THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Australian Institute for Bioengineering and Nanotechnology

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY
Science at the heart of medicine

Investigating Skewness to Understand Gene Expression Heterogeneity in Large Patient Cohorts

Benjamin Church, Henry Williams, Jessica Mar

University of Queensland
Brisbane, Australia
j.mar@uq.edu.au

ICIBM Conference, 10 June 2019

Australian Institute for Bioengineering & Nanotechnology

AIBN provides a multidisciplinary environment at the interface between the **biological, chemical** and **computing sciences.**

Located in a custom designed $73 M building opened in October 2006.

Brisbane, Australia

The distribution captures information about the underlying cell population

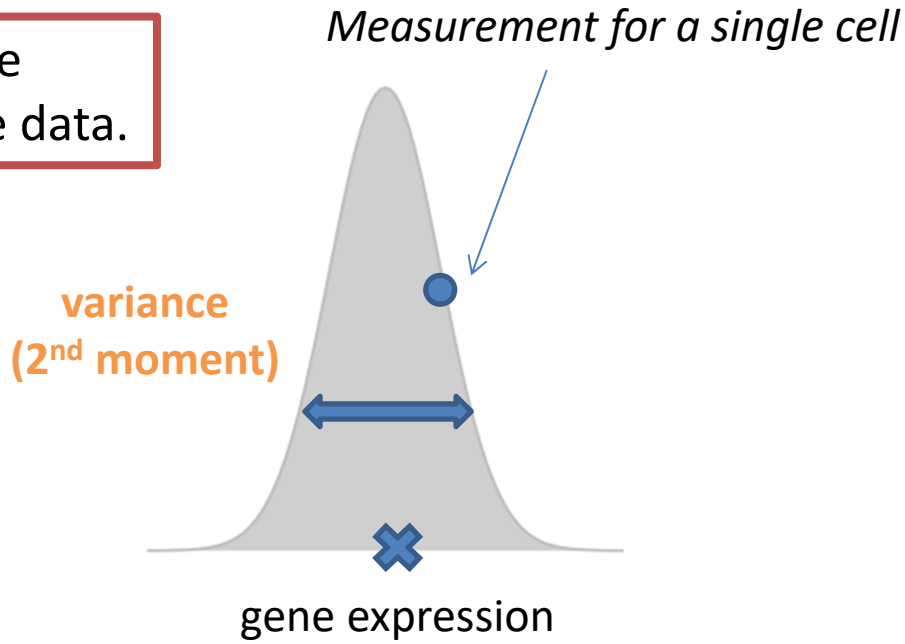
Consider the density of a gene's expression profile in a population of cells:

Statistical moments report on the underlying population structure of the data.

Other **higher moments**:
Skewness (3rd moment)
Kurtosis
Hyperskewness
Hyperflatness

Other features: bimodality, multi-modes.

Higher moments are good for identifying extreme values, outliers, and sub-populations.

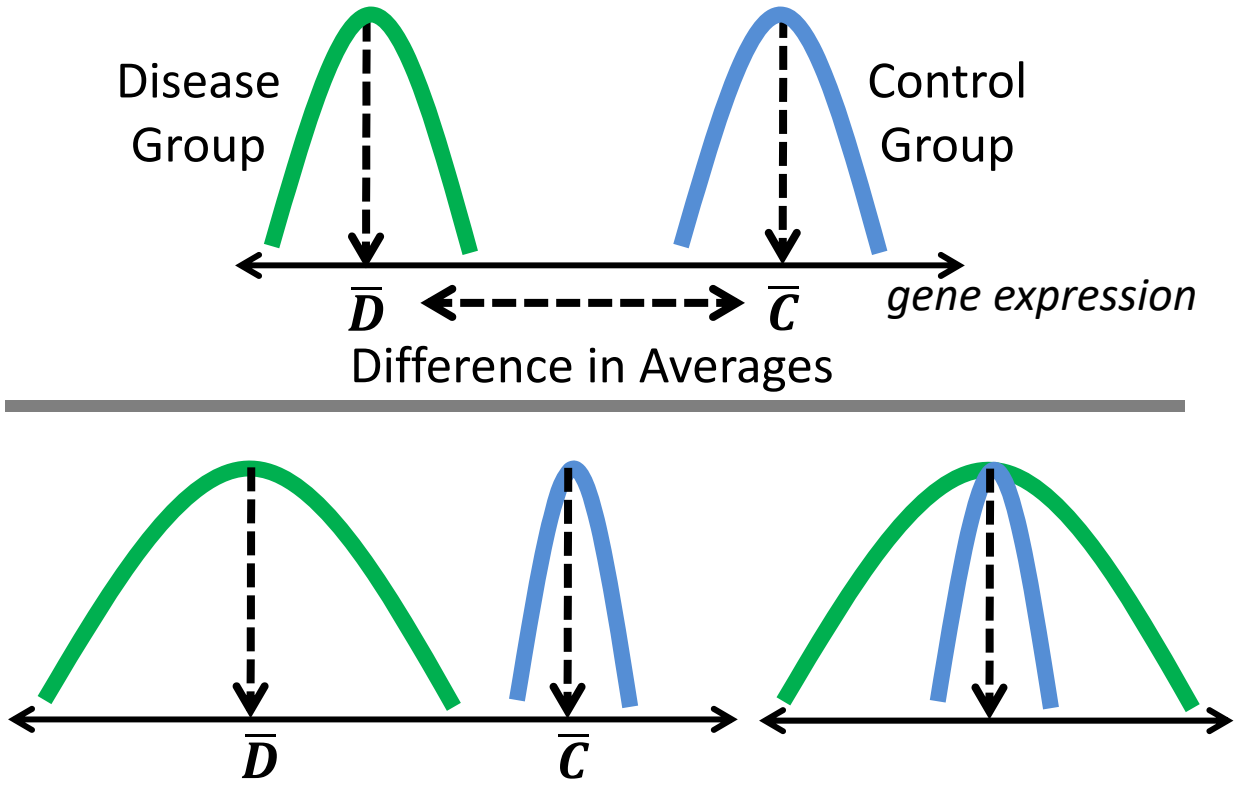


Bioinformatics methods typically focus only on the 1st moment.

Mar (2019). *Biophysical Reviews.*

Gene expression variance as a population-specific regulatory parameter

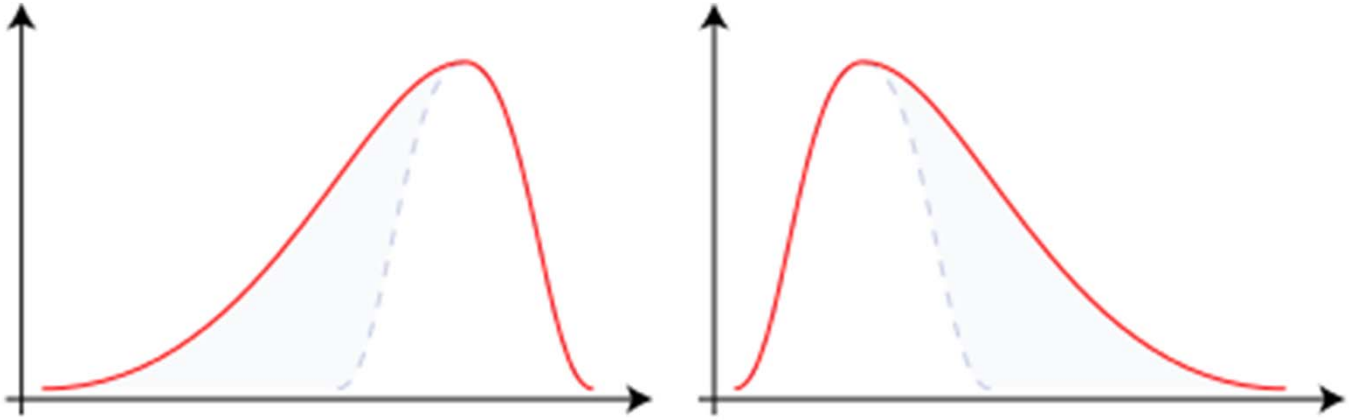
$$T_{(gene)} = \frac{\overline{D} - \overline{C}}{f(Var(D, C))}$$



Mar *et al.* (2011) **PLoS Genetics.**

Skewness as a Measure of Gene Expression Heterogeneity

Do changes in skew predict changes in biology?
Does expression skewness in genes reflect interesting differences in biology between cancer datasets?




Ben Church Henry Williams

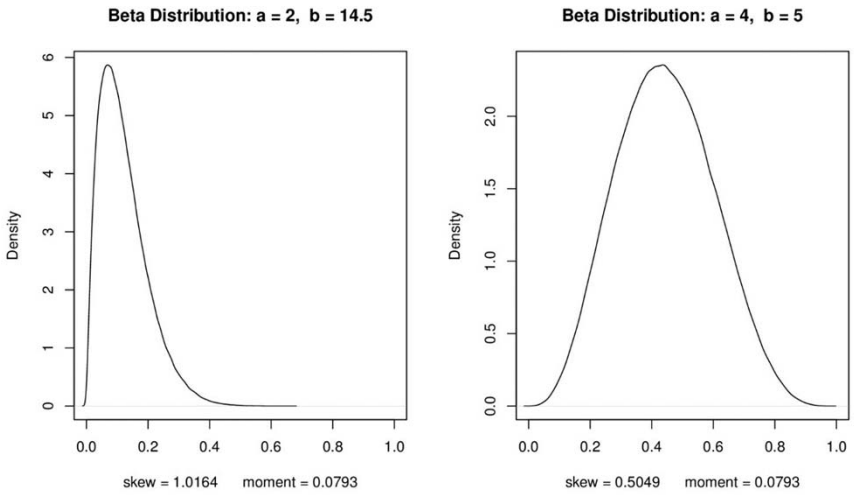
- Skewness is associated with the third statistical moment.
- It measures the degree of asymmetry in a distribution.

Measuring Gene Expression Skewness in a Cohort

For a gene's transcript expression (g) in a population |X|, the estimate of skewness is defined as :

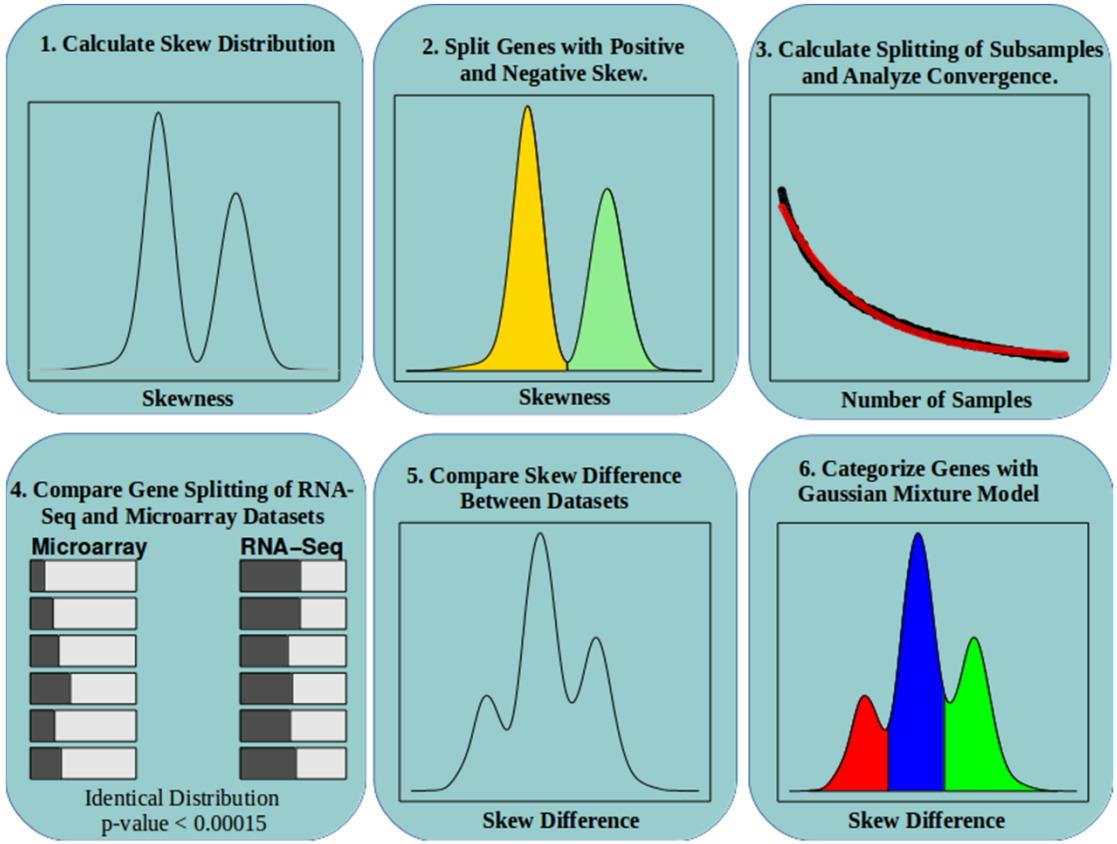
$$S_g(X) = \frac{\sqrt[3]{\frac{1}{|X|-1}\sum_{x \in X}(g_x - \mu_g)^3}}{\sigma_g}$$

- This is a biased estimator, with correction factor |X|/(|X|-2).
- However, for our data sets N ~ 500 so this correction is of order 0.2%



- This statistic was chosen to differentiate between wide slightly asymmetric distributions and narrow highly asymmetric distributions by normalizing be the standard deviation.
- (Right) the third moment cannot distinguish between these two qualitatively different distributions.

Investigating Skewness in Transcriptional Regulation of Different Tumor Types



Ben Church Henry Williams

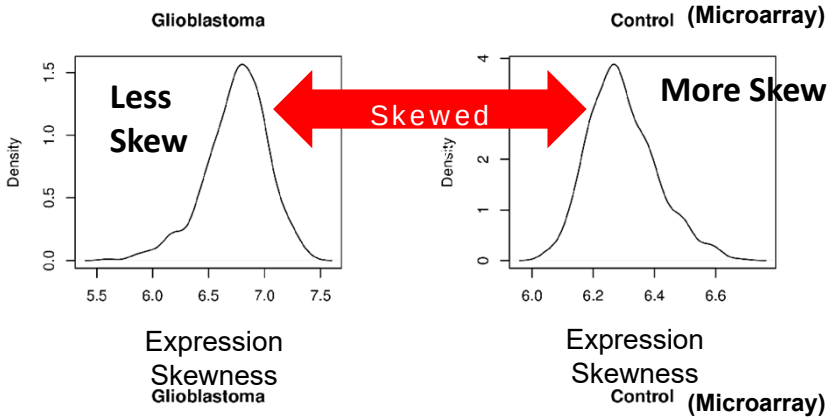
Comparing Changes in Gene Expression Skewness Between Two Different Cohorts

Microarray Data Sets

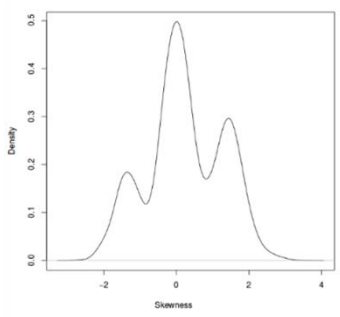
TCGA Ovarian Cancer [N = 568]
TCGA Glioblastoma [N = 548]
TCGA Breast Cancer (Luminal A) [N = 284]
AML – Over 60s [N= 461]
AML – Normal Karyotype [N = 251]
HapMap Control

RNA-Seq Data Sets

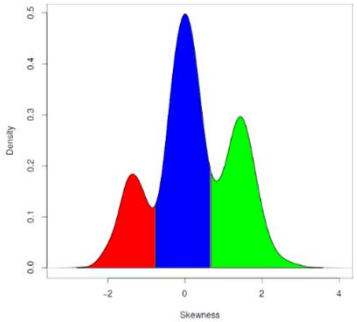
TCGA Melanoma [N = 470]
TCGA Head & Neck SC [N = 519]
TCGA Lower Grade Glioma [N = 514]
TCGA Lung Squamous Cell Carcinoma (LUSC) [N = 495]
TCGA Kidney (KIRC) [N = 531]
1000 Genomes/Geuvadis
LCLs [N = 465]



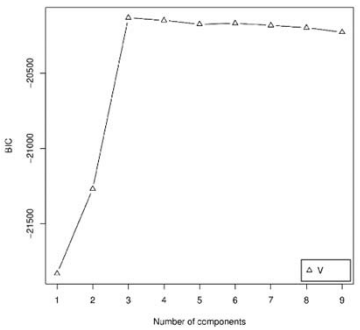
Difference in expression skewness between Glioblastoma and Control



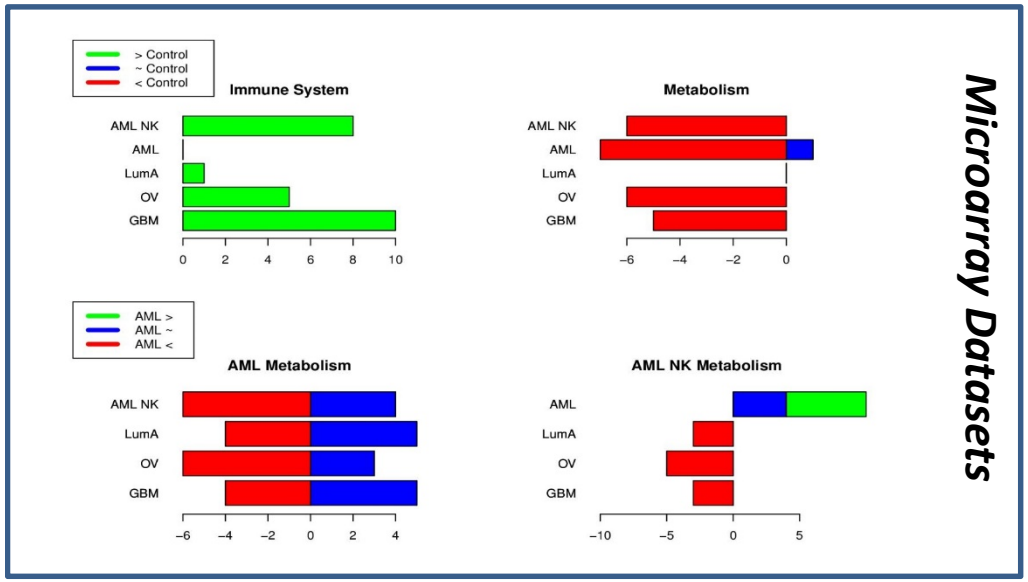
Three groups of expression skewness are evident in the comparison

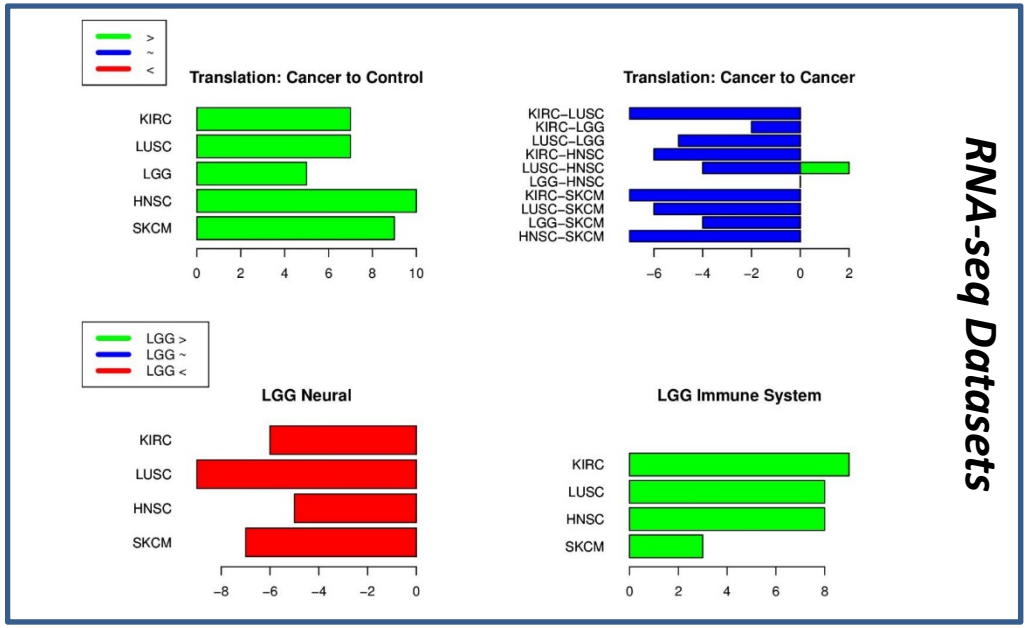


Mixture model used to identify the three groups of genes with different expression skew



Over-Representation Analysis Identified Immune-Related Pathways with Increased Skewness in Cancer versus Controls



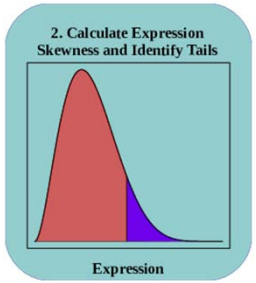


7. Assess Enrichment of Gene Function in Mixture Categories

Microarray	RNA-Seq
- Immune Processes are up-skewed in cancer relative to control	- Translational Pathways are up-skewed in cancer relative to control
- Metabolism Pathways are down-skewed in cancer relative to control	- Translational Pathways have consistent skewness across cancers
- Metabolism Pathways are down-skewed in AML relative to other cancers	- LGG has significant pathway differences with respect to other cancers

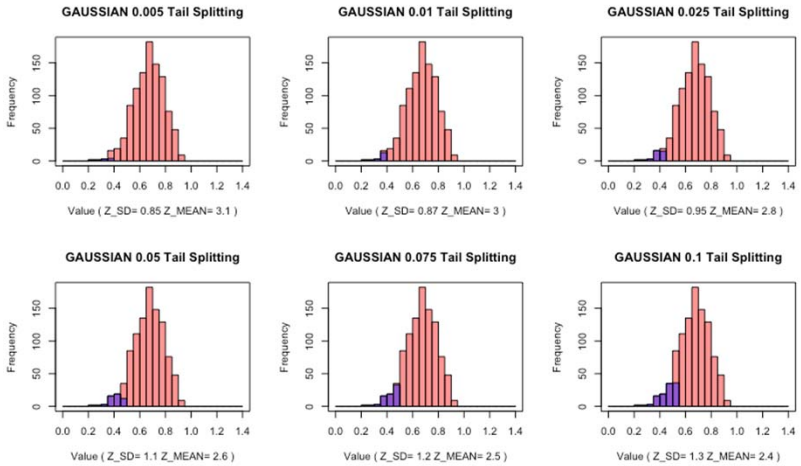
- Tissue-specific trends do exist and aren't always consistent for the different categories of pathways/functions.
- Cancer to control comparisons show greater changes in skew for translation pathways than cancer to cancer comparisons.

Identifying Patients with Extreme Expression based on Skewness

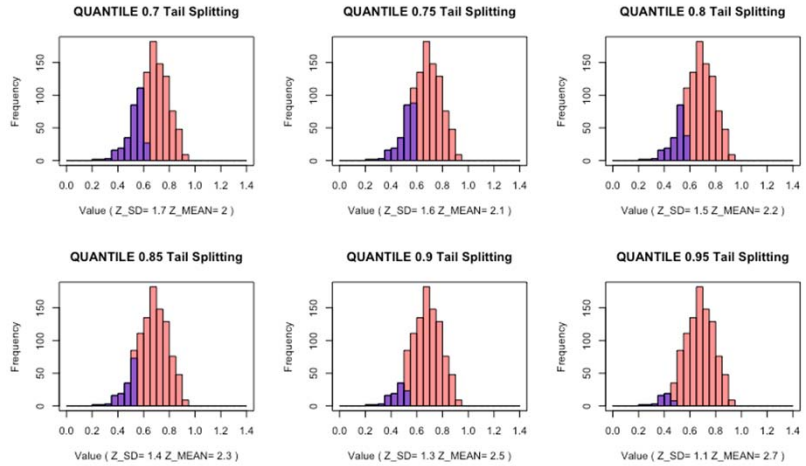


- Identifying patients that exist in tail versus non-tail regions of the gene expression distribution can be based on quantiles, or a Gaussian mixture model.
- Simulations showed these two methods produce roughly the same results.

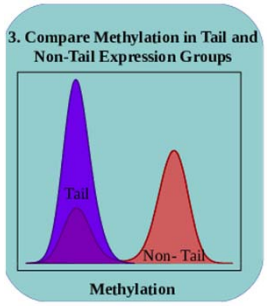
Gaussian Mixture Model



Quantile Splitting



- The top 500 significant genes with differential methylation for patients in tail versus non-tail regions of the expression distribution were identified and retained for further analysis.

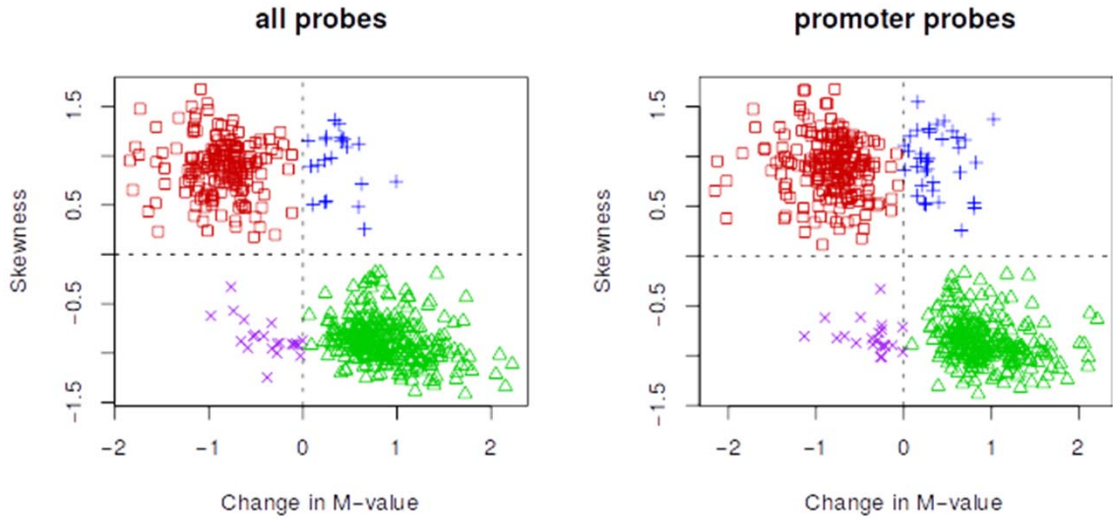


Relationship between DNA Methylation and Gene Expression Skewness

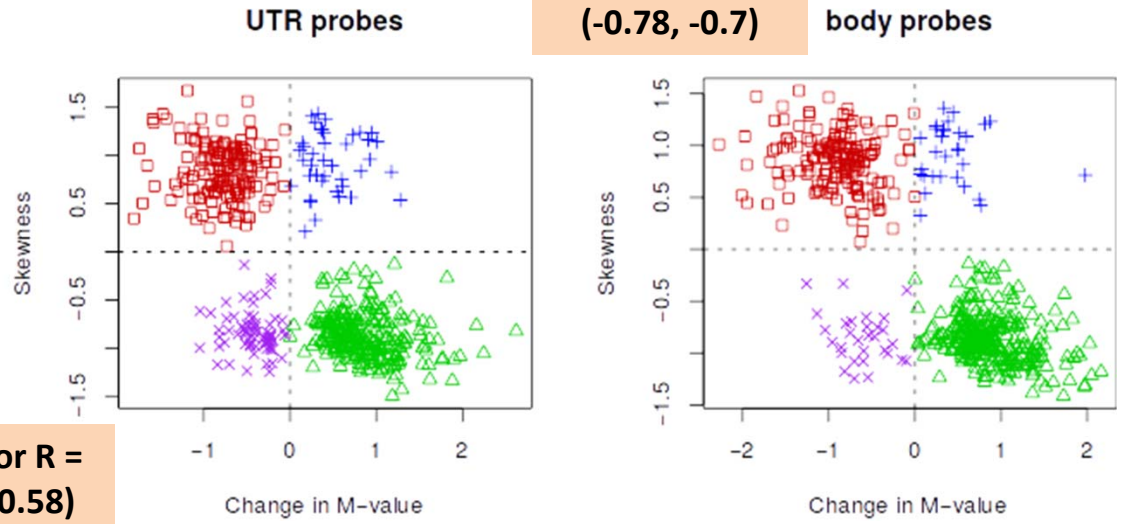
95% CI for R = (-0.82, -0.76)

- Methylation vs Skewness plots for 500 most significant genes in **TCGA-Kidney Renal Cell Carcinoma (KIRC)**.
- Data has been colored by quadrants (± skewness, ±ΔM-value).
- Distribution of genes in quadrants point to a negative association between expression skewness and DNA methylation (Fisher's exact test, P-value < 10^{-6})

R is Pearson correlation



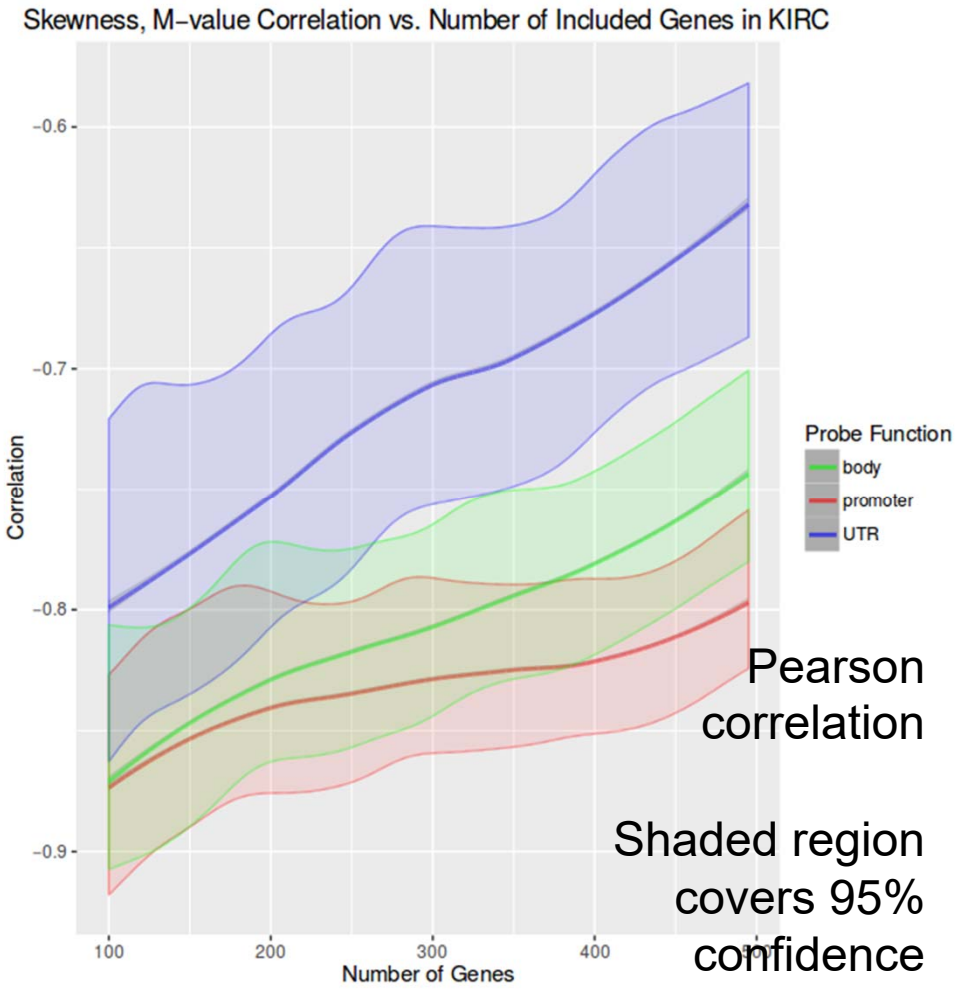
95% CI for R = (-0.78, -0.7)



95% CI for R = (-0.68, -0.58)

Negative correlation is robust to the number of significant genes

- To assess dependence of correlation results on the choice of 500 significant genes, we varied the number of significant genes.
- Negative correlation means that high methylation suppresses expression making leftward tail (lower expression).
- **Promoter probes show the greatest robustness compared to other regions & most negative correlation** – suggests a link between skewness and DNA methylation for the **top 500 genes with differential methylation**.
- The negative correlation for ***gene body*** is interesting.



Conclusions

- Gene expression skewness provides insight into understanding heterogeneity of patient cancer cohorts.
- There is a link between patients with extreme gene expression (tails of a skewed distribution) and differential promoter DNA methylation.
- These results suggest that future work on analyses that use gene expression moments beyond mean and variance may be useful.

Acknowledgements

Mar Lab @ Einstein

Dr Laurence de Torrenté (now NYGC)
Yu Hasegawa (now UC Davis)
Dr Abhi Ratnakumar (now MSKCC)
Daniel Piqué (now 3rd yr Med)
Raymund Bueno (now Cofactor Genomics)
Shuonan Chen (now Columbia)
Ameya Kulkarni
Sam Zimmerman (now Harvard)

Amazing Summer Students

Ben Church (now Columbia)
Henry Williams (now Columbia)
Marjorie Liebling (now Cornell)
Cassie Litchfield

Einstein Collaborators

Prof John Greally
Dr Masako Suzuki
Prof Nir Barzilai
Prof Yousin Suh
A/Prof Cristina Montagna

Children's Hospital Of Philadelphia/UPenn

Dr Deanne Taylor
Prof Maja Bucan
Dr Pichai Raman

Vijg Lab

Prof Jan Vijg
Moonsook Lee

FUNDING (USA)

DOD
NIH/NIGMS

Ben Church Henry Williams

EINSTEIN
Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY
Science at the heart of medicine

Acknowledgements

My Group

Dr Atefeh Taherian Fard
Othmar Korn
Ameya Kulkarni (Einstein PCI/Genetics)
Ebony Watson
Huiwen Zheng
Malindrie Dharmaratne
Sas Logan
Stephanie Kemp
Yuyang Wei
Jon Xu
Brad Balderson
Ariane Mora

Prof Alan Rowan (AIBN)
Prof Ernst Wolvetang (AIBN)
Prof Christine Wells (Uni Melbourne)
Prof Louise Ryan (UTS)

FUNDING (Australia)

National Stem Cell Foundation of Australia

Australian Government
Australian Research Council

Australian Government
National Health and Medical Research Council
NHMRC

Can I Please Have a More Specific Title? (It Can Be A Bit Fun)

Skewness as a Measure of Gene Expression

Outliers in the Balance: Uncovering the "Hidden Measure" of Biostatistics

Wait Just a Moment! A Consideration of Skewness in Biostatistics

Secret Biological Insights of the Third Moment

On the Bleeding Edge: What can a Statistical Study of Outliers Teach Biology?

What's Love Got to Skew with It?

The SKEW Files: What They Don't Want you to Know about Statistics

A Skew Paradigm in Biostatistics

Third is the Skew Moment

Orange is the Skew Black

Statistics Wars IV: A Skew Hope

A Whole Skew World

Gangs of Skew York

How Do You Skew?

Edge Cases: A Consideration of Skewness in Biology

Dr. Third Moment or How I Learned to Stop Worrying and Love the Skew