THE UNIVERSITY OF ARIZONA
Mel & Enid Zuckerman
College of Public Health

# Problems with Correlations in Relative Data and a Proposed Alternative

Dominic LaRoche

November 3, 2015

THE UNIVERSITY OF ARIZONA
Mel & Enid Zuckerman
College of Public Health

# The Papers

**Proportionality: A Valid Alternative to Correlation for Relative Data**
David Lovell, Vera Pawlowsky-Glahn, Juan Jos Egozcue, Samuel Marguerat, and Jrg Bhler
PLoS Computational Biology 11(3) (2015)

**Proportions, Percentages, PPM: Do the Molecular Biosciences Treat Compositional Data Right?**
Lovell DR, Muller W, Taylor JM, Zwart AB, Helliwell CA
Compositional data analysis: Theory and applications p193-207 (2011)

# Section 1

## Introduction to the Problem

# What is Relative Data? –Also called "Compositional Data"

- Compositional data are vectors of non-negative components showing the *relative* weight or importance of a set of *parts in a total*
- The total sum of a compositional vector is considered irrelevant, or an artifact of the sampling procedure.
- No individual component can be interpreted isolated from the other. A composition carries no absolute information on increment/decrement of mass.

# Is *my* data relative??

Relative data arises naturally in many biological measurements:

- Is your sample a fixed size?
  - 1 gram of tissue
  - 1 $\mu$g of total RNA
  - 1 $\mu$g of meta-genomic DNA
  - 1 mL of blood, etc.

# Is *my* data relative??

Relative data arises naturally in many biological measurements:

- Is your sample a fixed size?
  - 1 gram of tissue
  - 1 $\mu$g of total RNA
  - 1 $\mu$g of meta-genomic DNA
  - 1 mL of blood, etc.
- Is your data a constrained count?
  - How many total reads can your favorite platform handle?
  - Counts of codons or bases in a fixed length of DNA

# Is *my* data relative??

Relative data arises naturally in many biological measurements:

- Is your sample a fixed size?
  - 1 gram of tissue
  - 1 $\mu$g of total RNA
  - 1 $\mu$g of meta-genomic DNA
  - 1 mL of blood, etc.
- Is your data a constrained count?
  - How many total reads can your favorite platform handle?
  - Counts of codons or bases in a fixed length of DNA
- Is your data based on proportions?
  - Different k-mers in genomes
  - GO terms in samples
  - different reads in NGS sequencing runs

# Is *my* data relative??

Relative data arises naturally in many biological measurements:

- Is your sample a fixed size?
  - 1 gram of tissue
  - 1 $\mu$g of total RNA
  - 1 $\mu$g of meta-genomic DNA
  - 1 mL of blood, etc.
- Is your data a constrained count?
  - How many total reads can your favorite platform handle?
  - Counts of codons or bases in a fixed length of DNA
- Is your data based on proportions?
  - Different k-mers in genomes
  - GO terms in samples
  - different reads in NGS sequencing runs

**Then your data might be relative!**

# Should I *care* if my data is relative??

# Should I *care* if my data is relative??

**Yes**

# Should I *care* if my data is relative??

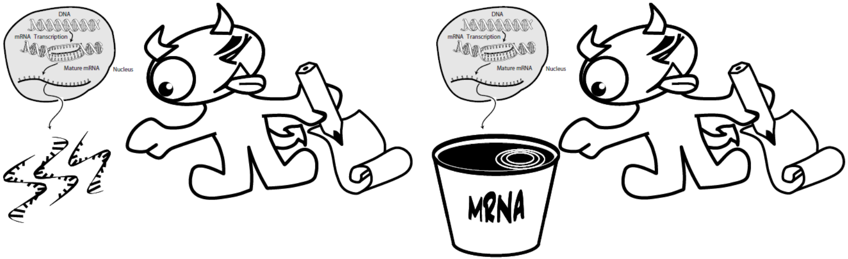
**Yes**

Long answer: It depends...

In certain cases it doesn't matter much but in others it matters a lot.

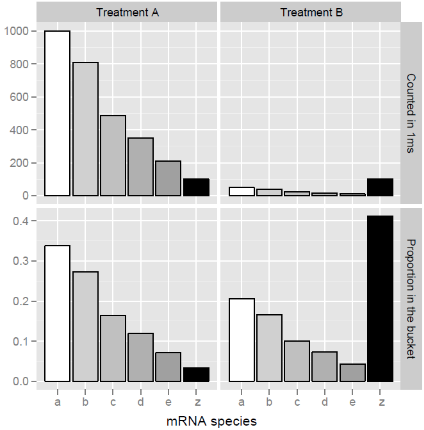
# The ‘Omics Imp’

Lovell et al.

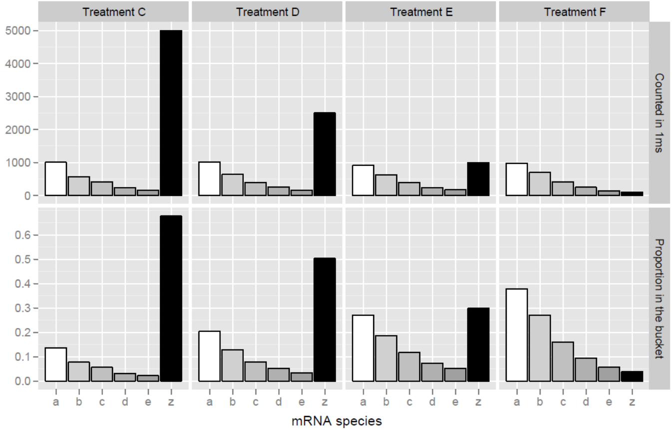



- On the left the imp tallies sequences as they are produced in a fixed time period
- On the right the imp counts the sequences in some fixed size bucket
    - Data on the right are parts of a total

# Relative data can be misleading



# Relative data can be misleading



# When is relative data likely to be misleading?

Take a 3 component example: $Total = c_1 + c_2 + c_3$

# When is relative data likely to be misleading?

Take a 3 component example: $Total = c_1 + c_2 + c_3$

- $c_1, c_2 \gg c_3$
  - As $c_1 \uparrow$ then $c_2 \downarrow$
  - Correlation is attenuated

# When is relative data likely to be misleading?

Take a 3 component example: $Total = c_1 + c_2 + c_3$

- $c_1, c_2 \gg c_3$
  - As $c_1 \uparrow$ then $c_2 \downarrow$
  - Correlation is attenuated
- $c_3 \gg c_1, c_2$
  - Here $var(c_3)$ dominates the composition
  - Correlation is biased high as $var(c_3) \uparrow$

# David Lovell's Take

Yes, I am going to show a slide show during a slide show.

David Lovell scaring you about correlating relative data

# A little background before the proposed alternative: Model 2 regression

Also known as Standardized Major Axis (SMA) Regression

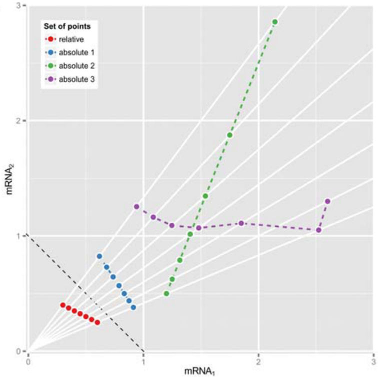
- Traditional regression has 1 random variable: $Y = \beta_0 + \beta X + \epsilon$
    - $X$ is considered "fixed" so has no random error ($\epsilon$)

# A little background before the proposed alternative: Model 2 regression

Also known as Standardized Major Axis (SMA) Regression

- Traditional regression has 1 random variable: $Y = \beta_0 + \beta X + \epsilon$
  - $X$ is considered "fixed" so has no random error ($\epsilon$)
- SMA regression gives the relationship between to random variables
  - Accounts for the random error in both variables
  - Slope estimate: $\beta = \frac{sd(Y_1)}{sd(Y_2)}$

# Correlations on Relative Data




Correlations on relative data tell us absolutely nothing about the relationship between the absolute abundances.

# The proposed alternative to correlation

The authors propose "proportionality", $\phi$, as a substitute for correlation.

- They start with Aitchison's log-ratio variance:

$$var(log(x/y)) = var(logx - logy)$$

# The proposed alternative to correlation

The authors propose “proportionality”, $\phi$, as a substitute for correlation.

- They start with Aitchison's log-ratio variance:

$$var(log(x/y)) = var(logx - logy)$$

- They factor log-ratio variance using the properties of variance:

$$= var(log(x)) \left(1 + \frac{var(log(y))}{var(log(x))} - 2\sqrt{\frac{var(log(y))}{var(log(x))}} \frac{cov(log(x), log(y))}{\sqrt{var(log(x))var(log(y))}}\right)$$

# The proposed alternative to correlation

The authors propose “proportionality”, $\phi$, as a substitute for correlation.

- They start with Aitchison’s log-ratio variance:

$$var(log(x/y)) = var(logx - logy)$$

- They factor log-ratio variance using the properties of variance:

$$= var(log(x))\left(1 + \frac{var(log(y))}{var(log(x))} - 2\sqrt{\frac{var(log(y))}{var(log(x))}}\frac{cov(log(x), log(y))}{\sqrt{var(log(x))var(log(y))}}\right)$$

- The identify the SMA slope estimate $\beta$ and the correlation:

$$= var(log(x))(1 + \beta^2 - 2\beta|r|)$$

# The proposed alternative to correlation

The authors propose "proportionality", $\phi$, as a substitute for correlation.

- They start with Aitchison's log-ratio variance:

$$var(log(x/y)) = var(logx - logy)$$

- They factor log-ratio variance using the properties of variance:

$$= var(log(x))\left(1 + \frac{var(log(y))}{var(log(x))} - 2\sqrt{\frac{var(log(y))}{var(log(x))}}\frac{cov(log(x), log(y))}{\sqrt{var(log(x))var(log(y))}}\right)$$

- The identify the SMA slope estimate $\beta$ and the correlation:

$$= var(log(x))(1 + \beta^2 - 2\beta|r|)$$

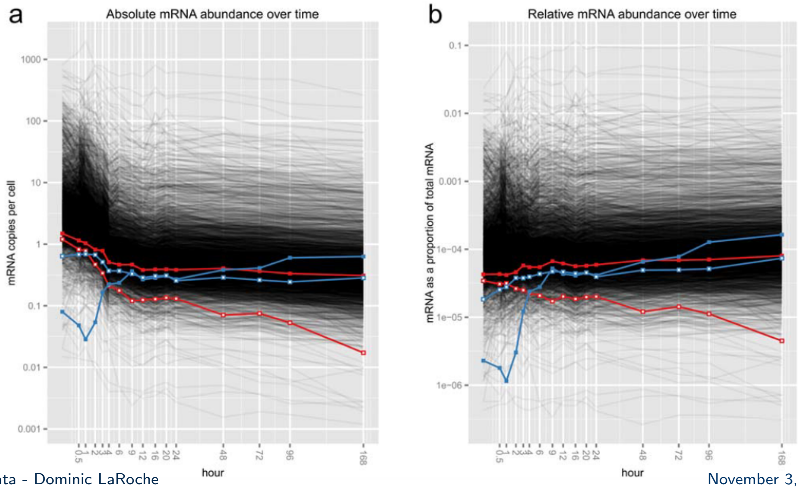
- They drop the unnecessary term to get $\phi$:

$$\phi(log(x), log(y)) = (1 + \beta^2 - 2\beta|r|)$$

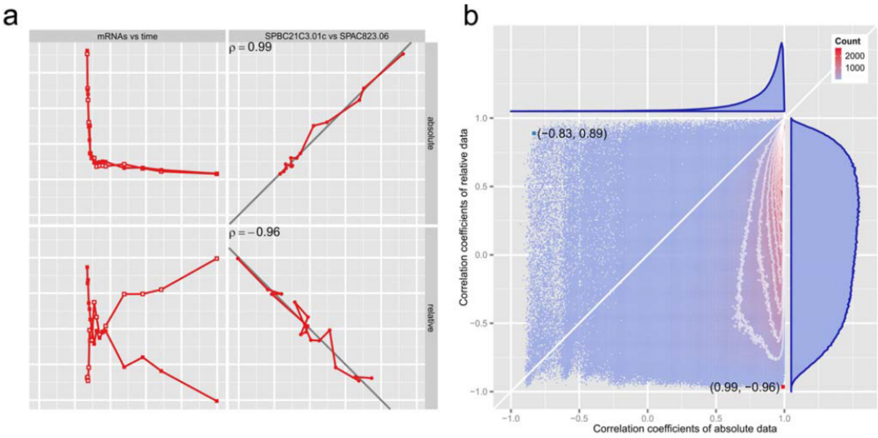
# Benefits of proportionality

- Derived from Aitchison's log-ratio variance
- Composed of two established metrics of association
- However, $\phi$ is not symmetric like $\rho$

# Yeast Example



# Yeast Example



# When can I mostly ignore the relative nature of my data?

Relative data aren't always a problem:

- Components of interest are relatively small parts of mixture samples that remain constant in size
- Only using univariate statistics (e.g. variance)
- log-transformation can *help* (due to the properties of the log)

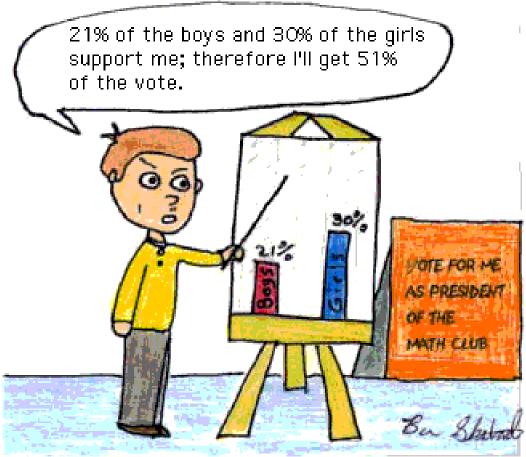
# When can I mostly ignore the relative nature of my data?

Relative data aren't always a problem:

- Components of interest are relatively small parts of mixture samples that remain constant in size
- Only using univariate statistics (e.g. variance)
- log-transformation can *help* (due to the properties of the log)

Do you feel lucky? -*Dirty Harry*

# Questions?


21% of the boys and 30% of the girls support me; therefore I'll get 51% of the vote.
21%
Boys
30%
Girls
VOTE FOR ME AS PRESIDENT OF THE MATH CLUB