Marcel Buczkowski 145159

Bioconductor raport – tRNA

Zadanie 1.

Wczytanie plików I pakietów:

```
BiocManager::install("tRNAscanImport")
library(tRNAscanImport)
library(Biostrings)
library(rtracklayer)
BiocManager::install("tRNA")
library(tRNA)


###zad1###
human_file <- system.file("extdata",
                          file = "human.tRNAscan",
                          package = "tRNAscanImport")

human_gr <- import.tRNAscanAsGRanges(human_file)
head(human_gr, 2)
istRNAGRanges(human_gr)

yeast_file <- system.file("extdata",
                          file = "yeast.tRNAscan",
                          package = "tRNAscanImport")

ecoli_file <- system.file("extdata",
                          file = "ecoli.tRNAscan",
                          package = "tRNAscanImport")

yeast_gr <- import.tRNAscanAsGRanges(yeast_file)
coli_gr <- import.tRNAscanAsGRanges(ecoli_file)
```

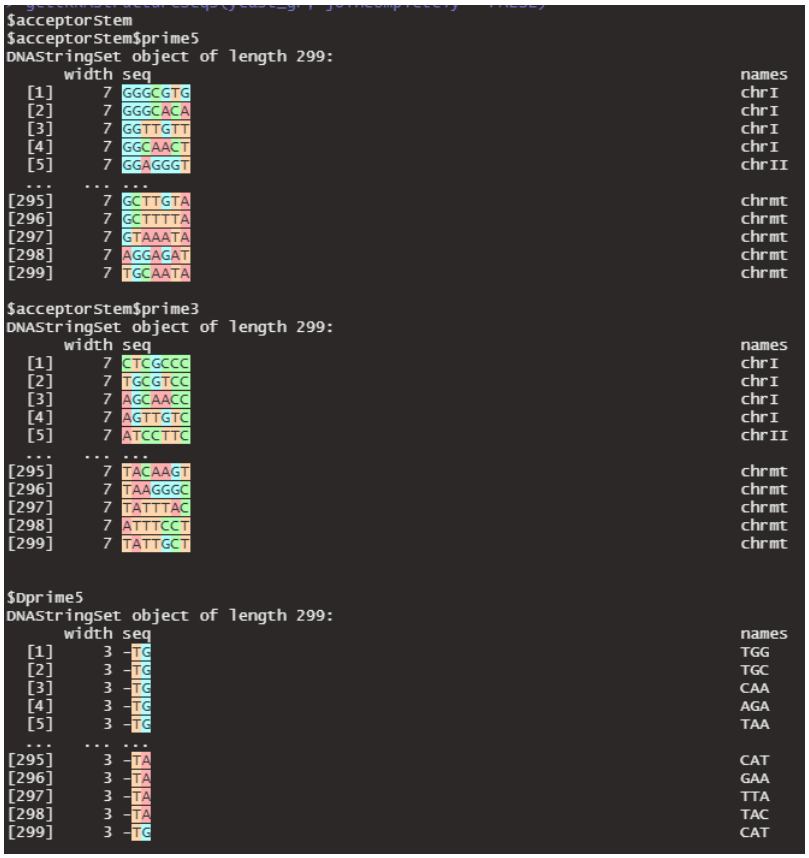
Funkcja getRNAstructureSeqs:

```
gettRNAstructureSeqs(yeast_gr, joinCompletely = TRUE)
```

```
DNAStringSet object of length 299:
      width seq
  [1]    97 GGGCGTG-TGGTC-TAGT--GGTA--T-GAT-TCTCGC-TT...------------CCCTGGG-TTCAATTCCCAGCTCGCCCC
  [2]    97 GGGCACA-TGGCGCAGTT--GGT---AGCGC-GCTTCC-CT...------------TCATCGG-TTCGATTCCGGTTGCGTCCA
  [3]    97 GGTTGTT-TGGCC-GAGC--GGTCTAA-GGC-GCCTGA-TT...GTAA-----GATGCAAGAG-TTCGAATCTCTTAGCAACCA
  [4]    97 GGCAACT-TGGCC-GAGT--GGTTA-A-GGC-GAAAGA-TT...TTT-----GCCCGCGCAGG-TTCGAGTCCTGCAGTTGTCG
  [5]    97 GGAGGGT-TGGCC-GAGT--GGTCTAA-GGC-GGCAGA-CT...GGTT----GTCCGCGCGAG-TTCGAACCTCGCATCCTTCA
  ...   ... ...
[295]    97 GCTTGTA-TAGTTTAATT--GGTT--AAAAC-ATTTGT-CT...------------TGTAAGG-TTCAATTCCTTCTACAAGTA
[296]    97 GCTTTTA-TAGCTTAGT---GGT---AAAGC-GATAAA-TT...------------ACATGTAGTTCGATTCTCATTAAGGGCA
[297]    97 GTAAATA-TAATTTAAT---GGT---AAAAT-GTATGTTTT...------------ATCTAAG-TTCAAATCTTAGTATTTACA
[298]    97 AGGAGAT-TAGCT-TAATT-GGTA--T-AGC-ATTCGT-TT...------------TTATAGG-TTCGAACCCTATATTTCCTA
[299]    97 TGCAATA-TGATGTAATT--GGTT--AACAT-TTTAGG-GT...------------ATATACG-TTCAAATCGTATTATTGCTA
```

Kiedy paramter joinCompletely jest ustawiony na true to zwracane są całe sekwencje tRNA. Kiedy ustawimy paramter na false mamy kilka list które reprezentują te same tRNA ale różne ich fragmenty strukturalne:

```
gettRNAstructureSeqs(yeast_gr, joinCompletely = FALSE)
```



```
$acceptorStem
$acceptorStem$prime5
DNAStringSet object of length 299:
      width seq                                                 names
  [1]     7 GGGCGTG                                             chrI
  [2]     7 GGGCACA                                             chrI
  [3]     7 GGTTGTT                                             chrI
  [4]     7 GGCAACT                                             chrI
  [5]     7 GGAGGGT                                             chrII
  ...   ... ...
[295]     7 GCTTGTA                                             chrmt
[296]     7 GCTTTTA                                             chrmt
[297]     7 GTAAATA                                             chrmt
[298]     7 AGGAGAT                                             chrmt
[299]     7 TGCAATA                                             chrmt

$acceptorStem$prime3
DNAStringSet object of length 299:
      width seq                                                 names
  [1]     7 CTCGCCC                                             chrI
  [2]     7 TGCGTCC                                             chrI
  [3]     7 AGCAACC                                             chrI
  [4]     7 AGTTGTC                                             chrI
  [5]     7 ATCCTTC                                             chrII
  ...   ... ...
[295]     7 TACAAGT                                             chrmt
[296]     7 TAAGGGC                                             chrmt
[297]     7 TATTTAC                                             chrmt
[298]     7 ATTTCCT                                             chrmt
[299]     7 TATTGCT                                             chrmt


$Dprime5
DNAStringSet object of length 299:
      width seq                                                 names
  [1]     3 -TG                                                 TGG
  [2]     3 -TG                                                 TGC
  [3]     3 -TG                                                 CAA
  [4]     3 -TG                                                 AGA
  [5]     3 -TG                                                 TAA
  ...   ... ...
[295]     3 -TA                                                 CAT
[296]     3 -TA                                                 GAA
[297]     3 -TA                                                 TTA
[298]     3 -TA                                                 TAC
[299]     3 -TG                                                 CAT
```

Parametr joinFeatures = TRUE odpowiada za to, że zwracane części sekwencji tRNA są połączone w jedną sekwencje jeśli należą one do tego samego typu. Parametr ten wyklucza się z parametrem joinCompletely. Np. łączone są części ramienia akceptora 3 i 5 w jedną sekwencje:

```
> gettRNAstructureSeqs(yeast_gr, joinFeatures = TRUE)
$acceptorStem
DNAStringSet object of length 299:
      width seq                                                 names
  [1]    17 GGGCGTG---CTCGCCC                                   TGG
  [2]    17 GGGCACA---TGCGTCC                                   TGC
  [3]    17 GGTTGTT---AGCAACC                                   CAA
  [4]    17 GGCAACT---AGTTGTC                                   AGA
  [5]    17 GGAGGGT---ATCCTTC                                   TAA
  ...   ... ...
[295]    17 GCTTGTA---TACAAGT                                   CAT
[296]    17 GCTTTTA---TAAGGGC                                   GAA
[297]    17 GTAAATA---TATTTAC                                   TTA
[298]    17 AGGAGAT---ATTTCCT                                   TAC
[299]    17 TGCAATA---TATTGCT                                   CAT

$Dprime5
DNAStringSet object of length 299:
      width seq                                                 names
  [1]     3 -TG                                                 TGG
  [2]     3 -TG                                                 TGC
  [3]     3 -TG                                                 CAA
  [4]     3 -TG                                                 AGA
  [5]     3 -TG                                                 TAA
  ...   ... ...
[295]     3 -TA                                                 CAT
[296]     3 -TA                                                 GAA
[297]     3 -TA                                                 TTA
```

Parametr structure przyjmujący określone argumenty ogranicza wynik tylko do zadanych rodzajów fragmentów struktury tRNA (np. Tloop)

```
gettRNAstructureSeqs(yeast_gr, structure="Tloop")
```

```
$Tloop
DNAStringSet object of length 299:
      width seq                                                 names
  [1]     7 TTCAATT                                             TGG
  [2]     7 TTCGATT                                             TGC
  [3]     7 TTCGAAT                                             CAA
  [4]     7 TTCGAGT                                             AGA
  [5]     7 TTCGAAC                                             TAA
  ...   ... ...
[295]     7 TTCAATT                                             CAT
[296]     7 TTCGATT                                             GAA
[297]     7 TTCAAAT                                             TTA
[298]     7 TTCGAAC                                             TAC
[299]     7 TTCAAAT                                             CAT
```

Parametr padSequences = TRUE pwoduje, że sekwencje tego samego typu zwracane są o tej samej długości.

```
gettRNAstructureSeqs(yeast_gr, structure="variableLoop", padSequences = FALSE)
gettRNAstructureSeqs(yeast_gr, structure="variableLoop", padSequences = TRUE)
```

```
> gettRNAstructureSeqs(yeast_gr, structure="variableLoop", padSequences = FALSE)
$variableLoop
DNAStringSet object of length 299:
      width seq                                                 names
  [1]     5 AGGCC                                               TGG
  [2]     5 AGGTC                                               TGC
  [3]    13 TATCGTAAGATGC                                       CAA
  [4]    14 TGGGCTTTGCCCGC                                      AGA
  [5]    15 TGGACGGTTGTCCGC                                     TAA
  ...   ... ...
[295]     5 TAATG                                               CAT
[296]     4 TTAC                                                GAA
[297]     4 TTAT                                                TTA
[298]     5 AGATT                                               TAC
[299]     4 TTAT                                                CAT

> gettRNAstructureSeqs(yeast_gr, structure="variableLoop", padSequences = TRUE)
$variableLoop
DNAStringSet object of length 299:
      width seq                                                 names
  [1]    20 AGG----------------CC                               TGG
  [2]    20 AGG----------------TC                               TGC
  [3]    20 TATC--GTAA-----GATGC                                CAA
  [4]    20 TGGGC-TTT-----GCCCGC                                AGA
  [5]    20 TGGAC-GGTT----GTCCGC                                TAA
  ...   ... ...
[295]    20 TAA----------------TG                               CAT
[296]    20 TT-----------------AC                               GAA
[297]    20 TT-----------------AT                               TTA
[298]    20 AGA----------------TT                               TAC
[299]    20 TT-----------------AT                               CAT
```

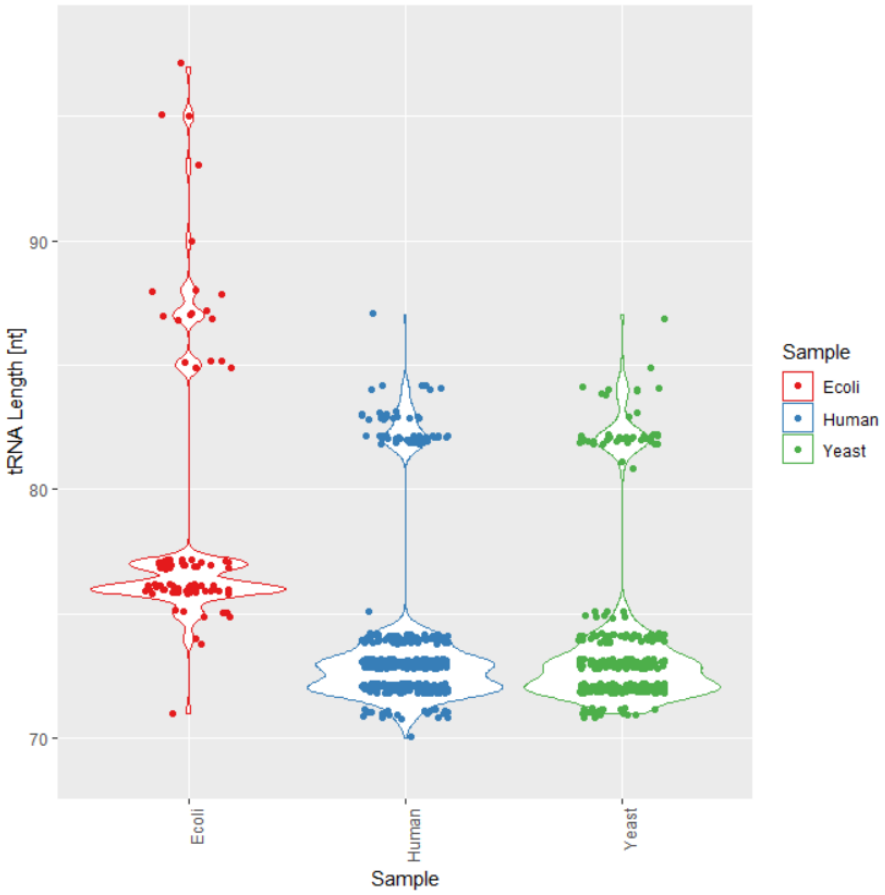
Następnie przeprowadzono wizualizacje:

```
grl <- GRangesList(Yeast = yeast_gr,
                   Human = human_gr,
                   Ecoli = coli_gr)

plots <- gettRNAFeaturePlots(grl)
```

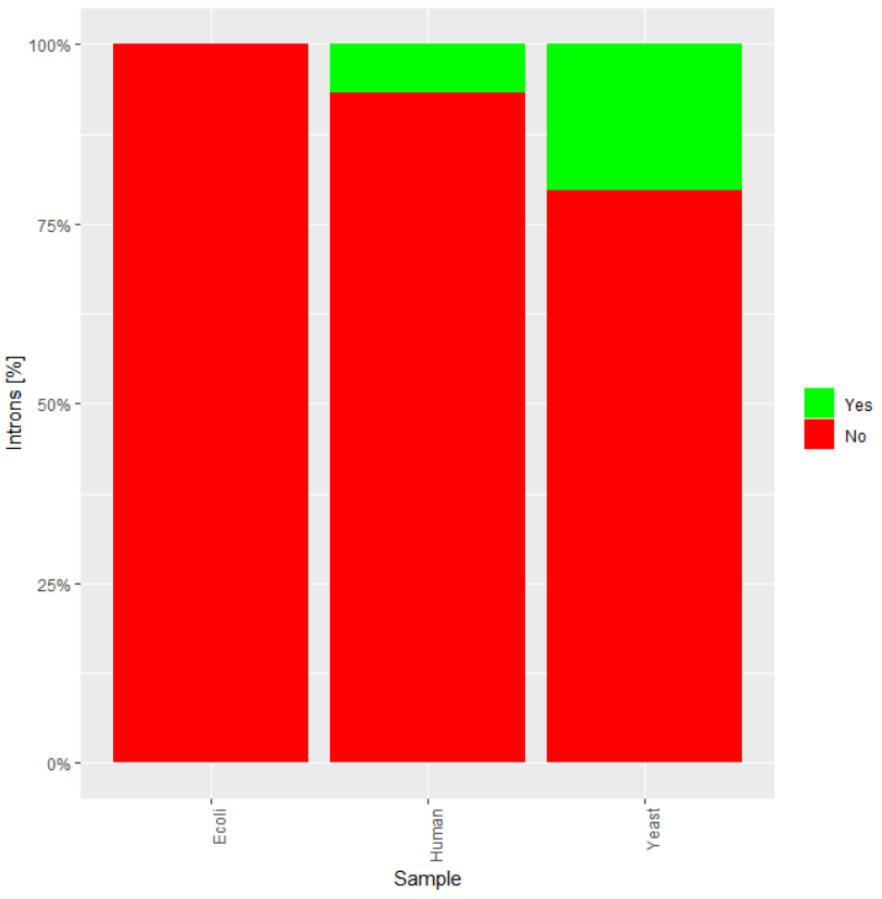
```
plots$length
```

Plot generuje 3 wykresy typu violin i kropkowe. Szerokie miejsca oznaczają pewne podgrupy danych które grupują w sobie większość obserwacji. Dla tych wykresów widać raczej dominację dwóch podgrup – zwłaszcza dla Człowieka i Drożdży. Trna tych gatunków mają bardzo zbliżone do siebie długości oraz rozkład danych. Mocno różnią się tutaj tRNA Ecoli, które na ogół są dłuższe oraz mogą występować u nich tRNA o długości nawet większe niż 90.

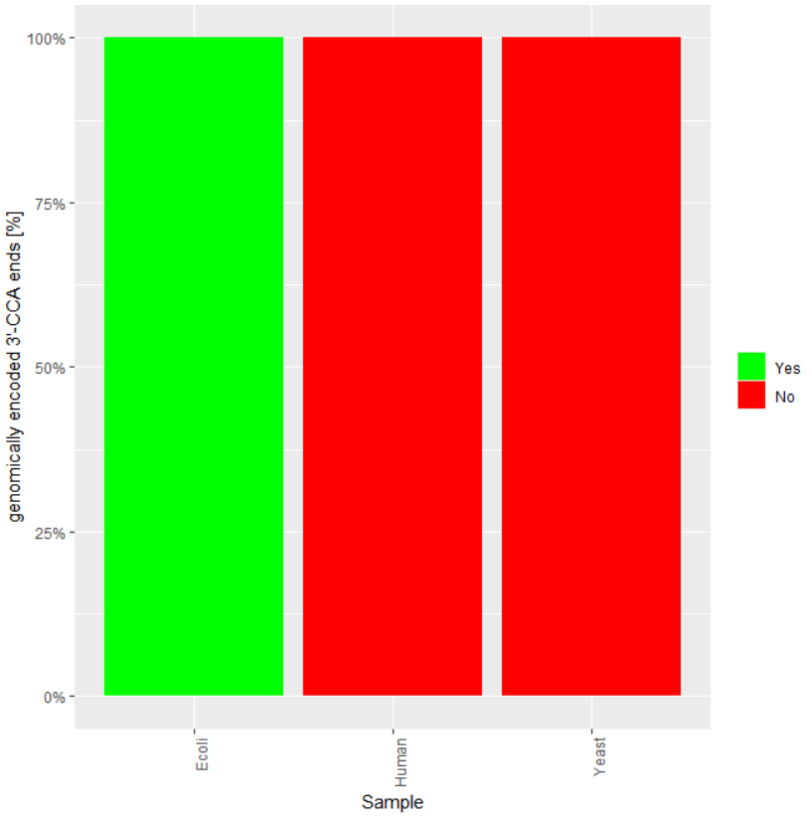



```
plots$tRNAscan_intron
```

Ten wykres przedstawia zawartość intronów w tRNA w zależności od gatunku. Widać, że próbki ecoli nie miały w sobie żadnych intronów, próbki człowieka ok 5% a próbki Droższy ok. 15% intronów.

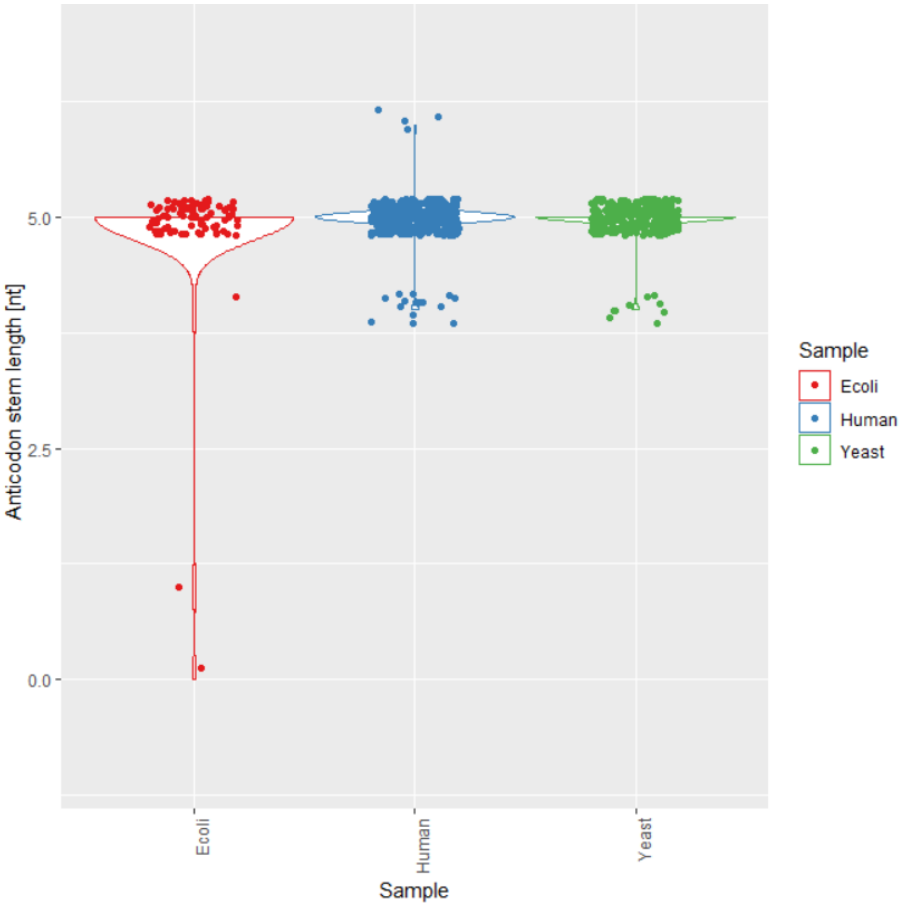


```
plots$cca
```



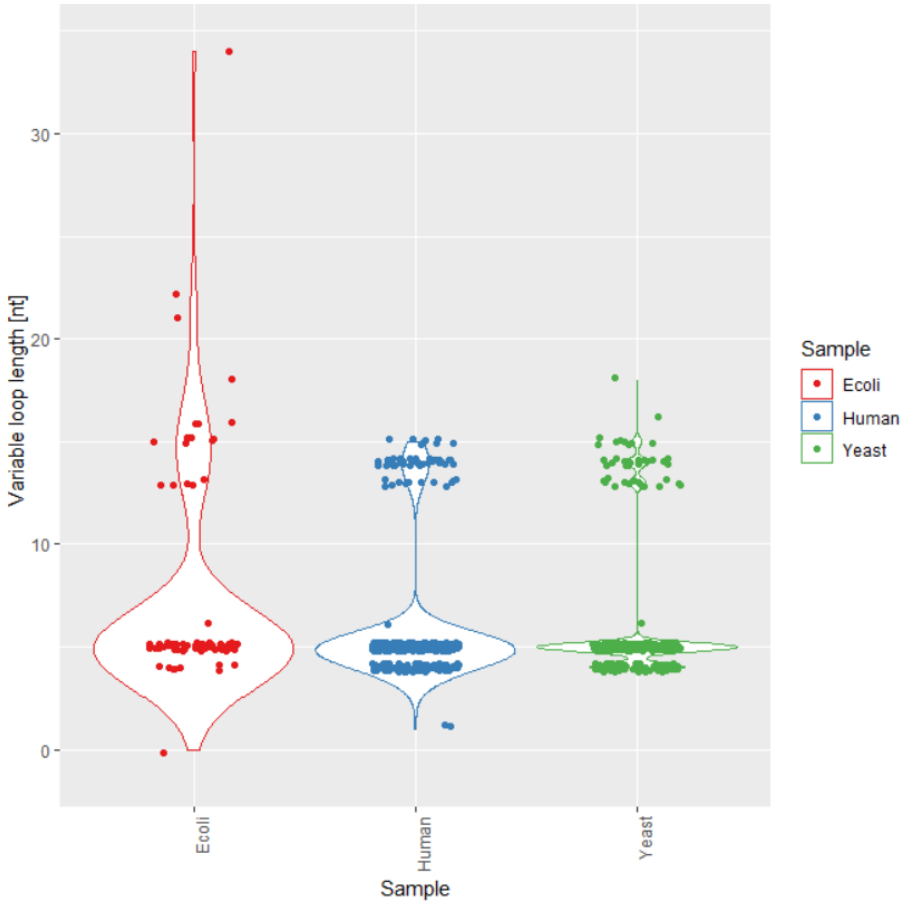

```
plots$anticodonStem_length
```

Długość ramienia antykodonu jest na ogół zawsze długości 5 chociaż widać też takie o długości 4. W dodatku dla człowieka występują takie o długości 6, co dla pozostałych gatunków się nie zdarza.
W dodatku u bakterii są 2 przypadki gdzie brakuje tego ramienia.



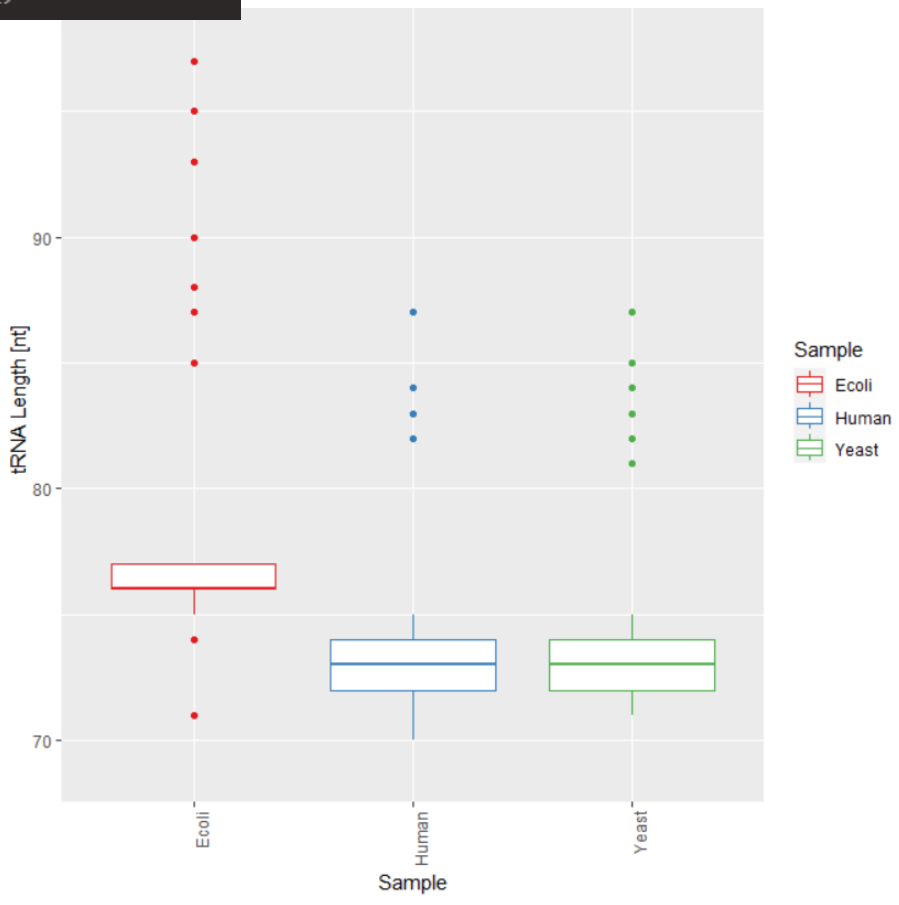
```
plots$variableLoop_length
```

Na tym wykresie pokazano długości pętli zmiennej w tRNA. Ponownie dla droższy i człowieka są to podobne długości a dla bakterii również są podobne, ale występuję pewne długie wyjątki.




```
plots$length$layers <- plots$length$layers[c(-1L,-2L)]
plots$length + ggplot2::geom_boxplot()
```

Można również wprowadzać modyfikacje ggplot2 do wykresów. Np. zamienić go na boxplot

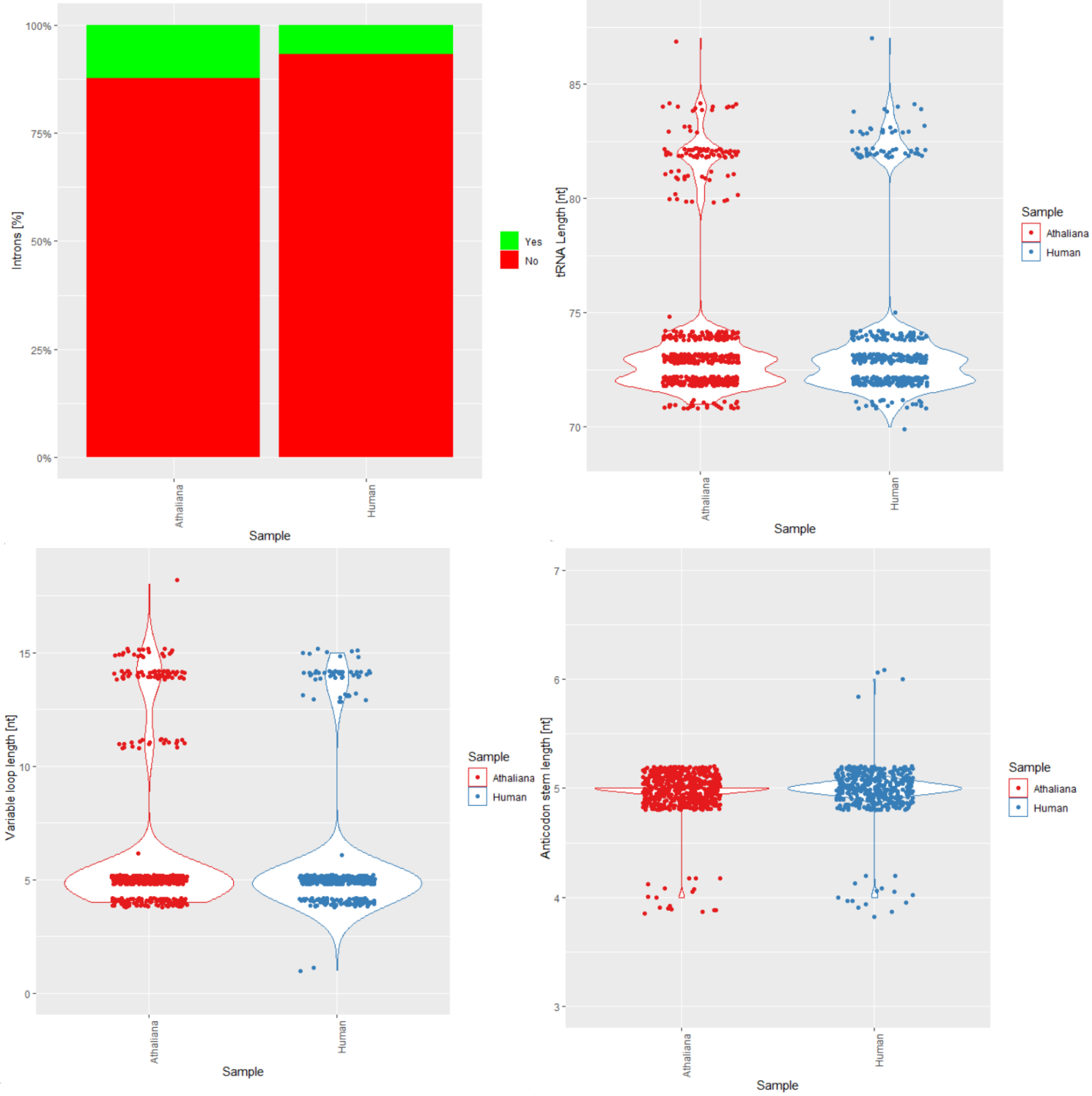


Marcel Buczkowski 145159

Zadanie 2.

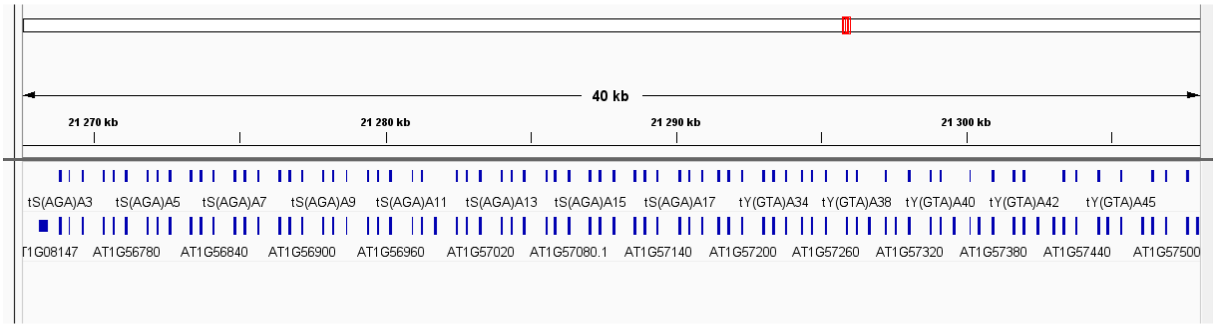
```
### ZAD 2 ###
Ath_gr <- import.tRNAscanAsGRanges("C:\\Users\\marce\\hakowanie\\programming_R\\bioconductor\\
Ath_gr
istRNAGRanges(Ath_gr)
grl <- GRangesList(Athaliana = Ath_gr,
                   Human = human_gr)

plots <- gettRNAFeaturePlots(grl)
plots$length
plots$tRNAscan_intron
plots$variableLoop_length
plots$anticodonStem_length
```



Nie ma istotnych różnic między tRNA człowieka i A.thaliana.

```r
gff <- tRNAscan2GFF(Ath_gr)
export.gff3(gff, con = "trna.gff")
```



W dużej mierze przewidywania trna-Scan są zgodne z aktualną adnotacją genomową. Na prawym końcu widać, że trna scan nie wykrył niektórych tRNA, które obecne są w adnotacji.

```r
library(rtracklayer)
gff = import.gff3("C:\\Users\\marce\\hakowanie\\programming_R\\bioconductor\\lab9\\annotation.gff3")
head(gff)
seqlevelsStyle(gff) <- "NCBI"
seqlevelsStyle(Ath_gr) <- "NCBI"

overlaps<-findOverlaps(Ath_gr, gff)
overlaps

hits <- queryHits(overlaps)
hits_u <- unique(hits)
length(hits_u)
```

```
> countOverlaps(Ath_gr, gff)
  [1] 3 3 5 3 3 3 3 3 3 4 3 3 3 3 6 3 3 3 3 3 3 3 3 3 8 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [77] 3 3 3 3 3 3 3 3 3 8 3 3 3 3 3 3 3 3 3 3 3 3 3 4 3 3 3 3 3 4 3 3 3 3 3 4 4 3 4 4 3 4 4 3 4 4 3 4 4 3 4 4 3 4 4 3 4 4 3 4 3 4 4 3 4 4 3 4 4 3 4 4 3 4 4 3
[153] 4 4 3 4 4 3 4 4 3 4 4 3 4 3 4 3 4 3 4 3 4 4 4 3 4 4 4 3 4 3 3 3 3 3 3 3 3 3 3 3 3 3 3 7 3 3 3 3 3 3 7 3 3 3 3 3 3 3 3 3 3 3 3 3 4 3 3 3 3 3 3 0 0 0 0
[229] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[305] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[381] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[457] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[533] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Widać, że faktycznie dla każdego trna z Ath_gr mamy kilka overlaps do pliku z adnotacją. Można również wyświetlić obiekt nałożen za pomocą funkcji findOverlaps

```
> overlaps
Hits object with 746 hits and 0 metadata columns:
        queryHits subjectHits
        <integer>   <integer>
    [1]         1         834
    [2]         1         835
    [3]         1         836
    [4]         2        2521
    [5]         2        2522
    ...       ...         ...
  [742]       223      177530
  [743]       223      177531
  [744]       224      178185
  [745]       224      178186
  [746]       224      178187
  -------
  queryLength: 584 / subjectLength: 179672
```

Obiekt heats przechowuje za pomocą funkcji queryHits ID wszystkich overlapów, wyświetlając unikalny set możemy bez powtórzeń znaleźć wszystkie dopasowane tRNA (jest ich 224):