RiboProfiling - Analysis and quality assessment of Ribosome-profiling data

Alexandra Popa[1*]

[1] Institut de Pharmacologie Moleculaire et Cellulaire (IPMC), Valbonne, FRANCE

[*]alexandra.mariela.popa@gmail.com

January 19, 2017

Contents

1 Introduction

Ribosome profiling, the recently developed high throughput sequencing technique, enabled the mapping of translated regions genome-wide. This technique takes advantage of the fact that ribosomes actively engaged in translation can protect their associated mRNA fragments against RNAse digestion. The sequencing of these protected fragments can reveal potentially translated sequences. *RiboProfiling* is a Bioconductor package that provides multiple types of ribo-seq data analysis, starting from BAM files alone:

- Quality assessment of read match size distribution.
- Read coverage around the TSS for the definition of an offset (shift value) of ribosome position.
- Calibrate reads by applying an offset to the read start positions along the transcript.
- Table of read start counts (shifted if specified) on the specified region (CDS, 5pUTR, 3pUTR).
- A graphical quality function for ribo-seq data, based on the previously obtained tables.
- The quantification of the frequency and coverage matrices for motifs of codons (1, 2 or 3 codons) in ORFs.
- Principal component analysis of codon motif coverage in ribosome-profiling.

The package also provides the `riboSeqFromBAM` function, that assembles in a global framework some of the above analyses, allowing a quick overview of the data.

2 Data Input

As input data, the *RiboProfiling* package can start with a path to a ribo-seq (or other type of) BAM to be analyzed. A list of multiple BAM files can also be provided. BAM files from either bowtie/tophat, STAR, and Lifescope (Solid), single- and paired-end have been validated. Many intermediate functions of the *RiboProfiling* package can be called for additional modifications on the data objects.

Example data files provided with the *RiboProfiling* package are based on chromosome 1 reads from human primary BJ fibroblasts data (PMID: 23594524, SRA: http://www.ebi.ac.uk/ena/data/view/SRP020544):

- BAM file Ctrl "http://genomique.info/data/public/RiboProfiling/ctrl.bam".
- BAM file Nutlin2h "http://genomique.info/data/public/RiboProfiling/nutlin2h.bam".

The data provided with the *RiboProfiling* package and accessible through the `data` function consists of:

- **ctrlGAlignments** - a GAlignments object corresponding to the "http://genomique.info/data/public/RiboProfiling/ctrll.bam" BAM.
- **codonIndexCovCtrl** - a list containing the number of reads for each codon in each ORF.
- **codonDataCtrl** - a list of 2 data.frames containing the frequency of codons for each ORF and, respectively, the read coverage for the same codons and ORFs.

3 Quick start

The fastest way to analyze a list of BAM files with the *RiboProfiling* package is to call the `riboSeqFromBAM` function. The only input needed are the paths to the BAM files and the genome version on which the mapping was done. Here is an example on how to use this wrapper function on 2 BAM files.

```
library(RiboProfiling)
listInputBam <- c(
    BamFile("http://genomique.info/data/public/RiboProfiling/ctrl.bam"),
    BamFile("http://genomique.info/data/public/RiboProfiling/nutlin2h.bam")
```

```
    )
covData <- riboSeqFromBAM(listInputBam, genomeName="hg19")
```

The following analyses and results are returned:

- **a histogram of read match lengths**: figures **??** and **??**.
- **a read start coverage plot around the TSS**: figures **??** and **??**.
- **an automatic estimation and application of the ribosome offset position (if no offset value is provided (listShiftValue parameter missing))**
- **a data.frame containing CDSs annotation and counts on the 5pUTR, CDS, and 3pUTR once the offset has been applied**
- **pairs and boxplots of the read counts**: figures **??** and **??**.
- **a list of per ORF per codon coverage**

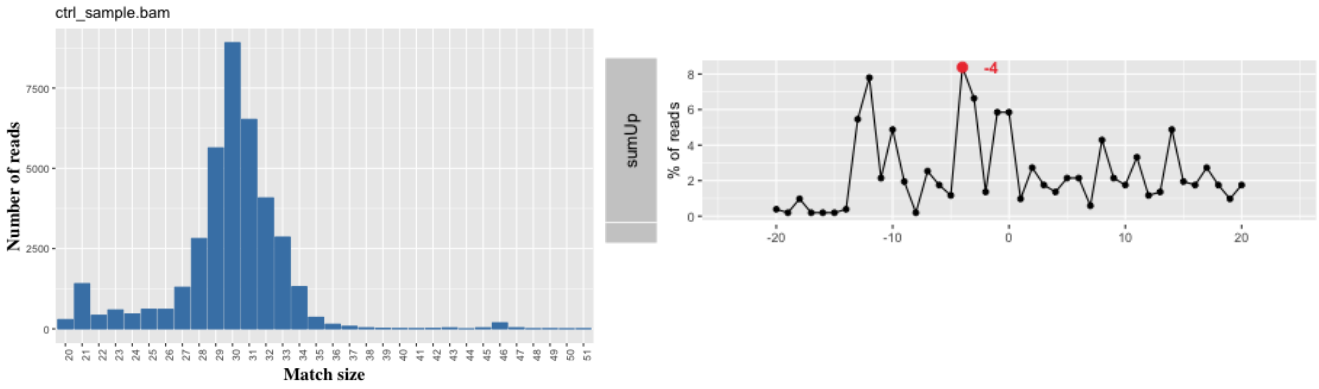

Figure 1: **Histogram of read match size and read start offset to ribosome P-site for ctrl.bam dataset.**

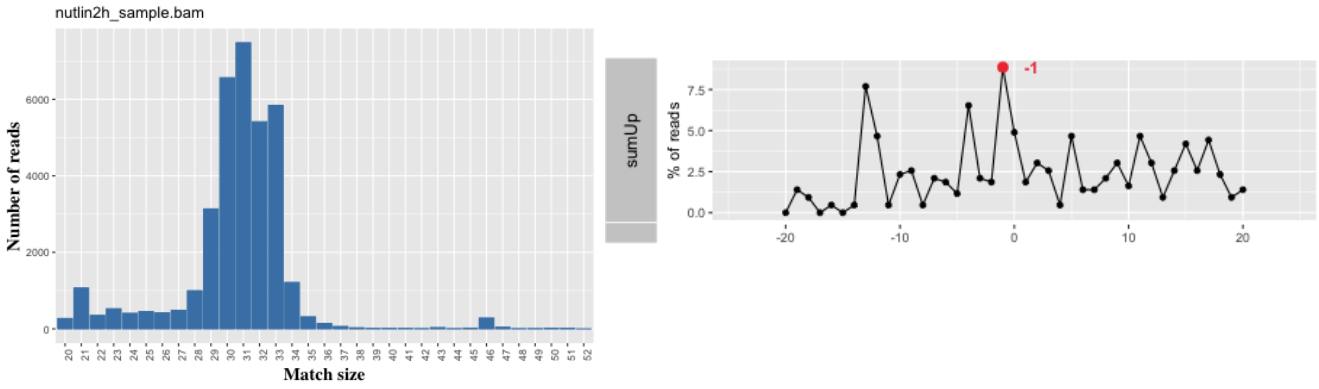

Figure 2: **Histogram of read match size and read start offset to ribosome P-site for nutlin2h.bam dataset.**

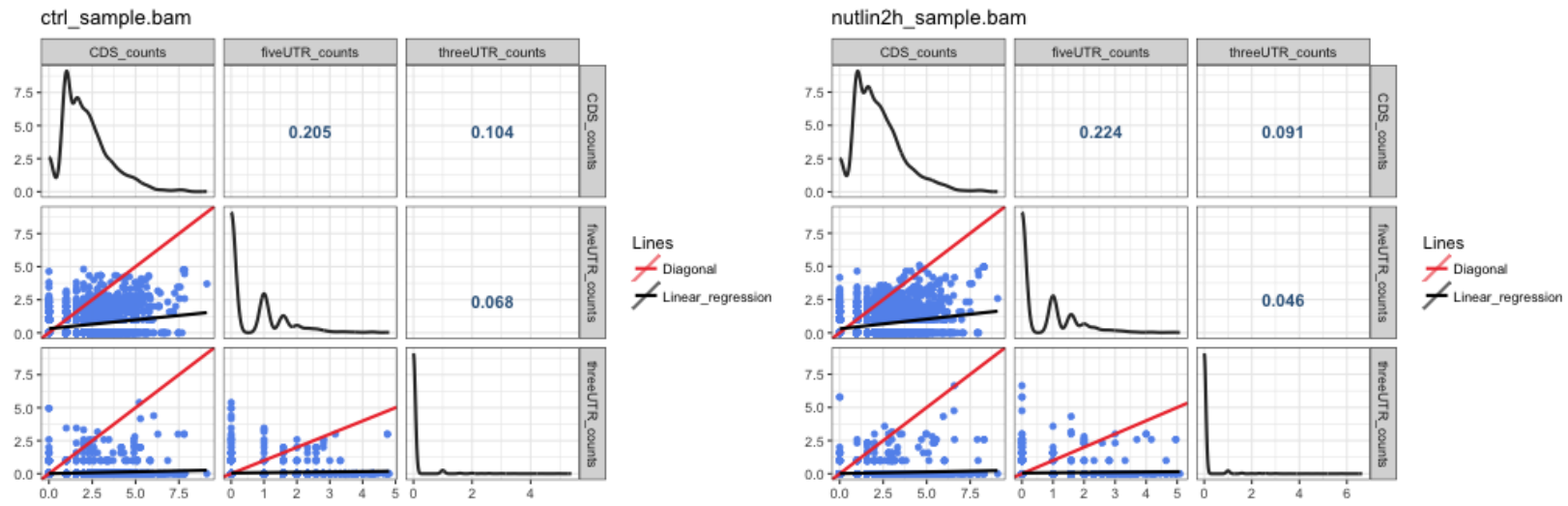

Figure 3: **Pairs of shifted read counts on 5pUTR, CDS, and 3pUTR for ctrl.bam and nutlin2h.bam dataset.**

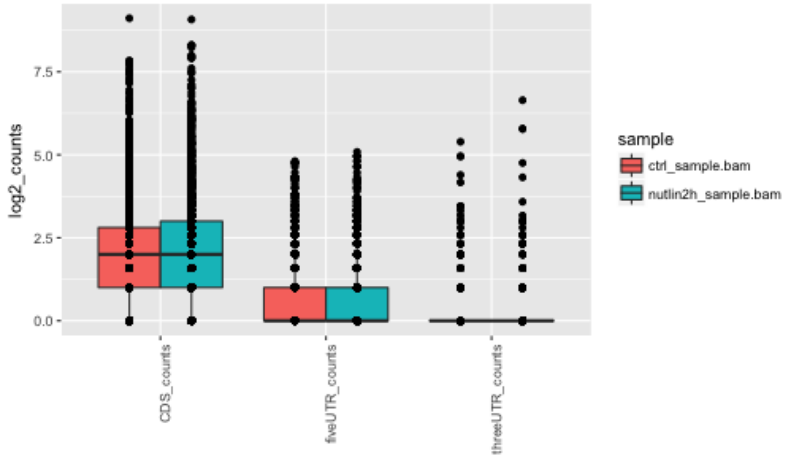

Figure 4: **Boxplot of shifted read counts on 5pUTR, CDS, and 3pUTR for ctrl.bam and nutlin2h.bam dataset.**

All these results will be explained in detail in the following sections.

This general function only works if your species model has an ensembl table in the UCSC database or if you provide the corresponding txdb object containing the annotations.

4 Histogram of read match length

This histogram represents both a quality control of the sequencing and an important tool to define the match sizes of reads corresponding to ribosome footprints (around 30bp). The *histMatchLength* function in the *RiboProfiling* package produces this histogram starting from a GAlignment object (the loaded records of a BAM) such as the *ctrlGAlignments* data example (figure **??**).

One can create a GAlignments from a BAM using the `readGAlignments` function from the *GenomicAlignments* package:

```
aln <- readGAlignments(
       BamFile("http://genomique.info/data/public/RiboProfiling/ctrl.bam")
```

```
    )
```

Or based on an already existing GAlignments object:

```
data(ctrlGAlignments)
aln <- ctrlGAlignments
matchLenDistr <- histMatchLength(aln, 0)
matchLenDistr[[2]]
```

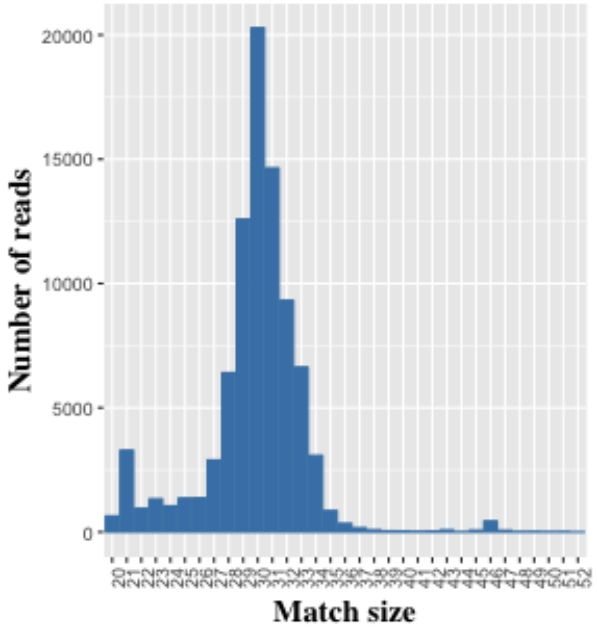

Figure 5: **Histogram of read match length for example data: ctrlGAlignments.**

5 Read start coverage plot around the TSS

The initiation of protein synthesis starts with the first codon in the P-site. In ribosome profiling experiments, the location of the P-site codon must be inferred in order to recalibrate the read start positions relative to the transcript. An offset is usually observed in Ribo-seq experiments between the starting of the reads and the AUG codons of protein coding sequences. Three functions allow the estimation of the read start position relative to the P-site codon:

- `aroundPromoter`: returns the genomic positions flanking the transcript start site (TSS) for the x% (3% default value) best expressed CDSs.
- `readStartCov`: returns the read start coverage around the TSS on the predefined CDSs.
- `plotSummarizedCov`: plots the summarized coverage in a specified range (e.g. around TSS) for the specified match sizes.

```
#transform the GAlignments object into a GRanges object
#(faster processing of the object)
alnGRanges <- readsToStartOrEnd(aln, what="start")
#txdb object with annotations
txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
oneBinRanges <- aroundPromoter(txdb, alnGRanges, percBestExpressed=0.001)
#the coverage in the TSS flanking region for the reads with match sizes 29:31
listPromoterCov <-
```

```
    readStartCov(
        alnGRanges,
        oneBinRanges,
        matchSize=c(29:31),
        fixedInterval=c(-20, 20),
        renameChr="aroundTSS",
        charPerc="perc"
    )
plotSummarizedCov(listPromoterCov)
```

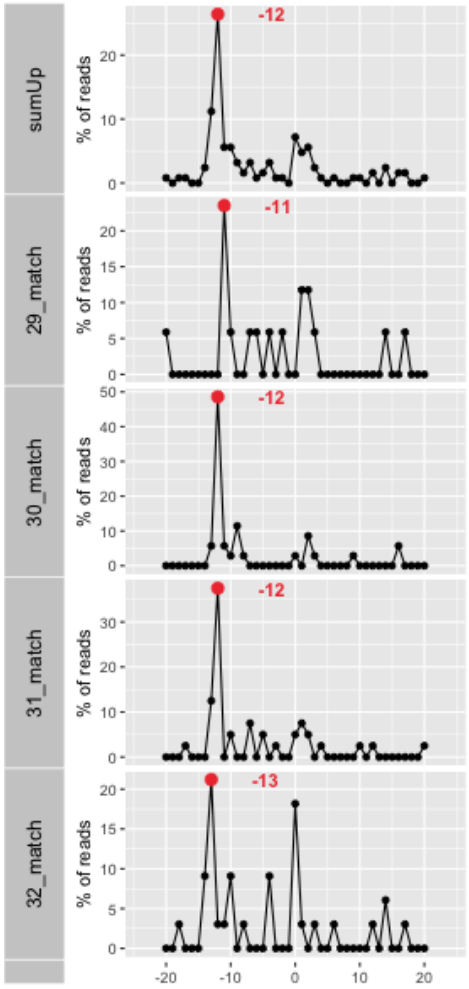

Figure 6: **Offset around TSS for the best expressed CDSs on a subset of match sizes.**

The first time you analyse a ribo-seq BAM file, it is advisable to make an offset graph for all the previously selected read match sizes, as the offset might be different depending on the read match size .

6 Count reads on features

The purpose of Ribosome Profiling experiments is mainly to identify RNAse resistant regions, which might indicate that these sequences are likely to be translated. The `countShiftReads` function quantifies the read start coverage on different sequence features. This function integrates the specificity of ribo-seq data of having shifted reads starts from the P-site codon. The `countShiftReads` function recalibrates the read start positions along the transcript with the specified offset (parameter *shiftValue*, default 0) and returns the coverage on the 5pUTR, CDS, 3pUTR, and a matrix of codon coverage for each ORF.

The counts can be used globally to see what is RNAse protected and what is not, but also for differential analysis between conditions. The `countsPlot` function provides some quality graphs for the ribo-seq data: pairs between counts on features and boxplots of counts between conditions.

As an example, we compute the read coverage on the **ctrlGAlignments** example data:

```
#keep only the match read sizes 30-33
alnGRanges <- alnGRanges[which(!is.na(match(alnGRanges$score,30:33)))]
#get all CDSs by transcript
cds <- cdsBy(txdb, by="tx", use.names=TRUE)
#get all exons by transcript
exonGRanges <- exonsBy(txdb, by="tx", use.names=TRUE)
#get the per transcript relative position of start and end codons
cdsPosTransc <- orfRelativePos(cds, exonGRanges)
#compute the counts on the different features
#after applying the specified shift value on the read start along the transcript
countsDataCtrl1 <-
    countShiftReads(
        exonGRanges=exonGRanges[names(cdsPosTransc)],
        cdsPosTransc=cdsPosTransc,
        alnGRanges=alnGRanges,
        shiftValue=-14
    )
head(countsDataCtrl1[[1]])
listCountsPlots <- countsPlot(
    list(countsDataCtrl1[[1]]),
    grep("_counts$", colnames(countsDataCtrl1[[1]])),
    1
)
listCountsPlots
```

```
## Warning in countShiftReads(exonGRanges[names(cdsPosTransc)], cdsPosTransc, :  Param motifSize
should be an integer!  Accepted values 3, 6 or 9.  Default value is 3.

## Warning:  Ignoring unknown aesthetics:  fill

##                 gene  chr strand transc_genomic_start transc_genomic_end
## uc001abw.1 uc001abw.1 chr1      +               861121             879961
## uc031pjl.1 uc031pjl.1 chr1      +               861302             879533
## uc031pjm.1 uc031pjm.1 chr1      +               861302             879533
##            transc_length orf_start orf_end orf_length CDS_counts fiveUTR_counts
## uc001abw.1          2554        81    2126       2046          2              0
## uc031pjl.1          2120        21    2120       2100          2              0
## uc031pjm.1          2084        21    2084       2064          2              0
##            threeUTR_counts
## uc001abw.1               0
## uc031pjl.1               0
```

```
## uc031pjm.1                0
```

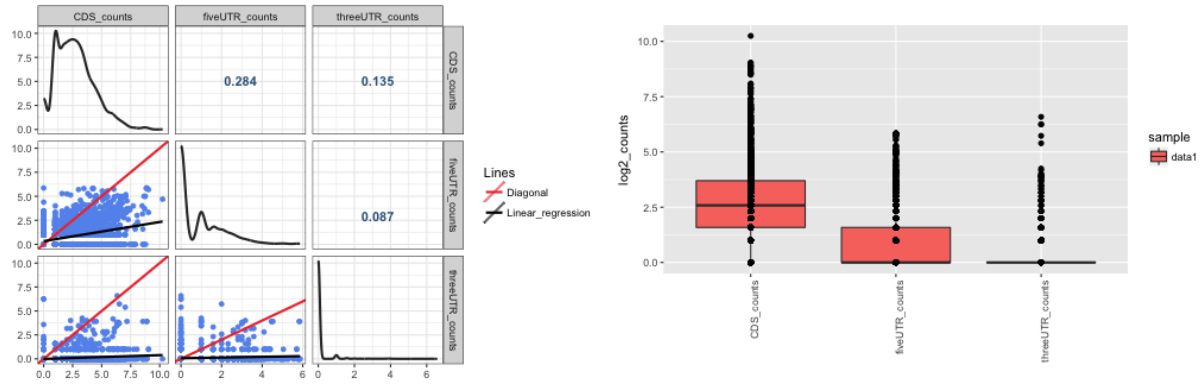

Figure 7: **Pairs and boxplots of read coverage on genomic features.**

7 Count reads on codon motifs

The previous `countShiftReads` function produces a second element in the list: the counts of reads per motifs of 1, 2, or 3 codons. For each ORF, for each motif in the ORF, the Ribo-seq coverage of the motif is reported as follows:

- for motifs of 3 nucleotides (1 codon) - the sum of read starts on the 3 nucleotides in the codon is reported.
- consecutive motifs of 6 nucleotides (2 consecutive codons) overlap on 3 nucleotides. The Ribo-seq coverage is reported as the coverage on the 1st codon in the motif considered as being in the P-site.
- consecutive motifs of 9 nucleotides (3 consecutive codons) overlap on 6 nucleotides. The Ribo-seq coverage is reported as the coverage on the 2nd codon in the motif considered as being in the P-site.

Codon motifs in each ORF are described as the index of their position in the ORF.

Here is an example:

```
data(codonIndexCovCtrl)
head(codonIndexCovCtrl[[1]], n=3)
```

```
##   codonID nbrReads
## 1       1        0
## 2       2        0
## 3       3        0
```

The `codonInfo` function associates the read coverage on codons with their corresponding codon type for each ORF. It returns a list of 2 matrices: one containing the frequency of each codon motif type in each ORF and the second containing the coverage of each codon motif type in each ORF.

```
listReadsCodon <- countsDataCtrl1[[2]]
#get the names of the expressed ORFs grouped by transcript
cds <- cdsBy(txdb, use.names=TRUE)
orfCoord <- cds[names(cds) %in% names(listReadsCodon)]
#chromosome names should correspond between the BAM,
#the annotations, and the genome
genomeSeq <- BSgenome.Hsapiens.UCSC.hg19
#codon frequency, coverage, and annotation
codonData <- codonInfo(listReadsCodon, genomeSeq, orfCoord)
```

```
## Warning in codonInfo(listReadsCodon, genomeSeq, orfCoord):  Param motifSize should be an integer!
Accepted values 3, 6 or 9.  Default value is 3.
```

```
## Warning:  Quoted identifiers should have class SQL, use DBI::SQL() if the caller performs the
quoting.
```

```
data("codonDataCtrl")
head(codonDataCtrl[[1]], n=3)
```

```
##             AAA AAC AAG AAT ACA ACC ACG ACT AGA AGC AGG AGT ATA ATC ATG ATT CAA
## uc010nxq.1   1   2   1   0   1   2   2   2   2   3   6   6   0   3   3   0   0
## uc001abv.1   0   4   6   0   1   3   0   1   3   8   1   1   2   5   3   1   0
## uc001abw.1   1   8  18   2   4  13   7   4   3  21  11   2   2   8   8   1   4
##             CAC CAG CAT CCA CCC CCG CCT CGA CGC CGG CGT CTA CTC CTG CTT GAA GAC
## uc010nxq.1   7   5   2   3   9   0   4   0   0   0   0   1   3   8   2   1   3
## uc001abv.1   5   5   2   1   5   4   2   4   3   5   1   0   2   6   1   1   6
## uc001abw.1  16  27   6  14  45  13  12   8  10  21   3   1  13  56   5   5  22
##             GAG GAT GCA GCC GCG GCT GGA GGC GGG GGT GTA GTC GTG GTT TAA TAC TAG
## uc010nxq.1   6   1   1   2   1   2   3   2   1   3   0   1   5   2   0   1   1
## uc001abv.1   5   2   1   5   1   0   0   6   4   2   0   3   3   1   0   0   0
## uc001abw.1  41   7   7  38  11  11   7  30  26   7   1   9  12   3   0   6   0
##             TAT TCA TCC TCG TCT TGA TGC TGG TGT TTA TTC TTG TTT
## uc010nxq.1   0   0   3   0   3   0   3   0   3   0   3   2   3
## uc001abv.1   0   0   5   0   2   0   4   1   2   0   3   1   0
## uc001abw.1   2   4  17   6   4   1   8   5   4   0  14   5   2
```

The codon coverage can be used to study the ribosome translation dynamics. One can test if codon motifs accumulating ribo-seq reads represent a stalling in the ribosome progression. In the *RiboProfiling* package we have implemented the `codonPCA` function that performs a PCA analysis on a matrix of read density on codons (figures **??**).

```
codonUsage <- codonData[[1]]
codonCovMatrix <- codonData[[2]]
#keep only genes with a minimum number of reads
nbrReadsGene <- apply(codonCovMatrix, 1, sum)
ixExpGenes <- which(nbrReadsGene >= 50)
codonCovMatrix <- codonCovMatrix[ixExpGenes, ]
#get the PCA on the codon coverage
codonCovMatrixTransp <- t(codonCovMatrix)
rownames(codonCovMatrixTransp) <- colnames(codonCovMatrix)
colnames(codonCovMatrixTransp) <- rownames(codonCovMatrix)
listPCACodonCoverage <- codonPCA(codonCovMatrixTransp, "codonCoverage")
listPCACodonCoverage[[2]]
```

```
## $`PC_1-2`
##
## $`PC_1-3`
##
## $`PC_1-4`
##
## $`PC_2-3`
##
## $`PC_2-4`
```

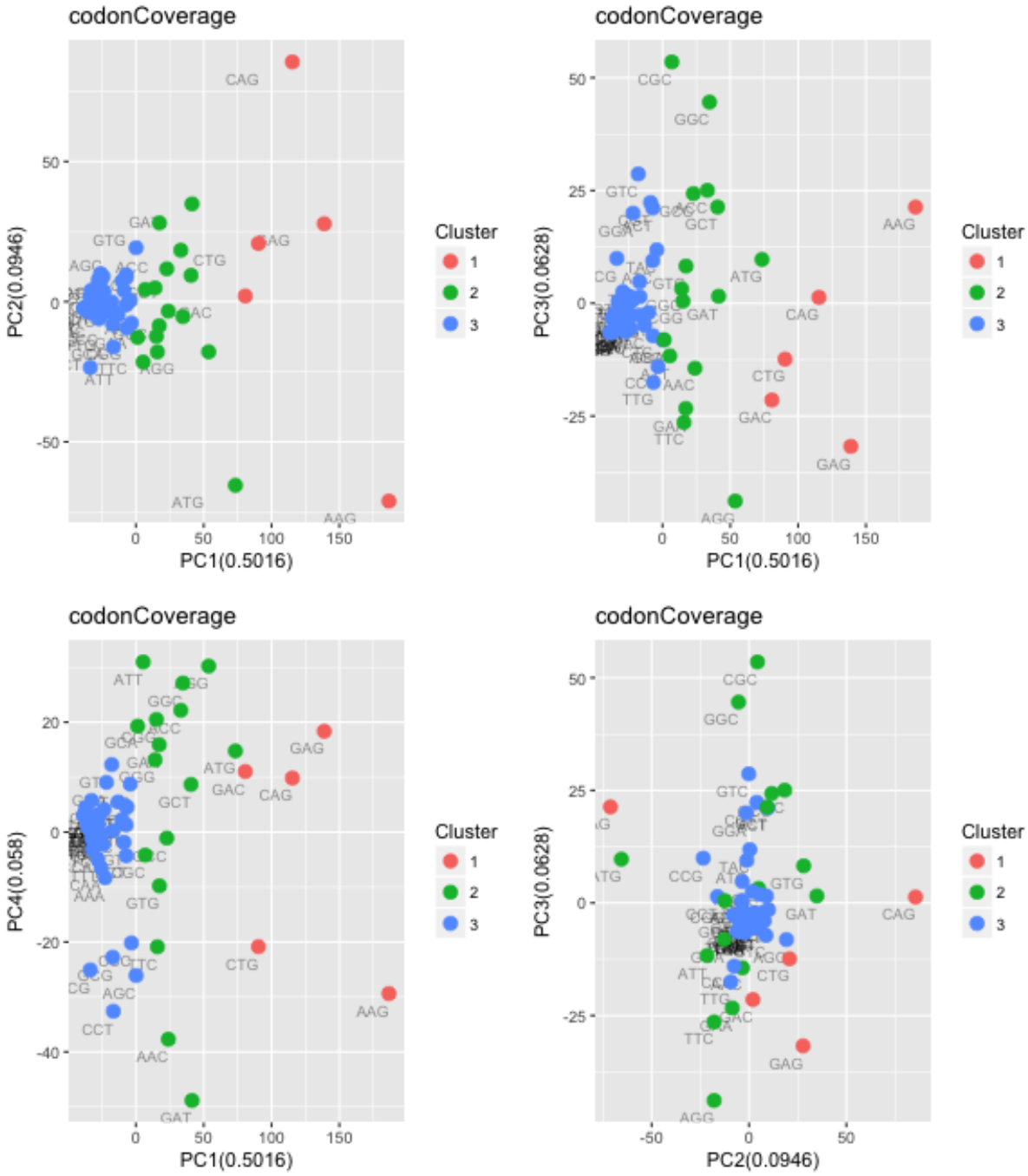

Figure 8: **PCA on codon coverage for the ctrlGAlignments example data.**

```
sessionInfo()
```

```
## R version 3.3.2 (2016-10-31)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: OS X Mavericks 10.9.5
##
## locale:
## [1] C/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
```

```
##  [1] tcltk     stats4    parallel  stats     graphics  grDevices utils
##  [8] datasets  methods   base
##
## other attached packages:
## [1] RSQLite_1.1-2       RiboProfiling_1.4.1 Biostrings_2.42.1
## [4] XVector_0.14.0      IRanges_2.8.1       S4Vectors_0.12.1
## [7] BiocGenerics_0.20.0 knitr_1.15.1
##
## loaded via a namespace (and not attached):
##  [1] Biobase_2.34.0
##  [2] httr_1.2.1
##  [3] AnnotationHub_2.6.4
##  [4] splines_3.3.2
##  [5] gsubfn_0.6-6
##  [6] Formula_1.2-1
##  [7] shiny_1.0.0
##  [8] assertthat_0.1
##  [9] interactiveDisplayBase_1.12.0
## [10] highr_0.6
## [11] latticeExtra_0.6-28
## [12] RBGL_1.50.0
## [13] BSgenome_1.42.0
## [14] Rsamtools_1.26.1
## [15] yaml_2.1.14
## [16] backports_1.0.5
## [17] lattice_0.20-34
## [18] biovizBase_1.22.0
## [19] chron_2.3-49
## [20] digest_0.6.11
## [21] GenomicRanges_1.26.2
## [22] RColorBrewer_1.1-2
## [23] checkmate_1.8.2
## [24] colorspace_1.3-2
## [25] ggbio_1.22.3
## [26] htmltools_0.3.5
## [27] httpuv_1.3.3
## [28] Matrix_1.2-7.1
## [29] plyr_1.8.4
## [30] OrganismDbi_1.16.0
## [31] XML_3.98-1.5
## [32] biomaRt_2.30.0
## [33] zlibbioc_1.20.0
## [34] xtable_1.8-2
## [35] scales_0.4.1
## [36] BiocParallel_1.8.1
## [37] htmlTable_1.8
## [38] tibble_1.2
## [39] ggplot2_2.2.1
## [40] sqldf_0.4-10
## [41] SummarizedExperiment_1.4.0
## [42] GenomicFeatures_1.26.2
## [43] nnet_7.3-12
## [44] lazyeval_0.2.0
```

```
## [45] proto_1.0.0
## [46] survival_2.40-1
## [47] magrittr_1.5
## [48] mime_0.5
## [49] memoise_1.0.0
## [50] evaluate_0.10
## [51] GGally_1.3.0
## [52] foreign_0.8-67
## [53] BSgenome.Hsapiens.UCSC.hg19_1.4.0
## [54] graph_1.52.0
## [55] BiocInstaller_1.24.0
## [56] tools_3.3.2
## [57] data.table_1.10.0
## [58] BiocStyle_2.2.1
## [59] stringr_1.1.0
## [60] munsell_0.4.3
## [61] cluster_2.0.5
## [62] AnnotationDbi_1.36.1
## [63] ensembldb_1.6.2
## [64] GenomeInfoDb_1.10.2
## [65] grid_3.3.2
## [66] RCurl_1.95-4.8
## [67] dichromat_2.0-0
## [68] VariantAnnotation_1.20.2
## [69] labeling_0.3
## [70] bitops_1.0-6
## [71] base64enc_0.1-3
## [72] gtable_0.2.0
## [73] DBI_0.5-1
## [74] reshape_0.8.6
## [75] reshape2_1.4.2
## [76] R6_2.2.0
## [77] GenomicAlignments_1.10.0
## [78] gridExtra_2.2.1
## [79] rtracklayer_1.34.1
## [80] Hmisc_4.0-2
## [81] TxDb.Hsapiens.UCSC.hg19.knownGene_3.2.2
## [82] stringi_1.1.2
## [83] Rcpp_0.12.9
## [84] rpart_4.1-10
## [85] acepack_1.4.1
```