getSeq(eval(parse(text = BSSPECIES)), NtoT)

# write to file
# writeFasta(NtoT_seq, file="/YourDirectories/fastaFiles/files/oflam_all.fa")
#

# align to genome using Rsubread
# align(index="/YourDirectories/Dm6",
#       readfile1 =  "/YourDirectories/oflam_all.fa",
#       output_file = "/YourDirectories/Rsubread/aligned/oflam_all.bam",
#       maxMismatches = 0, nthreads = 6, sortReadsByCoordinates = TRUE)

# load and identify mult-mapping reads
GA <- readGAlignments(file = "/YourDirectories/Rsubread/aligned/oflam_all.bam",
                      use.names = TRUE, param = ScanBamParam(tag = c("NH"), what = c("flag")))
GA.DT <- as.data.table(GA)
GA.DT$Name <- names(GA)
mcols(NtoT)[["MULTI"]] <- 1
mcols(NtoT[names(NtoT) %in% GA.DT[flag == 0][seqnames == "chrX"]
           [start %in% RangeVect][end %in% RangeVect][,Name]])[["MULTI"]] <- 0

# scale 'un-mappability' per length
Hmmm <- lapply(c(18:32),function(a){

  if(a %in% width(NtoT[NtoT$MULTI == 1])){
    GR <- NtoT[width(NtoT) == a & NtoT$MULTI == 1]
    GR.DT <- as.data.table(GR)
    GR.DT$Names <- names(GR)
    min(GR.DT$start)
    max(GR.DT$end)
    covDT <- as.data.table(coverage(GR))[!1:(min(RangeVect)-1),3]
```

```r
    covDT[,value := (value/a)*ReadDistrib[width == a][,V1]]
    covDT <- rbind(covDT, data.table(
      value = rep(0, times=((max(RangeVect)-min(RangeVect)+1)-nrow(covDT)))))
    return(covDT)
  }
  else{return(data.table(value = rep(0,(max(RangeVect)-min(RangeVect)+1))))} })

# sum up scaled un-mappabilities by length
Hmmm <- bind_cols(Hmmm)
colnames(Hmmm) <- paste0(c(18:32),"nt")
Hmmm[,Coverage := matrixStats::rowSums2(as.matrix(Hmmm))]
Hmmm[,Position := c(1:nrow(Hmmm))]

# get coverage and make final table
RIC.DT <- as.data.table(ReadsInClust)
RIC.DT <- RIC.DT[rep(seq.int(1,nrow(RIC.DT)), RIC.DT$MULT)]
RIC <- GRanges(seqnames = RIC.DT$seqnames,
               ranges = IRanges(start=RIC.DT$start, end = RIC.DT$end),
               strand = RIC.DT$strand)

as.data.table(coverage(RIC)[[]])  [!1:(min(RangeVect)-1)]

FinalCov <- as.data.table(coverage(RIC)[[6]])[!1:(min(RangeVect)-1)]
FinalCov$Position <- c(start(TheCluster):max(end(RIC)) )-(start(TheCluster)-1)
FinalCov <- merge(FinalCov,Hmmm[,c("Coverage","Position")],  all.y = T, by="Position")
FinalCov$Position <- c(start(TheCluster):end(TheCluster))
colnames(FinalCov) <- c("POSITION", "COVERAGE", "MAP")

# take moving average (we used 2001)
library(accelerometry)
FinalCov$MAP.TR <- movingaves(c(rep(0,1000), FinalCov$MAP, rep(0,1000)), 2001)
FinalCov$Cov.TR <- movingaves(c(rep(0,1000), FinalCov$COVERAGE, rep(0,1000)), 2001)
FinalCov$MultiCovTR <- movingaves(c(rep(0,1000), FinalCov$MultiCov, rep(0,1000)), 2001)

TRANS = "log10"
YLIMS.1 = c(10,1000)

ggplot() +
  theme_bw() +
  geom_segment(data = FinalCov, aes(x=POSITION, xend=POSITION,
                                    y= YLIMS.1[1] ,yend = YLIMS.1[2],
                                    color = MAP.TR)) +
  ## SCALES
  scale_color_gradient(low = "white",high = "grey80") +
  scale_y_continuous(trans = TRANS, limits= YLIMS.1) +
  scale_x_continuous(breaks = seq(21631000,21800000,10000)) +
  annotation_logticks(sides = "l") +

  ggtitle(paste0("ORIGINAL FLAM  ", "chrX:",
                 min(FinalCov[["POSITION"]]), " - ",
                 max(FinalCov[["POSITION"]]))) +
  xlab("Genomic position") + ylab("Seq. abundance (ppm)") +
  ## THEME
  theme(aspect.ratio = 0.3,
        legend.position = "bottom",
        panel.border = element_blank(),
        panel.grid = element_blank())
```

Additional Figures * ***Analysis of the piRNA sensors***

```
# NOTE: piRNA Sensors - prepare data

# Sensor information
# Flam_1st: chrX:21,635,292-21,635,521:+ (230bp)
# Flam_2nd: chrX:21,635,062-21,635,291:+ (230bp)
# Flam_100bp: chrX:21,635,292-21,635,391:+ (100bp)
# Flam_460bp: chrX:21,635,062-21,635,521:+ (460bp)
# CG17514: chr3L:27,970,220-27,970,449:- (230bp)
# Tj: chr2L:19,466,747-19,466,976:+ (230bp)
# Path:  chr3L:9,495,281-9,495,510:- (230bp)
# CycB3:  chr3R:24,868,353-24,868,582:-  (230bp)
# CycB3_5:  chr3R:24,870,599-24,870,828:- (230bp)

# Genomic Ranges components
CHR = c("chrX","chrX","chrX","chrX","chr3L","chr2L","chr3L","chr3R","chr3R")
START = c(21635292,21635062,21635292,21635062,27970220,19466747,9495281,24868353,24870599)
END = c(21635521,21635291,21635391,21635521,27970449,19466976,9495510,24868582,24870828)
STRAND = c("+","+","+","+","-","+","-","-","-")
NAME = c("Flam230a","Flam230b","Flam100","Flam460","CG17514","Tj","Path","CycB3","CycB35")
SI <- keepStandardChromosomes(seqinfo(eval(parse(text = "Dmelanogaster"))))

# Genomic ranges
SEN.GR <- GRanges(seqnames = CHR,
                  ranges = IRanges(start = START, end = END),
                  strand = STRAND,
                  seqinfo = SI)
mcols(SEN.GR)[["name"]] <- paste0(CHR,":",START,"-",END,":",STRAND,":",NAME)
SEN.GR

# Get sequence
SEN.SEQ <- getSeq(eval(parse(text = "Dmelanogaster")), SEN.GR)
names(SEN.SEQ) <- mcols(SEN.GR)[["name"]]

# Write individual fasta files
lapply(names(SEN.SEQ), function(n){
  A.SEQ <- SEN.SEQ[n]
  N <- gsub(":","_",n)
  writeFasta(object = A.SEQ, file = paste0(FA.DIR,N,".fa")) })

## In the bash on the server

##
##  1. Make genomes from each fasta for your aligner
##  2. Align piRNAs to each sensor and
##  3. Proced with bam files
```

```r
##

# find the new bam files
# BAM.PATH = "/Users/genzorp/Documents/GITHUB/piDiversity/Sensors/stranded_STAR/"
EXTENTION = ".Aligned.sortedByCoord.out.bam$"
FILE.NAMES = grep(EXTENTION, list.files(BAM.PATH), value = TRUE)
SAMPLE.NAMES = gsub(EXTENTION, "", FILE.NAMES)
SENSOR.NAMES <- nth(tstrsplit(SAMPLE.NAMES, split = "_"),4)
BAM.FILE.PATHS <- paste0(BAM.PATH,FILE.NAMES)
names(BAM.FILE.PATHS) <- SENSOR.NAMES
BAM.FILE.PATHS

# Load the bam files
source(paste0(FUN.DIR,"filterBam.R"))
SENSOR.BAM.L <- lapply(names(BAM.FILE.PATHS), function(s){

  message(paste0("Loading: ",s))
  SGR <- suppressMessages(filterBam(BAMFILE = BAM.FILE.PATHS[[s]],
                  BSSPECIES = "Dmelanogaster",
                  PERFECT.MATCH.ONLY = TRUE,
                  READ.SIZE.RANGE = c(24,29),
                  STANDARD.CONTIGS.ONLY = FALSE,
                  GET.ORIGINAL.SEQUENCE = TRUE,
                  SPLIT.NAME.BY = "_"))

  ## Filter by actual strand of piRNA
  mcols(SGR)[["rs"]] <- nth(tstrsplit(names(SGR),split = "_"),4)
  mcols(SGR)[["gs"]] <- nth(tstrsplit(seqnames(SGR),split = ":"),3)
  message(length(SGR))
  SGR <- SGR[mcols(SGR)[["rs"]] == mcols(SGR)[["gs"]],]
  message(length(SGR))

  ## Return
  return(SGR)})
names(SENSOR.BAM.L) <- names(BAM.FILE.PATHS)

# save image
save.image(file = "/Users/genzorp/Documents/GITHUB/piDiversity/sessions/revisions/sensor_mapped.RData")
```

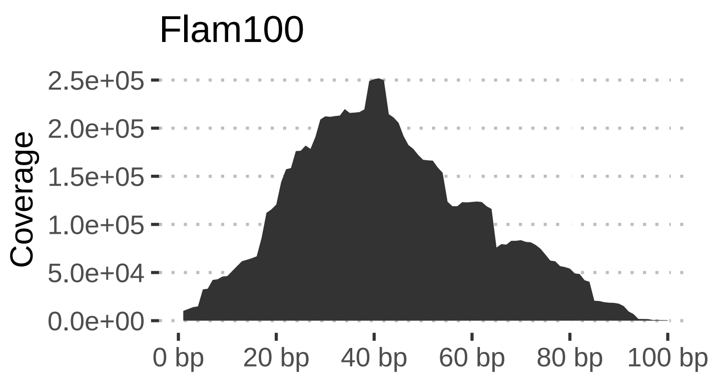
```r
# NOTE: piRNA SENSORS - analyze data

# libraries
suppressPackageStartupMessages({
  library(ggfortify); library(scales);
  library(ggbio); library(data.table)})

# load prepared data
load(file = "/Users/genzorp/Documents/GITHUB/LIVE/piRNA_Diversity/exampleRData/sensor_mapped.RData")

# Expand mapped data into reads
READS.BAM.L <- suppressWarnings(suppressMessages(
  lapply(names(SENSOR.BAM.L), function(s){
    message(paste0("Expanding to reads: ",s))
    A.GR <- SENSOR.BAM.L[[s]]
    DT.GR <- as.data.table(A.GR, keep.rownames = "aseq")
    setkey(DT.GR, cols = "aseq")
    DT.LONG <- DT.GR[rep(aseq, MULT)]
    R.GR <- makeGRangesFromDataFrame(DT.LONG, keep.extra.columns = TRUE)
    return(R.GR)}) ))
names(READS.BAM.L) <- names(SENSOR.BAM.L)

# plot piRNA read coverage
Flam100.COV <- ggbio::autoplot(READS.BAM.L[["Flam100"]], stat = "coverage", main="Flam100")
```

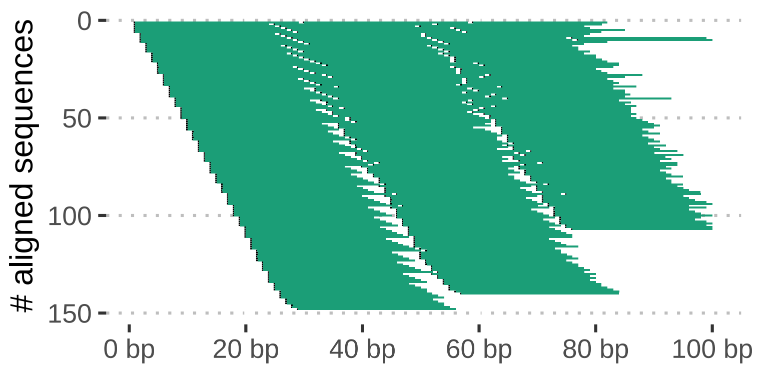
```
## Scale for 'x' is already present. Adding another scale for 'x', which will
## replace the existing scale.
```

```r
Flam100.COV <- Flam100.COV + theme_pubclean() +
  scale_y_continuous(labels = scientific_format()) +
  theme(aspect.ratio = 0.5)

# plot mapped piRNA sequences
Flam100.SEQ <-  ggbio::autoplot(object = SENSOR.BAM.L[["Flam100"]], aes(fill = strand))
Flam100.SEQ <- Flam100.SEQ + scale_y_reverse() + theme_pubclean() +
  ylab("# aligned sequences") +
  theme(aspect.ratio = 0.5, legend.position = "none")
```

```
## Scale for 'y' is already present. Adding another scale for 'y', which will
## replace the existing scale.
```

```r
# Arrange
SENSOR.MAPPING.PLOT <- ggarrange(plotlist = list(Flam100.COV@ggplot,
                                                 Flam100.SEQ@ggplot), nrow = 2)

SENSOR.MAPPING.PLOT
```





bash Scripts

These scripts were used to count the total number of calibrators in publicly available small RNA libraries in order to estimate correction factor (kcorr) **PMID:30193099**, and for quantification of miRNA and piRNA populations in data generated in this paper.

```
# NOTE: Bash script used to count calibrator sequences in public small RNA libraries
#       for calculation of correction factor (kcorr)
#       - mouse male germ cell data from PMID: 30193099

#!/bin/bash

# ABOUT
# Find the number of calibrator sequences by looking for their full sequence
# without mismatches

# inputs
input=$1
name=$(basename ${input%.fastq.gz})

echo "Counting the Calibrators"
printf "Calibrator\nA_CalibratorSequence" > "${name}"_Calibrator_seq.txt

echo "Counts" > "${name}"_Calibrator_Count.txt
zgrep 'CalibratorSequence' "${name}".fastq.gz | wc -l >> "${name}"_Calibrator_Count.txt

paste "${name}"_Calibrator_seq.txt \
      "${name}"_Calibrator_Count.txt > "${name}"_Calibrator_Counts.txt
```

```
# remove intermediate files
rm "${name}"_Calibrator_seq.txt
rm "${name}"_Calibrator_Count.txt

echo "Done"
```

```
# NOTE: Bash script used to count calibrator sequences in small RNA libraries
#       generated for this paper in order to quantify piRNA and miRNA populations.

#!/bin/bash
module load cutadapt
module load samtools
module load bowtie/1.2.3
module load fastqc

# Provide fastq file as first argument to this bash script
input=$1
name=$(basename ${input%.fastq})
mkdir "${name}"_Analysis

# Remove 3-prime adapter sequence, keep > 29-nt
echo "removing adaptor"
cutadapt  -a TGACTGTGGAATTCTCGGGTGCCAAGG \
          --overlap 10 \
          -m 29 \
          --untrimmed-output "${name}"_Untrimmed.fastq \
          -o "${name}"_trimmed.fastq "${name}".fastq \
          > "${name}"_Remove3Adaptor_Report.txt

mv "${name}"_Untrimmed.fastq "${name}"_Analysis
mv "${name}"_trimmed.fastq "${name}"_Analysis
mv "${name}"_Remove3Adaptor_Report.txt "${name}"_Analysis
cd "${name}"_Analysis

# Collapse by sequence / remove PCR duplicates
echo "removing PCR duplicates"
cat "${name}"_trimmed.fastq | awk 'NR%4==2' | sort | uniq -c | \
    awk '{OFS= "\n"; print ">"NR"-"$1,$2}' > "${name}"_trimmed_collapsed.fasta

# Remove UMIs (8N at 5'end & 2N at 3'end)
echo "removing UMIs"
cutadapt  -u 8 -u -2 \
          -o "${name}"_trimmed_collapsed_UMIremoved.fasta \
          "${name}"_trimmed_collapsed.fasta > \
          "${name}"_RemoveUMI_Report.txt

# Remove structural RNAs
echo "Align reads to structural RNAs"
bowtie  /YourStructuralIndexFolder/dm6_strRNAs_Bowtie1index \
        -v 1 -m 100 -p 12 --best --strata \
        --un NonStructural_"${name}"_trimmed_collapsed_UMIremoved.fasta \
        -f "${name}"_trimmed_collapsed_UMIremoved.fasta \
        -S 2> "${name}"_AlignToStractural_Report.txt | \
        samtools view -Sh > Structural_"${name}"_trimmed_collapsed_UMIremoved.sam

# Find the numbers of calibrators looking for their 25 first nt
echo "Counting the Calibrators"
printf  "Calibrator\n1_CALIBRATORSEQUENCEONE\n2_CALIBRATORSEQUENCETWO" > \
        "${name}"_Calibrator_seq.txt
echo Counts > "${name}"_Calibrator_Count.txt

grep \
  'CALIBRATORSEQUENCEONE' NonStructural_"${name}"_trimmed_collapsed_UMIremoved.fasta | \
  wc >> "${name}"_Calibrator_Count.txt
grep \
  'CALIBRATORSEQUENCETWO' NonStructural_"${name}"_trimmed_collapsed_UMIremoved.fasta | \
  wc >> "${name}"_Calibrator_Count.txt

paste "${name}"_Calibrator_seq.txt "${name}"_Calibrator_Count.txt > \
      "${name}"_Calibrator_Counts.txt

rm "${name}"_Calibrator_seq.txt
rm "${name}"_Calibrator_Count.txt

# Remove calibrators with cutadapt - use a first 25-nt of each calibrator
echo "Removing the Calibrators"
cutadapt  \
  -g file:/FolderWithCalibratorFasta/Calibrators_first25nt.fasta \
  --overlap 25 -m 20 -e 0 \
  -o NonStructural_"${name}"_trimmed_collapsed_UMI_noCalibrators.fasta \
  NonStructural_"${name}"_trimmed_collapsed_UMIremoved.fasta > \
  "${name}"_RemoveCalibrators_Report.txt

# Identify miRNAs
echo "Mapping against miRNAs"
bowtie \
  /YourMiRNAIndexFolder/dm_mature_miRNAs_Bowtie1index \
  -v 1 -m 100 -p 12 --best --strata \
  --al NonStructural_"${name}"_trimmed_collapsed_UMI_noCalibrators_miRNAs.fasta \
  --un NonStructural_"${name}"_trimmed_collapsed_UMI_noCalibrators_miRNAout.fasta \
  -f NonStructural_"${name}"_trimmed_collapsed_UMI_noCalibrators.fasta \
  -S 2> "${name}"_AlignTomiRNA_Report.txt | \
  samtools view -Sh > \
  Mapped_NonStructural_"${name}"_trimmed_collapsed_UMI_noCalibrators_miR_Aligned.sam

# Filter to keep 24-29-nt long reads only
echo "Selecting for 24-29nt long reads"
cutadapt  \
  -m 24 -M 29 \
  -o NonStructural_"${name}"_trimmed_collapsed_UMI_noCalibrators_miRNAout_24to29.fasta \
  NonStructural_"${name}"_trimmed_collapsed_UMI_noCalibrators_miRNAout.fasta > \
  "${name}"_keep24to29nt_Report.txt

# Map to Dm6 genome
echo "Aligning to Dm6 genome"
bowtie  \
  /YourGenomeIndexFolder/dm6_Bowtie1index \
  -v 1 -m 100 -p 12 --best --strata \
  --al Mapped_NonStructural_"${name}"_trimmed_collapsed_UMI_noCalibrators_miRNAout_24to29.fasta \
  --un Unmapped_NonStructural_"${name}"_trimmed_collapsed_UMI_noCalibrators_miRNAout_24to29.fasta \
  -f NonStructural_"${name}"_trimmed_collapsed_UMI_noCalibrators_miRNAout_24to29.fasta \
  -S 2> "${name}"_AlignToDm6_Report.txt | \
  samtools view -Sh > \
  NonStructural_"${name}"_trimmed_collapsed_UMI_noCalibrators_miRNAout_24to29_Dm6Aligned.sam
```