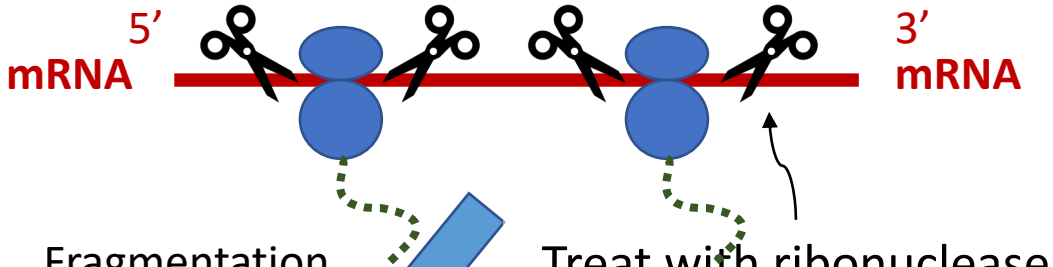
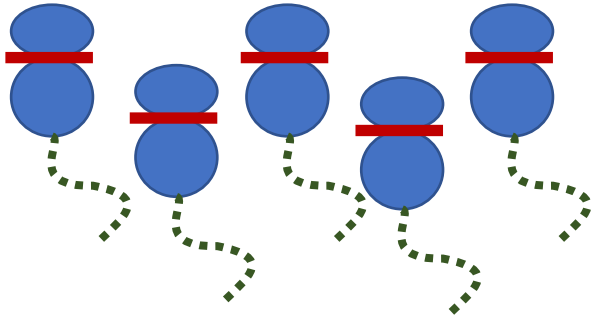
ribosomeProfilingQC

Jianhong Ou
Mariah Hoye



Fragmentation

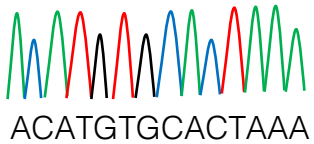
Treat with ribonuclease enzyme



Remove rRNA

Ribosome Protected Fragments (RPFs)

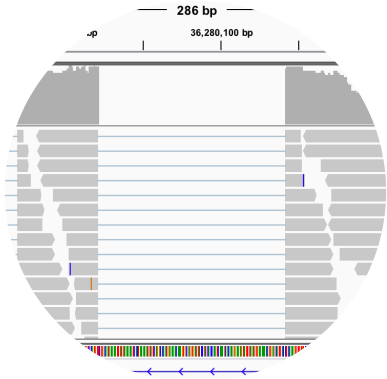
Determine the sequence of RPF



FASTQ Files

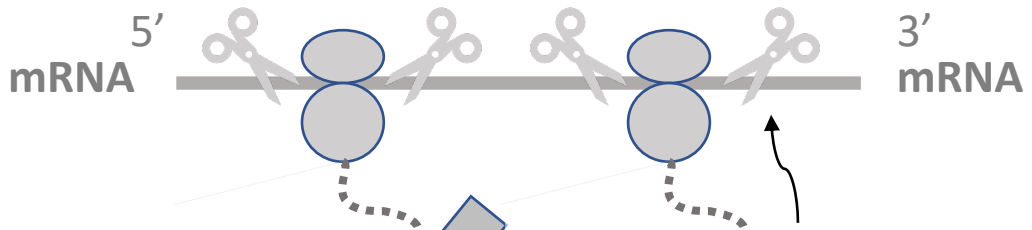
RIBOSOME PROFILING

ribosomeProfilingQC



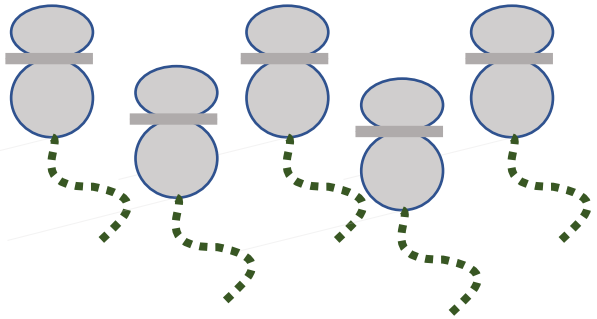
Map to the genome

Questions in QC



Fragmentation

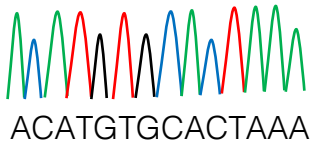
Treat with ribonuclease enzyme



Remove rRNA

Ribosome protected fragments (RPF)

Determine the sequence of RPF



FASTQ Files

1. Does RPFs preparation succeed?

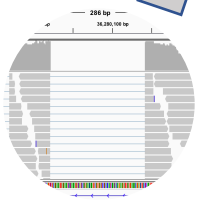
S: ribosomeProfilingQC

K: fragment size (around 28 nt.)
strand specifics (sense or antisense)
exon occupation (exon enriched)
whole transcript coverage
reading frame

2. Does the sequencing succeed?

S: FastQC & QualiMap 2

K: sequencing score (or mapping score)
library complexity (duplicates rate)



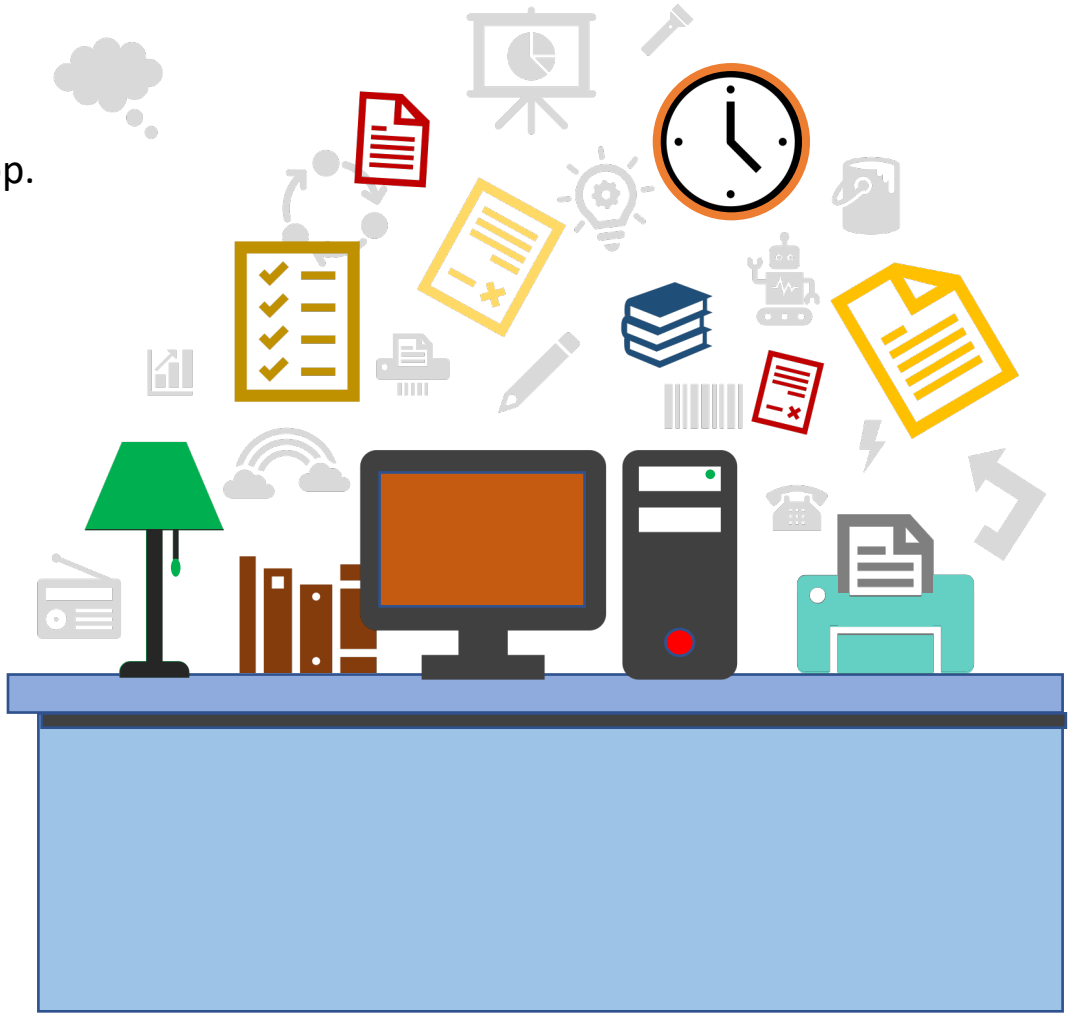
Map to the genome

Downstream analysis and Validation

Differential analysis:

- translation level
- codon level
- alternative splicing
- polyadenylation usage
- transcription start site

Main function list



Estimate P site

The *estimatePsite* function will search start/stop codons that occur in the reads. It can search best P site candidates from the 5’ end and 3’ end.

Bar plots

readsEndPlot: the end reads shifted from the start/stop.
summaryReadsLength: the fragment size distribution.
strandPlot: the strand percentage.
readsDistribution: genomic elements distribution.
plotFrameDensity: collapse all the RPFs in each frame.

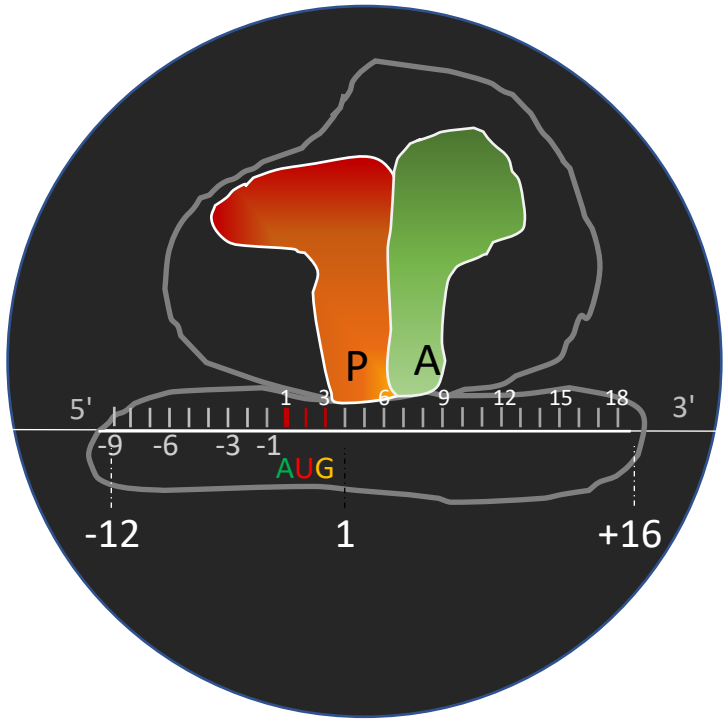
Line plots

The *metaPlot* function can indicate the reads distribution in 5’UTR, CDS and 3’UTR region.
The *plotTranscript* can be used to view the reading frame distribution for given transcripts.

Count table

The *countReads* function can be used to count multiple files of ribo-seq and RNA-seq data.
The *coverageDepth* function will calculate the coverage depth for gene level or transcript level.

Estimate P site position



- As it shown in the left figure, P site of the ribosome is in position 13 (if using RNase I). However, in different experiments, the P site may be shifted due to various experimental conditions such as the choice of enzyme and the cell type. The *estimatePsite* function can be used to check the P site.
- It has been shown that for certain enzymes, such as MNase, estimating the P site from the 3' end works much better. The estimatePsite can search P site from 5' or 3' end as user defined.

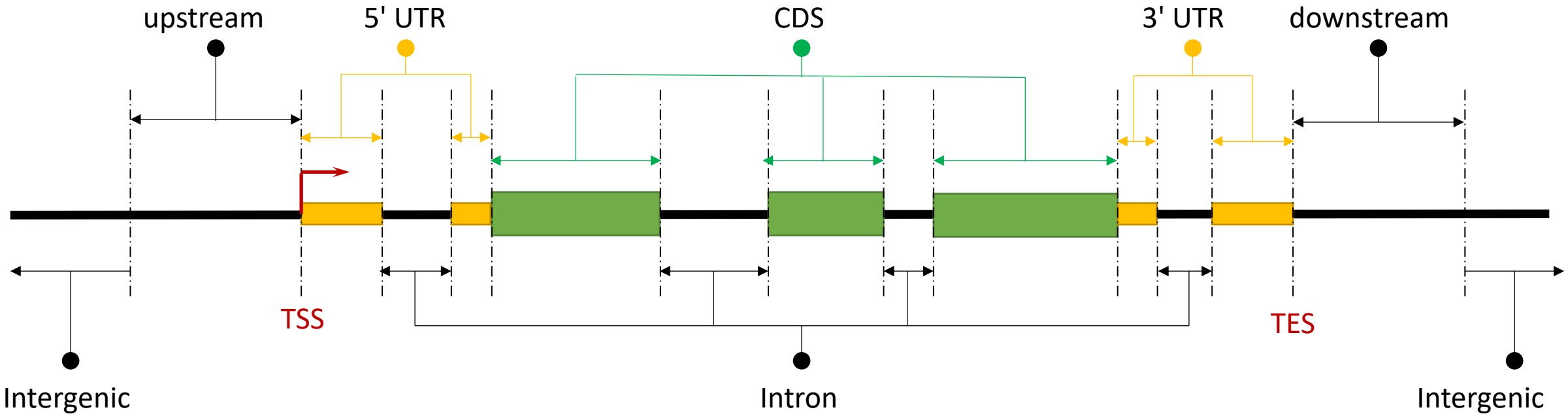
Output of prepareCDS

```
estimatePsite(bamfile, CDS, genome, anchor = "5end")
```

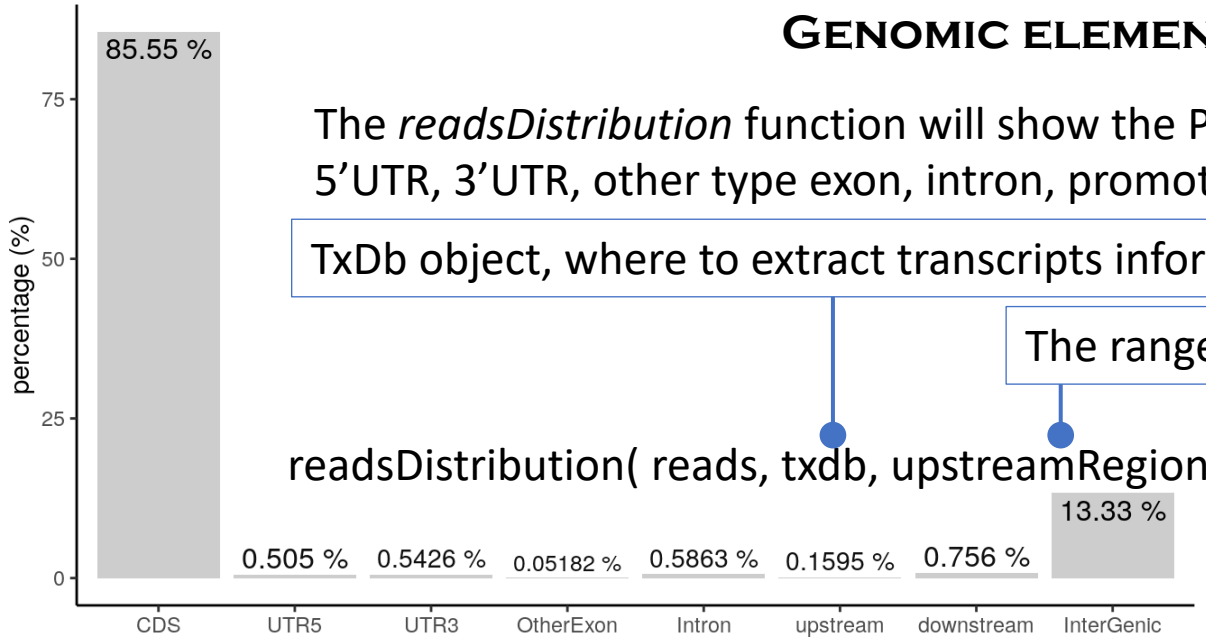
A BamFile object

5end or 3end. Default is 5end

Ref: Fuad et.al., DOI: 10.7554/eLife.42591



Genomic element distribution



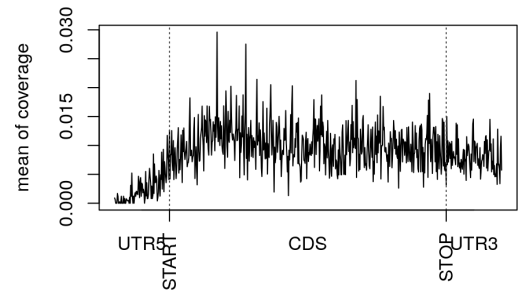
The *readsDistribution* function will show the P site locations in different genomic elements: CDS, 5'UTR, 3'UTR, other type exon, intron, promoter, downstream or intergenic region.

TxDb object, where to extract transcripts information

The range of upstream and downstream region defined in nt.

```
readsDistribution( reads, txdb, upstreamRegion = 3000, downstreamRegion = 3000, plot = TRUE, ... )
```

Metagene analysis plot for 5'UTR/CDS/3'UTR



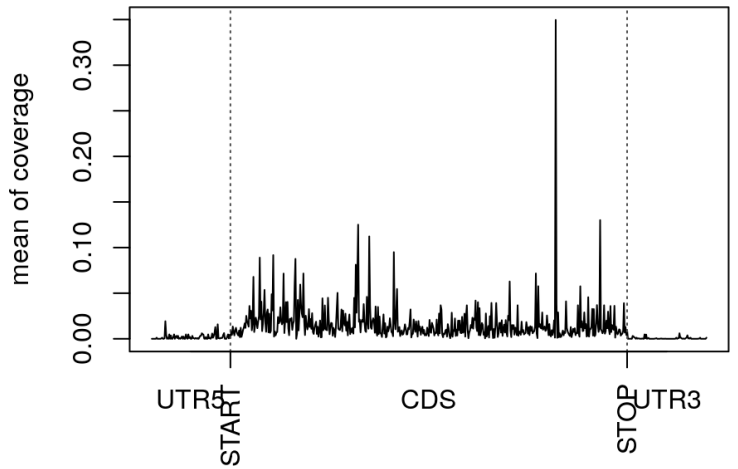
RNA-seq

- A metagene analysis is a quantitative summarization over aligned multiple genomic regions.

- coverageDepth(RPFs, RNAs, gtf,
- level = c("tx", "gene"),
- bestpsite = 13, readsLen = c(28, 29),
- anchor = "5end", region = "cds",
- ext = 5000, ...)

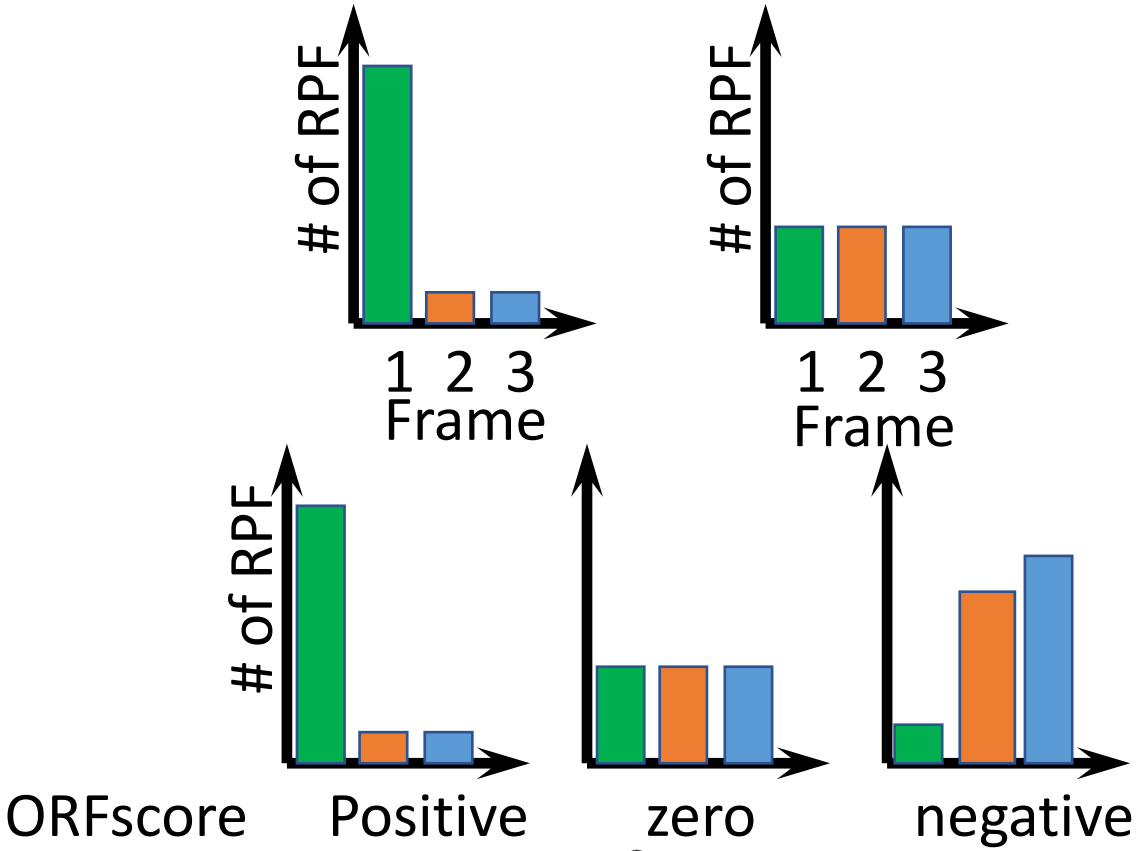
- metaPlot(UTR5coverage, CDScoverage, UTR3coverage,
- sample, xaxis = c("RPFs", "mRNA"),
- bins = c(UTR5 = 100, CDS = 500, UTR3 = 100), ...)

Ribo-seq



ORFscore VS. coverageRate

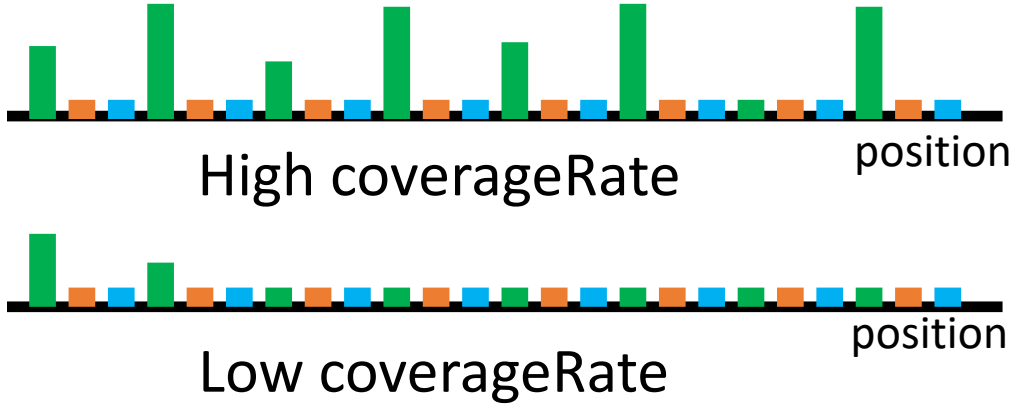
observed value VS. random variable

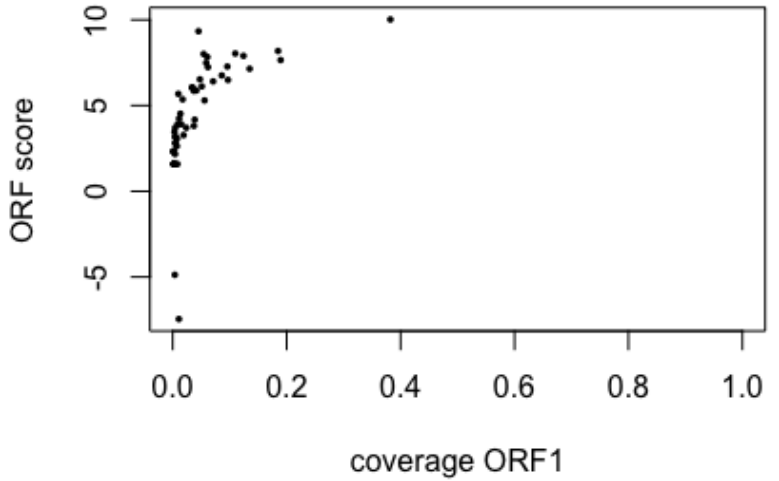


$$ORFscore = log_2((\sum_{n=1}^{3} \frac{(F_i - \bar{F})^2}{\bar{F}}) + 1)$$

If F1 is smaller than F2 or F3, ORF score = −1 * ORF score

(% of in-frame positions with reads)

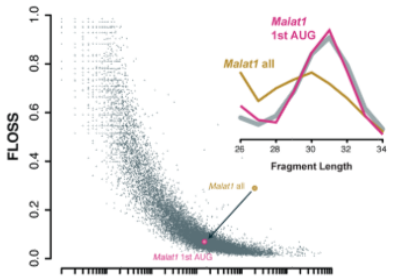





```
cvg <- frameCounts(pc.sub,
                   coverageRate=TRUE)
ORFscore <- getORFscore(pc.sub)
plot(cvg[names(ORFscore)],
     ORFscore)
```

0 <= coverageRate <=1

Modified from Bazzini et.al., doi:10.1002/embj.201488411

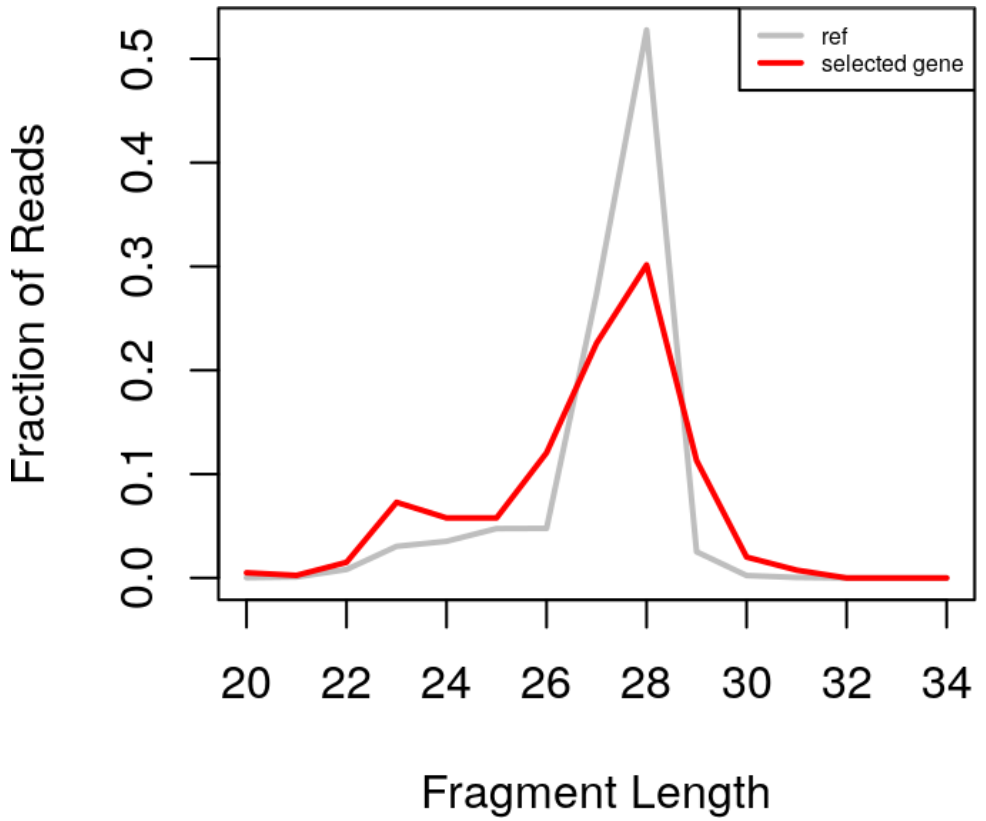


Fragment Length Organization Similarity Score

FLOSS

FLOSS measures the magnitude of disagreement between these two distributions, with lower scores reflecting higher similarity.

$$FLOSS = \sum \| f_{sel}(L) - f_{ref}(L) \|$$



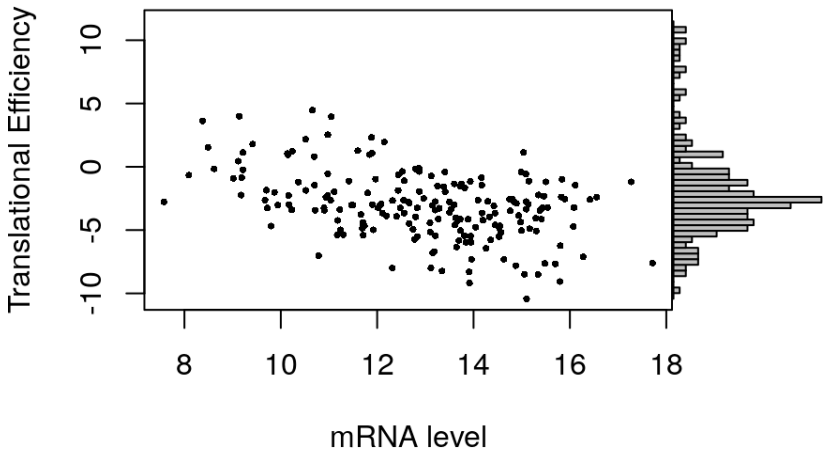
FLOSS(reads, ref, CDS, readLengths = c(26:34), level = c("tx", "gene"), draw = FALSE)

Ref: Nicholas et. al., doi: 10.1016/j.celrep.2014.07.045

	coverageDepth	countReads
Counting method	Coverage for RPFs P-site or coverage for 5'end of RNA-seq reads	findOverlaps for RPFs P-site and featureCounts for RNAseq
output	cvgd object with slot coverage and granges. In coverage slot, there is a list of RPFs and RNAs with RleList for each samples.	A list with element RPFs, RNAs and annotation. Counts table in RPFs or RNAs.
Aims	differential usage of alternative Translation Initiation Sites, alternative Polyadenylation Sites or alternative splicing sites	downstream differential analysis
Downstream	ribosomeProfilingQC:: spliceEvent	edgeR or DESeq2 for RPFs differential analysis. limma for differential Translational Efficiency (TE).

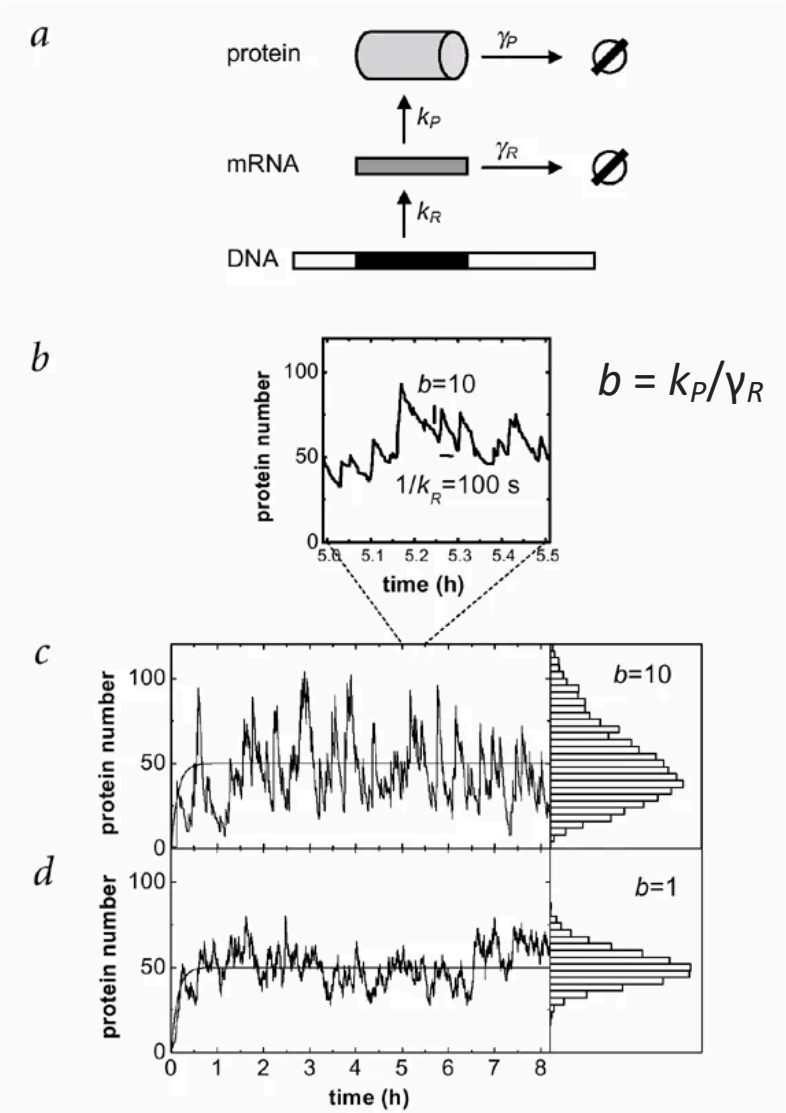
Translational Efficiency (TE)

TE is defined as the ratios of the absolute level of ribosome occupancy divided by RNA levels for transcripts.



Output of getFPKM

translationalEfficiency(x, window,
RPFsampleOrder, mRNAsampleOrder,
pseudocount = 1, log2 = FALSE,
normByLibSize = FALSE)

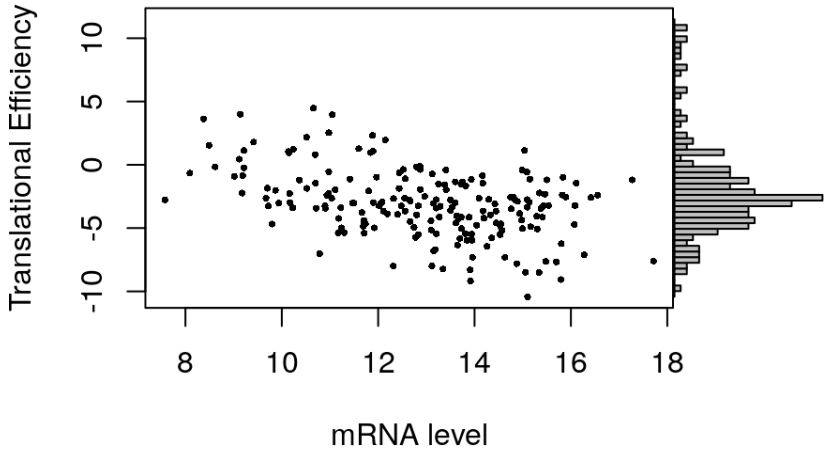


Modified from Ozbudak et al. Nature genetics (2002): 69-73.

Ref: Nicholas et.al., doi: 10.1126/science.1168978
Ozbudak et al. Nature genetics (2002): 69-73.

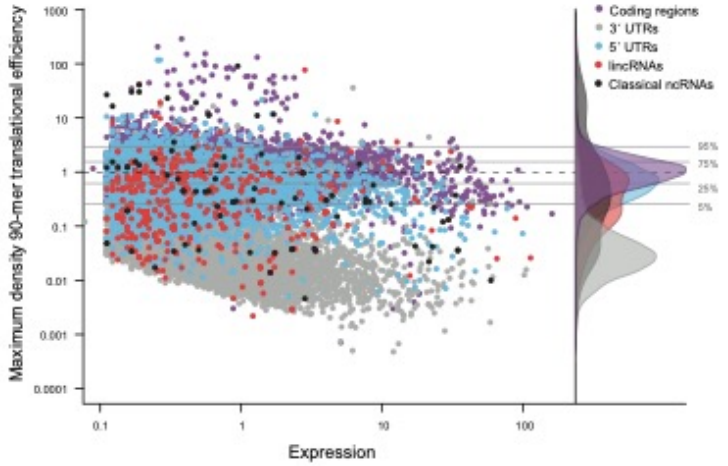
Translational Efficiency (TE)

TE is defined as the ratios of the absolute level of ribosome occupancy divided by RNA levels for transcripts.



Maximum N-mer translational efficiency

TE shows a bias that a higher value in lowly expressed transcripts. This issue can be fixed by calculating the maximum value in the most highly ribosome-occupied 90 nt window within a feature.



Output of getFPKM. if window is set, it must be output of coverageDep

Set to calculate maximum N-mer TE, otherwise, TE

```
translationalEfficiency( x, window,
                         RPFsampleOrder, mRNAsampleOrder,
                         pseudocount = 1, log2 = FALSE, normByLibSize = FALSE )
```


Ref: Nicholas et.al., doi: 10.1126/science.1168978

Mitchell et.al., doi: 10.1016/j.cell.2013.06.009

ribosomeProfilingQC CAN ...

help researcher to assess the quality of the RPFs library preparation by simple plots.

generate counts for downstream analysis and validation.

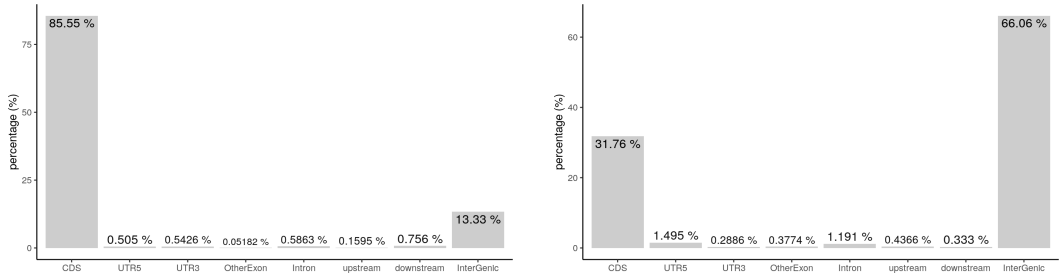
Downstream analysis and Validation

Differential analysis:

- translation level
- codon level
- alternative splicing
- polyadenylation usage
- transcription start site

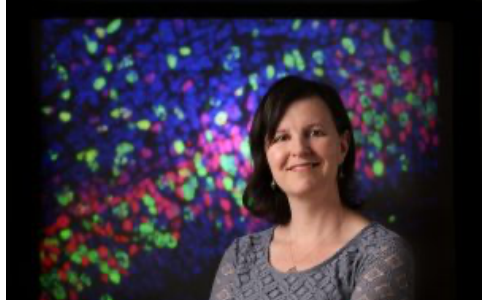
strand specifics?

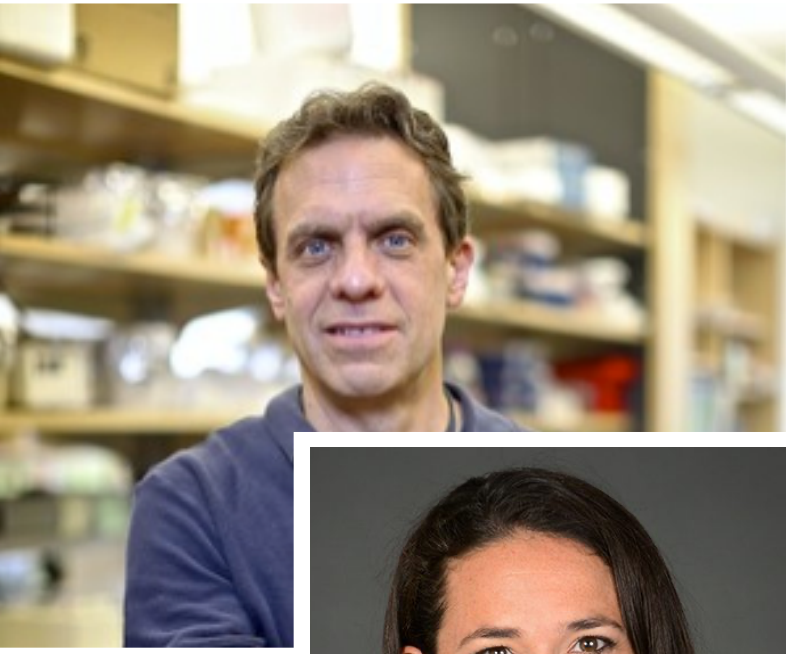
CDS enriched?

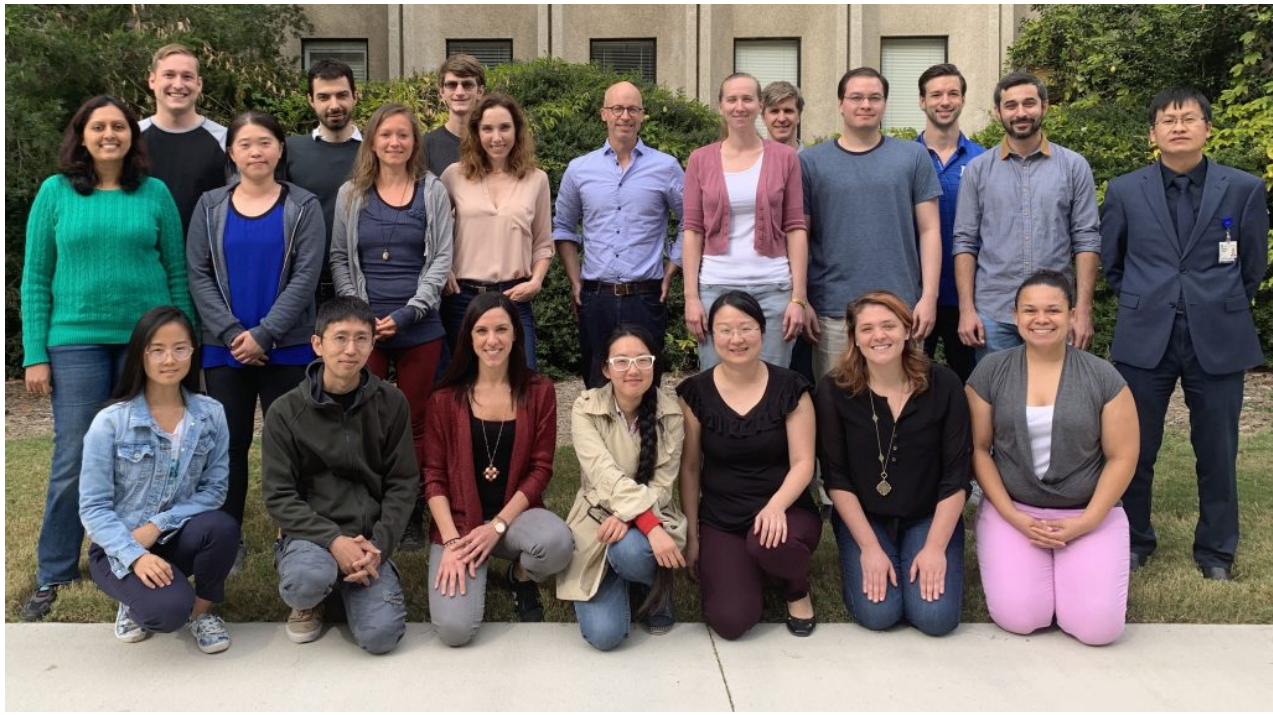


reading frame shifted?

...







Acknowledgement