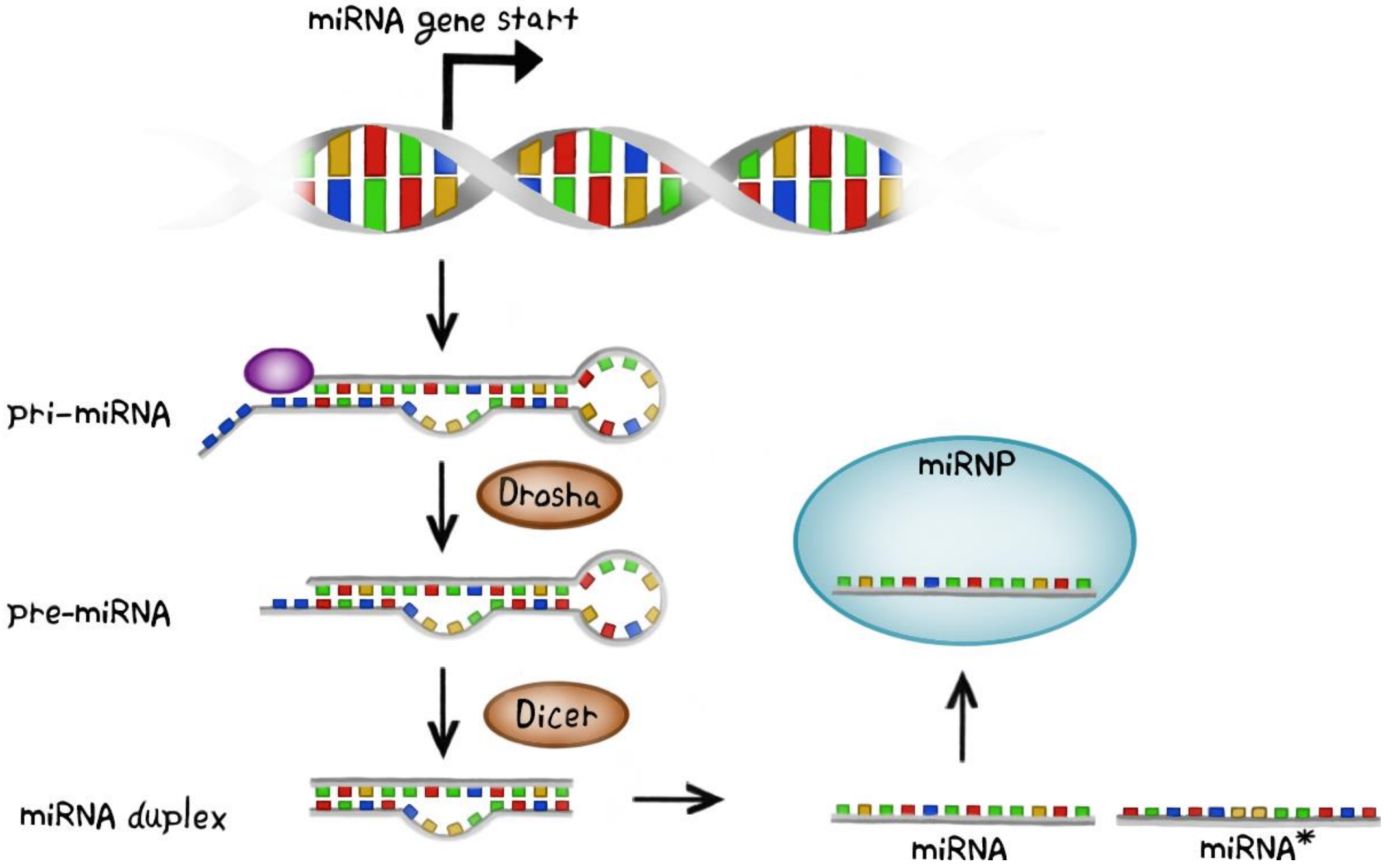
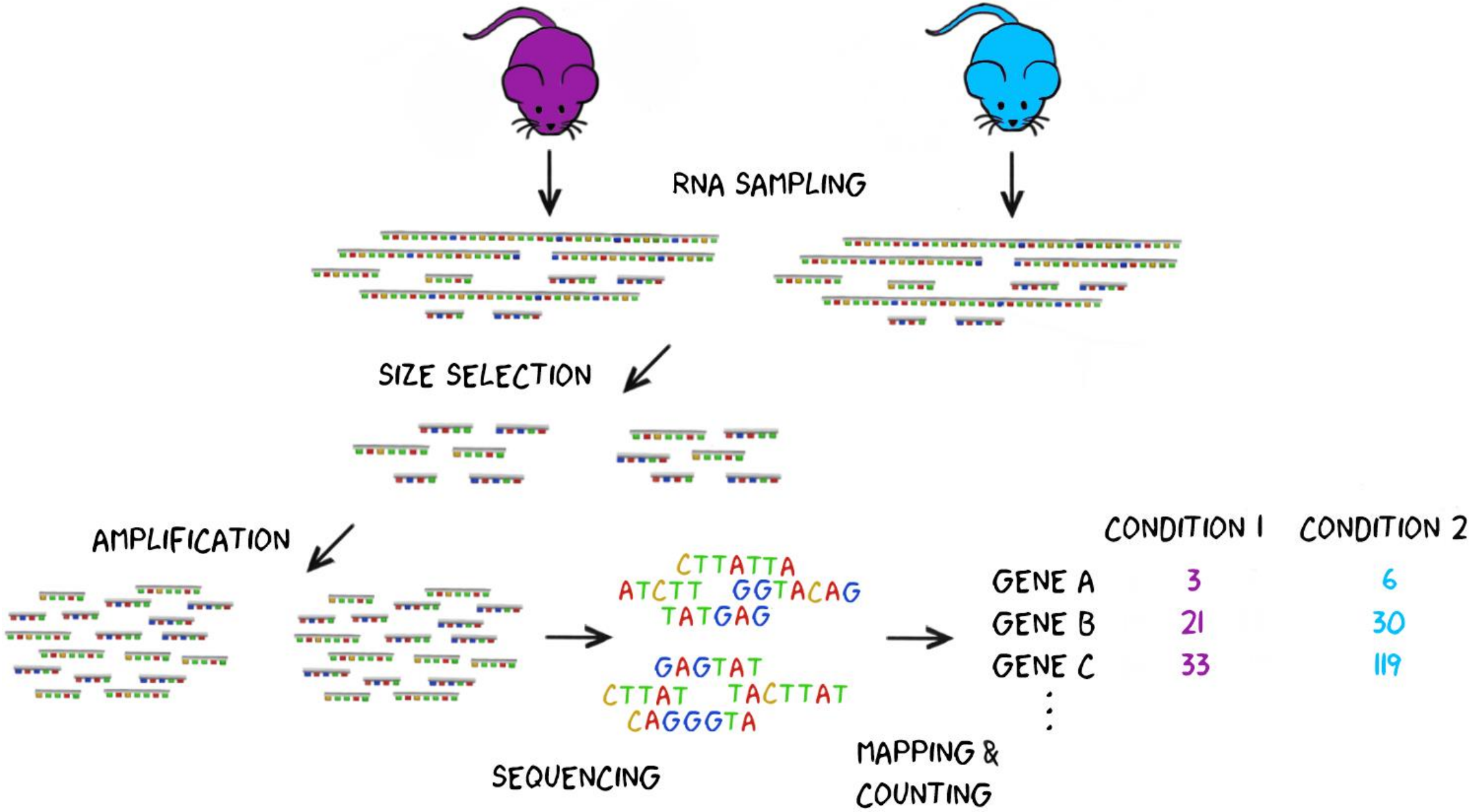
# A new statistical approach to identify differential expression in small RNA-Seq data

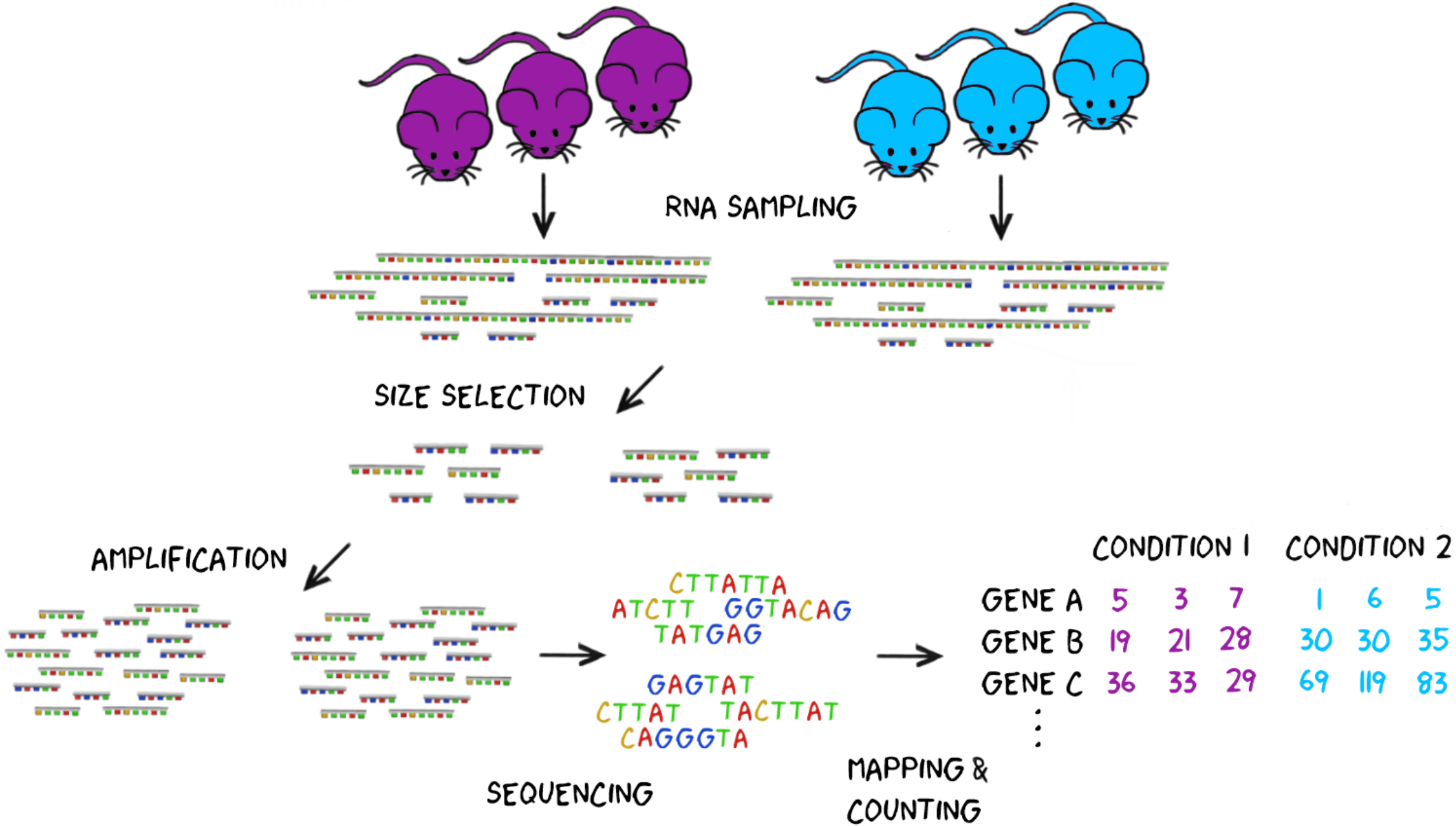
32nd TBI Winterseminar, 2017

Emanuel Barth


seit 1558


miRNA gene start
pri-miRNA
Drosha
pre-miRNA
Dicer
miRNA duplex
miRNP
miRNA
miRNA*

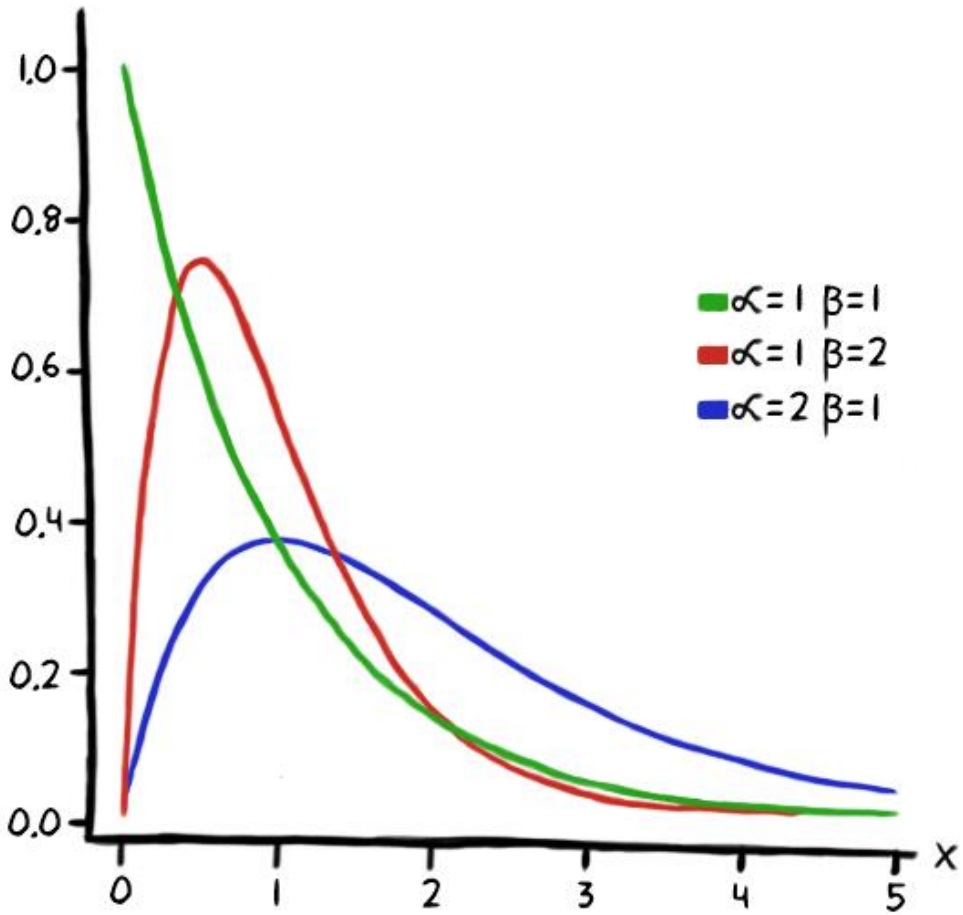

RNA SAMPLING
SIZE SELECTION
AMPLIFICATION
CTTATTA
ATCTT GGTACAG
TATGAG
GAGTAT
CTTAT TACTTAT
CAGGGTA
SEQUENCING
MAPPING &
COUNTING
CONDITION 1
CONDITION 2
GENE A 3 6
GENE B 21 30
GENE C 33 119


RNA SAMPLING
SIZE SELECTION
AMPLIFICATION
CTTATTA
ATCTT GGTACAG
TATGAG
GAGTAT
CTTAT TACTTAT
CAGGGTA
SEQUENCING
MAPPING &
COUNTING
CONDITION 1
CONDITION 2
GENE A 5 3 7 1 6 5
GENE B 19 21 28 30 30 35
GENE C 36 33 29 69 119 83


- Reads coming from small RNA-Seq are more likely to follow a Gamma distribution rather than a Poisson distribution

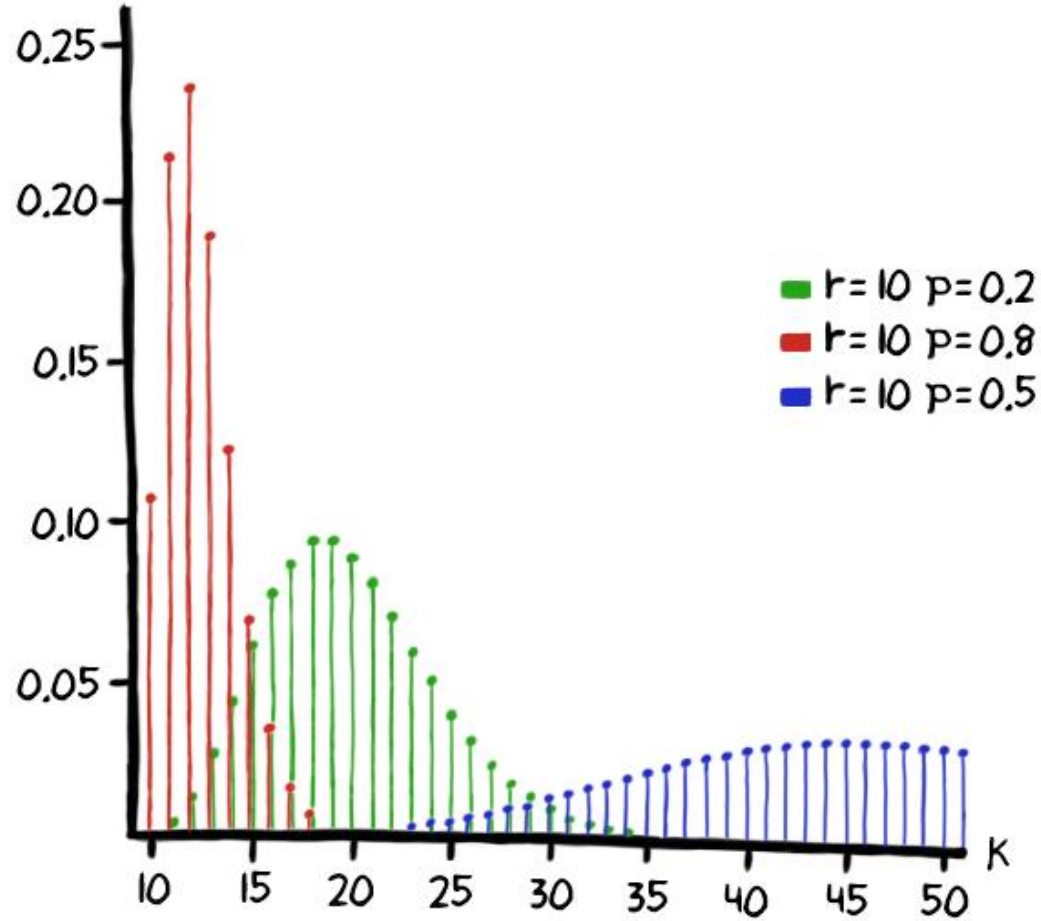
- Through a simple transformation of the count data, we can make use of much simpler statistical models for differential expression analysis

## GAMMA




$$f(x) = \begin{cases} \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

## NEGATIVE BINOMIAL




$$f(x) = \sum_{K=r}^{\lfloor x \rfloor} \binom{K-1}{r-1} p^{r}(1-p)^{K-r}$$

# MicroRNA Differential Expression analysis (MeRDE)

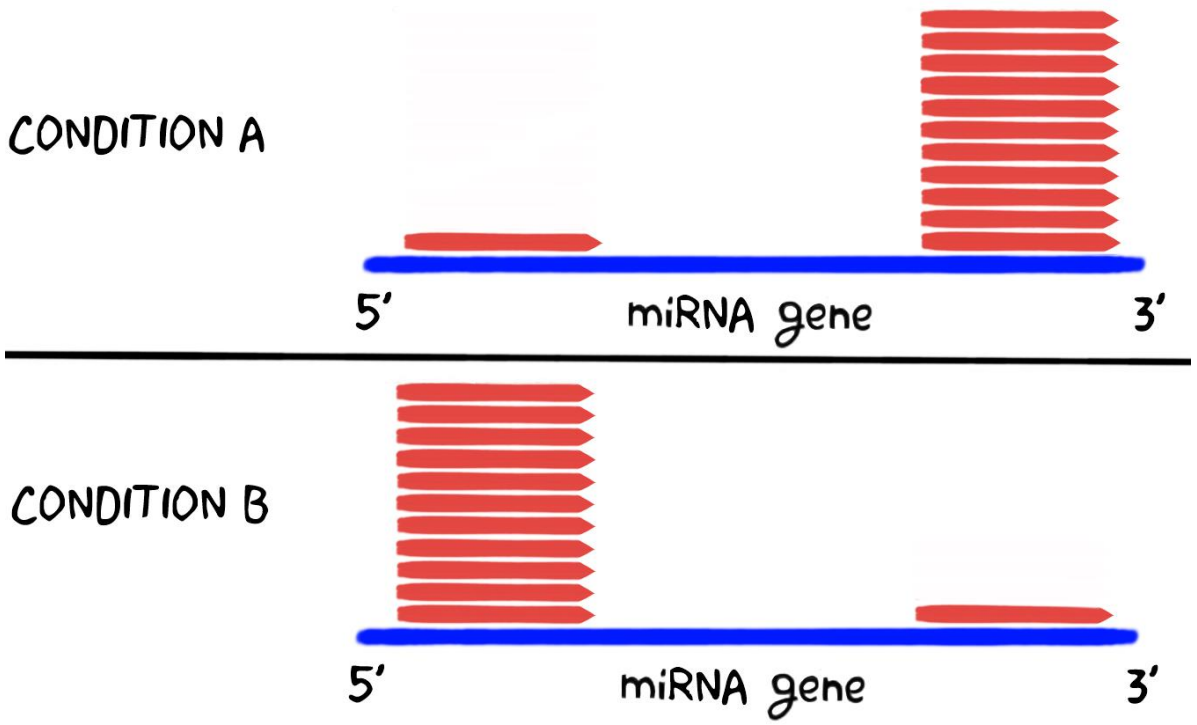
- Input: sam/bam mapping files or raw count file
- Output: differentially expressed miRNAs (and some nice pictures)

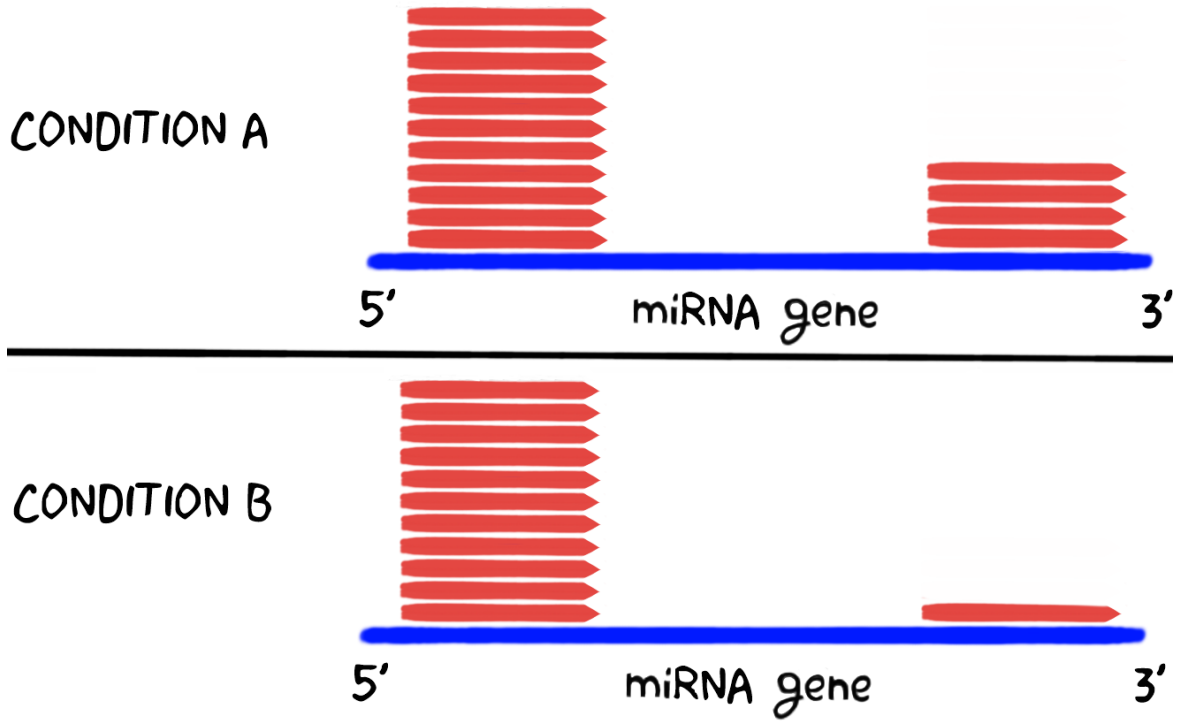
- Count normalization based on pseudo-reference scaling[1]
- Count transformation using the cubic root function
- Testing for differential expression with Welch‘s t-test

[1]Anders et al. *Differential expression analysis for sequence count data,* G Bio, 2010, 11:R106

# MicroRNA Differential Expression analysis (MeRDE)

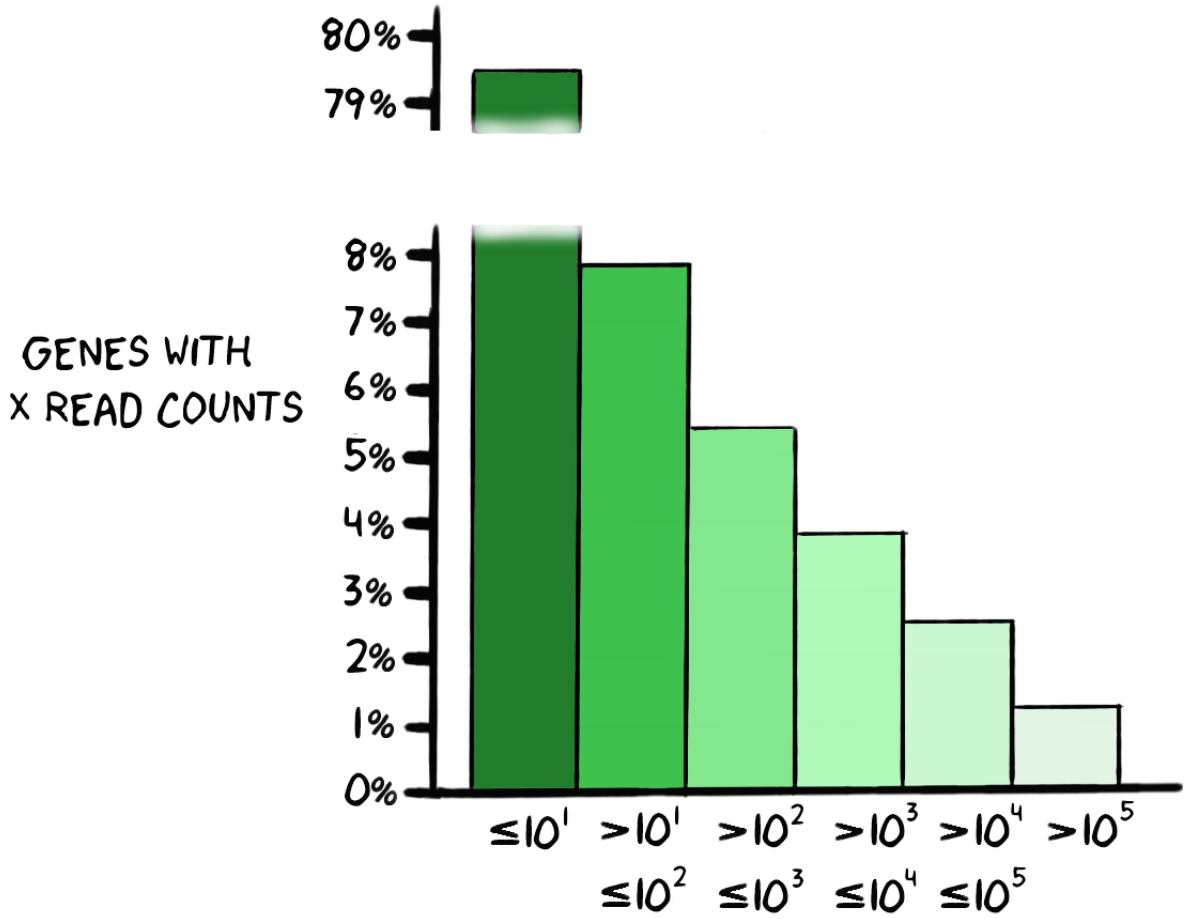
- Treats 5‘- and 3‘-ends independently



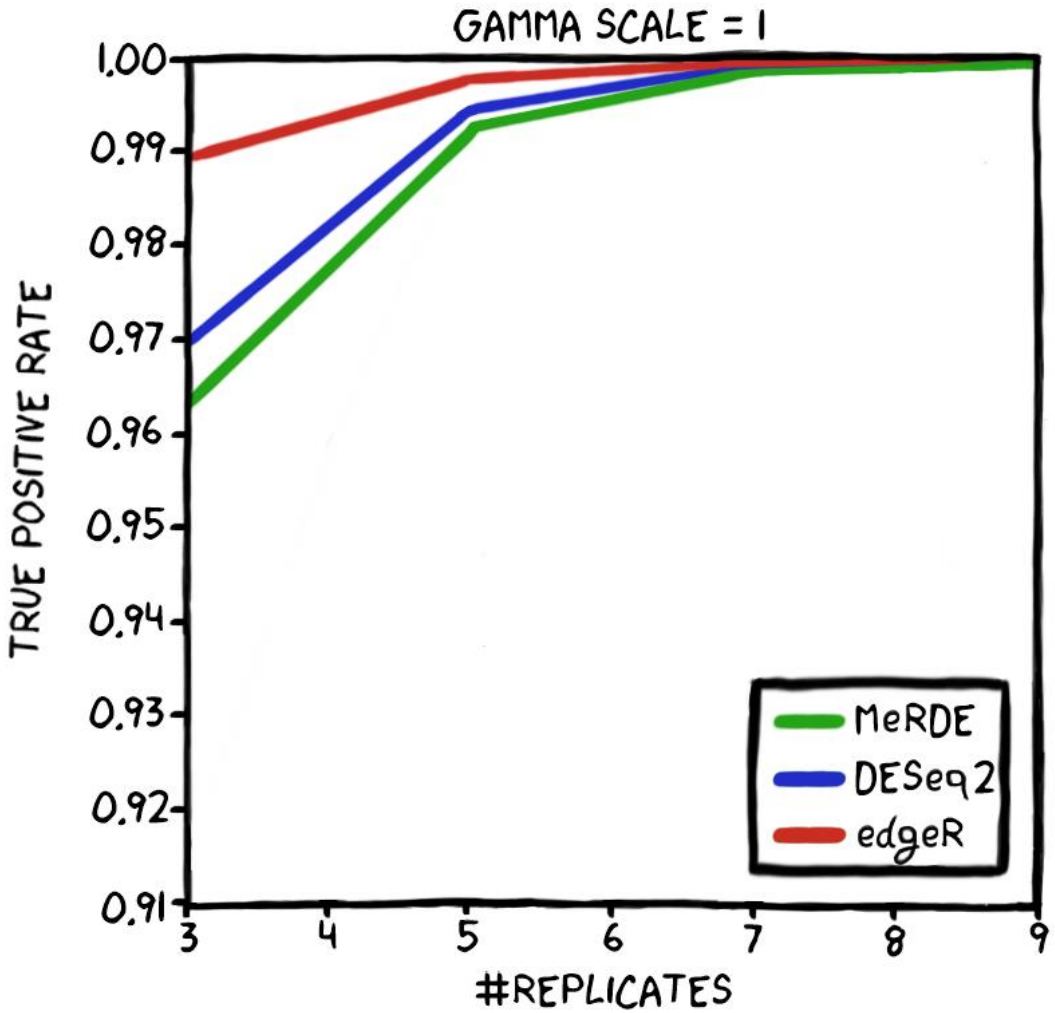



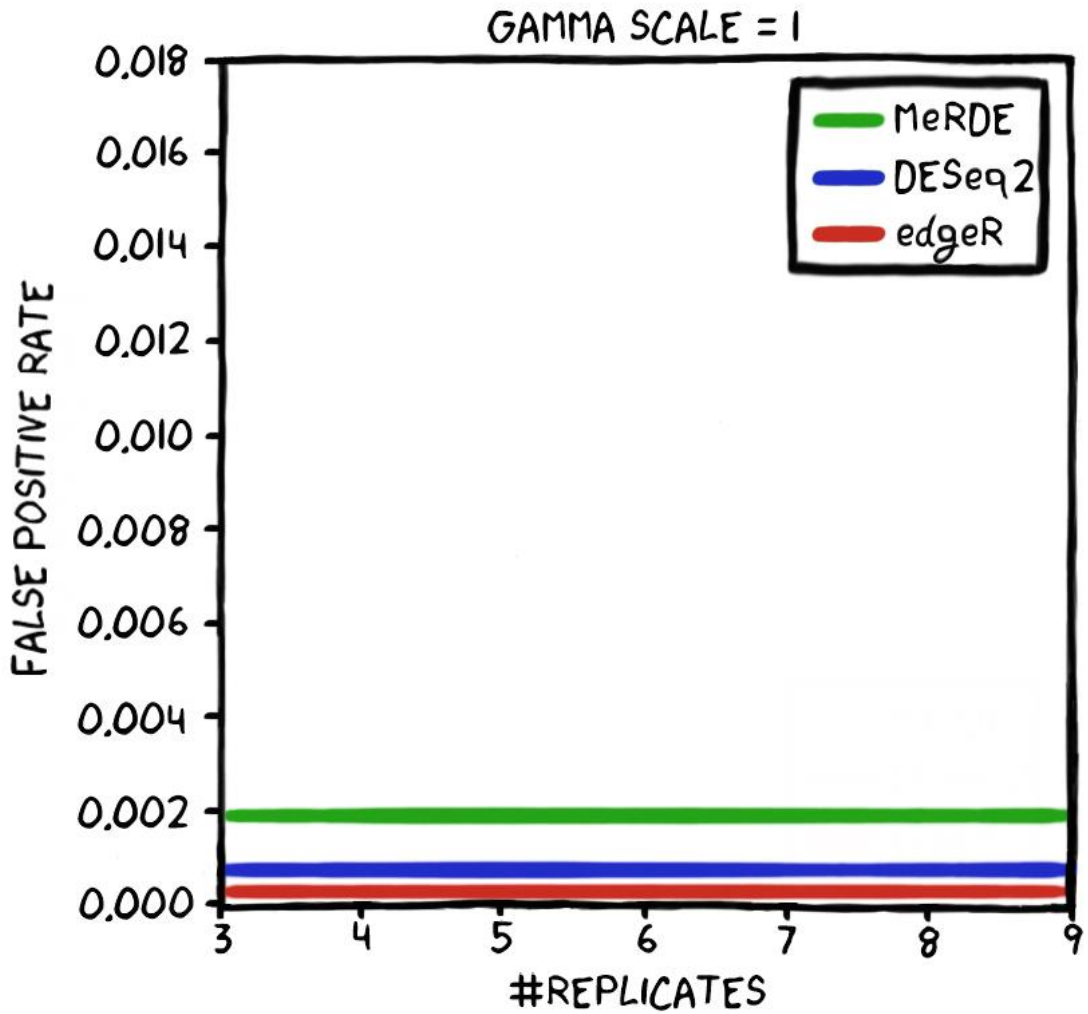
- small RNA-Seq data sets of three different species of different tissues between different ages

- Human: 60 Datasets
- Mouse: 74 Datasets
- *Notho. furzeri*: 160 Datasets



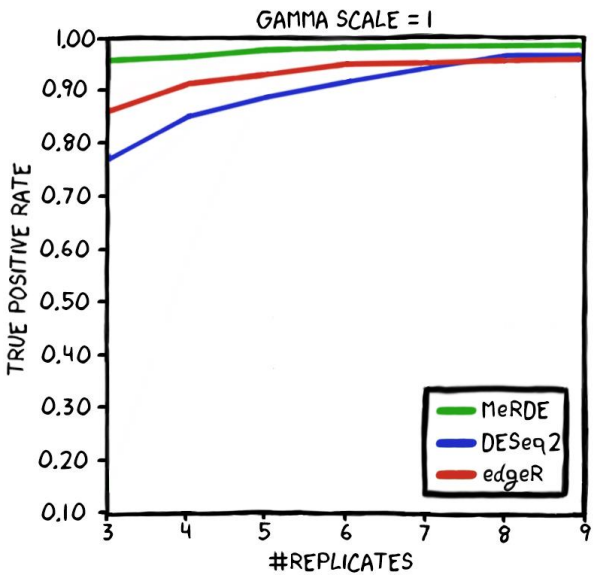
- Artificial data setup:
  - Gamma scaling factor $\beta = 1$
  - two conditions with 3 to 9 replicates
  - 1000 count files each containing 5000 simulated miRNAs
  - 1% - 5% of the genes are DE
  - DE factor between 2-fold and 5-fold up-/downregulated

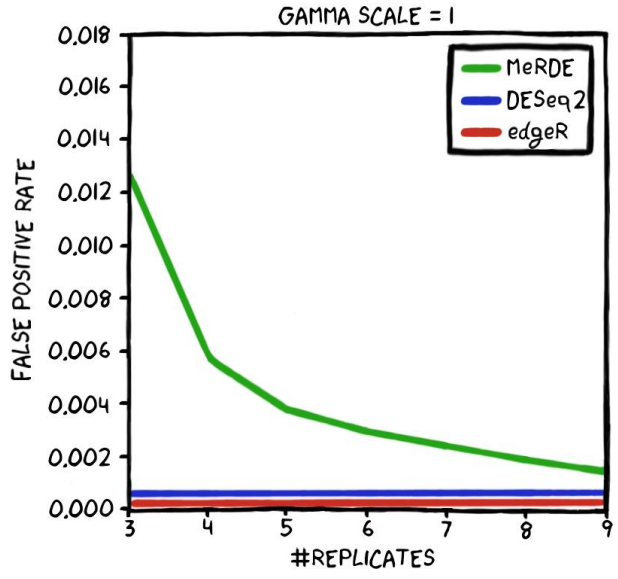

GAMMA SCALE = 1
TRUE POSITIVE RATE
1.00
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.91
3
4
5
6
7
8
9
#REPLICATES
MeRDE
DESeq2
edgeR

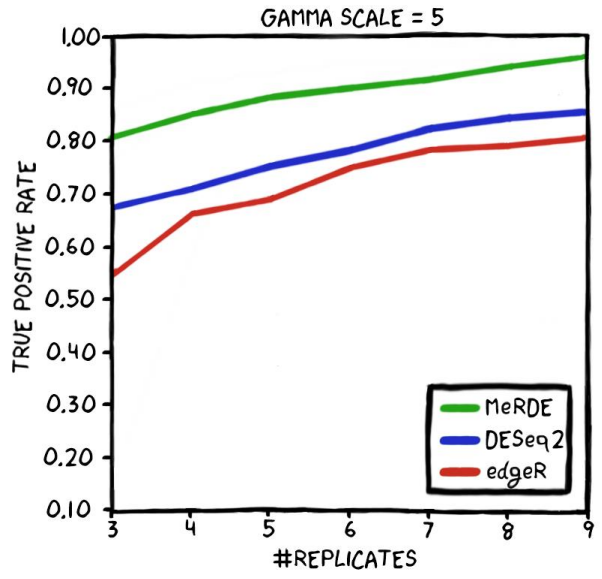

GAMMA SCALE = 1
FALSE POSITIVE RATE
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
3
4
5
6
7
8
9
#REPLICATES
MeRDE
DESeq2
edgeR

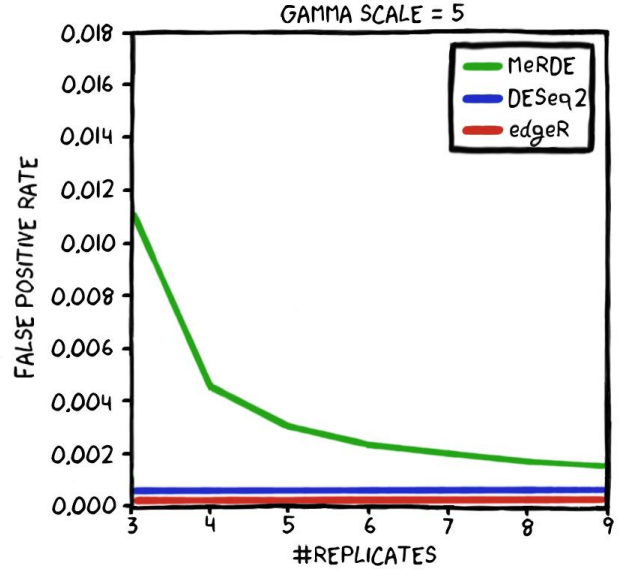
- Improved artificial data setup:
  - Gamma scaling factor $\beta$ =1, 5, 10, 20
  - two conditions with 3 to 9 replicates
  - 1000 count files each containing 5000 miRNAs
  - 1% - 5% of the genes are DE
  - DE factor between $f(x)$ and 2-fold up-/downregulated
  - outlier rate of 2%

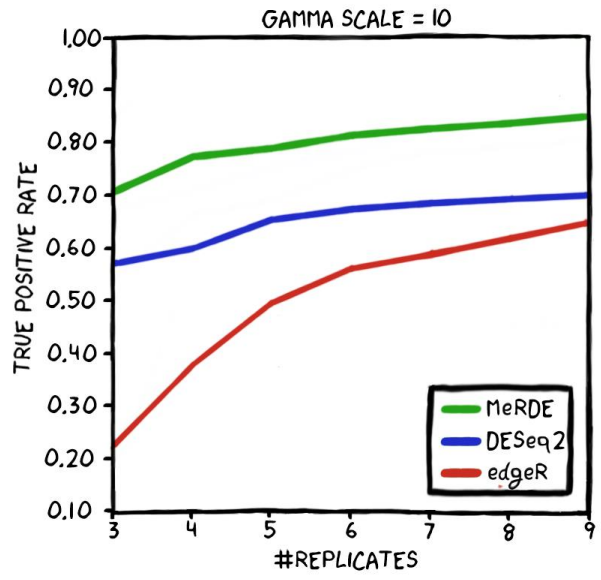
$$f(x) = \frac{1.322}{\log_{4.8} x} + 1.1 \text{ where } x \text{ is the base mean expression}$$

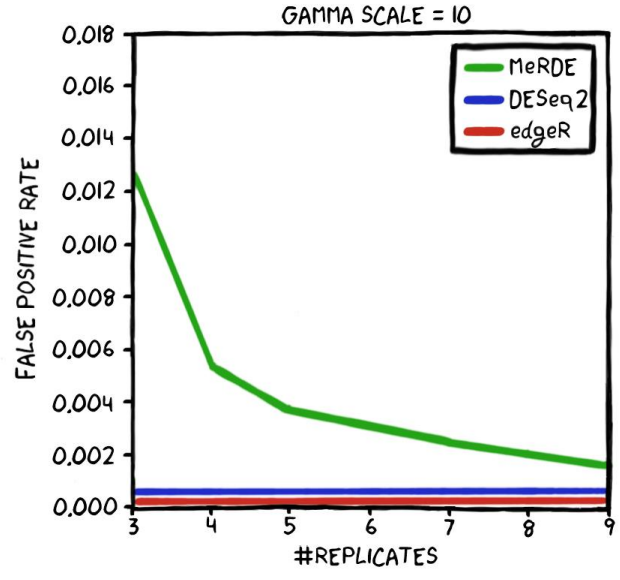

GAMMA SCALE = 1
TRUE POSITIVE RATE
#REPLICATES
MeRDE
DESeq2
edgeR

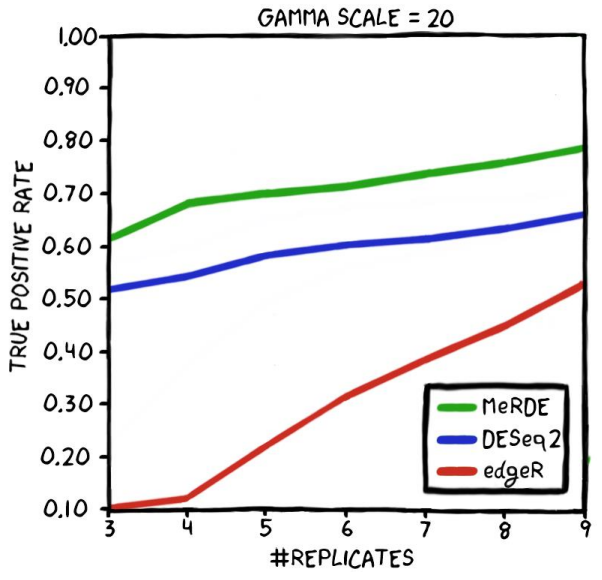

GAMMA SCALE = 1
FALSE POSITIVE RATE
#REPLICATES
MeRDE
DESeq2
edgeR

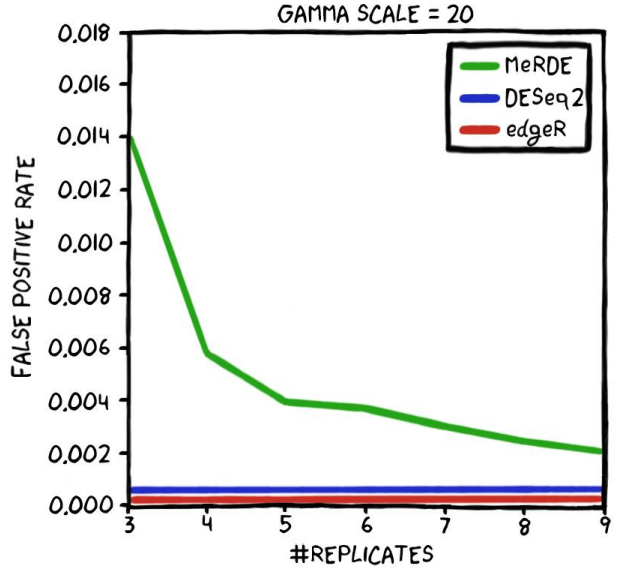

GAMMA SCALE = 5
TRUE POSITIVE RATE
#REPLICATES
MeRDE
DESeq2
edgeR


GAMMA SCALE = 5
FALSE POSITIVE RATE
#REPLICATES
MeRDE
DESeq2
edgeR


GAMMA SCALE = 10
TRUE POSITIVE RATE
#REPLICATES
MeRDE
DESeq2
edgeR

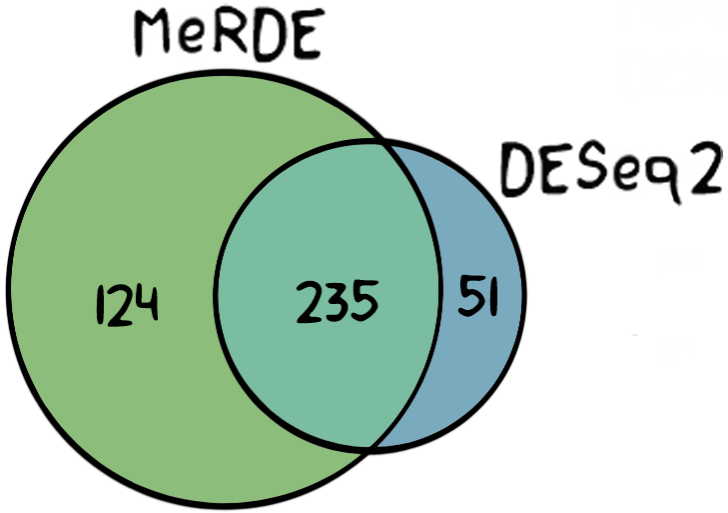

GAMMA SCALE = 10
FALSE POSITIVE RATE
#REPLICATES
MeRDE
DESeq2
edgeR


GAMMA SCALE = 20
TRUE POSITIVE RATE
#REPLICATES
MeRDE
DESeq2
edgeR

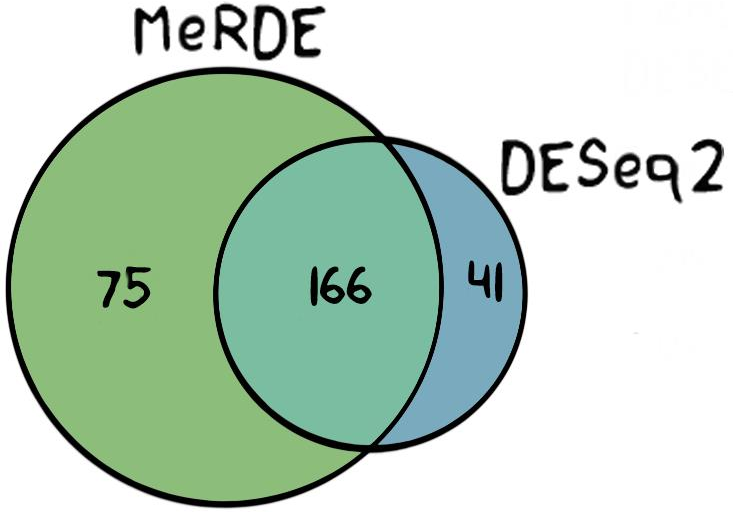

GAMMA SCALE = 20
FALSE POSITIVE RATE
#REPLICATES
MeRDE
DESeq2
edgeR

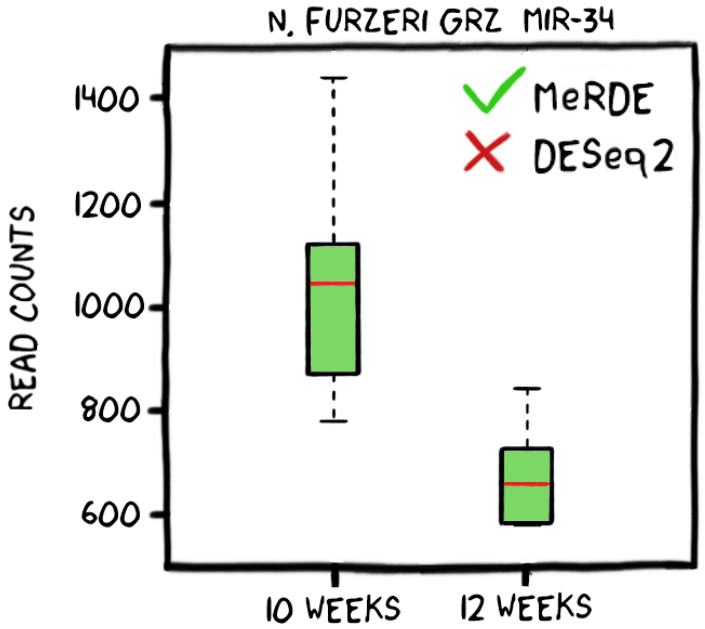
Real data results:

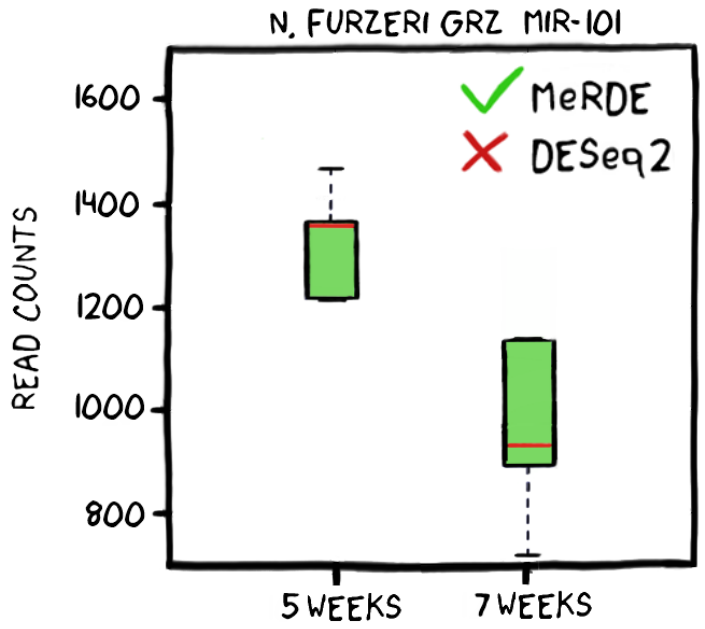
- Pairwise age comparison of two short-lived fish
  - *Nothobranchius furzeri GRZ*: 5, 7, 10, 12 and 14 weeks of age
  - ≈ 20% more DEGs compared to DESeq2

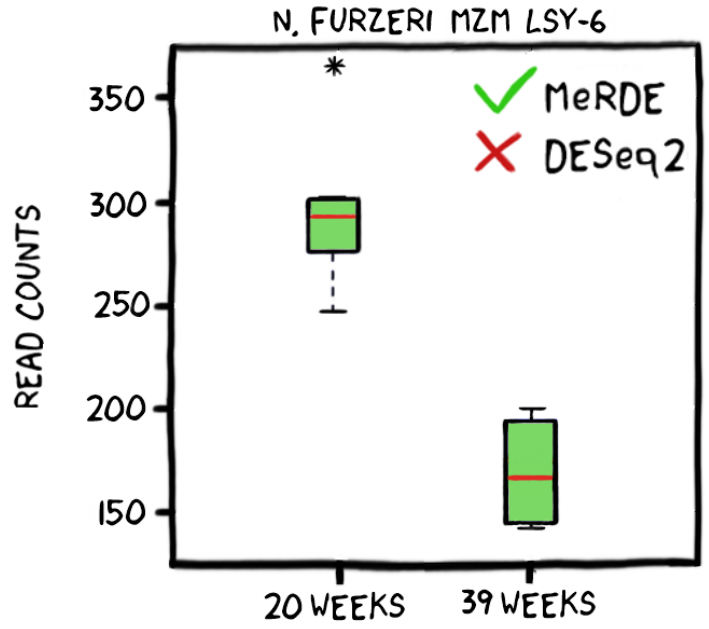


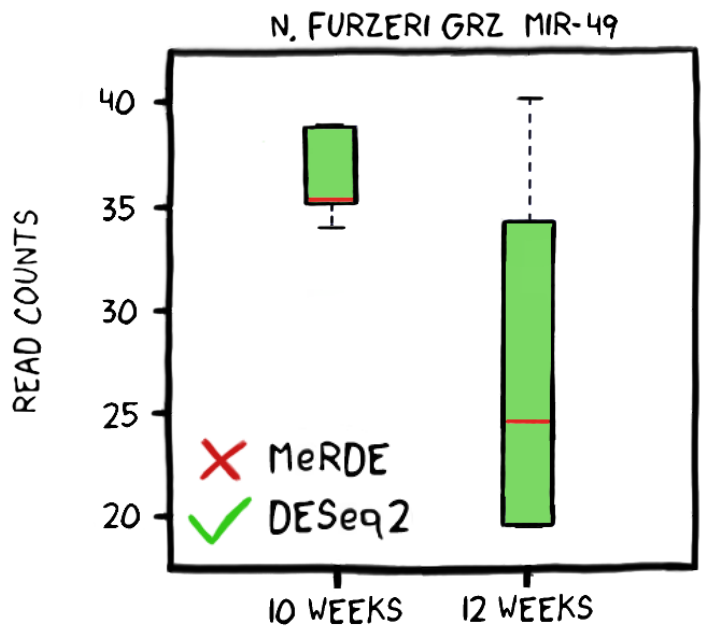

  - *Nothobranchius furzeri MZM* 5, 12, 20, 27 and 39 weeks of age
  - ≈ 14% more DEGs compared to DESeq2

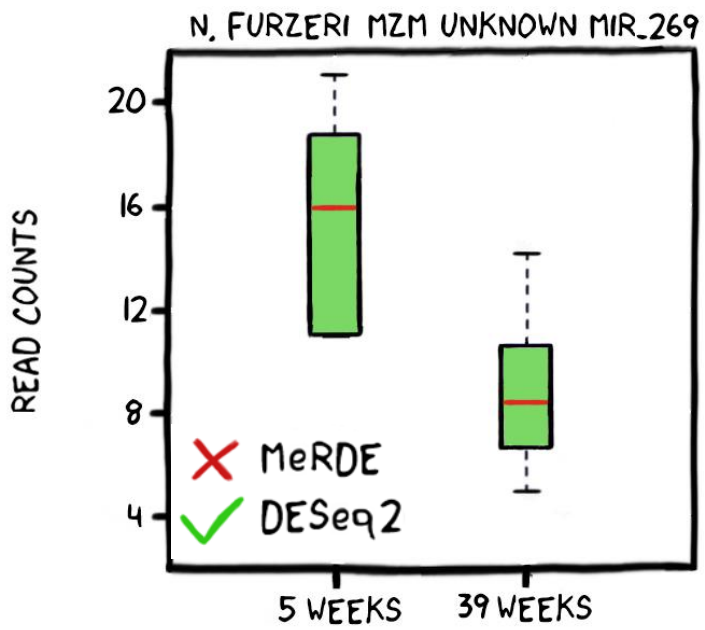


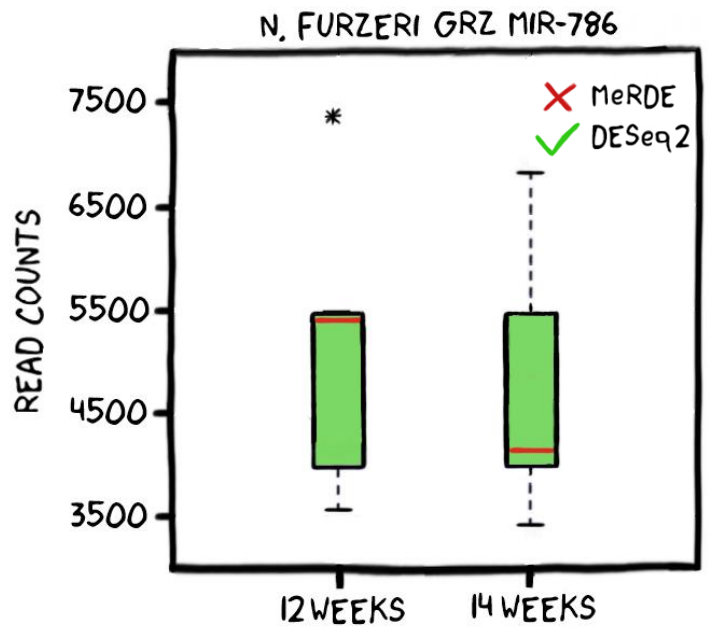

N. FURZERI GRZ MIR-34
READ COUNTS
1400
1200
1000
800
600
MeRDE
DESeq2
10 WEEKS
12 WEEKS


N. FURZERI GRZ MIR-101
READ COUNTS
1600
1400
1200
1000
800
MeRDE
DESeq2
5 WEEKS
7 WEEKS


N. FURZERI MZM LSY-6
READ COUNTS
350
300
250
200
150
MeRDE
DESeq2
20 WEEKS
39 WEEKS

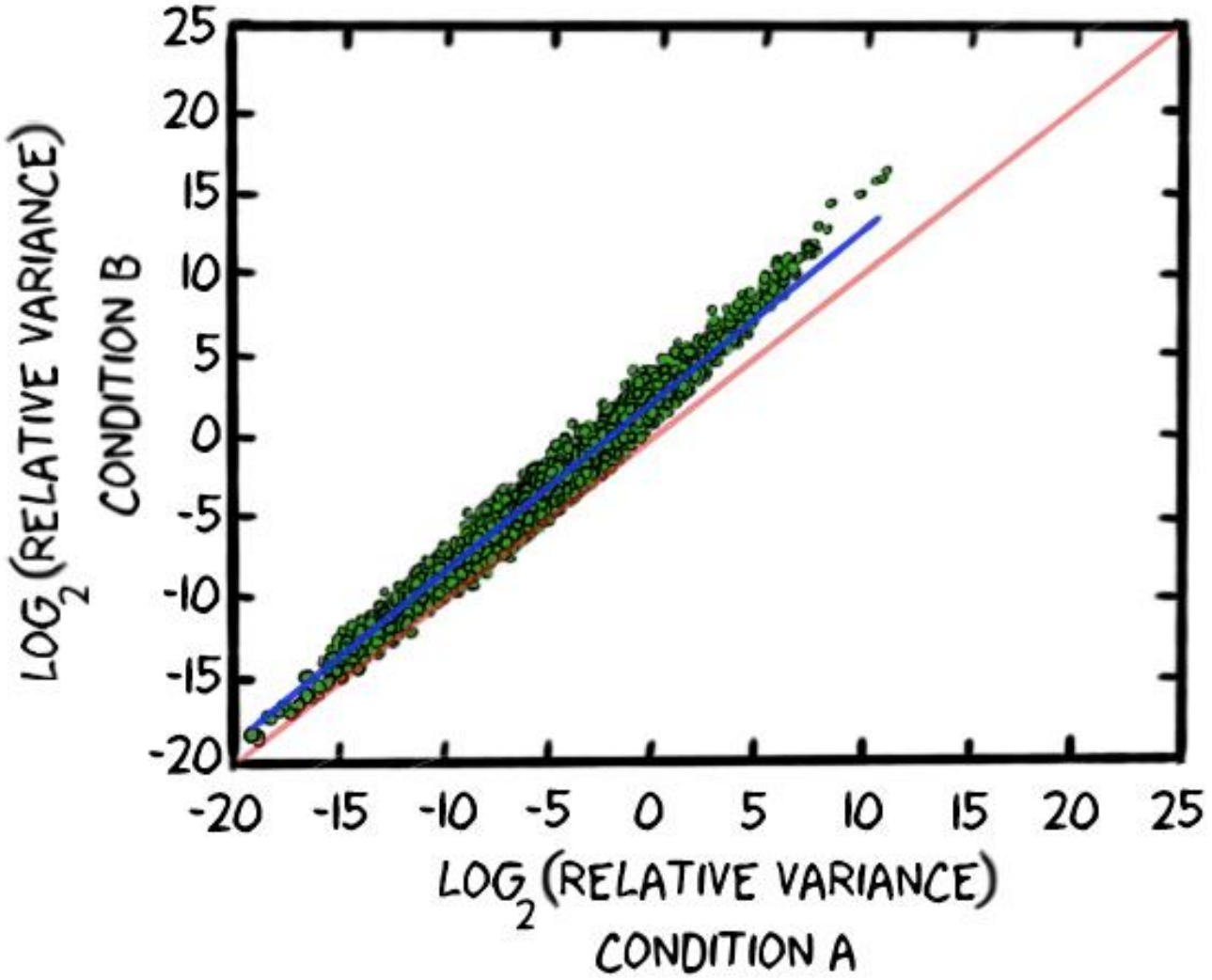

N. FURZERI GRZ MIR-49
READ COUNTS
40
35
30
25
20
MeRDE
DESeq2
10 WEEKS
12 WEEKS


N. FURZERI MZM UNKNOWN MIR_269
READ COUNTS
20
16
12
8
4
MeRDE
DESeq2
5 WEEKS
39 WEEKS


N. FURZERI GRZ MIR-786
READ COUNTS
7500
6500
5500
4500
3500
MeRDE
DESeq2
12 WEEKS
14 WEEKS

Still to do:

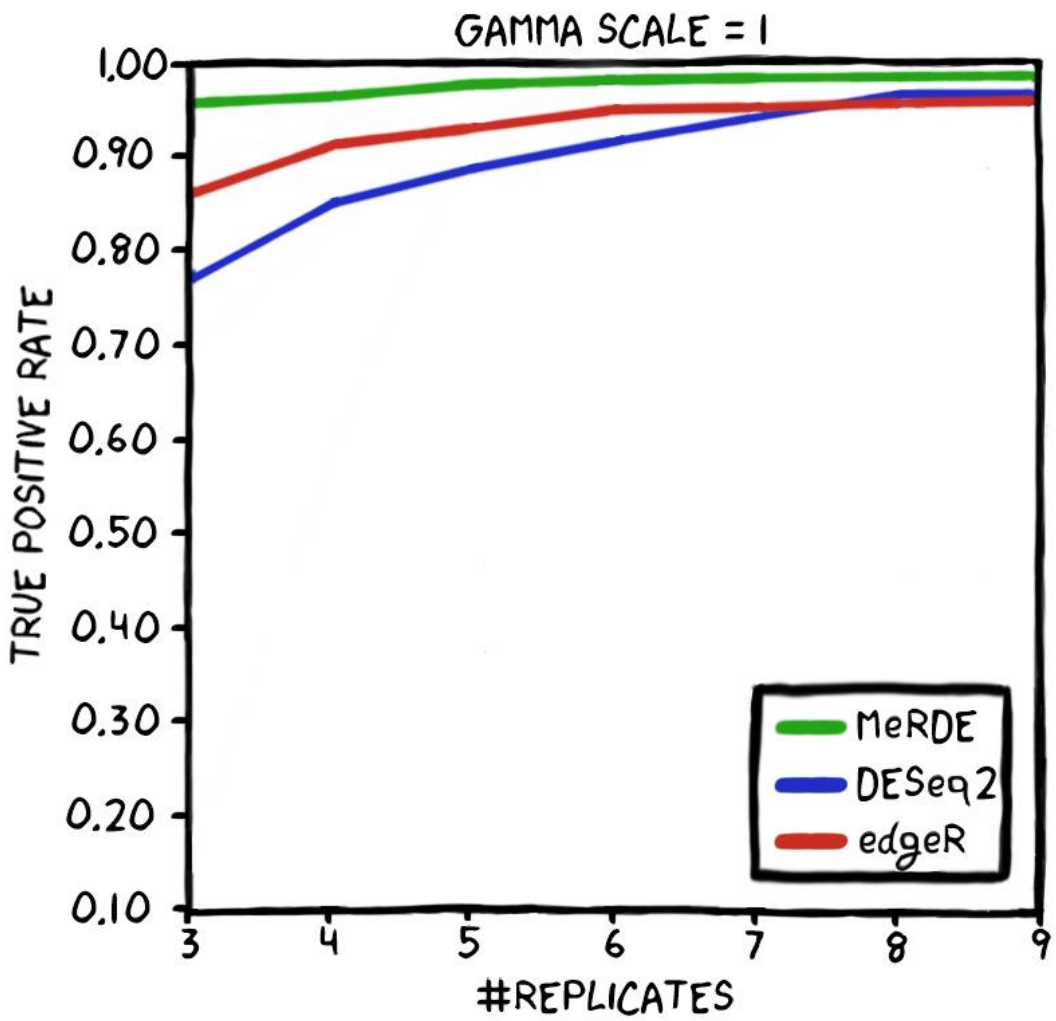
- Detecting and handling outliers

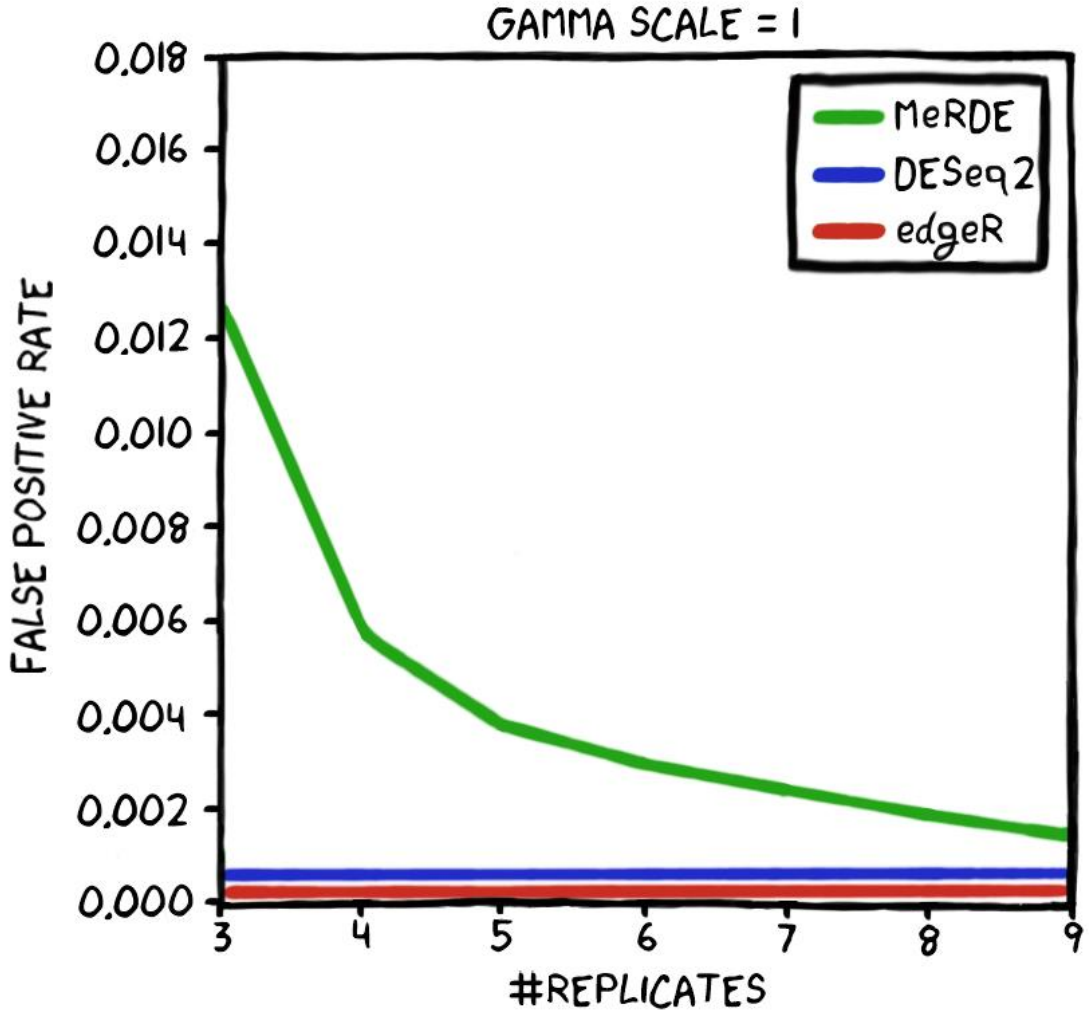


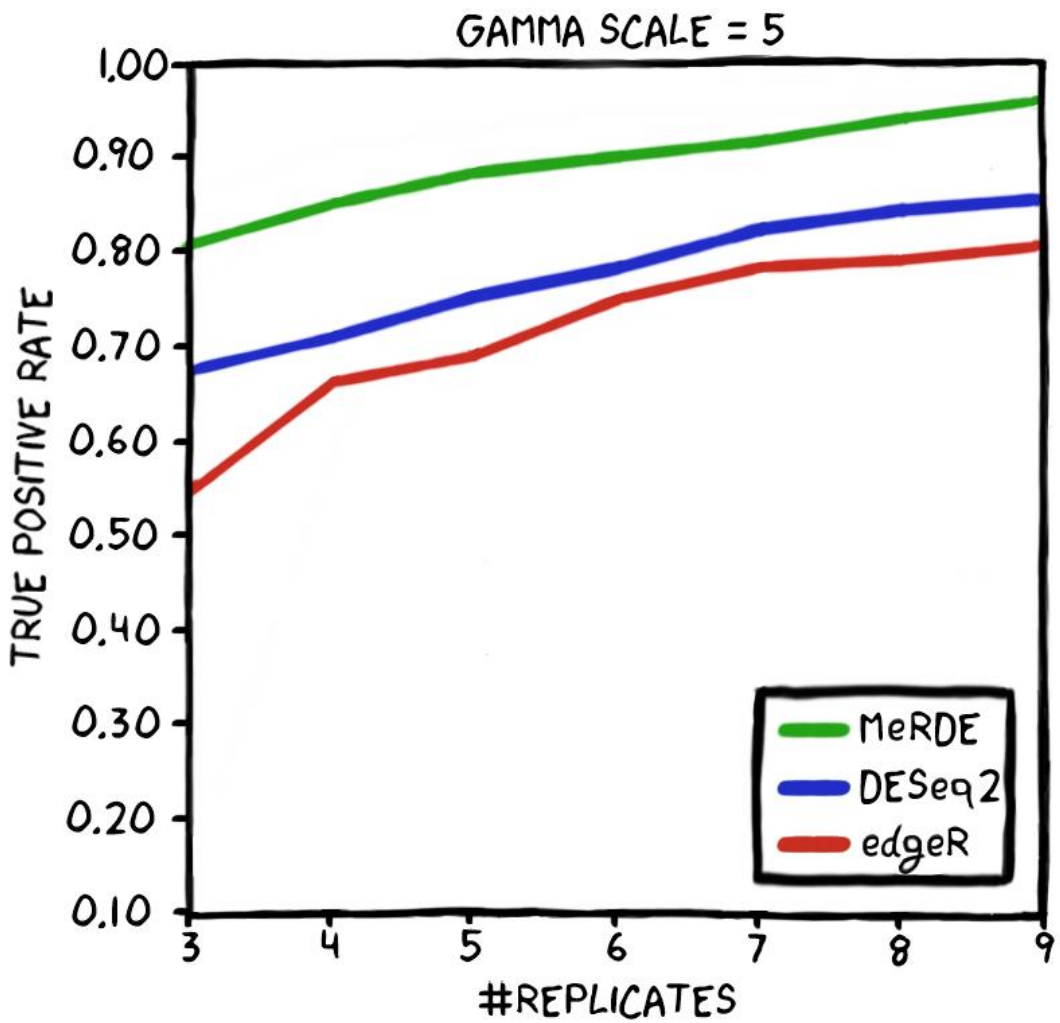
Still to do:

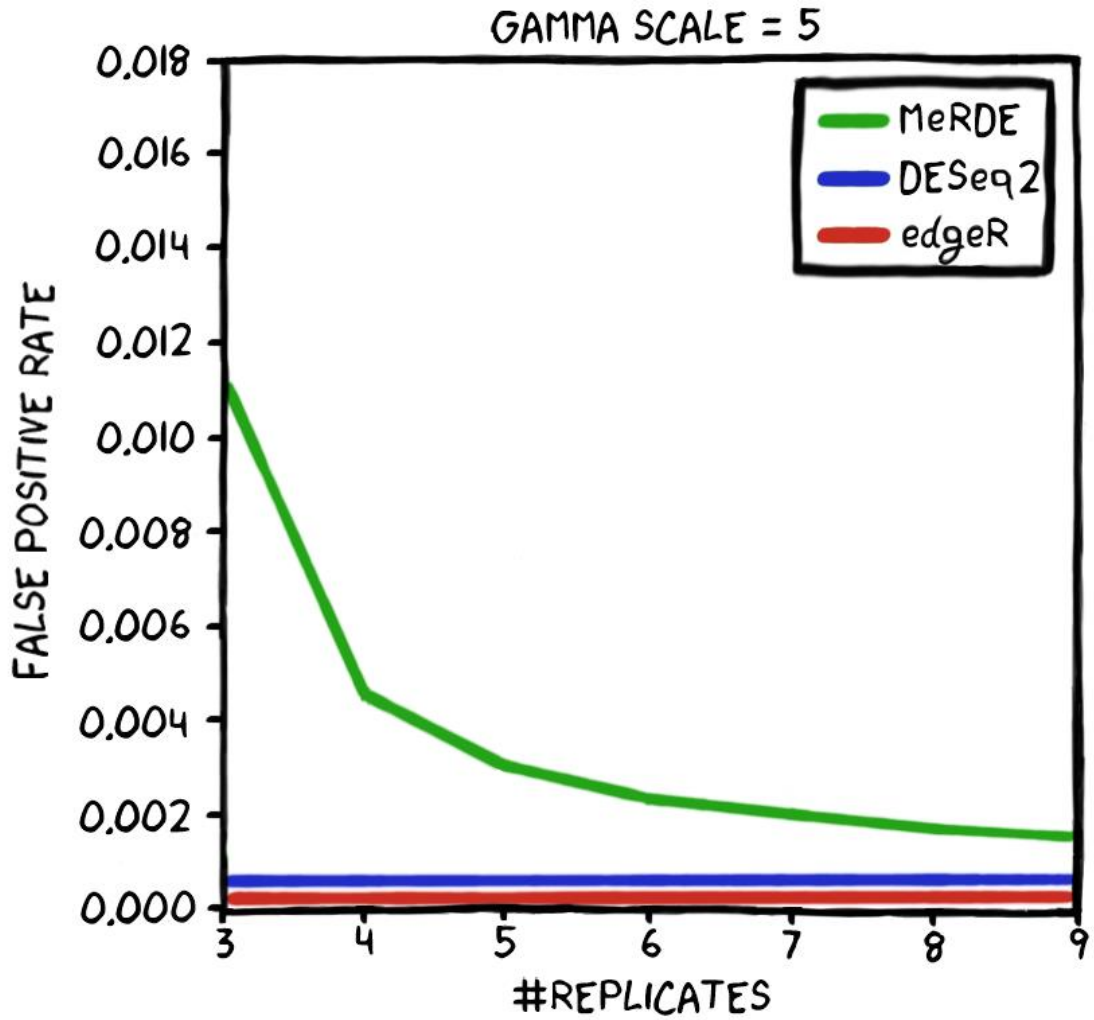
- Adjust transformation based on the genes estimated means and variances by applying a regularized cubic root function

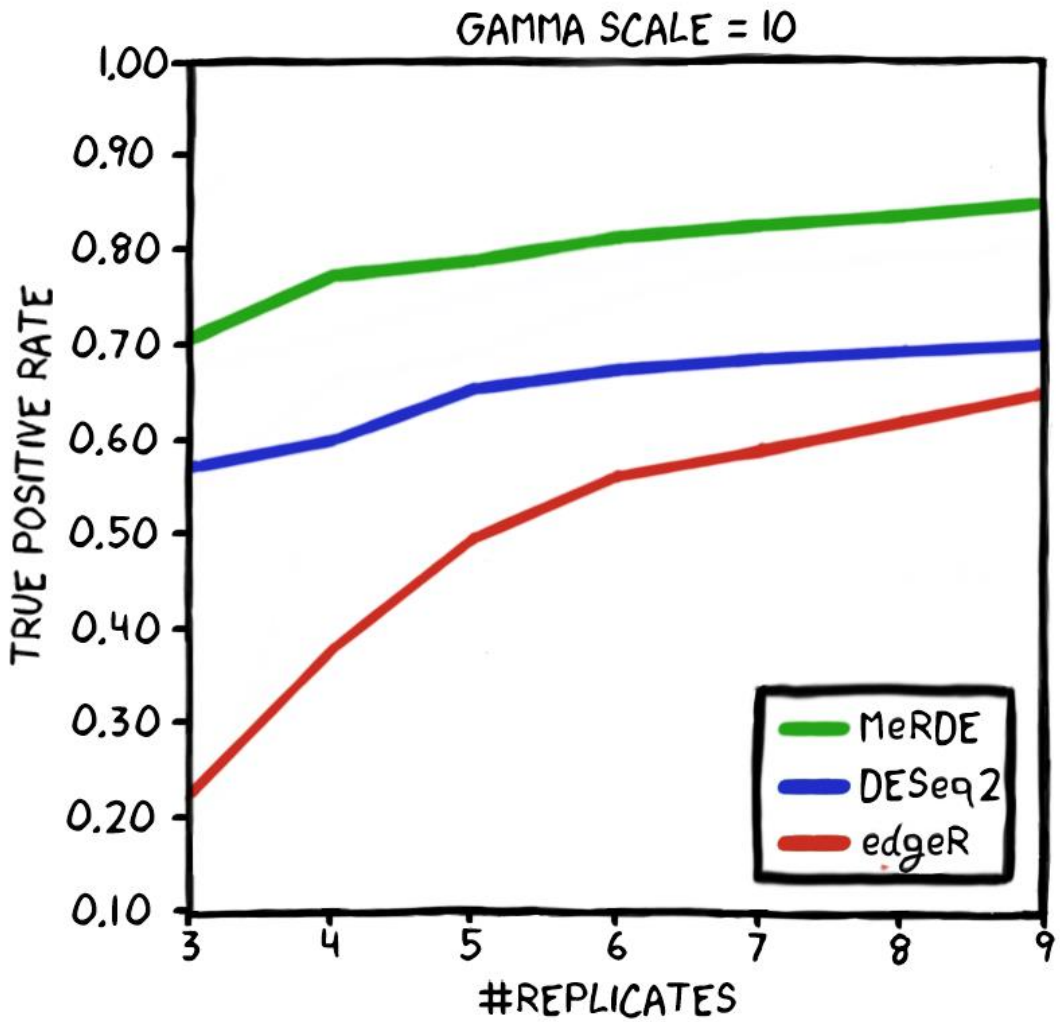
- Instead of count transformation, create a new model and perform hypothesis tests assuming gamma-distributed random variables

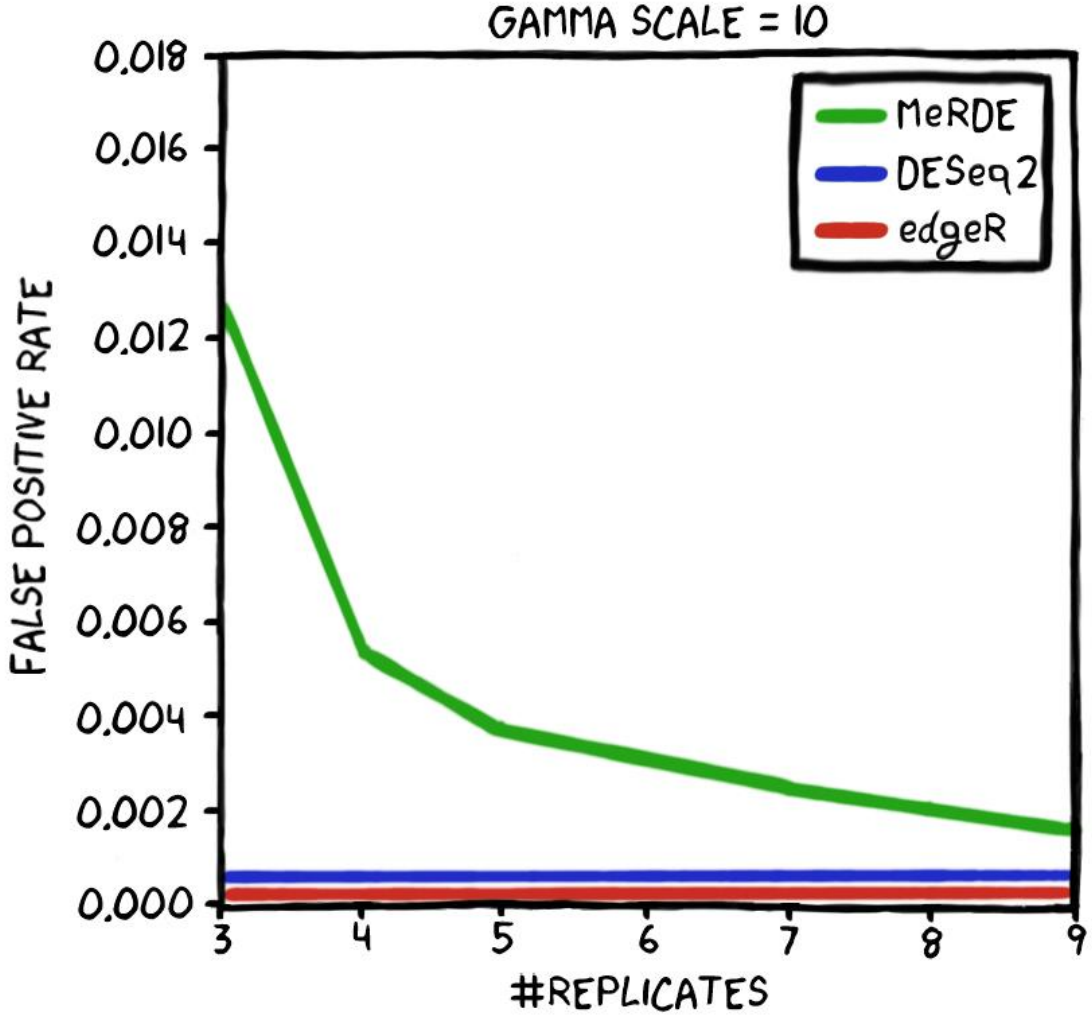

GAMMA SCALE = 1
TRUE POSITIVE RATE
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
3
4
5
6
7
8
9
#REPLICATES
MeRDE
DESeq2
edgeR

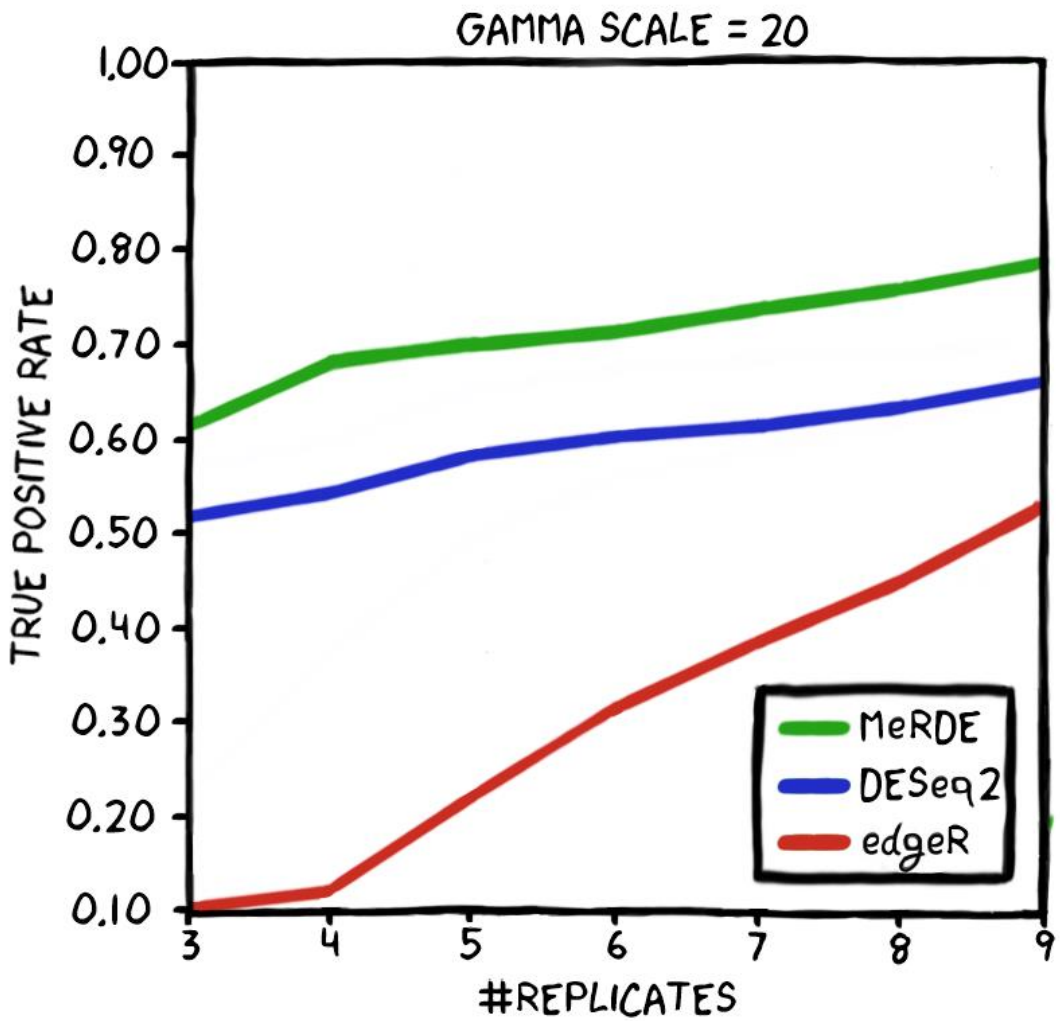

GAMMA SCALE = 1
FALSE POSITIVE RATE
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
3
4
5
6
7
8
9
#REPLICATES
MeRDE
DESeq2
edgeR

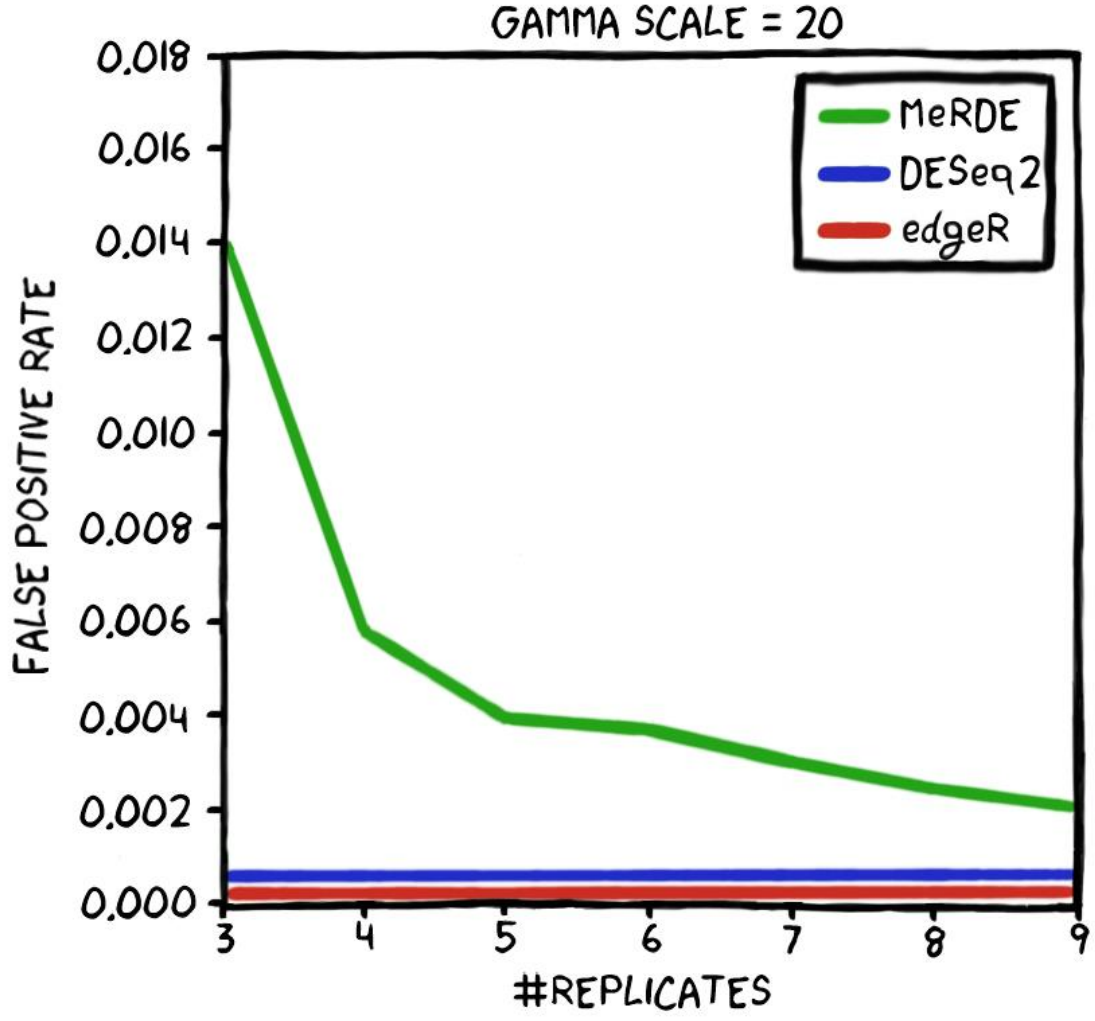

GAMMA SCALE = 5
TRUE POSITIVE RATE
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
3
4
5
6
7
8
9
#REPLICATES
MeRDE
DESeq2
edgeR


GAMMA SCALE = 5
FALSE POSITIVE RATE
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
3
4
5
6
7
8
9
#REPLICATES
MeRDE
DESeq2
edgeR


GAMMA SCALE = 10
TRUE POSITIVE RATE
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
3
4
5
6
7
8
9
#REPLICATES
MeRDE
DESeq2
edgeR

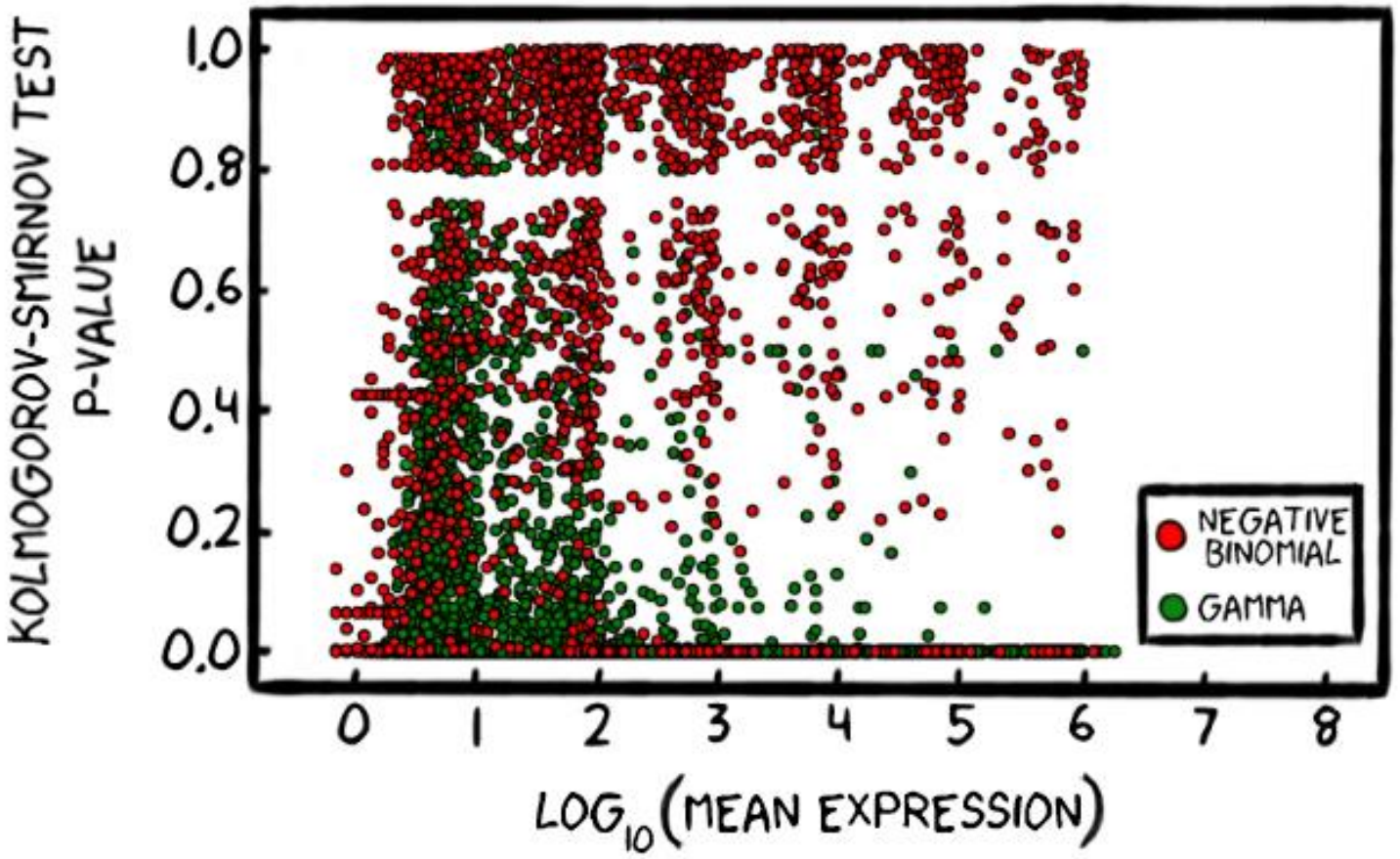

GAMMA SCALE = 10
FALSE POSITIVE RATE
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
3
4
5
6
7
8
9
#REPLICATES
MeRDE
DESeq2
edgeR

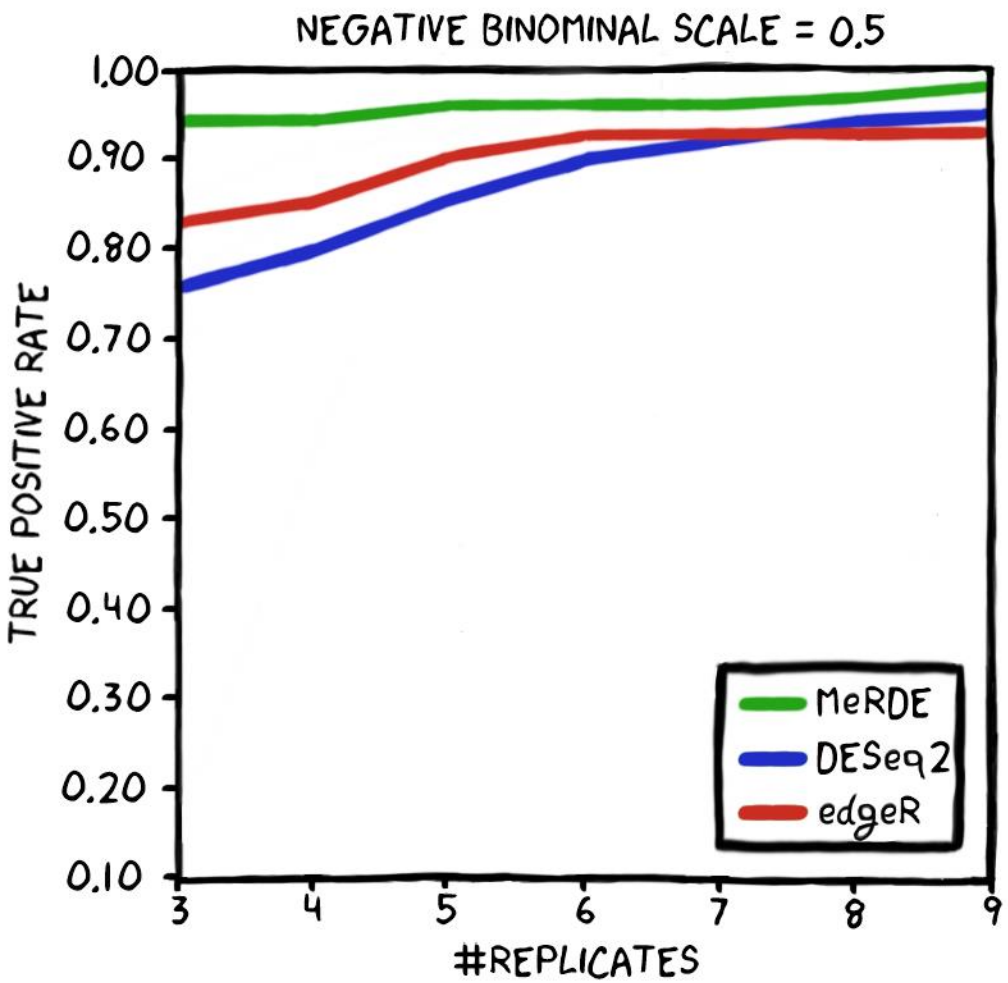

GAMMA SCALE = 20
TRUE POSITIVE RATE
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
3
4
5
6
7
8
9
#REPLICATES
MeRDE
DESeq2
edgeR

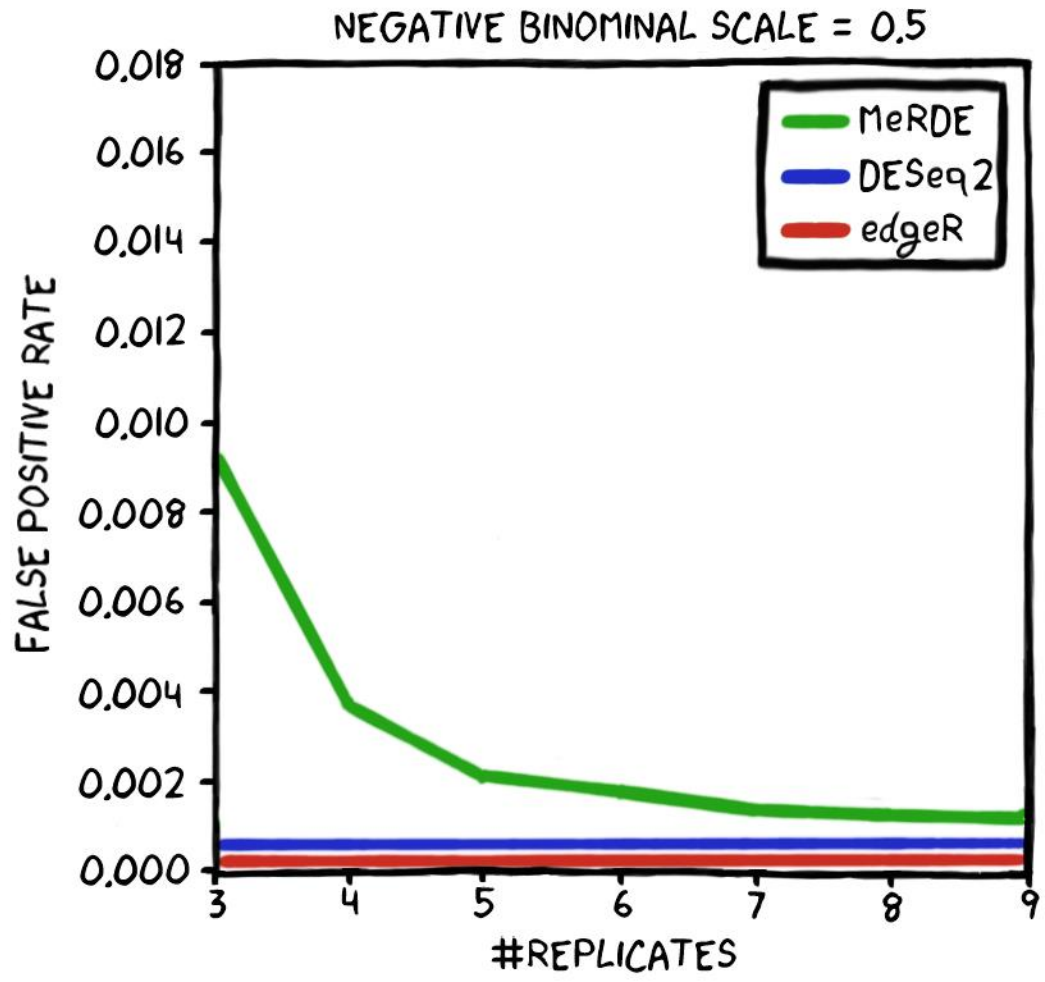

GAMMA SCALE = 20
FALSE POSITIVE RATE
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
3
4
5
6
7
8
9
#REPLICATES
MeRDE
DESeq2
edgeR

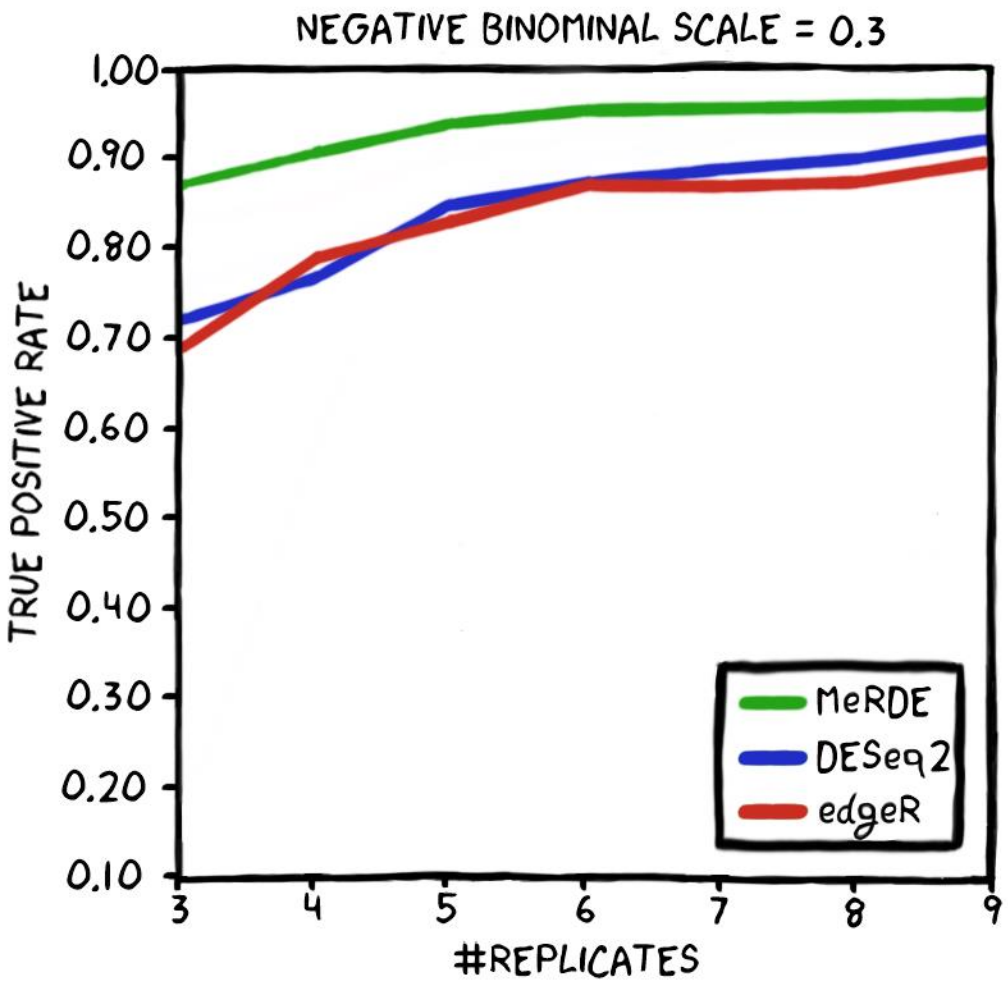
- small RNA-Seq data sets of three different species of different tissues between different ages

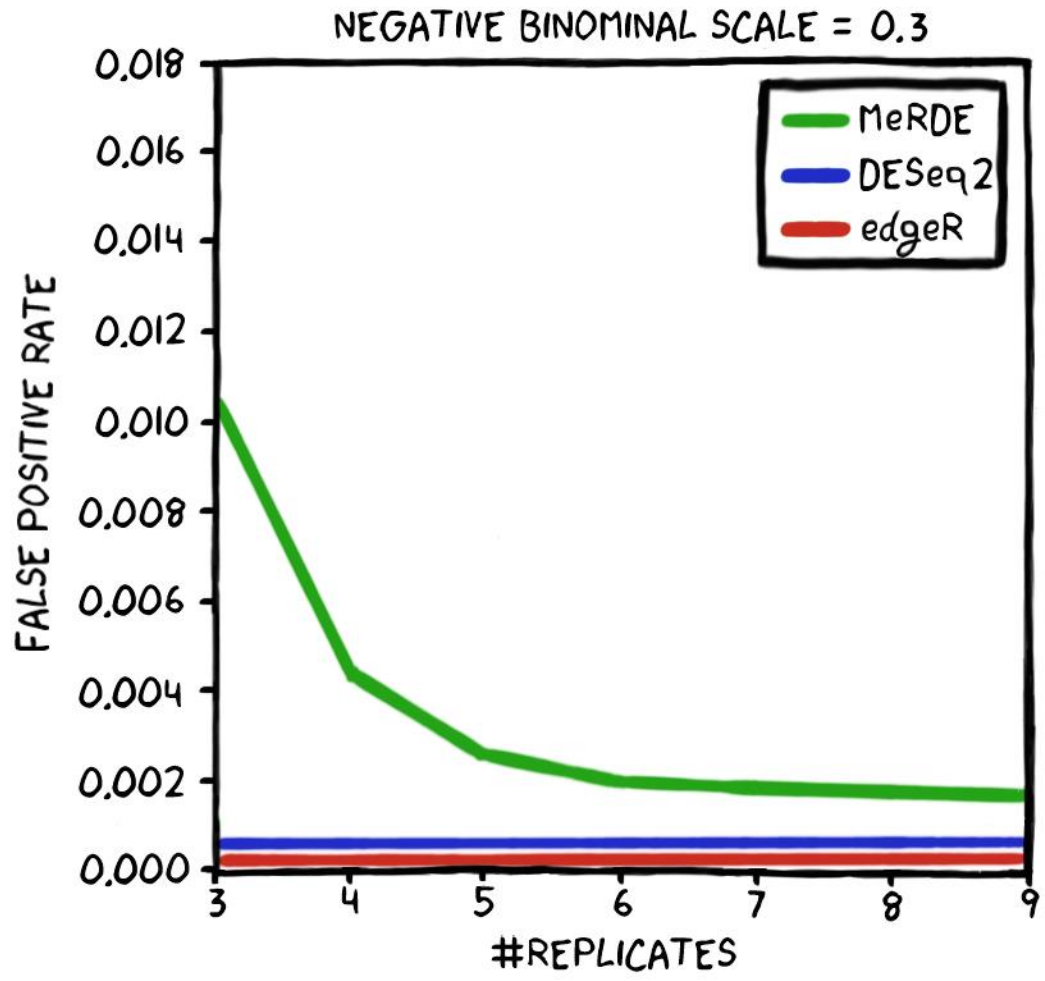
- Human: 60 Datasets
- Mouse: 74 Datasets
- *Notho. furzeri*: 160 Datasets

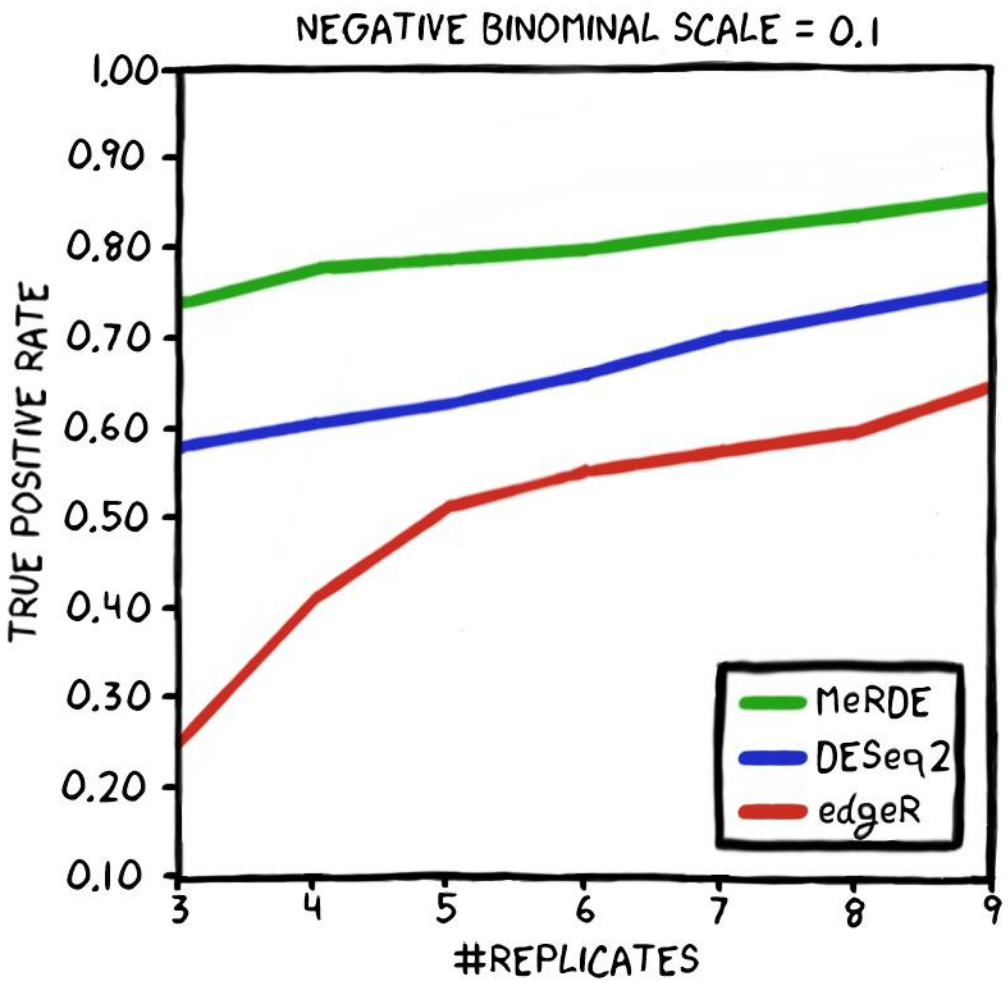


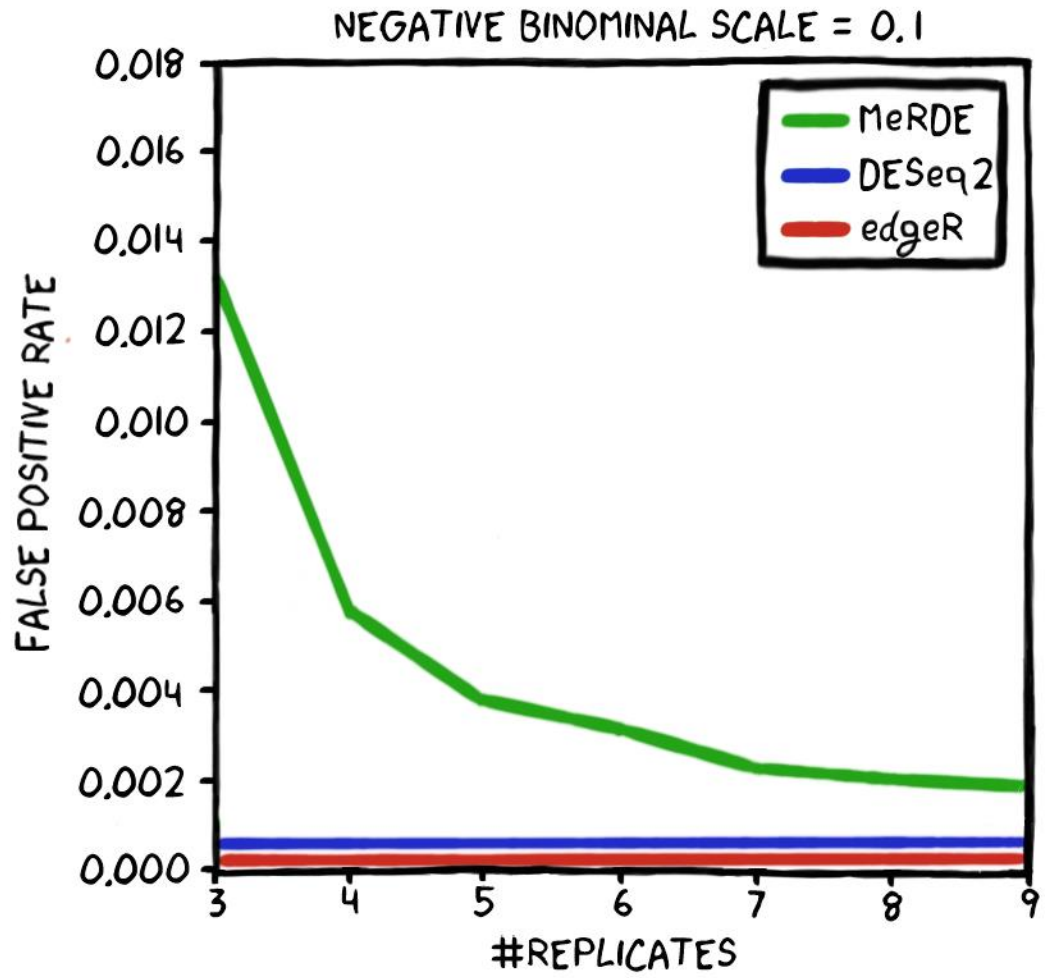

NEGATIVE BINOMINAL SCALE = 0.5
TRUE POSITIVE RATE
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
3
4
5
6
7
8
9
#REPLICATES
MeRDE
DESeq2
edgeR


NEGATIVE BINOMINAL SCALE = 0.5
FALSE POSITIVE RATE
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
3
4
5
6
7
8
9
#REPLICATES
MeRDE
DESeq2
edgeR


NEGATIVE BINOMINAL SCALE = 0.3
TRUE POSITIVE RATE
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
3
4
5
6
7
8
9
#REPLICATES
MeRDE
DESeq2
edgeR


NEGATIVE BINOMINAL SCALE = 0.3
FALSE POSITIVE RATE
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
3
4
5
6
7
8
9
#REPLICATES
MeRDE
DESeq2
edgeR


NEGATIVE BINOMINAL SCALE = 0.1
TRUE POSITIVE RATE
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
3
4
5
6
7
8
9
#REPLICATES
MeRDE
DESeq2
edgeR


NEGATIVE BINOMINAL SCALE = 0.1
FALSE POSITIVE RATE
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
3
4
5
6
7
8
9
#REPLICATES
MeRDE
DESeq2
edgeR