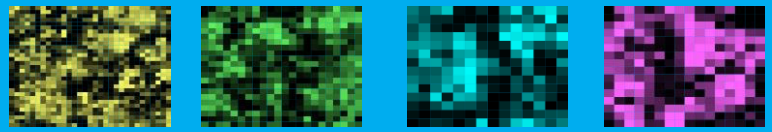
Data Normalization



Network Analysis in Systems Biology

Avi Ma'ayan, PhD

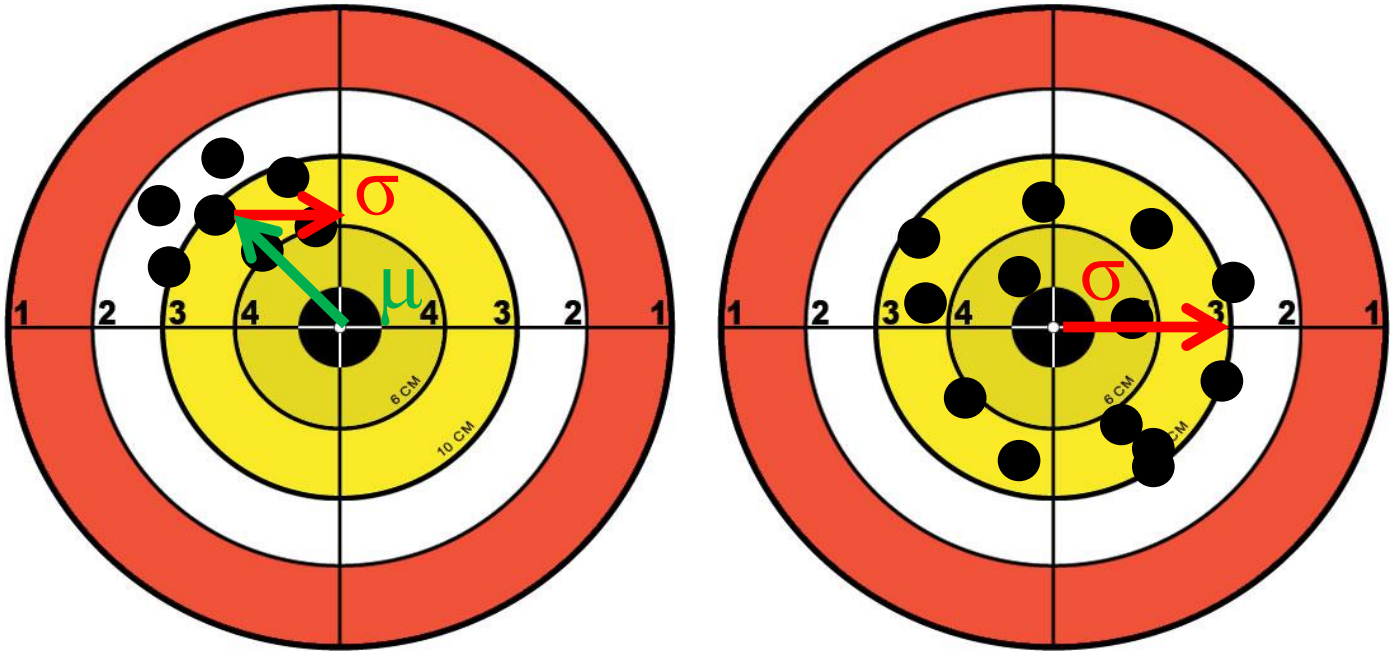
Associate Professor

Department of Pharmacology and Systems Therapeutics

Icahn School of Medicine at Mount Sinai, New York, NY 10029

Mount Sinai

systematic vs. non-systematic measurement errors



The Mean and Standard Deviation

Mean

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Standard Deviation

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}$$

Z-score Normalization Example

Original			Mean and STDDEV		Subtract the mean			Divide by the Standard deviation		
2	4	4	3.33	1.15	-1.3	0.6	0.6	-1.2	0.6	0.6
5	4	14	7.66	5.50	-2.6	-3.6	6.3	-0.5	-0.6	1.1
4	6	8	6	2	-2	0	2	-1	0	1
3	5	8	5.33	2.51	-2.3	-0.3	2.6	-0.9	-0.1	1.1
3	3	9	5	3.46	-2	-2	4	-0.6	-0.6	1.2

After Z-score normalization the mean becomes 0 and the standard deviation becomes 1

Quantile normalization example

Original

2	4	4
5	4	14
4	6	8
3	5	8
3	3	9

Ranked

2	3	4
3	4	8
3	4	8
4	5	9
5	6	14

Averaged

3	3	3
5	5	5
5	5	5
6	6	6
8	8	8

Re-ordered

3	5	3
8	5	8
6	8	5
5	6	5
5	3	6

After quantile normalization the columns adds up to the same value

Median Polish Normalization

4	3	6	4	7	4
8	1	10	5	11	8
6	2	7	8	8	7
9	4	12	9	12	9
7	5	9	6	10	7

Row medians

0	-1	2	0	3
0	-7	2	-3	3
-1	-5	0	1	1
0	-5	3	0	3
0	-2	2	-1	3

Matrix after removing row medians

Median Polish Normalization

0	-1	2	0	3
0	-7	2	-3	3
-1	-5	0	1	1
0	-5	3	0	3
0	-2	2	-1	3
0	**-5**	**2**	**0**	**3**

Column medians

0	4	0	0	0
0	-2	0	-3	0
-1	0	-2	1	-2
0	0	1	0	0
0	3	0	-1	0

Matrix after subtracting
Column medians

Median Polish Normalization

0	4	0	0	0	0
0	-2	0	-3	0	0
-1	0	-2	1	-2	-1
0	0	1	0	0	0
0	3	0	-1	0	0

0	4	0	0	0
0	-2	0	-3	0
0	1	-1	2	-1
0	0	1	0	0
0	3	0	-1	0

Matrix after removing row medians

Median Polish Normalization

0	4	0	0	0
0	-2	0	-3	0
0	1	-1	2	-1
0	0	1	0	0
0	3	0	-1	0

0	1	0	0	0

0	3	0	0	0
0	-3	0	-3	0
0	0	-1	2	-1
0	-1	1	0	0
0	2	0	-1	0

Matrix after removing column medians

Median Polish Normalization

0	3	0	0	0	0
0	-3	0	-3	0	0
0	0	-1	2	-1	0
0	-1	1	0	0	0
0	2	0	-1	0	0
0	0	0	0	0	

Row and column medians are 0. The median polish algorithm converged. The matrix left is the residual matrix. We subtract the residual matrix from the original matrix.

Median Polish Normalization

4	3	6	4	7
8	1	10	5	11
6	2	7	8	8
9	4	12	9	12
7	5	9	6	10

−

0	3	0	0	0
0	-3	0	-3	0
0	0	-1	2	-1
0	-1	1	0	0
0	2	0	-1	0

=

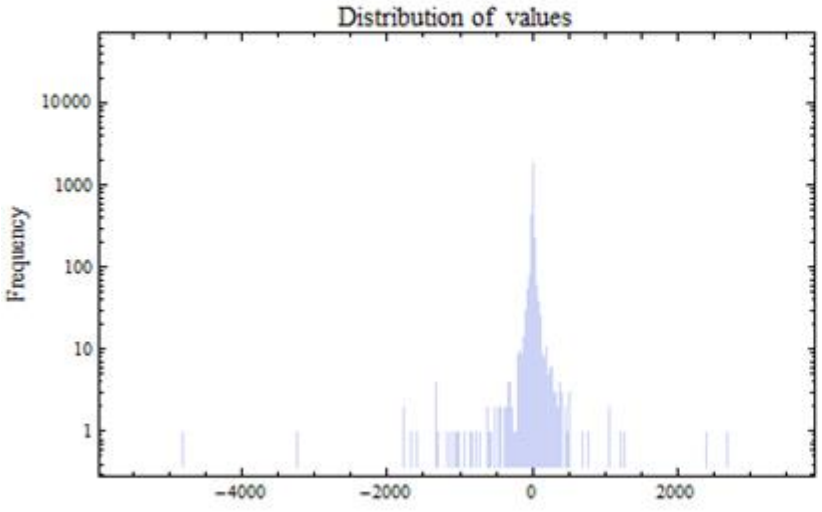
4	0	6	4	7
8	4	10	8	11
6	2	8	6	9
9	5	11	9	12
7	3	9	7	10

4.2
8.2
6.2
9.2
7.2

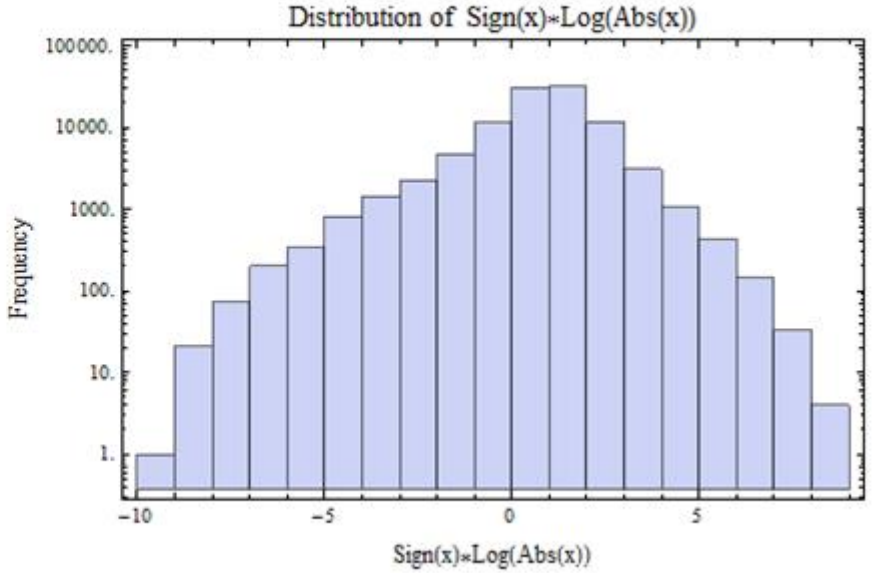
In green is the robust mean RMA expression for the probe set.

Log Transformation of Data

Original

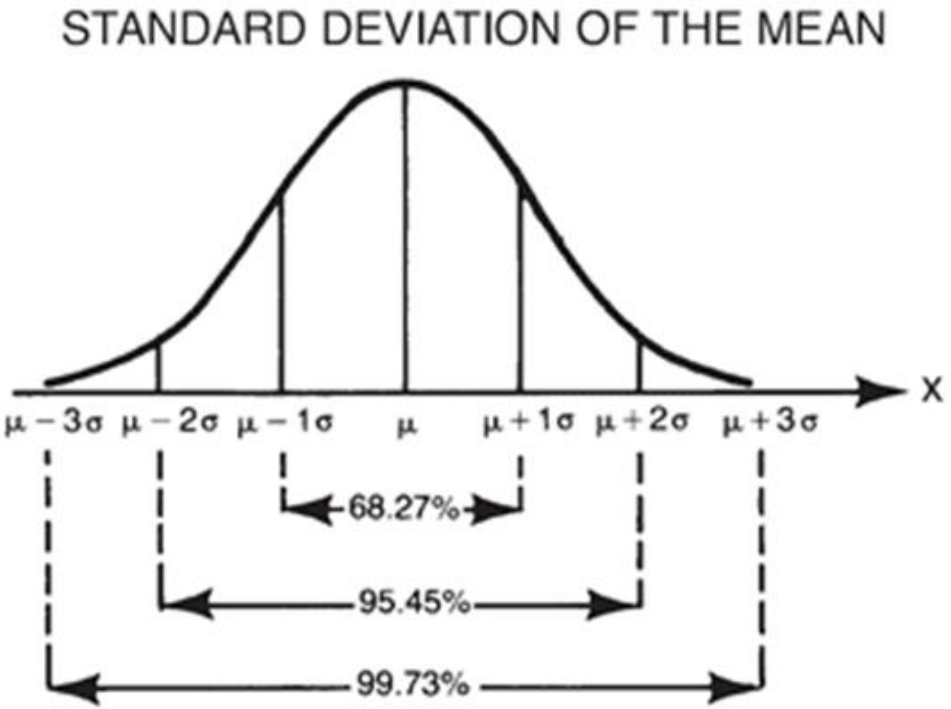


Log transformed

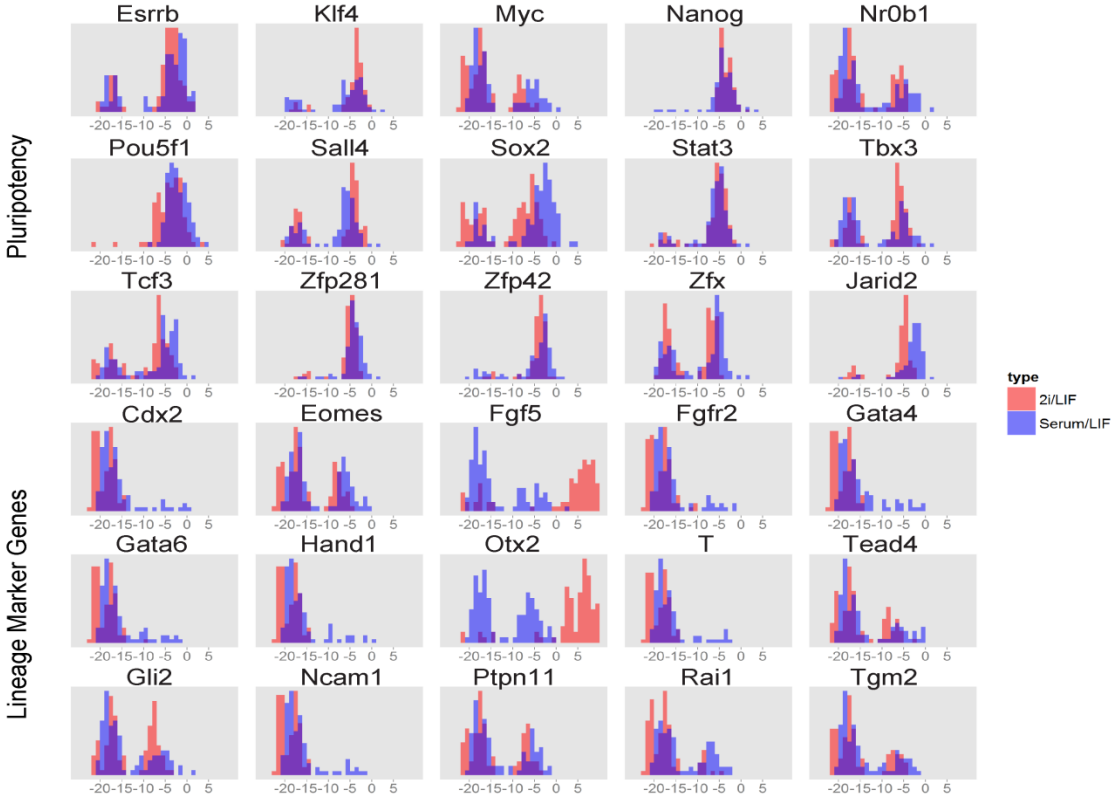


Attempt to lessen the dominance of the extreme values using log transformation

Data Distributions - Normal



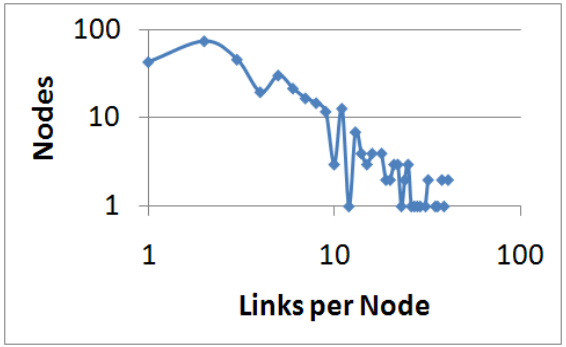
Data Distributions – Bimodal

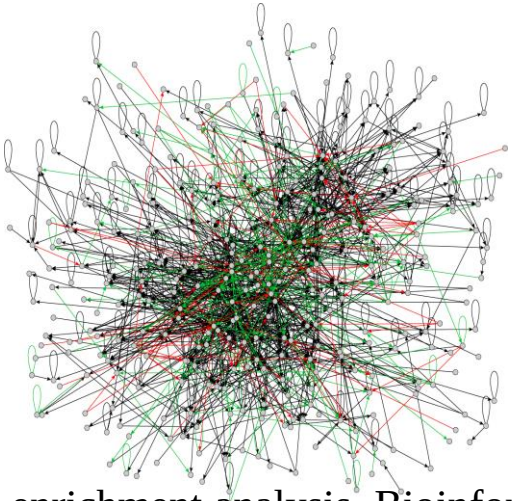


Xu H, Ang Y-S, Sevilla A, Lemischka IR, Ma'ayan A (2014) Construction and Validation of a Regulatory Network for Pluripotency and Self-Renewal of Mouse Embryonic Stem Cells. PLoS Comput Biol 10(8): e1003777

The Human Kinome Network

This mammalian kinase-kinase subnetwork extracted from the kinase-substrate network consists of 356 kinases and phosphatases (331 kinases and 25 phosphatases) 1380 interactions extracted from 1072 papers 1322 phosphorylations and 58 dephsphorylations. The average link per node is 7.15 whereas the connectivity distribution fits a power-law.





Lachmann A, Ma'ayan A. KEA: kinase enrichment analysis. Bioinformatics. 2009 Mar 1;25(5):684-6.

References

https://www.coursera.org/course/getdata

http://en.wikipedia.org/wiki/Standard_score

http://en.wikipedia.org/wiki/Quantile_normalization

http://en.wikipedia.org/wiki/Median_polish

http://www.r-statistics.com/2013/05/log-transformations-for-skewed-and-wide-distributions-from-practical-data-science-with-r/

Xu H, Ang Y-S, Sevilla A, Lemischka IR, Ma'ayan A (2014) Construction and Validation of a Regulatory Network for Pluripotency and Self-Renewal of Mouse Embryonic Stem Cells. PLoS Comput Biol 10(8): e1003777

Lachmann A, Ma'ayan A. KEA: kinase enrichment analysis. Bioinformatics. 2009 Mar 1;25(5):684-6.