Bioinformatics Institute
BII

SIMPLE STATISTICAL LAWS FOR HIGHER RESOLUTION OMICS DATA ANALYSIS

Kumar Selvarajoo, PhD

Senior Principal Investigator, Computational Biology & Omics
Bioinformatics Institute, A*STAR
Singapore Institute of Food & Biotechnology Innovation, A*STAR
Adjunct Assoc Professor, NUS Yong Loo Lin School of Medicine
Adjunct Assoc Professor, SBS, NTU

29 Mar 2022

COMPUTATIONAL BIOLOGY & OMICS

Members

Dr Mohamed Helmy
Dr Hock Chuan Yeo
Olga Sirbu
Dr Jasmeet Kaur (SIFBI)

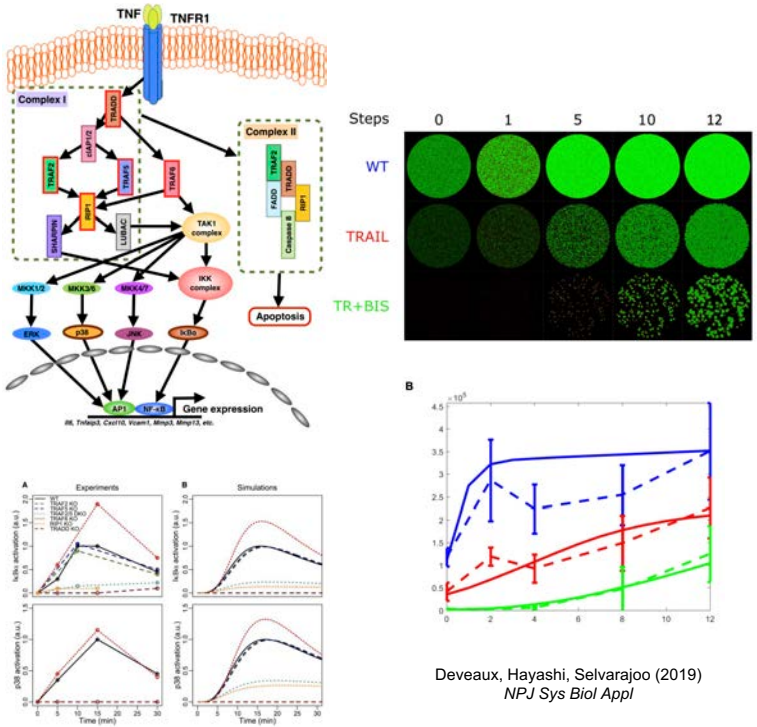
Co-Members & Interns

Shi Mun Lee (ASRL)
Clarence Sim (NTU)
Vivien Paitimusa (NUS)
Vissale Srinivasan (India)
Parian Hatami (Iran)

CREATING GROWTH, ENHANCING LIVES

COMPUTATIONAL BIOLOGY & OMICS

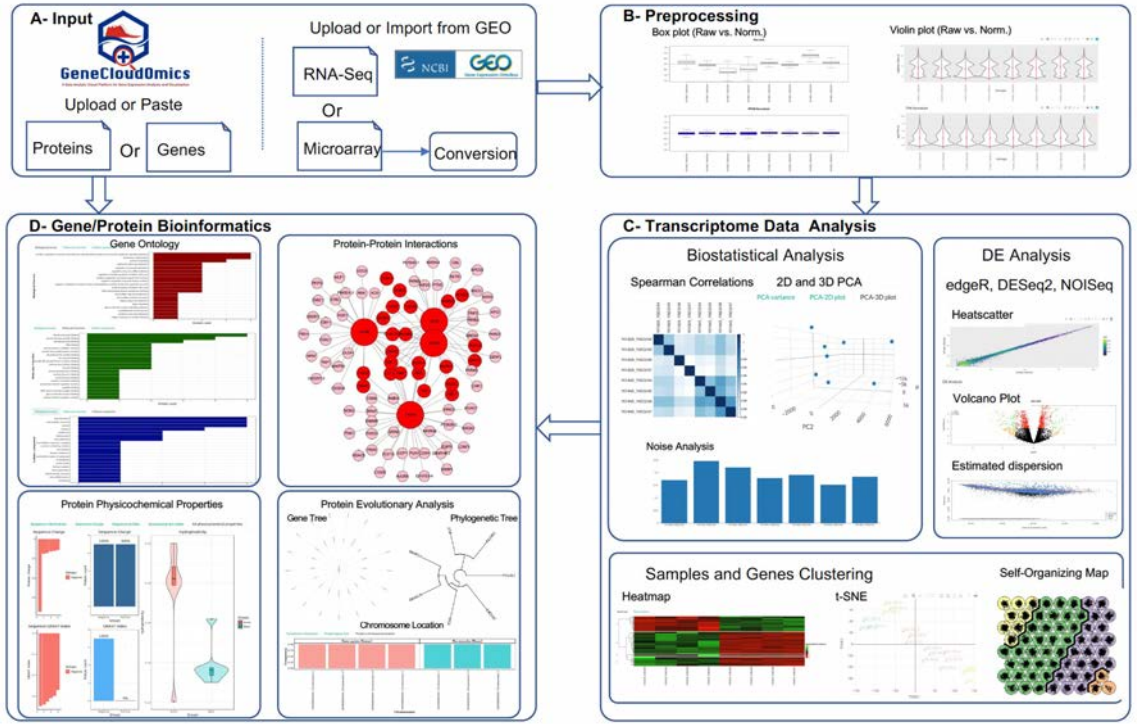
Dynamic Modeling



Hayashi,.., Selvarajoo (2013) *Cell Comm Signal*

Deveaux, Hayashi, Selvarajoo (2019)
NPJ Sys Biol Appl

Data Analytics & Machine Learning



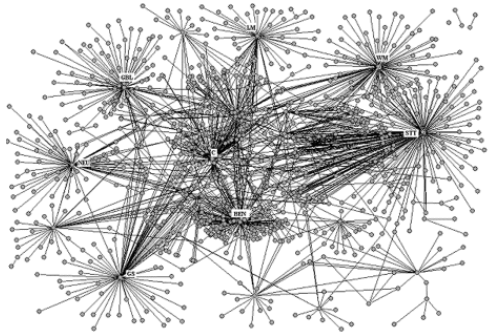
Helmy,.., Selvarajoo (2021) *Front Bioinfor*

LARGE SCALE NETWORKS ORGANIZATION

Important to understand statistical structures

United Airlines.com

Skype



Koppel M (2006)

Williams S (2012) Science

CREATING GROWTH, ENHANCING LIVES

SCALE-FREE NETWORKS

Emergence of Scaling in Random Networks

Albert-László Barabási* and Réka Albert

Systems as diverse as genetic networks or the World Wide Web are best described as networks with complex topology. A common property of many large networks is that the vertex connectivities follow a scale-free power-law distribution. This feature was found to be a consequence of two generic mechanisms: (i) networks expand continuously by the addition of new vertices, and (ii) new vertices attach preferentially to sites that are already well connected. A model based on these two ingredients reproduces the observed stationary scale-free distributions, which indicates that the development of large networks is governed by robust self-organizing phenomena that go beyond the particulars of the individual systems.

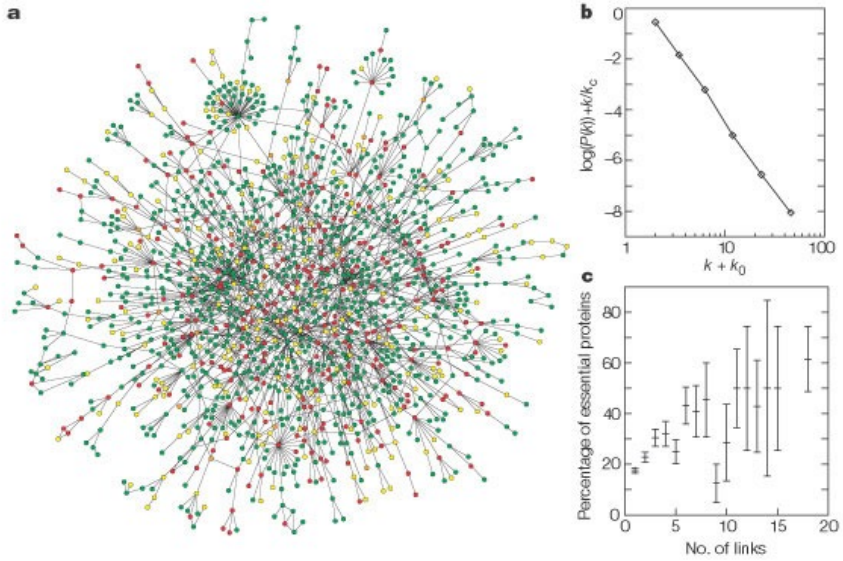
The inability of contemporary science to describe systems composed of nonidentical elements that have diverse and nonlocal interactions currently limits advances in many disciplines, ranging from molecular biology to computer science (*1*). The difficulty of describing these systems lies partly in their topology: Many of them form rather complex networks whose vertices are the elements of the system and whose edges represent the interactions between them. For example, liv-

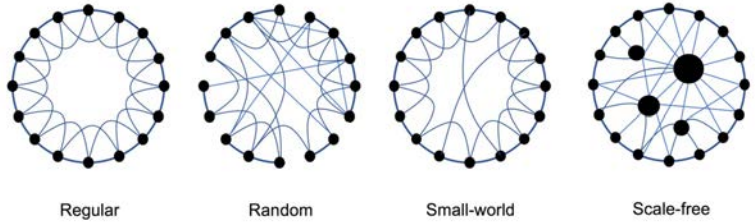
Department of Physics, University of Notre Dame, Notre Dame, IN 46556, USA.

*To whom correspondence should be addressed. E-mail: alb@nd.edu

encemag.org SCIENCE VOL 286 15 OCTOBER 1999

Yeast PPI



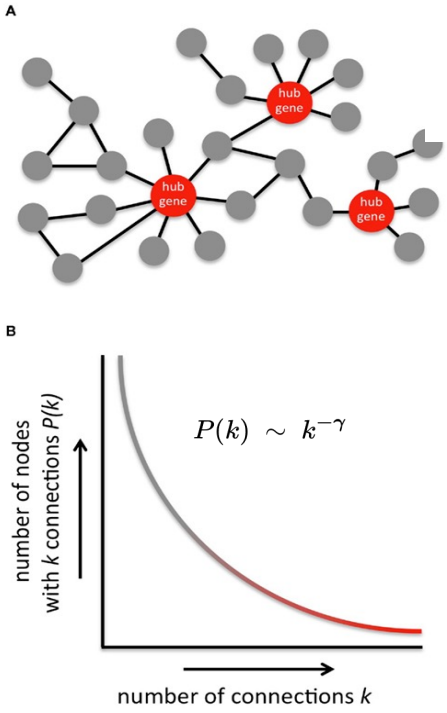


Jeong,…,Barabasi (2001) Nature

CREATING GROWTH, ENHANCING LIVES

SCALE-FREE NETWORKS IN BIOLOGY

A scale-free network follows a power-law degree distribution asymptotically.



van Kesteren Re *et al* (2011) *Front. Mol. Neurosic.*

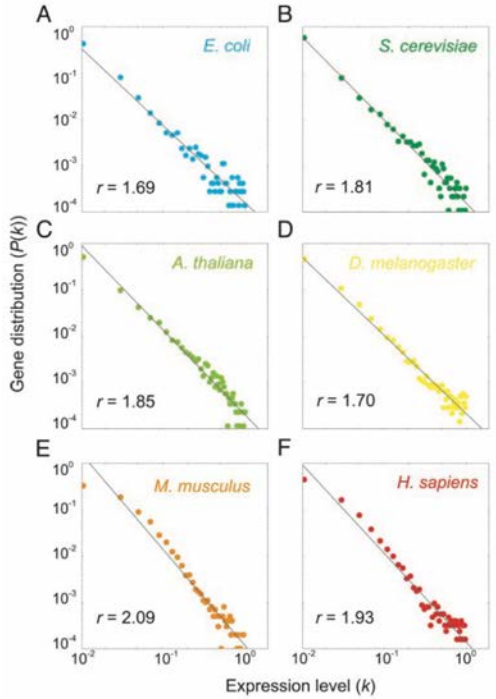
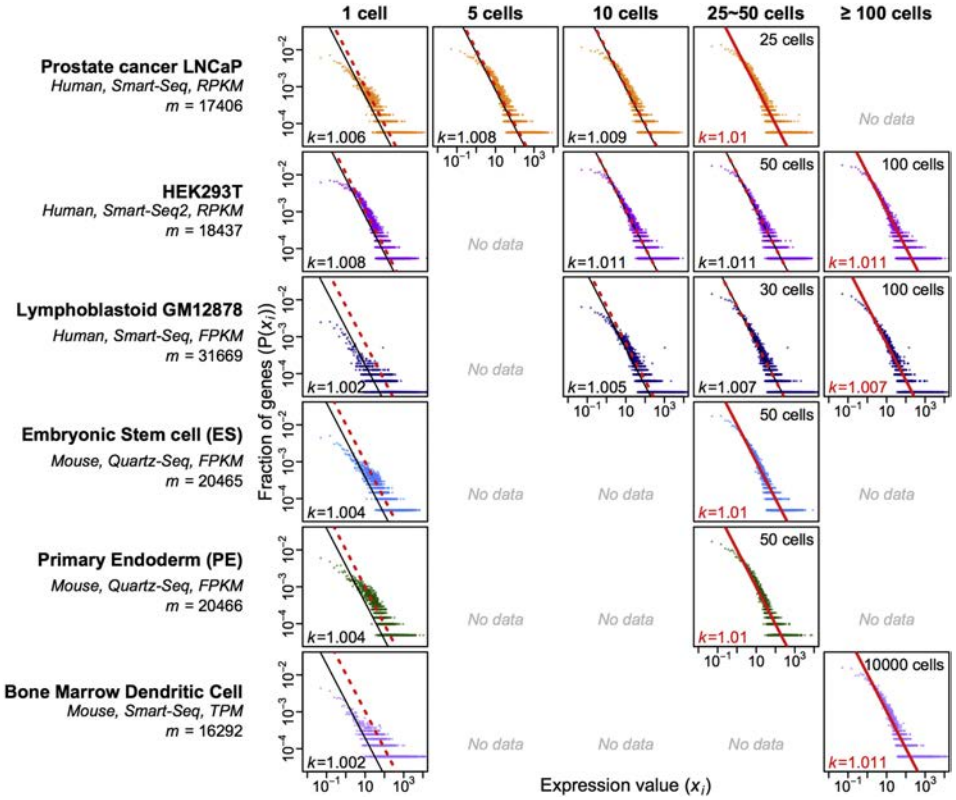


Fig. 1. Evolutional conservation of transcriptional organization. The distributions of gene expression levels in *E. coli* (*A*), *S. cerevisiae* (*B*), *A. thaliana* (*C*), *D. melanogaster* (*D*), *M. musculus* (*E*), and *H. sapiens* (*F*) exhibit a power-law distribution in which the probability that a gene has an expression level *k*, decays as a power law, $P(k) \propto k^{-r}$. A straight line in each panel represents the estimated power-law distribution. The estimated value of exponent *r* is indicated in the lower left corner of each panel.

Ueda HR *et al* (2004) *PNAS*

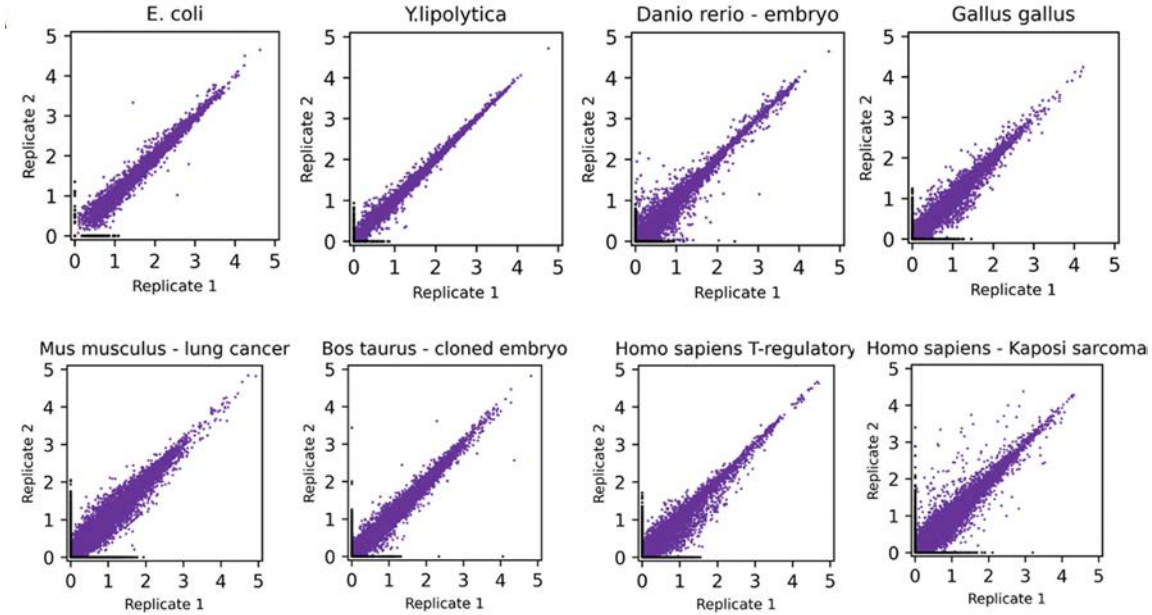
FITTING POWER LAW INTO GENE EXPRESSIONS



Piras & Selvarajoo (2015) *Genomics*

CREATING GROWTH, ENHANCING LIVES

TRANSCRIPTOME-WIDE REPLICATE SCATTER PLOTS



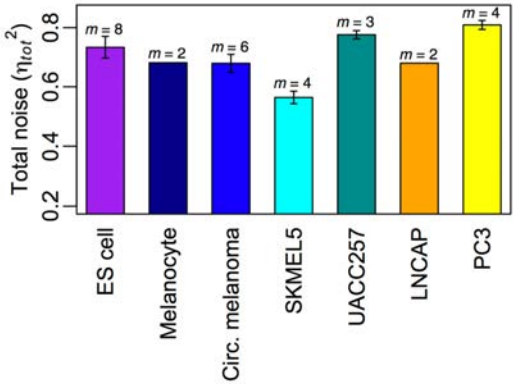
Giuliani, Bui, Helmy & Selvarajoo K (2022) *Genomics*

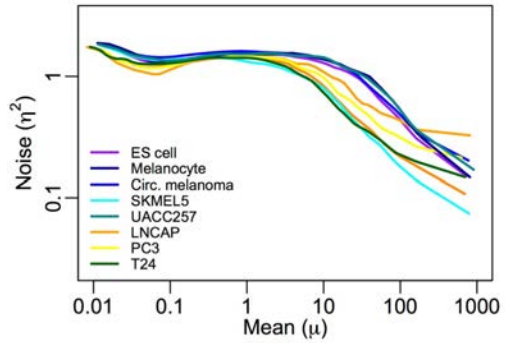
HOW TO BE OBJECTIVE IN SELECTING SUBSET OF GENES FOR FURTHER ANALYSIS?

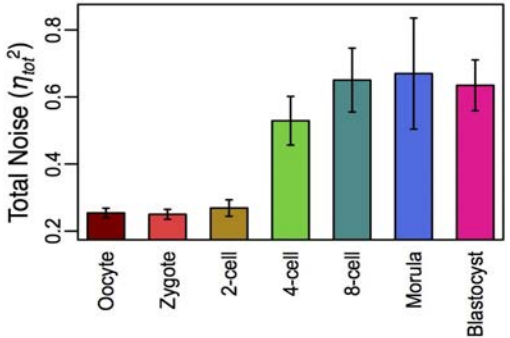
CREATING GROWTH, ENHANCING LIVES

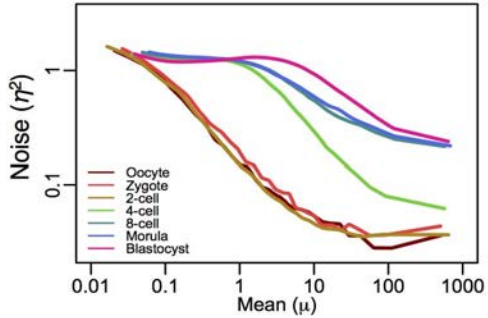
QUANTIFYING TRANSCRIPTOME-WIDE SCATTER

$$\eta^2_{ijk} = \sigma^2_{ijk} \Big/ \mu^2_{ijk} \qquad \eta^2_i = \frac{2}{m(m-1)} \sum_{j=1}^{m-1} \sum_{k=j+1}^{m} \eta^2_{ijk} \qquad \eta^2_{tot} = \frac{1}{n} \sum_{i=1}^{n} \eta^2_i$$



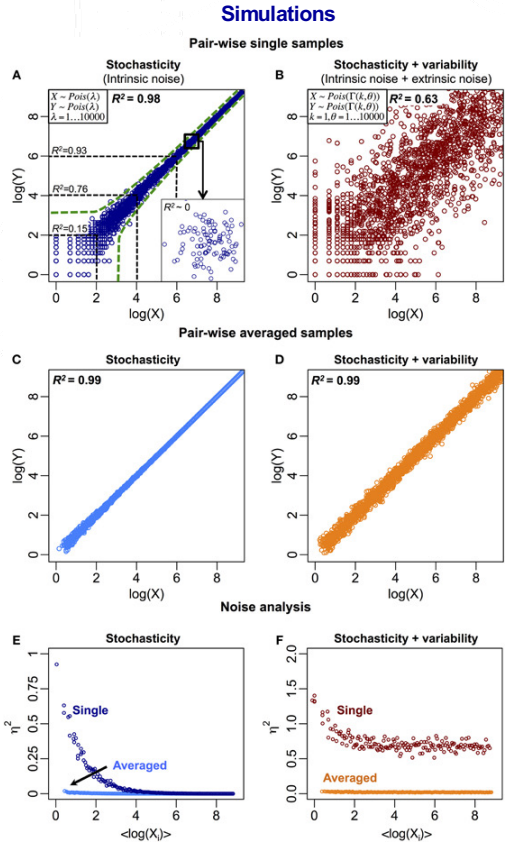






Piras, Tomita & Selvarajoo (2014) *Sci Rep*

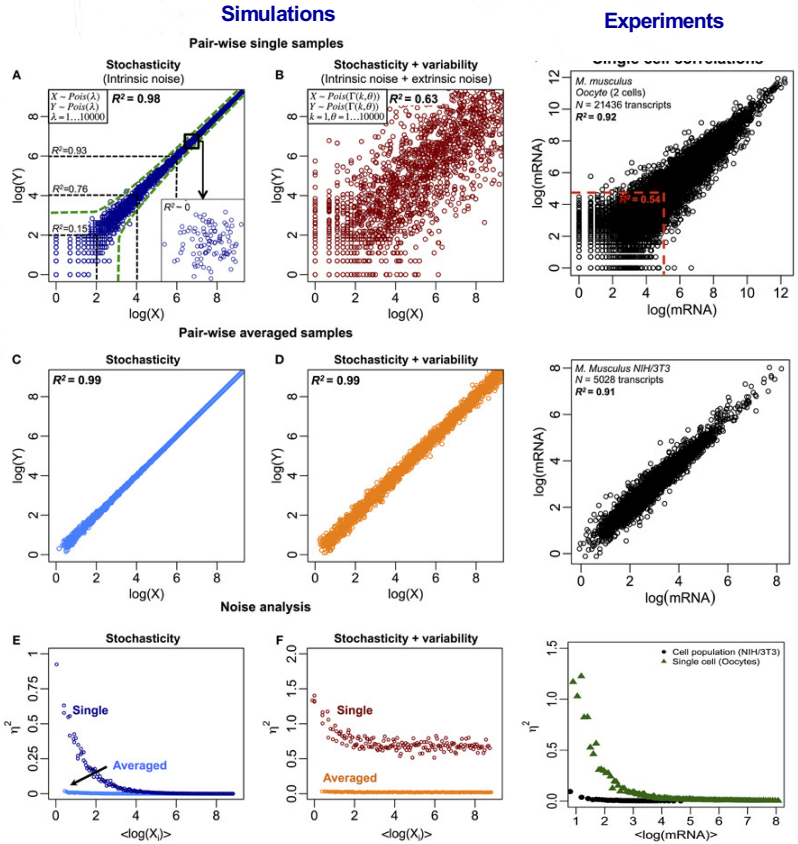
ARITIFICIAL DATA GENERATION TO UNDERSTAND TRANSCRIPTOME-WIDE SCATTER



Piras, Tomita & Selvarajoo (2012) *Front Physiol*

STOCHASTIC & VARIABLE NOISE REDUCE AT POPULATION SCALE

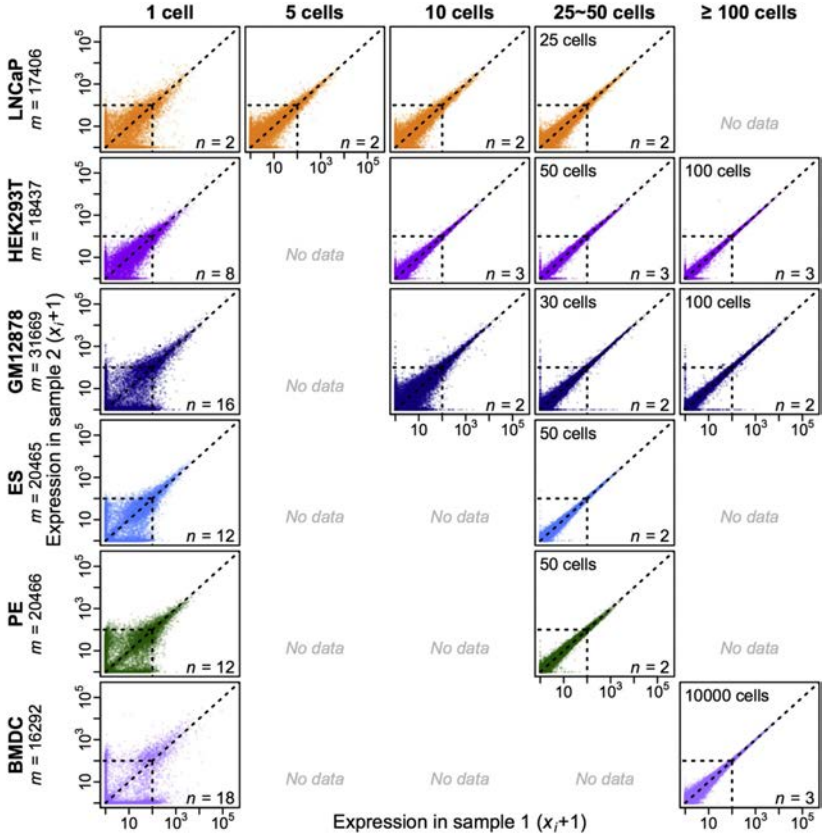
ARITIFICIAL DATA GENERATION TO UNDERSTAND TRANSCRIPTOME-WIDE SCATTER



Piras, Tomita & Selvarajoo (2012) *Front Physiol*

STOCHASTIC & VARIABLE NOISE REDUCE AT POPULATION SCALE

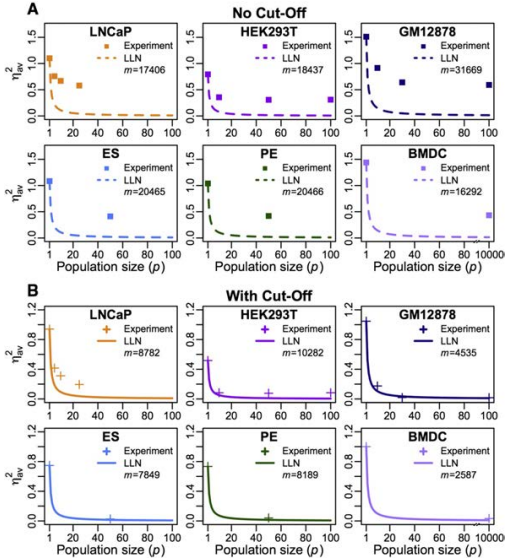
REDUCTION OF GENE EXPRESSION VARIABILITY FROM SINGLE CELLS TO POPULATIONS



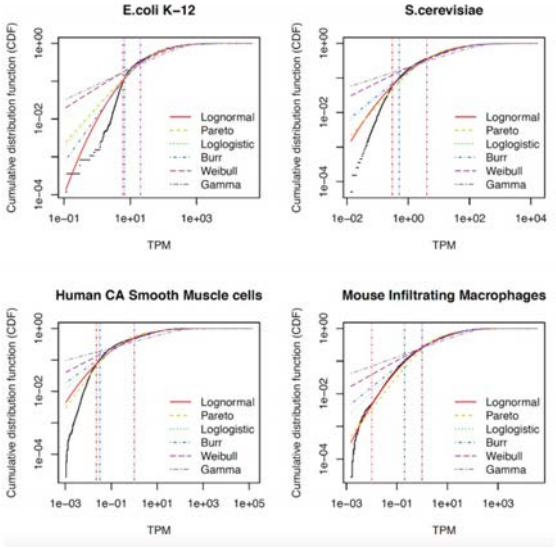
Piras & Selvarajoo (2015) *Genomics*

A	Population size (p)				
Pearson (R)	1	5	10	25-50	≥100
LNCaP	0.701	0.964	0.958	0.986	
HEK293T	0.791		0.994	0.989	0.988
GM12878	0.831		0.977	0.997	0.997
ES	0.944			0.997	
PE	0.924			0.989	
BMDC	0.583				0.998

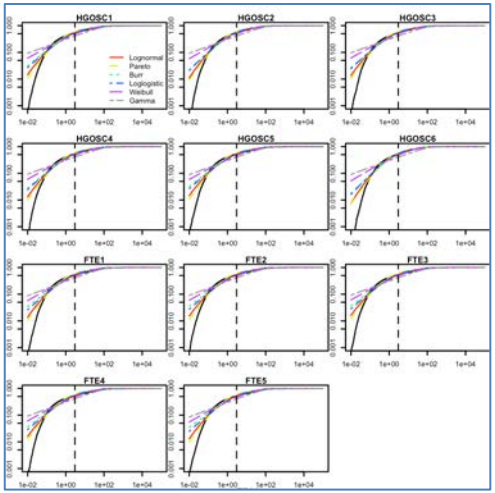
B	Population size (p)				
Spearman (ρ)	1	5	10	25-50	≥100
LNCaP	0.707	0.859	0.885	0.916	
HEK293T	0.824		0.954	0.960	0.962
GM12878	0.590		0.824	0.873	0.882
ES	0.738			0.952	
PE	0.780			0.951	
BMDC	0.599				0.951

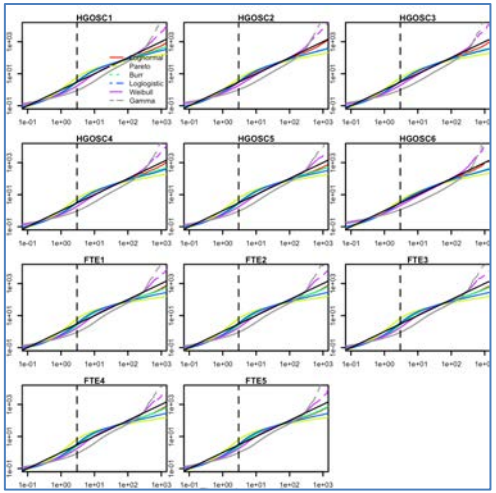


STATISTICAL DISTRIBUTIONS IN GENE EXPRESSIONS



Bui, Giuliani & Selvarajoo (2018) *Organisms*





$$\mathrm{AIC} = 2k - 2\ln(\hat{L})$$

Lognormal	Pareto	Burr	Loglogistic	Weibull	Gamma	min_AIC	cutoff_value
182334.622305679	185138.3729234	184511.298174322	184546.199345831	186248.437610793	195172.952049535	Lognormal	3
186068.689254044	188442.448569136	188056.183232653	188055.699224736	191097.489423336	202390.453780023	Lognormal	3
170006.092819634	172450.294874584	171894.437204904	171898.99358787	174244.334934714	184525.205882947	Lognormal	3
173541.933762983	176114.138569288	175477.739683717	175499.450644427	177576.34646291	187132.842822571	Lognormal	3
181239.704839802	184121.469005915	183362.205200763	183430.700526238	184774.55598799	192385.714712317	Lognormal	3
168669.324374807	170766.996176221	170293.423915209	170295.886610194	173407.015843268	186164.73005373	Lognormal	3
193842.63724364	197047.632290774	196153.484528826	196288.873144072	197512.468743641	205910.268064475	Lognormal	3
189942.234820299	193260.443278287	192146.691768559	192320.366637287	193516.280740822	202160.301488673	Lognormal	3
188822.803491237	192107.494800117	191068.519064182	191227.059418503	192458.130870089	201200.576306569	Lognormal	3
191184.058113789	194439.168957514	193535.325188924	193688.320724645	194883.64872641	203529.412248312	Lognormal	3
193412.496131042	196824.198953224	195579.717472085	195825.686218319	196725.146452413	204812.236581635	Lognormal	3

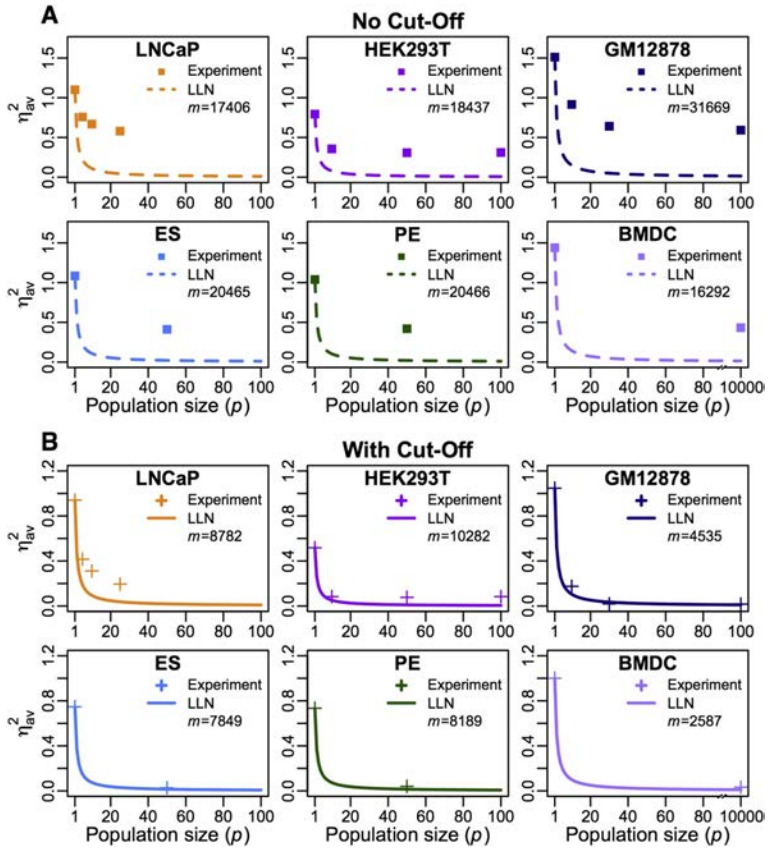
GeneCloudOmics
A Data Analytic Cloud Platform for Gene Expression Analysis and Visualization

Helmy,.., Selvarajoo (2021) *Front Bioinfor*

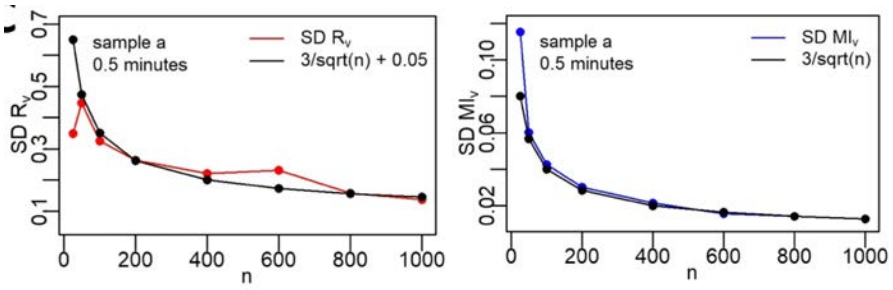
CREATING GROWTH, ENHANCING LIVES

CREATING GROWTH, ENHANCING LIVES

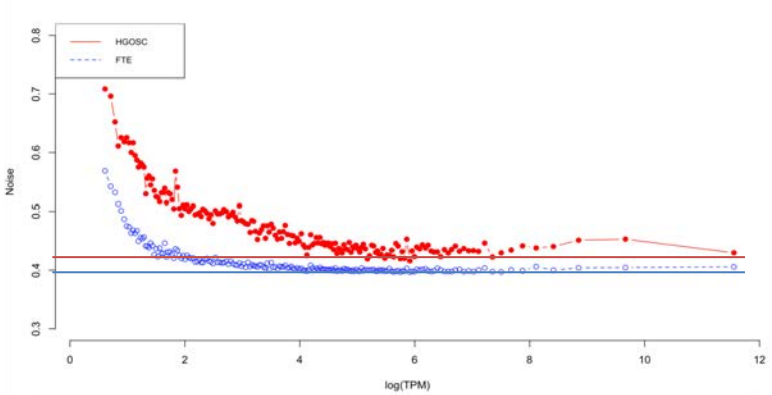
LAW OF LARGE NUMBERS OR INVERSE SQUARE ROOT



Piras & Selvarajoo (2015) *Genomics*



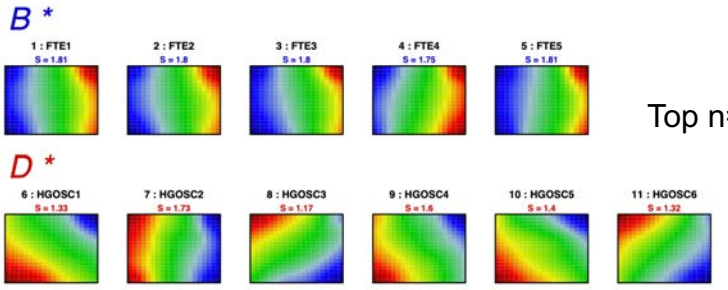
Bui & Selvarajoo (2020) *Sci Rep*



Olga, Helmy, Selvarajoo (in preparation)

CREATING GROWTH, ENHANCING LIVES

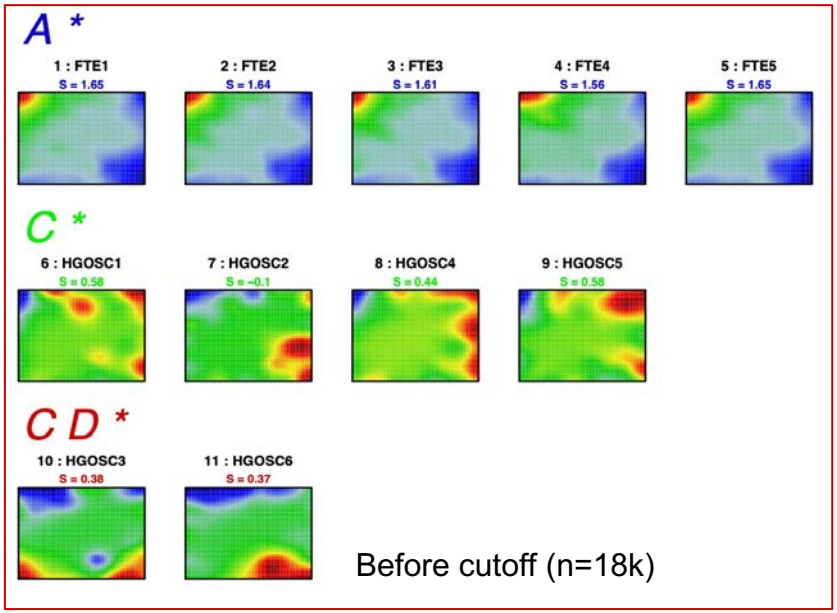
SELF-ORGANIZING MAPS FOR OVARIAN CANCERS



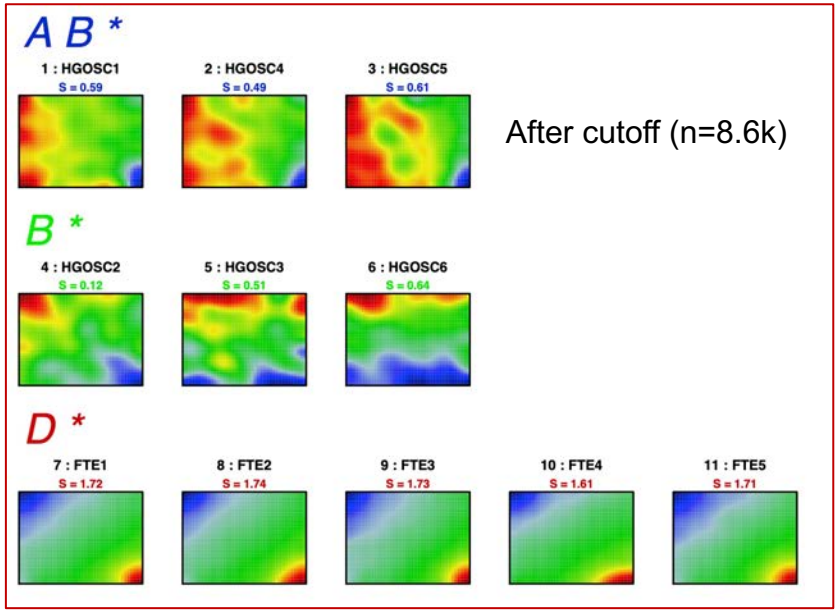
Top n=500

HGOSC: high grade ovarian serous cancer
FTE: Fallopian Tube Cells

GSE190688



Before cutoff (n=18k)

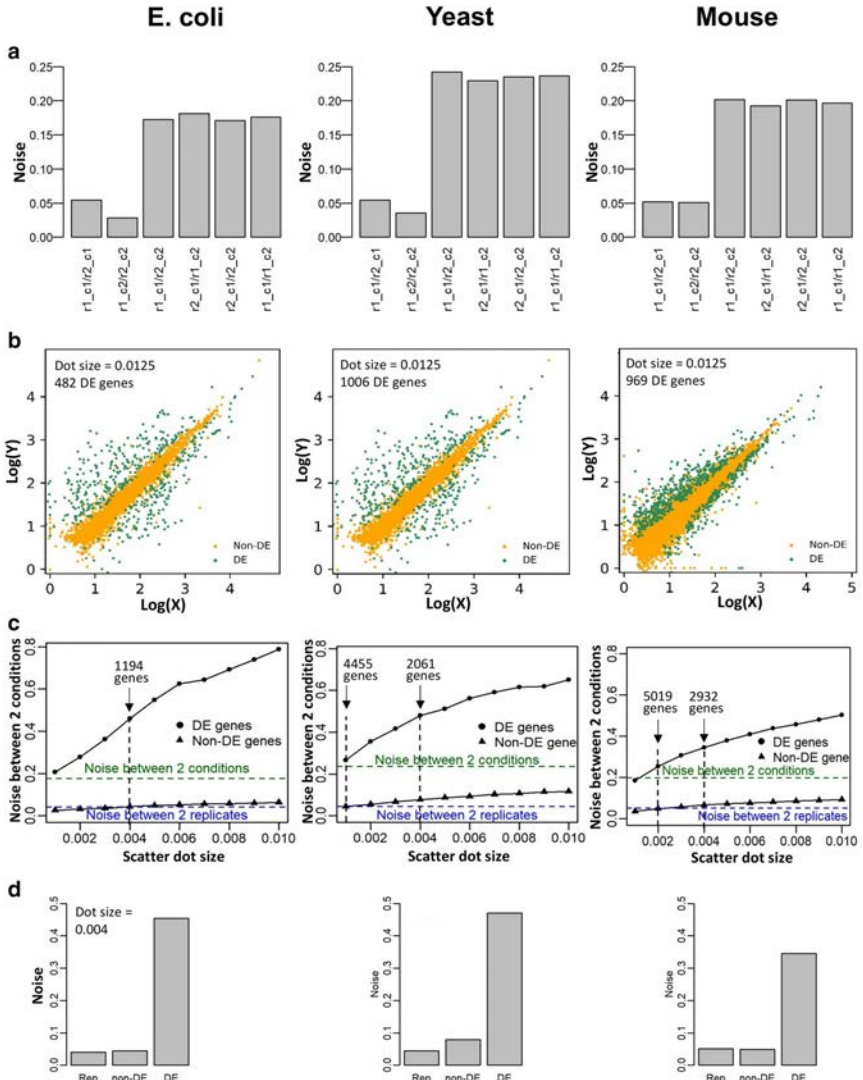


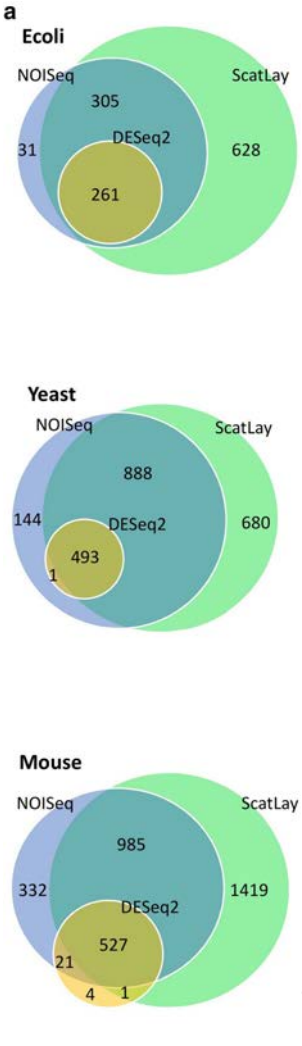
After cutoff (n=8.6k)

Olga, Helmy, Selvarajoo (in preparation)

CREATING GROWTH, ENHANCING LIVES

SCATLAY:
UTILIZING NOISE TO EXTRACT DIFFERENTIAL EXPRESSED GENES





Bui, Lee, Selvarajoo (2020) *Sci Rep*

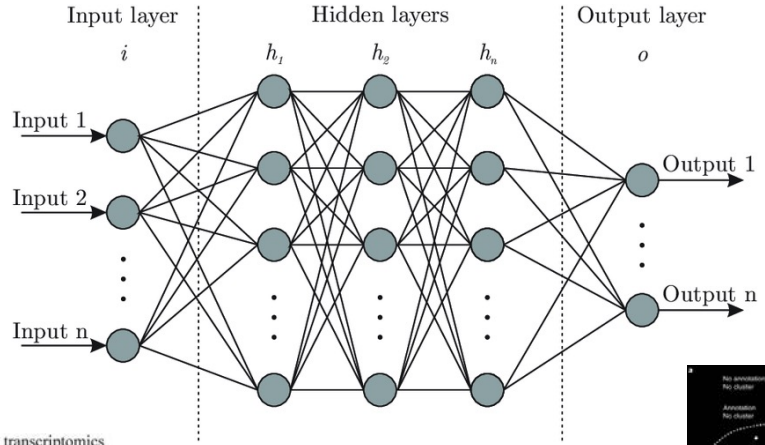
SYNTHETIC DATA & MACHINE LEARNING

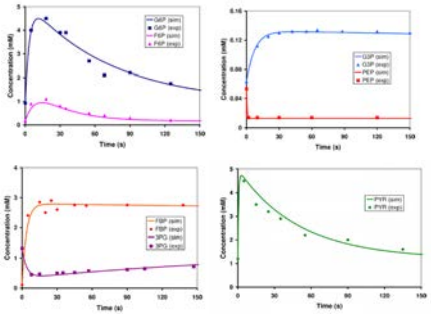
Dynamic/Spatial Gene Expression Data	Hidden Layers: Data Analytics & Neural Networks	Dynamic/Spatial Metabolites Concentrations

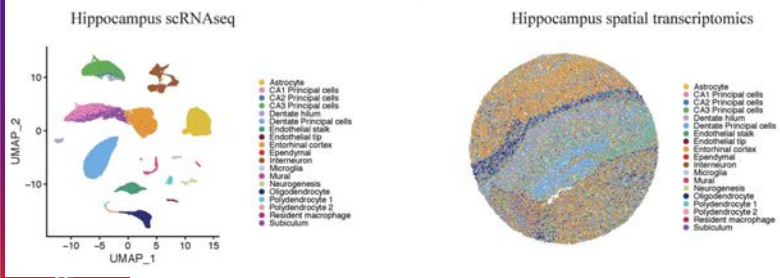
Gene names

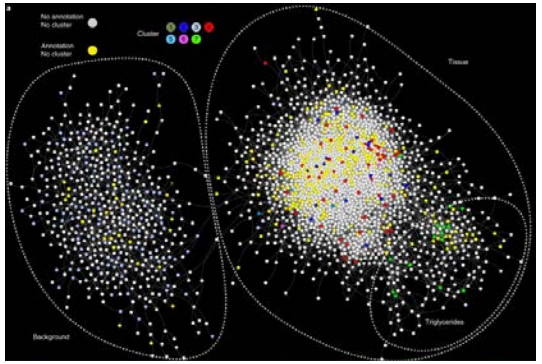
Raw read counts; Replicates of the same condition should be placed together

	a1 / t0	b1 / t0	c1 / t0	a2 / t1	b2 / t1	c2 / t1
15S_rRNA	12	0	8	17	6	1
21S_rRNA	195	180	113	268	93	125
HRA1	2	3	0	0	1	0
ICR1	98	148	197	84	161	540
LSR1	142	54	70	170	33	69
NME1	10	5	1	9	4	9
NTS1-2	3399	139	219	3362	106	241
PWR1	0	0	0	0	2	0









Song & Su (2020) *Brief Bioinfor*

Gieer et al (2020) *Nat Methods*

SUMMARY

- Gene expression distribution follows power-law or lognormal distributions
- Gene expression noise reduces similarly to law of large numbers
- Distribution fitting & Scatlay utilize statistical principles to provide wider coverage
- Using these information we can obtain a more objective selection of genes for further investigation
- Statistical Laws are useful for generating Synthetic Data for ML applications

CREATING GROWTH, ENHANCING LIVES

Bioinformatics Institute
BII

Singapore Institute of Food and Biotechnology Innovation
SIFBI

Nestlé

SOPHIE'S BIONUTRIENTS

KIKKOMAN

THANK YOU

SynCTI

tbi
Toulouse Biotechnology Institute
Bio & Chemical Engineering

twb
White Biotechnology
center of excellence

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Dr Vincent Piras
Evotec, France

Thuy Tien Bui
Northeastern, US

Prof Alessandro Giuliani, ISS, Italy