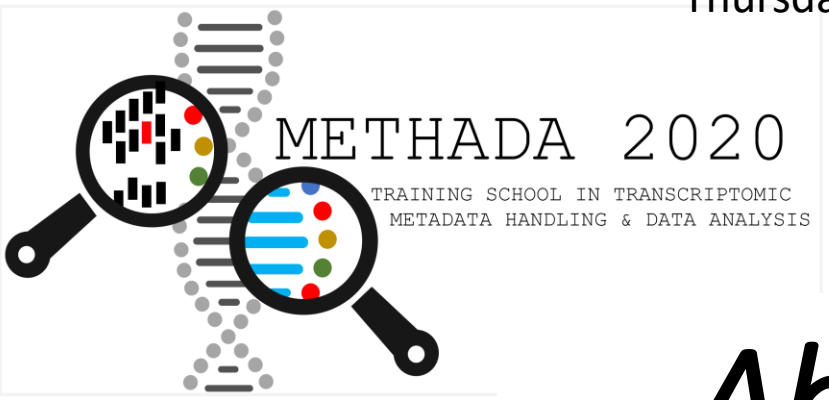
Thursday February 6, 2020, I2SysBio, Universitat de Valencia



About TomExpress

From samples, through raw counts and normalization, and beyond

Elie Maza
Laboratory of Genomics and Biotechnology of Fruits
Institut National Polytechnique de Toulouse
University of Toulouse, France

Plan

I. Introduction
II. About TomExpress
III. Normalization, a crucial step
IV. Hands-on session on TomExpress
V. Reference genes: a meta-analysis with TomExpress

Plan

I. Introduction
II. About TomExpress
III. Normalization, a crucial step
IV. Hands-on session on TomExpress
V. Reference genes: a meta-analysis with TomExpress

RNA-Seq beginnings

Second generation sequencing

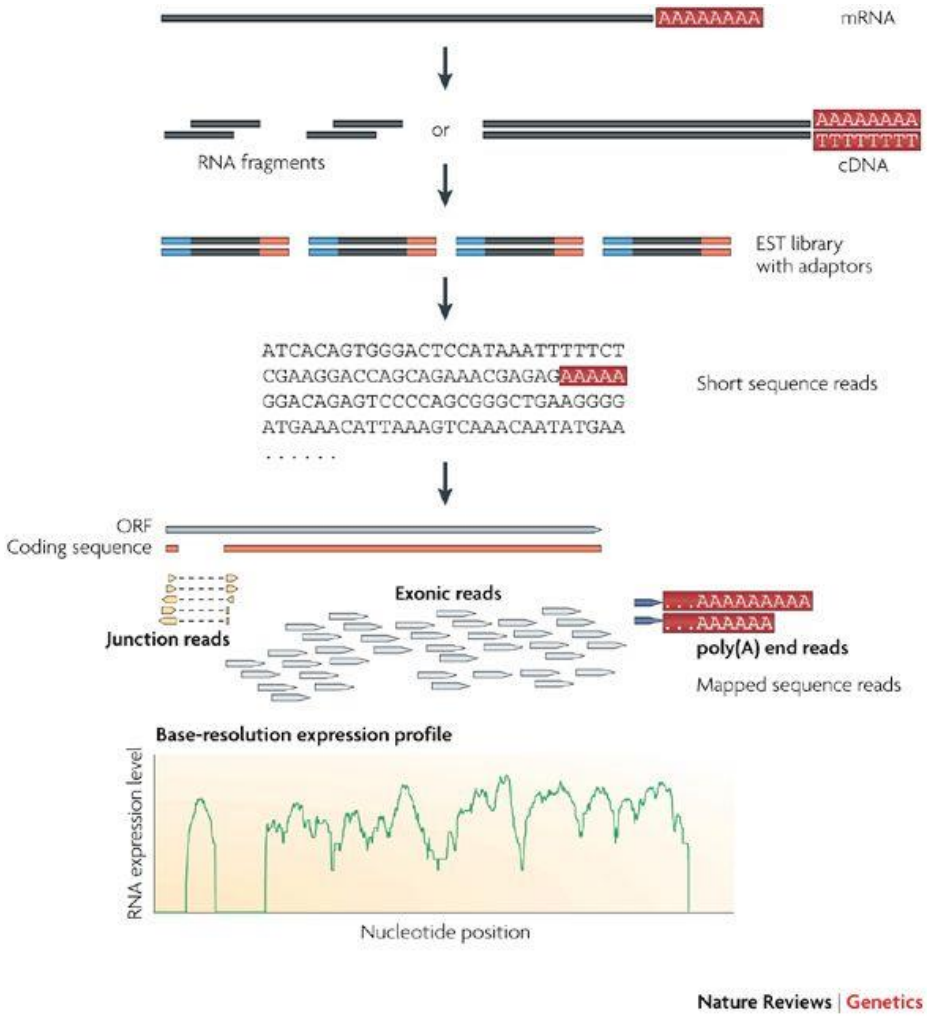
Next Generation Sequencing (NGS)

High Throughput Sequencing (HTS)

Shotgun sequencing

Whole Transcriptome Sequencing (WTS)

Wang et al. (2009)



- Z. Wang, M. Gerstein, and M. Snyder (2009) ***RNA-Seq: a revolutionary tool for transcriptomics***, *Nature Reviews Genetics*, 10(1):57-63.

FAIR principles: an important turning point!

- Explosion of publicly available transcriptomic data sets
- Hidden high scientific potential
    - Reanalysis
    - Integration
- Metadata handling issues
    - sample/experiment annotations
    - standardized protocols
    - …

FAIR principles: an important turning point!

- Explosion of publicly available transcriptomic data sets
- Hidden high scientific potential

  - Reanalysis
  - Integration

- Metadata handling issues

  - sample/experiment annotations
  - standardized protocols
  - …

Plan

I. Introduction
II. About TomExpress
III. Normalization, a crucial step
IV. Hands-on session on TomExpress
V. Reference genes: a meta-analysis with TomExpress

TomExpress v20

- 38 projects
- 433 biological conditions
- 1201 samples

http://tomexpress.gbfwebtools.fr

- M. Zouine, E. Maza, A. Djari, M. Lauvernier, P. Frasse, A. Smouni, J. Pirrello, and M. Bouzayen (2017) ***TomExpress, a unified tomato RNA-Seq platform for visualization of expression data, clustering and correlation networks***, *Plant J*, 92:727-735.

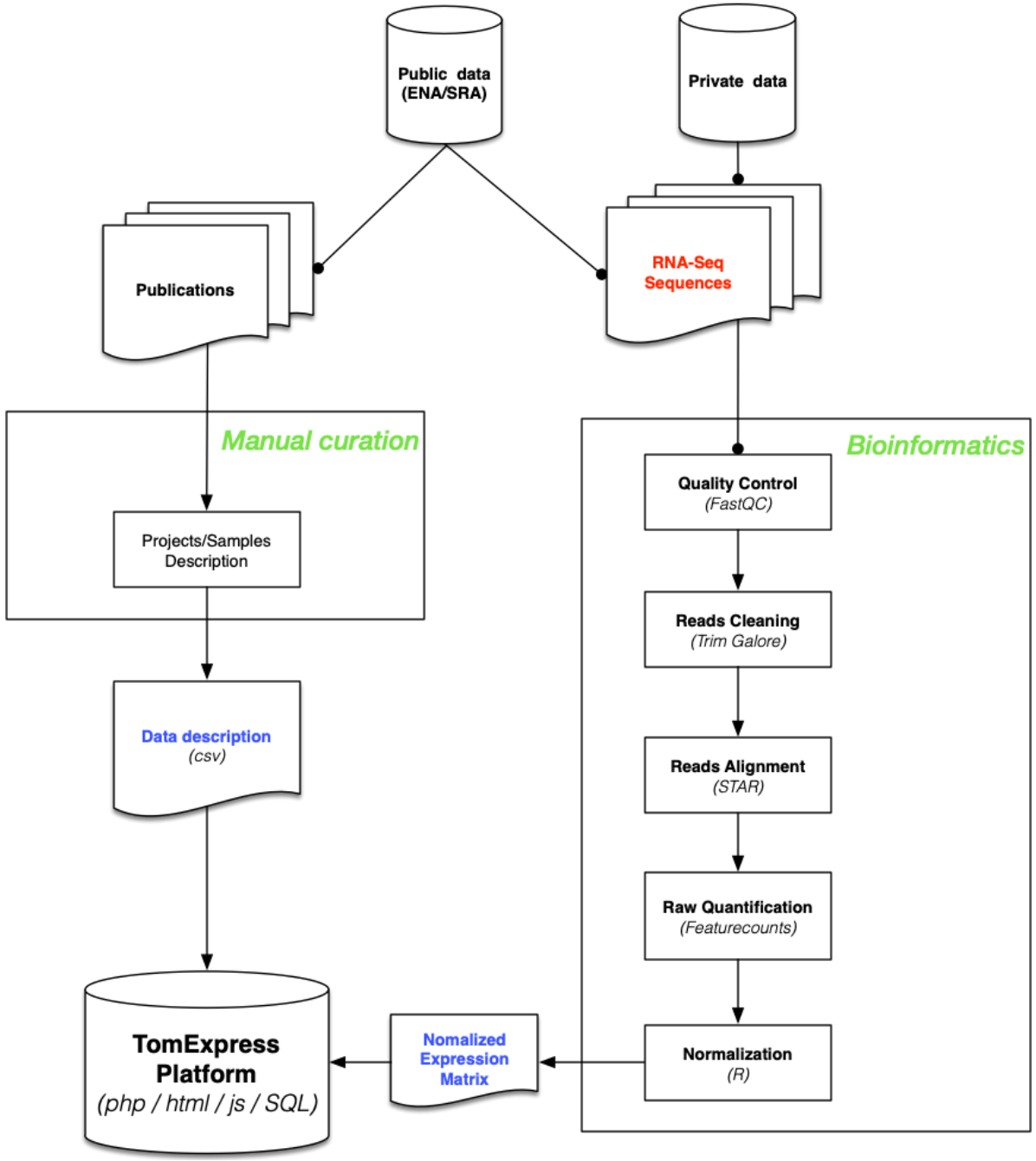
Goals and constraints

Goals:

- Observe the expression of a gene of interest along chosen organs, tissues, etc.
- Observe multiple gene expressions
- Build heatmaps of these expressions
- Find correlated genes
- Build gene co-expression networks
- Etc.

Constraints:

- Metadata handling for a homogeneous data description
- A unique gene expression quantification process
- A common normalization



Plan

I. Introduction
II. About TomExpress
III. Normalization, a crucial step
IV. Hands-on session on TomExpress
V. Reference genes: a meta-analysis with TomExpress

Normalization – Notations

Let X_{gk} the raw counts and μ_{gk} the unknown amount of transcript g per cell in condition k. Let L_g the size of transcript g. Then, library and transcriptome sizes can be written as follows:

$$N_k = \sum_{g=1}^{G} X_{gk} \quad \text{and} \quad S_k = \sum_{g=1}^{G} \mu_{gk} L_g$$

Then

$$E(X_{gk}) = \frac{\mu_{gk}\, L_g}{S_k} \times N_k$$

- M. D. Robinson, and A. Oshlack (2010) ***A scaling normalization method for differential expression analysis of RNA-seq data***, *Genome Biology*, 11(3):R25.

Normalization – Notations

Let X_{gk} the raw counts and μ_{gk} the unknown amount of transcript g per cell in condition k. Let L_g the size of transcript g. Then, library and transcriptome sizes can be written as follows:

$$N_k = \sum_{g=1}^{G} X_{gk} \quad \text{and} \quad S_k = \sum_{g=1}^{G} \mu_{gk} L_g$$

Then

$$E(X_{gk}) = \frac{\mu_{gk} \boxed{L_g}}{\boxed{S_k}} \times \boxed{N_k}$$

- M. D. Robinson, and A. Oshlack (2010) ***A scaling normalization method for differential expression analysis of RNA-seq data***, *Genome Biology*, 11(3):R25.

Normalization methods

i. $\mathrm{RPM}_{gk} = \frac{X_{gk}}{N_k} \times 10^6$ → $\mathrm{RPKM}_{gk} = \frac{X_{gk}}{N_k L_g} \times 10^9$ (FPKM)

ii. $\mathrm{RPK}_{gk} = \frac{X_{gk}}{L_g} \times 10^3$ → $\mathrm{TPM}_{gk} = \frac{X_{gk}/L_g}{\sum_{g=1}^{G} X_{gk}/L_g} \times 10^9$

iii. Upper quartile, Median, Quantile normalization

iv. TMM (edgeR) , RLE (DESeq2) , MRN

v. Housekeeping genes, Spike-ins

by library sizes

by adjustment of distributions

by relative transcriptome sizes

by controls

- C. Evans, J. Hardin, and D. M. Stoebel (2017) ***Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions***, *Briefings in Bioinformatics*, bbx008.

Normalization – How do TMM, LRE or MRN work?

We have $$\mathrm{E}(\mathrm{X}_{\mathrm{gk}}) = \frac{\mu_{\mathrm{gk}} \mathrm{L}_{\mathrm{g}}}{\mathrm{S}_{\mathrm{k}}} \times \mathrm{N}_{\mathrm{k}}$$

- E. Maza, P. Frasse, P. Senin, M. Bouzayen, and M. Zouine (2013) ***Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments***, *Commun Integr Biol.,* 6(6):e25849.
- E. Maza (2016) ***In Papyro Comparison of TMM (edgeR), RLE (DESeq2), and MRN Normalization Methods for a Simple Two-Conditions-Without-Replicates RNA-Seq Experimental Design***, *Front. Genet.* 7:164.

Normalization – How do TMM, LRE or MRN work?

We have $\mathrm{E}(\mathrm{X_{gk}}) = \frac{\mu_{gk} L_g}{S_k} \times N_k$ then $\frac{X_{gk}/N_k}{X_{g1}/N_1} \approx \frac{\mu_{gk}}{\mu_{g1}} \times \boxed{\frac{S_1}{S_k}}$

- E. Maza, P. Frasse, P. Senin, M. Bouzayen, and M. Zouine (2013) ***Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments***, *Commun Integr Biol.,* 6(6):e25849.
- E. Maza (2016) ***In Papyro Comparison of TMM (edgeR), RLE (DESeq2), and MRN Normalization Methods for a Simple Two-Conditions-Without-Replicates RNA-Seq Experimental Design***, *Front. Genet.* 7:164.

Normalization – How do TMM, LRE or MRN work?

We have $$E(X_{gk}) = \frac{\mu_{gk} L_g}{S_k} \times N_k$$ then $$\frac{X_{gk}/N_k}{X_{g1}/N_1} \approx \frac{\mu_{gk}}{\mu_{g1}} \times \boxed{\frac{S_1}{S_k}}$$

Assumption: Less than 50% of genes are up-regulated and less than 50% are down-regulated. Then

$$\underset{g=1,\ldots,G}{\text{median}} \left(\frac{X_{gk}/N_k}{X_{g1}/N_1} \right) \approx \underbrace{\underset{g=1,\ldots,G}{\text{median}} \left(\frac{\mu_{gk}}{\mu_{g1}} \right)}_{=1} \times \frac{S_1}{S_k} \approx \frac{S_1}{S_k}$$

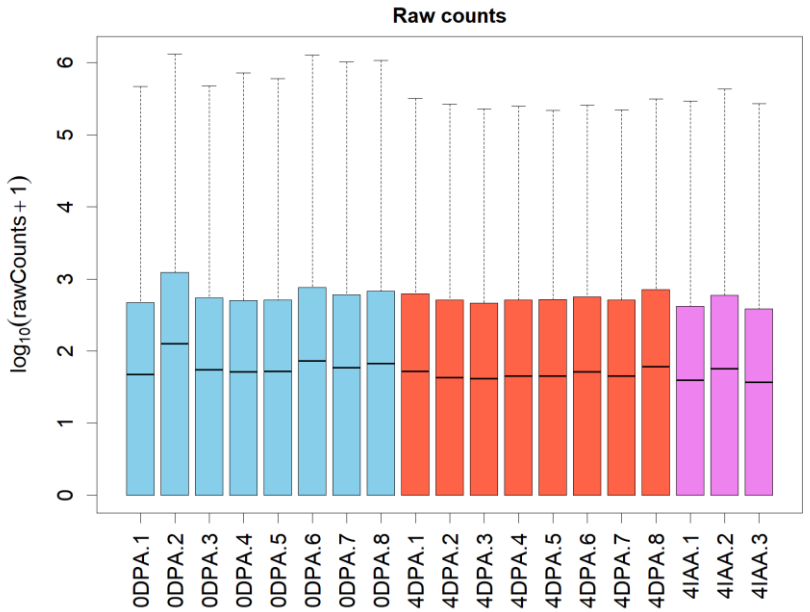
- E. Maza, P. Frasse, P. Senin, M. Bouzayen, and M. Zouine (2013) ***Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments***, *Commun Integr Biol.*, 6(6):e25849.
- E. Maza (2016) ***In Papyro Comparison of TMM (edgeR), RLE (DESeq2), and MRN Normalization Methods for a Simple Two-Conditions-Without-Replicates RNA-Seq Experimental Design***, *Front. Genet.* 7:164.

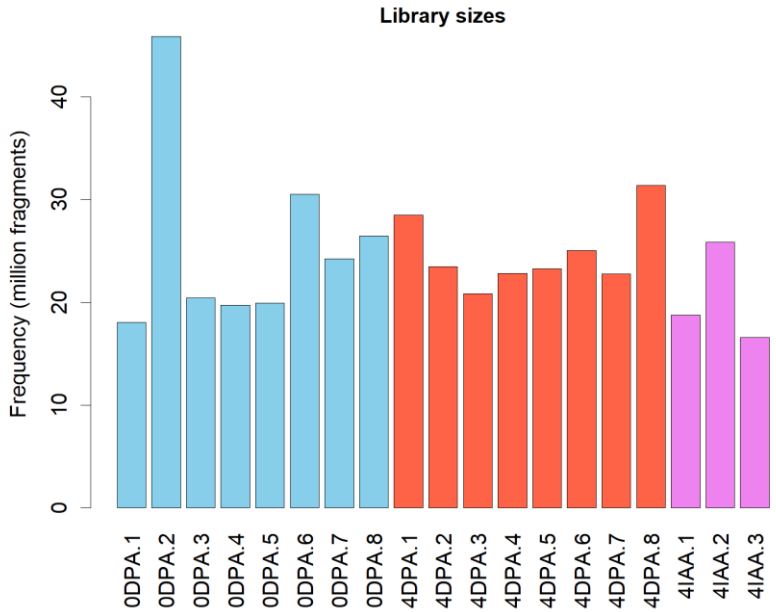
Normalization – Example 1: TOGE data

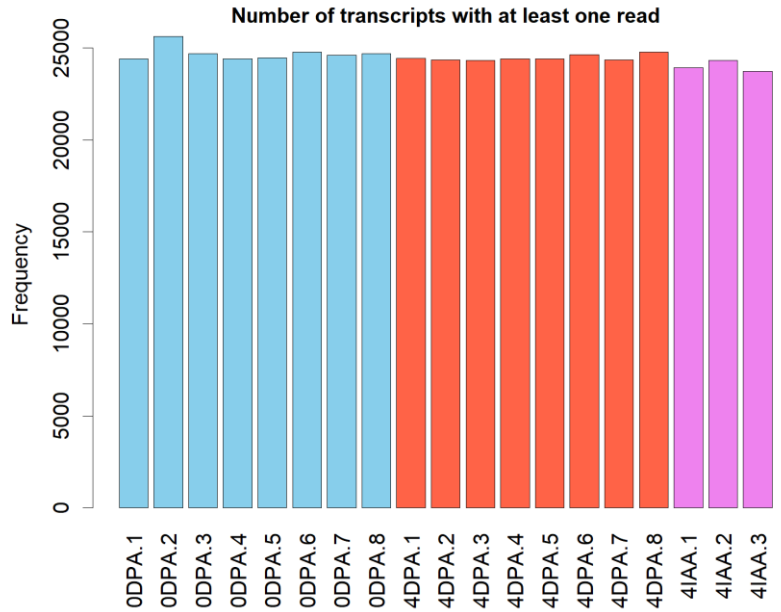
- Tomato plants (*Solanum lycopersicum L.* cv Micro-Tom) were grown in a culture chamber
- Three biological conditions or stages:
  - The ovary at anthesis stage (0 DPA) → 8 replicates
  - The young fruit after natural pollination (4 DPA) → 8 replicates
  - The young fruit after emasculation and Auxin treatment (4 DPA)→ 3 replicates
- mRNA sequencing: HiSeq 2500 System (2x125 bp paired-end sequences)

- S. Lamarre, P. Frasse, M. Zouine, D. Labourdette, E. Sainderichin, G. Hu, V. Le Berre-Anton, M. Bouzayen, and E. Maza (2018) ***Optimization of an RNA-Seq Differential Gene Expression Analysis Depending on Biological Replicate Number and Library Size***, *Front. Plant Sci*. 9:108.

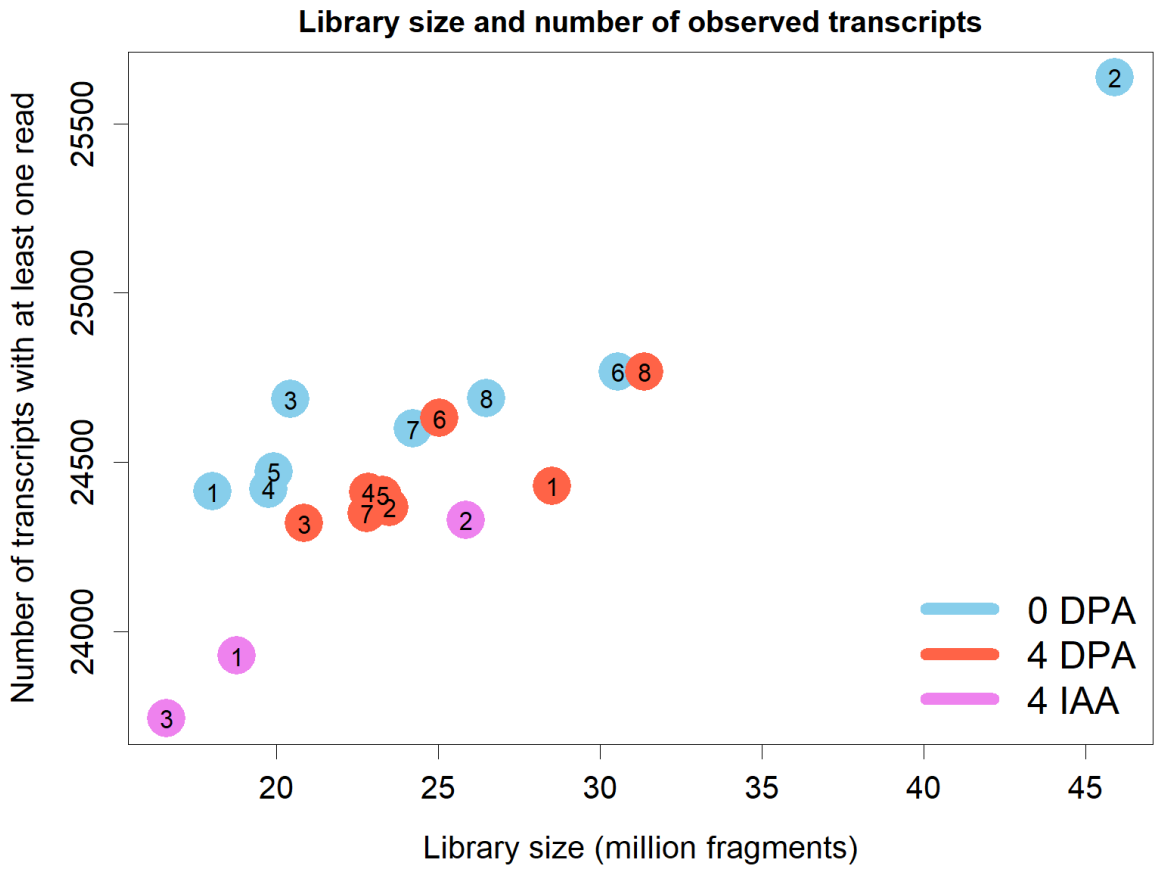
Normalization – Example 1: TOGE data

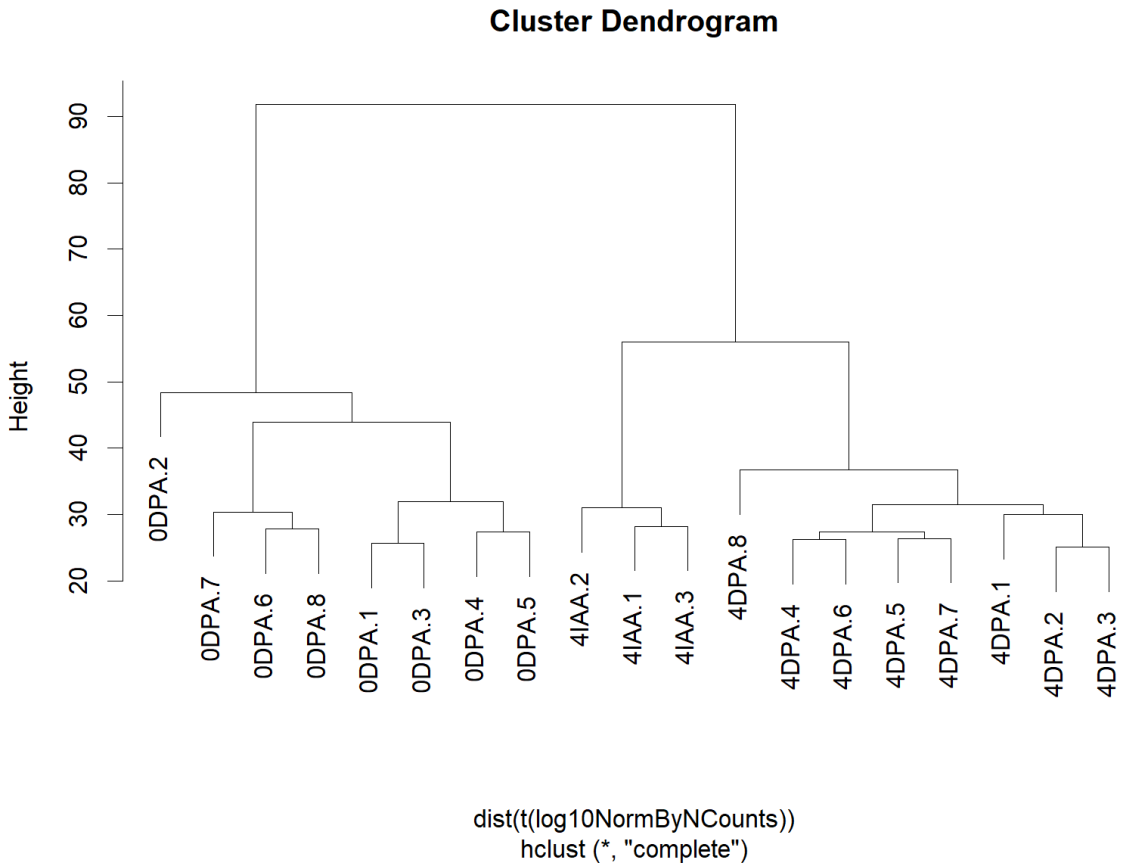






Normalization – Example 1: TOGE data



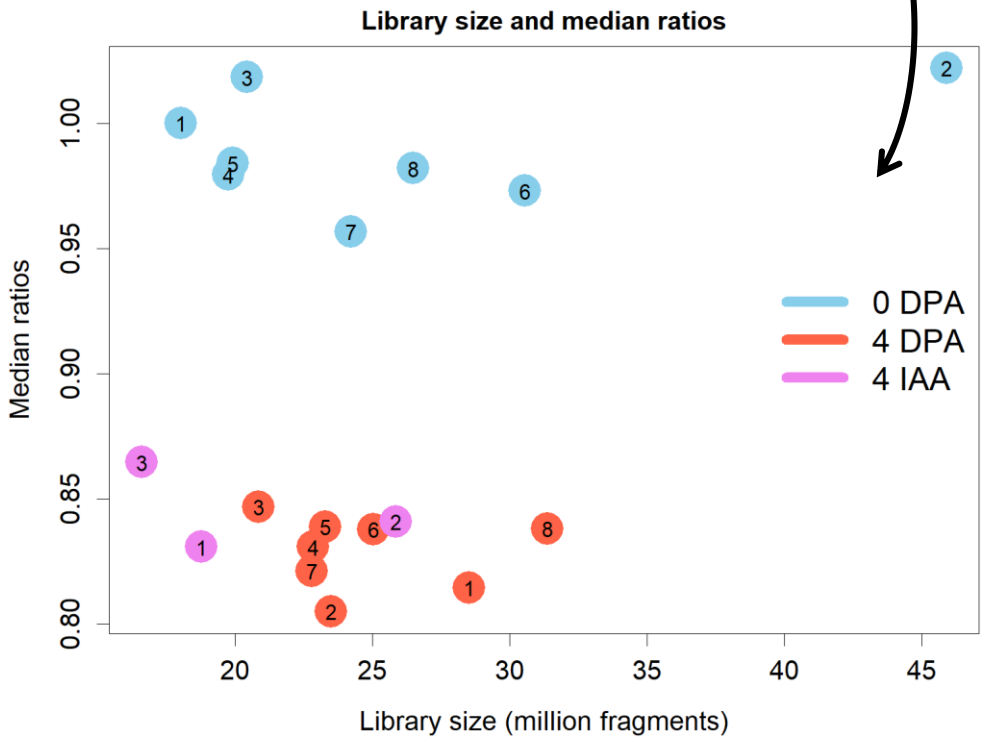


Normalization – Example 1: TOGE data

$$\underset{g=1,\ldots,G}{\text{median}}\left(\frac{X_{gk}/N_k}{X_{g1}/N_1}\right) \approx \frac{S_1}{S_k}$$

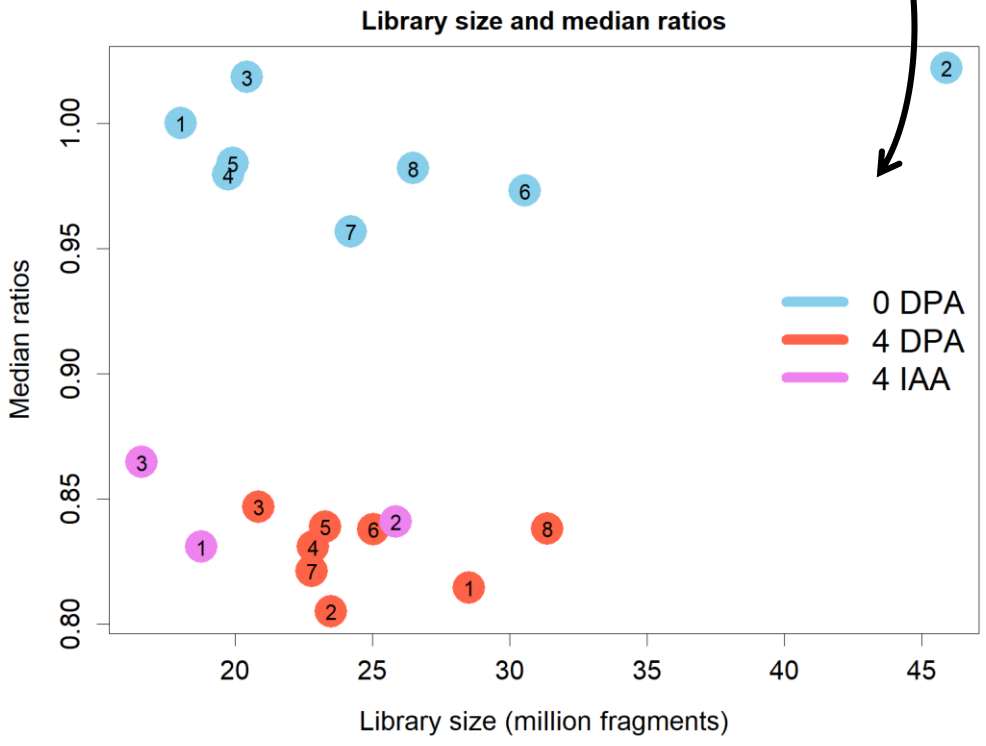
Normalization – Example 1: TOGE data

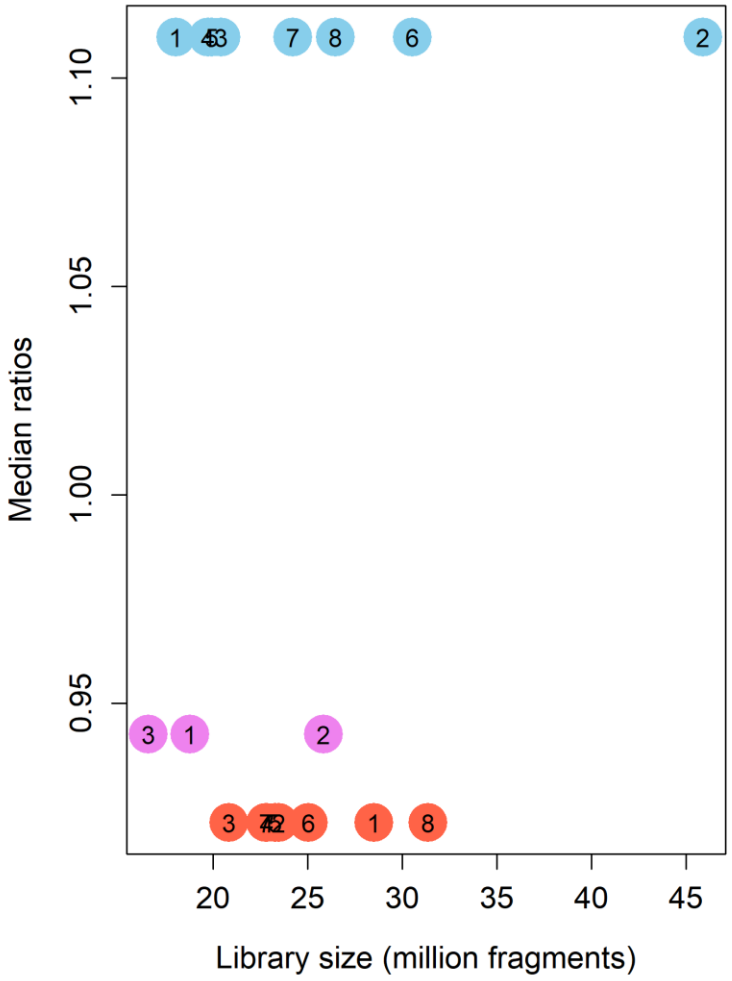
$$\underset{g=1,\ldots,G}{\text{median}}\left(\frac{X_{gk}/N_k}{X_{g1}/N_1}\right) \approx \frac{S_1}{S_k}$$

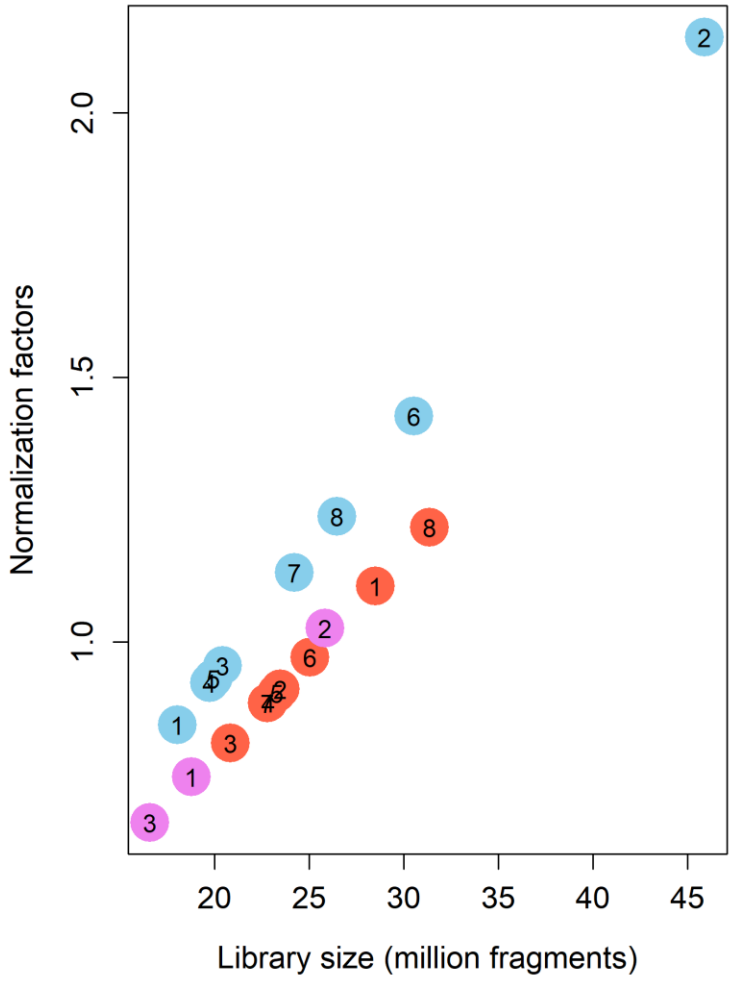


Normalization – Example 1: TOGE data

$$\underset{g=1,\ldots,G}{\text{median}}\left(\frac{X_{gk}/N_k}{X_{g1}/N_1}\right) \approx \frac{S_1}{S_k}$$







Normalization – Example 2: Ploidy data

- Cherry tomato (*Solanum lycopersicum Mill.* cv Wva106)
- Nuclei from pericarp cells of 30 DPA fruits
- 4 ploidy levels: 4C, 8C, 16C and 32C
- 3 replicates per ploidy level

- J. Pirrello, C. Deluche, N. Frangne, F. Gévaudant, E. Maza, A. Djari, M. Bourge, J.-P. Renaudin, S. Brown, C. Bowler, M. Zouine, C. Chevalier, and N. Gonzalez (2018) ***Transcriptome profiling of sorted endoreduplicated nuclei from tomato fruits: how the global shift in expression ascribed to DNA ploidy influences RNA-Seq data normalization and interpretation***, *Plant J*, 93:387-398.

Normalization – Example 2: Ploidy data

« I found no DE genes by performing a DE analysis with DESeq2!? »

Pirrello et al. (2018)

	4C	8C	16C	32C
4C	0	47	254	325
8C	45	0	26	71
16C	525	69	0	0
32C	715	270	12	0

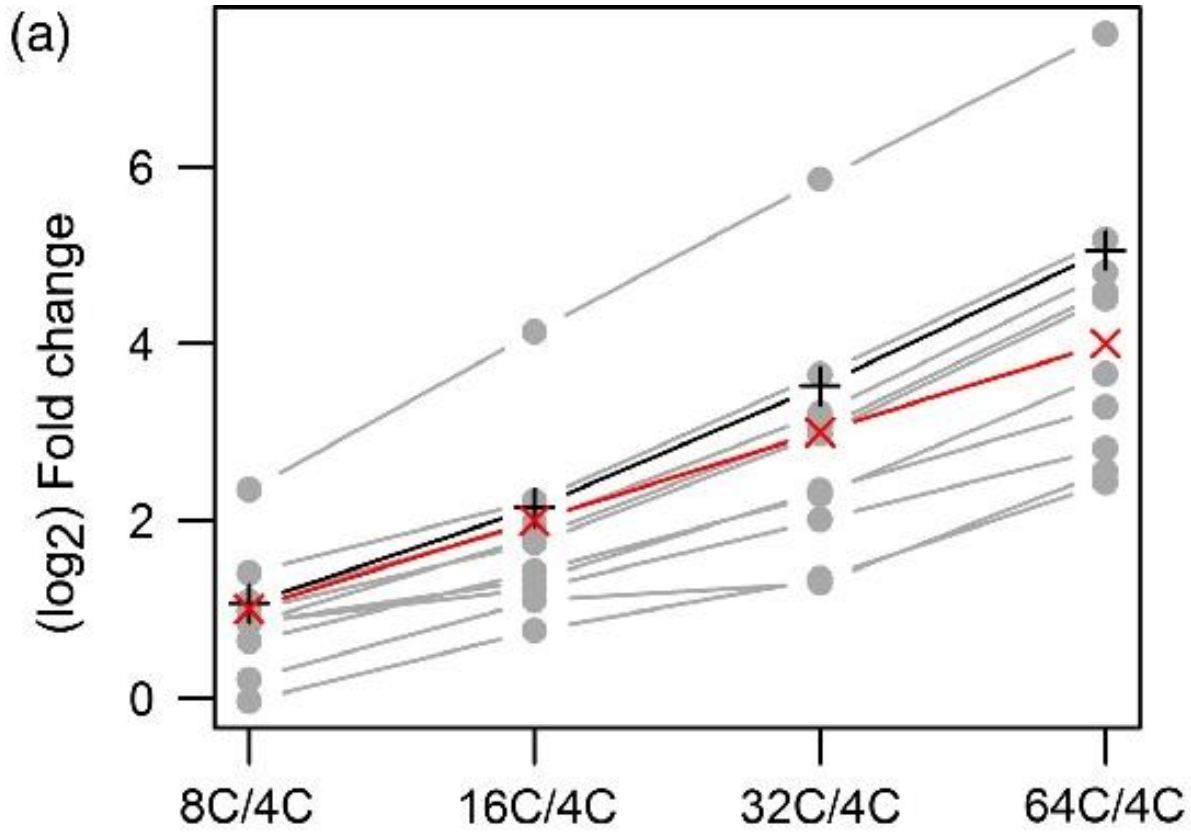
Normalization – Example 2: Ploidy data

« I found no DE genes by performing a DE analysis with DESeq2!? »

Pirrello et al. (2018)

	4C	8C	16C	32C
4C	0	47	254	325
8C	45	0	26	71
16C	525	69	0	0
32C	715	270	12	0

Pirrello et al. (2018)



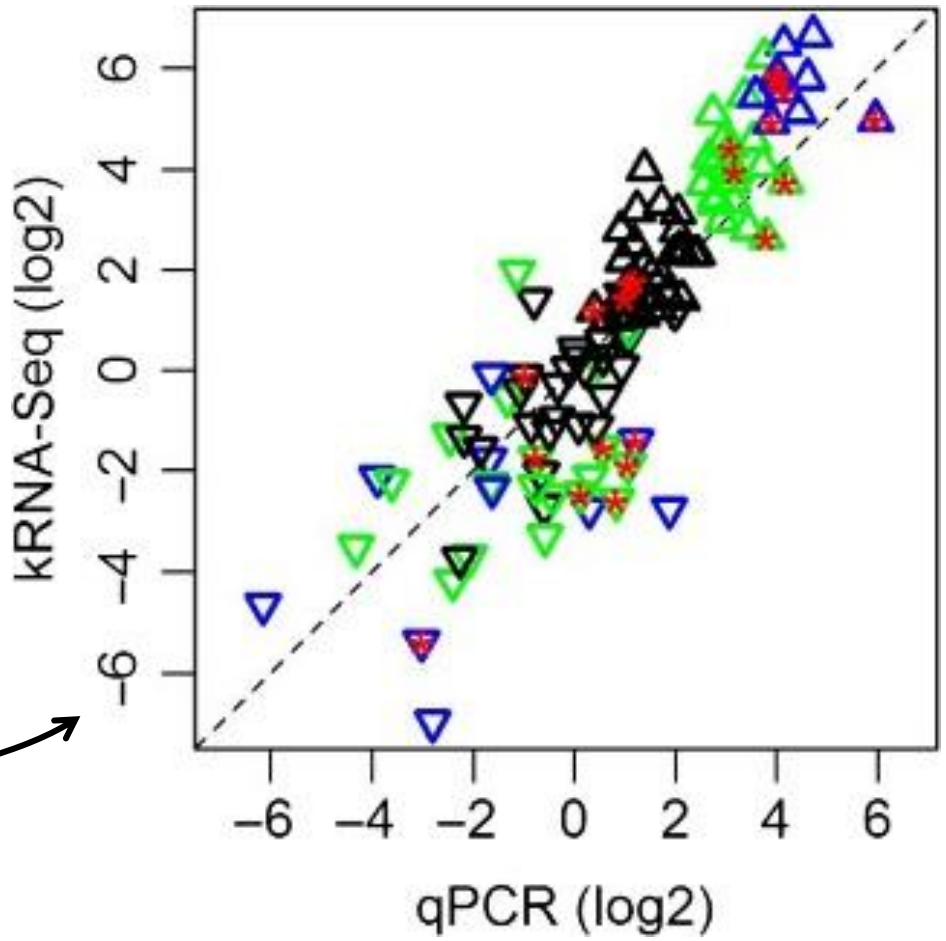
Normalization – Example 2: Ploidy data

Gene expression fold change between levels $2^{p}C$ and $2^{q}C$ with q>p:

Pirrello et. al (2018)

$$\frac{\mu_{gq}}{\mu_{gp}} = 2^{q-p} \times NFC_g$$

Pirrello et. al (2018)



Plan

I. Introduction
II. About TomExpress
III. Normalization, a crucial step
IV. Hands-on session on TomExpress
V. Reference genes: a meta-analysis with TomExpress

Example 1

Genes of interest

Gene name (ITAG, SGN)	Description	Symbol
Solyc03g118290	Auxin response factor	SlARF2A
Solyc12g042070	Auxin response factor	SlARF2B
Solyc03g031860	Phytoene synthase 1	PSY1
Solyc05g012020	Ripening inhibitor	RIN
Solyc02g077850	Colorless non-ripening	CNR
Solyc10g006880	Non ripening	NOR
Solyc07g055920	Tomato AGAMOUS-like 1	TAGL1
Solyc09g007870	Ethylene insensitive 2	EIN2
Solyc09g075440	Never ripe, Ethylene receptor	NR, ETR3
Solyc01g104340	Green ripe	GR

Example 2

- Choose a gene of interest from you plant of interest (grape)
- Find the ortholog on tomato (Sequence similarity)
- Observe its expression profile on TomExpress
- Find correlated genes
- Build heatmaps/networks of these genes

Plan

I. Introduction
II. About TomExpress
III. Normalization, a crucial step
IV. Hands-on session on TomExpress
V. Reference genes: a meta-analysis with TomExpress

Ref. genes – Background

Reference gene criteria:

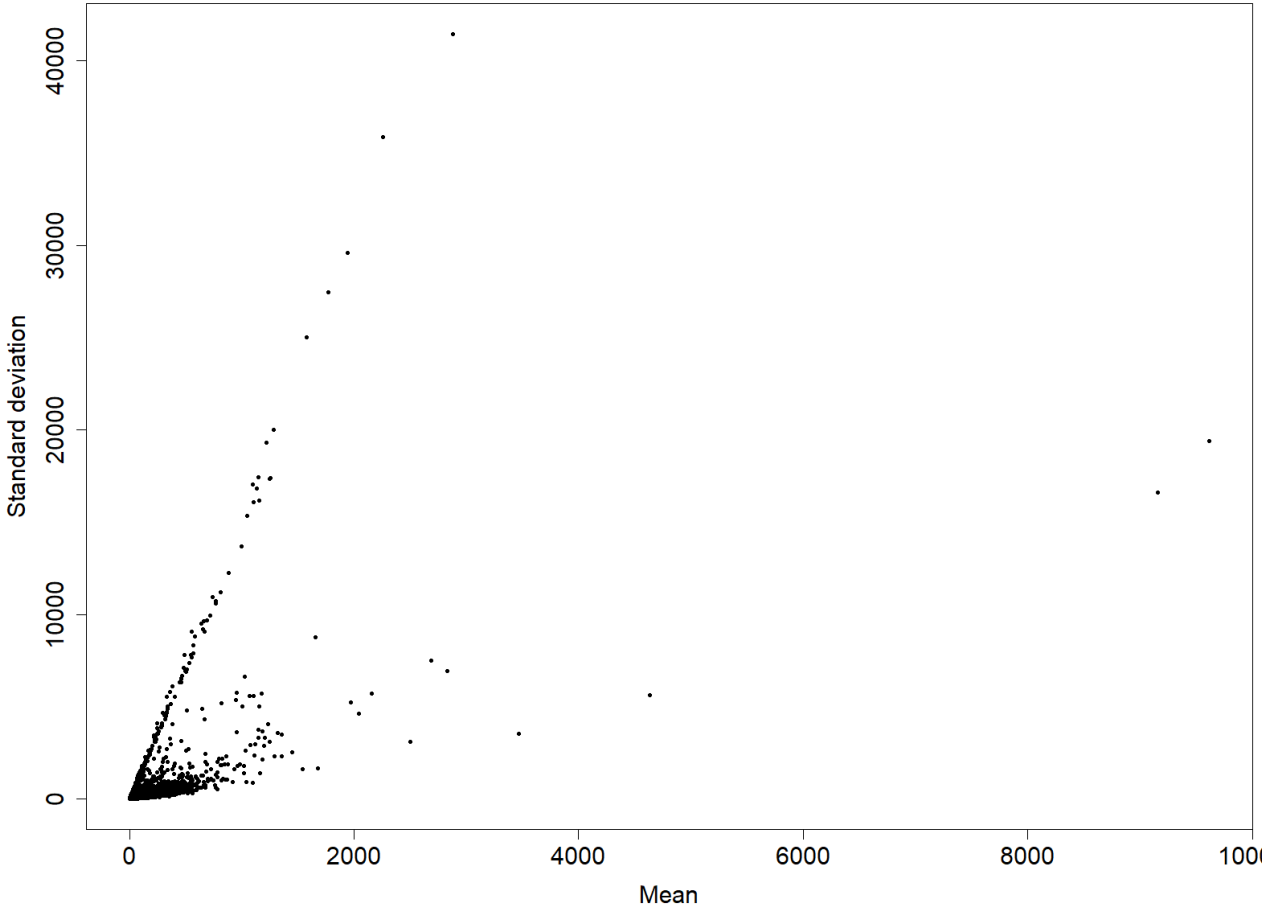
i. An expression level unaffected by experimental factors.
ii. Minimal variability between tissues and physiological states.
iii. A similar threshold cycle with a gene of interest.

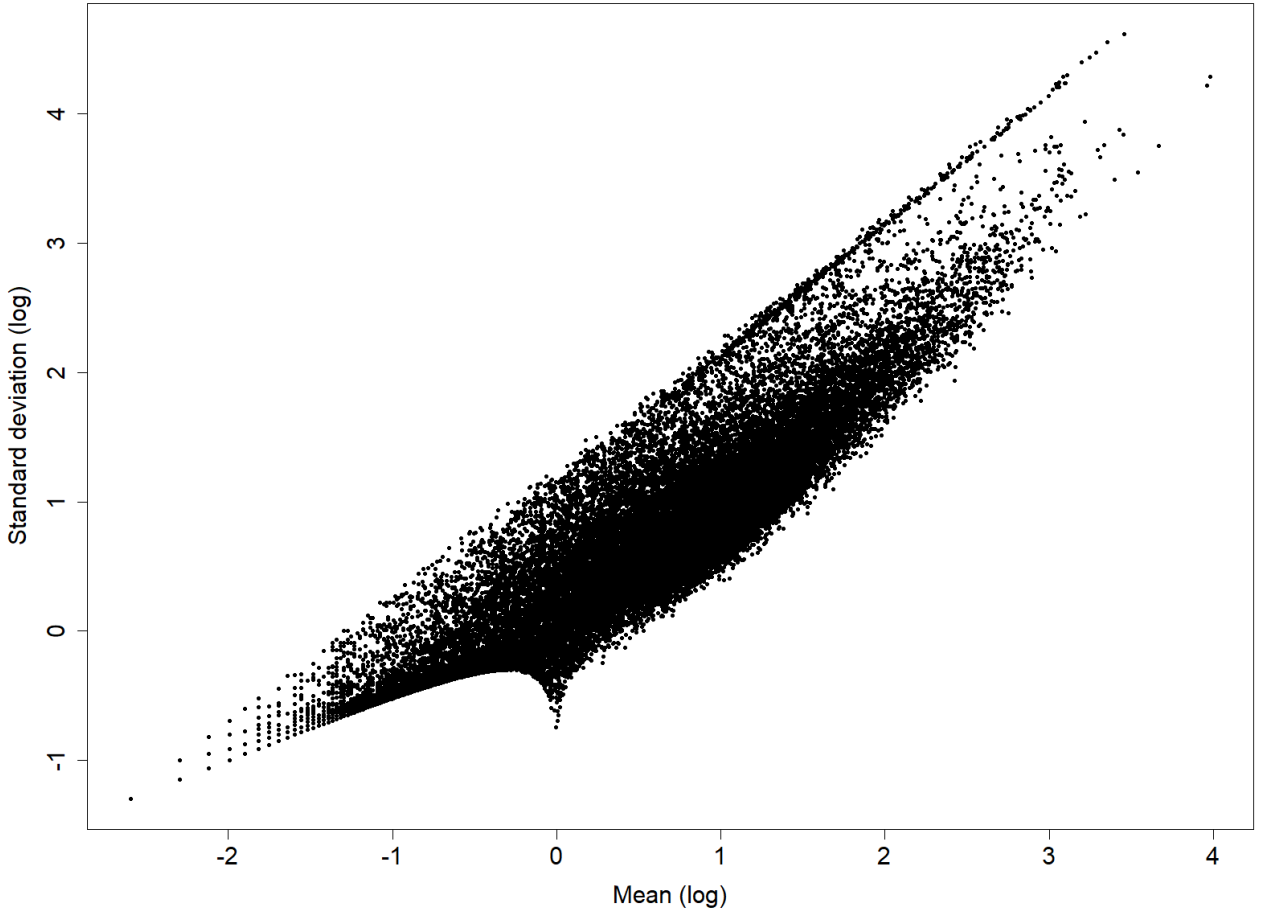
“ In many reports can be found a fundamental rule that there is no universal reference gene and when analyzing dozens of cited examples for expression variability between the tissues, caused by stress factors or tumors and diseases, it is difficult to disagree with this statement. ”

- Does TomExpress reflect this fact?
- Can we propose better reference genes than classical ones using TomExpress?

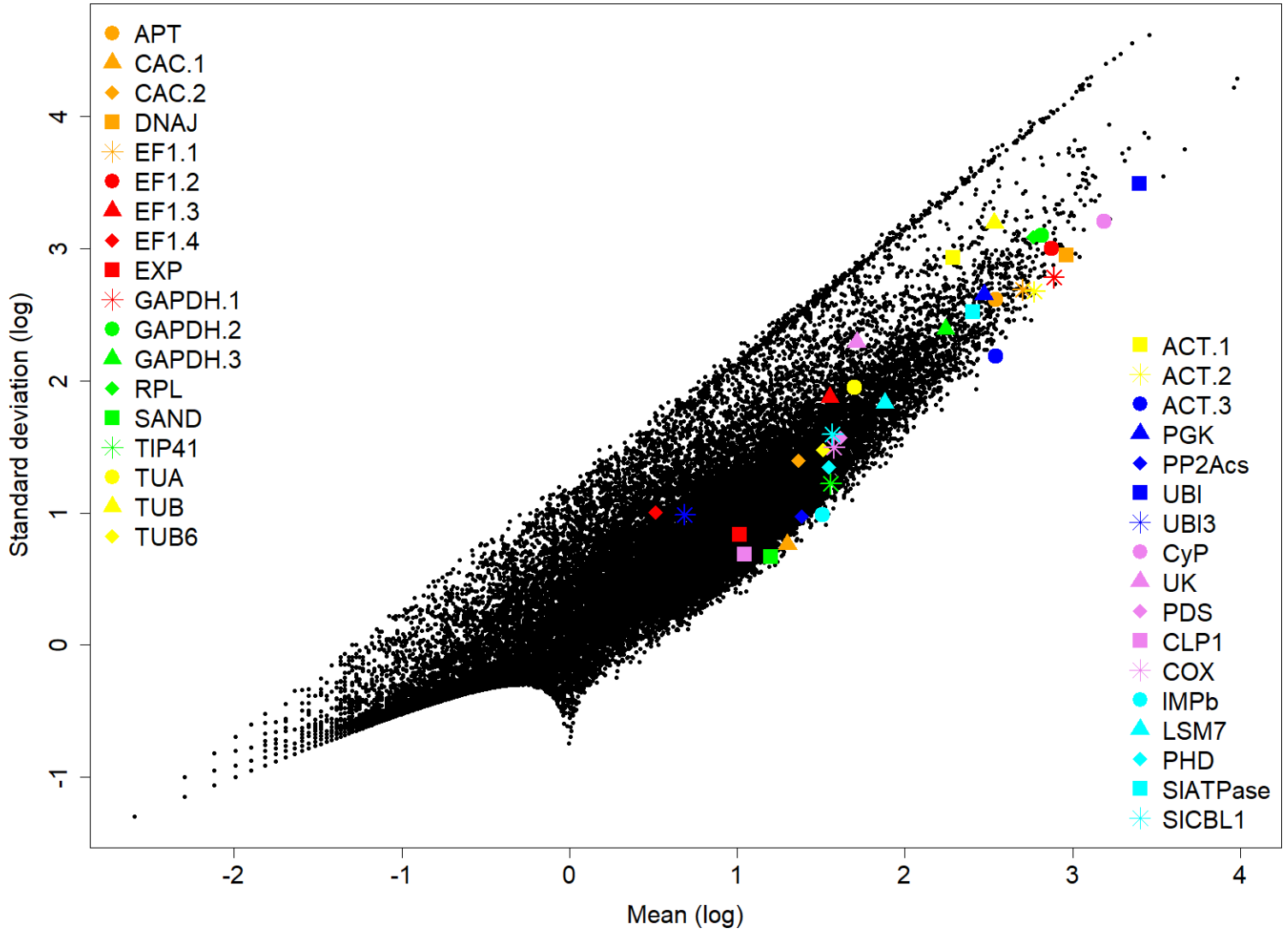
- B. Kozera, and M. Rapacz (2013) ***Reference genes in real-time PCR***, *Journal of Applied Genetics*, 54(4):391-406.

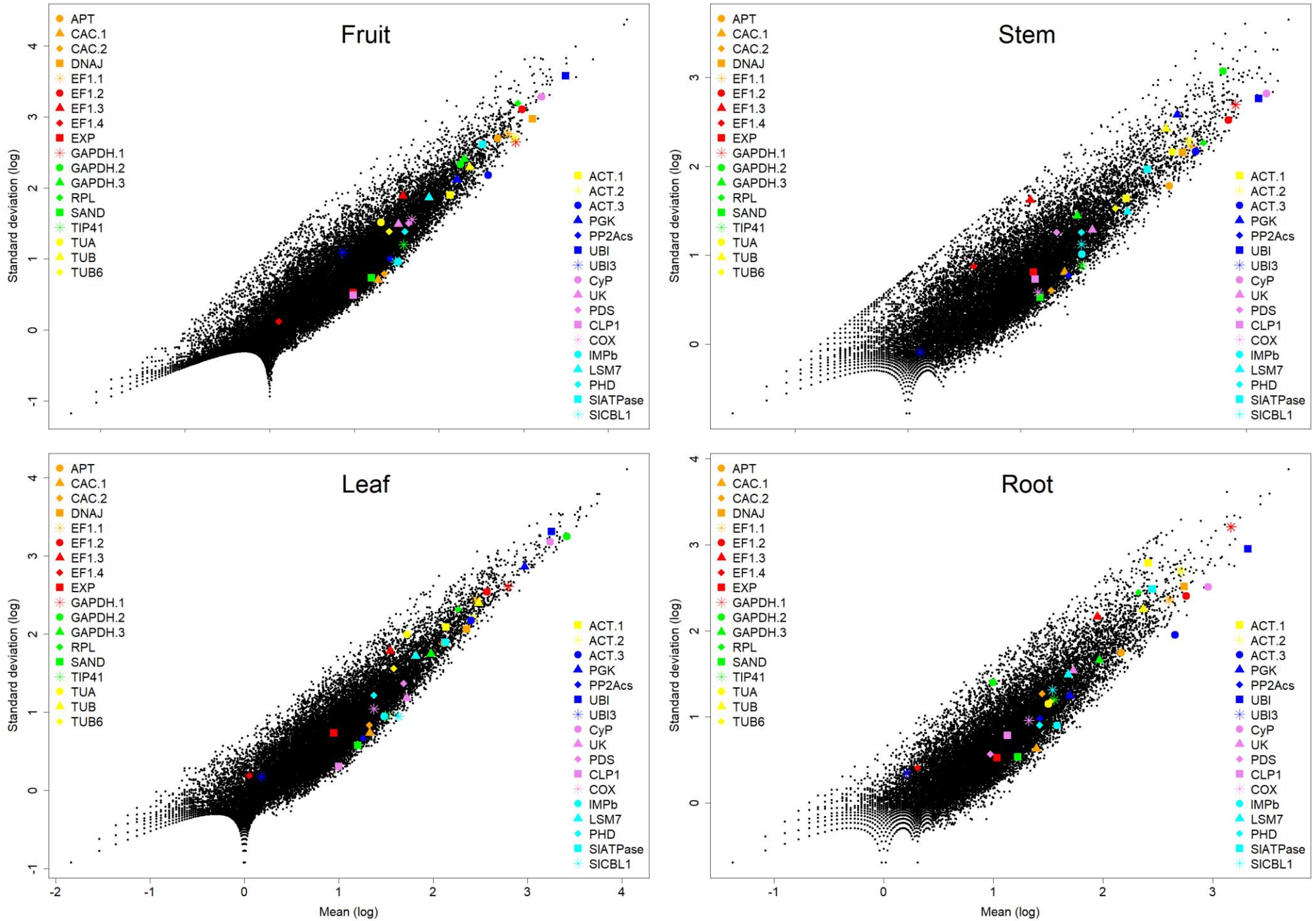
Ref. genes – Gene means and variances



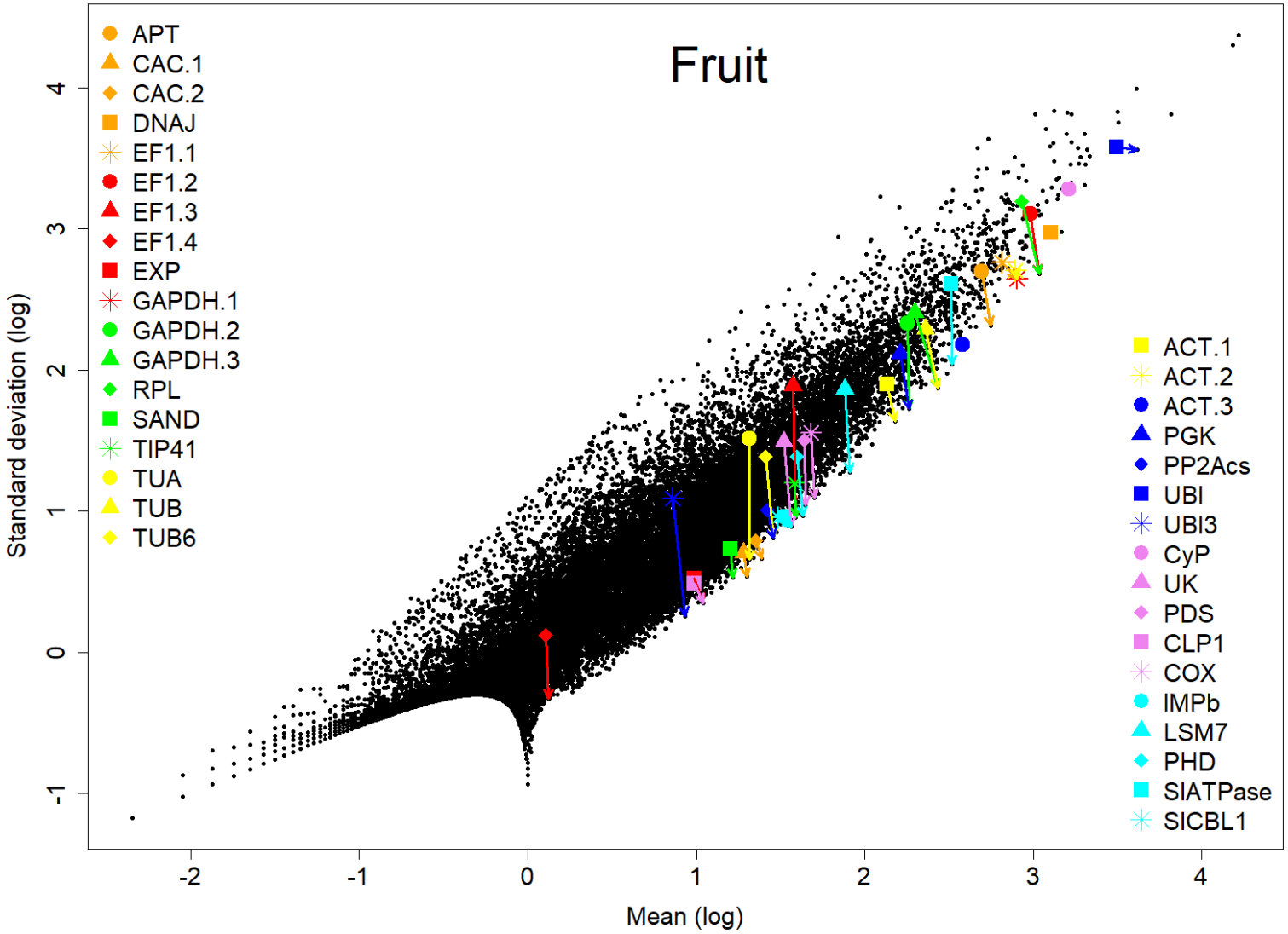


Ref. genes – Classical reference genes

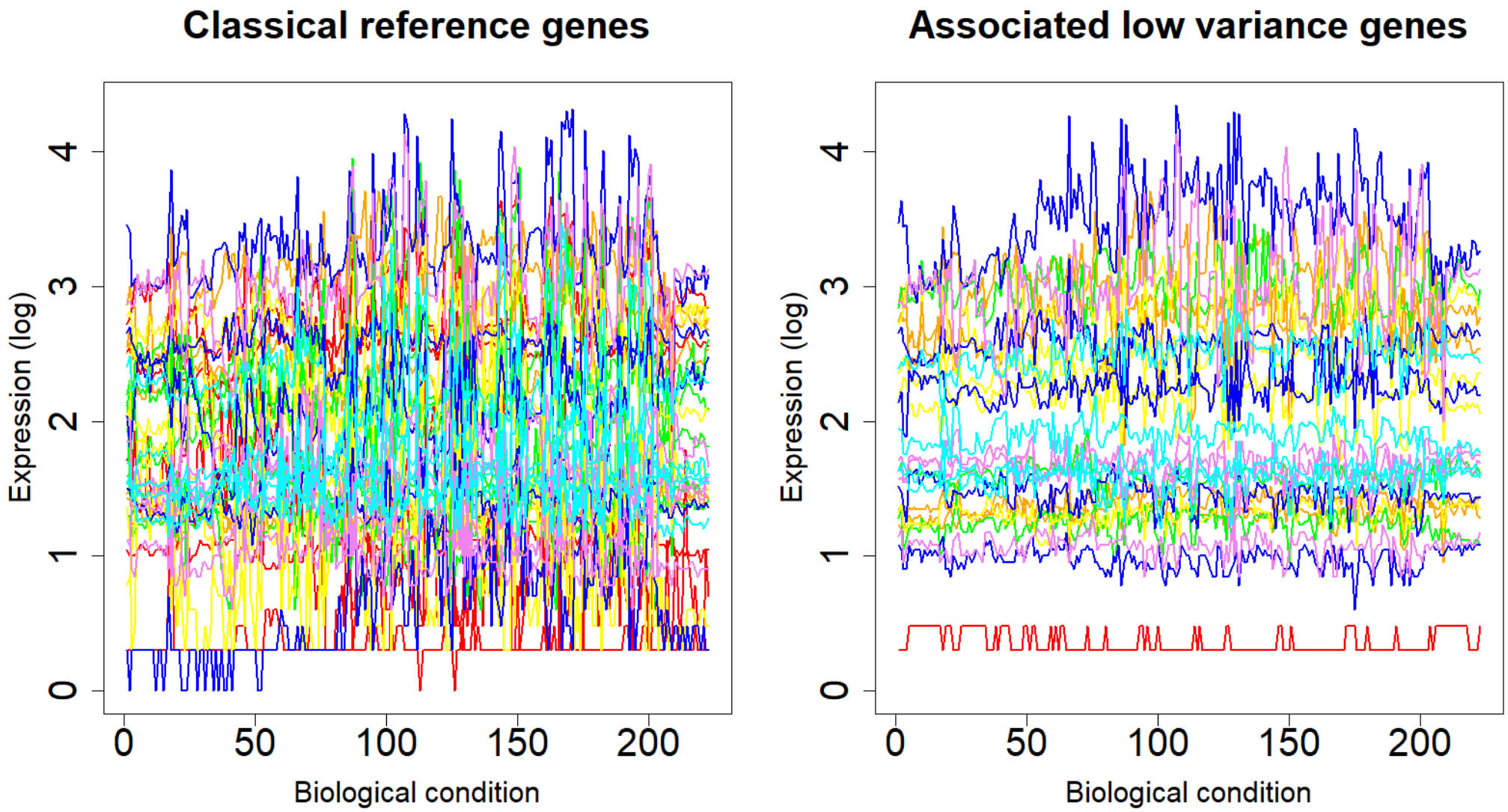



Fruit
Stem
Leaf
Root
Standard deviation (log)
Mean (log)
APT
CAC.1
CAC.2
DNAJ
EF1.1
EF1.2
EF1.3
EF1.4
EXP
GAPDH.1
GAPDH.2
GAPDH.3
RPL
SAND
TIP41
TUA
TUB
TUB6
ACT.1
ACT.2
ACT.3
PGK
PP2Acs
UBI
UBI3
CyP
UK
PDS
CLP1
COX
IMPb
LSM7
PHD
SIATPase
SICBL1

Ref. genes – Lowest variance genes



Ref. genes – Lowest variance genes



GBF lab members

Technicians

Dominique Saint-Martin

Lydie Tessarotto Lemonnier

Post-docs

Guojian Hu

Baowen Huang

PhD Students

Yi Chen

Professors and associate professors

Mondher Bouzayen (Pr, Director)

Christian Chervin (Pr)

Jean-Claude Pech (Pr)

Anne Bernadac (MCF)

Elie Maza (MCF)

Julien Pirrello (MCF)

Benoît Van-Der-Rest (MCF)

Mohamed Zouine (MCF, Vice director)

Engineers

Pierre Frasse (IR)

Isabelle Mila (AI)

Anis Djari (IE)

ETHYLENE 2020 : XIIth Ethylene symposium

FRANCE, Toulouse June 29th to July 3rd 2020

Thanks

www.ethylene2020.com

Local Organisers

Julien PIRRELLO
Christian CHERVIN
Mondher BOUZAYEN

RNA-Seq – R packages for DE analysis

Lamarre et al. (2018)

R-package or method	Reference	Dec 2013	Jan 2015	Jan 2016	Jan 2017	Oct 2017	Oct 2017 (%)	Distribution	Normalization	Bayesian
edgeR	(Robinson et al., 2010)	430	982	1854	3040	4450	22,00	negative binomial	TMM	no
Cufflinks (*Cuffdiff* *)	(Trapnell et al., 2010)	861	1648	2446	3300	4283	21,17	Poisson	FPKM (geometric)	yes
DESeq	(Anders and Huber, 2010)	607	1395	2299	3167	4157	20,55	negative binomial	RLE	no
DESeq2	(Love et al., 2014)			83	282	1899	9,39	negative binomial	RLE	yes
vst or *QN + limma*	(Ritchie et al., 2015)					1276	6,31	Gaussian	vst or QN	yes
Cuffdiff 2	(Trapnell et al., 2013)			421	699	950	4,70	beta negative binomial	geometric	no
DEGSeq	(Wang et al., 2010)	178	297	458	636	850	4,20	Poisson	total count	no
voom + limma	(Law et al., 2014)					493	2,44	Gaussian	log-CPM	yes
NOISeq	(Tarazona et al., 2011)	65	164	263	377	473	2,34	nonparametric	CPM	no
baySeq	(Hardcastle and Kelly, 2010)	72	109	178	232	302	1,49	negative binomial	total count	yes
EBSeq	(Leng et al., 2013)	5	31	93	170	270	1,33	negative binomial	RLE	yes
Myrna	(Langmead et al., 2010)	57	88	112	117	149	0,74	Poisson or Gaussian	3rd quartile	no
SAMseq	(Li and Tibshirani, 2013)	0	22	52	91	129	0,64	nonparametric	trimmed total count	no
GFOLD	(Feng et al., 2012)			41	73	93	0,46	hierarchical Poisson	RLE	yes
PoissonSeq	(Li et al., 2012)	4	19	43	63	88	0,44	Poisson	trimmed total count	no
DSS	(Wu et al., 2013)			31	44	61	0,30	gamma-Poisson	3rd quartile	yes
BBSeq	(Zhou et al., 2011)	15	21	30	40	50	0,25	beta-binomial	total count	no
QuasiSeq	(Lund et al., 2012)					42	0,21	negative binomial	3rd quartile	no
TSPM	(Auer and Doerge, 2011)	8	12	16	26	39	0,19	two-stage Poisson	total count	no
ShrinkSeq	(Wiel et al., 2013)	5	14	18	28	33	0,16	zero-inflated negative binomial	none	yes
GENE-counter	(Cumbie et al., 2011)	9	14	20	26	30	0,15	negative binomial	total count	no
NBPSeq	(Di et al., 2011)	11	14	21	23	28	0,14	negative binomial	total count	no
sSeq	(Yu et al., 2013)					27	0,13	negative binomial	RLE	no
Polyfit	(Burden et al., 2014)					15	0,07	negative binomial	RLE	no
NPEBseq	(Bi and Davuluri, 2013)	0	4	11	12	14	0,07	gamma-Poisson	TMM	yes
BMDE	(Lee et al., 2011)	4	5	8	8	10	0,05	binomial (position-level)	total count	yes
LFCseq	(Lin et al., 2014)					6	0,03	nonparametric	trimmed total count	no
CEDER	(Wan and Sun, 2012)	0	1	2	4	5	0,02	negative binomial	RLE	no
ShrinkBayes	(van de Wiel et al., 2014)			1	3	5	0,02	zero-inflated negative binomial	none	yes

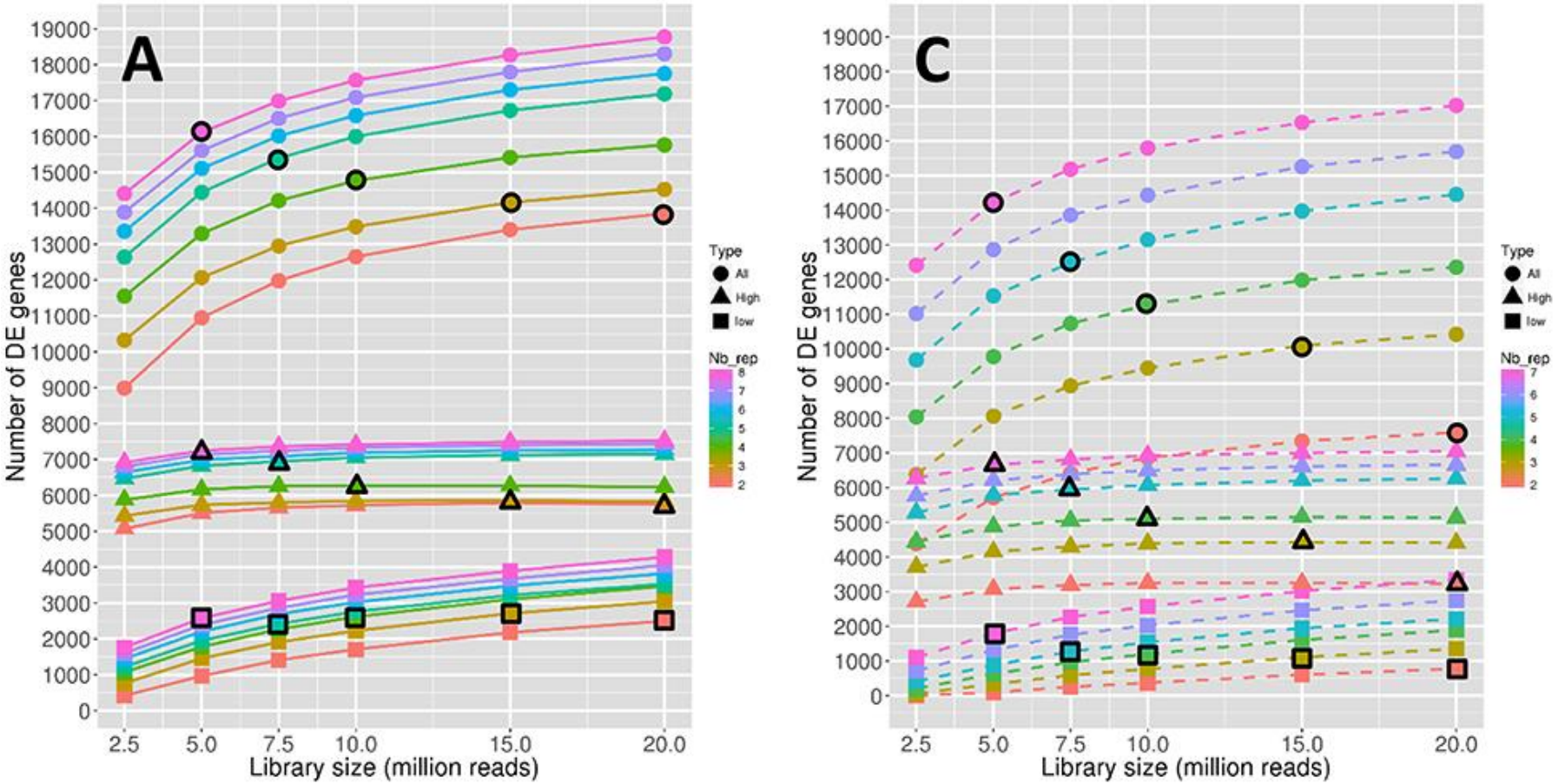
Replicates – Data

- TOGE experiment
  - 2 conditions
  - 8 replicates per condition
- TomExpress v16
  - 16 experiments
  - up to 18 conditions
  - up to 5 replicates per condition

- S. Lamarre, P. Frasse, M. Zouine, D. Labourdette, E. Sainderichin, G. Hu, V. Le Berre-Anton, M. Bouzayen, and E. Maza (2018) ***Optimization of an RNA-Seq Differential Gene Expression Analysis Depending on Biological Replicate Number and Library Size***, *Front. Plant Sci.* 9:108.

Replicates – Number of DE genes and power

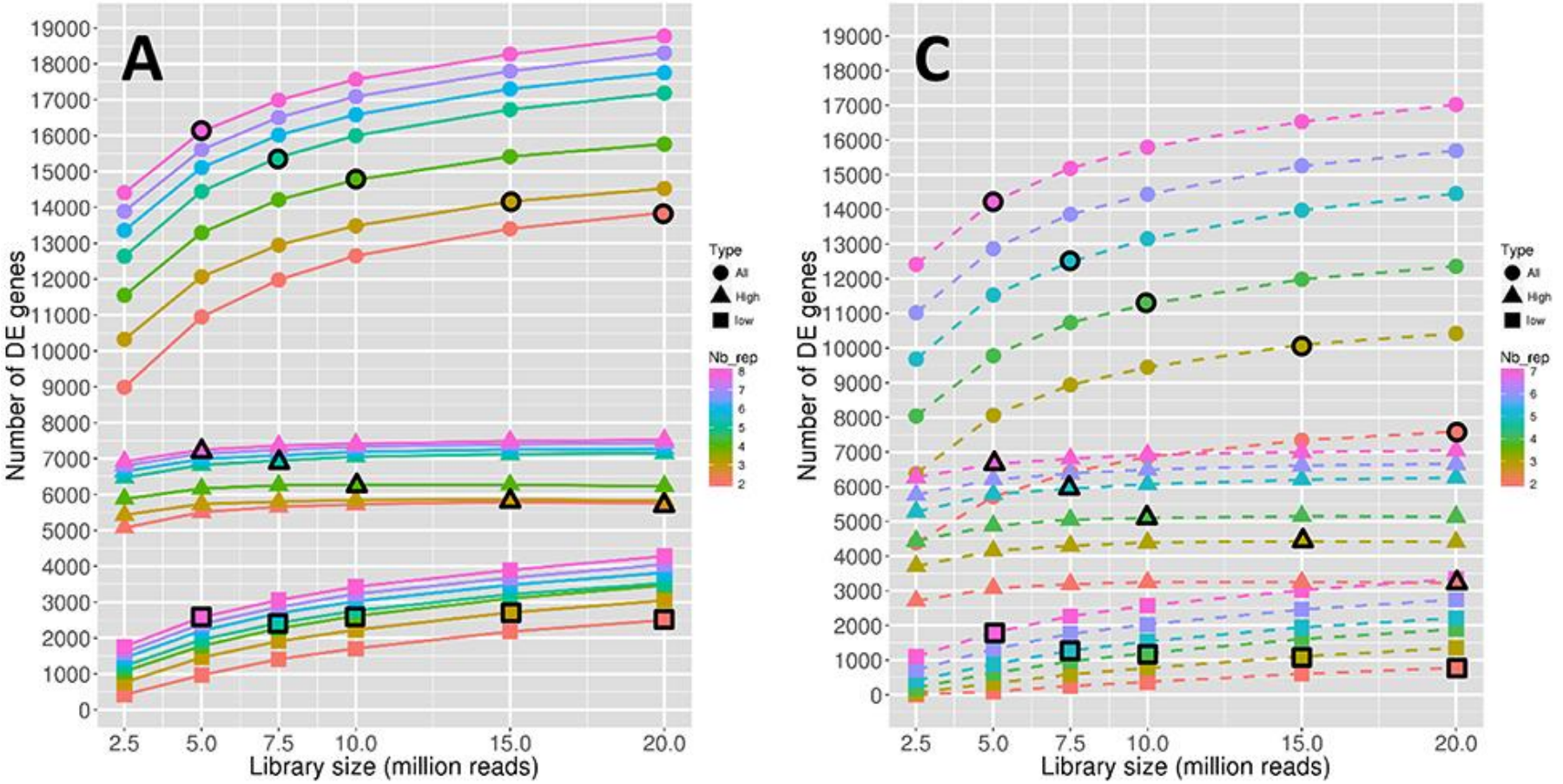
Lamarre et al. (2018)



Replicates – Number of DE genes and power

Replicate number >> library size

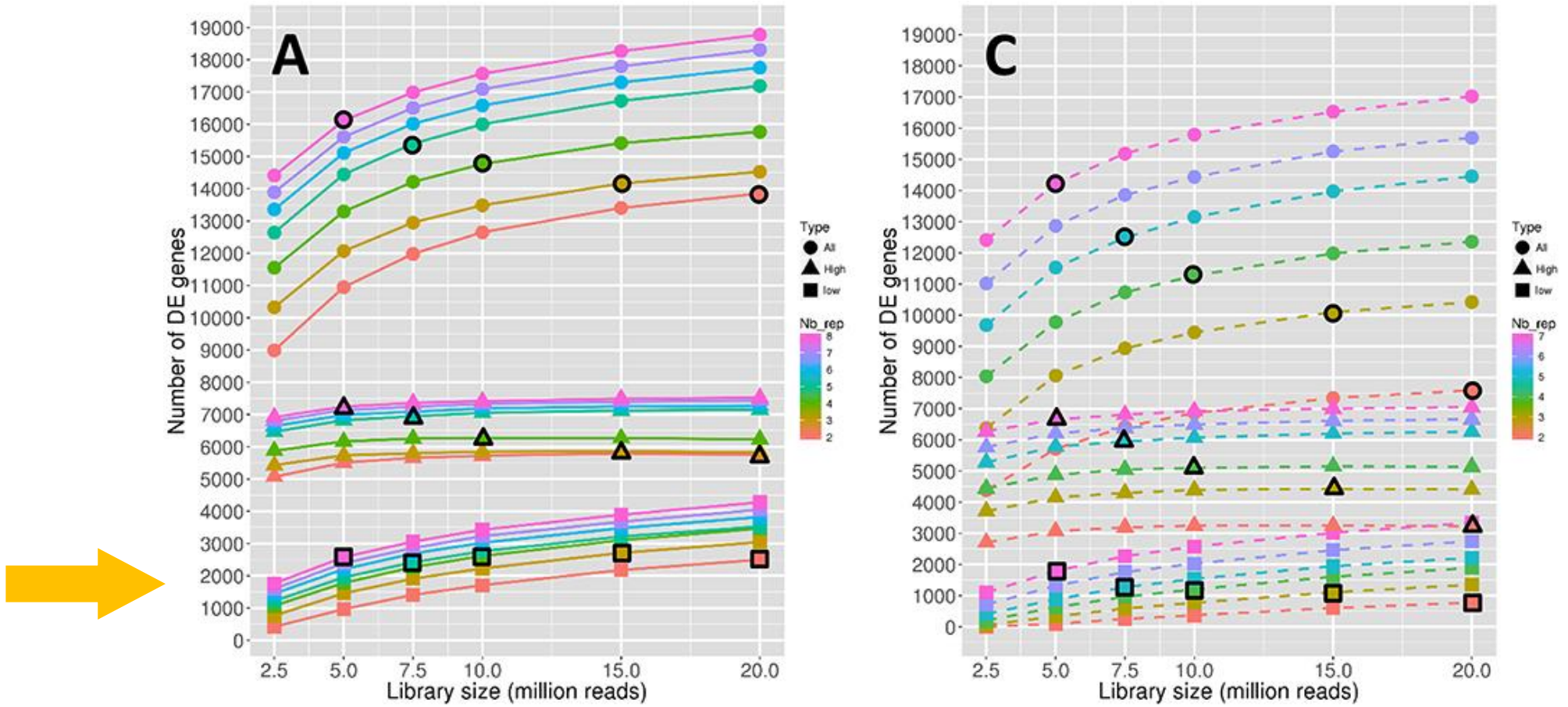
Lamarre et al. (2018)



Replicates – Number of DE genes and power

Replicate number >> library size

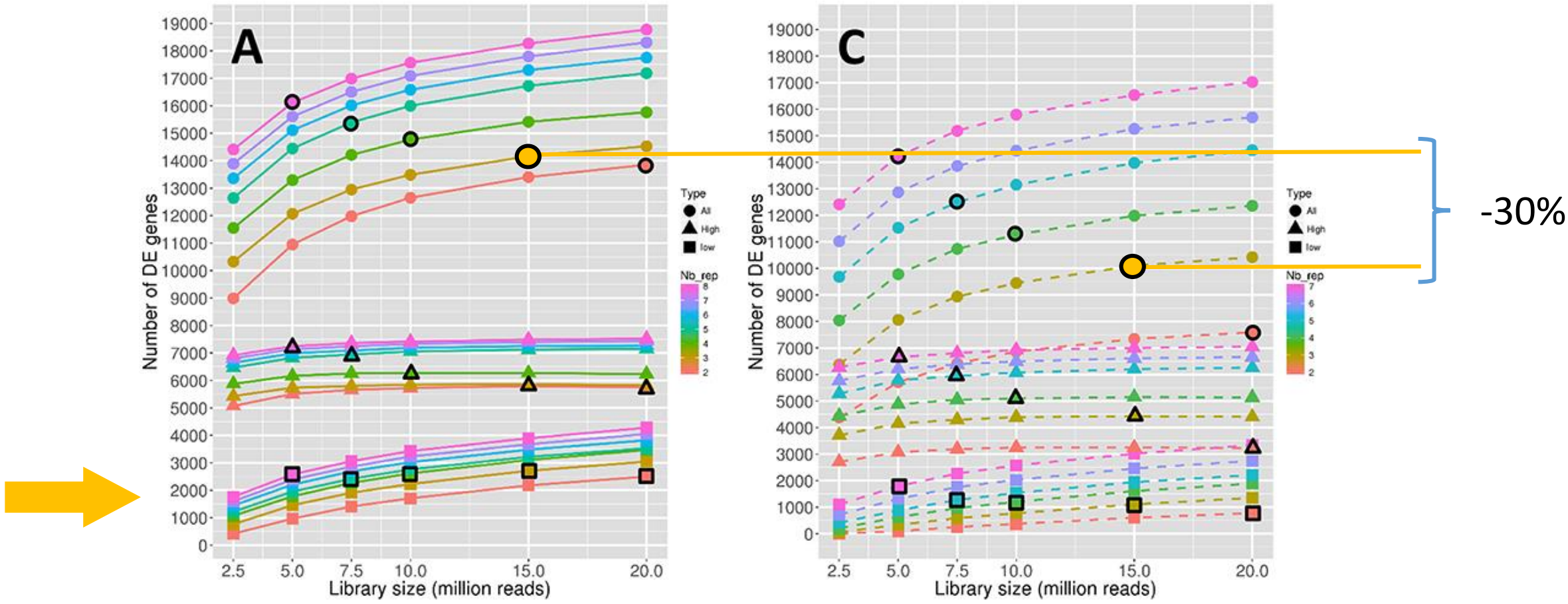
Lamarre et al. (2018)



Replicates – Number of DE genes and power

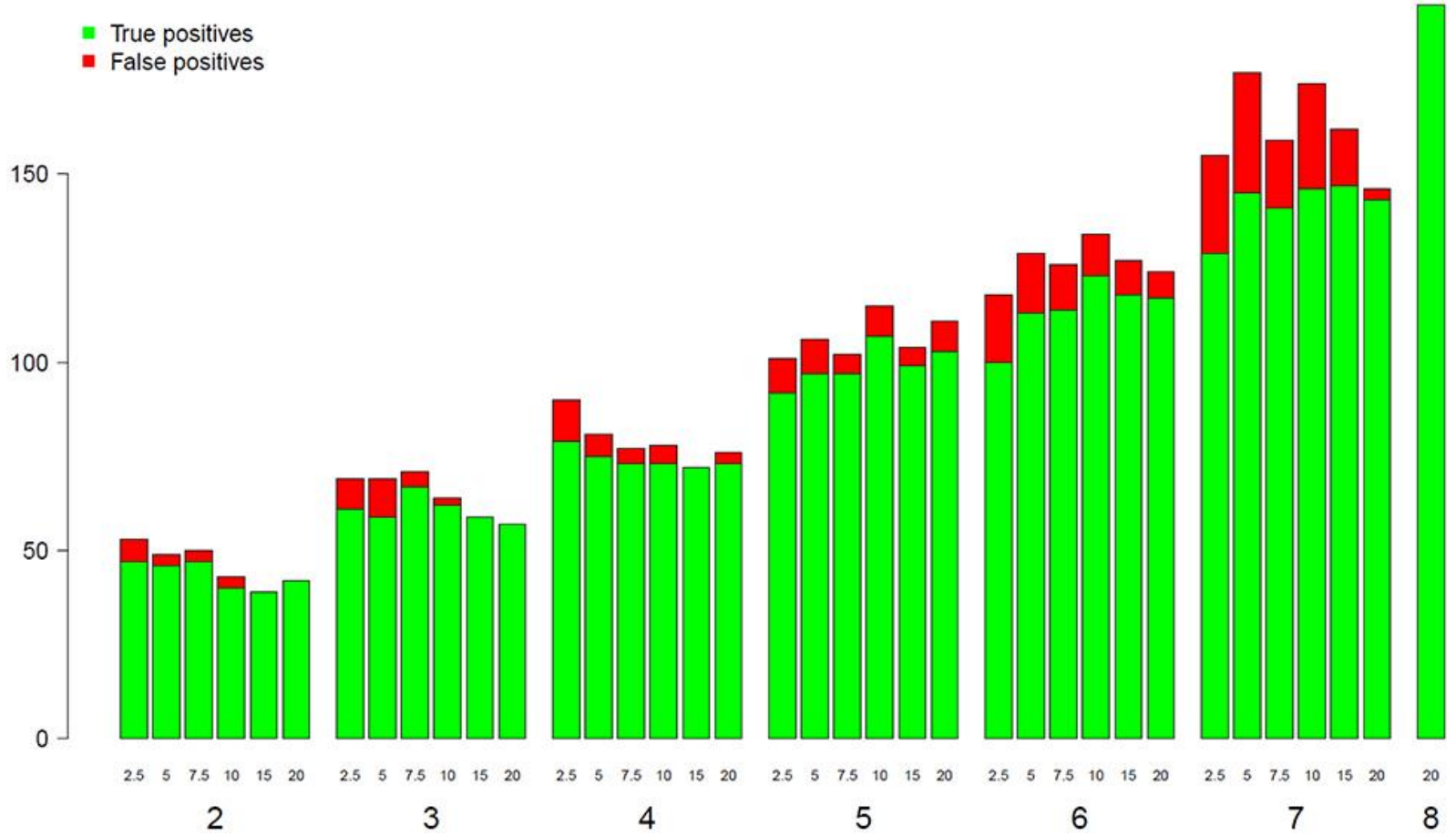
Replicate number >> library size

Lamarre et al. (2018)

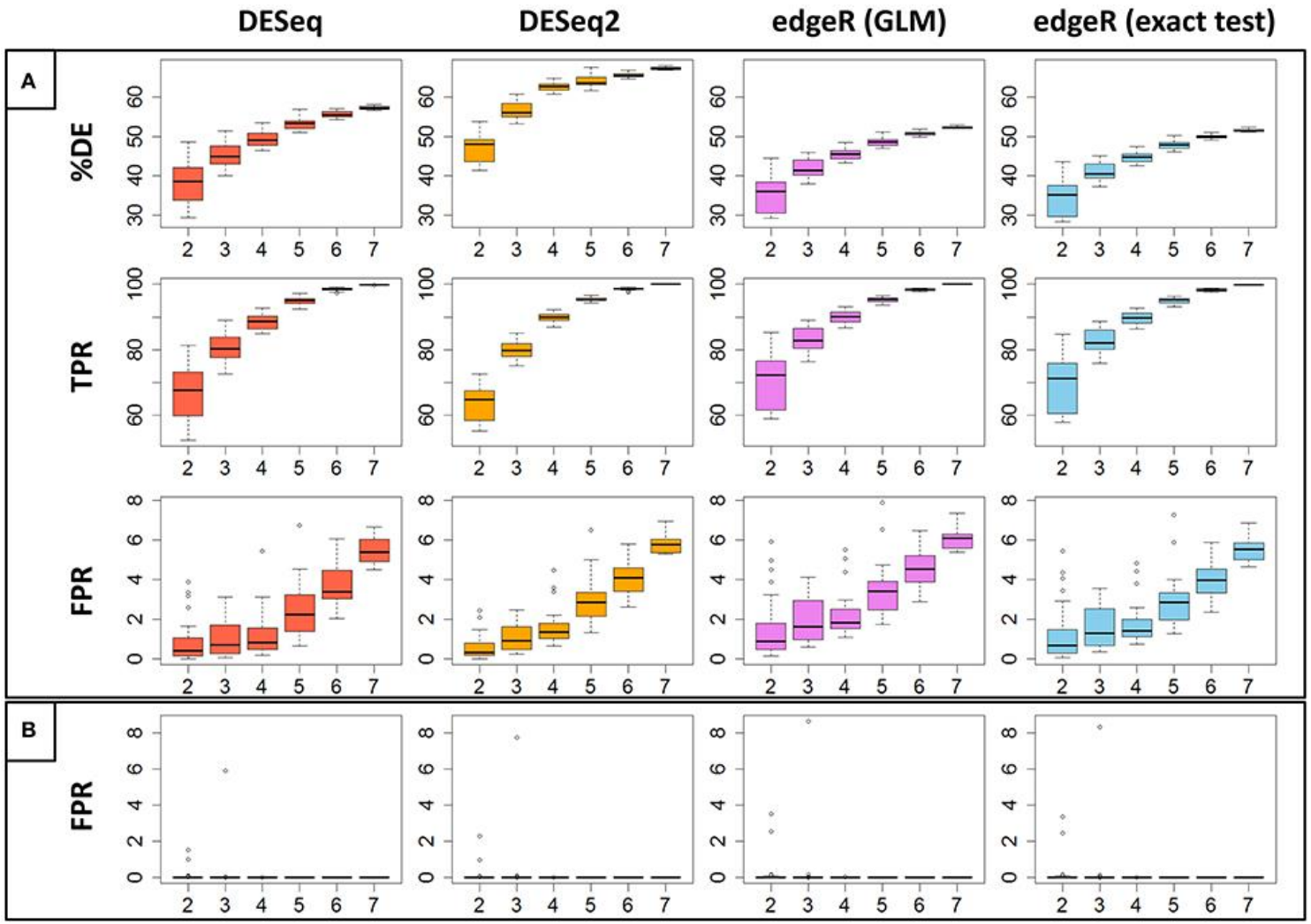


Replicates – GO enrichment analysis

Lamarre et al. (2018)



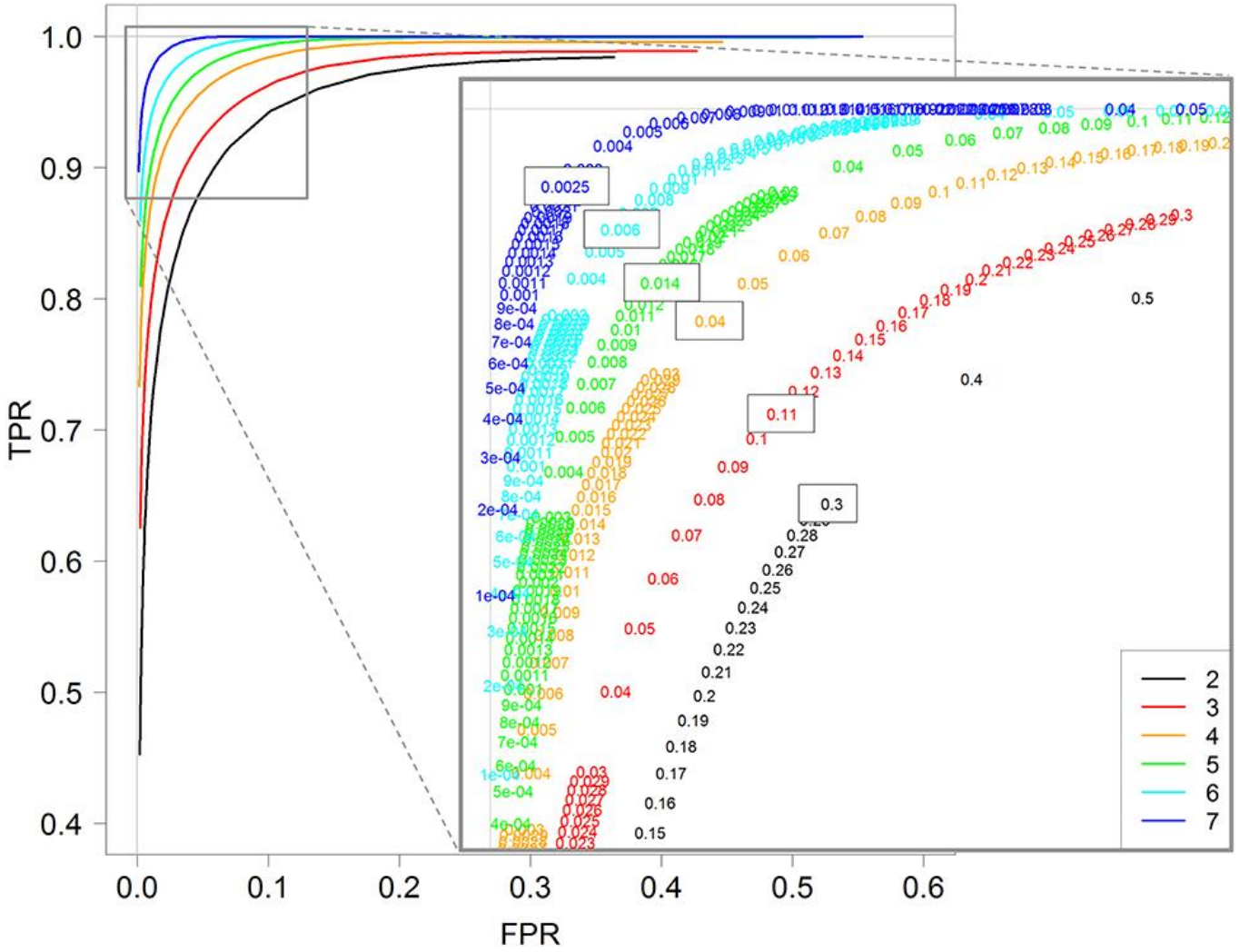
Replicates – Sensitivity and Specificity



Lamarre et al. (2018)

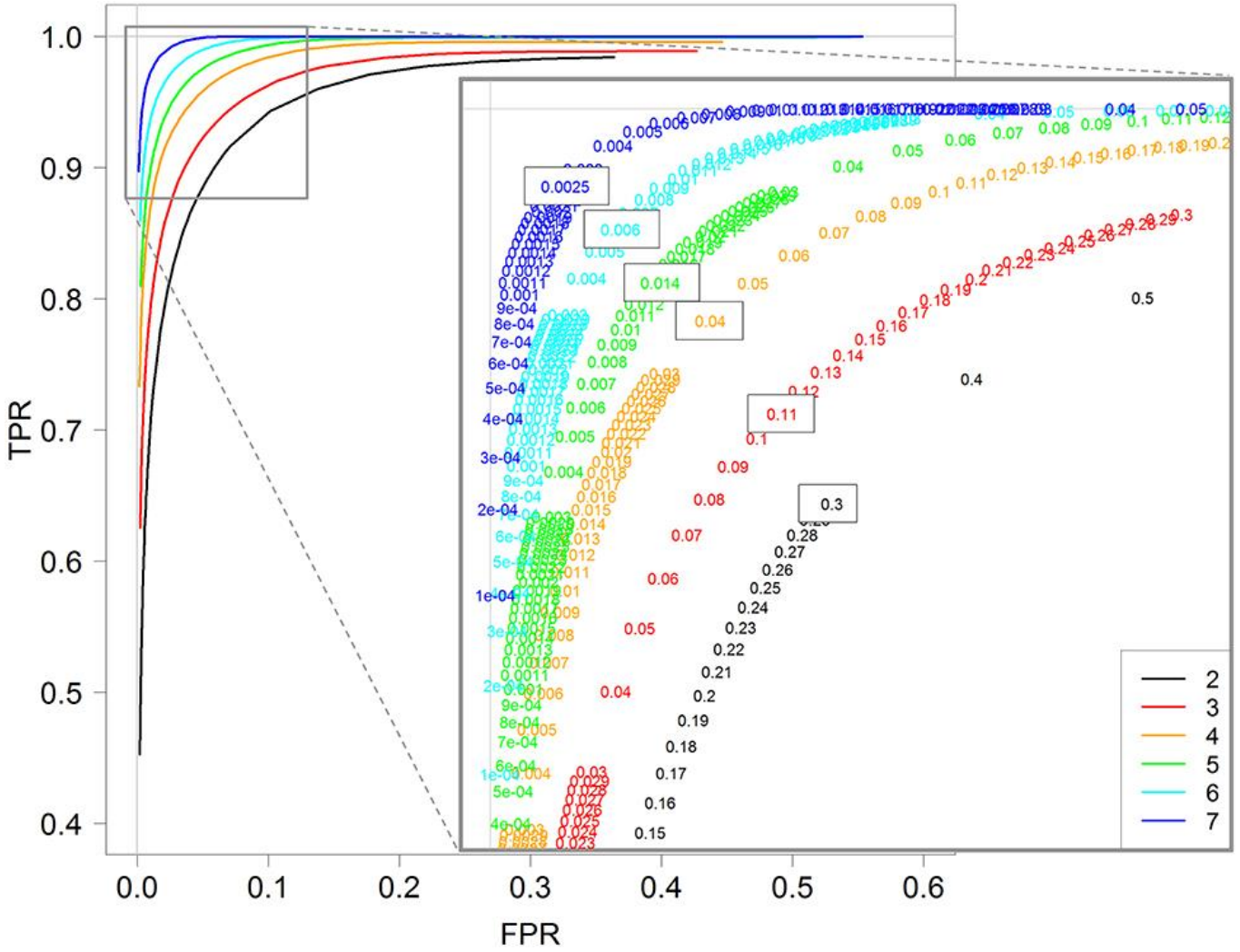
Replicates – Optimal threshold to control FDR

Lamarre et al. (2018)



Replicates – Optimal threshold to control FDR

Lamarre et al. (2018)



Opt. threshold $\approx 2^{-r}$

Replicate number (r)	Opt. threshold
2	0.25
3	0.12
4	0.06
5	0.03

Replicates – Meta-analysis (TomExpress)

Lamarre et al. (2018)

5565 condition pairs
× 20 replicate numbers
× 5 library sizes
× 3 repetitions
= 1,752,975 pairwise DE analyses

4 replicates & 20 M reads