CoDaSeq: Analyzing HTS using compositional data analysis

Greg Gloor

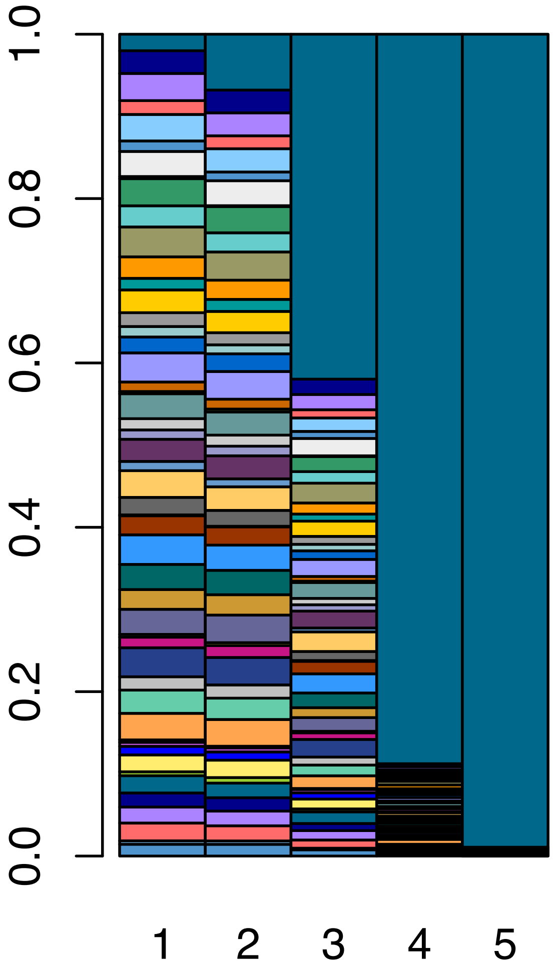
Department of Biochemistry

University of Western Ontario

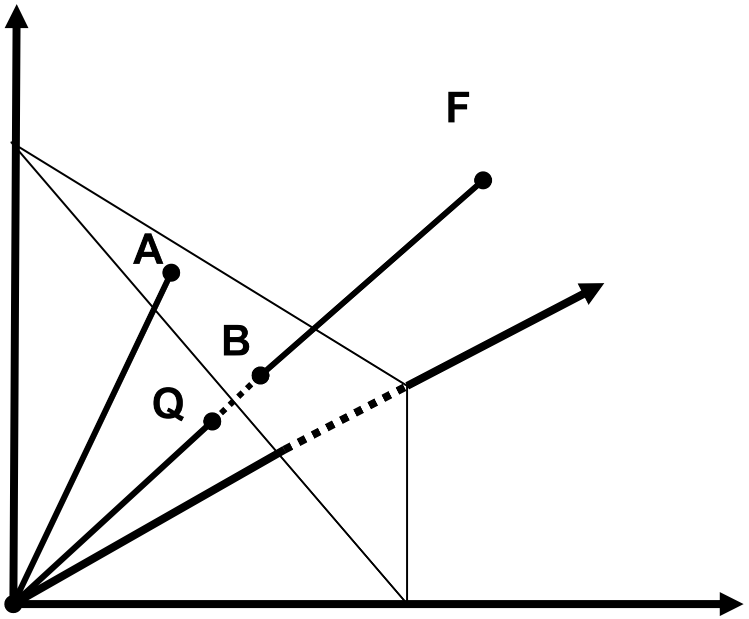
ggloor.github.io

GLBio 2016

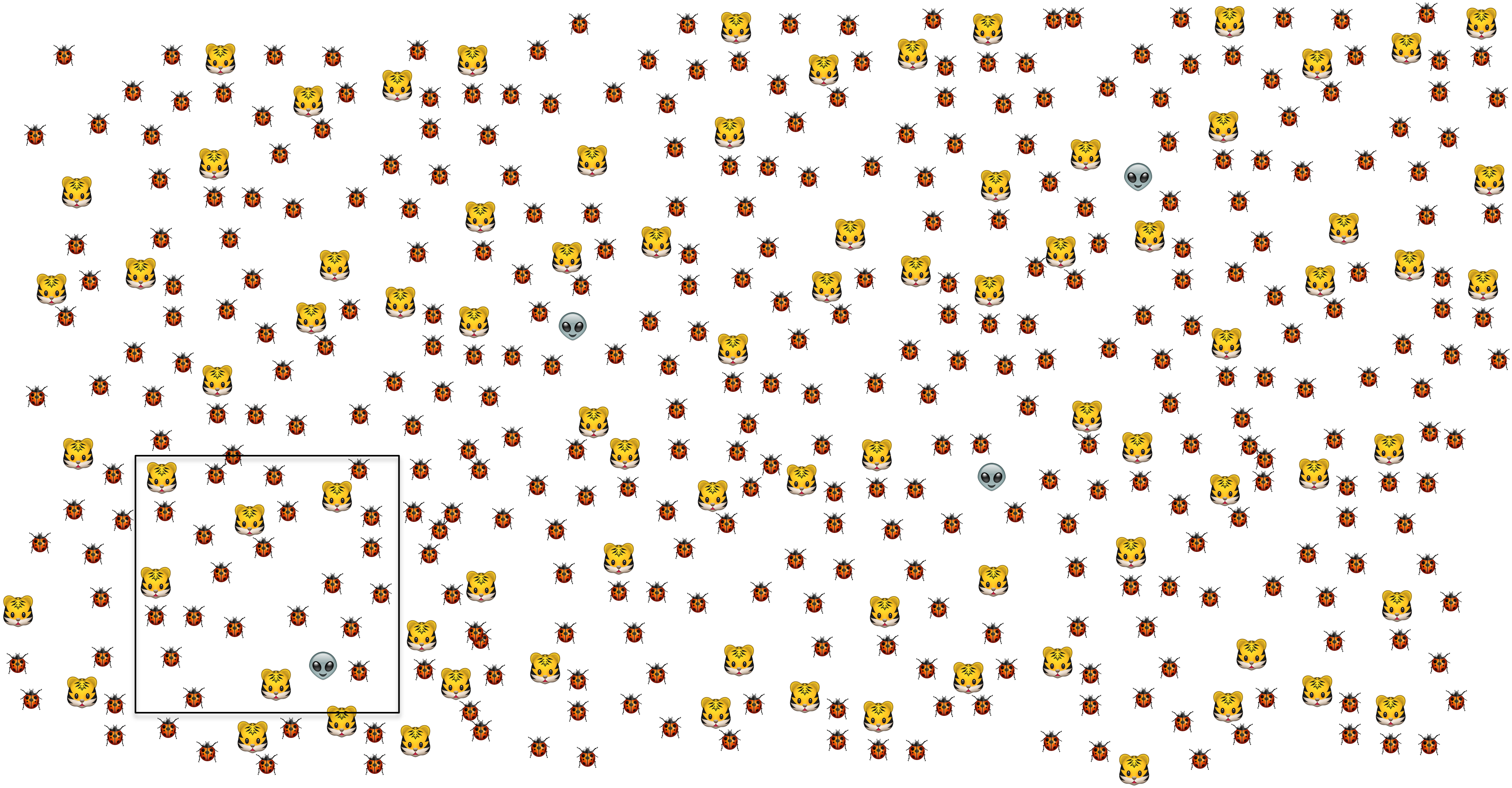
Geometry is key



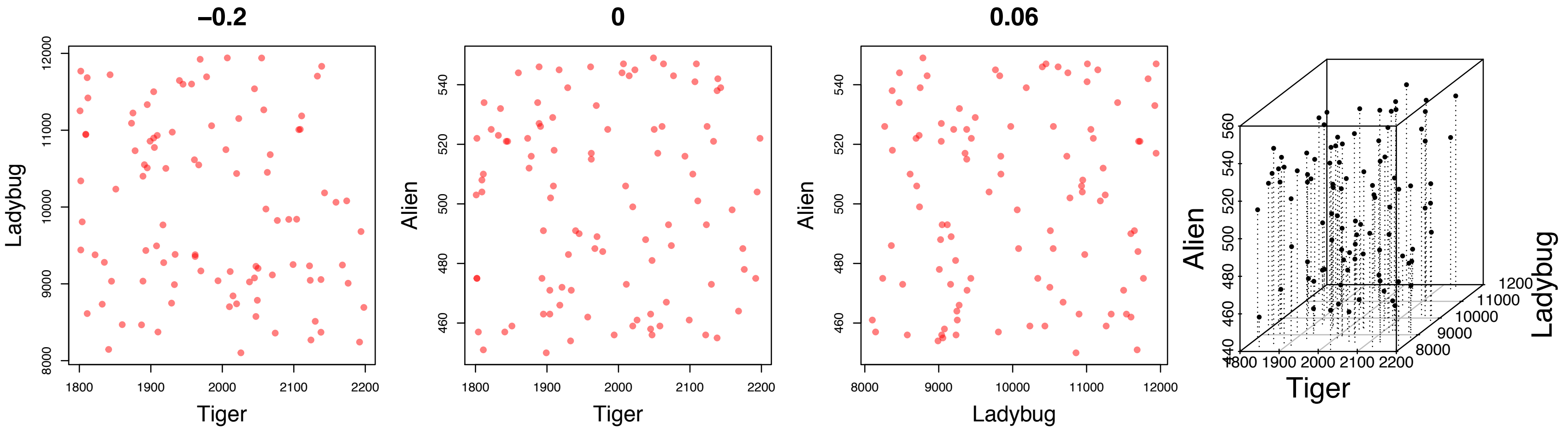
≠



Random sample of environment



Counting our things



🐯
🐞
👽

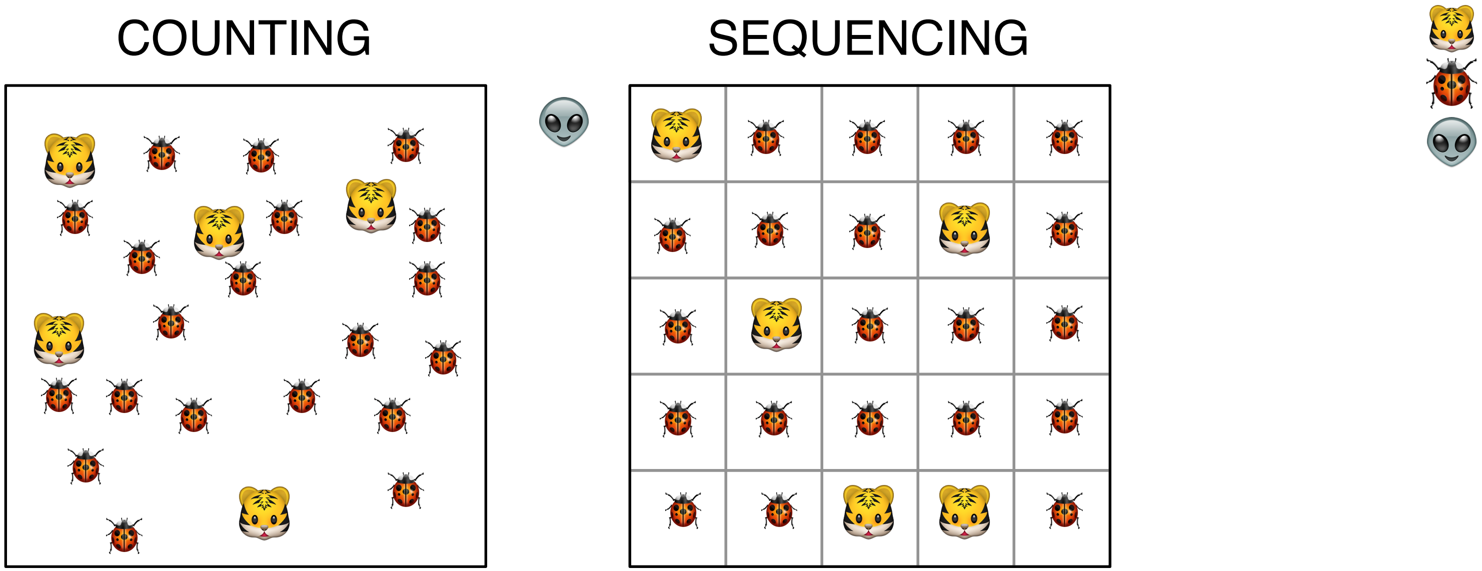
- <u>**Example 100 random sample sets**</u>

🐯 range=1800-2200

🐞 range= 8000-12000

👽 range=450-550

HTS is not counting



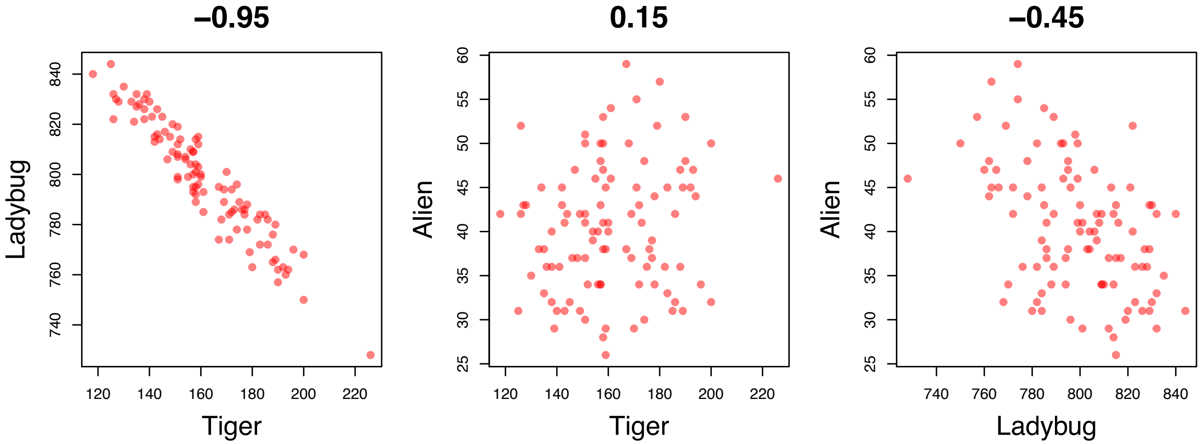
- Sequencing is a constant-sum operation
    - We only get the number of reads that the machine can deliver

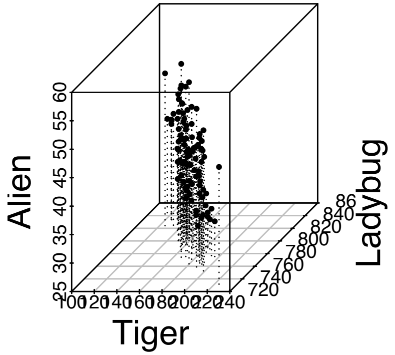
- Any constant sum is equivalent

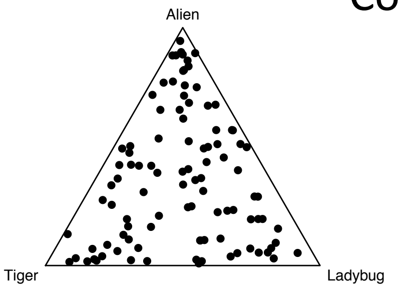
Gloor, et al. 2016. Ann Epidemiology
Gloor, et al. 2016. Can J. Micro

Effect of a constant sum?

🐯
🐞
👽







Constant sum of 1000

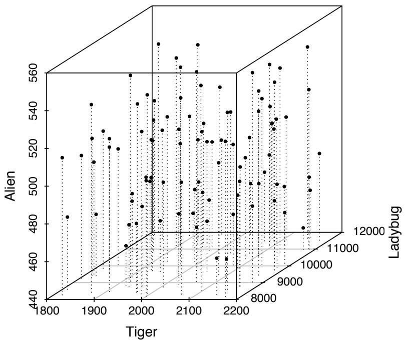
Constant sum operations:

- Count normalization
- Rarefaction
- Proportion
- percentage, relative abundance
- RNA-seq, metagenomics, tag-sequencing

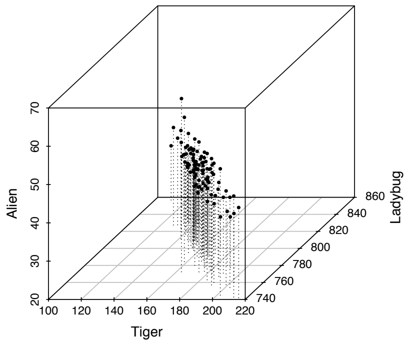
Gloor, et al. 2016. Ann Epidemiology
Gloor, et al. 2016. Can J. Micro

We have CoDa

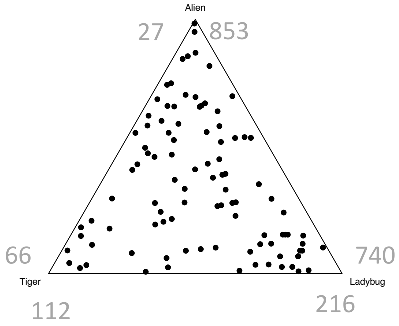
Count



Sequencing



Simplex



🐯
🐞
👽

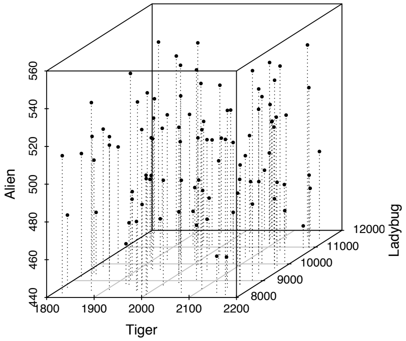
- Working on a Simplex
  - Addition and subtraction are not useful operations
  - Subsetting and aggregating are problematic (Subcompositions)
  - Correlation and covariation are unreliable
  - Scale dependence (Scale invariance)
  - Sparse data becomes an issue
  - Measurement error greatest at low count margin
- Problem remains regardless of dimension
  - It just 'looks' OK

Aitchison 1986. Stat. Anal. Comp Data
Pawlowsky-Glahn, 2015. Mod. Anal. CoDa

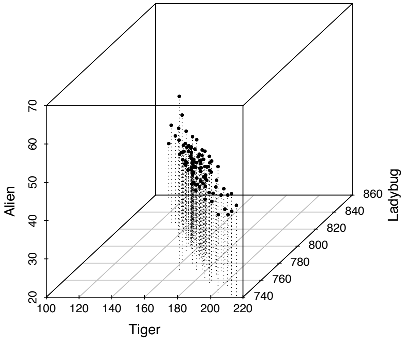
Only have ratio information

🐯
🐞
👽

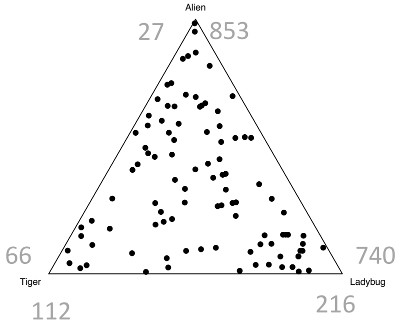
Count



Sequencing



Simplex



$$X = [\ x_1, x_2, \ldots x_D\ ],\ g_X = \text{geometric mean of } X$$

$$\mathbf{clr(x) = [\ \log(x_1/g_X), \log(x_2/g_X), \ldots \log(x_D/g_X)\ ]}$$

- Measurements are converted to ratios between parts
  - Abundance is not directly represented in the output
  - Values are now unconstrained
- The clr correction is scale invariant
- Must delete, estimate or replace 0 values

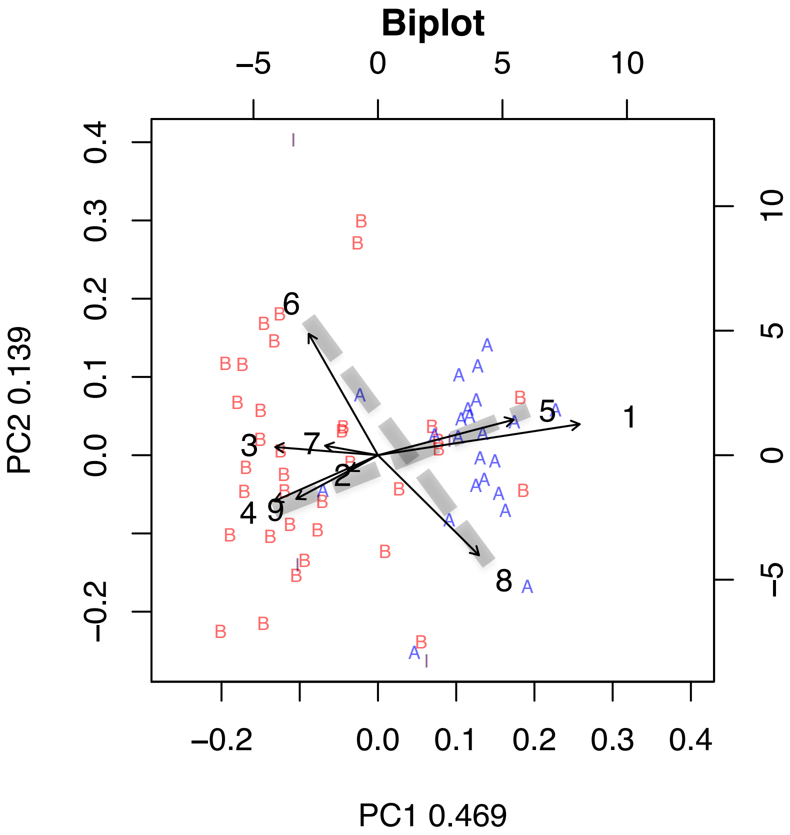
Aitchison 1986. Stat. Anal. Comp Data
Pawlowsky-Glahn, 2015. Mod. Anal. CoDa

Analysis tools based on **variance of the ratios** between parts

🐯
🐞
👽

- Compositional 0 replacement strategies
  - Prior to clr transformation
  - Best approaches are Bayesian but an open problem
- Outliers
- Exploratory data analysis
- Differential abundance
- Compositional association

Exploration: CoDa PCA biplot



■ OTU
🟦 A
🟥 B

Data Structure

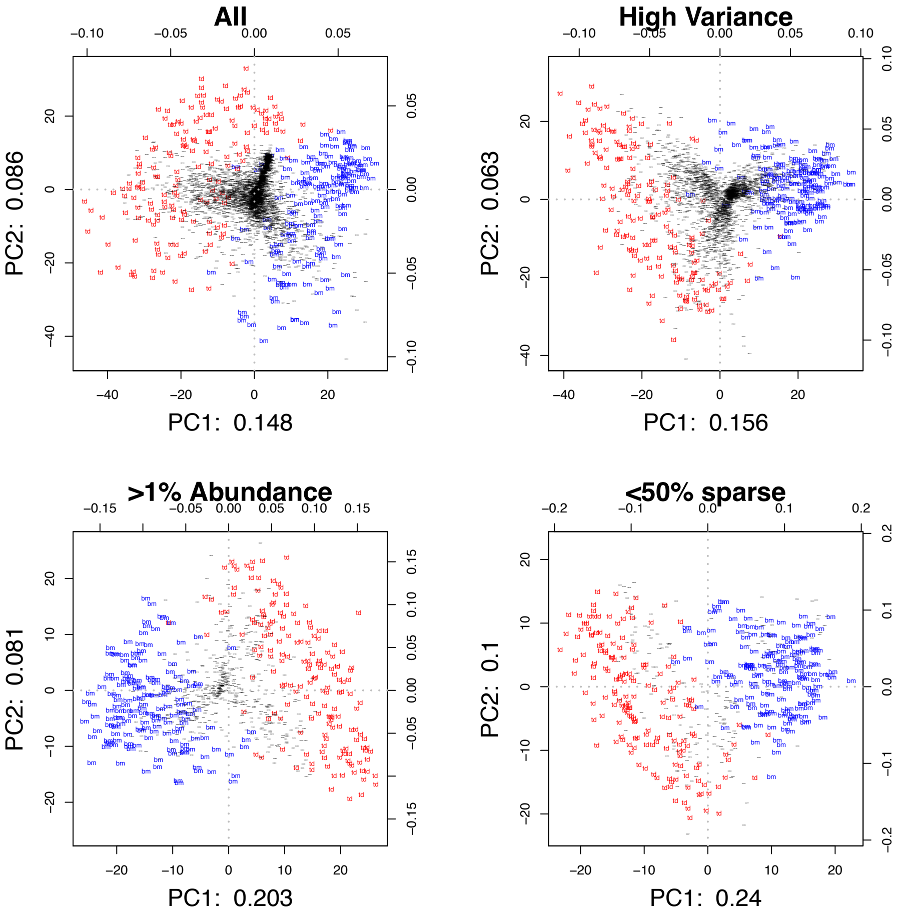
🐯
🐞
👽

- Exploratory Tool
- Samples + Variables after clr
- SVD is legal
  - But PCA is interpreted by ratios

1. Distance from origin ~ SD
2. Links ~ ratio abundance
3. Links with multiple tips ~ linear ratio dependence
4. Orthogonal links means ratios of parts are not related

Gloor, et al. 2016. Ann Epidemiology
Gloor, et al. 2016. Can J. Micro

Generally robust to filtering



⬛ OTU

🟦 BM

🟥 TD

HMP dataset, BM vs. TD

4776 OTUs in 366 samples

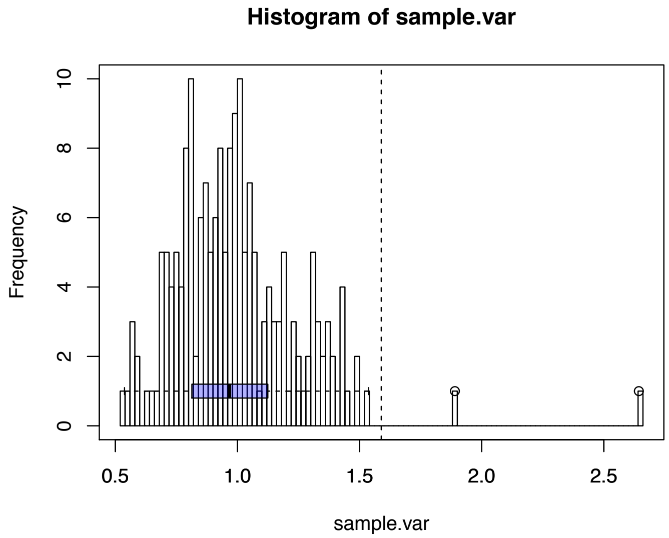
187 tongue, 179 cheek

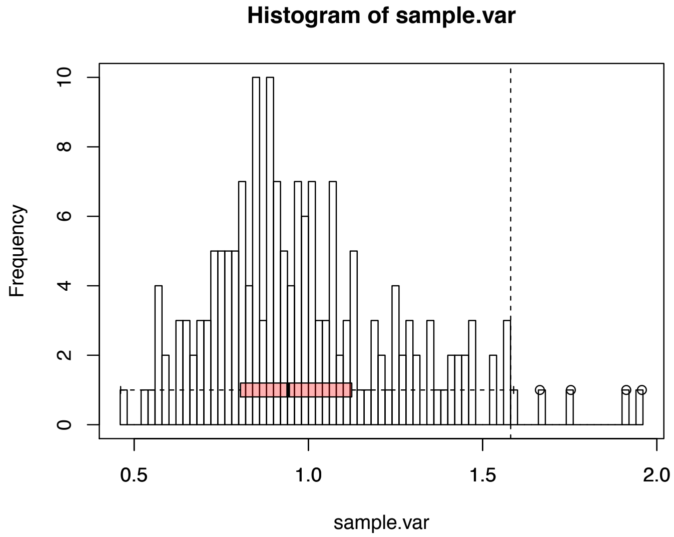
Filtering to remove rare or sparse variables is common

Variance ratios between remaining taxa are constant across filtering methds

Subcompositionally coherent

Outliers



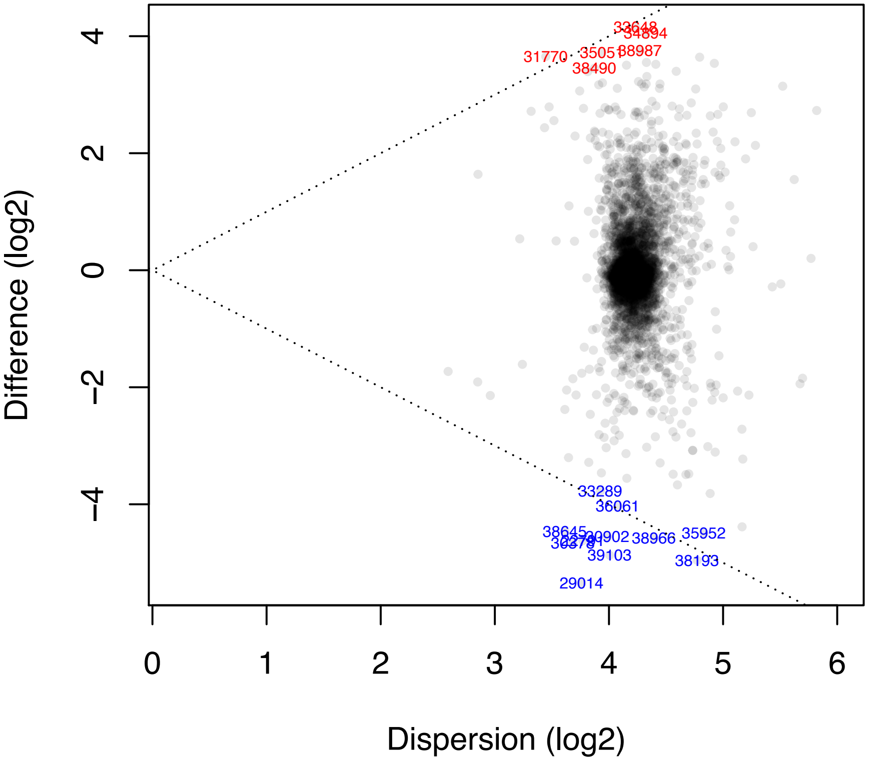


🐯
🐞
👽

- Similar to method developed by Barton lab for RNA-seq
    - (Schurch, RNA 2015)
- Samples that contribute > median + 2*IQR defined as outliers

- Generally best to discard outliers

Pairwise difference by effect



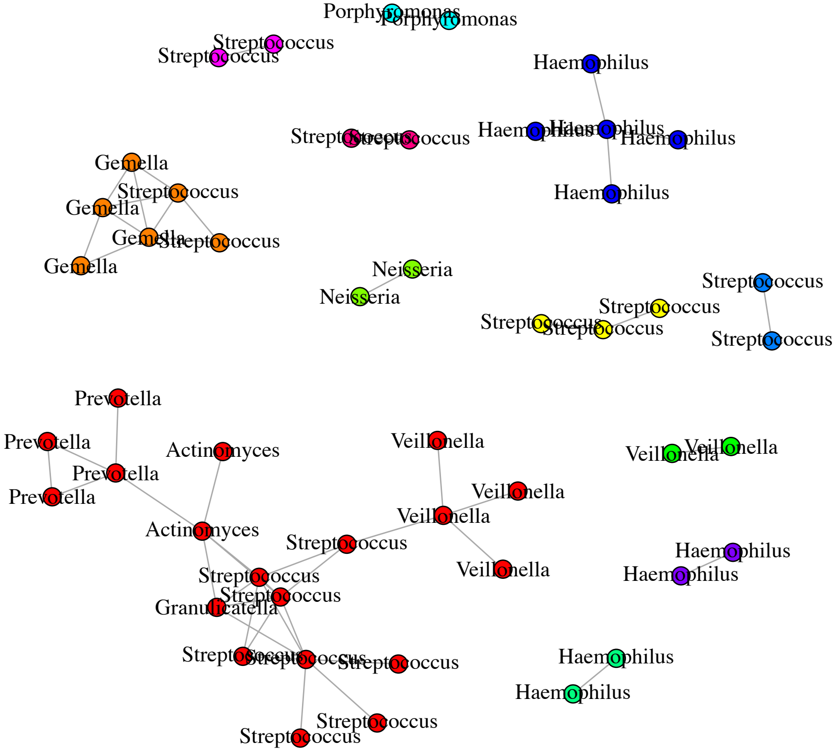
Effect plot: ALDEx2

Effect = Difference / Dispersion

- Bayesian estimate of clr values by Monte-Carlo sampling
  - Identifies OTUs where the difference between groups is robust to inferred technical replication
- Most values are seen not to be different between groups and so are non-discriminatory
- OTUs with an Abs(effect) > 0.8 colored by group

Fernandes, et al. 2013. PLoS ONE
Fernandes, et al. 2014. Microbiome
Gloor et al. 2016. Aus. J. Statistics
Gloor et al. 2016. J. Comp. Graph. Stat.

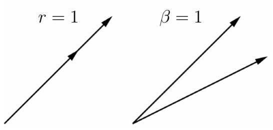
Association using Ø metric



🐯 🐞 👽

Every measure of correlation is affected by CoDa

“in the absence of any other information or assumptions, correlation of relative abundances is just wrong”



$Ø(x_{ij})$ = 1 + ß2 -2ßr =

(var(clr x_i) – var(clr x_j)) /(var(clr x_i)+ var(clr x_j))

- If variances are equal, Ø(x,y) = 0
- Measures constancy of proportion of OTUs across samples
- Enforces proper interpretation of associations

Lovell et al, PLoS Comp. Bio. 2015

Lovell, et al. 2015. PLoS Comp Bio
Friedrich & Alm 2012. PLoS Comp Bio

CoDaSeq

- Tools to analyze data in correct geometry 🐯🐞👽

  (16S rRNA geneseq, RNA-seq, metagenomics, ChIP-seq, SELEX, etc)

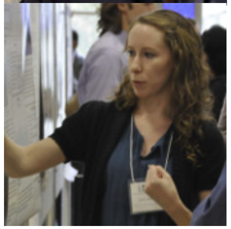
  (Jean Macklaim, Metagenomics Talks)

  – Data filtering
  – Outlier detection
  – Exploratory Data Analysis
  – Differential Abundance
  – Association and Correlation
- To be available on Bioconductor
  – Progress at ggloor.github.io

Acknowledgments

Canadian Centre for Human Microbiome and Probiotic Research

Andrew Fernandes



Jean Macklaim

Gregor Reid

I'D LIKE TO THANK MY DIRECTOR, MY FRIENDS AND FAMILY, AND— OF COURSE—THE WRITHING MASS OF GUT BACTERIA INSIDE ME.
I MEAN, THERE'S LIKE ONE OR TWO PINTS OF THEM IN HERE; THEIR CELLS OUTNUMBER MINE!
ANYWAY, THIS WAS A REAL TEAM EFFORT.

CoDa

Vera Pawlowsky-Glahn
Juan Jose Egozcue

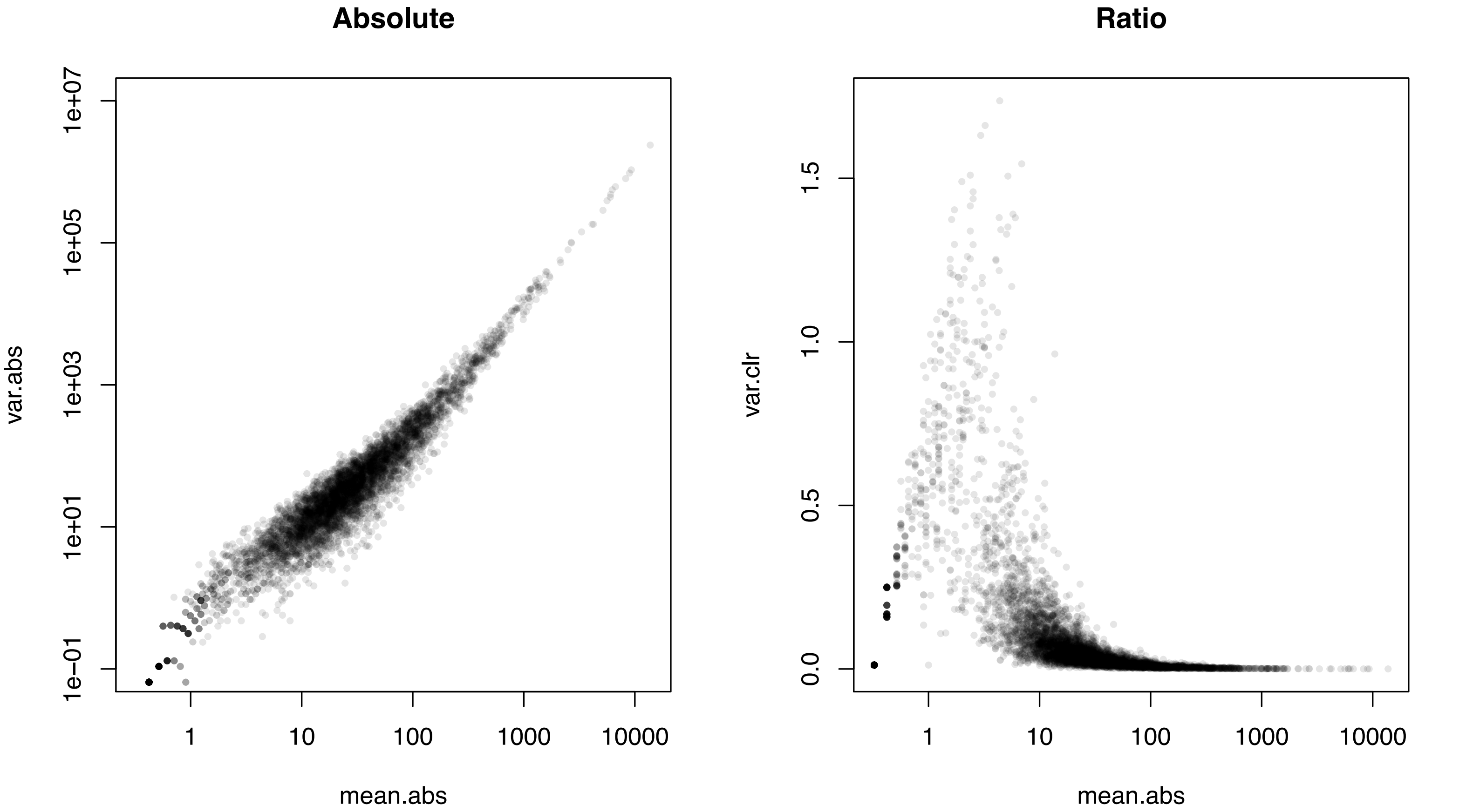
Justin Silverberg 🐯🐞👽

NSERC CRSNG People. Discovery. Innovat

Vogue vaginal microbiome group initiative
Advancing Women's Health through Microbiome Research

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

Non-Linear measurement error

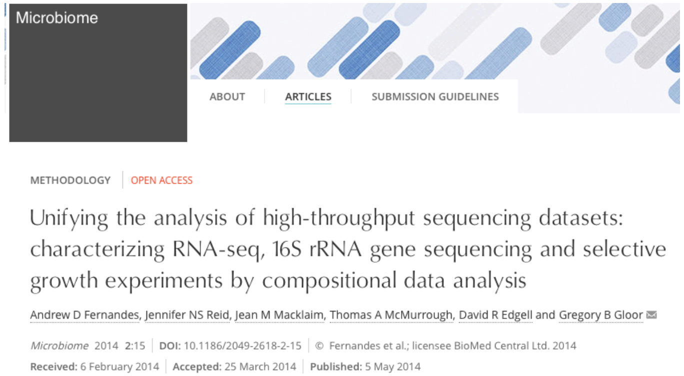


We are working in the wrong space

Constant sum == CoDa

- Correlation/covariation
  - Ordination (PCA), clustering, networks
- Subcompositonal incoherence
  - Normalization, rarefaction, subsetting, aggregation
- Noise is greatest at low count margin
  - Often 'most significant' is least abundant

ggloor.github.io



Microbiome

ABOUT | ARTICLES | SUBMISSION GUIDELINES

METHODOLOGY | OPEN ACCESS

Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis

Andrew D Fernandes, Jennifer NS Reid, Jean M Macklaim, Thomas A McMurrough, David R Edgell and Gregory B Gloor

Microbiome 2014 2:15 | DOI: 10.1186/2049-2618-2-15 | © Fernandes et al.; licensee BioMed Central Ltd. 2014

Received: 6 February 2014 | Accepted: 25 March 2014 | Published: 5 May 2014

PLOS | COMPUTATIONAL BIOLOGY

Browse | Publish | About

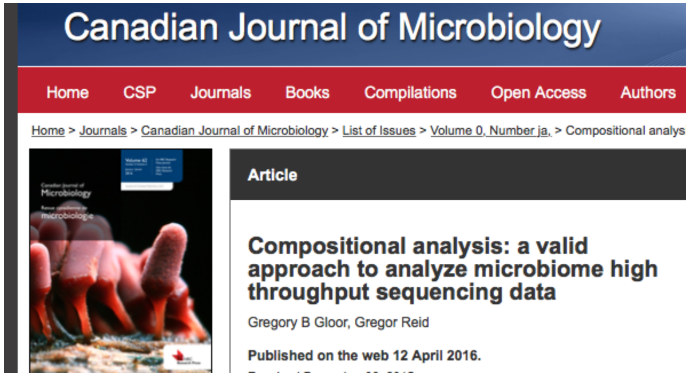
OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Proportionality: A Valid Alternative to Correlation for Relative Data

David Lovell, Vera Pawlowsky-Glahn, Juan José Egozcue, Samuel Marguerat, Jürg Bähler

Published: March 16, 2015 • http://dx.doi.org/10.1371/journal.pcbi.1004075



Canadian Journal of Microbiology

Home | CSP | Journals | Books | Compilations | Open Access | Authors

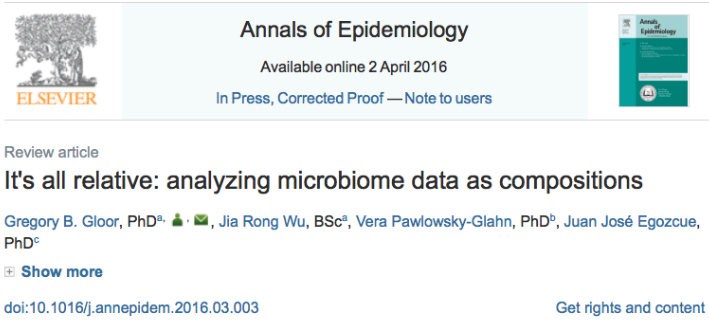
Home > Journals > Canadian Journal of Microbiology > List of Issues > Volume 0, Number ja, > Compositional analys

Article

Compositional analysis: a valid approach to analyze microbiome high throughput sequencing data

Gregory B Gloor, Gregor Reid

Published on the web 12 April 2016.



ELSEVIER

Annals of Epidemiology

Available online 2 April 2016

In Press, Corrected Proof — Note to users

Review article

It's all relative: analyzing microbiome data as compositions

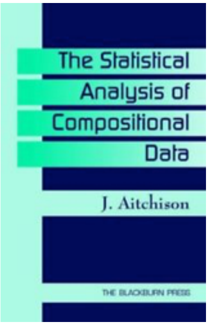
Gregory B. Gloor, PhD[a], Jia Rong Wu, BSc[a], Vera Pawlowsky-Glahn, PhD[b], Juan José Egozcue, PhD[c]

⊞ Show more

doi:10.1016/j.annepidem.2016.03.003

Get rights and content

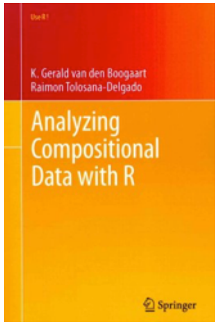
CoDa



The Statistical Analysis of Compositional Data

J. Aitchison

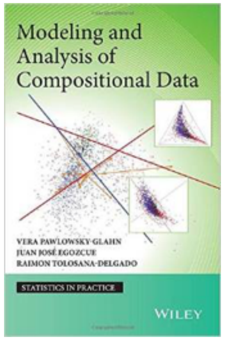
THE BLACKBURN PRESS



K. Gerald van den Boogaart
Raimon Tolosana-Delgado

Analyzing Compositional Data with R

Springer



Modeling and Analysis of Compositional Data

VERA PAWLOWSKY-GLAHN
JUAN JOSÉ EGOZCUE
RAIMON TOLOSANA-DELGADO

STATISTICS IN PRACTICE

WILEY