Now you have the data

Elmer A. Fernández (PhD)
Guadalupe Nibeyro
CONICET
DataLab @ FPM - CONICET

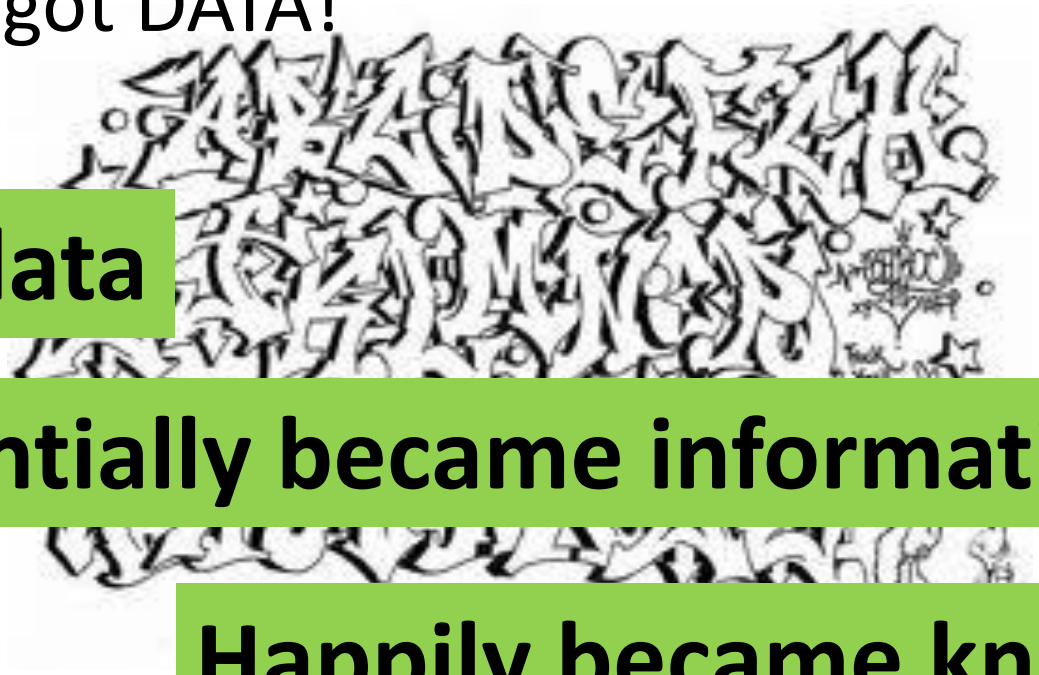
Your experiment is running

- You understand your problem, did the brainstorming meeting involving all the staff (technician, scientist, bioinformatics, statistician,…,aa, yes yes and YOU.)

- You already did all the wet lab
  - Under a SOP
- The sequencer ran and you have gotten the data.

Remember!

- You only got DATA!



You have data

Potentially became information

Happily became knowledge

Gene Annotation data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	PeakID	Chr	Start	End	Strand	Peak Sco	Focus Ra	Annotation	Detailed Anno	Distance to T	Nearest Pror	PromoterID	Nearest Unig	Nearest Refs	Nearest Ense	Gene Name	Gene Alias	Gene Descrip
2	chr18-1	chr18	69007968	69008268	+	593	0.939	intron (NR_03	intron (NR_03	74595	NR_034133	400655	Hs.579378	NR_034133		LOC400655	-	hypothetical
3	chr9-1	chr9	88209966	88210266	+	531.9	0.946	Intergenic	Intergenic	-50894	NM_001185(	79670	Hs.597057	NM_001185(	ENSG000000	ZCCHC6	DKFZp666B1	zinc finger, C
4	chr14-1	chr14	62337073	62337373	+	505.4	0.918	intron (NM_17	intron (NM_17	244485	NM_172375	27133	Hs.27043	NM_139318	ENSG000001	KCNH5	EAG2|H-EAG	potassium vo
5	chr17-1	chr17	5076243	5076543	+	492.1	0.936	intron (NR_03	intron (NR_03	2414	NM_207103	388325	Hs.462080	NM_207103	ENSG000001	C17orf87	FLJ32580|M	chromosome
6	chr17-2	chr17	47851714	47852014	+	476.2	0.824	Intergenic	Intergenic	-259488	NM_001082	56934	Hs.463466	NM_001082	ENSG000001	CA10	CA-RPX|CAR	carbonic anh
7	chr10-1	chr10	98420680	98420980	+	474.9	0.967	intron (NM_15	intron (NM_15	49439	NM_152309	118788	Hs.310456	NM_152309	ENSG000001	PIK3AP1	BCAP|RP11-3	phosphoinos
8	chr9-2	chr9	81294389	81294689	+	456.3	0.957	Intergenic	Intergenic	-82159	NM_007005	7091	Hs.444213	NM_007005	ENSG000001	TLE4	BCE-1|BCE1|	transducin-li
9	chr14-2	chr14	36817736	36818036	+	452.3	0.757	intron (NM_13	intron (NM_13	81017	NM_001195.	145282	Hs.660396	NM_001195.	ENSG000001	MIPOL1	DKFZp313M2	mirror-image
10	chr18-2	chr18	20049825	20050125	+	449.7	0.853	intron (NM_08	intron (NM_08	56219	NM_018030	114876	Hs.370725	NM_018030	ENSG000001	OSBPL1A	FLJ10217|OF	oxysterol bin
11	chr7-1	chr7	12226829	12227129	+	445.7	0.901	intron (NM_01	intron (NM_01	9606	NM_001134	54664	Hs.396358	NM_001134	ENSG000001	TMEM106B	FLJ11273|M	transmembra
12	chr14-3	chr14	88712188	88712488	+	443.1	0.844	intron (NM_00	intron (NM_00	240869	NM_005197	1112	Hs.621371	NM_001085	ENSG000000	FOXN3	C14orf116|C	forkhead box
13	chr18-3	chr18	62951924	62952224	+	443.1	0.947	Intergenic	Intergenic	-382689	NR_033921	643542	Hs.652901	NR_033921		LOC643542	-	hypothetical
14	chr3-1	chr3	32196769	32197069	+	443.1	0.87	Intergenic	Intergenic	-58256	NM_178868	152189	Hs.154986	NM_178868	ENSG000001	CMTM8	CKLFSF8|CKL	CKLF-like MA
15	chr11-1	chr11	110685448	110685748	+	425.8	0.907	Intergenic	Intergenic	-9849	NR_034154	399948	Hs.729225	NR_034154		C11orf92	DKFZp781P1	chromosome
16	chr4-1	chr4	81755366	81755666	+	423.2	0.908	intron (NM_15	intron (NM_15	279618	NM_152770	255119	Hs.527104	NM_152770	ENSG000001	C4orf22	MGC35043	chromosome

Raw data

Nombre ↑

- c52af53d-part-1468006400-1572863999
- SQ24017291-R24005611LR01-Br1384_combined_R1.trimmed.fastq.gz
- SQ24017291-R24005611LR01-Br1384_combined_R1.trimmed.fastq.gz.md5
- SQ24017291-R24005611LR01-Br1384_combined_R2.trimmed.fastq.gz

Quantified data

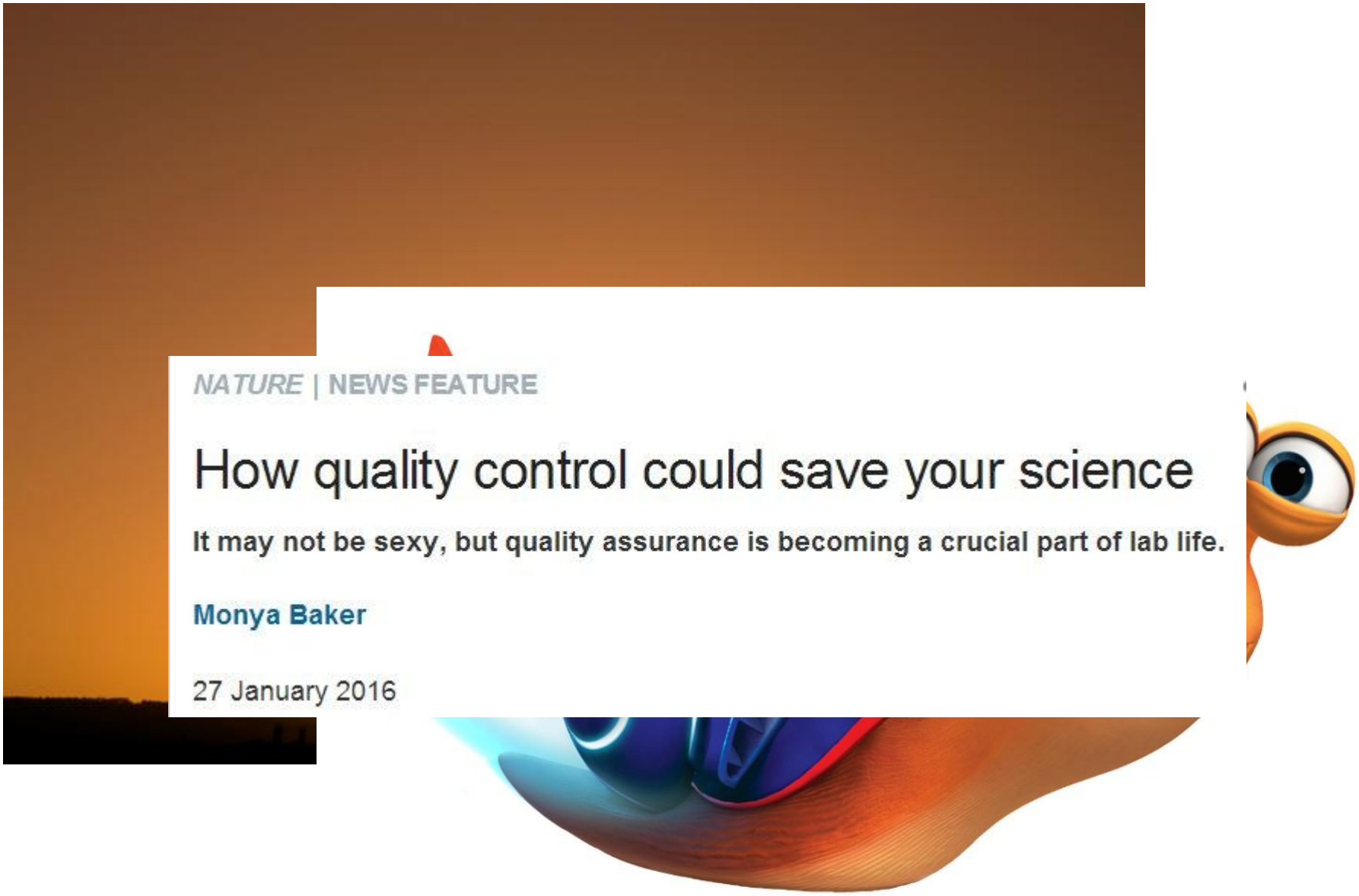
	C	D	E	F	G
		HCW5_Mg	HCW6_Mg	HCW1_not_induced	HCW2_not_induced
	KP00002	218	180	168	130
	KP00003	82	91	67	82
	KP00004	151	177	251	173
	KP00005	985	992	926	841
	KP00006	14	9	8	19
	KP00007	63	49	31	26
	KP00008	52	53	37	31
	KP00010	555	494	521	388
	KP00011	332	293	318	243
	KP00012	82	63	55	46
	KP00014	263	214	126	94
	KP00015	233	193	195	148
	KP00016	509	449	449	379

Meta data

L10

	A	B	C	D	E	F	G	H	I	J	K
1	**Idcode**	**Var1**	**Var2**	**Var3**	**Date**	**Subtipo**	**MTT**	**Secondary Tumor**	**Age**	**Gender**	**Ancestry**
2	XX133	86.40	3.00	33.00	4/19/2018	Luminal HER2	MET-	YES	34	F	AM
3	XX231	76.30	2.40	50.00	6/21/2018	Luminal B	MET-	NO	45	F	AF
4	XX378	61.60	2.40	53.00	8/8/2018	Triple Negativo	MET-	NO	66	M	HI
5	XX329	69.90	2.20	37.00	7/15/2016	Luminal HER2	MET-	YES	47	F	HI
6											

So, when you face a data analytics problem



NATURE | NEWS FEATURE

How quality control could save your science

It may not be sexy, but quality assurance is becoming a crucial part of lab life.

Monya Baker

27 January 2016

What does mean QA & DS Analysis

Some examples of QA

We performed a target sequencing experiment of three samples

Target sequencing experiment is a genomics technique that focuses on sequencing specific regions of the genome, rather than the entire genome. The **purpose** of this method is to selectively analyze certain genes or regions of interest, typically chosen because of their relevance to a particular disease, biological process, or research question.

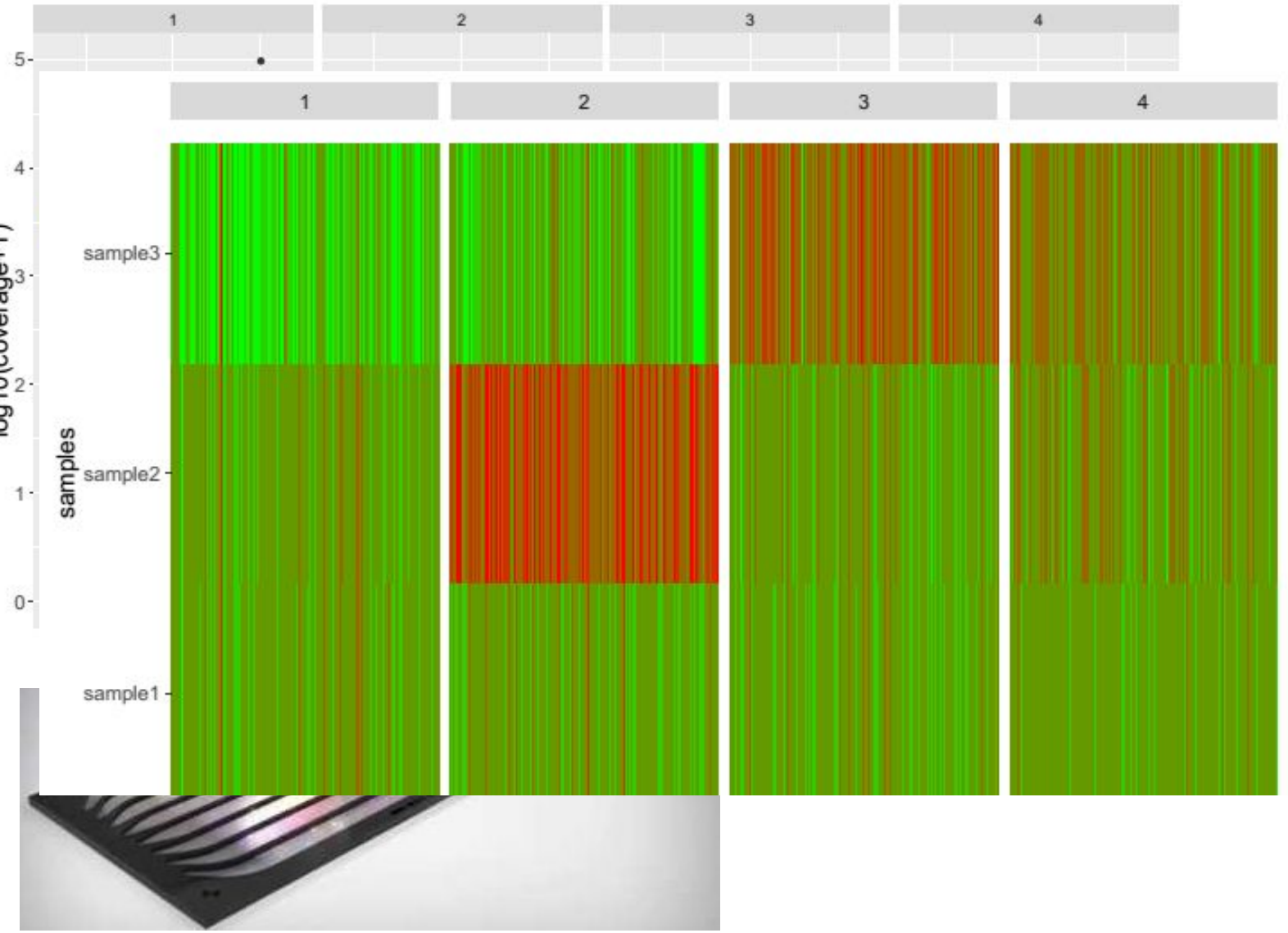
So, what should be thinking about?

Some examples of QA

that's make sense?

What should I be expecting?

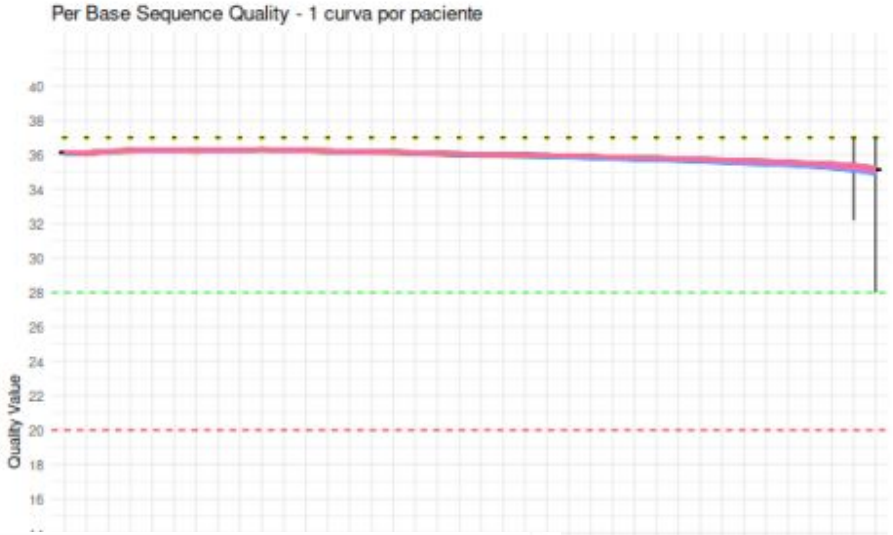
Some examples of QA



Some examples of QA on sequencing Data

RNAseq Data

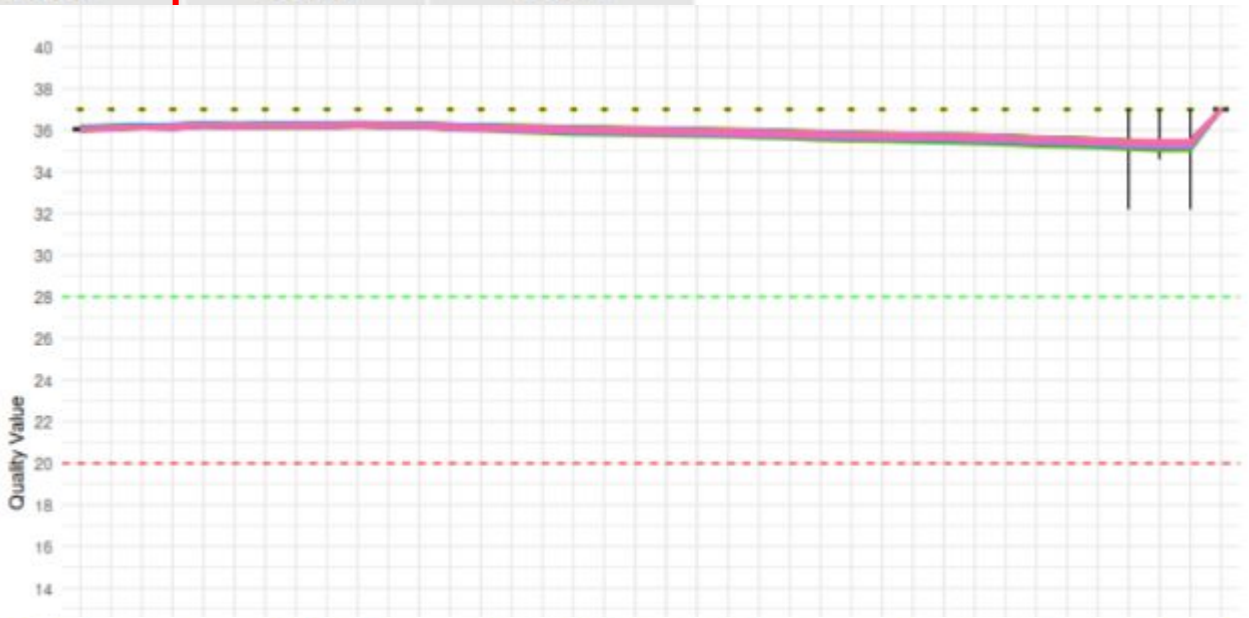
Estadístico	Q media R1	Q30 R1 (%)	Q media R2	Q30 R2 (%)
Min	35.87	97.53	35.69	96.68
Q1	35.92	97.72	35.80	97.36
Mean	35.95	97.87	35.83	97.48
Q3	35.99	98.06	35.86	97.61
Max	36.04	98.27	35.93	98.01
SD	0.05	0.21	0.06	0.26



Estadístico	Lecturas input (N°)	Largo medio input	Lecturas mapeadas (N°)	Lecturas mapeadas (%)	Largo medio mapeadas
Min	10120782	294.00	9564608	92.09	294.12
Q1	11732190	295.00	11022414	94.15	295.31
Mean	14602004	295.77	13808721	94.55	295.75
Q3	16497719	296.00	15671648	95.12	296.36
Max	16596816	297.00	15884138	96.18	297.13
SD	2611903	0.86	2488324	0.94	0.73

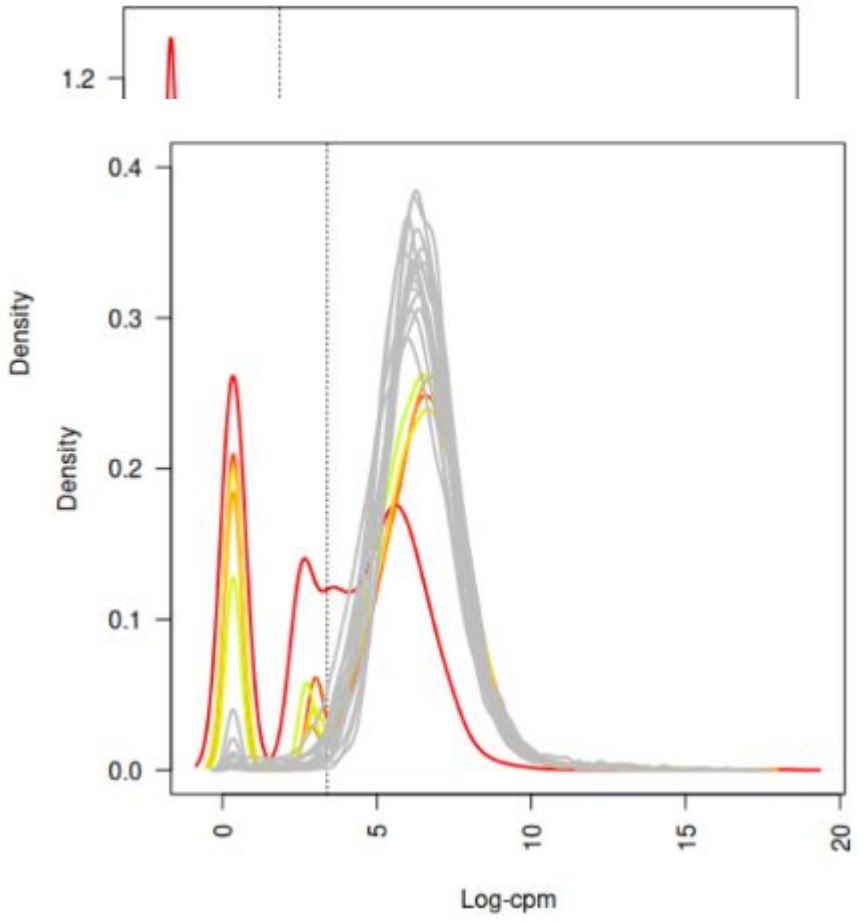
Some examples of QA on sequencing Data

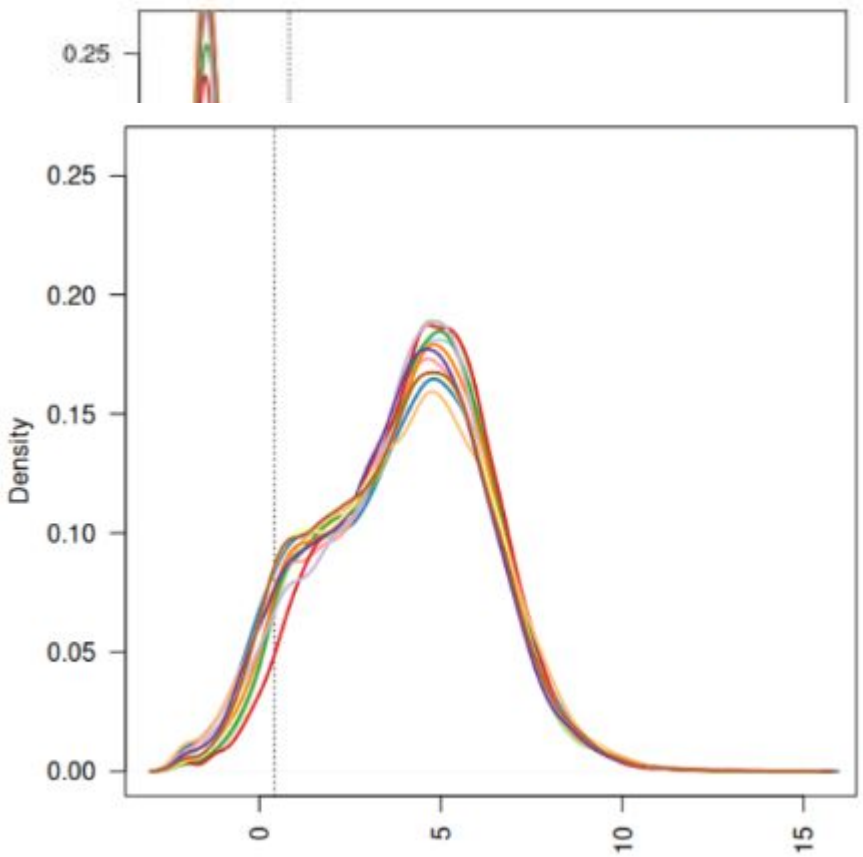
	Statistic	Q_Mean_R1	%_>=Q30_R1	Q_Mean_R2	%_>=Q30_R2
1	Min	35.25	94.82	35.21	94.84
2	Q1	35.44	96.02	35.39	95.43
3	Mean	35.60			
4	Q3	35.75			
5	Max	35.82			
6	SD	0.17			



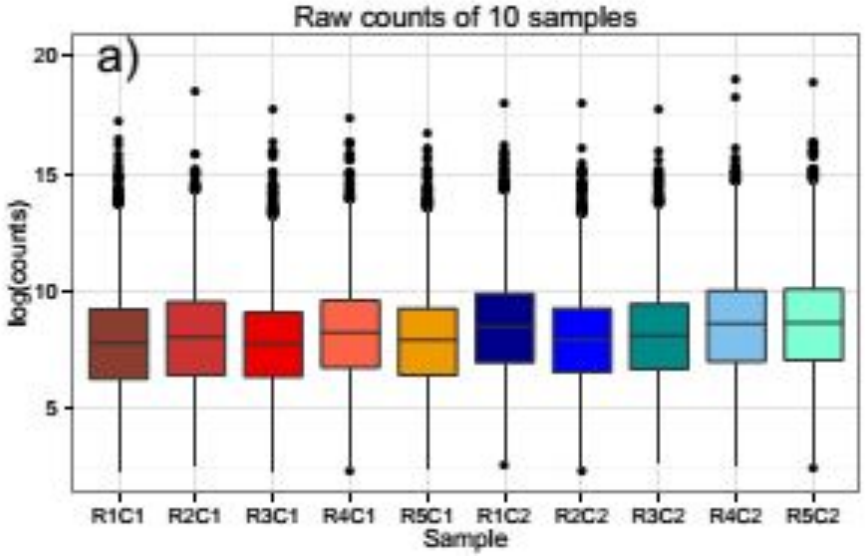
	Statistic	Number.of.input.reads	Average.input.read.length	Uniquely.mapped.reads.number	Uniquely.mapped.reads.%	Average.mapped.length
1	Min	374861	250.00	218820	58.37	253.20
2	Q1	1323983	259.00	1011945	82.13	260.76
3	Mean	5654087	264.68	4938357	83.28	266.30
4	Q3	6927377	271.00	6024305	87.45	273.22
5	Max	25289078	280.00	22759790	90.00	280.08
6	SD	6372973	9.02	5741510	7.83	8.50

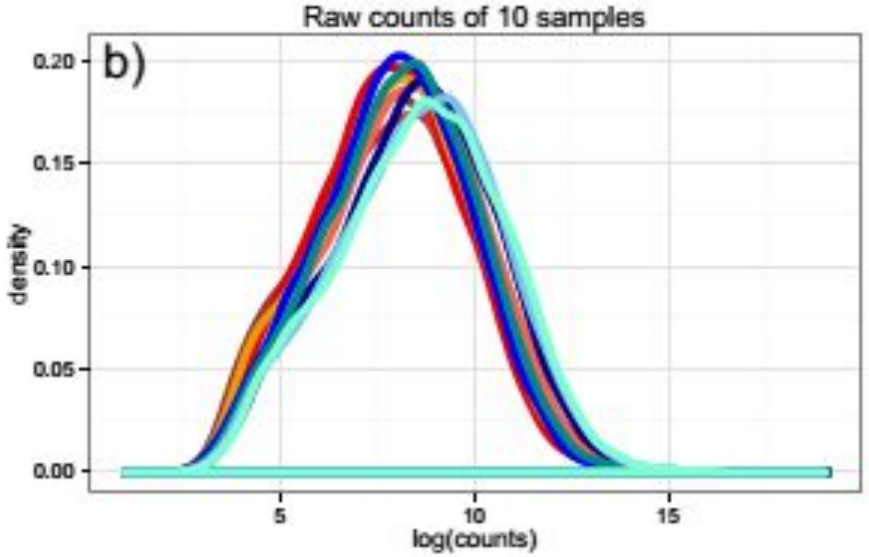
We already have the data, let's explore them

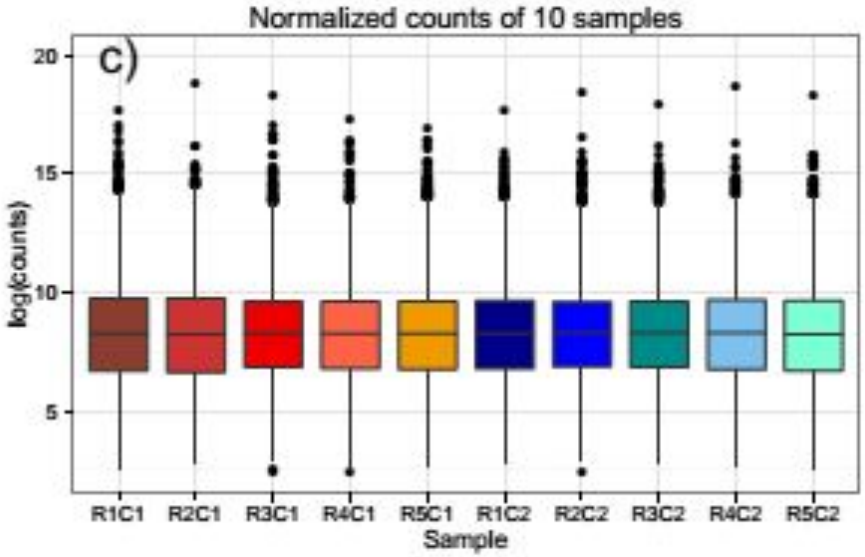


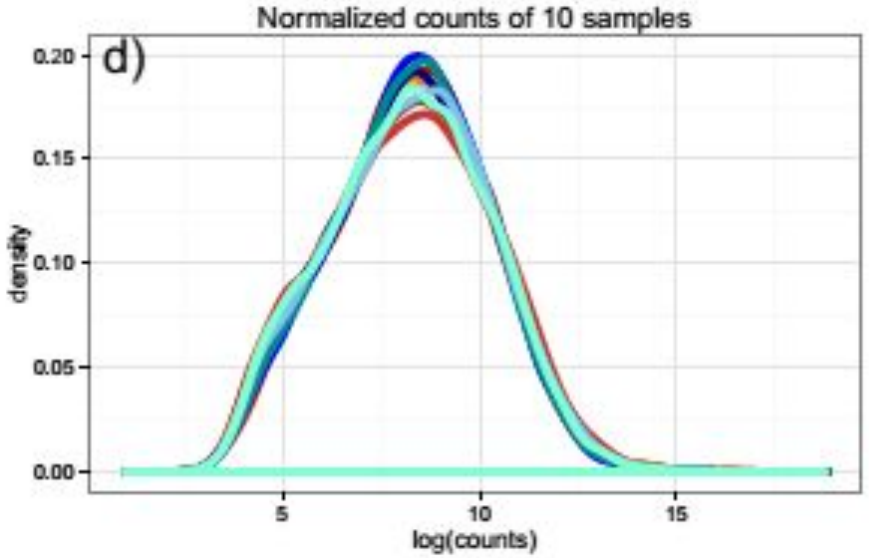


Another example









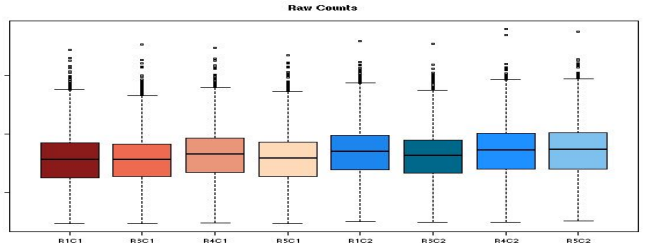
Normalization

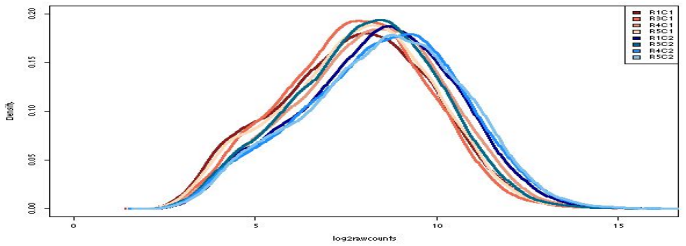
- There are many tipes of normalization, depending on the problem at hand (tecnical problems, distributional,problems, assumptions…).
- It is not necessary a problem
- It is ussually done in statistics
- Specially in multivariate analysis
- Let the data tells you what to do, according to your knowldege.
- It is cruccial to understand the problem to un order derstand the possible variation sources
- It should be applied with care.

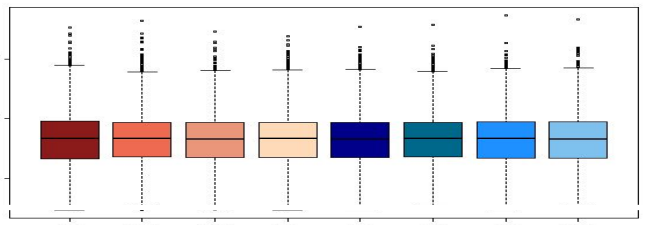
Normalization

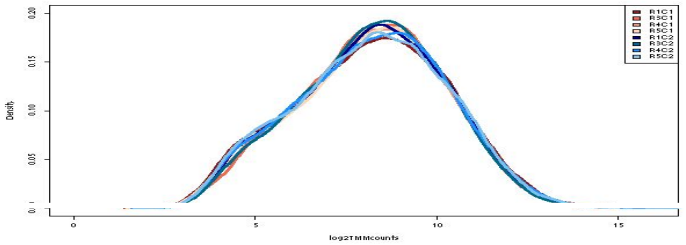
- In genomics/proteomics, the aim of normalization is to remove systematic technical effects tha occur in the data to ensure that technical bias has minimal impact on results.
- According to Bulard et al: the greatest impact on DE detection is the choice of the normalization procedure.

- Wha



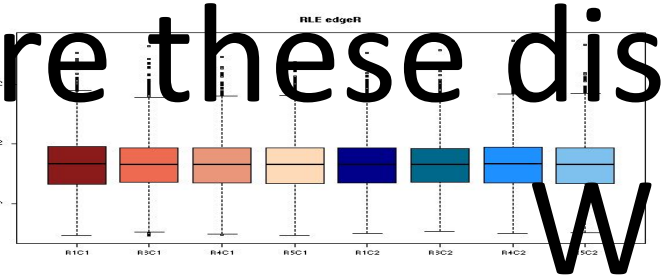


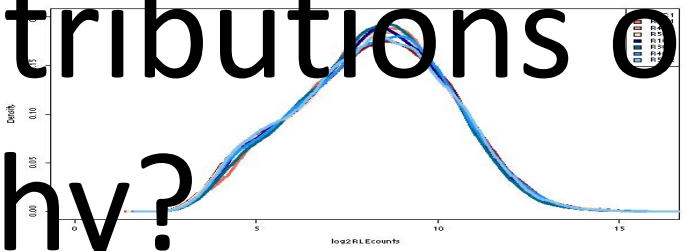


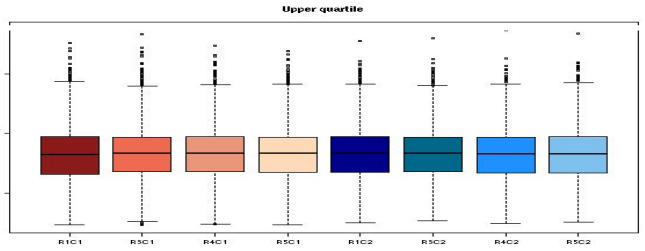


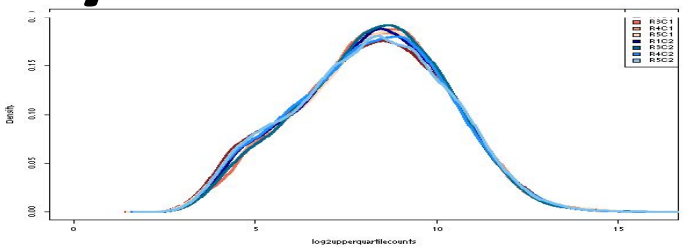
Are these distributions ok?

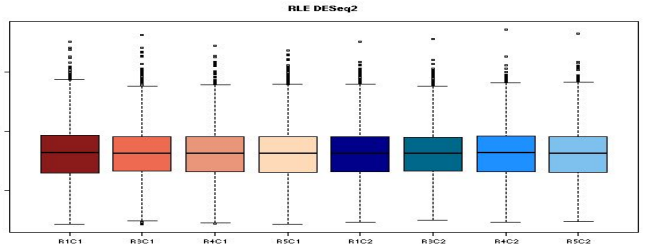
Why?

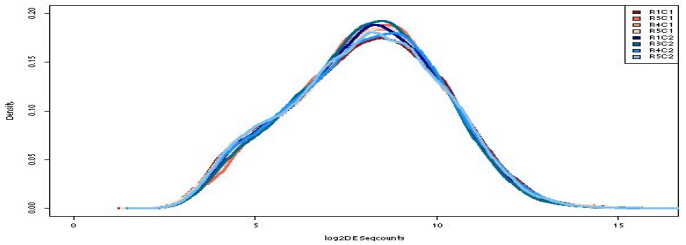




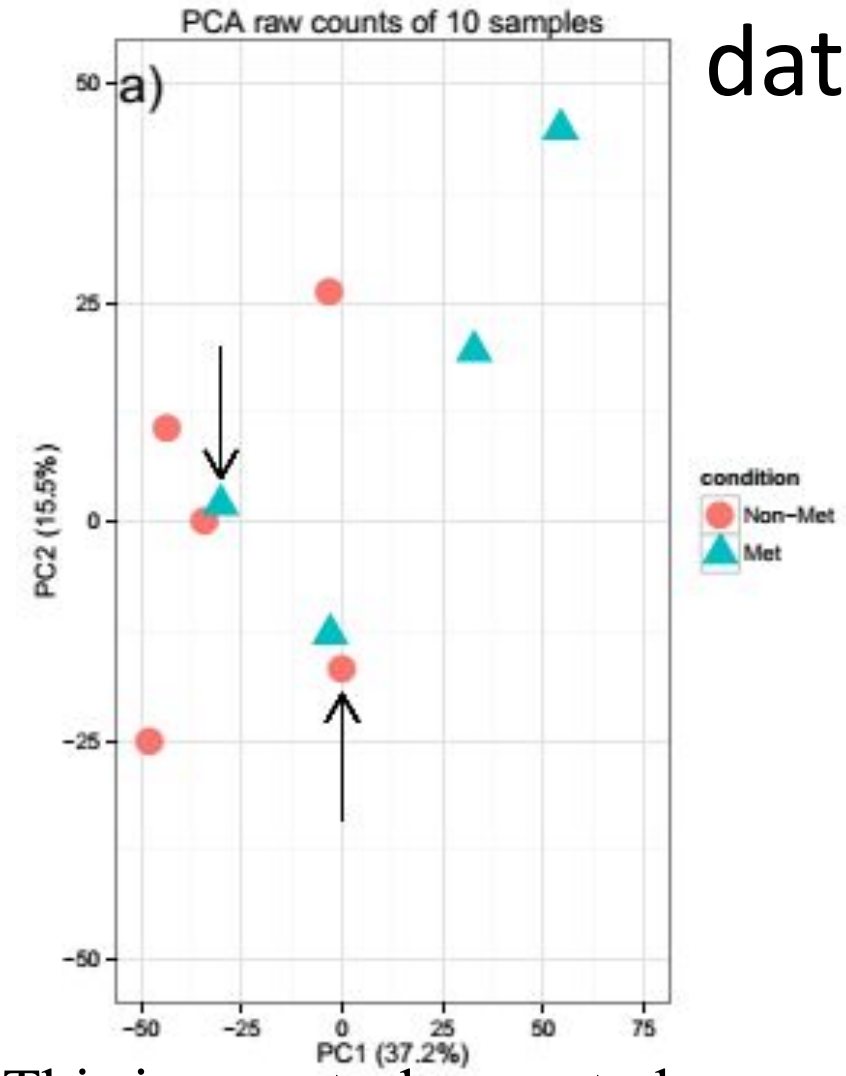


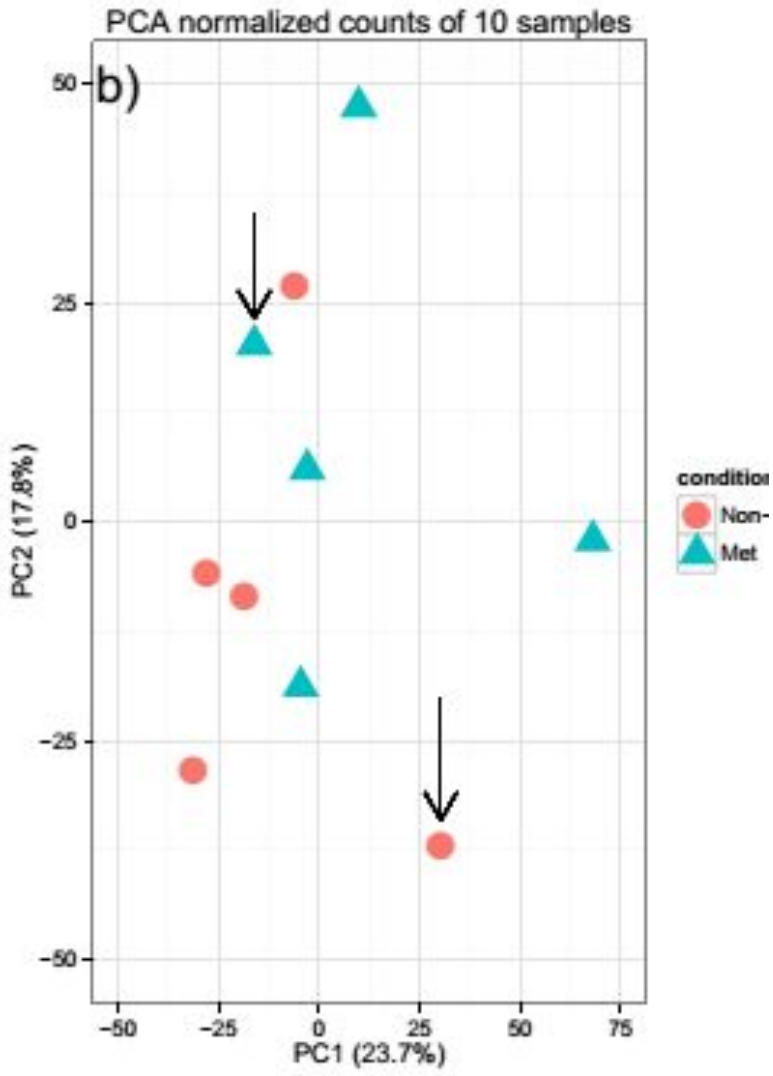






Take your time, and get used to the data

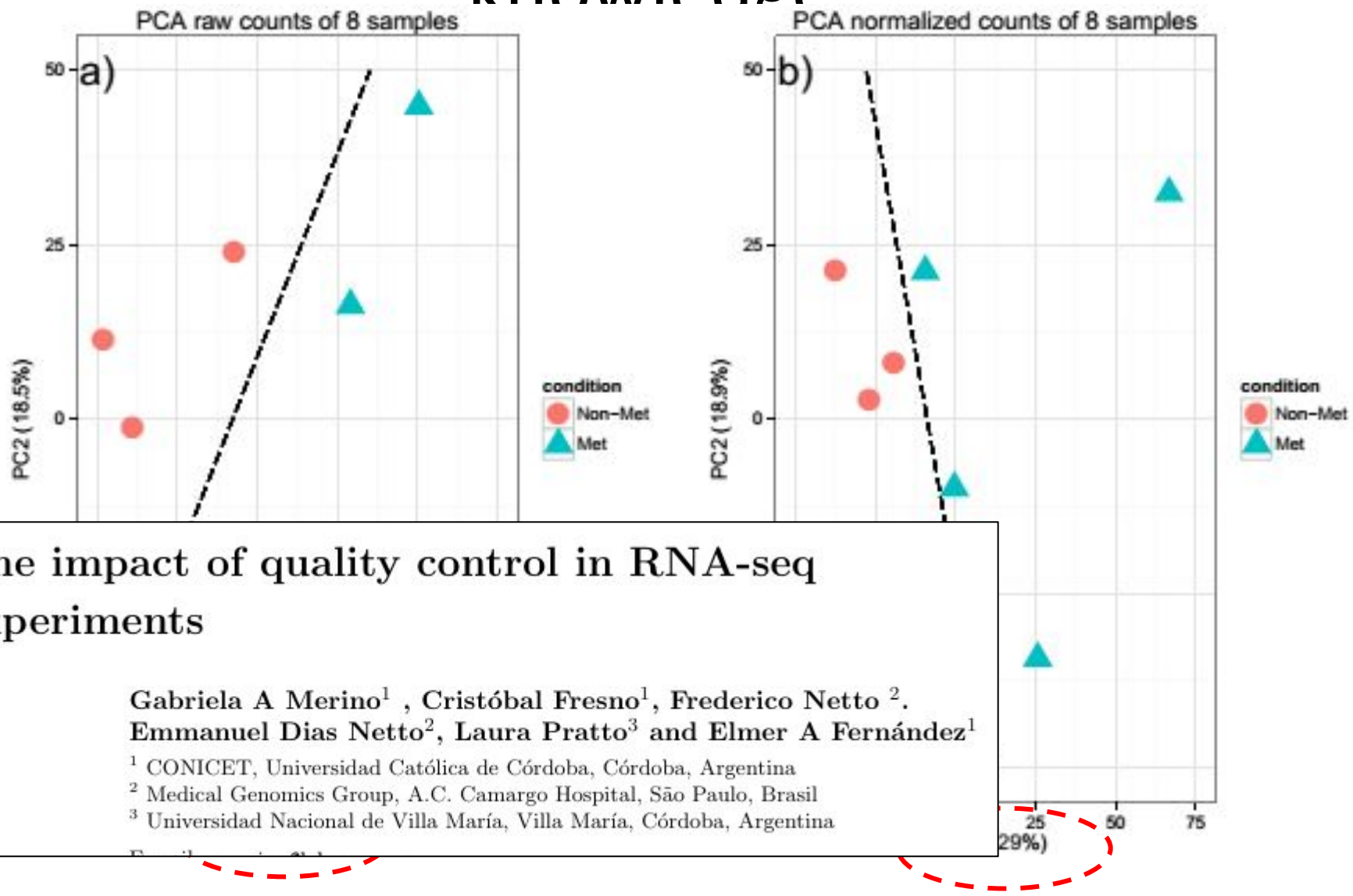




This is a control-case study

What do you expect from these data?

Analyze and use your assumptions and knowledge

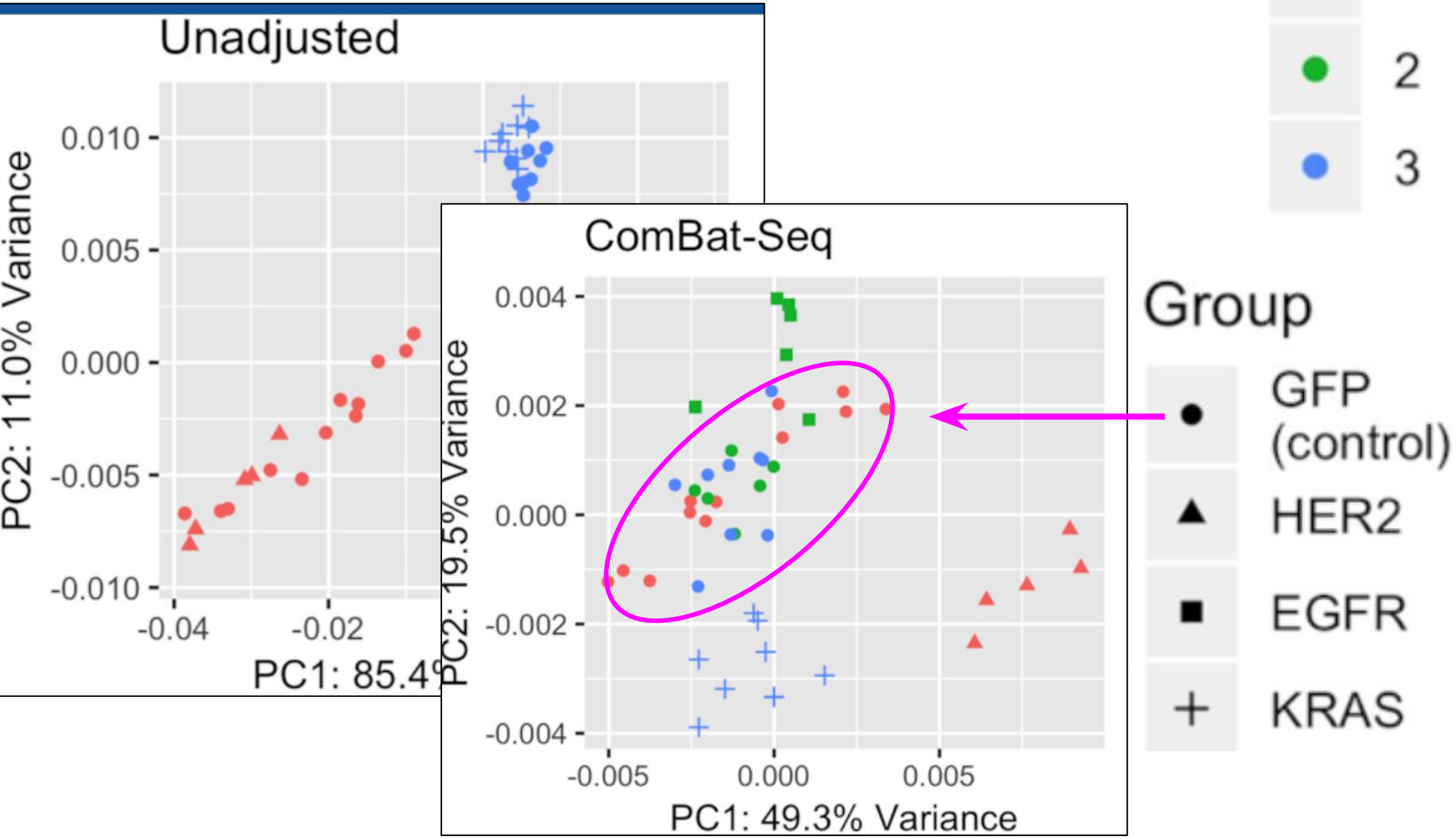


The impact of quality control in RNA-seq experiments

Gabriela A Merino[1] , Cristóbal Fresno[1], Frederico Netto [2]. Emmanuel Dias Netto[2], Laura Pratto[3] and Elmer A Fernández[1]

[1] CONICET, Universidad Católica de Córdoba, Córdoba, Argentina
[2] Medical Genomics Group, A.C. Camargo Hospital, São Paulo, Brasil
[3] Universidad Nacional de Villa María, Villa María, Córdoba, Argentina

Batch effects



Once you are satisfied

- Do DGE
  - edgeR, DESeq, what ever you like.
  - I do not suggest to use cufflink

Method

Highly accessed

Open Acc

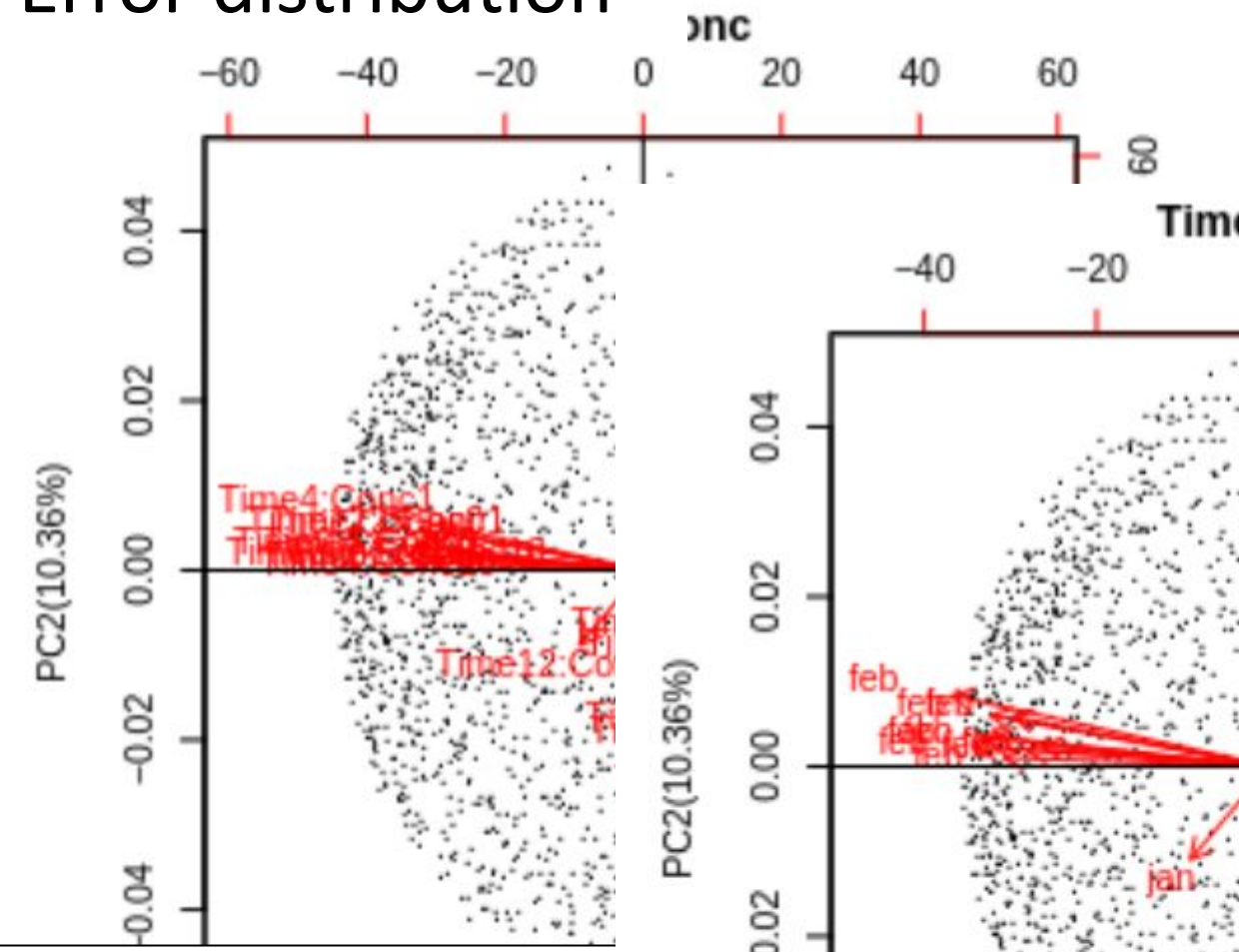
voom: precision weights unlock linear model analysis tools for RNA-s read counts

Charity W Law[1,2], Yunshun Chen[1,2], Wei Shi[1,3] and Gordon K Smyth[1,4]*

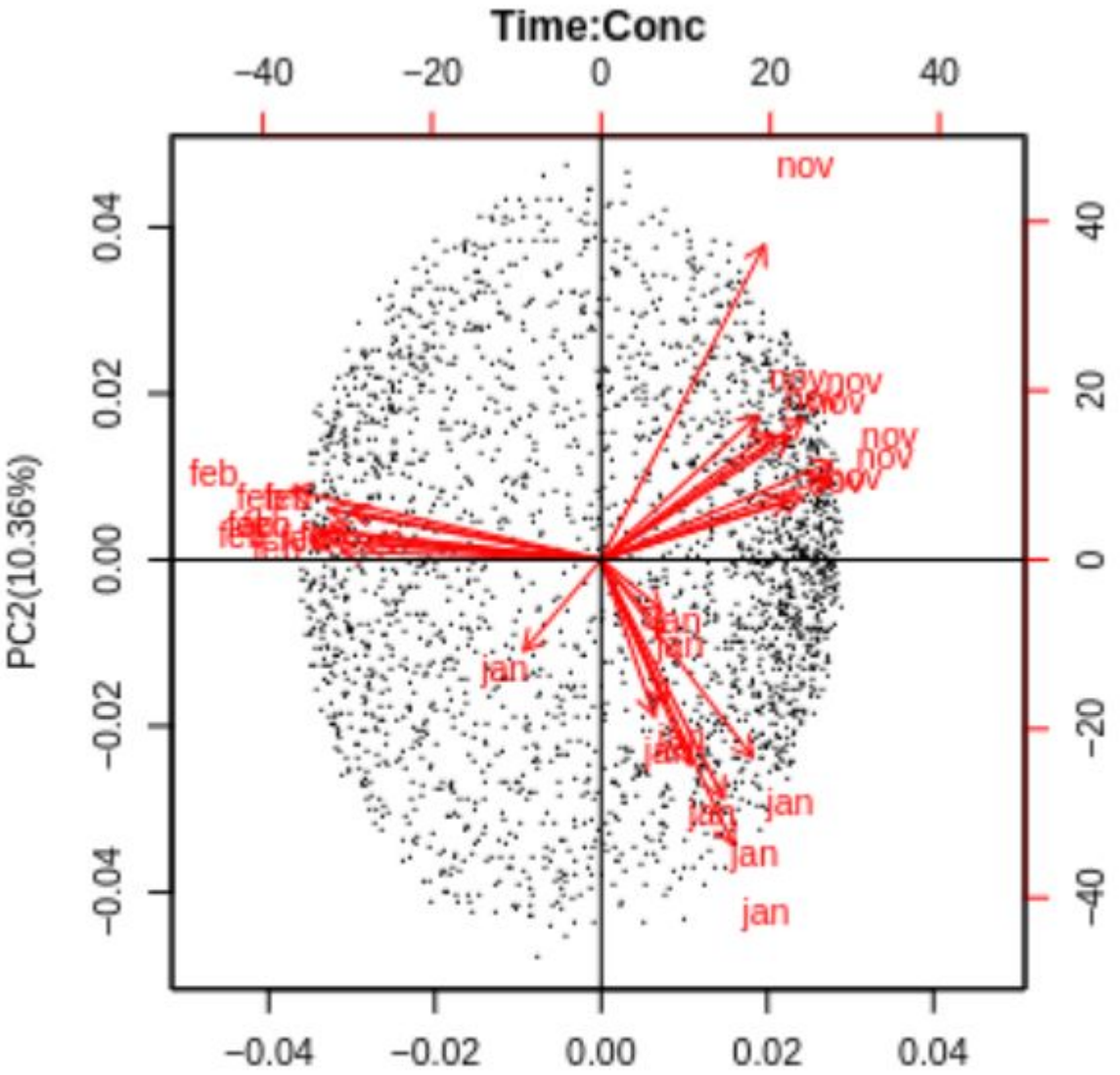
take into account that: edgeR, DESeq works over count data, but you can move towards normal based models through the voom method.


Results

Error distribution



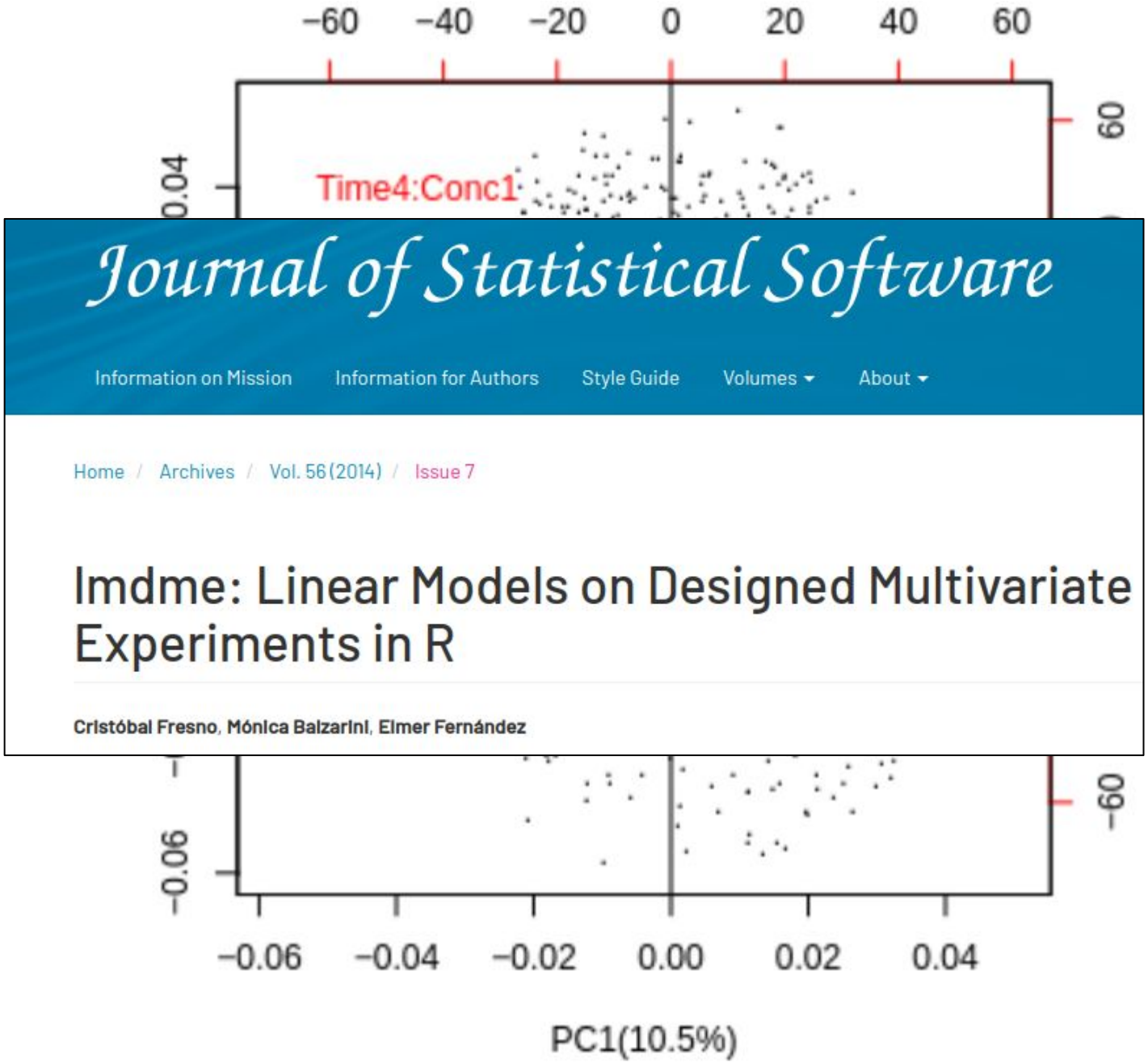
This is a



= **μ**+Trat+Conc+Tra

here

≈ ***N***(0,***σ***)



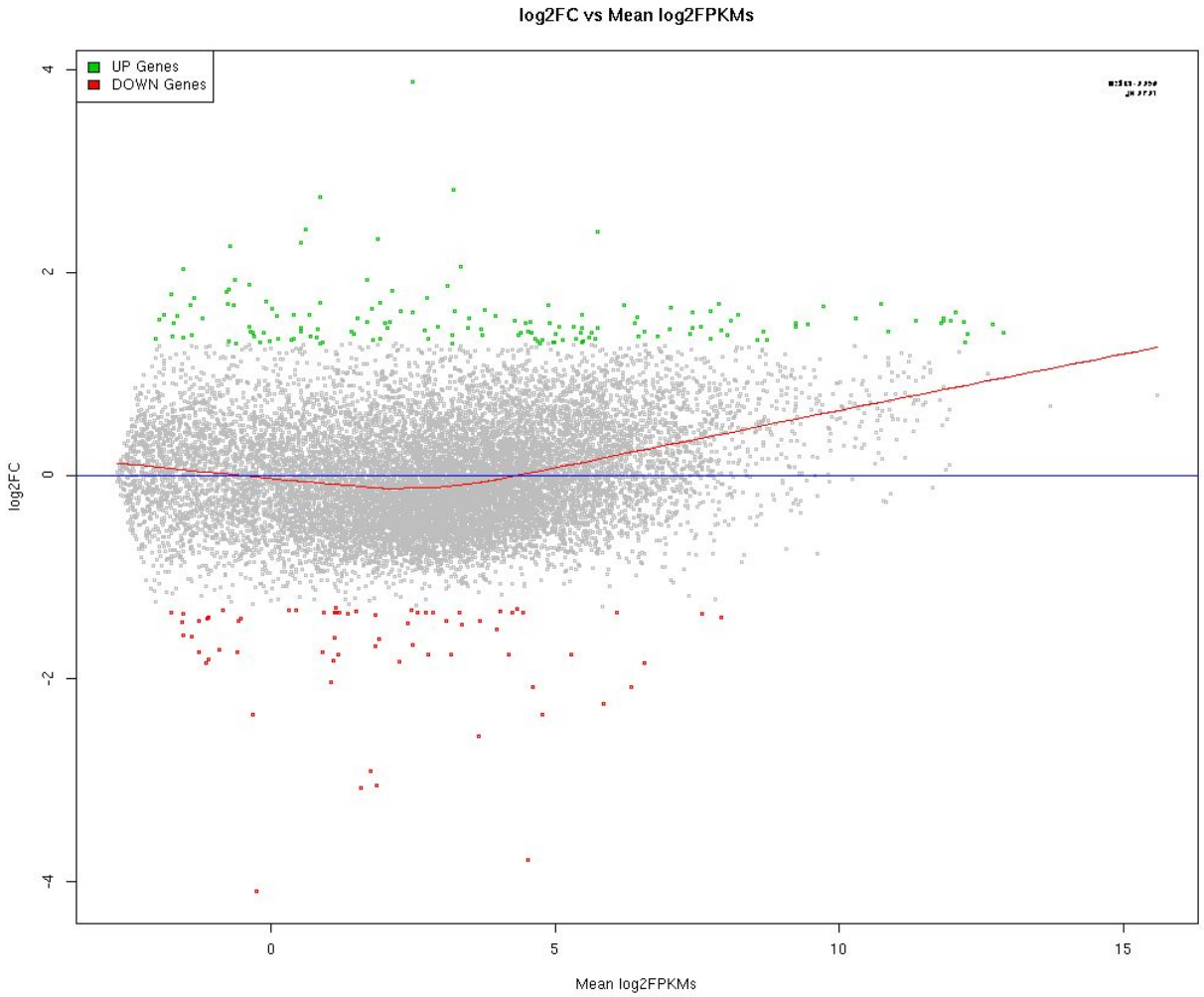
Journal of Statistical Software

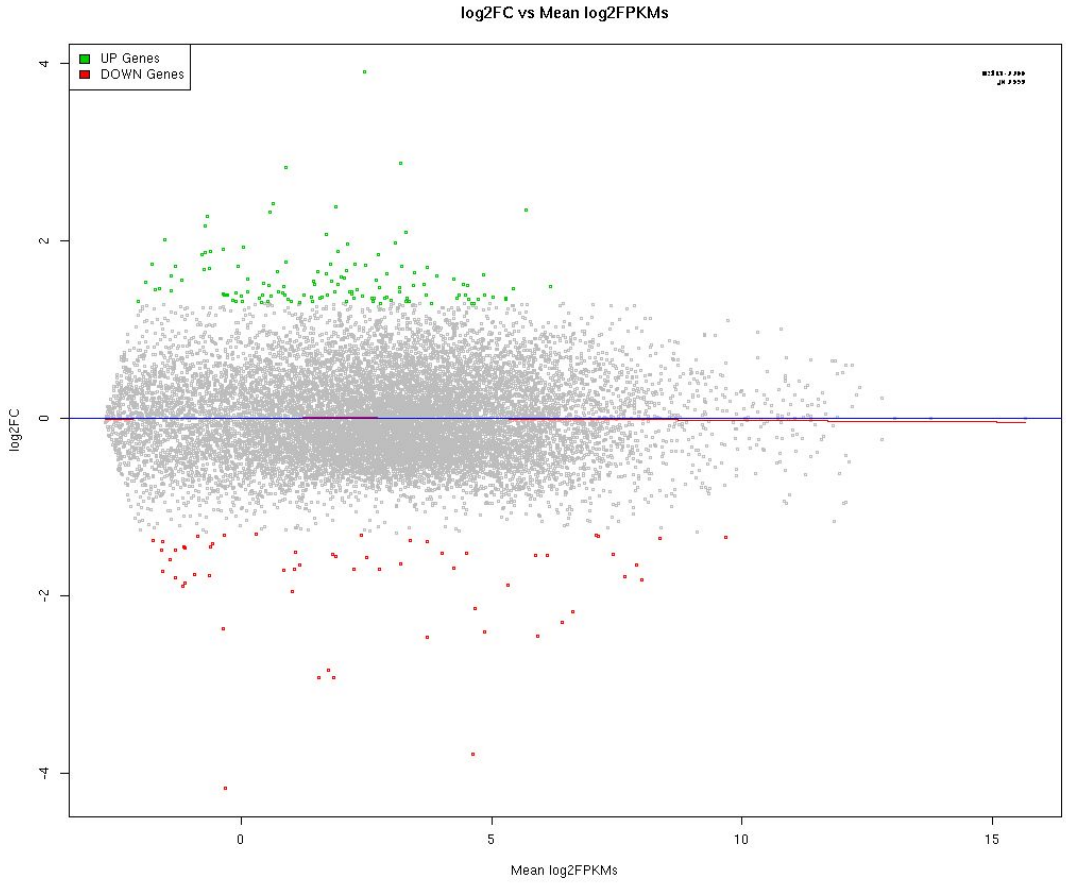
Information on Mission | Information for Authors | Style Guide | Volumes | About

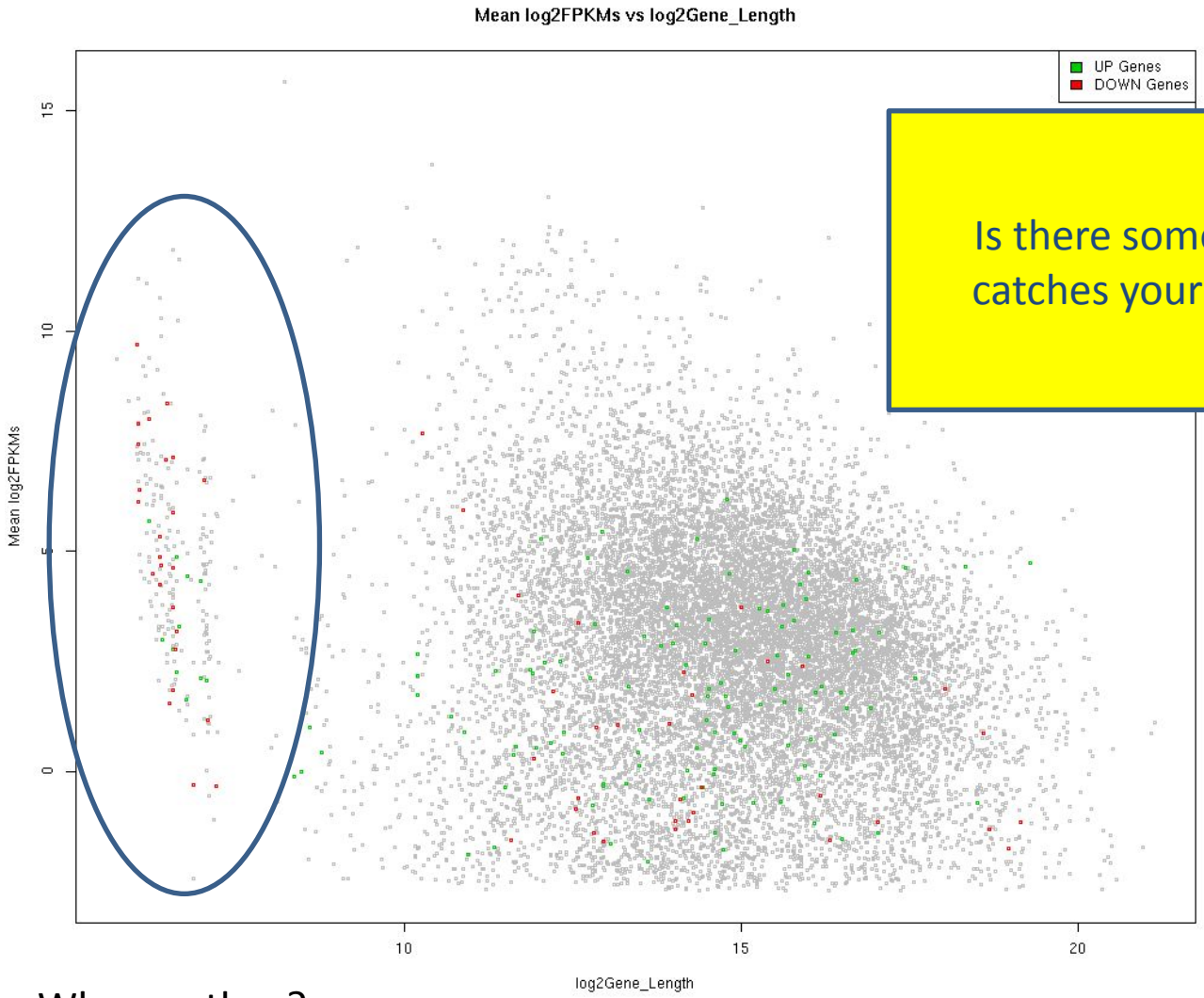
Home / Archives / Vol. 56 (2014) / Issue 7

lmdme: Linear Models on Designed Multivariate Experiments in R

Cristóbal Fresno, Mónica Balzarini, Elmer Fernández


log2FC vs Mean log2FPKMs
UP Genes
DOWN Genes
log2FC
Mean log2FPKMs
4
2
0
-2
-4
0
5
10
15


log2FC vs Mean log2FPKMs
UP Genes
DOWN Genes
log2FC
Mean log2FPKMs
4
2
0
-2
-4
0
5
10
15



Is there something that catches your attention?

Who are they?

So

- I did the alignment, normalization, summarization, DE…

To be continued…