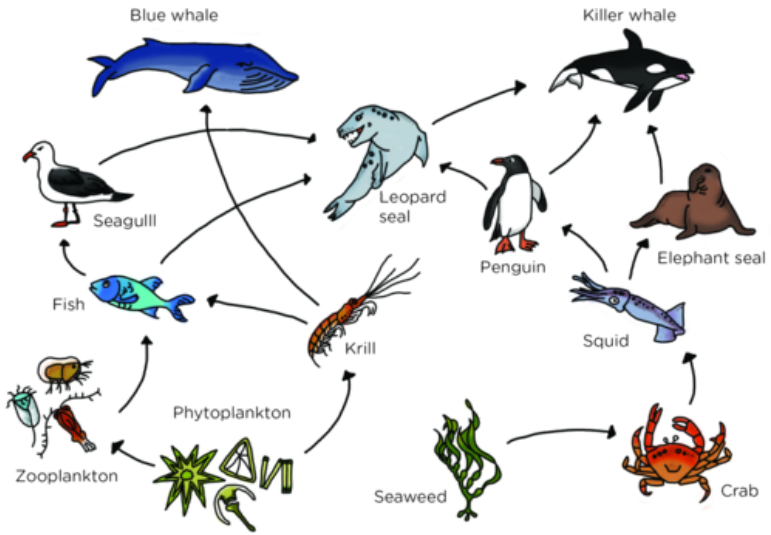
Microbiome datasets are compositional: and this is not optional

Greg Gloor

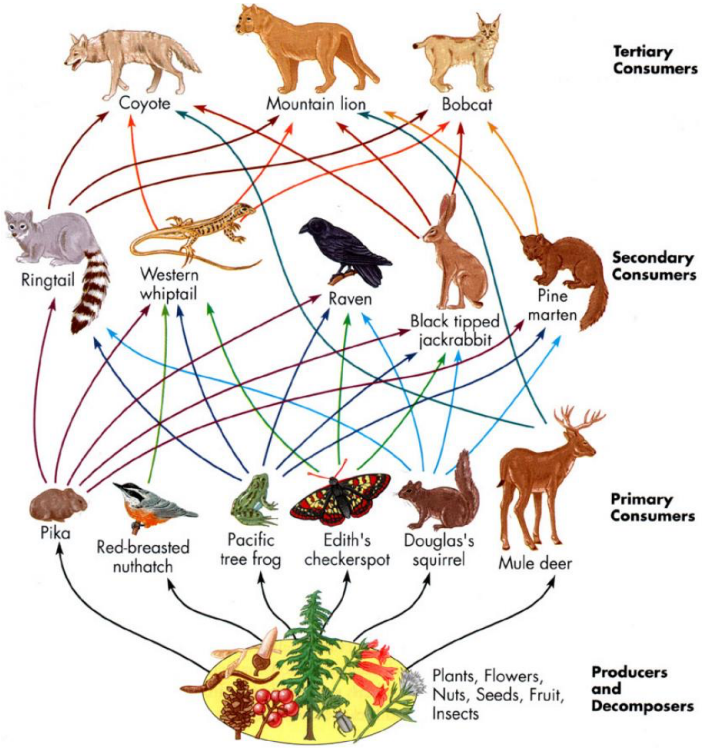
Dep't of Biochemistry

University of Western Ontario

Let’s be clear



vs.



‘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 also things we think we know that we don’t know**

- 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 constant sum
  - Therefore we have compositional data

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 constant sum
  - everything correlates with everything
- **Most of us are unaware of the problems**

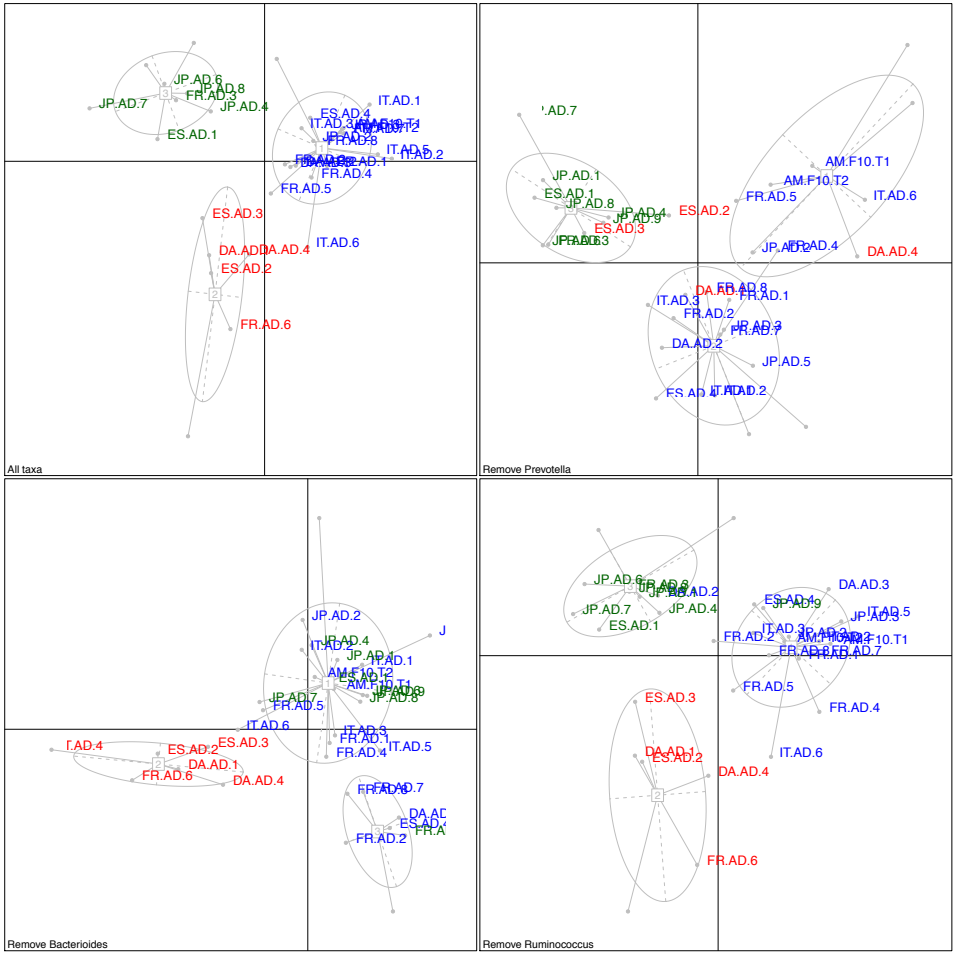
ARTICLE

doi:10.1038/nature09944

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.

An Enterotype is not the community



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

Following their recipe exactly, we can reproduce the result

But …

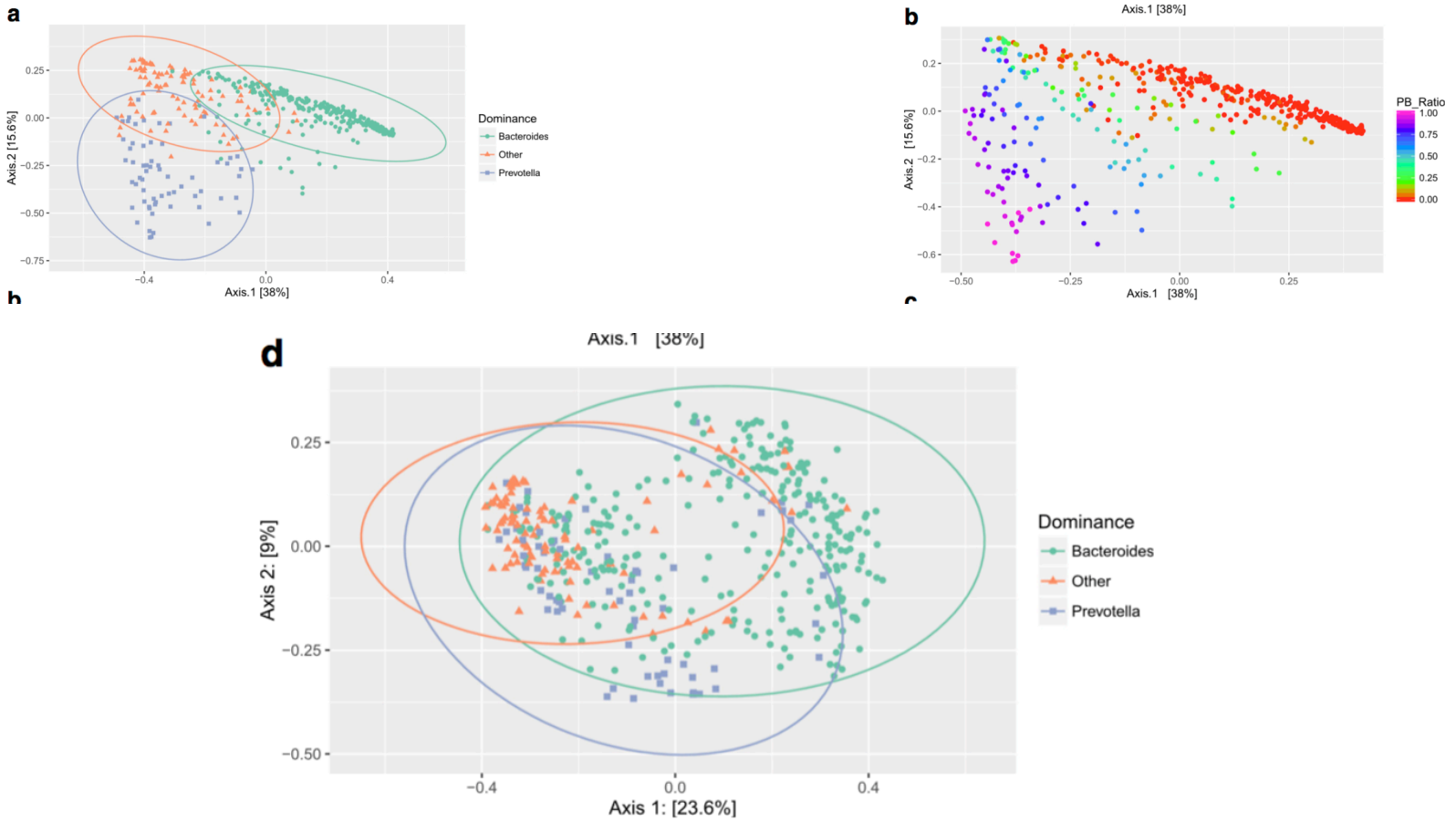
- Dropping one or more taxa shows that the conclusions are very fragile
- Can *draw* three groups even if only two present
- Even changing how we normalize the data affects the result

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

RESEARCH Open Access

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

Anastassia Gorvitovskaia[1], Susan P. Holmes[2*] and Susan M. Huse[3]



Autism

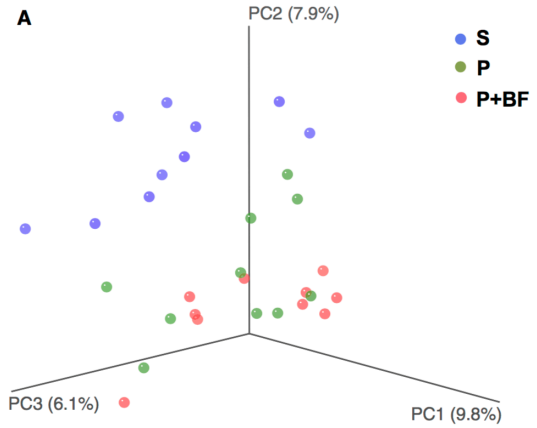
Microbiota Modulate Behavioral and Physiological Abnormalities Associated with Neurodevelopmental Disorders

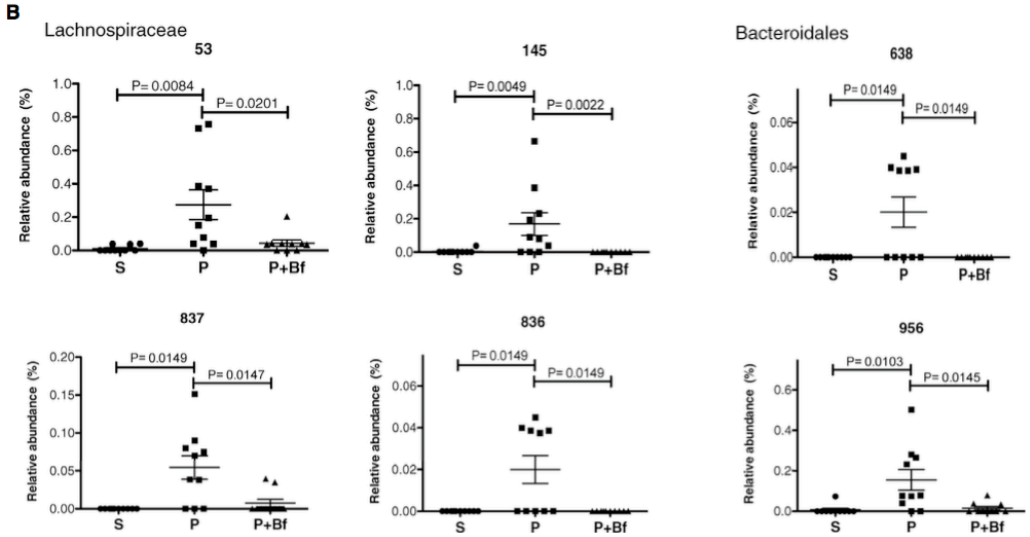
Hsiao … Mazmanian. Cell 155:1451 (2013)

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).**

If we have 50 taxa that have only 0 or 1 counts across 10 Control and 10 experimental, expect 1 in 50 to partition as shown with the number of OTUs in the system.

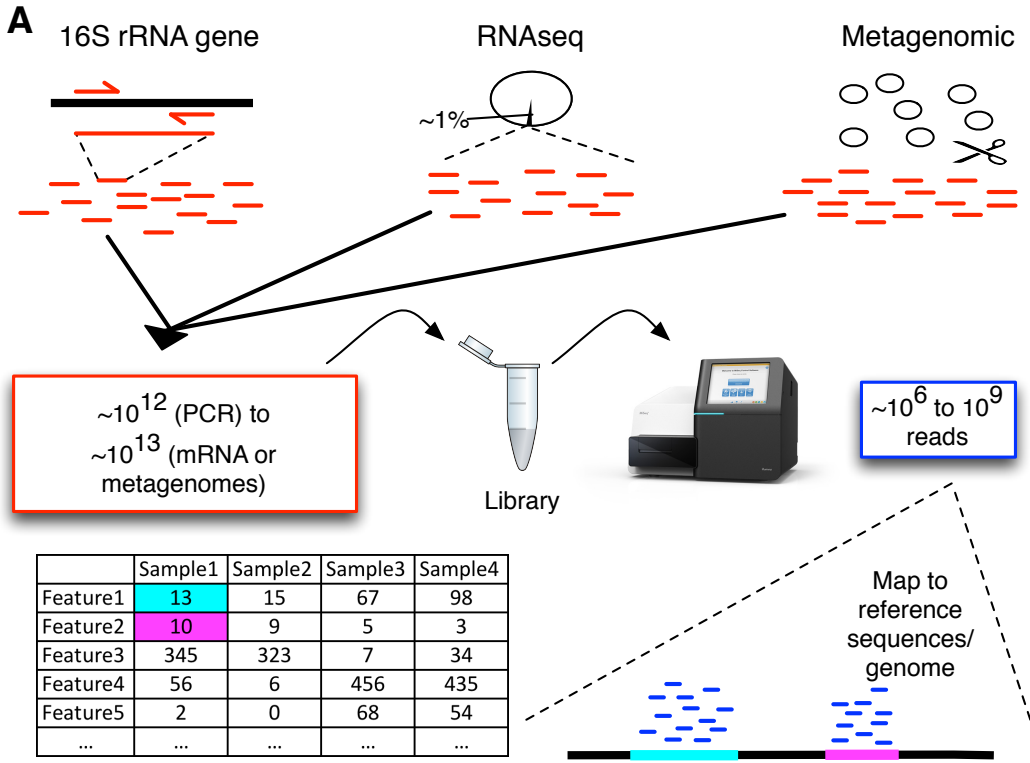
No multiple test correction

So we have …

- Multivariate group differences driven by the most abundant
- Between group univariate differences are usually the rarest

Philosoraptor is confused

Back to basics



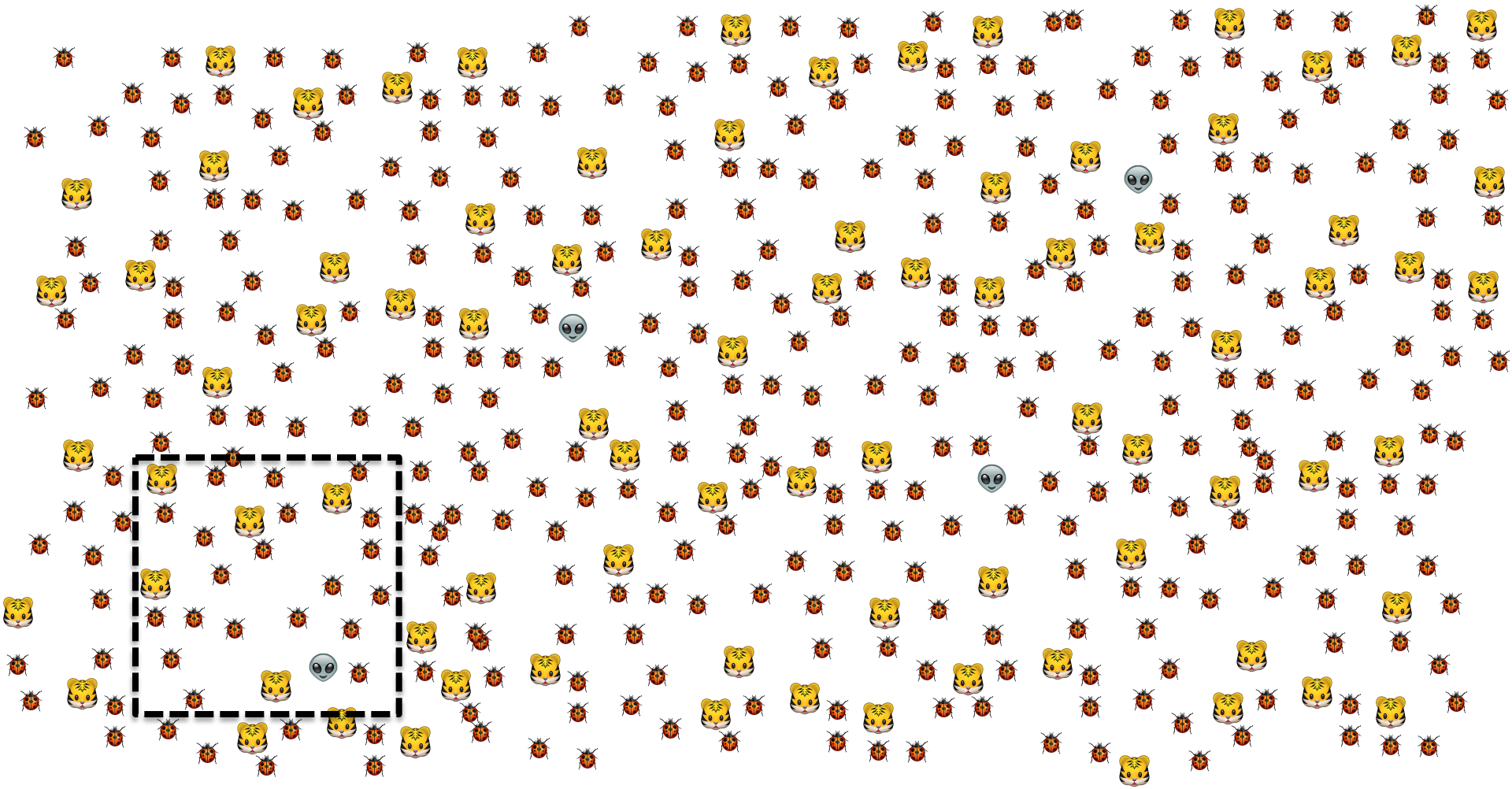
	Sample1	Sample2	Sample3	Sample4
Feature1	13	15	67	98
Feature2	10	9	5	3
Feature3	345	323	7	34
Feature4	56	6	456	435
Feature5	2	0	68	54
...	...	...	...	...

- DNA or RNA fragments from the environment
- A random sample used to make a library
- A random sample of the library is sequenced and mapped
- This generates a table of counts per feature (OTU, gene) in each sample

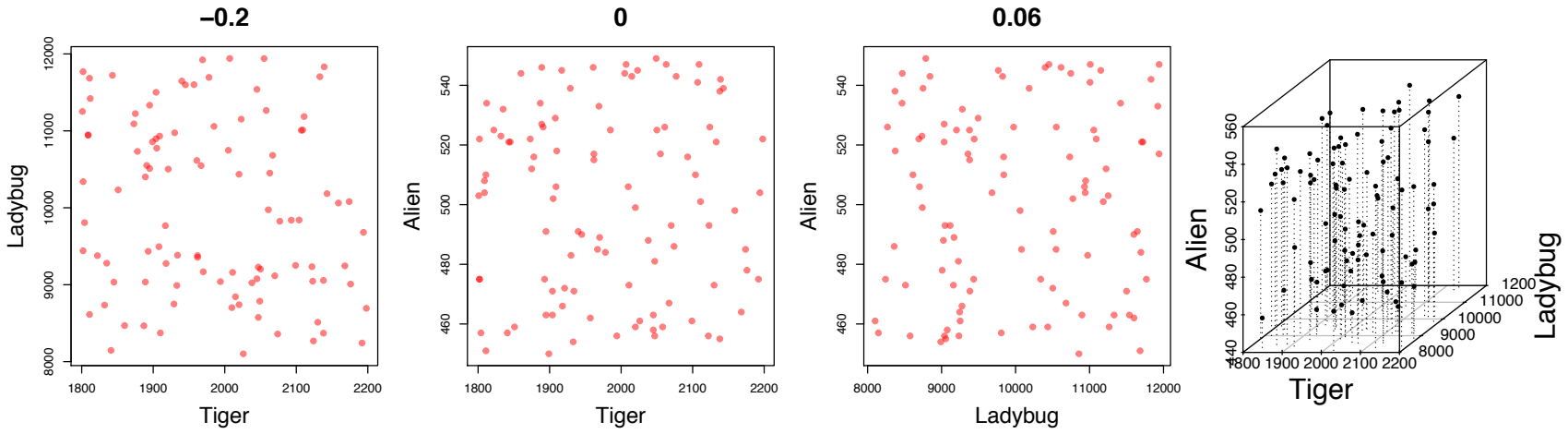
- **The number of reads is determined by the machine!**

Gloor et al. 2016. Ann. Epidem.
Gloor et al. 2016. Can. J. Micro.
Gloor et al. 2016 Aus. J. Stats

Random sample of environment



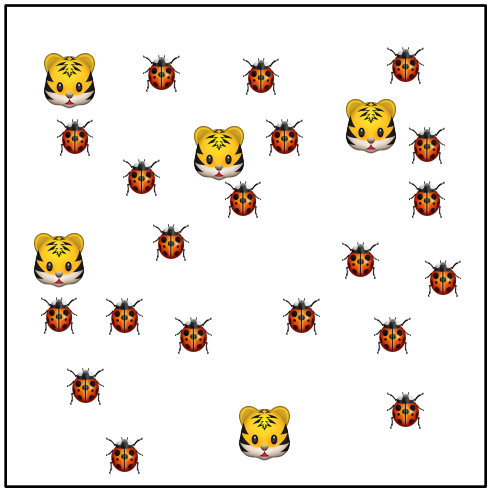
Counting our things



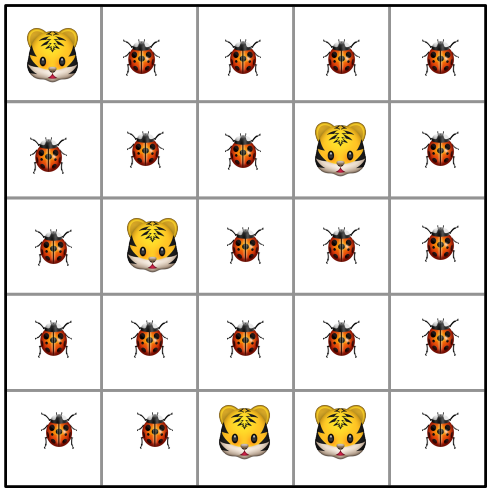
- **<u>Example 100 random sample sets</u>**
  - range=1800-2200
  - range= 8000-12000
  - range=450-550

HTS is not counting

COUNTING



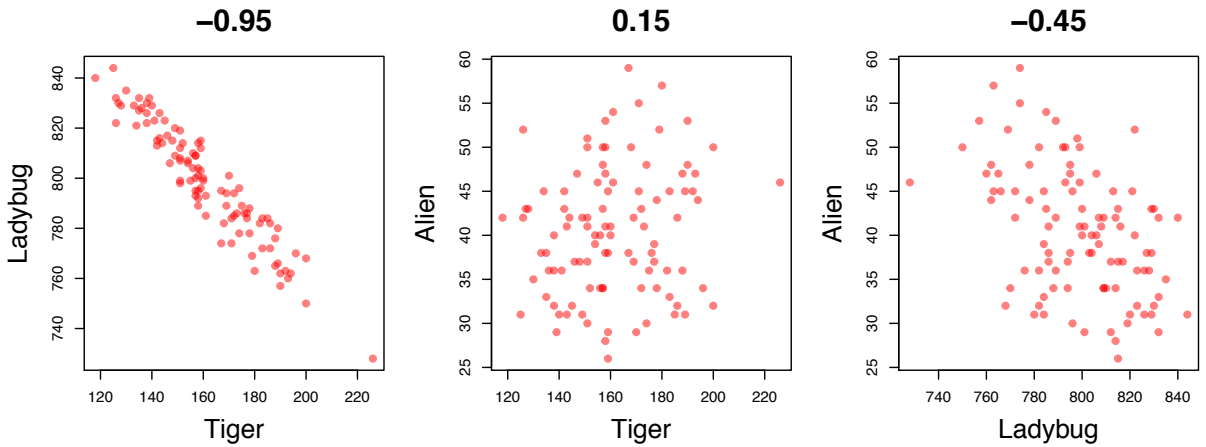
SEQUENCING

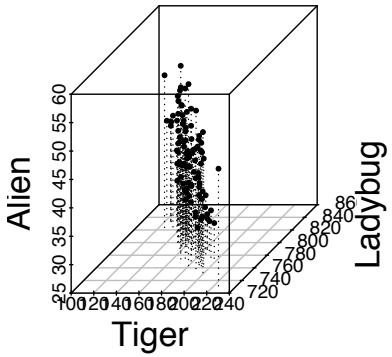


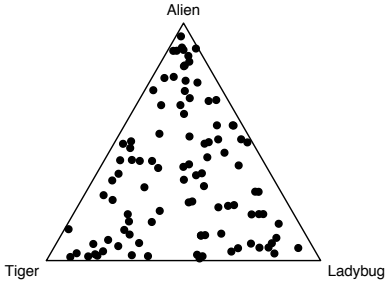
- Sequencing is a constant-sum operation
  - We only get the number of reads that the machine can deliver
- Any constant sum is equivalent

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

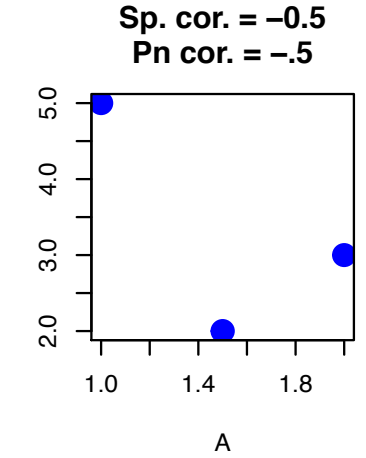
Gloor, et al. 2016. Ann Epidemiology
Gloor, et al. 2016. Can J. Micro

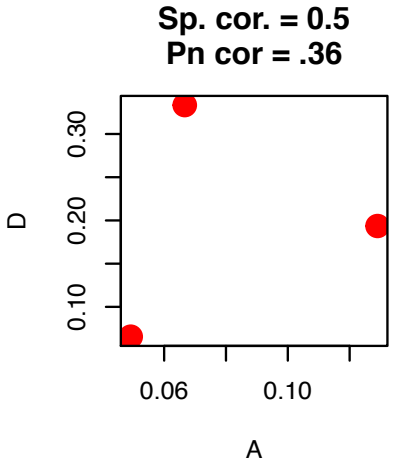
Fundamental problem

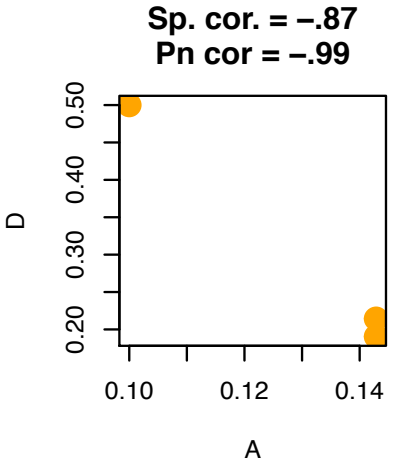
	A	B	C	D	E
S1	1	2	2	5	5
S2	1.5	4	3	2	20
S3	2	8	1	3	1.5

	A	B	C	D	E
S1	.067	.133	.133	.333	.333
S2	.049	.131	.098	.066	.656
S3	.129	.516	.065	.194	.097

	A	B	C	D	E
S1	.1	.2	.2	.5	-
S2	.143	.381	.286	.190	-
S3	.143	.571	.071	.214	-







Spurious correlation

- The correlation observed is not the same for the numerical and proportional data
- The correlation changes again when the proportional data are subset
- **this is spurious correlation** and is an unpredictable correlation observed between two variables whenever they share a common denominator (or a constant sum).
- Spurious correlation arises in **compositional data**

Compositional Data

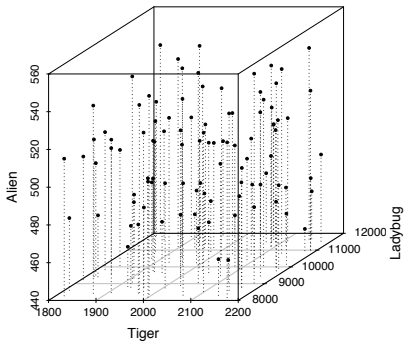
- **Compositional data are logically inconsistent with most tools in common use (Pearson 1897)**
  - correlation, PCA etc.
  - negative binomial, rarefaction,

- **The solution is to work in terms of relative difference (ratios)**
  - Aitchison and others
  - Conceptually similar to qPCR

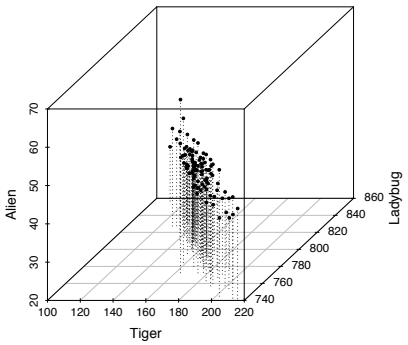
A univariate intuition

CoDa properties & problems

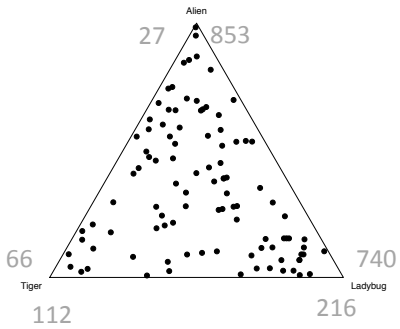
Count



Sequencing



Simplex

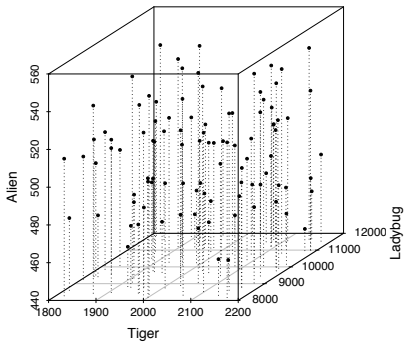


- Simplex: one fewer dimensions than datapoints
  - Scale dependence - hence 'normalization'
  - Correlation and covariation are unreliable
  - Addition and subtraction are not useful operations
    - Subsetting and aggregating are problematic
  - Sparse data is an issue
- Problem remains regardless of dimension
  - It just 'looks' OK

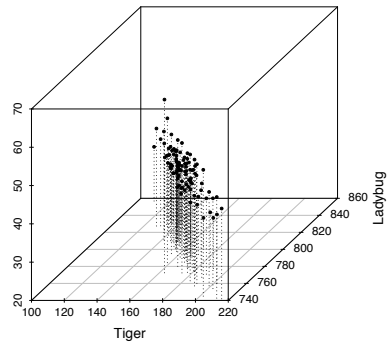
Aitchison 1986. Stat. Anal. Comp Data
Pawlowsky-Glahn, 2015. Mod. Anal. CoDa

Only have ratio information

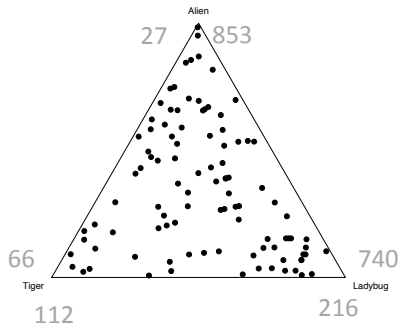
Count



Sequencing



Simplex



$X = [x_1, x_2, \ldots x_D]$, g_X = 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 now logarithms of ratios between parts
  - Abundance is not directly represented
  - Values are now unconstrained, but still on the simplex
- The clr transformation is scale invariant
- Ratios are linearly related
- Must delete, estimate or replace 0 values

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

But my data are not affected …

- Transcriptome
  - Remove rRNA, tRNA, snRNA, etc.
  - Or only mRNA, snRNA, tRNA, etc.
- 16S rRNA 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

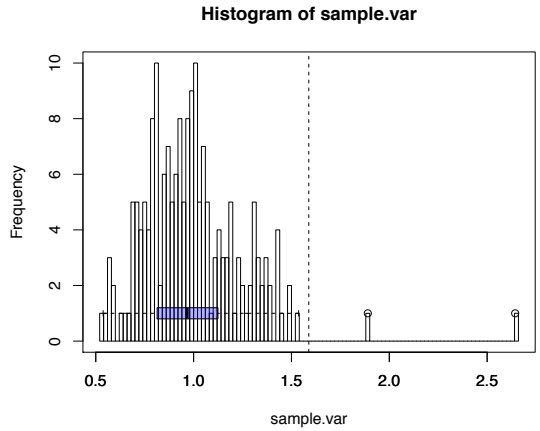
Applying analysis tools based on **variance of the ratios** between parts

- Outliers
  - CoDaSeq (github.com/ggloor)
- Exploratory data analysis
  - Compositional biplots
- Differential abundance
  - ALDEx2 (Bioconductor)
- Compositional association
  - propr (CRAN)
  - CoDaSeq (github.com/ggloor)

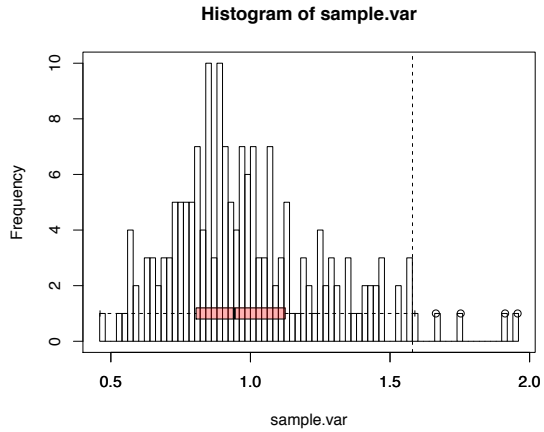
The 0 in the room

- 0 count can be real – or a non-detect
- Sequencing instruments do not deliver counts
  - Observed value is a single estimate of the probability of observing the value given the sequencing depth and the frequency of the molecule in the sample
- Our approach
  - Generate a distribution of probabilities, clr transform, perform analyses and report expected values
  - Dramatically reduces false positives with little or no loss of sensitivity for essentially any seq'omics dataset

Outliers in gene tag sequencing data

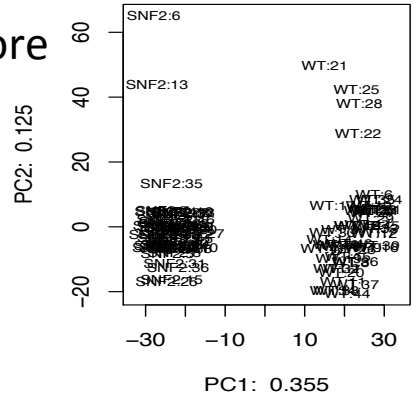


BM

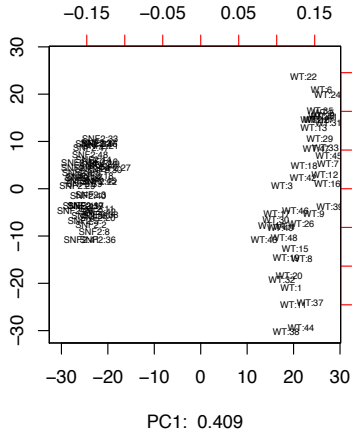


TD

before



after



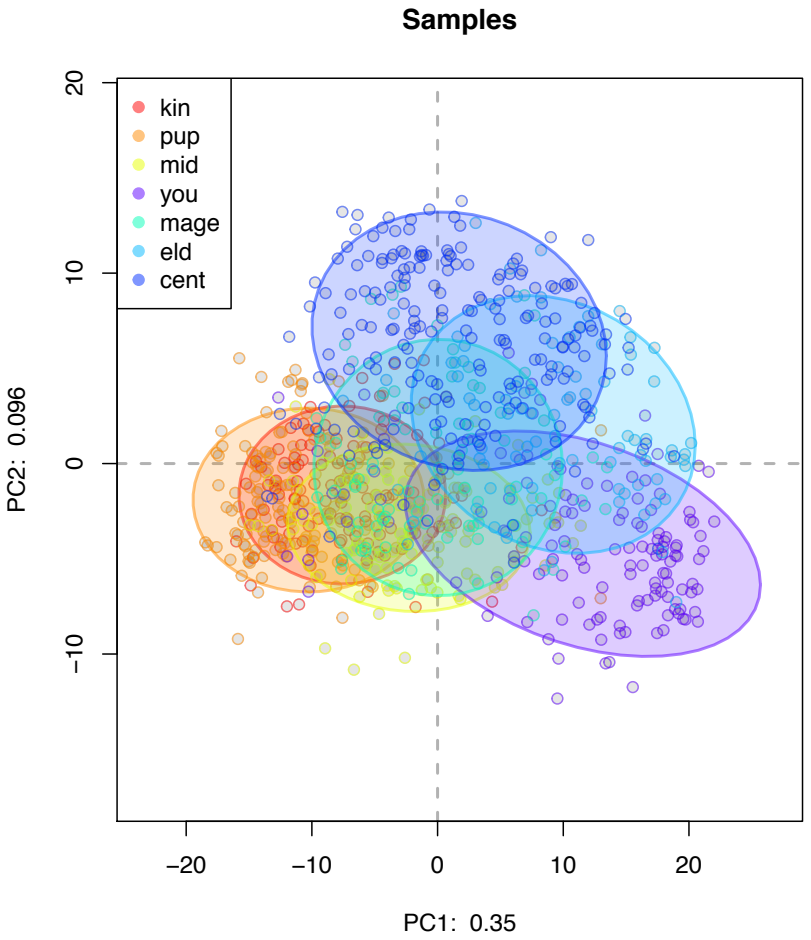
Advantages of tools based on **variance of the ratios** between parts

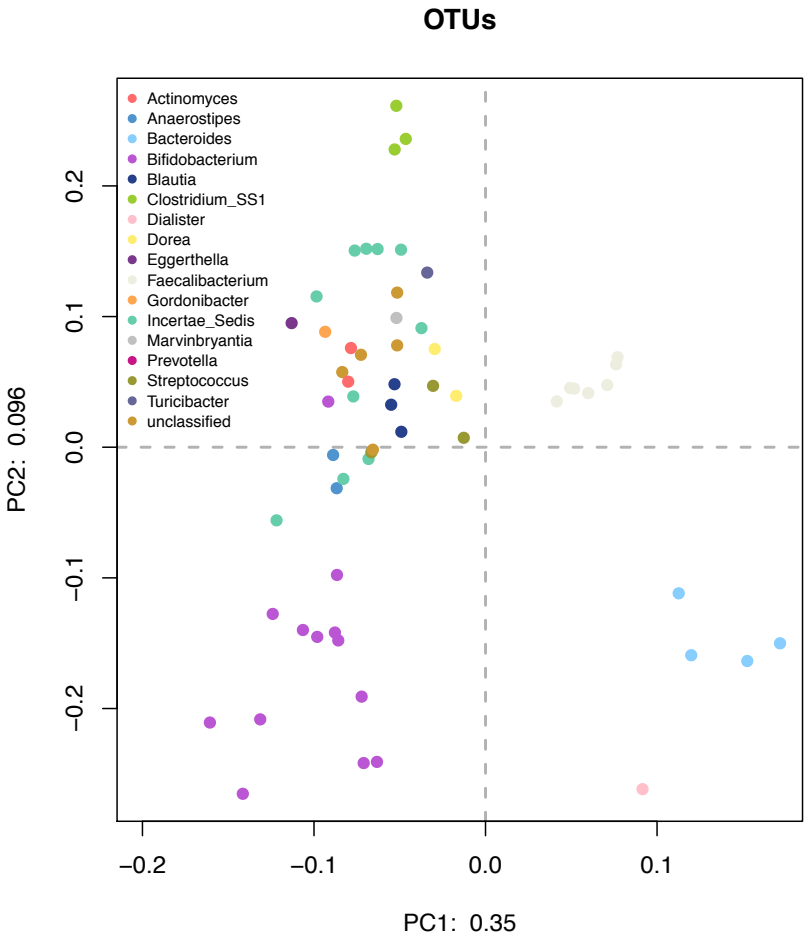
- Examine variance not abundance
- Easily interpret relationships between samples and features
- Robust to perturbations
  - Exploratory data analysis
  - Differential abundance
  - Compositional association

The gut microbiota of healthy aged Chinese is similar to the healthy young

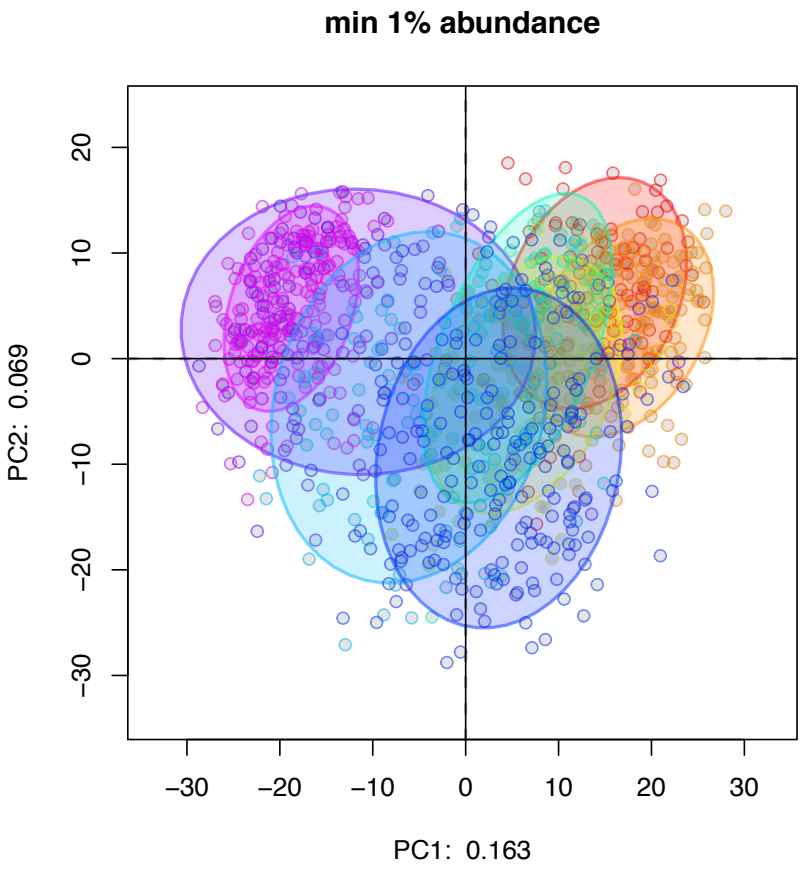
- Examined gut microbiota of over 1000 inter-generationally healthy Chinese
  - Community collected
  - Processed at Tianyi Health Sciences Institute
- Exclusion/Inclusion criteria:
  - Smoking, drinking, mood disorders, no chronic disease, no prescription medication or antibiotics for last 3 months, no personal or family disease history (cardiovascular, gastrointestinal, metabolic, neurological, respiratory, cancer)
  - Willing to donate sample, parents and grandparents lived to >80 yo (if under age 30)

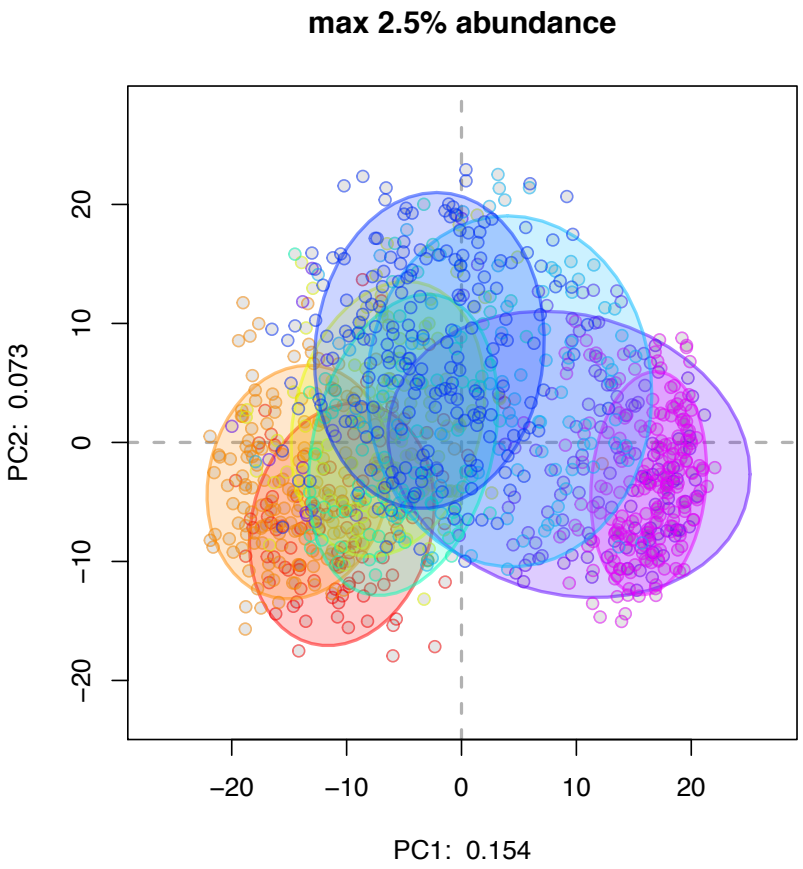
The gut microbiota of healthy aged Chinese is similar to the healthy young



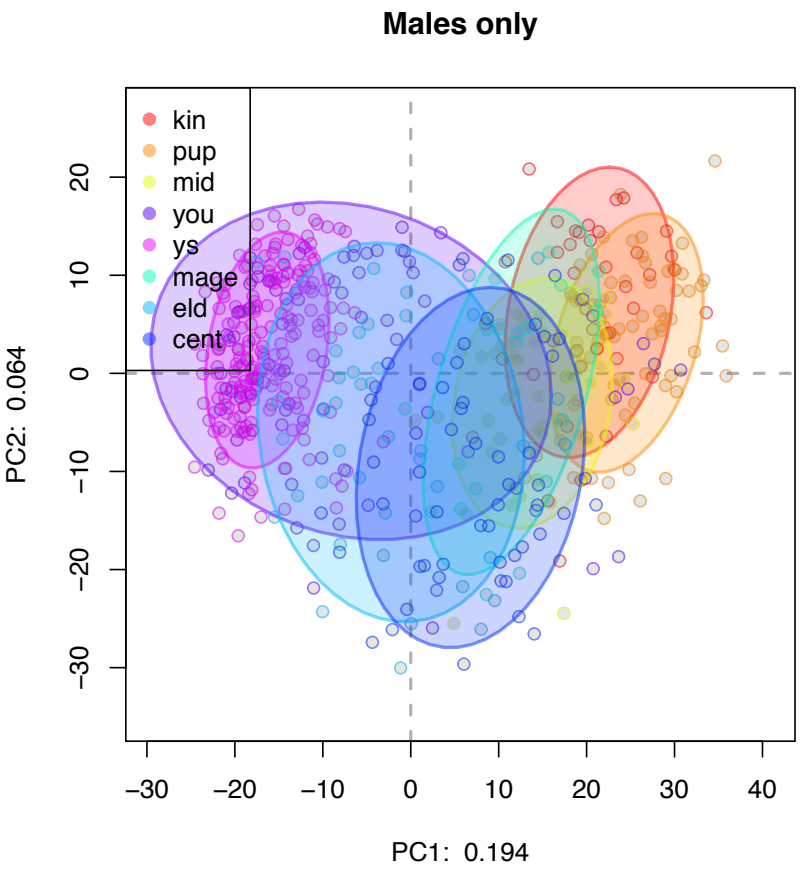


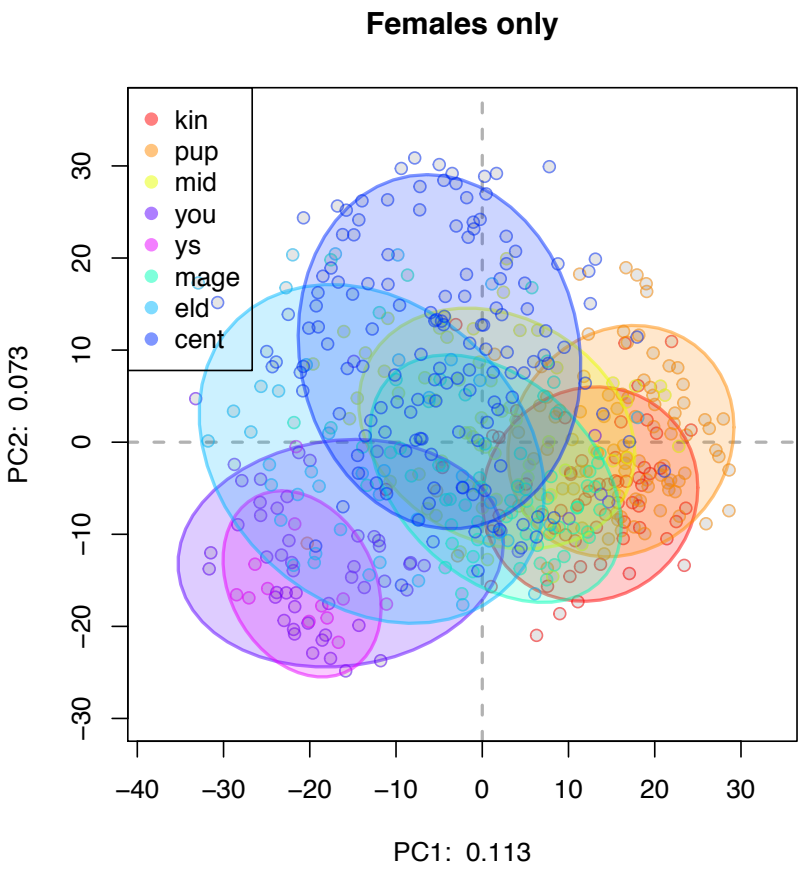
Imperturbable





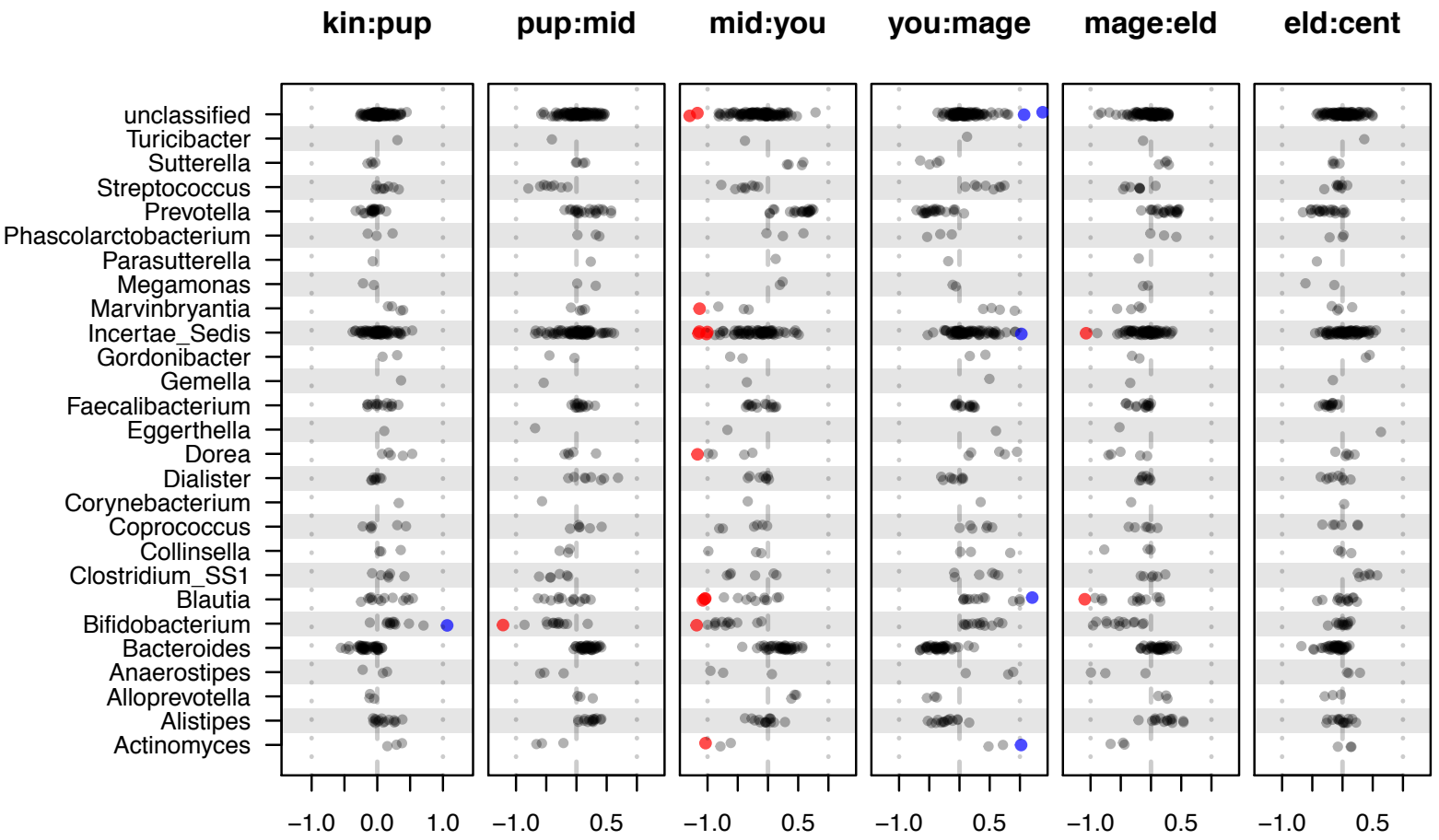
Sex independent



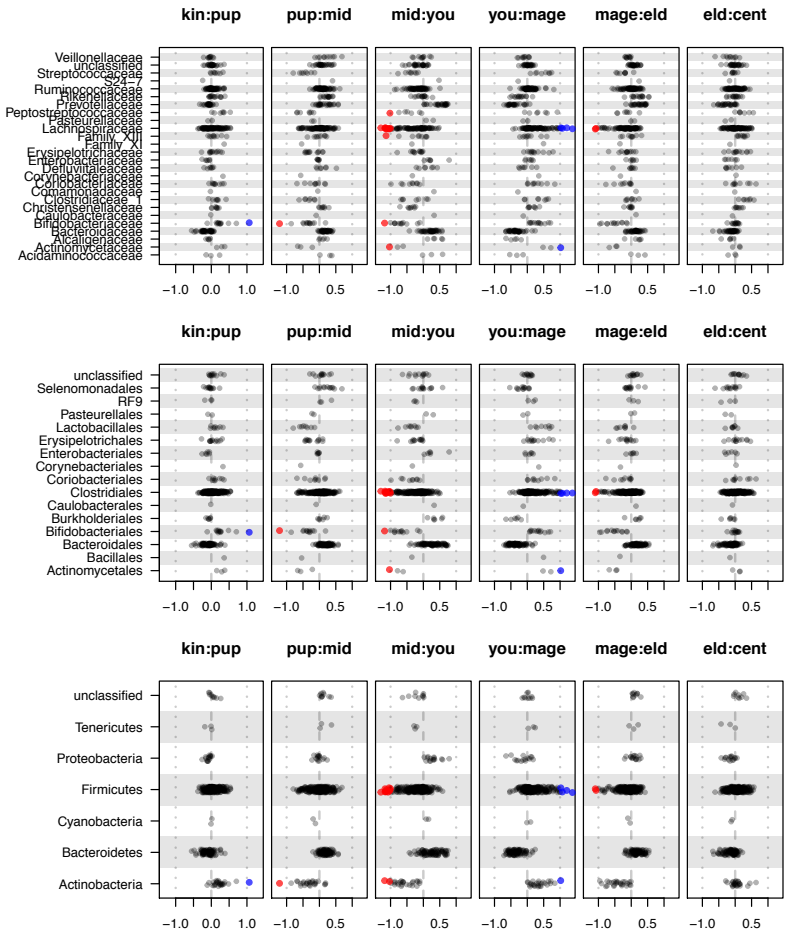


OTUs vs. Genera

ALDEx2



OTUs vs. other levels



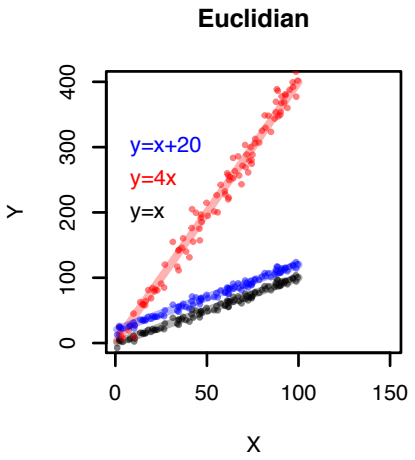
Association between individual OTUs and group membership generally weaker with larger group

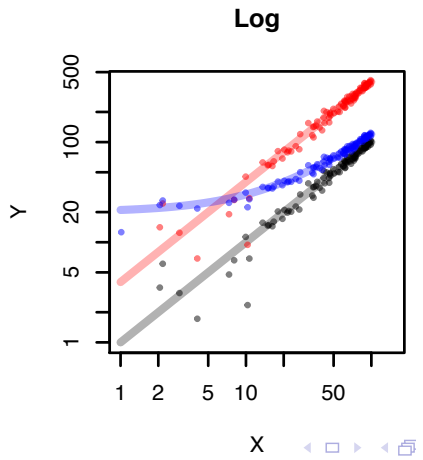
Bacteriodetes and Actinobacteria most cohesive phyla, and in opposite direction. Largely *Bacteriodes* + *Prevotella* and *Bifidobacteria* moving the pile.

Fake Correlations!!!

In addition to spurious correlation

False positive –ve correlations because …
what is the correlation between heads and tails?

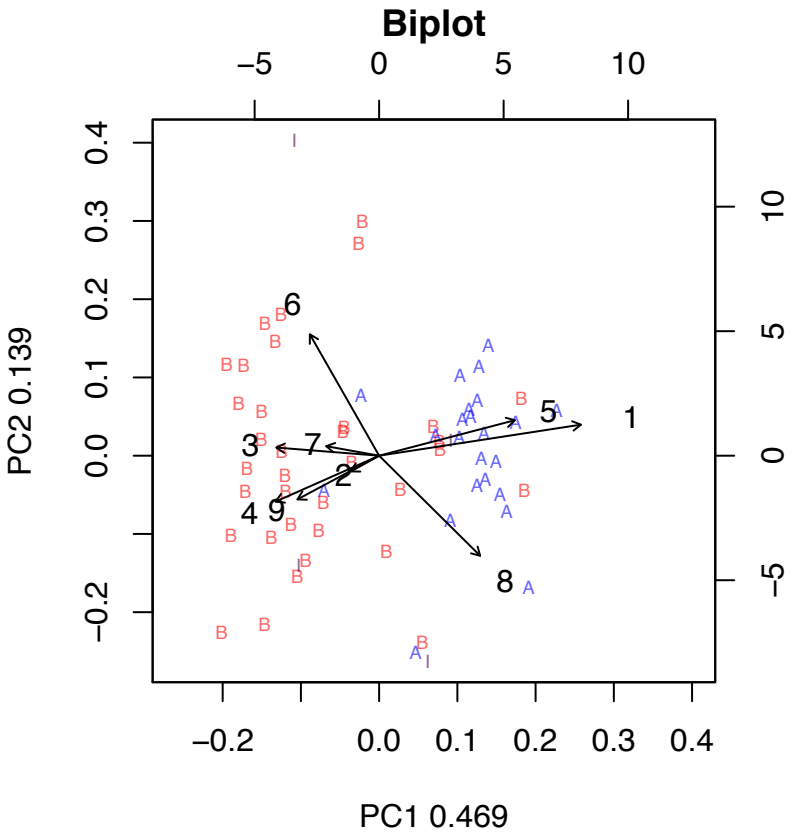


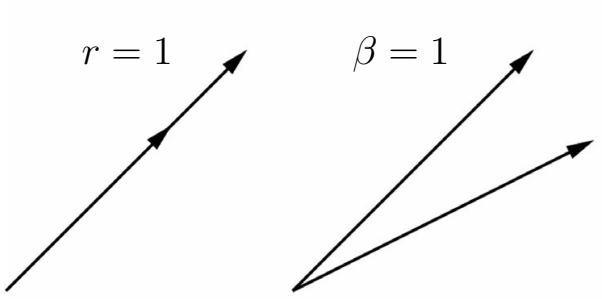


False positive +ve correlations because associations must maintain a constant ratio …

The Ø metric

propr, CoDaSeq



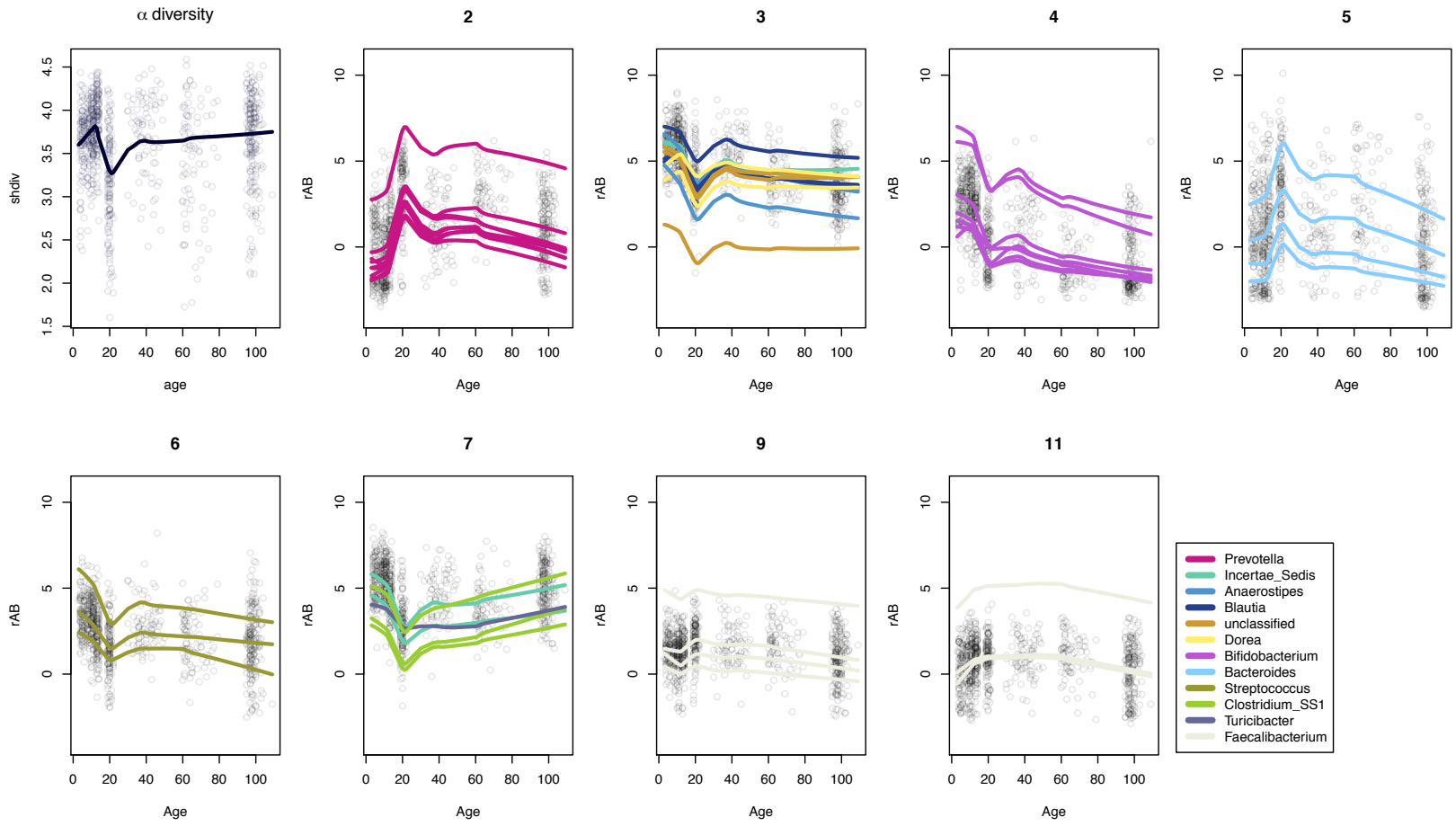


- $Ø = 1 + ß^2 - 2ß|r|$
- r=1 when two vectors in same direction
- ß=1 when the two vectors have the same variance
- So Ø = 0 iff r=2 & ß=1
  - Collapsing two measures into one number

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

Association

Ø<0.23



Summary

Analysis	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 least variance
Interpretation	Seems simple but is not	Seems hard but is not

Conclusions

- We only have ratio 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
  - Compositional PCA ~ ALDEx2 ~ Ø
- 0 replacement is a theoretical, but not a practical impediment

Reading & Sources

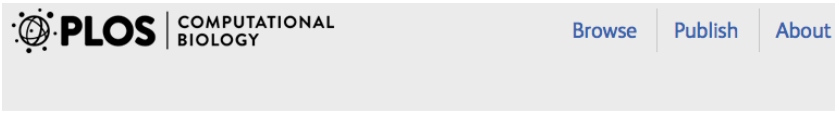
METHODOLOGY | OPEN ACCESS

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



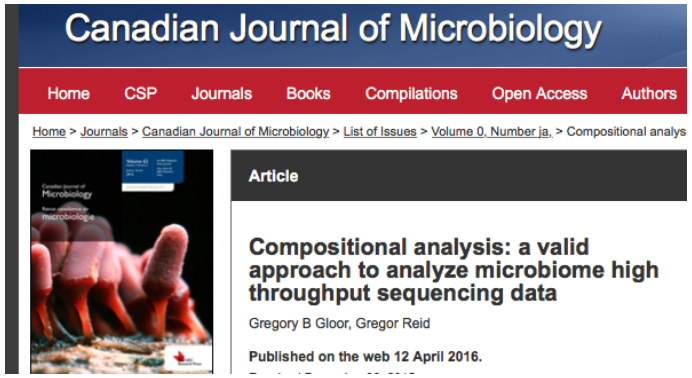
OPEN ACCESS PEER-REVIEWED

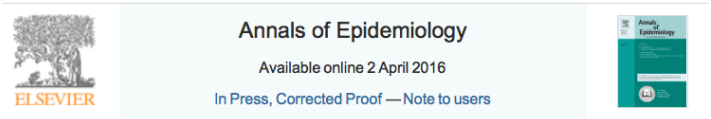
RESEARCH ARTICLE

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





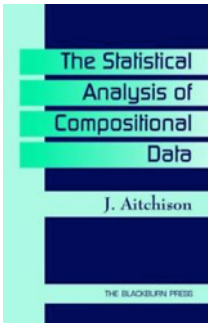
Review article

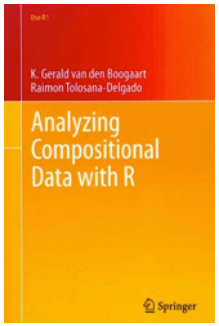
It's all relative: analyzing microbiome data as compositions

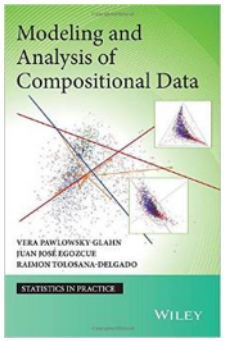
Gregory B. Gloor, PhD[a], Jia Rong Wu, BSc[a], Vera Pawlowsky-Glahn, PhD[b], Juan José Egozcue, PhD[c]

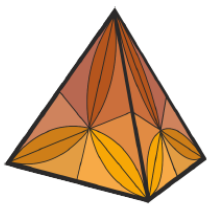
Show more

doi:10.1016/j.annepidem.2016.03.003 Get rights and content



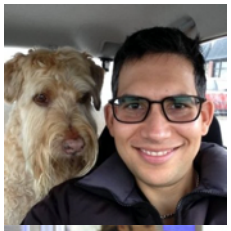




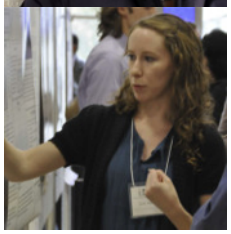


Acknowledgments

Canadian Centre for Human Microbiome and Probiotic Research



Andrew Fernandes



Jean Macklaim

CoDa

Vera Pawlowsky-Glahn
Juan Jose Egozcue

Justin Silverberg

Gregor Reid
Jeremy Burton
Michael Vu
Jia Rong Wu

NSERC CRSNG

People. Discovery. Innovat

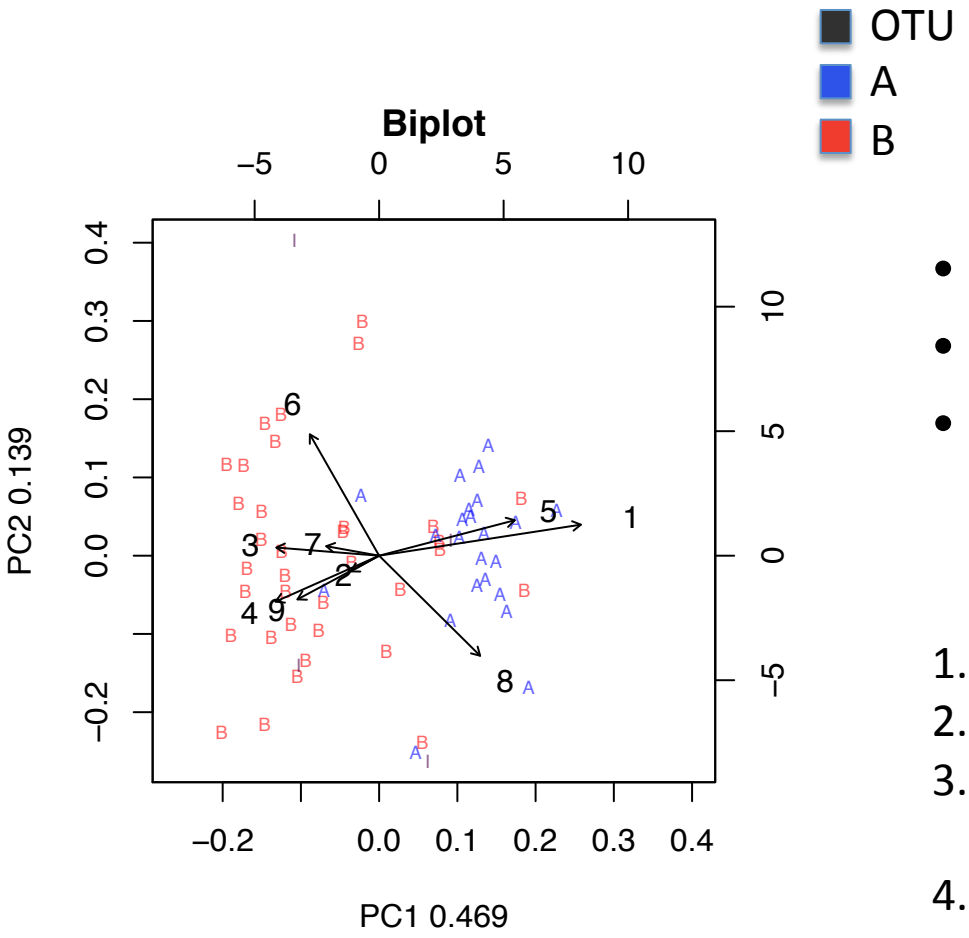
Vogue
vaginal microbiome group initiative
Advancing Women's Health through Microbiome Research

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

We are working in the wrong geometry

- 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

Exploration: CoDa PCA biplot



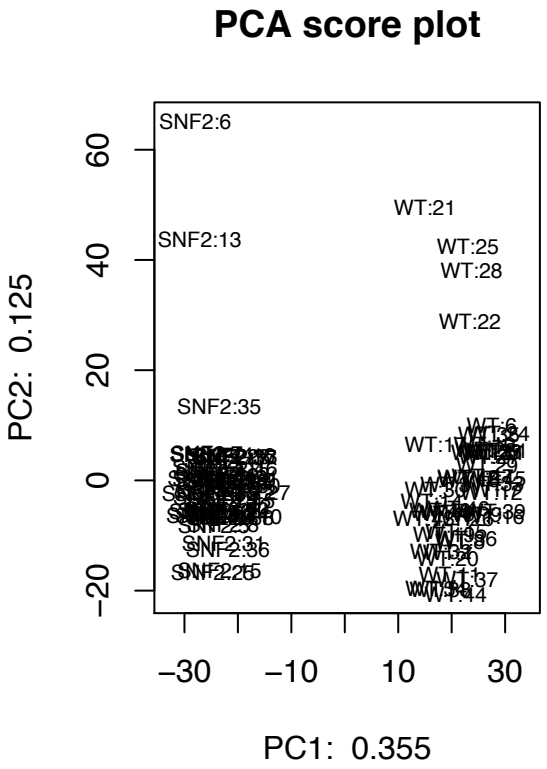
Data Structure

- Exploratory Tool
- Samples + Variables after clr
- SVD is now legal
  - But is on ratios
  - Links are informative

1. Distance from origin ~ SD
2. Links ~ ratio abundance
3. Links with multiple tips ~ linear dependence
4. Cos angle between links ~ correlation of the parts

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

Outliers

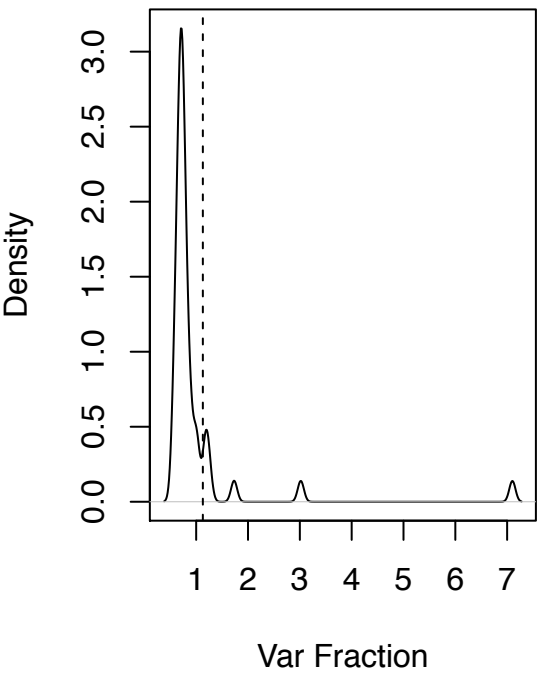


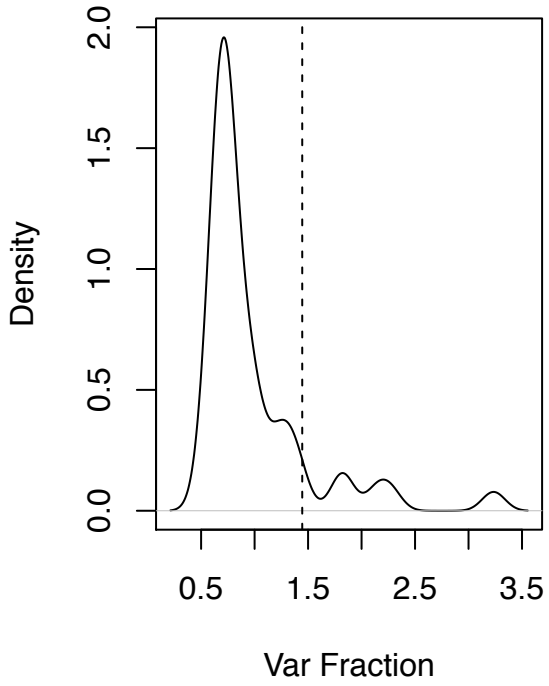
Yeast transcriptome dataset

- 2 conditions WT, ΔSNF2
  - 48 replicates
  - Tightly controlled conditions
  - All should be very similar
- PCA plot of scores shows two major things
  - The two conditions split
  - There is considerable spread within conditions
- Finding and removing outliers may improve our resolution

Data from Schurch, 2015

Outliers

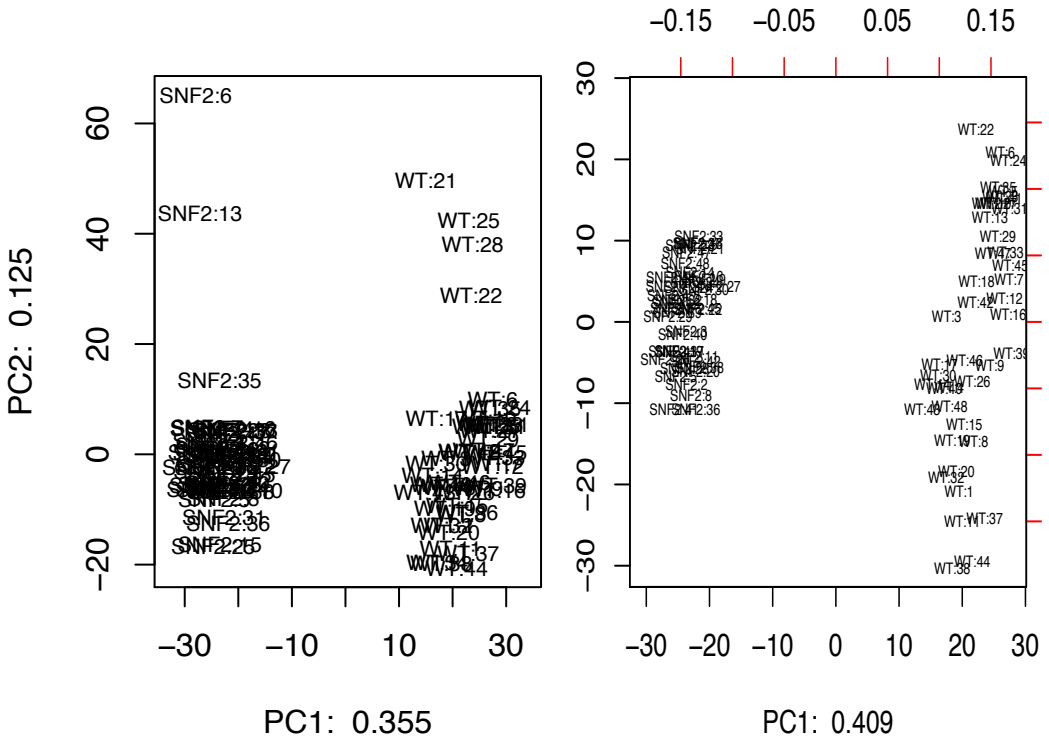




Yeast transcriptome dataset

- Contribution to total variance?
  - Density shows approximate normal with tails
  - Outliers often defined as those beyond twice mean + IQR
  - Removes 7 SNF2, 5 WT samples
    - Same set as removed by Schurch et al, with a more complex linear combination model of correlation, X^2, and fraction of 'bad' gene levels
  - Simple, quick approach

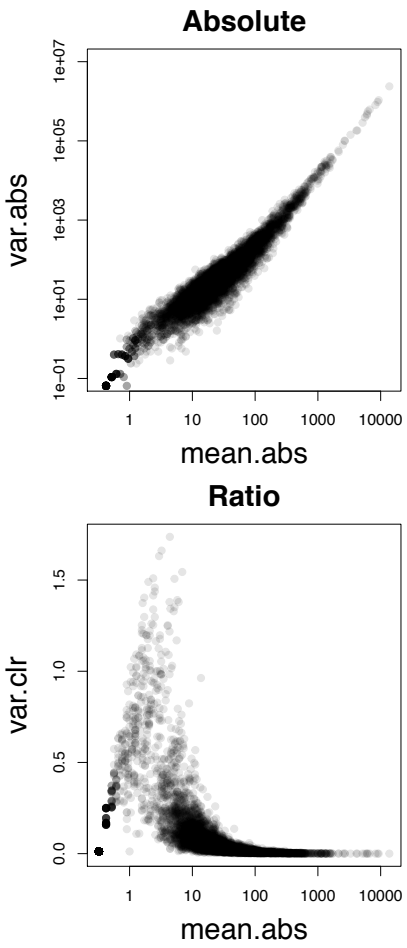
Outliers



Yeast transcriptome dataset

- More variance explained on component 1, less on component 2.
  - Reduced component 2 variance by half
- True spread of the data is easier to see
  - WT is more variable than SNF2

Where is the noise?

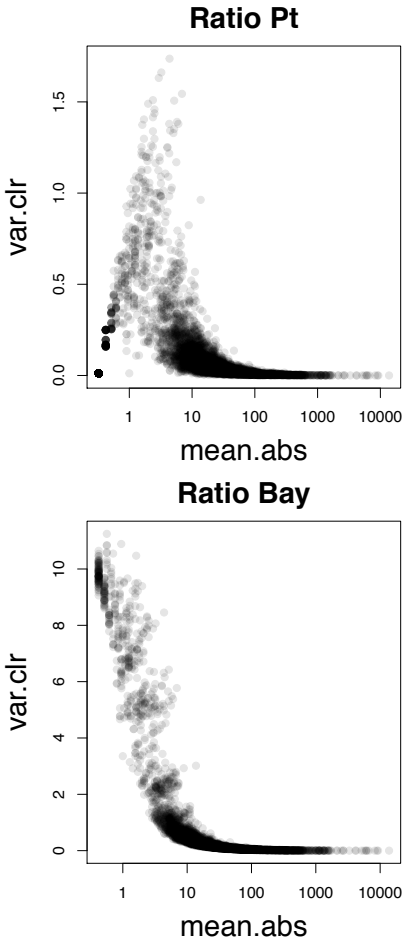


Variance vs. mean for data on the simplex

- Transcriptome datset
- 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

Fernandes, et al. 2013. PLoS ONE
Gloor et al. 2016. Aus. J. Statistics

Bayesian estimate

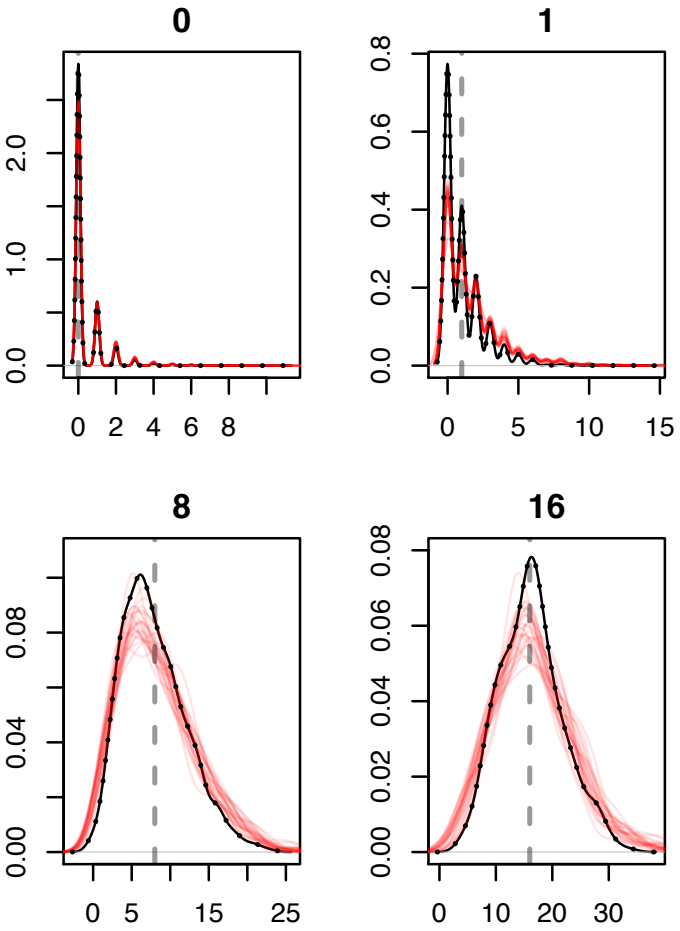


Variance vs. mean for data on the simplex

- 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

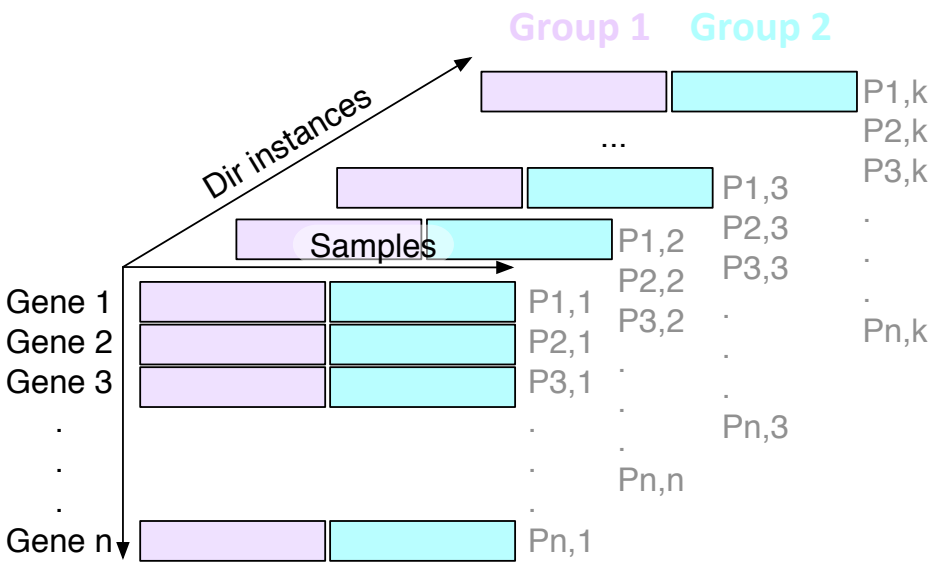


Distribution of values in replicate B for replicate A values

- 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



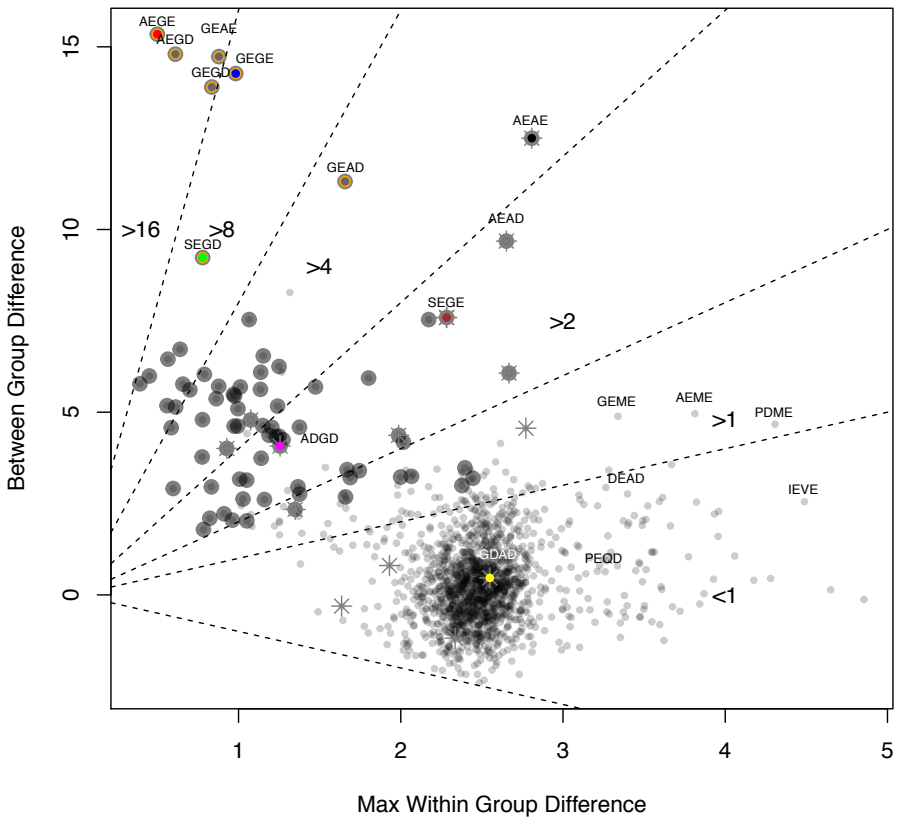
ALDEx2

- 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

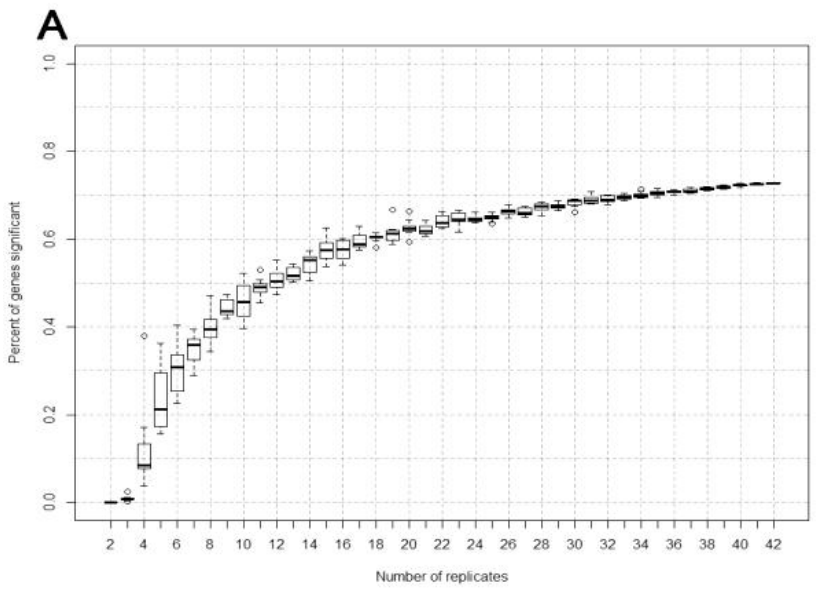
Effect-sizes

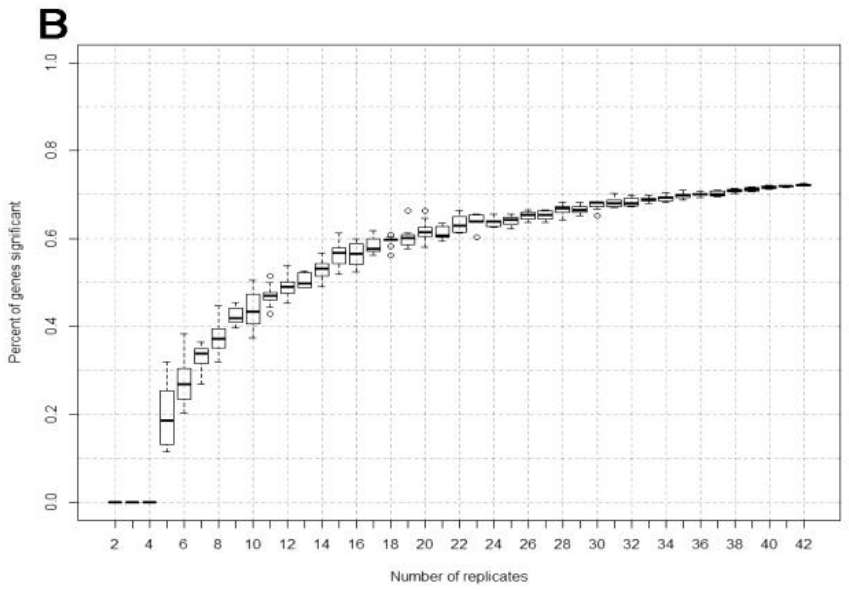


- E=Difference/Dispersion
- SELEX dataset
- Convenient way to summarize univariate differences in multivariate data
- X axis - maximum variation within either group for each feature
- Y axis – Difference between groups for each feature
- Lines indicate effect size boundaries
  - D=1, V=2 E=0.5
  - D=1 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

Gloor et al. 2016. J. Comp. Graph. Stat.

P-values are not reliable





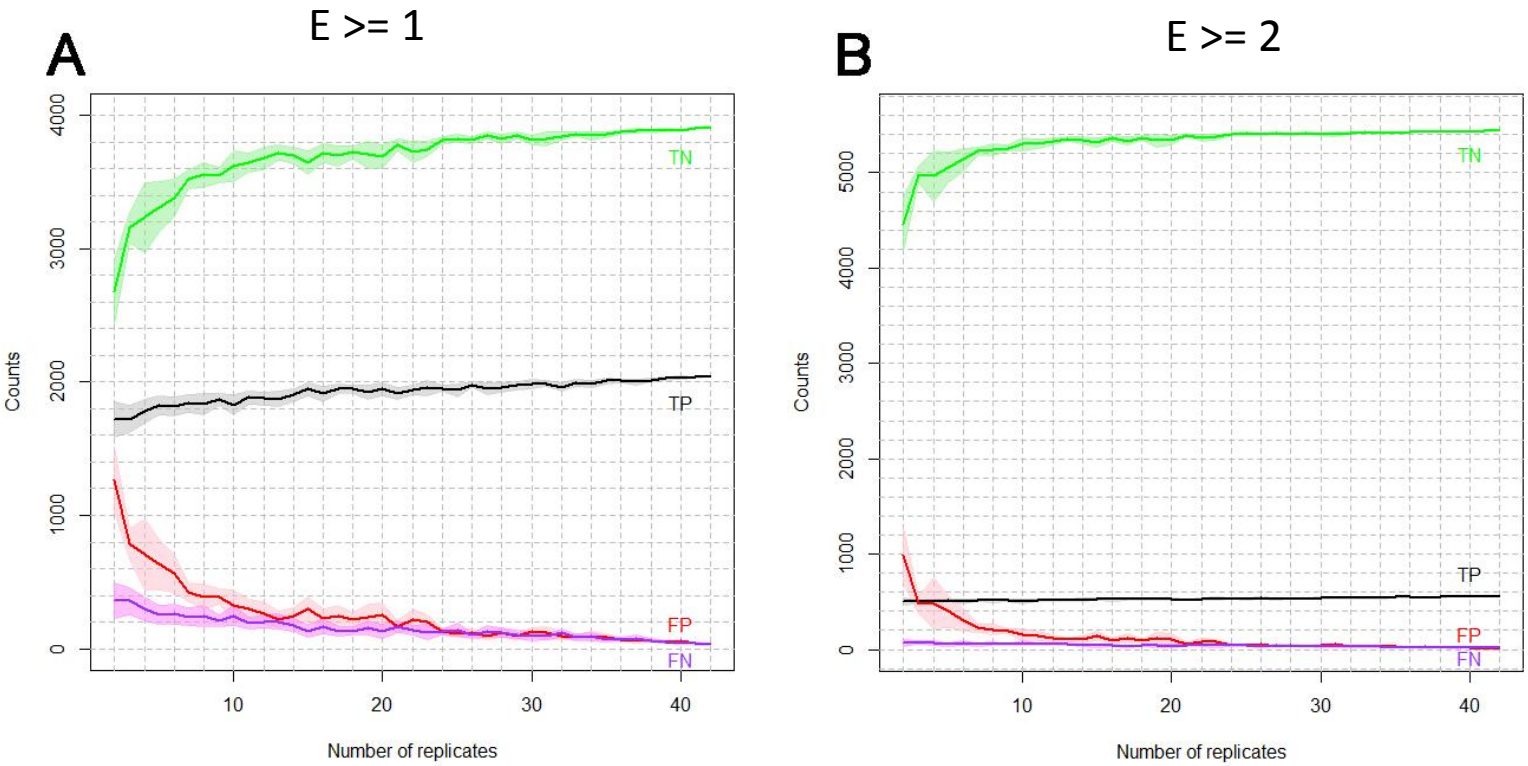
Yeast transcriptome dataset

- 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

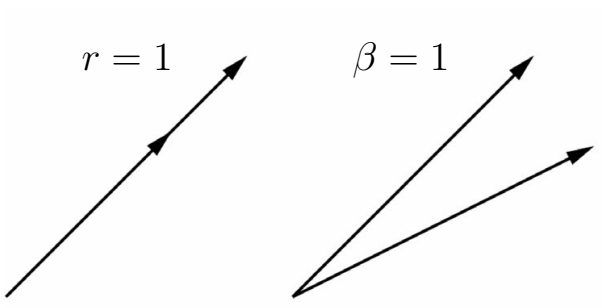
Effect sizes are stable estimates



Halsey, Nature Meth. 2015

The Ø metric

- Ø = 1 + ß2 -2ßr



  - x_i = clr(x_i), x_j = clr($_{xj}$)

  - $Ø_{ij}$ = (var(x_i) – var(x_j)) / (var(x_i) + var(x_j))

Lovell, PLoS Comp. Bio. 2015