

# An Introduction to *Guitar* Package

Xiao Du

Modified: 26 April, 2019. Compiled: November 11, 2019

## 1 Quick Start with Guitar

This is a manual for Guitar package. The Guitar package is aimed for RNA landmark-guided transcriptomic analysis of RNA-related genomic features.

The Guitar package enables the comparison of multiple genomic features, which need to be stored in a name list. Please see the following example, which reads 1000 RNA m6A methylation sites into R for detection. Of course, in actual data analysis, features may come from multiple sets of resources.

```
library(Guitar)

## Loading required package: GenomicFeatures
## Loading required package: BiocGenerics
## Loading required package: parallel
##
## Attaching package: 'BiocGenerics'
## The following objects are masked from 'package:parallel':
##
##   clusterApply, clusterApplyLB, clusterCall,
##   clusterEvalQ, clusterExport, clusterMap,
##   parApply, parCapply, parLapply, parLapplyLB,
##   parRapply, parSapply, parSapplyLB
## The following objects are masked from 'package:stats':
##
##   IQR, mad, sd, var, xtabs
## The following objects are masked from 'package:base':
##
##   Filter, Find, Map, Position, Reduce,
##   anyDuplicated, append, as.data.frame, basename,
##   cbind, colnames, dirname, do.call, duplicated,
##   eval, evalq, get, grep, grepl, intersect,
##   is.unsorted, lapply, mapply, match, mget, order,
##   paste, pmax, pmax.int, pmin, pmin.int, rank,
##   rbind, rownames, sapply, setdiff, sort, table,
##   tapply, union, unique, unsplit, which, which.max,
##   which.min
## Loading required package: S4Vectors
## Loading required package: stats4
##
## Attaching package: 'S4Vectors'
## The following object is masked from 'package:base':
##
##   expand.grid
```

```

## Loading required package: IRanges
##
## Attaching package: 'IRanges'
## The following object is masked from 'package:grDevices':
##
##     windows
## Loading required package: GenomeInfoDb
## Loading required package: GenomicRanges
## Loading required package: AnnotationDbi
## Loading required package: Biobase
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view
##     with 'browseVignettes()'. To cite Bioconductor,
##     see 'citation("Biobase")', and for packages
##     'citation("pkgname)".
## Loading required package: rtracklayer
## Loading required package: magrittr
## Loading required package: ggplot2
## Loading required package: dplyr
##
## Attaching package: 'dplyr'
## The following object is masked from 'package:AnnotationDbi':
##
##     select
## The following object is masked from 'package:Biobase':
##
##     combine
## The following objects are masked from 'package:GenomicRanges':
##
##     intersect, setdiff, union
## The following object is masked from 'package:GenomeInfoDb':
##
##     intersect
## The following objects are masked from 'package:IRanges':
##
##     collapse, desc, intersect, setdiff, slice, union
## The following objects are masked from 'package:S4Vectors':
##
##     first, intersect, rename, setdiff, setequal,
##     union
## The following objects are masked from 'package:BiocGenerics':
##
##     combine, intersect, setdiff, union
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
##
## Attaching package: 'Guitar'

```

```
## The following object is masked from 'package:BiocGenerics':
##
##      normalize

# genomic features imported into named list
stBedFiles <- list(system.file("extdata", "m6A_mm10_exomePeak_1000peaks_bed12.bed",
                             package="Guitar"))
```

With the following script, we may generate the transcriptomic distribution of genomic features to be tested, and the result will be automatically saved into a PDF file under the working directory with prefix "example". With the `GuitarPlot` function, the gene annotation can be downloaded from internet automatically with a genome assembly number provided; however, this feature requires working internet and might take a longer time. The toy `Guitar` coordinates generated internally should never be re-used in other real data analysis.

```
count <- GuitarPlot(txGenomeVer = "mm10",
                   stBedFiles = stBedFiles,
                   miscOutFilePrefix = NA)
```

In a more efficient protocol, in order to re-use the gene annotation and *Guitar coordinates*, you will have to build `Guitar Coordinates` from a `txdb` object in a separate step. The `transcriptDb` contains the gene annotation information and can be obtained in a number of ways, e.g, download the complete gene annotation of species from UCSC automatically, which might takes a few minutes. In the following analysis, we load the `Txdb` object from a toy dataset provided with the `Guitar` package. Please note that this is only a very small part of the complete hg19 transcriptome, and the `Txdb` object provided with `Guitar` package should not be used in real data analysis. With a `Txdb` object that contains gene annotation information, we in the next build *Guitar coordinates*, which is essentially a bridge connects the transcriptomic landmarks and genomic coordinates.

```
txdb_file <- system.file("extdata", "mm10_toy.sqlite",
                        package="Guitar")
txdb <- loadDb(txdb_file)
guitarTxdb <- makeGuitarTxdb(txdb = txdb, txPrimaryOnly = FALSE)

## [1] "There are 2946 transcripts of 2946 genes in the genome."
## [1] "total 2946 transcripts extracted ..."
## [1] "total 2719 transcripts left after ambiguity filter ..."
## [1] "total 2719 transcripts left after check chromosome validity ..."
## [1] "total 1342 mRNAs left after component length filter ..."
## [1] "total 307 ncRNAs left after ncRNA length filter ..."
## [1] "generate components for all tx"
## [1] "generate components for mRNA"
## [1] "generate components for lncRNA"
## [1] "generate chiped transcriptome"
## [1] "generate coverage checking ranges for tx"
## [1] "generate coverage checking ranges for mrna"
## [1] "generate coverage checking ranges for ncRNA"

# Or use gff. file to generate guitarTxdb
# Or use getTxdb() to download TxDb from internet:
# txdb <- getTxdb(txGenomeVer="hg19")
# guitarTxdb <- makeGuitarTxdb(txdb)
```

You may now generate the `Guitar` plot from the named list of genome-based features.

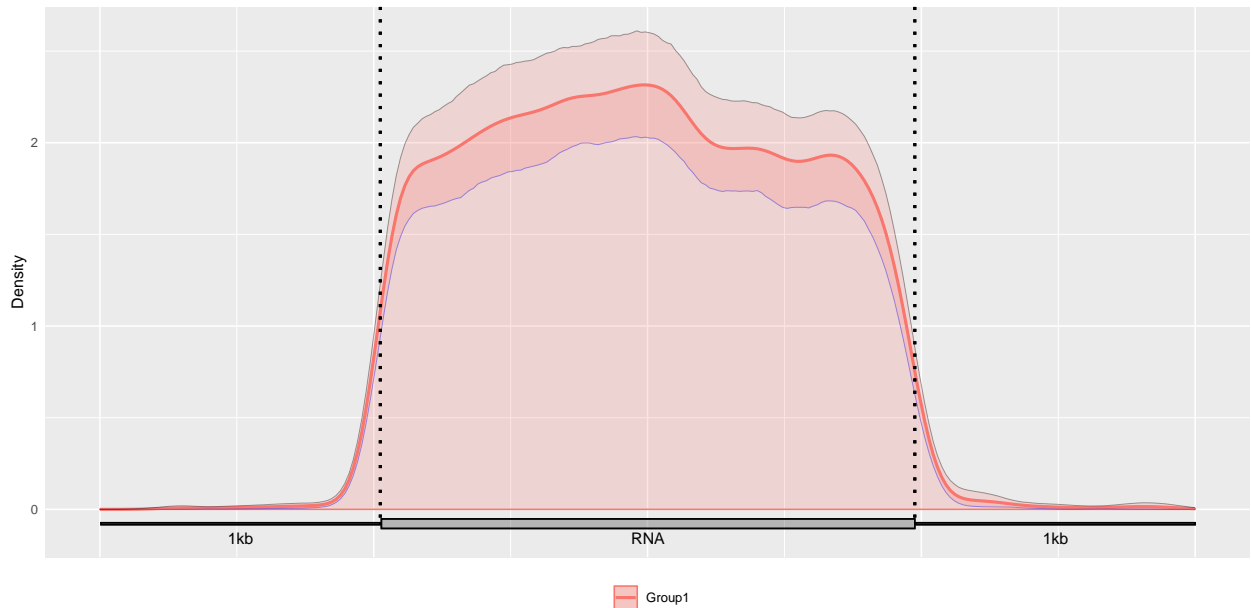
```
GuitarPlot(txTxdb = txdb,
           stBedFiles = stBedFiles,
           miscOutFilePrefix = "example")

## [1] "20191111001110"
## [1] "There are 2946 transcripts of 2946 genes in the genome."
## [1] "total 2946 transcripts extracted ..."
## [1] "total 2719 transcripts left after ambiguity filter ..."
## [1] "total 2719 transcripts left after check chromosome validity ..."
## [1] "total 1342 mRNAs left after component length filter ..."
## [1] "total 307 ncRNAs left after ncRNA length filter ..."
## [1] "generate components for all tx"
## [1] "generate components for mRNA"
## [1] "generate components for lncRNA"
## [1] "generate chiped transcriptome"
## [1] "generate coverage checking ranges for tx"
## [1] "generate coverage checking ranges for mrna"
## [1] "generate coverage checking ranges for ncrna"
## [1] "20191111001121"
## [1] "import BED file C:/Users/biocbuild/bbs-3.11-bioc/tmpdir/RtmpyKLOSK/Rinst18a46fd468af/Guitar/exto
## [1] "sample 10 points for Group1"
## [1] "start figure plotting for tx ..."
## [1] "start figure plotting for mrna ..."
## [1] "start figure plotting for ncrna ..."
```

Alternatively, you may also optionally include the promoter DNA region and tail DNA region on the 5' and 3' side of a transcript in the plot with parameter `headOrtail = TRUE`.

```
GuitarPlot(txTxdb = txdb,
           stBedFiles = stBedFiles,
           headOrtail = TRUE)

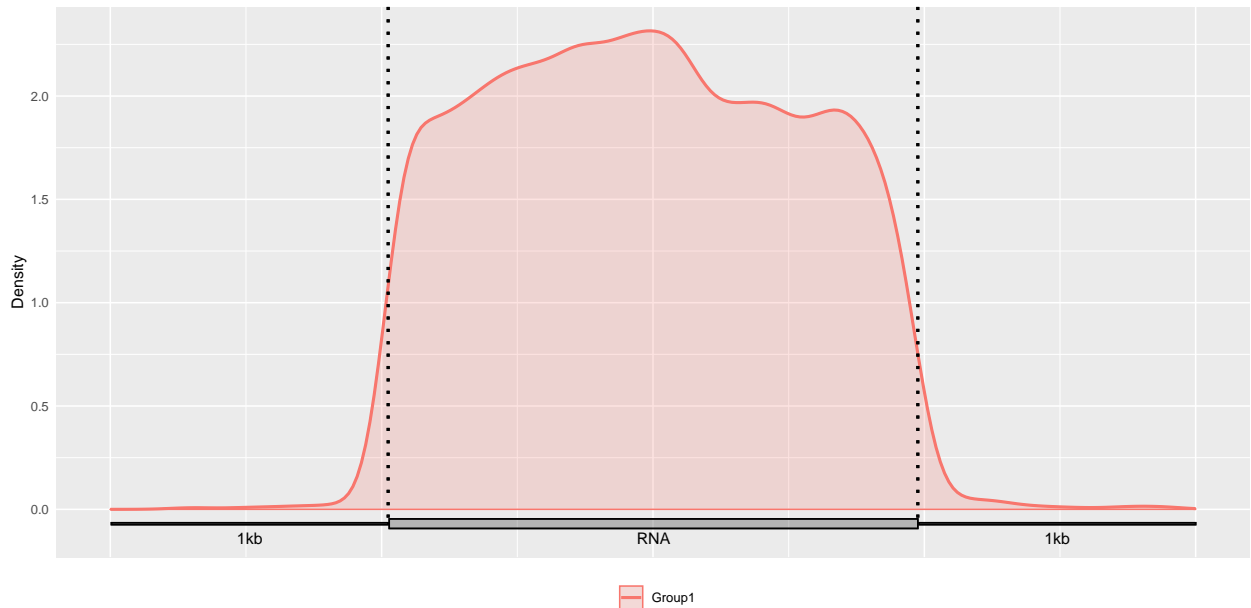
## [1] "20191111001151"
## [1] "There are 2946 transcripts of 2946 genes in the genome."
## [1] "total 2946 transcripts extracted ..."
## [1] "total 2719 transcripts left after ambiguity filter ..."
## [1] "total 2719 transcripts left after check chromosome validity ..."
## [1] "total 1342 mRNAs left after component length filter ..."
## [1] "total 307 ncRNAs left after ncRNA length filter ..."
## [1] "generate components for all tx"
## [1] "generate components for mRNA"
## [1] "generate components for lncRNA"
## [1] "generate chiped transcriptome"
## [1] "generate coverage checking ranges for tx"
## [1] "generate coverage checking ranges for mrna"
## [1] "generate coverage checking ranges for ncrna"
## [1] "20191111001202"
## [1] "import BED file C:/Users/biocbuild/bbs-3.11-bioc/tmpdir/RtmpyKLOSK/Rinst18a46fd468af/Guitar/exto
## [1] "sample 10 points for Group1"
## [1] "start figure plotting for tx ..."
```



Alternatively, you may also optionally include the Confidence Interval for guitar plot with parameter `enableCI = FALSE`.

```
GuitarPlot(txTxdb = txdb,
            stBedFiles = stBedFiles,
            headOrtail = TRUE,
            enableCI = FALSE)

## [1] "20191111001220"
## [1] "There are 2946 transcripts of 2946 genes in the genome."
## [1] "total 2946 transcripts extracted ..."
## [1] "total 2719 transcripts left after ambiguity filter ..."
## [1] "total 2719 transcripts left after check chromosome validity ..."
## [1] "total 1342 mRNAs left after component length filter ..."
## [1] "total 307 ncRNAs left after ncRNA length filter ..."
## [1] "generate components for all tx"
## [1] "generate components for mRNA"
## [1] "generate components for lncRNA"
## [1] "generate chiped transcriptome"
## [1] "generate coverage checking ranges for tx"
## [1] "generate coverage checking ranges for mrna"
## [1] "generate coverage checking ranges for ncrna"
## [1] "20191111001233"
## [1] "import BED file C:/Users/biocbuild/bbs-3.11-bioc/tmpdir/RtmpyKLOSK/Rinst18a46fd468af/Guitar/exto"
## [1] "sample 10 points for Group1"
## [1] "start figure plotting for tx ..."
```



## 2 Supported Data Format

Besides BED file, Guitar package also supports GRangesList and GRanges data structures. Please see the following examples.

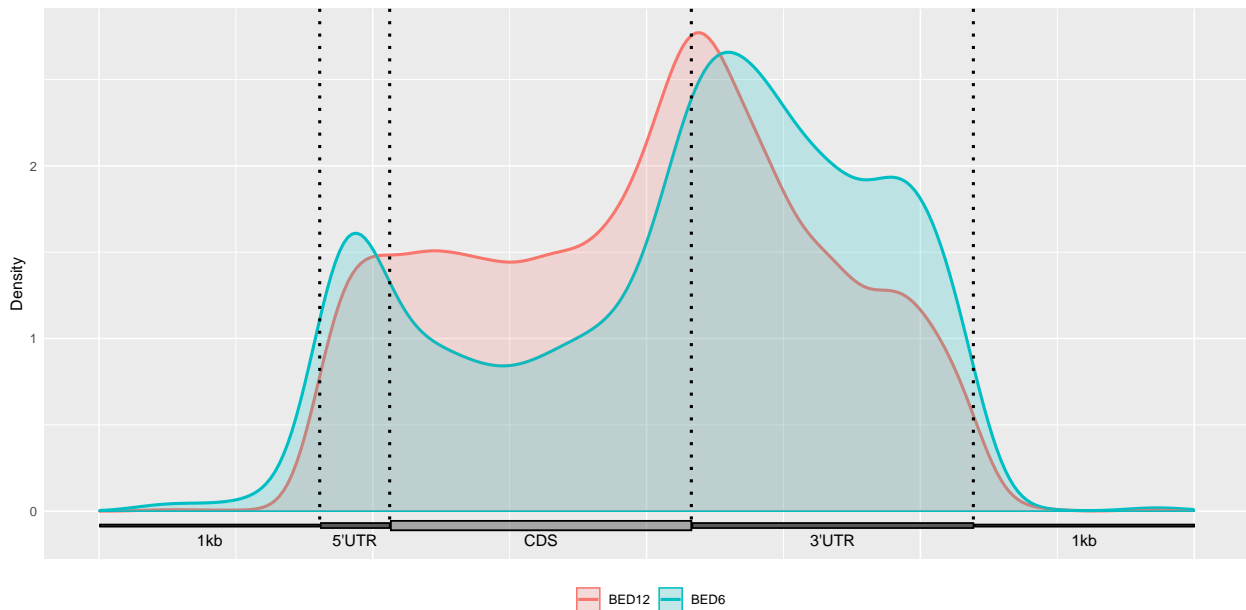
```
# import different data formats into a named list object.
# These genomic features are using mm10 genome assembly
stBedFiles <- list(system.file("extdata", "m6A_mm10_exomePeak_1000peaks_bed12.bed",
                             package="Guitar"),
                  system.file("extdata", "m6A_mm10_exomePeak_1000peaks_bed6.bed",
                             package="Guitar"))

# Build Guitar Coordinates
txdb_file <- system.file("extdata", "mm10_toy.sqlite",
                        package="Guitar")
txdb <- loadDb(txdb_file)

# Guitar Plot
GuitarPlot(txTxdb = txdb,
            stBedFiles = stBedFiles,
            headOrtail = TRUE,
            enableCI = FALSE,
            mapFilterTranscript = TRUE,
            pltTxType = c("mrna"),
            stGroupName = c("BED12", "BED6"))

## [1] "20191111001235"
## [1] "There are 2946 transcripts of 2946 genes in the genome."
## [1] "total 2946 transcripts extracted ..."
## [1] "total 2719 transcripts left after ambiguity filter ..."
## [1] "total 2719 transcripts left after check chromosome validity ..."
## [1] "total 1342 mRNAs left after component length filter ..."
## [1] "total 307 ncRNAs left after ncRNA length filter ..."
```

```
## [1] "generate components for mRNA"
## [1] "generate chiped transcriptome"
## [1] "generate coverage checking ranges for mrna"
## [1] "20191111001247"
## [1] "import BED file C:/Users/biocbuild/bbs-3.11-bioc/tmpdir/RtmpyKLOSK/Rinst18a46fd468af/Guitar/ext"
## [1] "import BED file C:/Users/biocbuild/bbs-3.11-bioc/tmpdir/RtmpyKLOSK/Rinst18a46fd468af/Guitar/ext"
## [1] "sample 10 points for BED12"
## [1] "sample 10 points for BED6"
## [1] "start figure plotting for mrna ..."
```



### 3 Processing of sampling sites information

We can select parameters for site sampling.

```
stGRangeLists = vector("list", length(stBedFiles))
sitesPoints <- list()
for (i in seq_len(length(stBedFiles))) {
  stGRangeLists[[i]] <- blocks(import(stBedFiles[[i]]))
}
for (i in seq_len(length(stGRangeLists))) {
  sitesPoints[[i]] <- samplePoints(stGRangeLists[i],
    stSampleNum = 10,
    stAmblguity = 5,
    pltTxType = c("mrna"),
    stSampleModle = "Equidistance",
    mapFilterTranscript = FALSE,
    guitarTxdb = guitarTxdb)
}
```

## 4 Guitar Coordinates - Transcriptomic Landmarks Projected on Genome

The `guitarTxdb` object contains the genome-projected transcriptome coordinates, which can be valuable for evaluating transcriptomic information related applications, such as checking the quality of MeRIP-Seq data. The `Guitar` coordinates are essentially the genomic projection of standardized transcript-based coordinates, making a viable bridge between the landmarks on transcript and genome-based coordinates.

It is based on the `txdb` object input, extracts the transcript information in `txdb`, selects the transcripts that match the parameters according to the component parameters set by the user, and saves according to the transcript type (`tx`, `mrna`, `ncrna`).

```
guitarTxdb <- makeGuitarTxdb(txdb = txdb,
                             txAmbiguity = 5,
                             txMrnaComponentProp = c(0.1,0.15,0.6,0.05,0.1),
                             txLncrnaComponentProp = c(0.2,0.6,0.2),
                             pltTxType = c("tx","mrna","ncrna"),
                             txPrimaryOnly = FALSE)

## [1] "There are 2946 transcripts of 2946 genes in the genome."
## [1] "total 2946 transcripts extracted ..."
## [1] "total 2719 transcripts left after ambiguity filter ..."
## [1] "total 2719 transcripts left after check chromosome validity ..."
## [1] "total 1342 mRNAs left after component length filter ..."
## [1] "total 307 ncRNAs left after ncRNA length filter ..."
## [1] "generate components for all tx"
## [1] "generate components for mRNA"
## [1] "generate components for lncRNA"
## [1] "generate chipped transcriptome"
## [1] "generate coverage checking ranges for tx"
## [1] "generate coverage checking ranges for mrna"
## [1] "generate coverage checking ranges for ncrna"
```

## 5 Check the Overlapping between Different Components

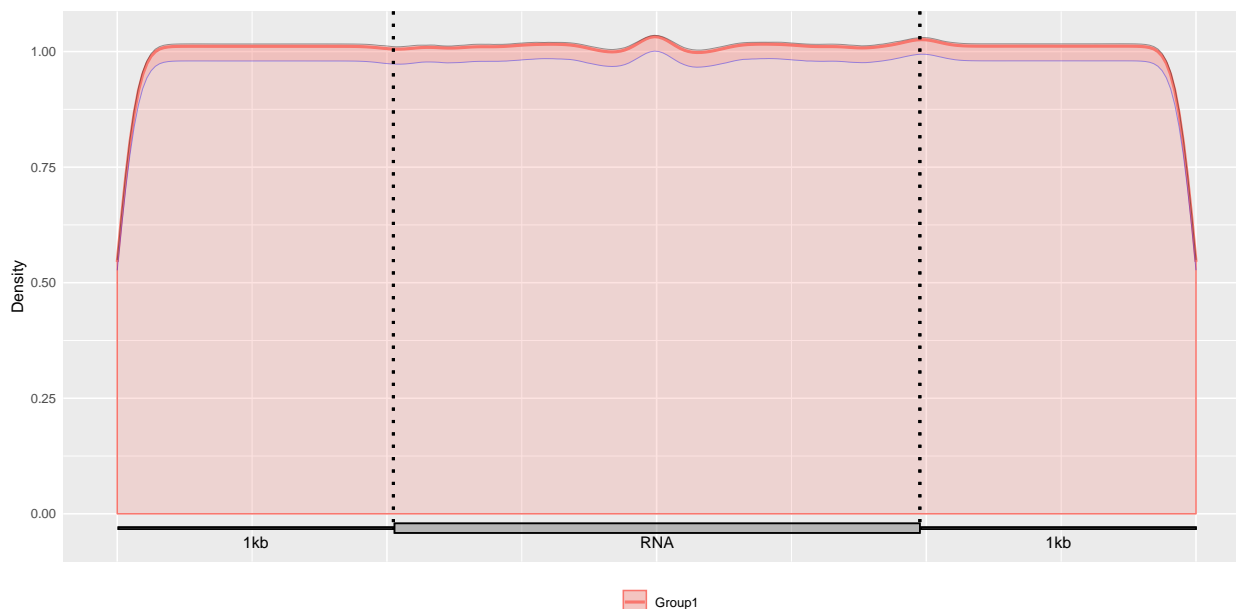
We can also check the distribution of the `Guitar` coordinates built.

```
gcl <- list(guitarTxdb$tx$tx)
GuitarPlot(txTxdb = txdb,
            stGRangeLists = gcl,
            stSampleNum = 200,
            enableCI = TRUE,
            pltTxType = c("tx"),
            txPrimaryOnly = FALSE
            )

## [1] "20191111001302"
## [1] "There are 2946 transcripts of 2946 genes in the genome."
## [1] "total 2946 transcripts extracted ..."
## [1] "total 2719 transcripts left after ambiguity filter ..."
## [1] "total 2719 transcripts left after check chromosome validity ..."
## [1] "total 1342 mRNAs left after component length filter ..."
## [1] "total 307 ncRNAs left after ncRNA length filter ..."
## [1] "generate components for all tx"
## [1] "generate chipped transcriptome"
```



```
## [1] "generate coverage checking ranges for tx"
## [1] "20191111001313"
## [1] "sample 200 points for Group1"
## [1] "start figure plotting for tx ..."
```

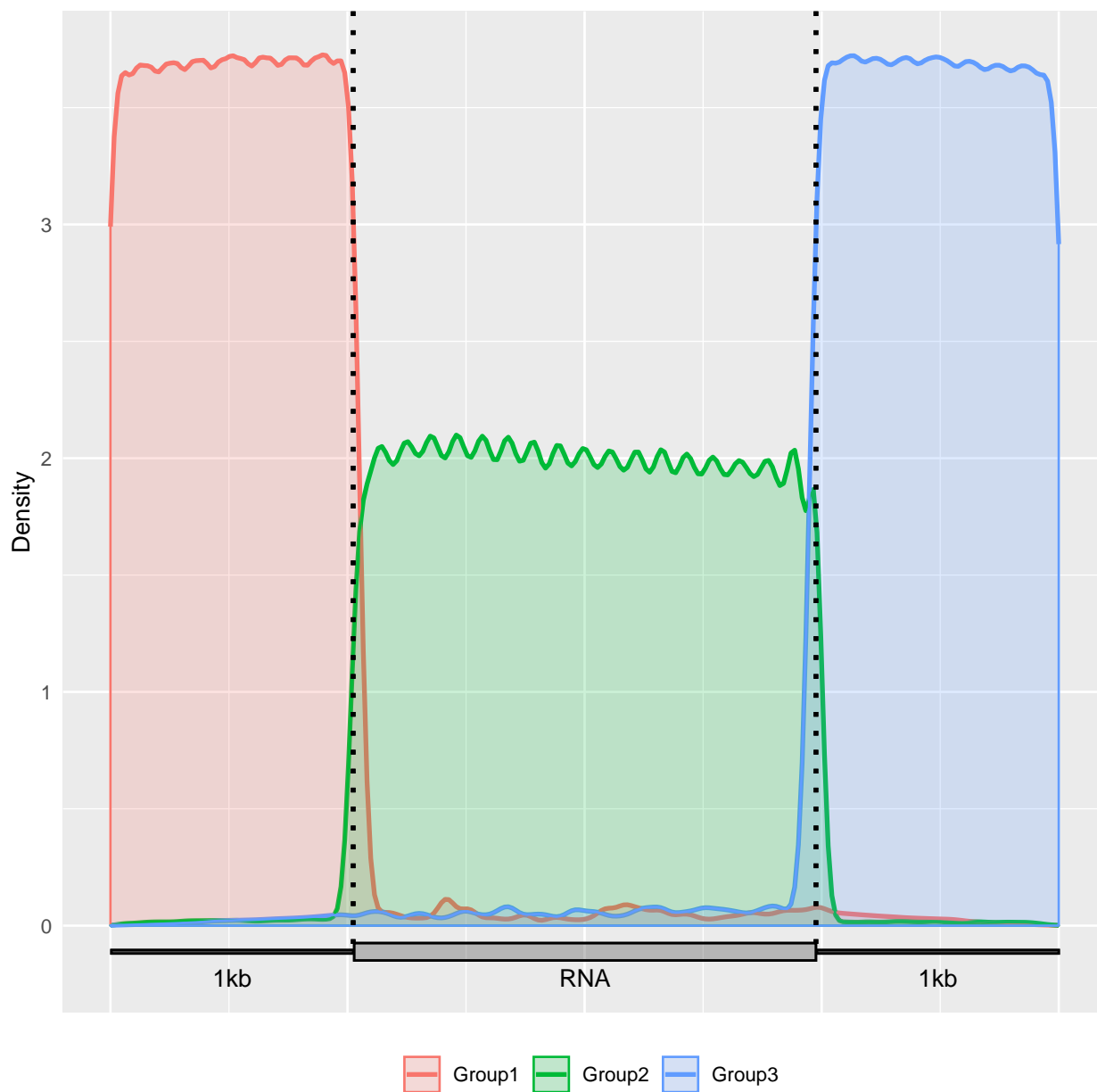


Alternatively, we can extract the RNA components, check the distribution of tx components in the transcriptome

```
GuitarCoords <- guitarTxdb$tx$txComponentGRange
type <- paste(mcols(GuitarCoords)$componentType,mcols(GuitarCoords)$txType)
key <- unique(type)
landmark <- list(1,2,3,4,5,6,7,8,9,10,11)
names(landmark) <- key
for (i in 1:length(key)) {
  landmark[[i]] <- GuitarCoords[type==key[i]]
}
GuitarPlot(txTxdb = txdb ,
            stGRangeLists = landmark[1:3],
            pltTxType = c("tx"),
            enableCI = FALSE
)

## [1] "20191111002231"
## [1] "There are 2946 transcripts of 2946 genes in the genome."
## [1] "total 2946 transcripts extracted ..."
## [1] "total 2719 transcripts left after ambiguity filter ..."
## [1] "total 2719 transcripts left after check chromosome validity ..."
## [1] "total 1342 mRNAs left after component length filter ..."
## [1] "total 307 ncRNAs left after ncRNA length filter ..."
## [1] "generate components for all tx"
## [1] "generate chiped transcriptome"
## [1] "generate coverage checking ranges for tx"
## [1] "20191111002241"
## [1] "sample 10 points for Group1"
```

```
## [1] "sample 10 points for Group2"
## [1] "sample 10 points for Group3"
## [1] "start figure plotting for tx ..."
```

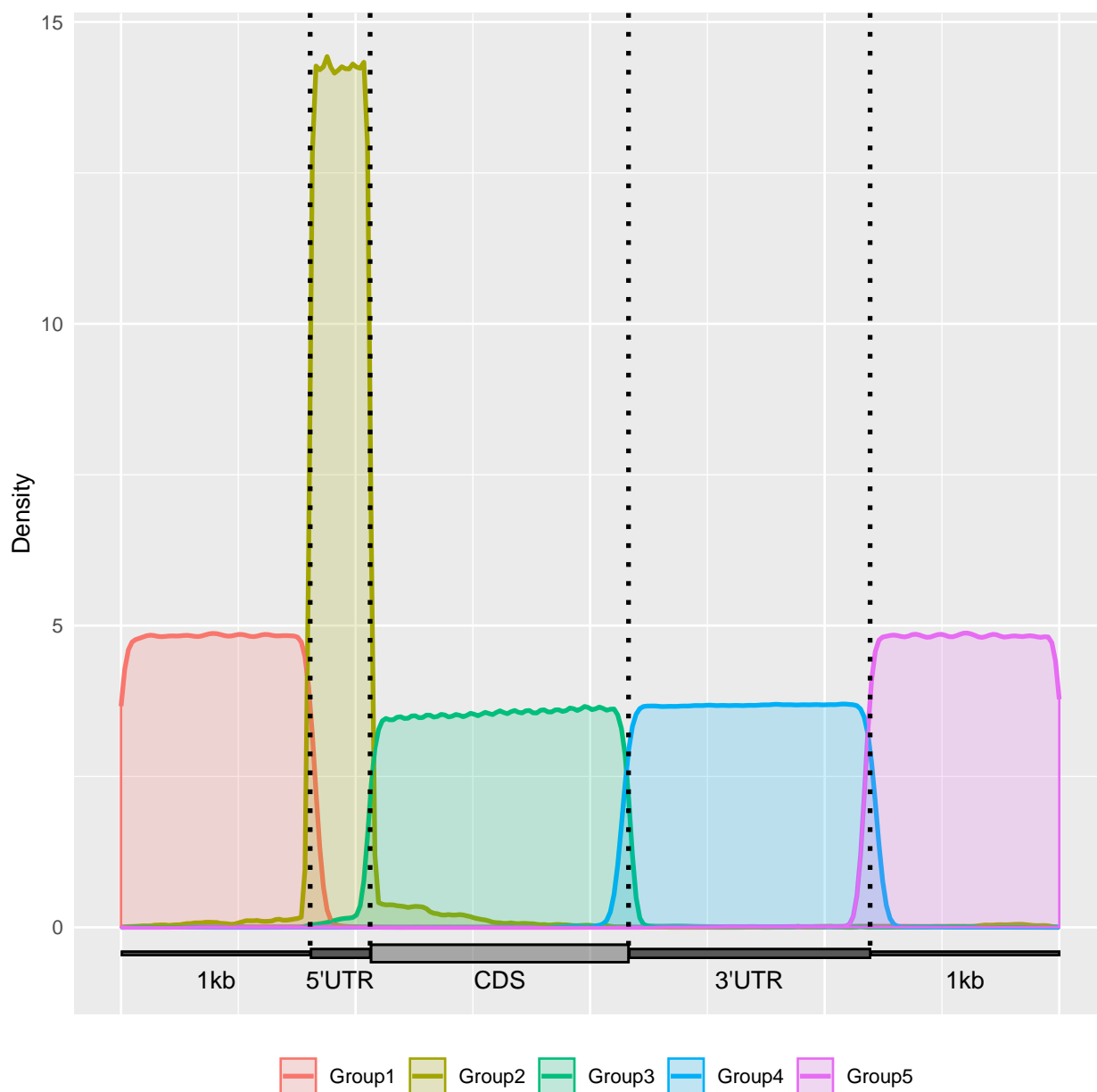


Check the distribution of mRNA components in the transcriptome

```
GuitarPlot(txTxdb = txdb ,
            stGRangeLists = landmark[4:8],
            pltTxType = c("mrna"),
            enableCI = FALSE
)

## [1] "20191111002303"
## [1] "There are 2946 transcripts of 2946 genes in the genome."
## [1] "total 2946 transcripts extracted ..."
```

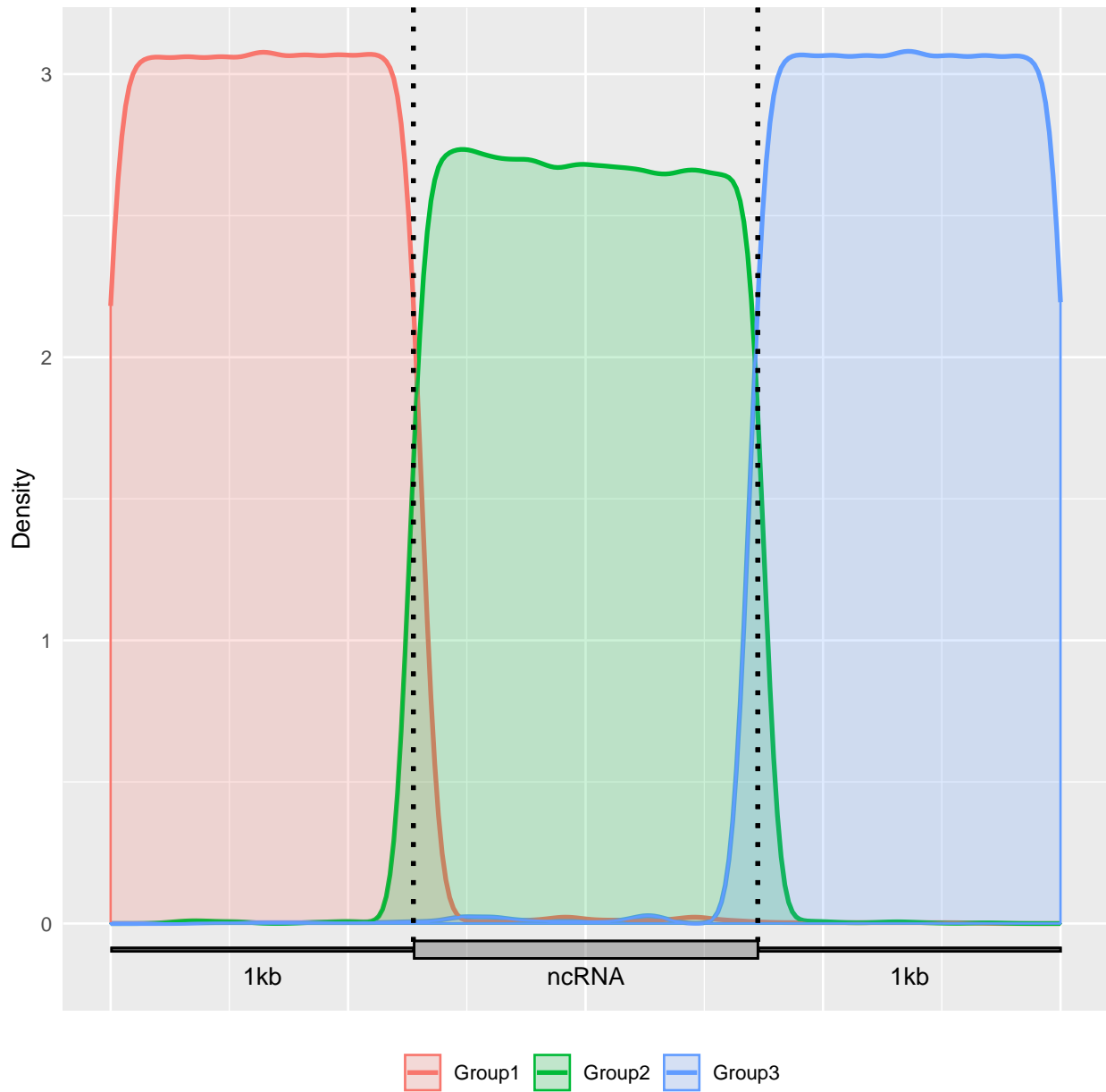
```
## [1] "total 2719 transcripts left after ambiguity filter ..."
## [1] "total 2719 transcripts left after check chromosome validity ..."
## [1] "total 1342 mRNAs left after component length filter ..."
## [1] "total 307 ncRNAs left after ncRNA length filter ..."
## [1] "generate components for mRNA"
## [1] "generate chiped transcriptome"
## [1] "generate coverage checking ranges for mrna"
## [1] "20191111002315"
## [1] "sample 10 points for Group1"
## [1] "sample 10 points for Group2"
## [1] "sample 10 points for Group3"
## [1] "sample 10 points for Group4"
## [1] "sample 10 points for Group5"
## [1] "start figure plotting for mrna ..."
```



Check the distribution of lncRNA components in the transcriptome

```
GuitarPlot(txTxdb = txdb ,
           stGRangeLists = landmark[9:11],
           pltTxType = c("ncrna"),
           enableCI = FALSE
)

## [1] "20191111002328"
## [1] "There are 2946 transcripts of 2946 genes in the genome."
## [1] "total 2946 transcripts extracted ..."
## [1] "total 2719 transcripts left after ambiguity filter ..."
## [1] "total 2719 transcripts left after check chromosome validity ..."
## [1] "total 1342 mRNAs left after component length filter ..."
## [1] "total 307 ncRNAs left after ncRNA length filter ..."
## [1] "generate components for lncRNA"
## [1] "generate chiped transcriptome"
## [1] "generate coverage checking ranges for ncRNA"
## [1] "20191111002341"
## [1] "sample 10 points for Group1"
## [1] "sample 10 points for Group2"
## [1] "sample 10 points for Group3"
## [1] "start figure plotting for ncRNA ..."
```



## 6 Session Information

```
sessionInfo()

## R Under development (unstable) (2019-11-04 r77367)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows Server 2012 R2 x64 (build 9600)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=C
```

```

## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats4      parallel  stats      graphics  grDevices
## [6] utils       datasets  methods    base
##
## other attached packages:
## [1] Guitar_2.3.0          dplyr_0.8.3
## [3] ggplot2_3.2.1         magrittr_1.5
## [5] rtracklayer_1.47.0    GenomicFeatures_1.39.0
## [7] AnnotationDbi_1.49.0  Biobase_2.47.0
## [9] GenomicRanges_1.39.1  GenomeInfoDb_1.23.0
## [11] IRanges_2.21.1        S4Vectors_0.25.0
## [13] BiocGenerics_0.33.0   knitr_1.25
##
## loaded via a namespace (and not attached):
## [1] Rcpp_1.0.3            lattice_0.20-38
## [3] prettyunits_1.0.2     Rsamtools_2.3.2
## [5] Biostrings_2.55.0     assertthat_0.2.1
## [7] zeallot_0.1.0         digest_0.6.22
## [9] BiocFileCache_1.11.2  R6_2.4.0
## [11] backports_1.1.5       RSQLite_2.1.2
## [13] evaluate_0.14         httr_1.4.1
## [15] highr_0.8             pillar_1.4.2
## [17] zlibbioc_1.33.0       rlang_0.4.1
## [19] progress_1.2.2        lazyeval_0.2.2
## [21] curl_4.2              blob_1.2.0
## [23] Matrix_1.2-17         labeling_0.3
## [25] BiocParallel_1.21.0   stringr_1.4.0
## [27] RCurl_1.95-4.12       bit_1.1-14
## [29] biomaRt_2.43.0        munsell_0.5.0
## [31] DelayedArray_0.13.0   compiler_4.0.0
## [33] xfun_0.10             pkgconfig_2.0.3
## [35] askpass_1.1           openssl_1.4.1
## [37] tidyselect_0.2.5     SummarizedExperiment_1.17.0
## [39] tibble_2.1.3          GenomeInfoDbData_1.2.2
## [41] matrixStats_0.55.0    XML_3.98-1.20
## [43] withr_2.1.2          crayon_1.3.4
## [45] dbplyr_1.4.2          GenomicAlignments_1.23.0
## [47] bitops_1.0-6         rappdirs_0.3.1
## [49] grid_4.0.0           gtable_0.3.0
## [51] DBI_1.0.0            scales_1.0.0
## [53] stringi_1.4.3        XVector_0.27.0
## [55] vctrs_0.2.0          tools_4.0.0
## [57] bit64_0.9-7          glue_1.3.1
## [59] purrr_0.3.3          hms_0.5.2
## [61] colorspace_1.4-1     memoise_1.1.0

```