

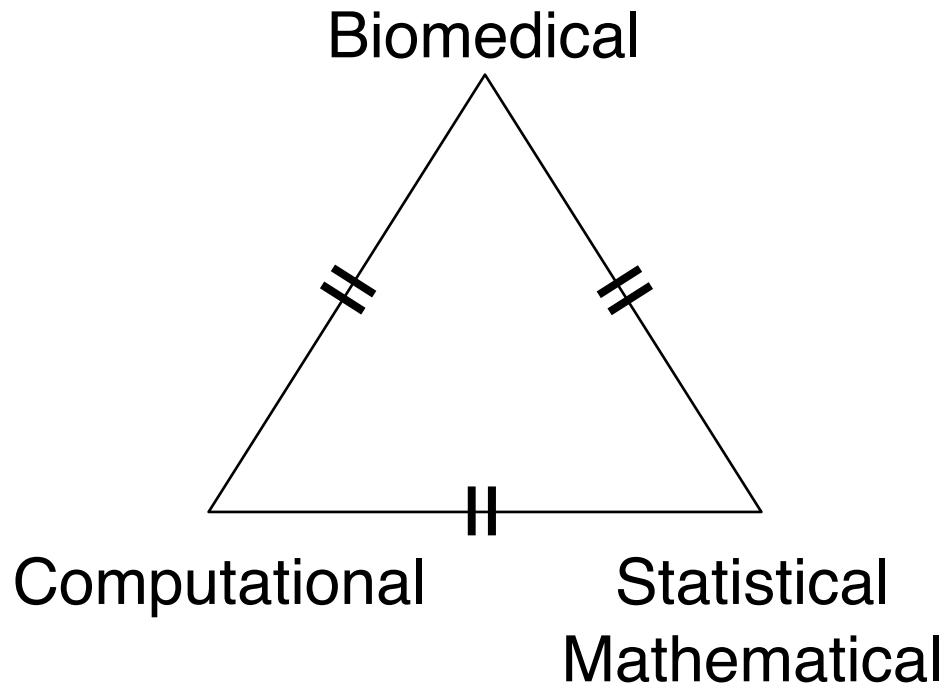
# Enterotypes, Autism, Aliens and the Elderly: the microbiome as a composition

Greg Gloor

Dep't of Biochemistry

University of Western Ontario

# We need to talk more ...



# 'Donalds' give the best quotes

- There are known knowns. These are things we know that we know.
- There are known unknowns. That is to say, there are things that we know we don't know.
- But there are also unknown unknowns. There are things we don't know we don't know.
- **There are unknown knowns. These are things we think we know that we actually don't know.**



# HWHAP

HOUSTON WE HAVE A PROBLEM

- Sequencing data are high-dimensional
  - Therefore statistical analyses can be wildly optimistic
- Sequencing data are sparse
  - Therefore we need to estimate many of our values
- Sequencing data are compositions
  - Applies to any ?-Seq experiment

# The biggest problem is -

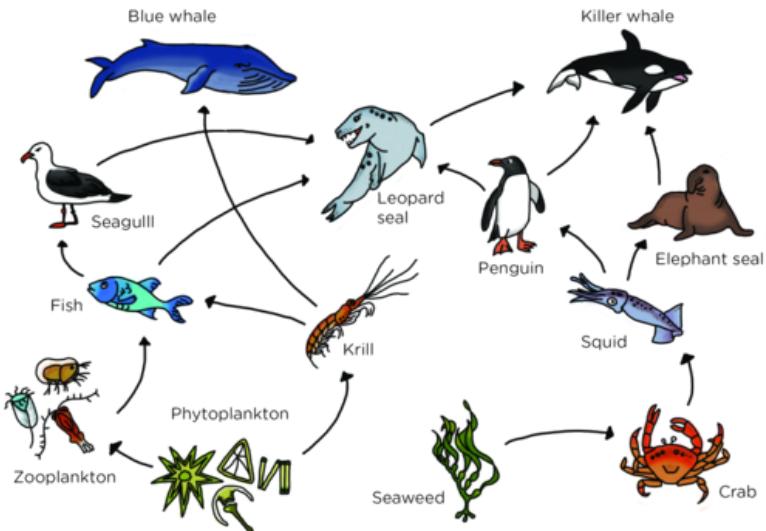
- Sequencing data are high-dimensional
  - Therefore statistical analyses can be wildly optimistic
- Sequencing data are sparse
  - Therefore we need to estimate many of our values
- Sequencing data are compositions
  - Applies to any ?-Seq experiment
- **Most of us are unaware of the implications**

# Enterotypes of the human gut microbiome

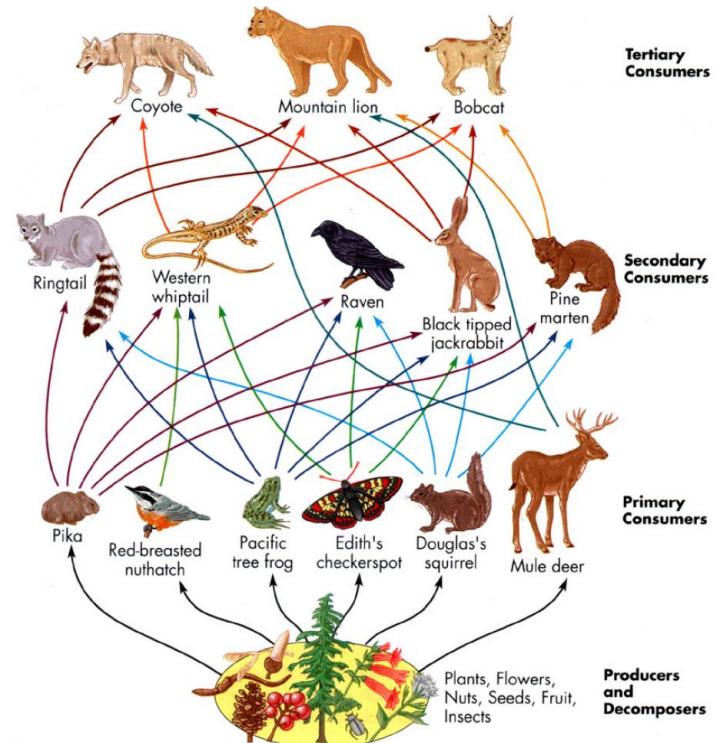
- It has been drawn to our attention that the **methods** described in the main text and the Supplementary Information of this Article have been **considered by some researchers to be insufficient to enable them to identify enterotypes** in their own data sets. Enterotypes were originally defined in this Article (page 177) as “**densely populated areas in a multi-dimensional space of community composition**” and should not be seen as discrete clusters, but as a way of stratifying samples to reduce complexity. Additionally, the Fig. 2 legend should not imply that between-class analysis is simply a method of visualizing principal component analysis (PCA); rather, it is a supervised rather than an unsupervised analysis of data because it incorporates the outcome of clustering of data.  
**To simplify enterotype identification in the original and other data sets, we have developed a comprehensive tutorial** at <http://enterotype.embl.de>—which is a website on enterotypes that will be updated as methods improve. We thank Ivica Letunic and Paul Costea from EMBL for setting up the tutorial.

Arumugam, Nature 2011 *Enterotypes of the human gut microbiome*

# Let's be clear

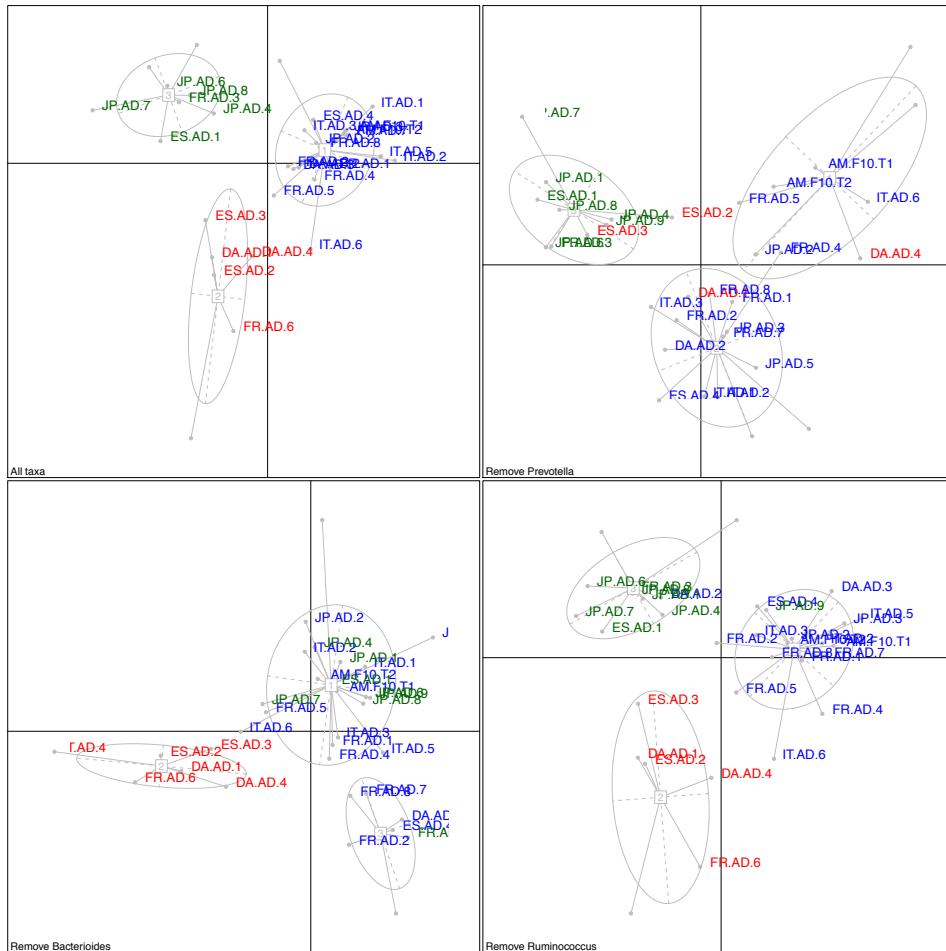


vs.



# An ‘enterotype’ is not the community

<http://enterotype.embl.de/enterotypes.html>



Following their recipe exactly, we can reproduce the result

But ...

Dropping one or more taxa  
shows that the conclusions are  
very fragile

Can *draw* three groups even if  
only two present

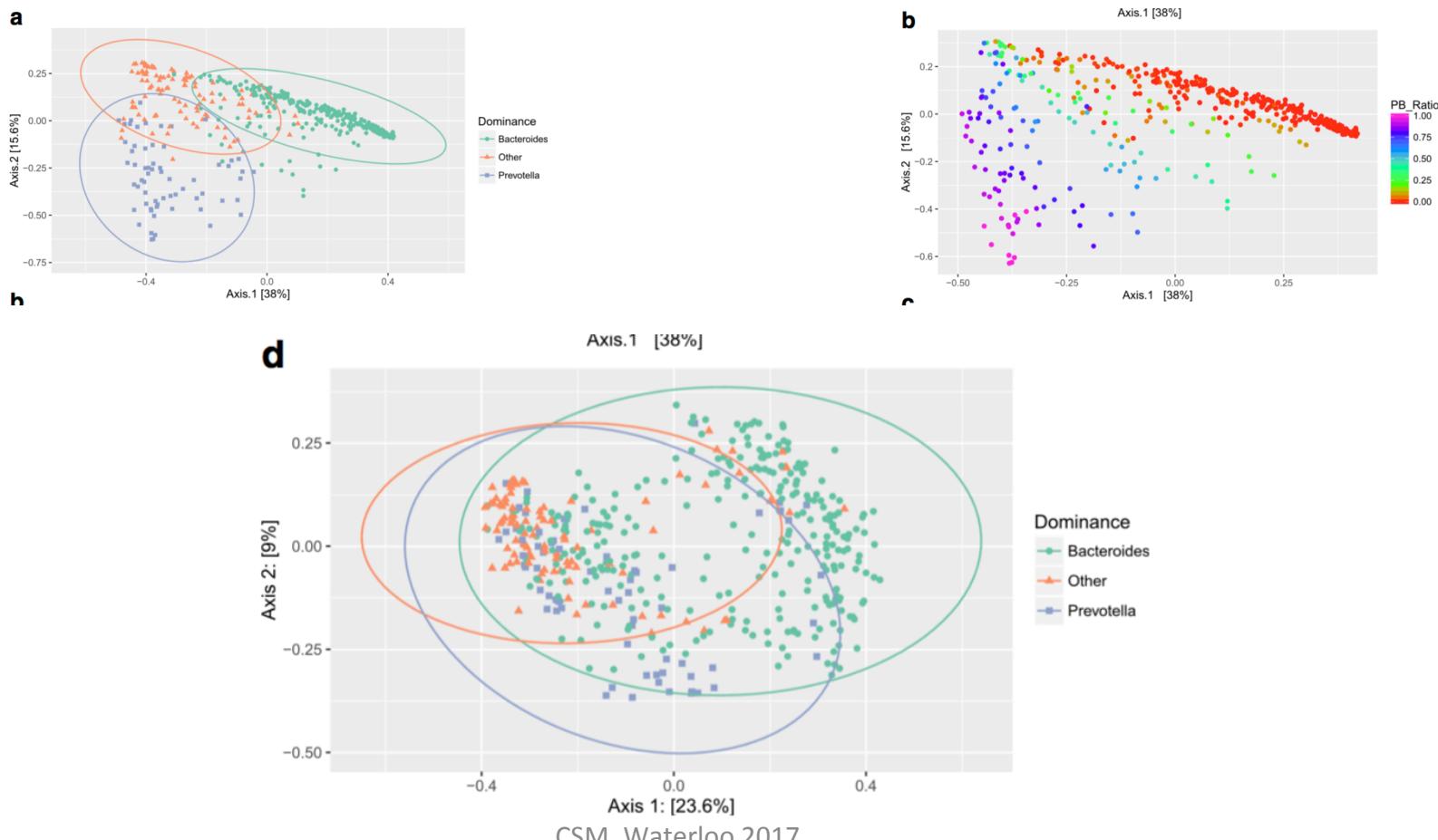
Even changing how we normalize  
the data affects the result

So are enterotypes real? Or an artifact  
of the analysis?



# Interpreting *Prevotella* and *Bacteroides* as biomarkers of diet and lifestyle

Anastassia Gorvitovskaya<sup>1</sup>, Susan P. Holmes<sup>2\*</sup> and Susan M. Huse<sup>3</sup>



# Autism

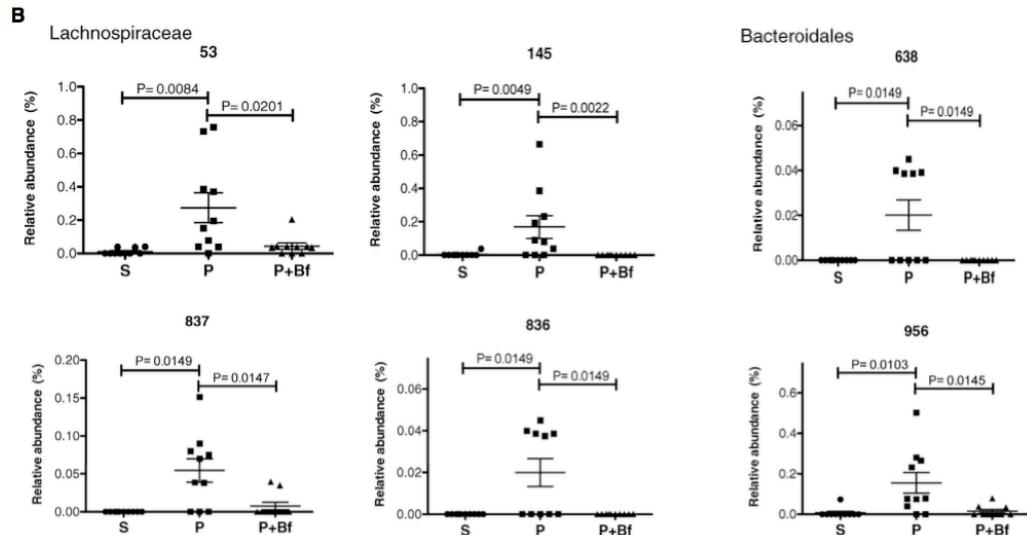
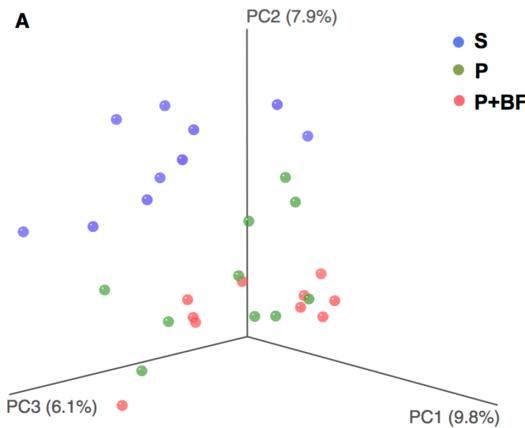
## Microbiota Modulate Behavioral and Physiological Abnormalities Associated with Neurodevelopmental Disorders

Hsiao ... Mazmanian. Cell 155:1451 (2013)

708 citations, top 1% of all social media attention ever

Neurodevelopmental disorders, including **autism spectrum disorder (ASD)**, are defined by core behavioral impairments; however, **subsets of individuals display a spectrum of gastrointestinal (GI) abnormalities**. We demonstrate GI barrier defects and micro- biota alterations in the maternal immune activation (MIA) mouse model that is known to display features of ASD. **Oral treatment of MIA offspring with the human commensal *Bacteroides fragilis*** corrects gut permeability, **alters microbial composition**, and ameliorates defects in communicative, stereotypic, anxiety-like and sensorimotor behaviors. MIA offspring display an altered serum metabolomic profile, and *B. fragilis* modulates levels of several metabolites. Treating naive mice with a metabolite that is increased by MIA and restored by *B. fragilis* causes certain behavioral abnormalities, suggesting that gut bacterial effects on the host metabolome impact behavior. Taken together, these findings support a gut-microbiome-brain connection in a mouse model of ASD and identify a potential probiotic therapy for GI and particular behavioral symptoms in human neurodevelopmental disorders.

# Bf did not change the microbiota!



No significant differences are observed following *B. fragilis* treatment of MIA offspring by PCoA (ANOSIM  $R = 0.0060$   $p = 0.4470$ ; Table S3), in microbiota richness (PD:  $p = 0.2980$ , Observed Species:  $p = 0.5440$ ) and evenness (Gini:  $p = 0.6110$ , Simpson Evenness:  $p = 0.5600$ ; Figures 4A, S2A and S2B), or in relative abundance at the class level (Figure S2C). However, evaluation of key OTUs that discriminate adult MIA offspring from controls reveals that *B. fragilis* treatment significantly alters levels of 35 OTUs (Table S2).

Combinatorics alone indicate we expect ‘significant’ random partitioning in this number of OTUs if we have 50 OTUs that have only 0 or 1 counts across 10 Control and 10 experimental samples

No multiple test correction

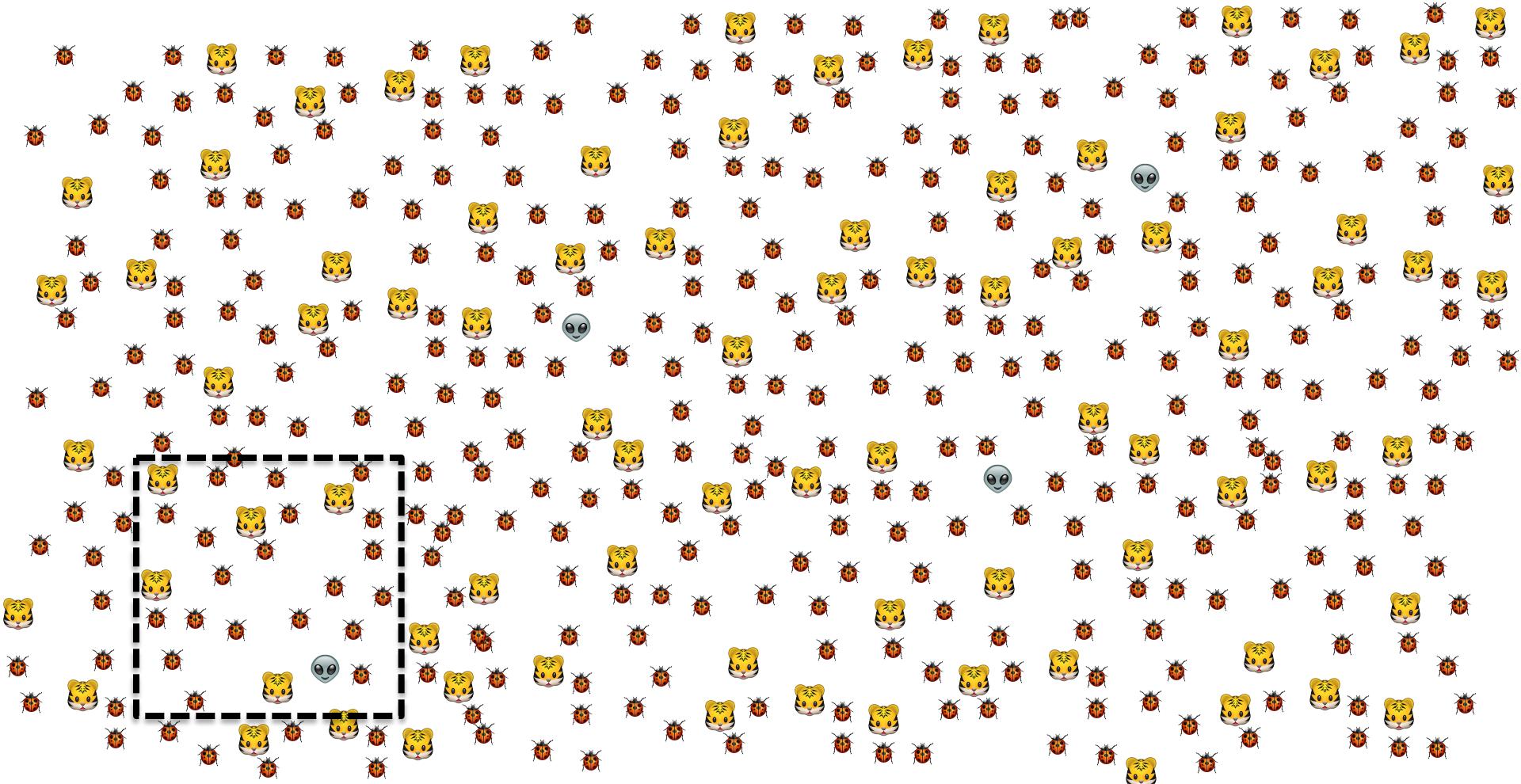
# So we have ...

- Multivariate group differences driven by the most abundant
  - Ordination driven by most abundant not most variable
- Between group univariate differences are usually the rarest\
  - FP rates driven by sparsity (Thorsen Microbiome 2016)

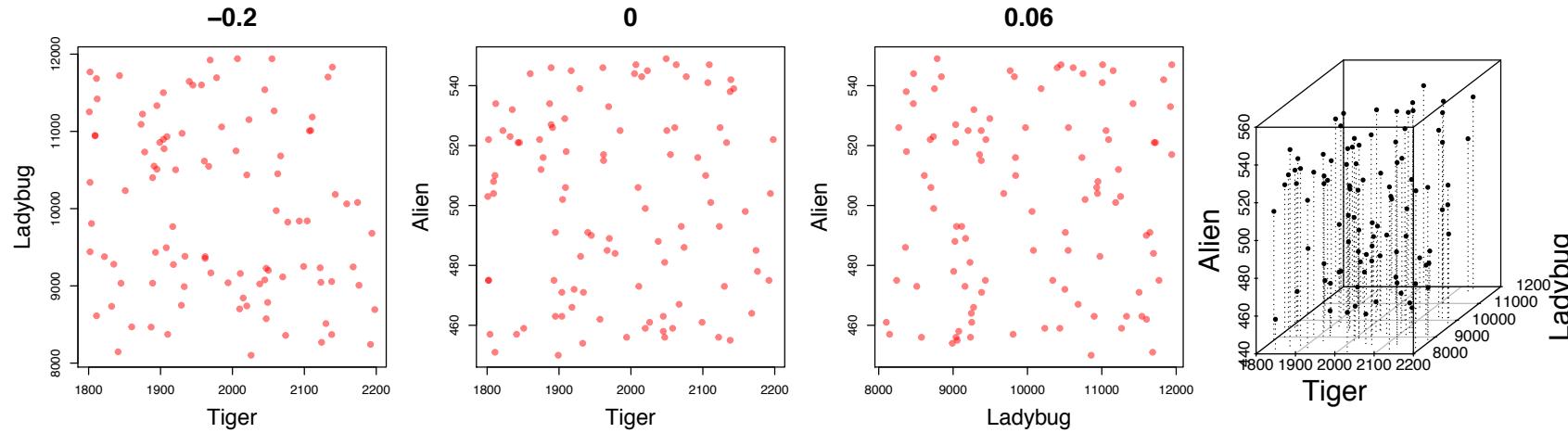


Philosoraptor is confused

# We imagine



# Counting our things



- Example 100 random sample sets



range=1800-2200



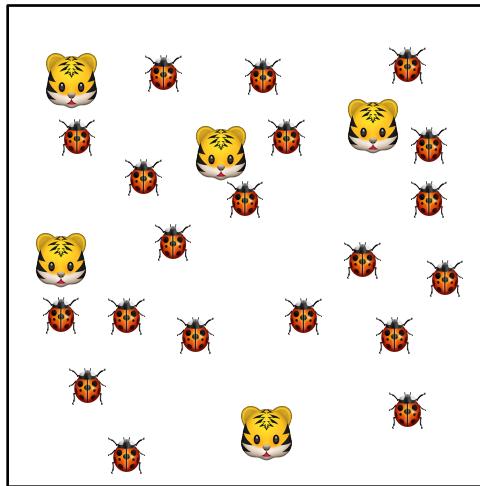
range= 8000-12000



range=450-550

# HTS is not counting

COUNTING



SEQUENCING

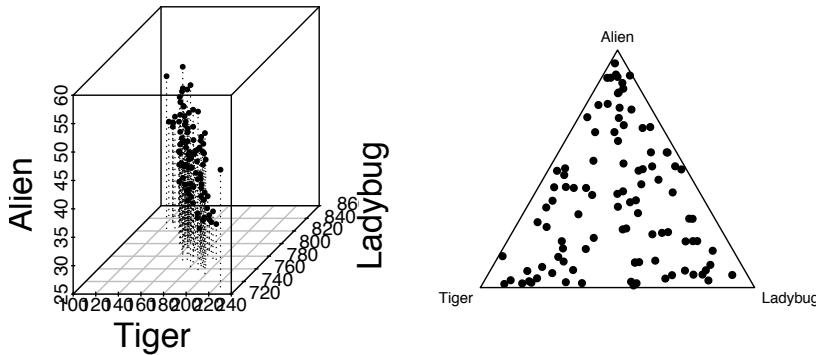
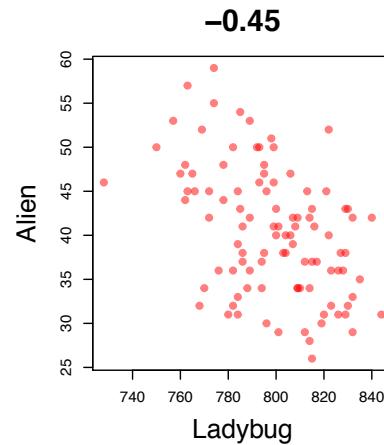
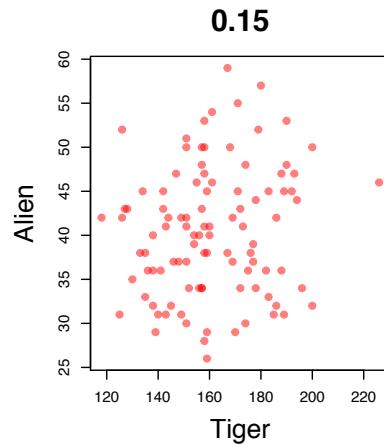
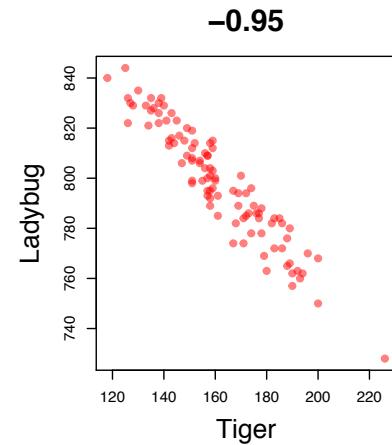
🐯	🐞	🐞	🐞	🐞
🐞	🐞	🐞	🐯	🐞
🐞	🐯	🐞	🐞	🐞
🐞	🐞	🐞	🐞	🐞
🐞	🐞	🐯	🐯	🐞



- Sequencing machines have fixed capacity
  - Cannot just ‘add one more’ forever
  - They do not deliver counts
  - Values are ‘constant sum’ constrained
  - Any constant sum is equivalent to any other

Friedman 2012 PLoS Comp Bio  
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



# Spurious correlation in CoDa

- The correlation observed is not the same for the numerical and proportional data, or for a subset of that data
- **this is spurious correlation** and is an unpredictable correlation observed between two variables whenever they share a common denominator (or a constant sum).



# But my data are not affected ...

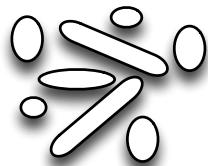
- Transcriptome
  - Remove rRNA, tRNA, snRNA, etc.
  - Or only mRNA, snRNA, tRNA, etc.
  - Different amounts of RNA in the system
- Marker gene sequencing
  - Primer bias
  - Remove low-counts, singletons, sparsity, chimeras, etc.
  - Total bacterial load varies
  - Asymmetry of occurrence
- Metagenome
  - Total bacterial load varies
  - DNA isolation bias
  - Asymmetry of occurrence



# The sequencing meat grinder



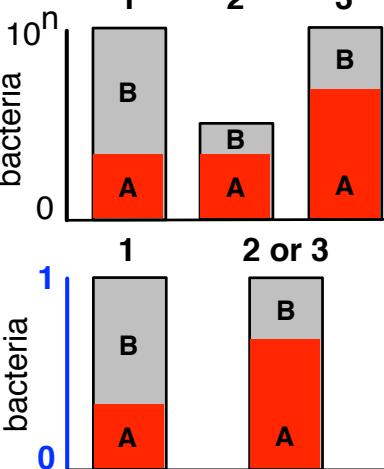
A  
Bacterial population



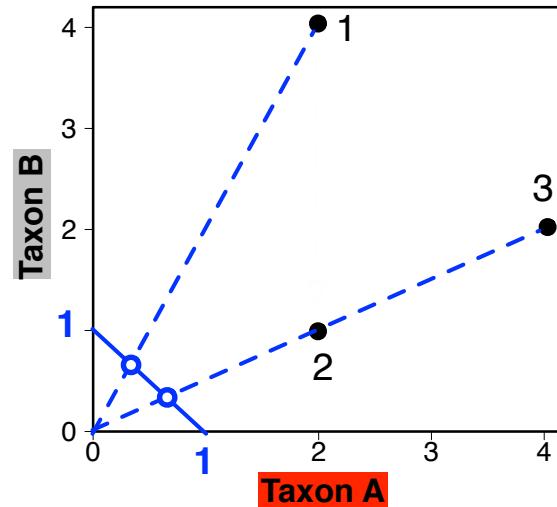
Sequence counts



B  
Total bacteria  
Proportion bacteria



C



D

		Change in population			Perceived change in proportion			
		1 → 2	1 → 3	2 → 3	1 → 2	1 → 3	2 → 3	
		Taxon A	—	↑	↑	↑	—	↑ Increase
		Taxon B	↓	↓	↑	↓	—	↓ Decrease

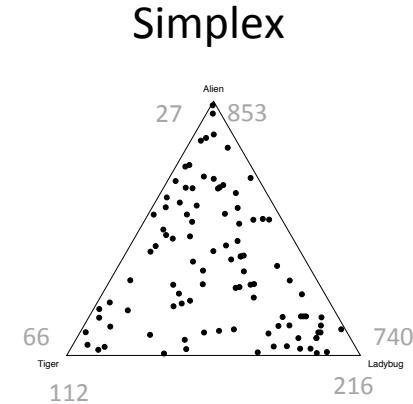
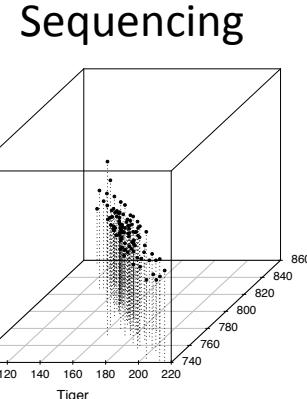
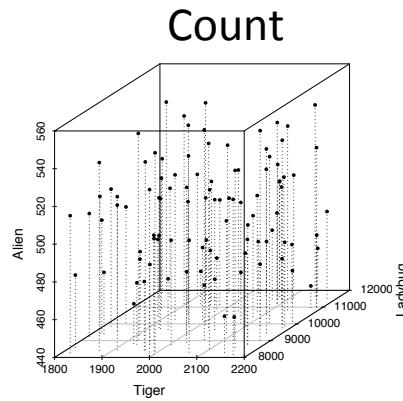
— No change

# CoDa

- **Compositional data are logically inconsistent with most tools in common use (Pearson 1897)**
  - correlation, PCA etc.
  - negative binomial, rarefaction, ZIG
- **The solution is to work in ratios**
  - Aitchison and others
  - Conceptually similar to qPCR



# CoDa properties and the clr



- Simplex: one fewer dimensions than variables
- Problems remain regardless of dimension

$X = [x_1, x_2, \dots, x_D]$ ,  $g_X$  = geometric mean of  $X$

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

- clr transformed data are scale invariant
- Formally, all pairs of ratios measured
- Ratios are linearly related

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

# The 0 in the room

- 0 count can be real – or a non-detect
- Sequencing instruments do not deliver counts
$$X = [x_1, x_2, x_i \dots x_D] = [p_1, p_2, p_i \dots p_D]\alpha ; \alpha = \Sigma X; \Sigma P = 1$$
$$x_i = (P_i | f_i)\alpha$$
- Our approach (ALDEx2)
  - $p_i = 0$  is very strong assumption;  $[P_i > 0, P_i \rightarrow 0]$
  - Generate a Bayesian posterior distribution of probabilities for each count in each sample; clr transform the distribution; perform analyses on each element in the distribution; report expected values
  - Dramatically reduces false positives with little or no loss of sensitivity for essentially any seq'omics dataset



# Examine variance of the ratio between parts



- Examine variance not abundance
- Easily interpret relationships between samples and features
- Robust to perturbations
  - Exploratory data analysis
  - Differential abundance
  - Compositional association

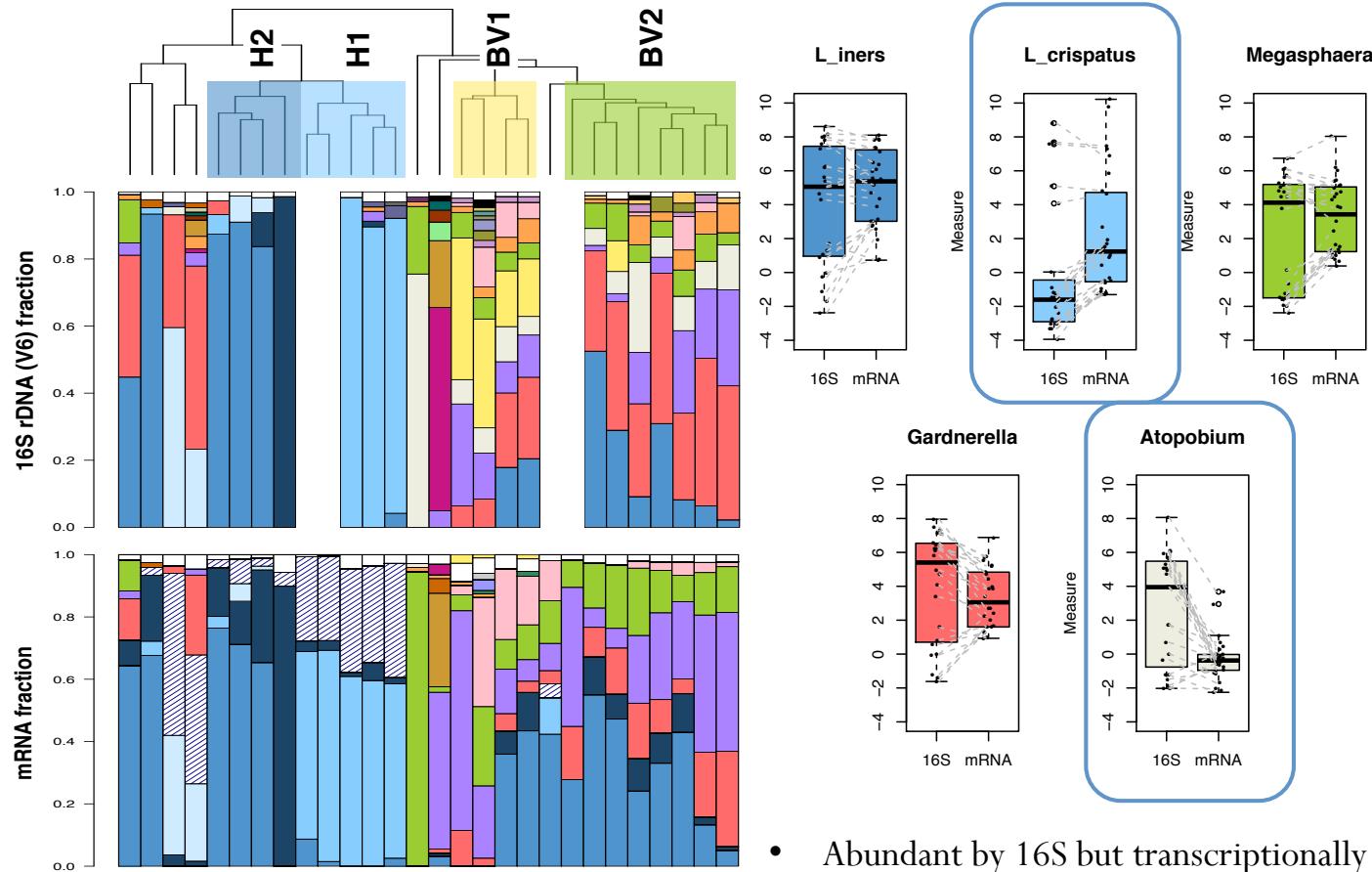
# Applying analysis tools based on variance of the ratios between parts

- Outliers
  - CoDaSeq ([github.com/ggloor](https://github.com/ggloor))
- Exploratory data analysis
  - Compositional biplots
  - omicplotR (Dan Giguere: [github.com/dgiguer/omicplotR/wiki](https://github.com/dgiguer/omicplotR/wiki))
- Differential abundance
  - ALDEx2 (Bioconductor)
- Compositional association
  - propr (CRAN)
  - CoDaSeq ([github.com/ggloor](https://github.com/ggloor))





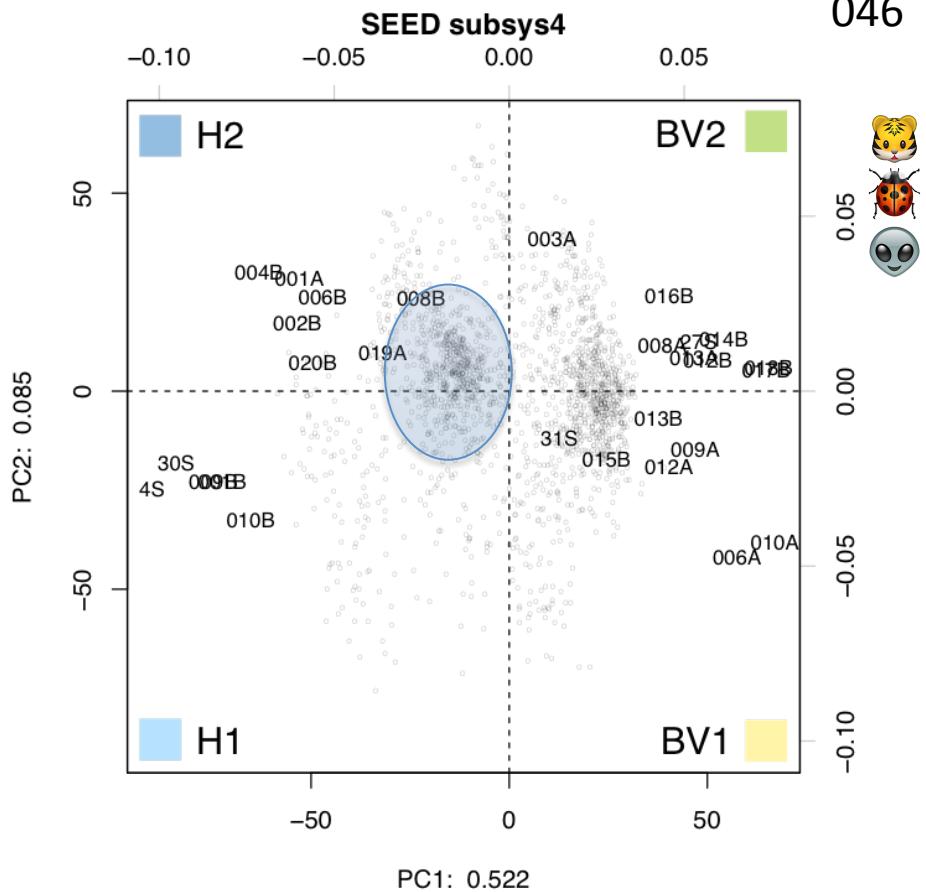
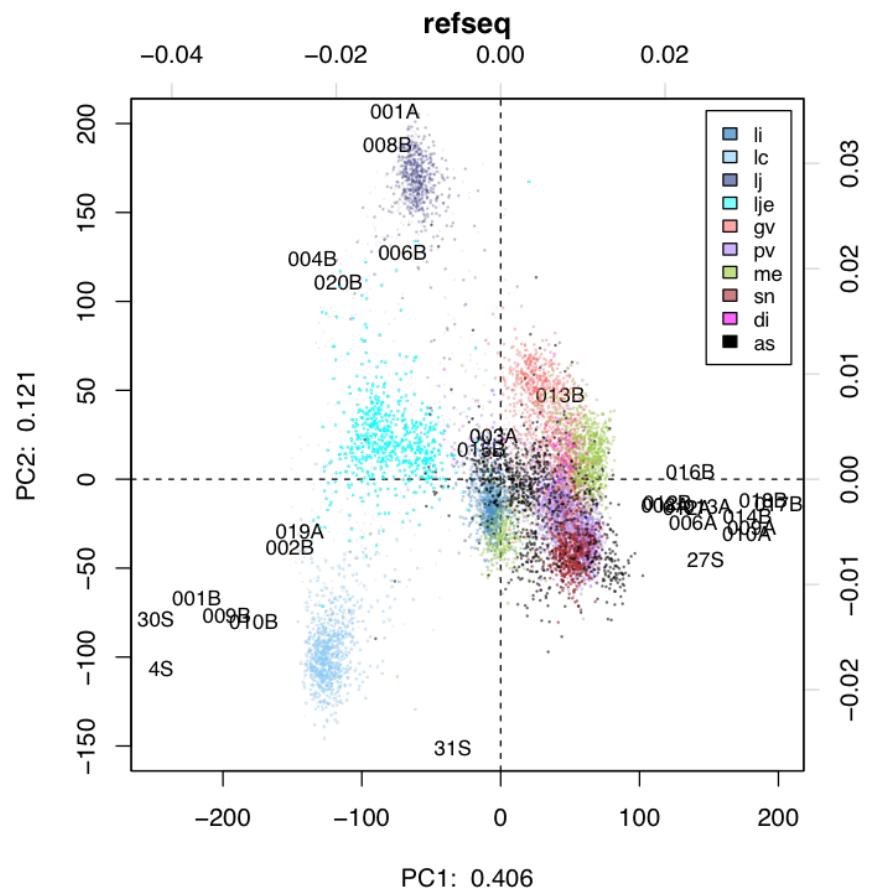
# Meta-transcriptome of H vs. BV



- Abundant by 16S but transcriptionally rare

Macklain, McMillan- PC  
McMillan, Sci. Rep, 2015

# Gene or function



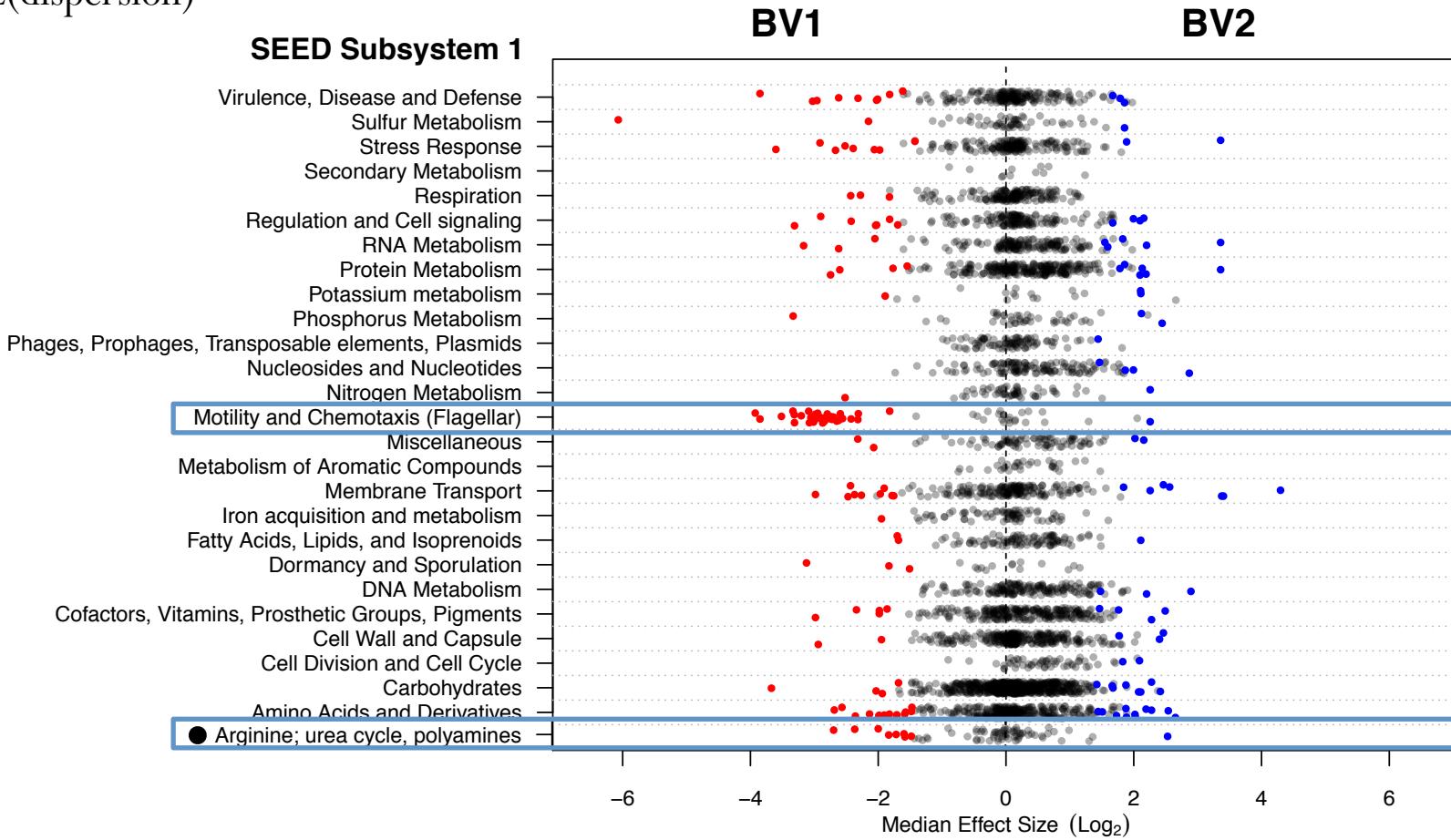
- Gene level shows species or strain level contribution
  - Function level is redundant - e.g. housekeeping functions collapse

# What separates BV subgroups?

- ALDEEx2 Effect size:

E(difference)

E(dispersion)



# Summary

## Analysis

Beta Diversity

Clustering

Differential abundance

Correlation

Interpretation

	Standard	CoDa
Beta Diversity	Driven by most abundant taxon or gene	Variance of ratios between taxa or genes
Clustering	Driven by most abundant taxon or gene	Variance of ratios between taxa or genes
Differential abundance	Usually rarest taxon or gene – most variable within and between groups	Variance of ratios between taxa or genes –most variable between groups
Correlation	Just wrong – false positives	Pairs of taxa or genes that have common variance
Interpretation	Seems simple but is not	Seems hard but is not



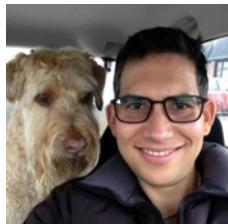
# Conclusions

- We only have relative information
  - Need to recast our questions. How are things changing relative to each other?
  - It takes a shift in thinking to interpret the data correctly
- Variances of the ratios is informative
  - Same features are identified through multivariate ordination, univariate differential abundance, compositional association
- 0 replacement is a theoretical, but not a practical impediment



# Acknowledgments

Canadian Centre for Human Microbiome  
and Probiotic Research



Andrew  
Fernandes



Jean  
Macklaim

Gregor Reid  
Jeremy Burton  
Amy McMillan  
Daniel Giguere  
Jia Rong Wu

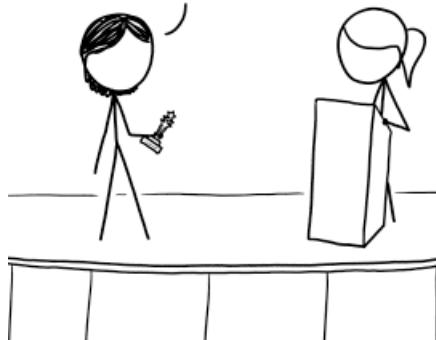


People. Discovery. Innovat

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.



Vera Pawlowsky-Glahn  
Juan Jose Egozcue



Justin Silverman  
philr (CRAN)

**V**igue  
vaginal microbiome group initiative  
Advancing Women's Health through  
Microbiome Research  
CSM\_Waterloo 2017

**CIHR IRSC**



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

# Reading & Sources



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 1a, > Compositional analysis

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.



## Annals of Epidemiology

Available online 2 April 2016

In Press, Corrected Proof — Note to users



Review article

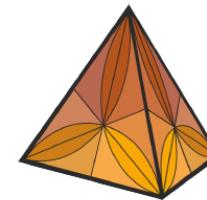
## It's all relative: analyzing microbiome data as compositions

Gregory B. Gloor, PhD<sup>a</sup>, , Jia Rong Wu, BSc<sup>a</sup>, Vera Pawlowsky-Glahn, PhD<sup>b</sup>, Juan José Egozcue,  
PhD<sup>c</sup>

[Show more](#)

[doi:10.1016/j.annepidem.2016.03.003](https://doi.org/10.1016/j.annepidem.2016.03.003)

[Get rights and content](#)



# CoDa

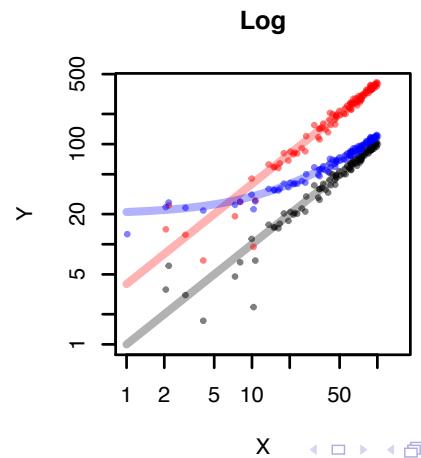
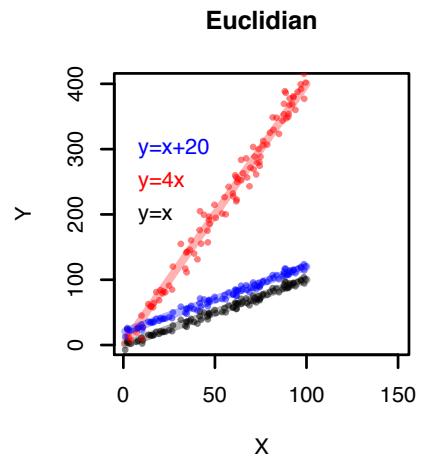


# Fake Correlations!!!



In addition to spurious correlation

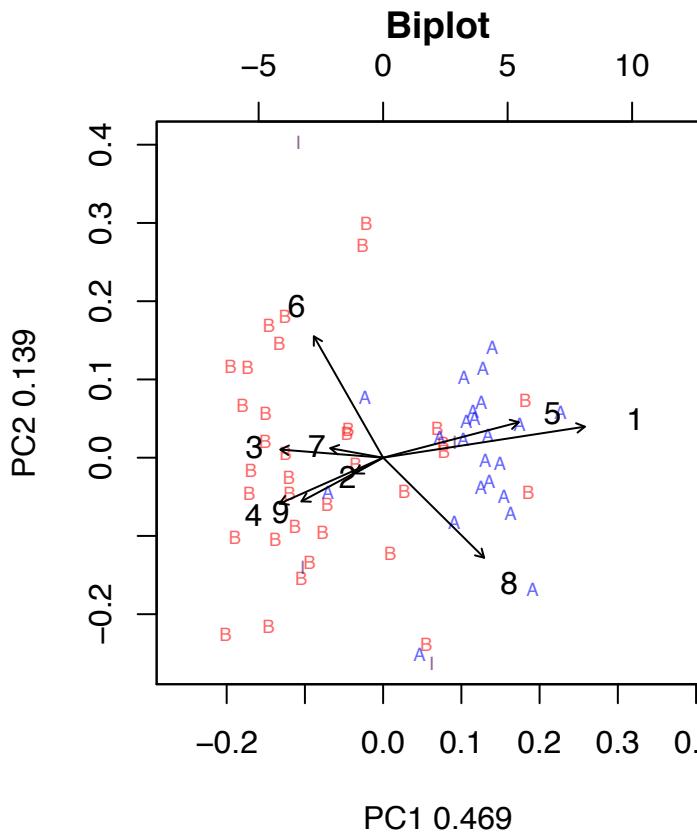
False positive -ve correlations because ...  
what is the correlation between heads and tails?

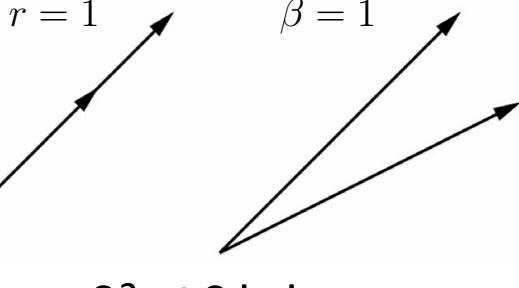


False positive +ve correlations  
because associations must  
maintain a constant ratio ...

# The $\emptyset$ metric

propr, CoDaSeq

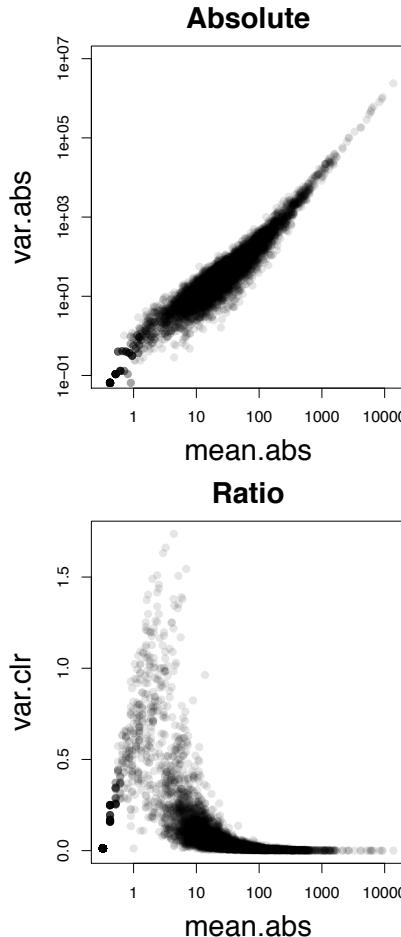


- 
- $\emptyset = 1 + \beta^2 - 2\beta|r|$
  - $r=1$  when two vectors in same direction
  - $\beta=1$  when the two vectors have the same variance
  - So  $\emptyset = 0$  iff  $r=2$  &  $\beta=1$ 
    - Collapsing two measures into one number

Lovell, PLoS Comp. Bio. 2015  
Erb, Theor. Biosci. 2016



# Where is the noise?

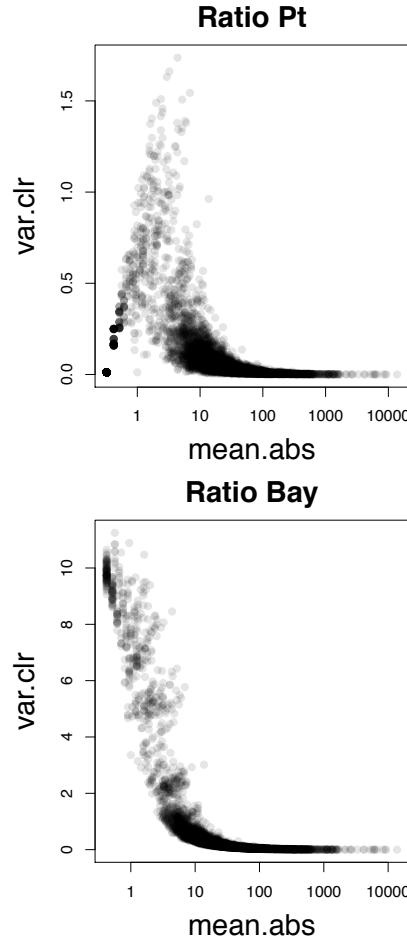


Variance vs. mean for data on the simplex

- Transcriptome dataset
- Top shows absolute data, with near linear dependence
  - Modeled by negative binomial
  - Lowest and highest counts have least variance
  - 0 is 0!
  - Inflates statistical significance at low count margin
- Bottom shows relative data, non-linear dependence on ratios
  - Follows approximate Dirichlet
  - Lowest and highest counts have least variance, but 0 is 0!
  - Inflates statistical significance at low count margin



# Bayesian estimate

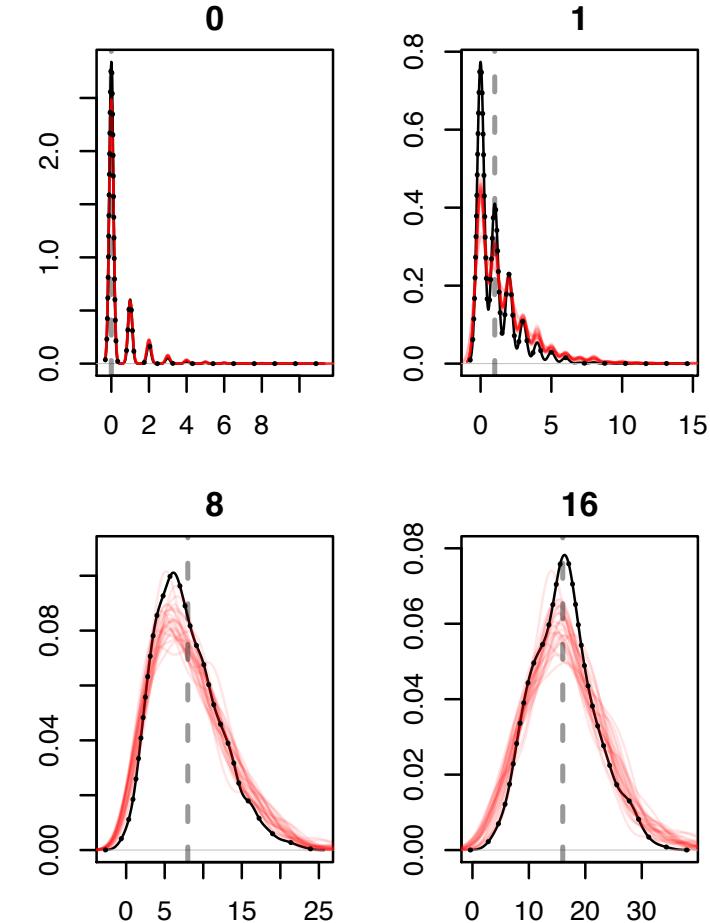


Variance vs. mean for data on the simplex

- Top shows point estimate
  - Follows approximate Dirichlet
  - Lowest and highest counts have least variance, but 0 is 0!
  - Inflates statistical significance at low count margin
- Bottom shows Bayesian estimate
  - Probabilities not counts
  - Drawn from Dirichlet
  - Values replaced by estimated probability of observing variable given total
  - Highest counts have least variance
  - Low counts must be very different to be significant at low count margin



# Modeling technical variation



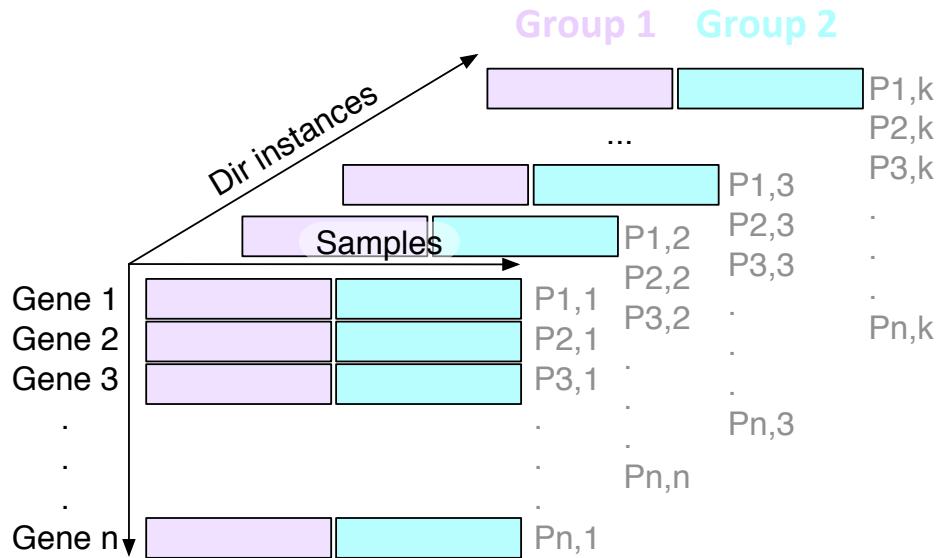
**Distribution of values in replicate B for replicate A values**

- Transcriptome dataset
- True technical variation in black
- Inferred technical variation in red



Gloor et al. 2016. Aus. J. Statistics

# Find test values not affected by random sampling

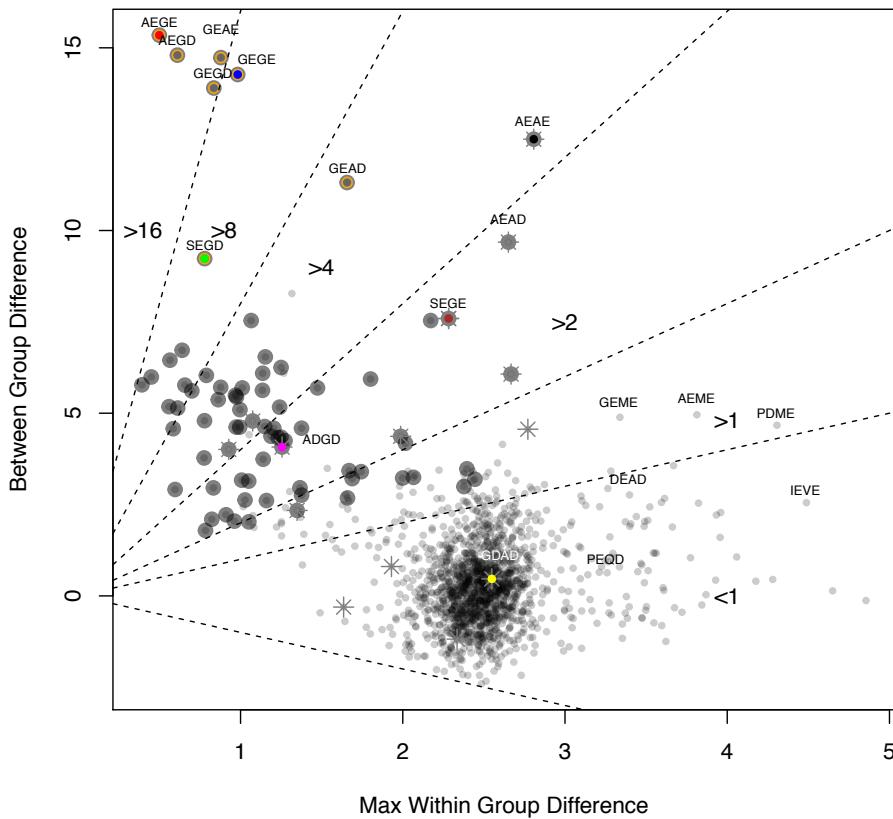


## ALDEX2

- Generate posterior estimates of the data consistent with the observed data and the chosen prior(s)
  - Dirichlet instances
- clr transform instances
- Calculate univariate test values on each instance
- Correct for FDR
- **Report the expected value of each test across instances**

Fernandes, et al. 2013. PLoS ONE  
Fernandes, et al. 2014. Microbiome

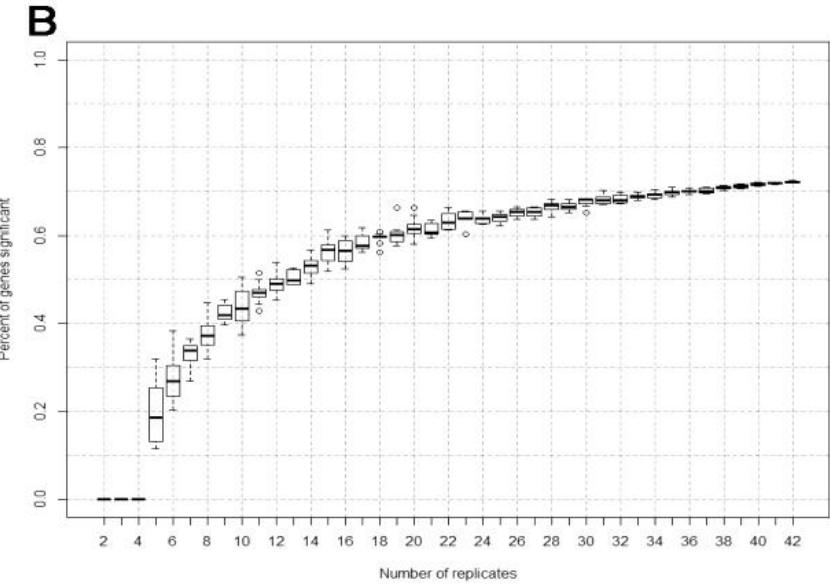
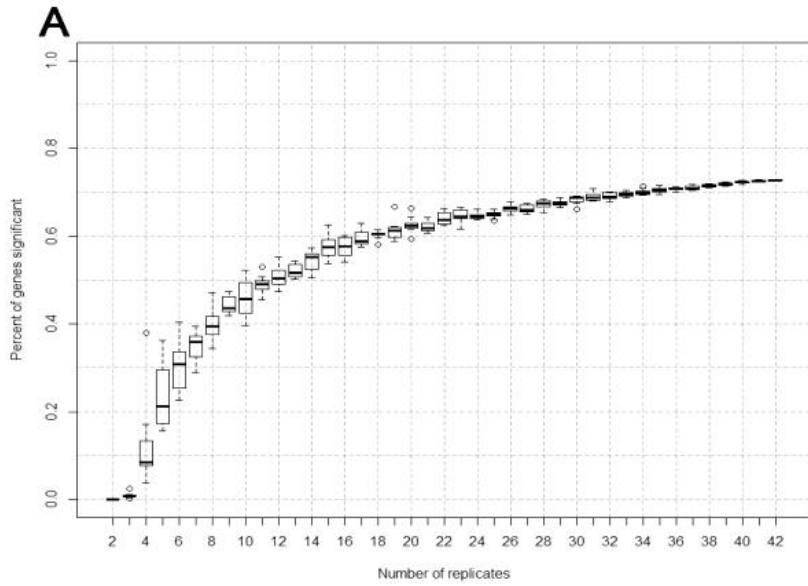
# Effect-sizes



- $E = \text{Difference}/\text{Dispersion}$
- SELEX dataset
- Convenient way to summarize univariate differences in multivariate data
- X axis - maximum variation within either group for each feature
- Y axis – Difference between groups for each feature
- Lines indicate effect size boundaries
  - $D=1, V=2$   $E=0.5$
  - $D=1, V=0.5$   $E=2$
  - $D=2, V=2$   $E=1$
- Read the plot in sectors
- Most 16S rRNA sequencing experiments have  $E < 1$
- Transcriptome experiments  $E > 2$



# P-values are not reliable



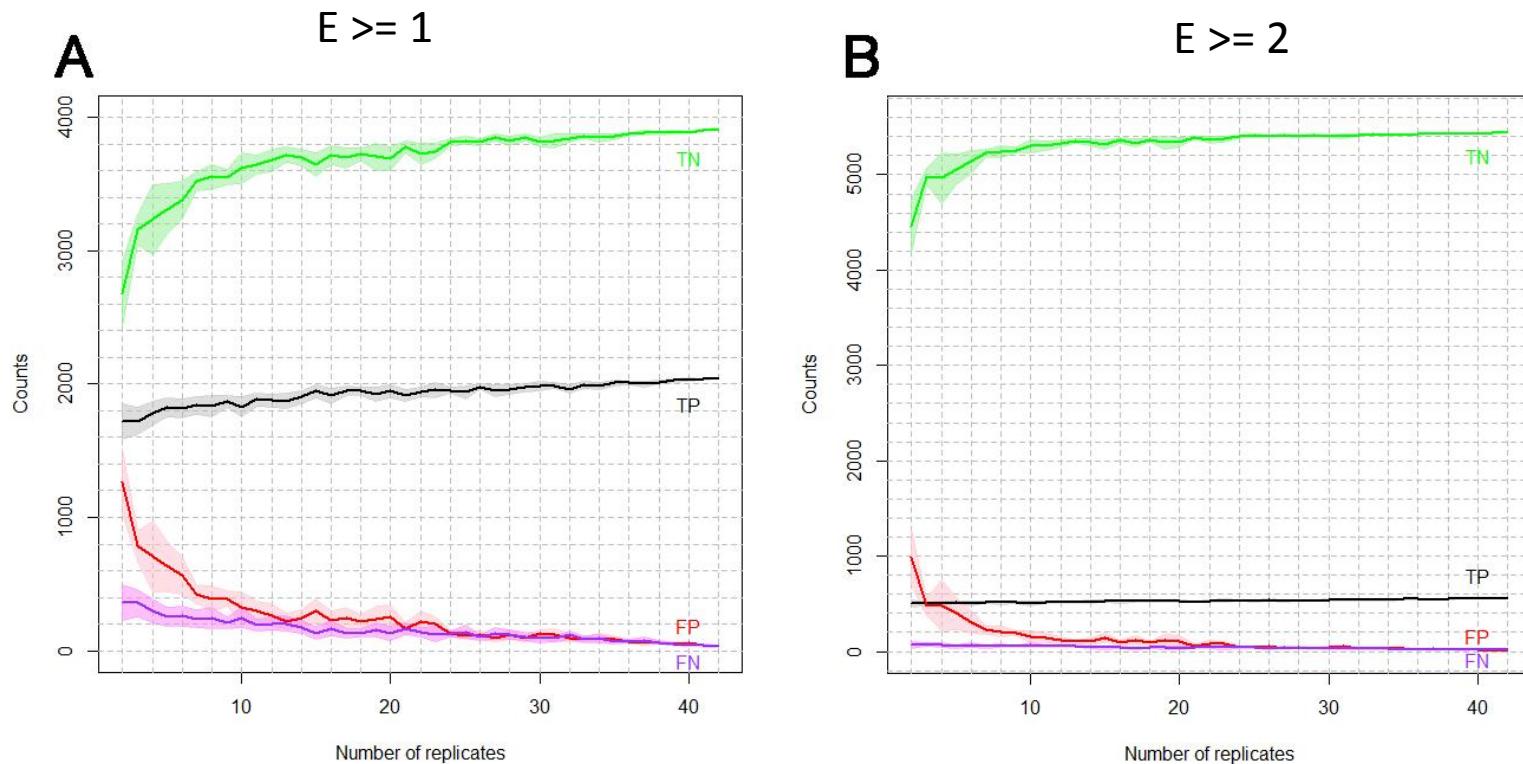
## Yeast transcriptome dataset

- 48 replicates, 2 conditions
- Randomly choose n replicates
- Calculate P and corrected P (q)  
‘Truth’ is q < 0.05 of 48 replicates

Halsey, Nature Meth. 2015



# Effect sizes are stable estimates

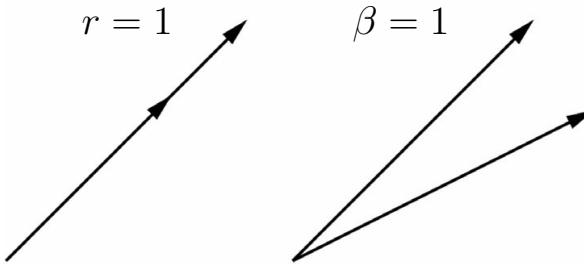


Halsey, Nature Meth. 2015

# The $\emptyset$ metric

- $\emptyset = 1 + \beta^2 - 2\beta r$

- $x_i = \text{clr}(x_i)$ ,  $x_j = \text{clr}(x_j)$



- $\emptyset_{ij} = (\text{var}(x_i) - \text{var}(x_j)) / (\text{var}(x_i) + \text{var}(x_j))$



Lovell, PLoS Comp. Bio. 2015