

# RNA-Seq Normalization in Babelomics 5

Marta R. Hidalgo

September 29th, 2016



**GDA**

International Course on  
Genomic Data Analysis



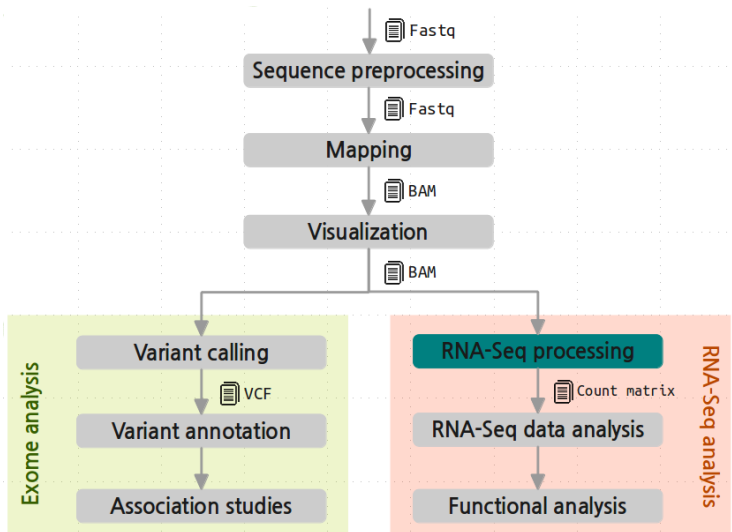
**PRINCIPE FELIPE**  
CENTRO DE INVESTIGACION

# Outline

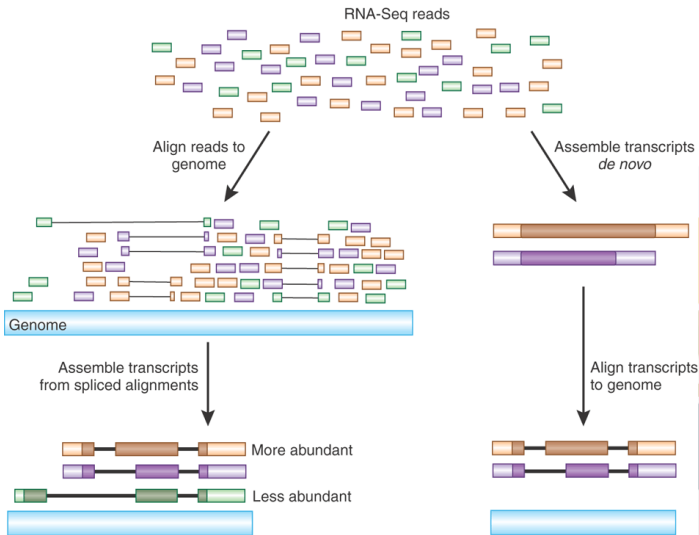
- 1 Introduction
- 2 Biases
- 3 Normalization methods
- 4 Normalization in Babelomics 5
- 5 Exercises



# Introduction



# Introduction



# Introduction

**What do we get?** A counts matrix (integer data)

A1BG	203	698	643	176	177	247	100	125
A1CF	0	0	0	0	0	0	0	1
A2BP1	398	245	263	540	7	1	1	13
A2LD1	89	149	81	265	312	823	217	803
A2M	55336	76480	49882	16376	67193	21941	14414	10123
A2ML1	67	3	6	444	170	28	84	17
A4GALT	59	870	206	326	72	344	458	2109
A4GNT	2	1	0	1	0	2	0	0
AAA1	2	0	0	0	1	0	0	0
AAAS	759	1061	2607	2129	1151	8130	1649	3447
AACS	784	566	1168	639	643	4281	383	1756
AACSL	1	2	1	0	1	0	0	0
AADAC	0	1	0	1	0	84	300	264

# Introduction

What do we get? A counts matrix (integer data)

A1BG	203	698	643	176	177	247	100	125
A1CF	0	0	0	0	0	0	0	1
A2BP1	398	245	263	540	7	1	1	13
A2LD1	89	149	81	265	312	823	217	803
A2M	55336	76480	49882	16376	67193	21941	14414	10123
A2ML1	67	3	6	444	170	28	84	17
A4GALT	59	870	206	326	72	344	458	2109
A4GNT	2	1	0	1	0	2	0	0
AAA1	2	0	0	0	1	0	0	0
AAAS	759	1061	2607	2129	1151	8130	1649	3447
AACS	784	566	1168	639	643	4281	383	1756
AACSL	1	2	1	0	1	0	0	0
AADAC	0	1	0	1	0	84	300	264

# Introduction

What do we get? A counts matrix (integer data)

A1BG	203	698	643	176	177	247	100	125
A1CF	0	0	0	0	0	0	0	1
A2BP1	398	245	263	540	7	1	1	13
A2LD1	89	149	81	265	312	823	217	803
A2M	55336	76480	49882	16376	67193	21941	14414	10123
A2ML1	67	3	6	444	170	28	84	17
A4GALT	59	870	206	326	72	344	458	2109
A4GNT	2	1	0	1	0	2	0	0
AAA1	2	0	0	0	1	0	0	0
AAAS	759	1061	2607	2129	1151	8130	1649	3447
AACS	784	566	1168	639	643	4281	383	1756
AACSL	1	2	1	0	1	0	0	0
AADAC	0	1	0	1	0	84	300	264

# Introduction

What do we get? A counts matrix (integer data)

A1BG	203	698	643	176	177	247	100	125
A1CF	0	0	0	0	0	0	0	1
A2BP1	398	245	263	540	7	1	1	13
A2LD1	89	149	81	265	312	823	217	803
A2M	55336	76480	49882	16376	67193	21941	14414	10123
A2ML1	67	3	6	444	170	28	84	17
A4GALT	59	870	206	326	72	344	458	2109
A4GNT	2	1	0	1	0	2	0	0
AAA1	2	0	0	0	1	0	0	0
AAAS	759	1061	2607	2129	1151	8130	1649	3447
AACS	784	566	1168	639	643	4281	383	1756
AACSL	1	2	1	0	1	0	0	0
AADAC	0	1	0	1	0	84	300	264





# Introduction

## Why normalizing?

- The technology introduces different biases
- We need to remove them to compare
  - Among genes in a sample
  - Among samples



# Biases



# Biases

- 1 Gene length
- 2 Library depth
- 3 RNA composition
- 4 Others

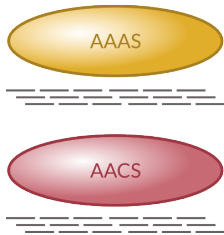
## Gene length

Larger genes get more reads

A1BG	203	698	643	176	177	247	100	125
A1CF	0	0	0	0	0	0	0	1
A2BP1	398	245	263	540	7	1	1	13
A2LD1	89	149	81	265	312	823	217	803
A2M	55336	76480	49882	16376	67193	21941	14414	10123
A2ML1	67	3	6	444	170	28	84	17
A4GALT	59	870	206	326	72	344	458	2109
A4GNT	2	1	0	1	0	2	0	0
AAA1	2	0	0	0	1	0	0	0
AAAS	759	1061	2607	2129	1151	8130	1649	3447
AACS	784	566	1168	639	643	4281	383	1756
AACSL	1	2	1	0	1	0	0	0
AADAC	0	1	0	1	0	84	300	264

# Biases

- 1 Gene length
- 2 Library depth
- 3 RNA composition
- 4 Others



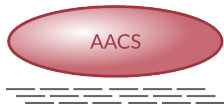
## Gene length

Larger genes get more reads



# Biases

- 1 Gene length
- 2 Library depth
- 3 RNA composition
- 4 Others



## Gene length

Larger genes get more reads

length = 1

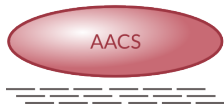
AAAS

length = 3

AACS

# Biases

- 1 Gene length
- 2 Library depth
- 3 RNA composition
- 4 Others



## Gene length

Larger genes get more reads

length = 1

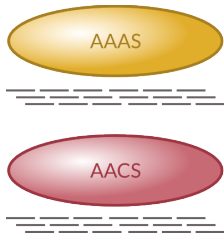
AAAS

length = 3

AACS

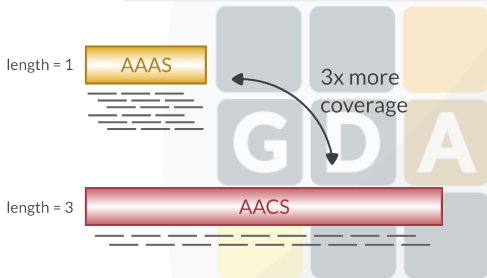
# Biases

- 1 Gene length
- 2 Library depth
- 3 RNA composition
- 4 Others



## Gene length

Larger genes get more reads

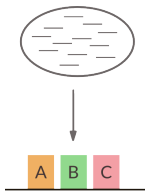


# Biases

- 1 Gene length
- 2 Library depth
- 3 RNA composition
- 4 Others

## Library depth

Deeper libraries give more reads



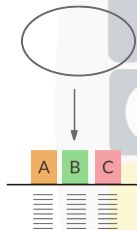
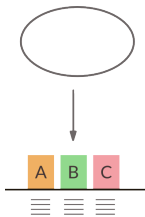


# Biases

- 1 Gene length
- 2 Library depth
- 3 RNA composition
- 4 Others

## Library depth

Deeper libraries give more reads



# Biases

- 1 Gene length
- 2 Library depth
- 3 RNA composition
- 4 Others



## RNA composition

A greedy gene steals reads from the others

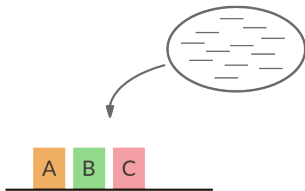


# Biases

- 1 Gene length
- 2 Library depth
- 3 RNA composition
- 4 Others

## RNA composition

A greedy gene steals reads from the others

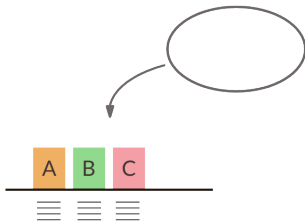


# Biases

- 1 Gene length
- 2 Library depth
- 3 RNA composition
- 4 Others

## RNA composition

A greedy gene steals reads from the others



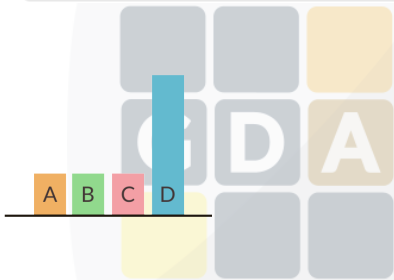
# Biases

- 1 Gene length
- 2 Library depth
- 3 RNA composition
- 4 Others



## RNA composition

A greedy gene steals reads from the others

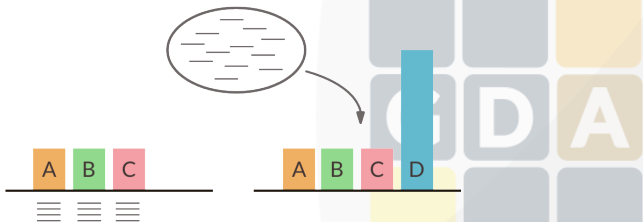


# Biases

- 1 Gene length
- 2 Library depth
- 3 RNA composition
- 4 Others

## RNA composition

A greedy gene steals reads from the others

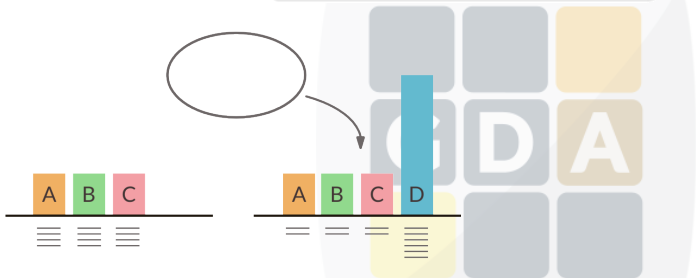


# Biases

- 1 Gene length
- 2 Library depth
- 3 RNA composition
- 4 Others

## RNA composition

A greedy gene steals reads from the others



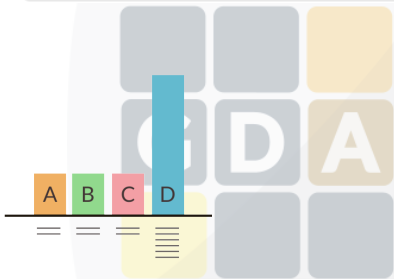
# Biases

- 1 Gene length
- 2 Library depth
- 3 RNA composition
- 4 Others



## RNA composition

A greedy gene steals reads from the others



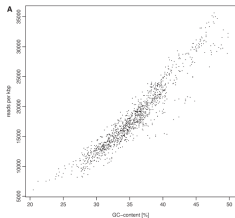


# Biases

- 1 Gene length
- 2 Library depth
- 3 RNA composition
- 4 Others

## Others

- GC-content
- Dinucleotide frequencies





Many biases

**We need to normalize!!**



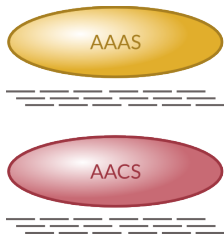


# Normalization methods



# Normalization methods

- 1 Gene length
- 2 Library depth
- 3 RPKM
- 4 TMM
- 5 Quantiles



## Gene length

Divide by gene length

length = 1

AAAS

length = 3

AACs



# Normalization methods

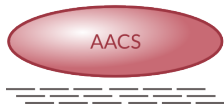
- 1 Gene length
- 2 Library depth
- 3 RPKM
- 4 TMM
- 5 Quantiles

Gene length

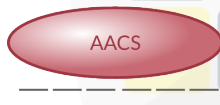
Divide by gene length



→  
divide by  
length = 1



→  
divide by  
length = 3



# Normalization methods

- 1 Gene length
- 2 Library depth
- 3 RPKM
- 4 TMM
- 5 Quantiles

## Gene length

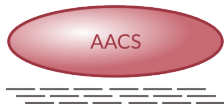
Divide by gene length



→  
divide by  
length = 1



←  
3x more  
coverage



→  
divide by  
length = 3

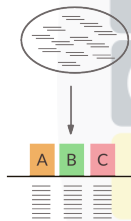
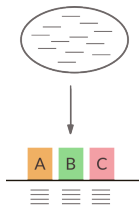


# Normalization methods

- 1 Gene length
- 2 Library depth
- 3 RPKM
- 4 TMM
- 5 Quantiles

Library depth

Divide by library depth

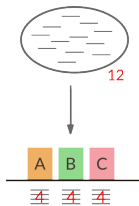


# Normalization methods

- 1 Gene length
- 2 Library depth
- 3 RPKM
- 4 TMM
- 5 Quantiles

Library depth

Divide by library depth



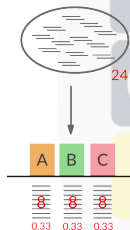
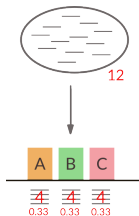


# Normalization methods

- 1 Gene length
- 2 Library depth
- 3 RPKM
- 4 TMM
- 5 Quantiles

Library depth

Divide by library depth



# Normalization methods

- 1 Gene length
- 2 Library depth
- 3 RPKM
- 4 TMM
- 5 Quantiles

## RPKM

- Reads per Kilobase per Million
- Remove gene length and library depth biases

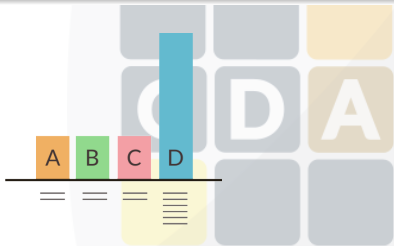
$$\text{RPKM} = \frac{\text{total exon reads}}{\text{mapped reads (millions)} * \text{exon length (KB)}}$$

# Normalization methods

- 1 Gene length
- 2 Library depth
- 3 RPKM
- 4 TMM
- 5 Quantiles

## TMM

- Trimmed Means of M-values
- Assumes only a few genes are DE
- Changes library depth

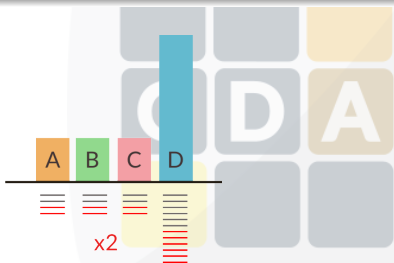


# Normalization methods

- 1 Gene length
- 2 Library depth
- 3 RPKM
- 4 TMM
- 5 Quantiles

## TMM

- Trimmed Means of M-values
- Assumes only a few genes are DE
- Changes library depth

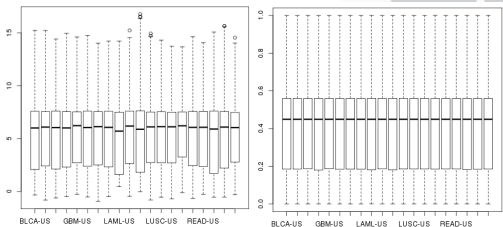



# Normalization methods

- 1 Gene length
- 2 Library depth
- 3 RPKM
- 4 TMM
- 5 Quantiles

## Quantiles

Makes all sample distributions the same





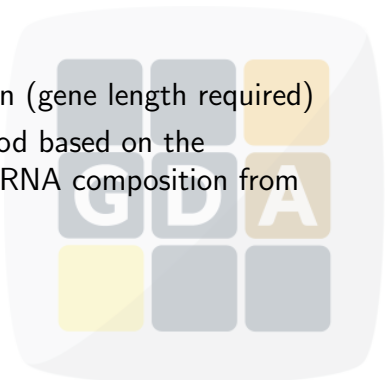
# Normalization in Babelomics 5



# Normalization in Babelomics 5

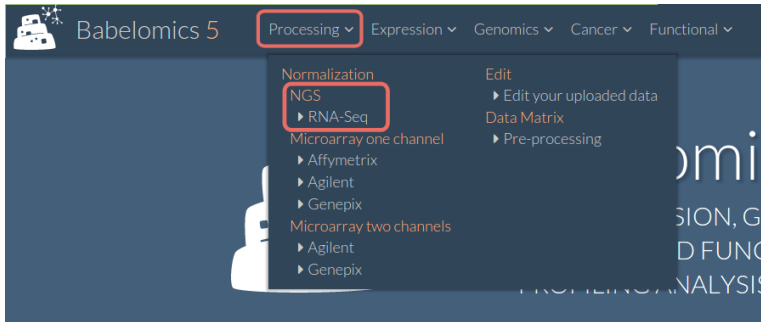
Available normalization methods in Babelomics 5

- 1 RPKM (gene length required)
- 2 TMM
- 3 TMM with gene length correction (gene length required)
- 4 Automatic selection of the method based on the diagnostic test for differences in RNA composition from NOISeq



# Normalization in Babelomics 5

Where can we find RNA-Seq normalization in Babelomics 5?




The screenshot displays the Babelomics 5 web application interface. The top navigation bar includes the Babelomics logo, the text 'Babelomics 5', and a series of dropdown menus: 'Processing', 'Expression', 'Genomics', 'Cancer', and 'Functional'. The 'Processing' dropdown menu is open, showing a list of options: 'Normalization', 'NGS', 'Microarray one channel', and 'Microarray two channels'. The 'NGS' option is highlighted with a red box, and its sub-menu is also visible, listing 'RNA-Seq' as the primary option, which is also highlighted with a red box. Other sub-options include 'Affymetrix', 'Agilent', and 'Genepix' under both 'Microarray one channel' and 'Microarray two channels'. To the right of the 'Processing' menu, there are sections for 'Edit' (with a sub-option 'Edit your uploaded data') and 'Data Matrix' (with a sub-option 'Pre-processing').



# Filling in the formular

## Select your data


The files must be on the server to select them.

You can upload files using the button  inside file browser.

WorkSpace/

## Select gene length file

The files must be on the server to select them.

You can upload files using the button  inside file browser.

WorkSpace/


## Normalization method

- Choose automatically the normalization method
- Choose manually the normalization method
  - TMM
  - RPKM

# Filling in the formular

## Job information

### Output folder

You can create folders using the button  + inside file browser.

File browser

WorkSpace/analysis ✕

### Job name

JobName

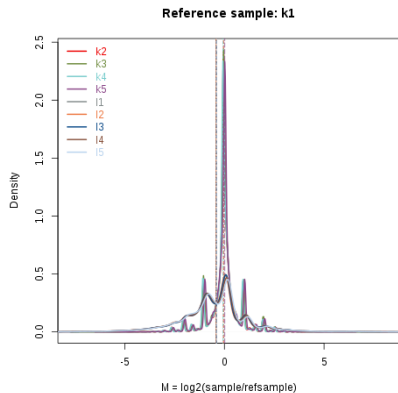
### Description

Job info...

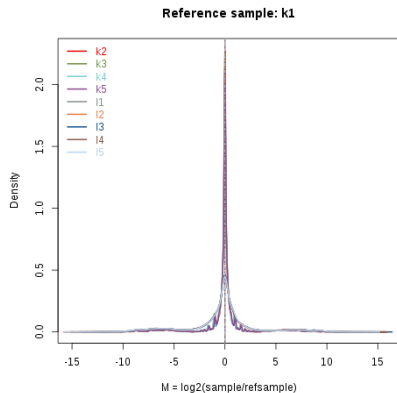
# The results

## RNA composition

RNA composition before normalization



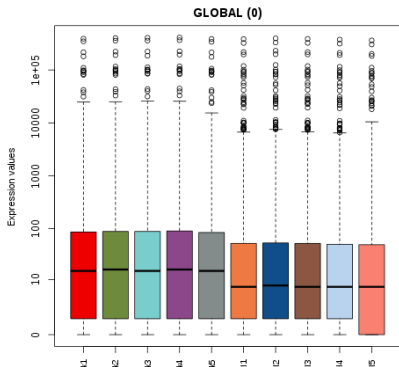
RNA composition after normalization



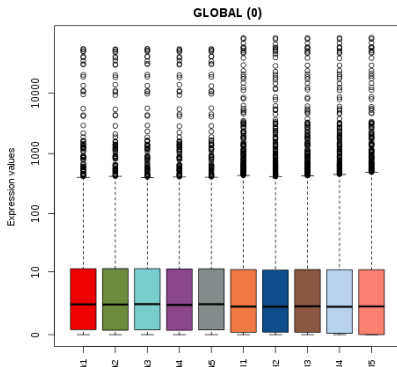
# The results

## Distribution of Expression values

Boxplot expression values before normalization



Boxplot expression values after normalization



# The results

## Table of Normalized values

File [normalized\\_results.txt](#)

#NAMES	k1	k2	k3	k4	k5
TSPAN6	42.11	39.49	39.02	34.59	42.55
TNMD	0	0	0	0.22	0
DPM1	13.17	16.31	17.22	15.5	16.21
SCYL3	1.54	2.12	2.21	1.88	2.31
C1orf112	1	1.15	0.77	1.04	1.82
FGR	2.25	3.19	2.07	2.24	1.2
FUCA2	37.84	41.24	39.91	38.24	33.78
GCLC	25.88	25.39	21.51	23.51	23.06
NFYA	4.62	4.59	4.03	4.16	4.69
STPG1	5.82	7.08	6.81	5.03	8

29405 Results

 Send to edit



# Exercises



# Normalization exercises

Go to **Babelomics 5**: <http://courses.babelomics.org/>

## Exercise 1

Run the Normalization Example (first button in the formular).  
Try all possible normalization methods:

- TMM with gene length
- TMM without gene length
- RPKM
- Automatic selection of the method

Compare the results. Which is the best normalization method?

For help, ask or visit the [normalization tutorial](#)

# Normalization exercises

## Exercise 2

Perform a normalization of the breast cancer data in the file *brca\_demo\_counts\_4babelomics.txt*

## Exercise 3

We will use a Kidney Renal Clear Cell carcinoma (KIRC) dataset from the TCGA

- 1 Go to the [GDA 2016 wiki](#)
- 2 Download the *kirc\_demo\_counts\_4babelomics.txt*
- 3 Upload this file to Babelomics 5
- 4 Normalize the data

For help, ask or visit the [normalization tutorial](#)