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

Problems with Correlations in Relative Data and a Proposed Alternative

Dominic LaRoche

November 2, 2015

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 oa fixed size?
  - 1 gram of tissue
  - 1 μg of total RNA
  - 1 μg of metagenomic DNA
  - 1 mL of blood, etc.

Is *my* data relative??

Relative data arises naturally in many biological measurements:

- Is your sample oa fixed size?
  - 1 gram of tissue
  - 1 μg of total RNA
  - 1 μg of metagenomic 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 oa fixed size?
  - 1 gram of tissue
  - 1 μg of total RNA
  - 1 μg of metagenomic 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 oa fixed size?
  - 1 gram of tissue
  - 1 μg of total RNA
  - 1 μg of metagenomic 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??

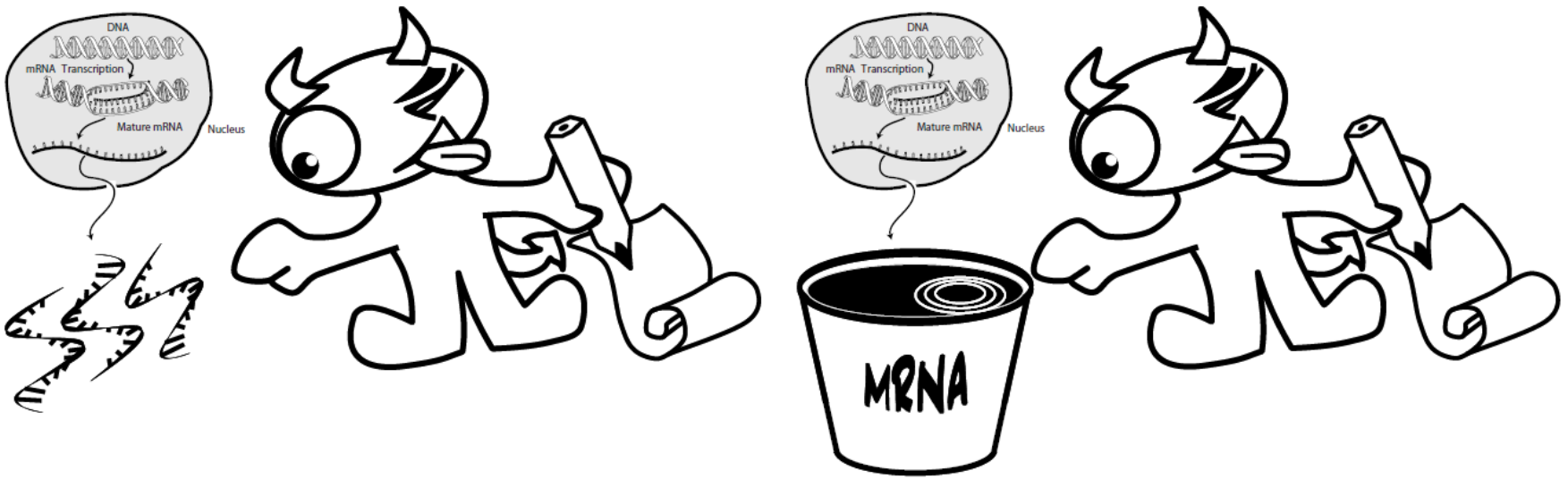
Yes

Long answer: It depends...

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

The 'Omics Imp'

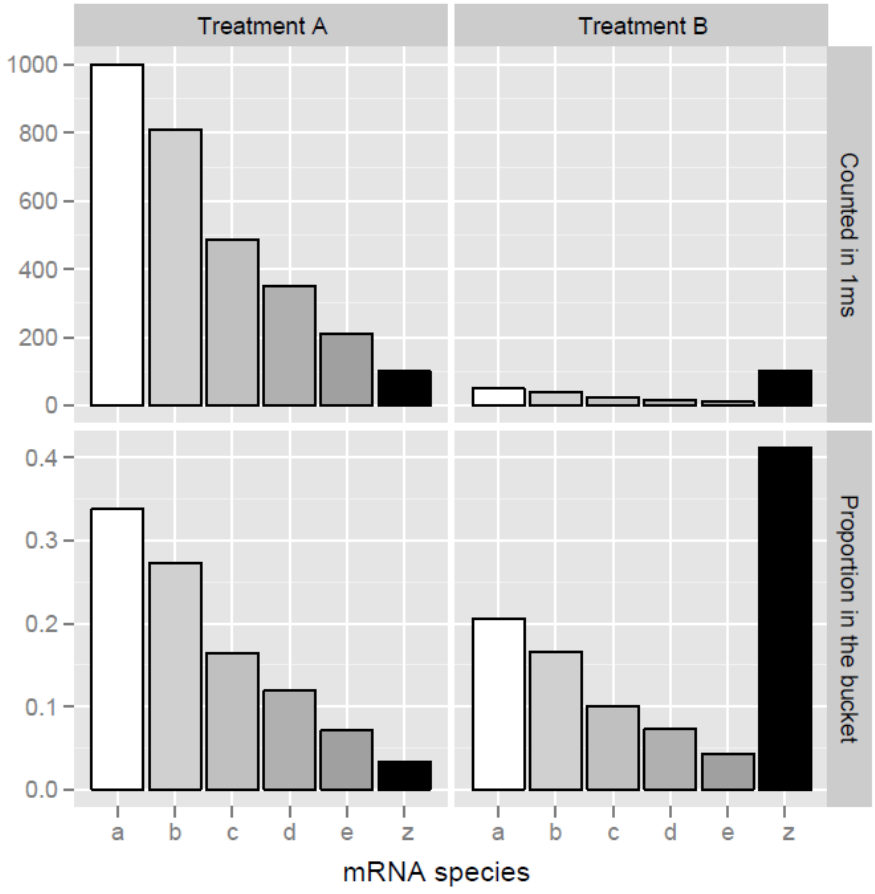
Lovell et al.



- 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

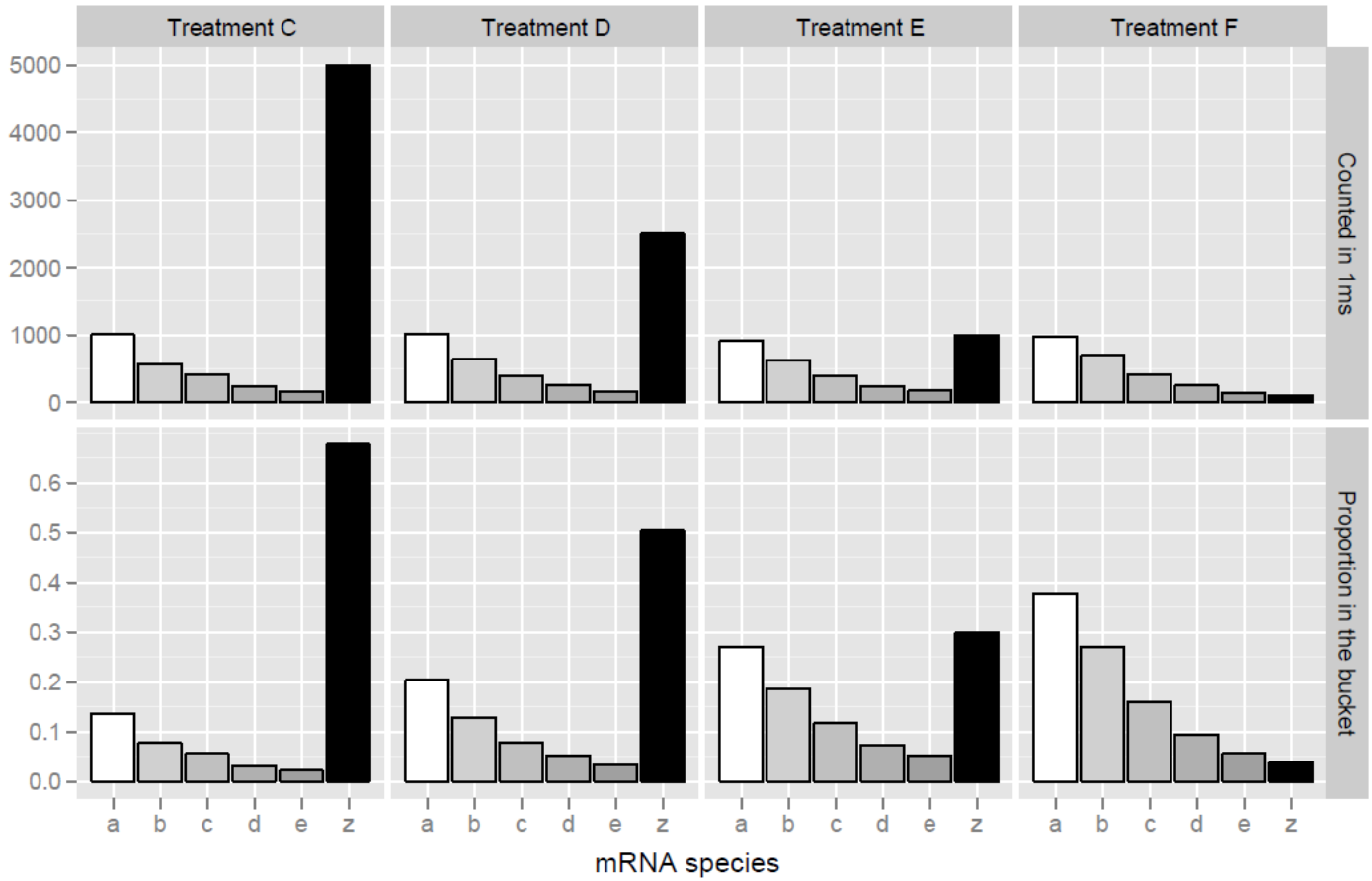
Relative data can be misleading

Example 1.



Relative data can be misleading

Example 2.



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 $var\,(c_3) \uparrow$

Correlations on Relative Data

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

David Lovell scaring you about correlating proportions

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