Accounting for asymmetry and batch effects in meta-transcriptomics

Greg Gloor
U. Western Ontario
Department of Biochemistry
ggloor@uwo.ca : @gbgloor
github.com/ggloor/compositions/presentations

Outline

- The process of high-throughput sequencing
- The data we get vs. what we think we have
- An introduction to ALDEx2
- Example analysis of a meta-transcriptome

Acknowledgements

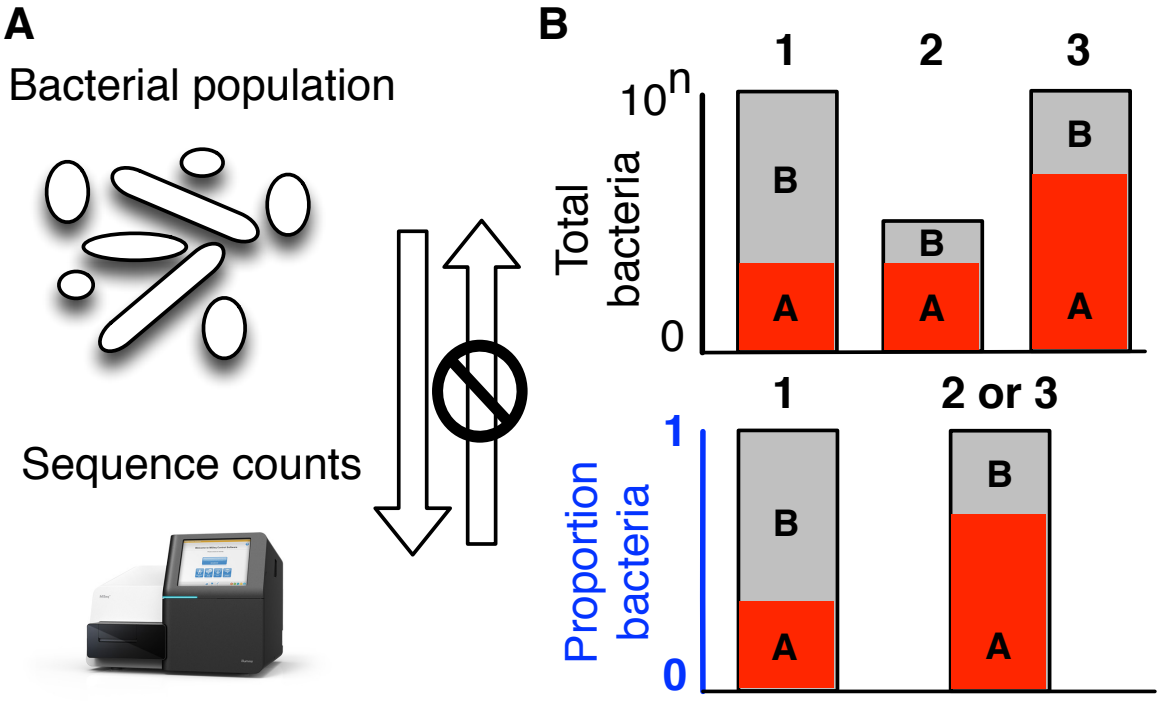
Andrew Fernandes
Lead Data Scientist
FICO

Jean Macklaim
Bioinformagician
DNA Genotek

Jia R. Wu
PhD student
U. Waterloo

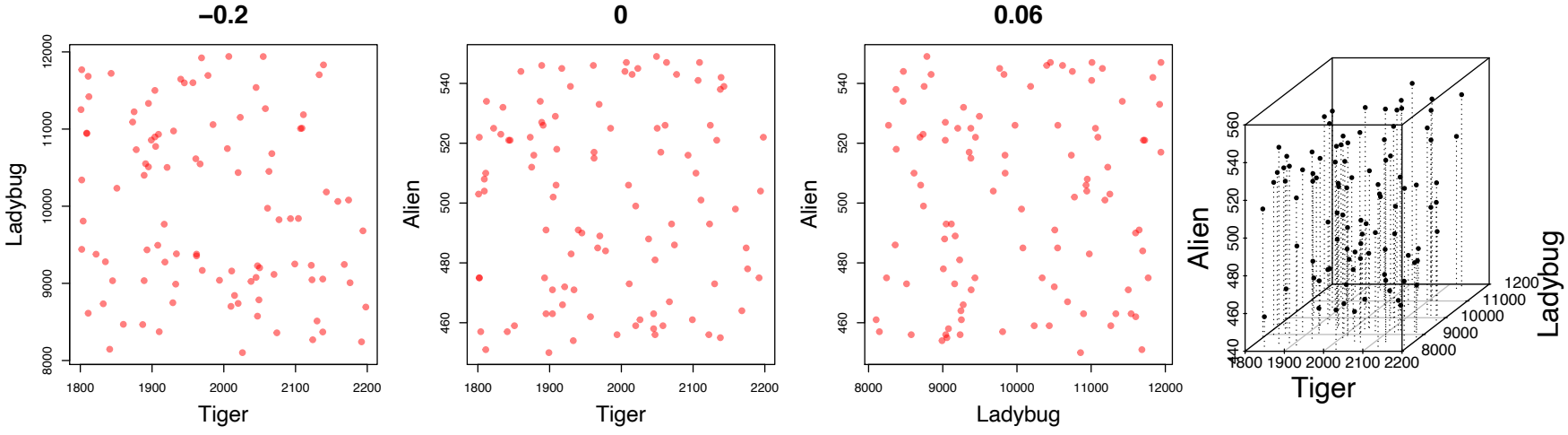
Thom Quinn
PhD student
Deakin U.

The sequencing meat grinder



Gloor et al Front. Micro. 2017

The machine delivers ‘counts’

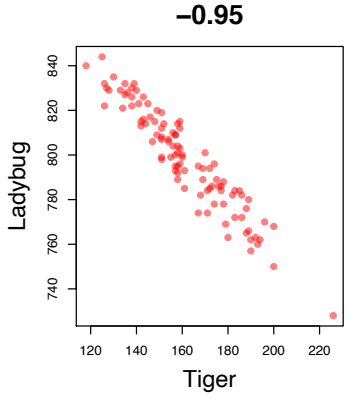


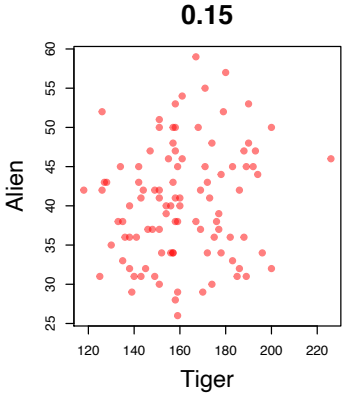
Example 100 random sample sets

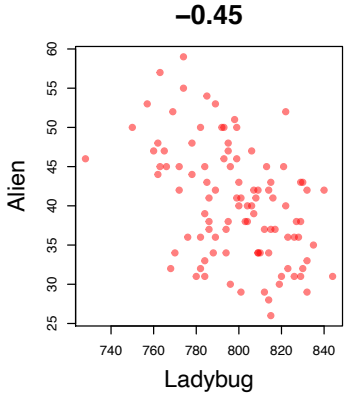
- 🐯 range=1800-2200
- 🐞 range= 8000-12000
- 👽 range=450-550

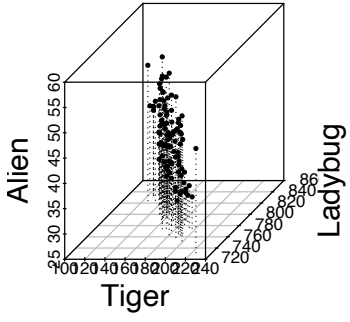
Gloor et al Front. Micro. 2017

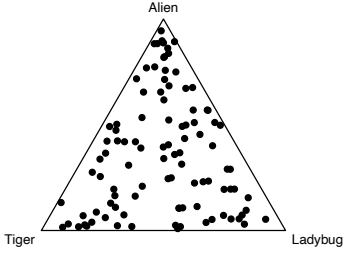
But these ‘counts’ have a maximum sum











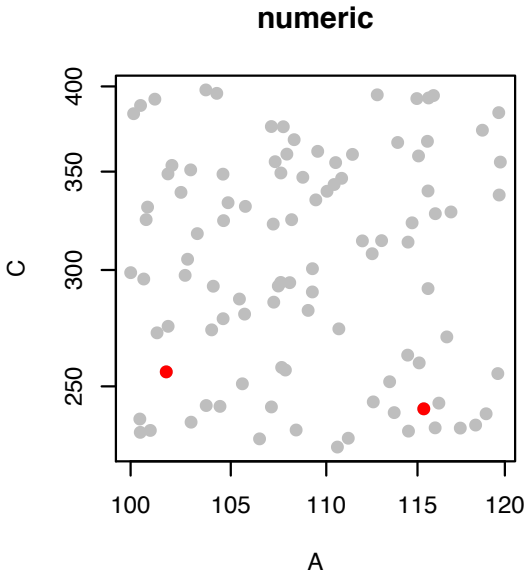
<u>Constant sum of 1000</u>

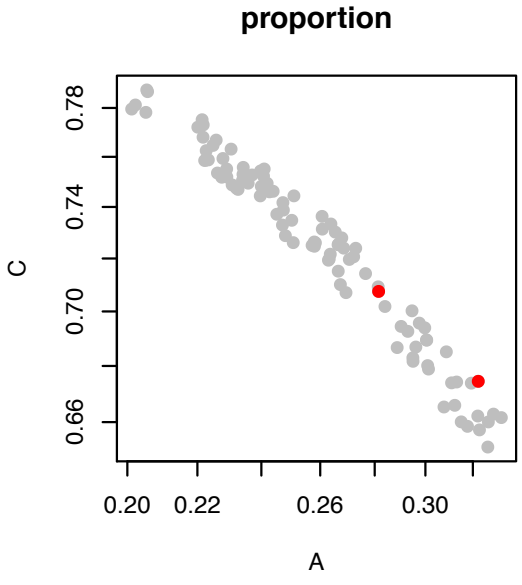
Constant sum operations:

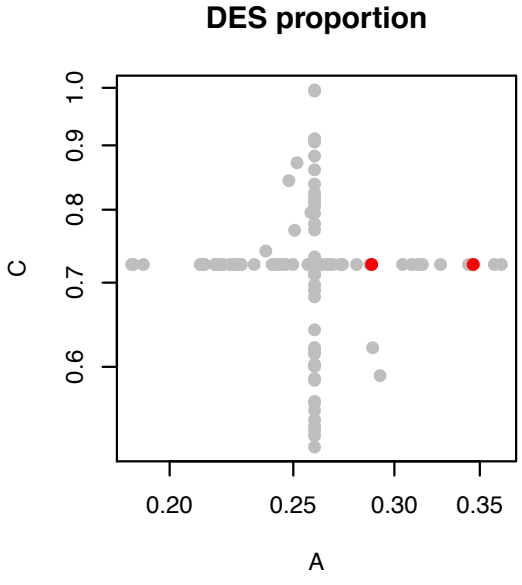
- Count normalization
- Rarefaction
- Proportion
- percentage, relative abundance
- Sequencing

Gloor et al Front. Micro. 2017

Normalization is no help at all



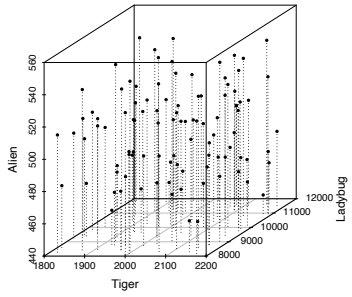




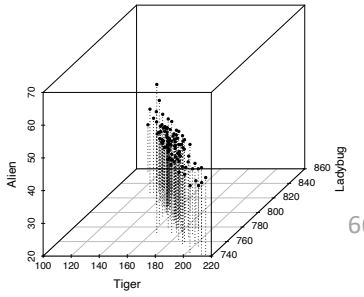
Once data are on the simplex, they cannot be removed without additional information

Log-ratios for CoDa

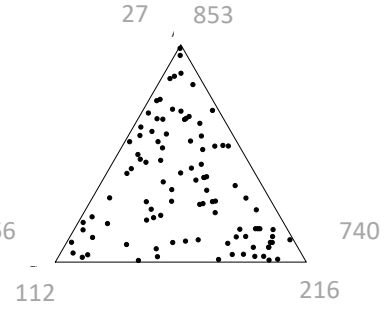
Count



Sequencing



Simplex



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

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

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

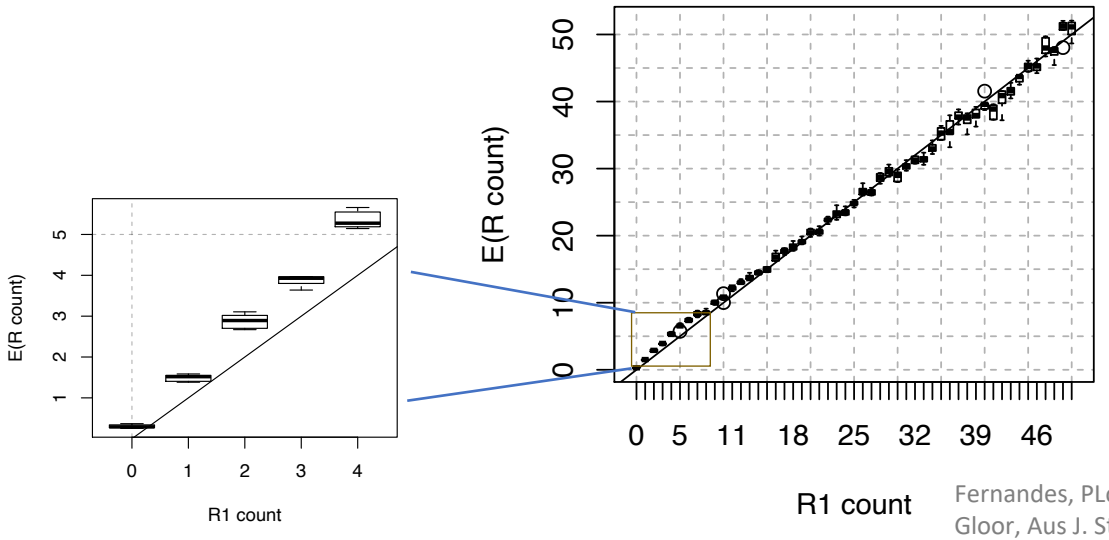
Aitchison 1986. Stat. Anal. Comp Data
Pawlowsky-Glahn, 2015. Mod. Anal. CoDa

The 0 in the room

- 0 count can be real – or a non-detect
- 0 count is often not 0 probability to detect the count
- Sequencing instruments deliver integer scaled probabilities: ‘counts’

$X=[x_1,x_2,x_i \dots x_D] = [p_1,p_2,p_i \dots p_D]\alpha$; $\alpha=\mathbf{\Sigma}X$; $\mathbf{\Sigma}P=1$

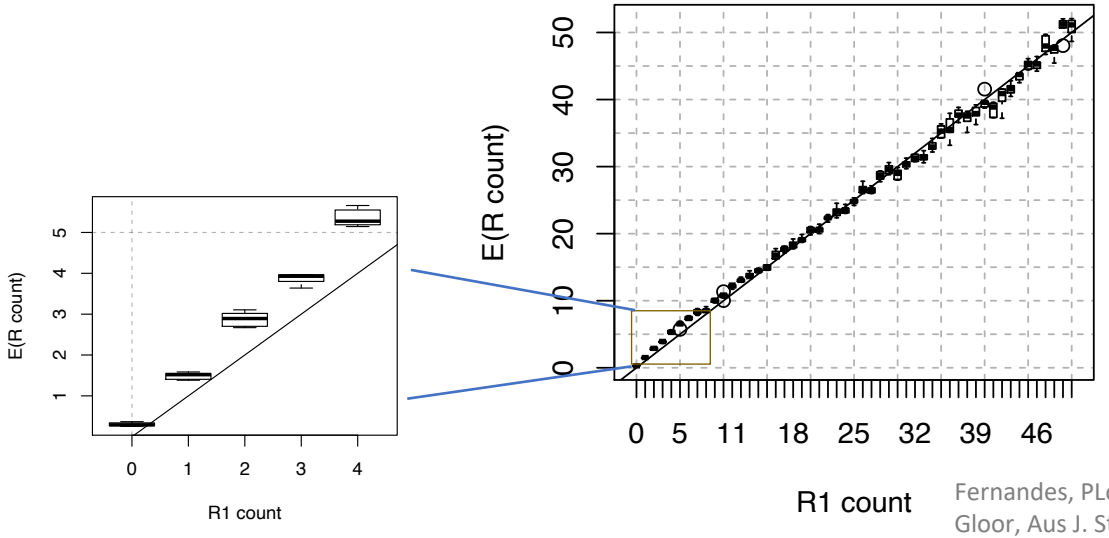
$x_i = (P_i \mid \mathcal{f}_i)\alpha$



Fernandes, PLoS ONE 2013
Gloor, Aus J. Stat. 2016

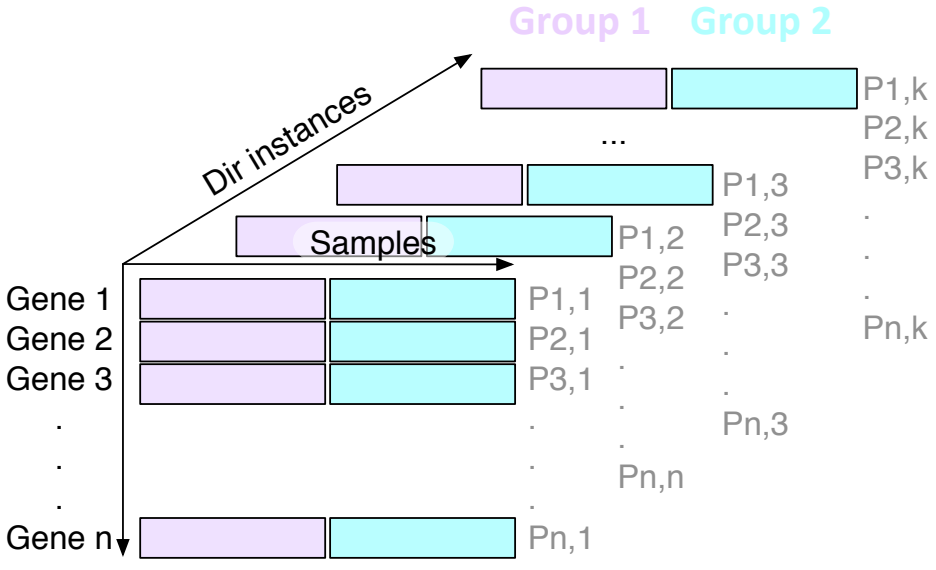
ALDEx2

- $p_i = 0$ is strong, invalid, assumption
- $[P_i > 0, P_i \rightarrow 0]$
- Model the expected value



Fernandes, PLoS ONE 2013
Gloor, Aus J. Stat. 2016

ALDEx2: probabilistic CoDa

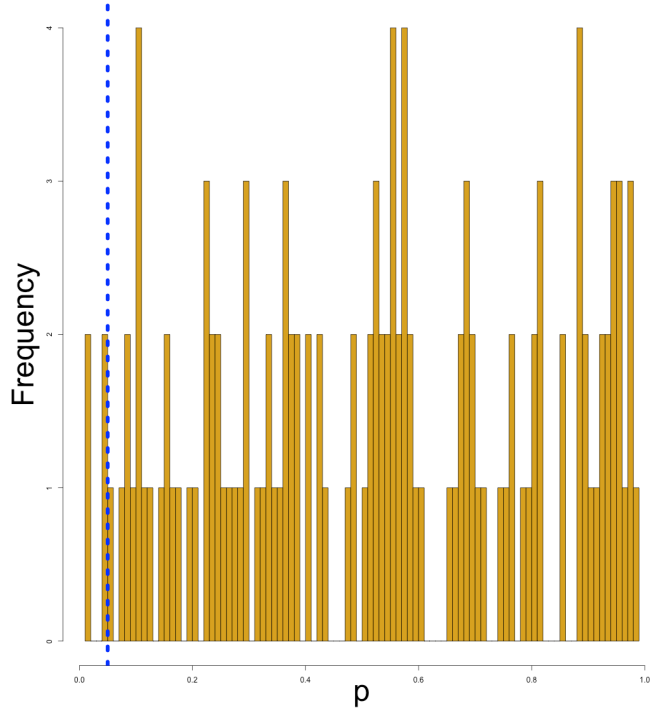


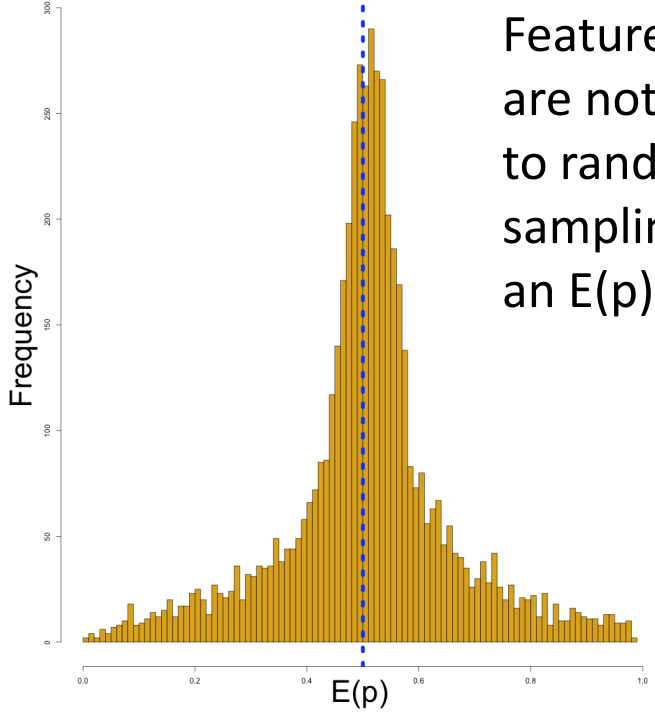
- Generate posterior estimates of the data consistent with the observed data and the chosen prior(s)
  - Dirichlet instances
- clr transform instances
- Calculate test values on each instance
- Correct for FDR

- **Report the expected value of each test for all instances**
- Dramatically reduces false positives with little or no loss of sensitivity for essentially any seq*omics dataset

Fernandes, et al. 2013. PLoS ONE
Fernandes, et al. 2014. Microbiome
Thorsen, Microbiome 2016

ALDEx2 Expected value





Features that are not robust to random sampling have an E(p) of ~0.5

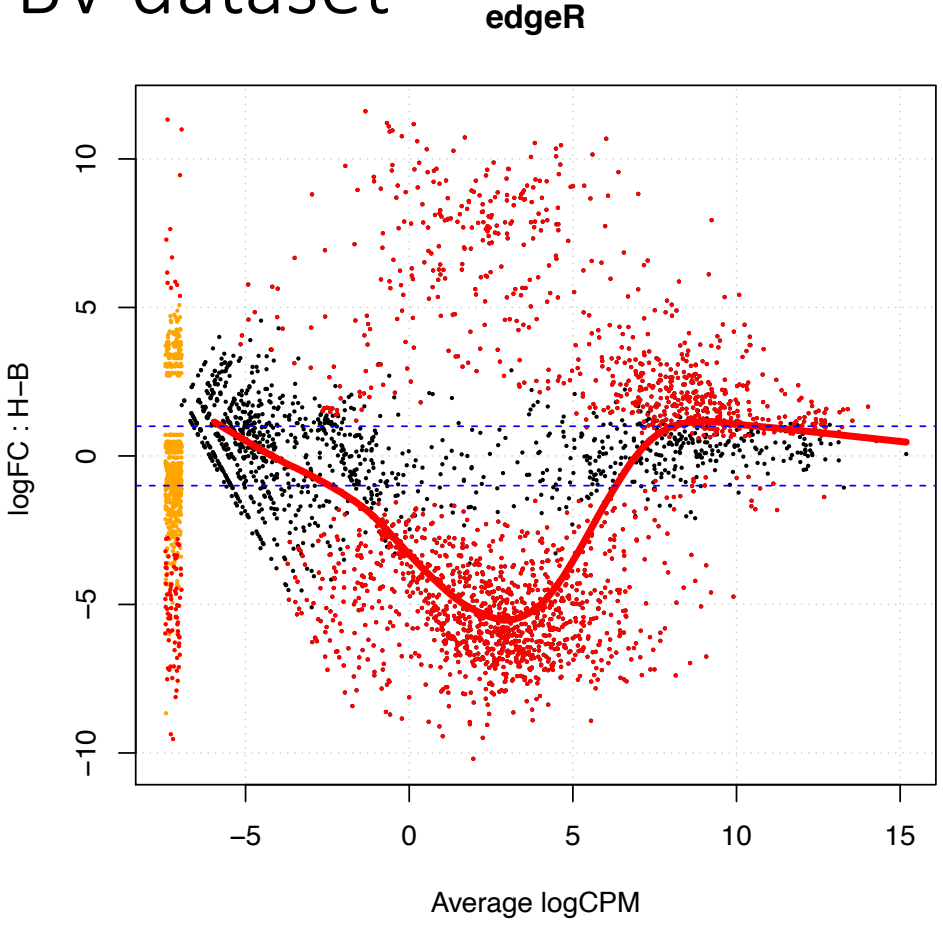
Meta-transcriptomics of an ecosystem

- Sequencing all the mRNA in a collection of bacteria
- Often unbalanced
  - Different conditions can have different taxonomic compositions
  - Both absolute and relative abundance of the taxa and their transcripts can change
- Relative abundance does not necessarily reflect environmental reality in sequencing data

Bacterial vaginosis

- Most common vaginal dysbiosis
  - H is predominantly *Lactobacillus* sp.
  - BV is mixed bag of anaerobes with *L. iners*
- Marked asymmetry in composition
  - Asymmetrically sparse functions
  - Asymmetrically expressed functions
  - Group genes to functional level (SEED, KEGG)
- If everything is different, then nothing is important
- We *must assume* something is invariant

BV dataset



Bland-Altmann Plot

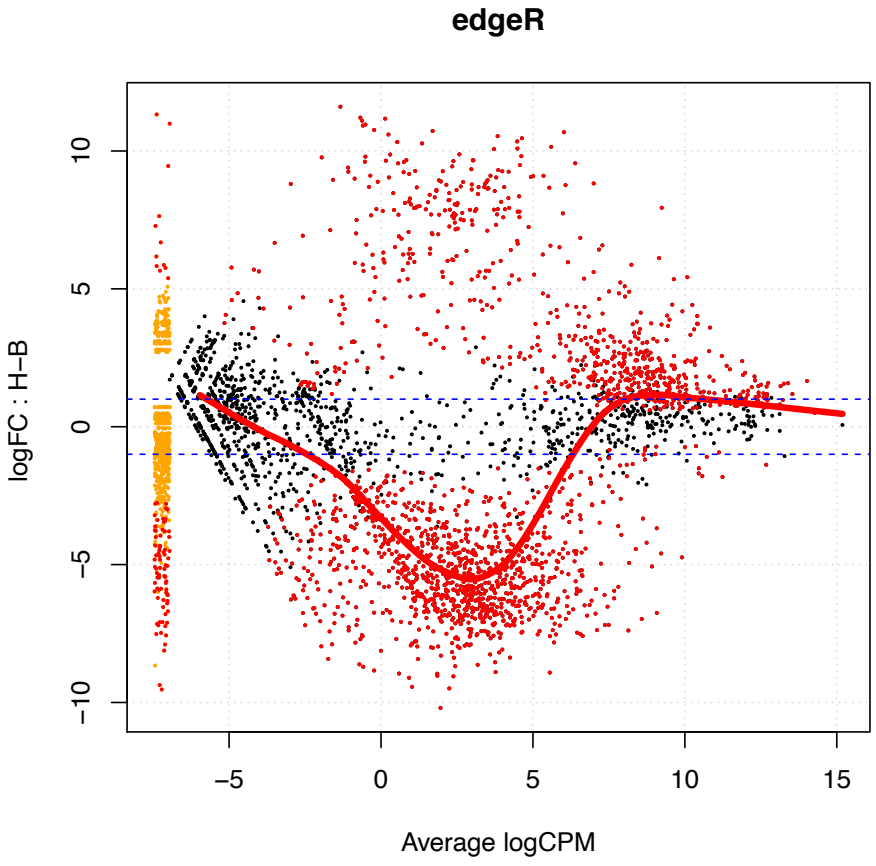
- **Each function is a point**
- **Red are 'DE'**

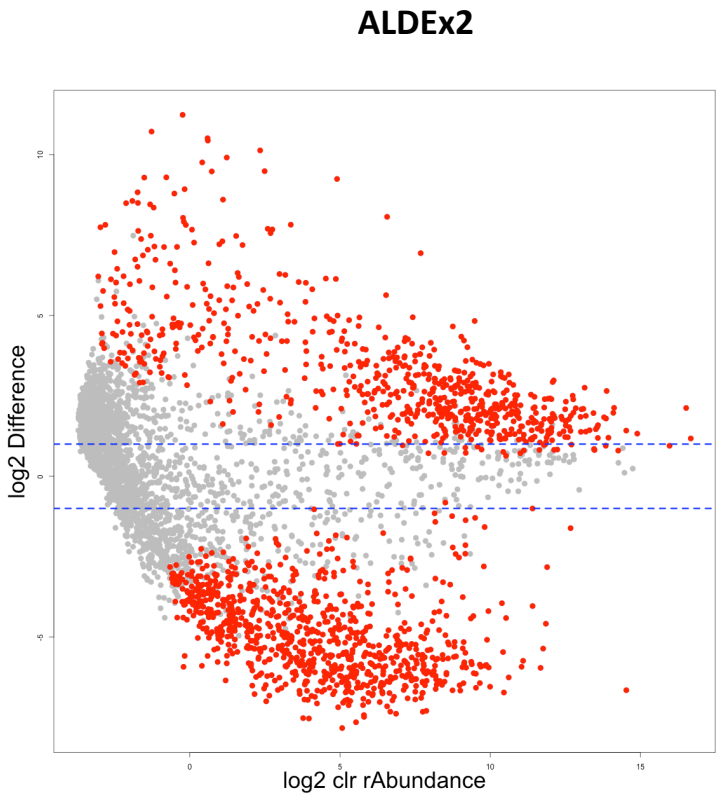
Obviously unbalanced dataset

What assumptions do not fit?

- Majority invariant?
- Dispersion ~ abundance?
- Counts?

So let's try probabilistic CoDa





Strategies to center the data

$$\mathbf{x}_{clr} = log(\frac{x_i}{G(\mathbf{x})})_{i=1...D}$$ All

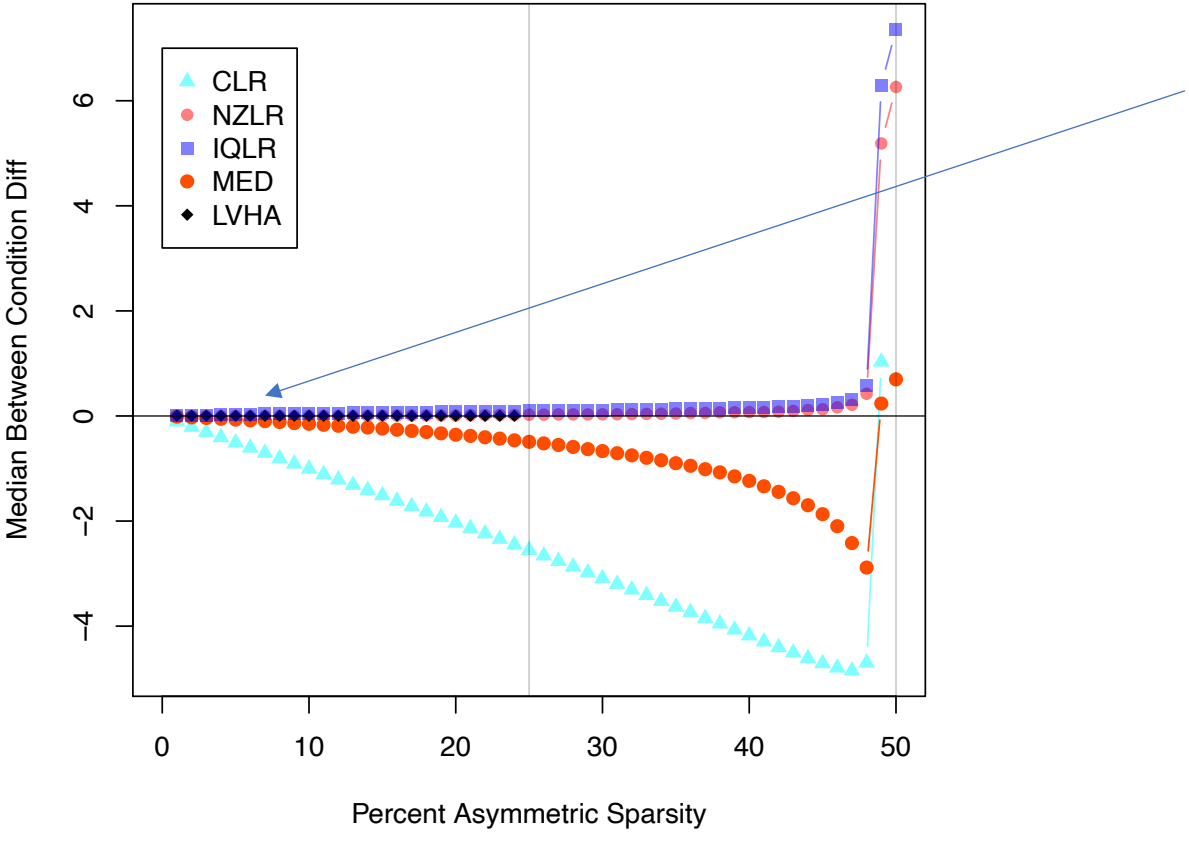
$$\mathbf{x}_{alr} = log(\frac{x_i}{x_D})_{i=1...D-1}$$ One

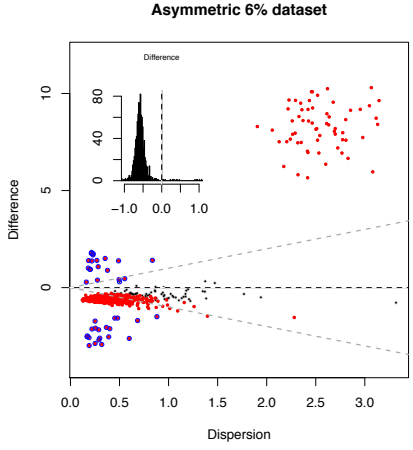
$$\mathbf{x_{i,}}_{IQLR} = \mathbf{log}(\frac{\mathbf{x_{i,j=1...D}}}{\mathbf{G}(IQVF)})$$ Robust midpoint variance

$$\mathbf{x_{i,}}_{LVHA} = \mathbf{log}(\frac{\mathbf{x_{i,j=1...D}}}{\mathbf{G}(LVHA)})$$ Low variance, high abundance

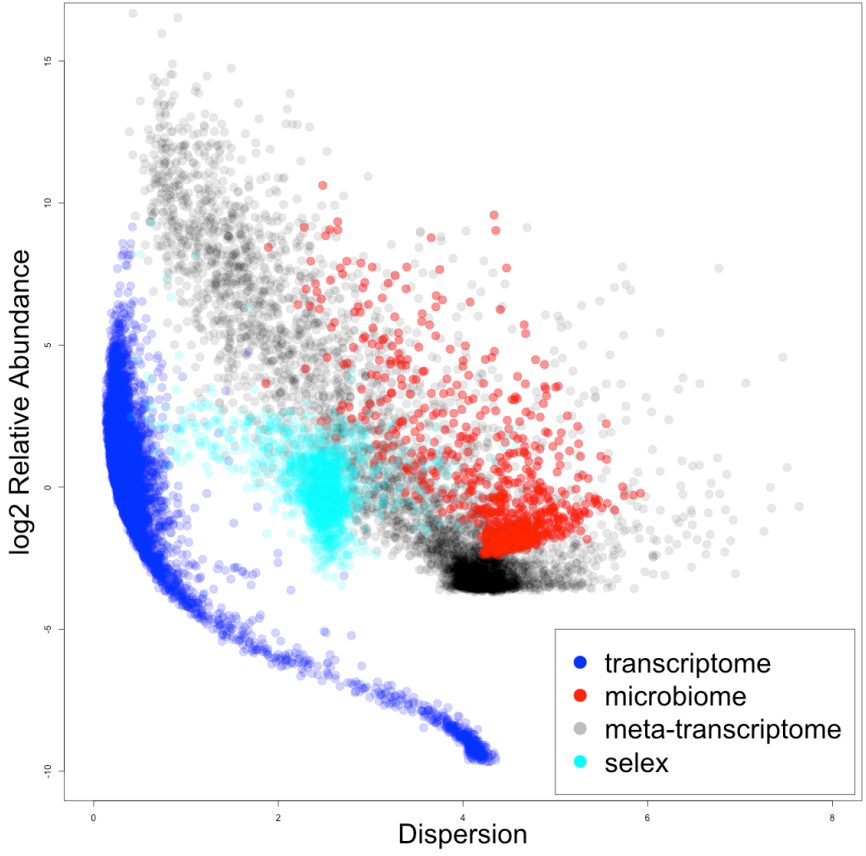
$$\mathbf{x_{i,MED}} = \mathbf{log}(\mathbf{x_{i,j=1...D}}) - \mathrm{MED}(\mathbf{log}(\mathbf{x_i}))$$ Median

Properties in simulated data





Dispersion v. rAbundance



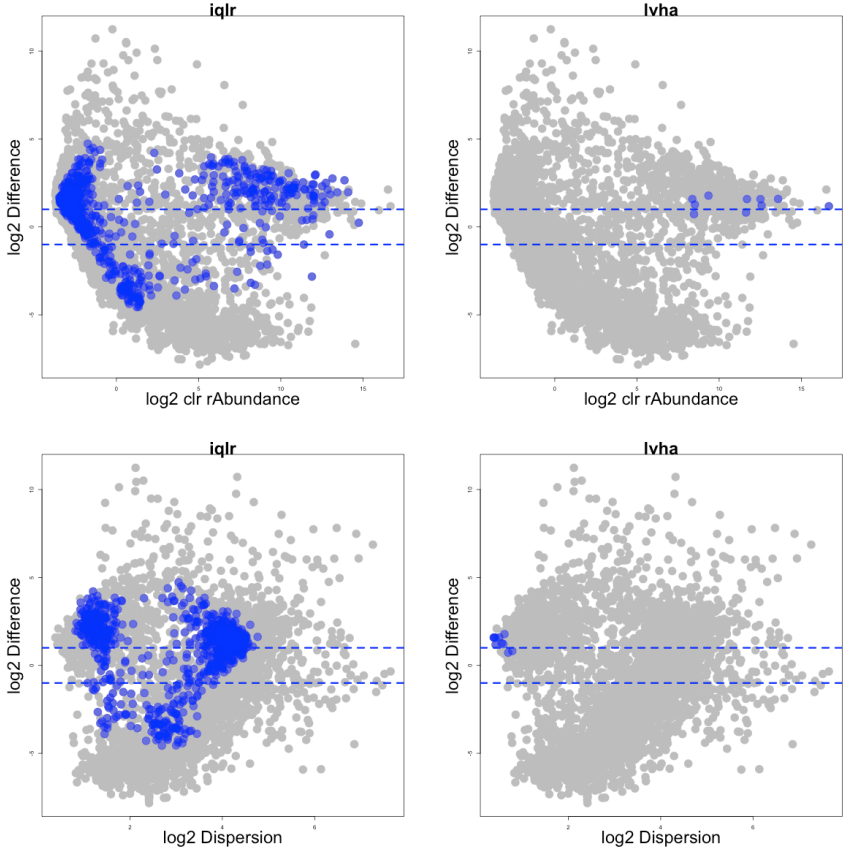
iqlr high variance datasets and low reproducibility

- tag-sequencing

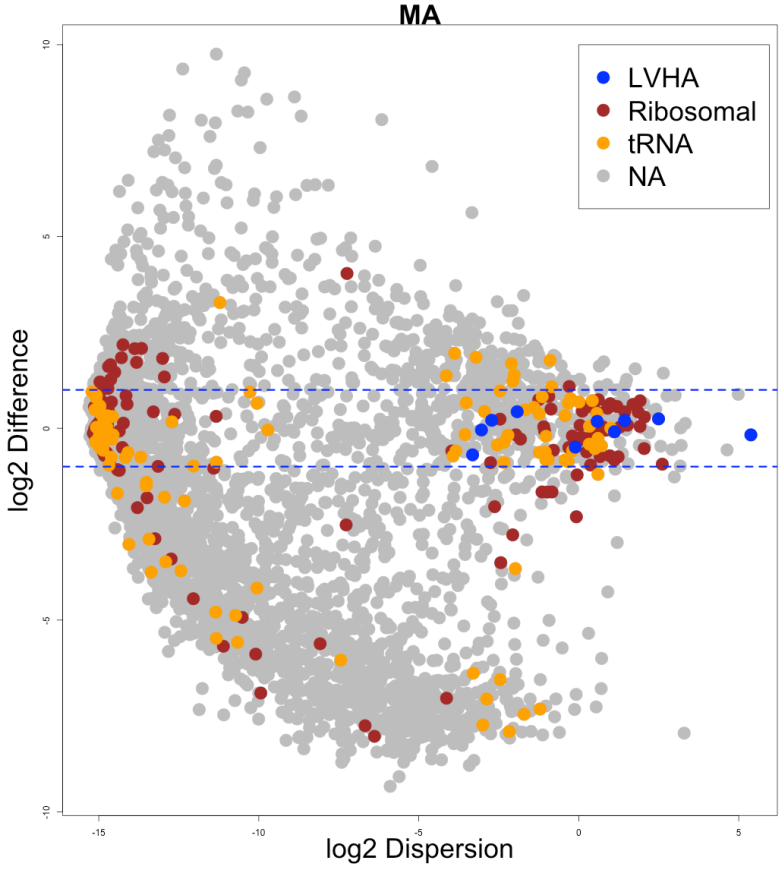
lvha low variance datasets and high reproducibility

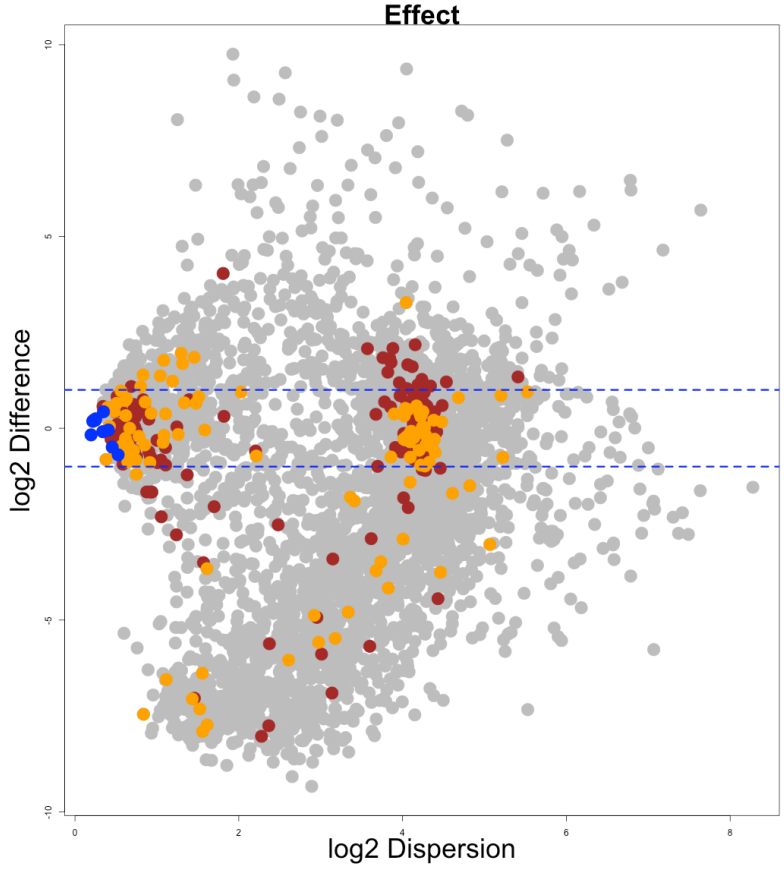
- transcriptomes

In meta-transcriptome data

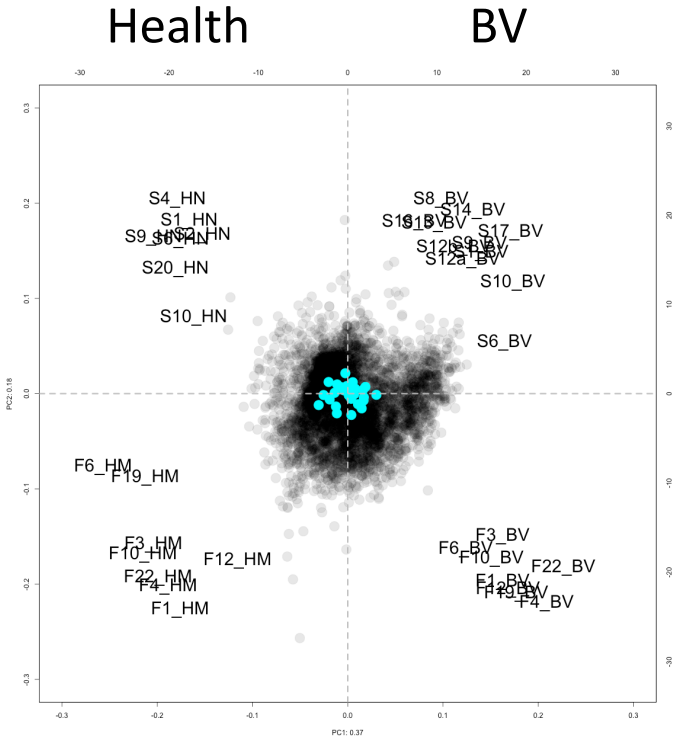


We have centered the housekeeping genes





Two disparate datasets



Swab

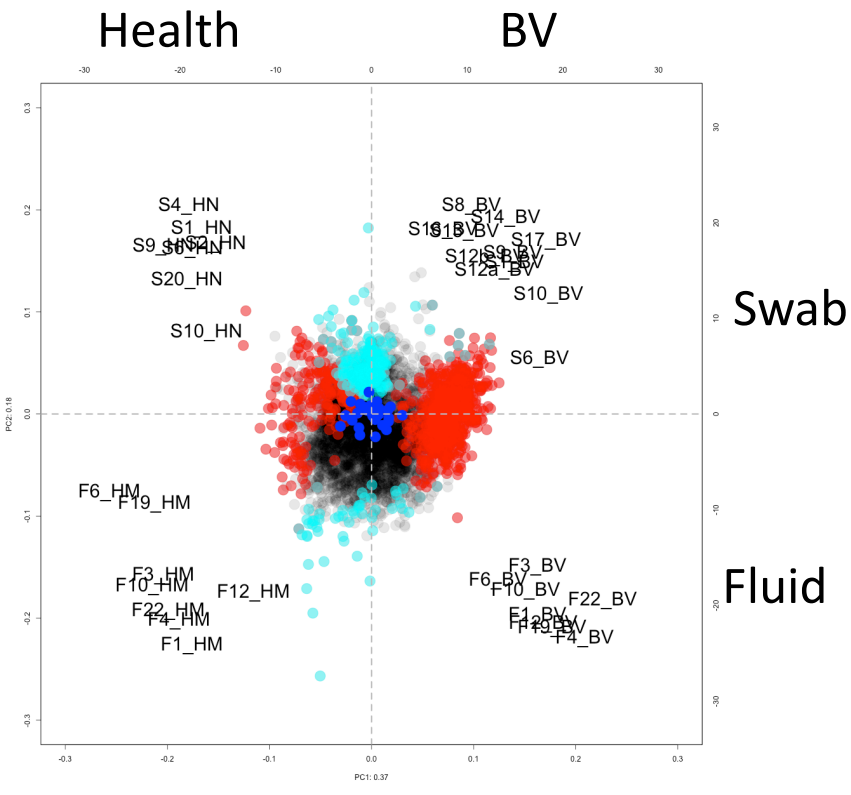
Fluid

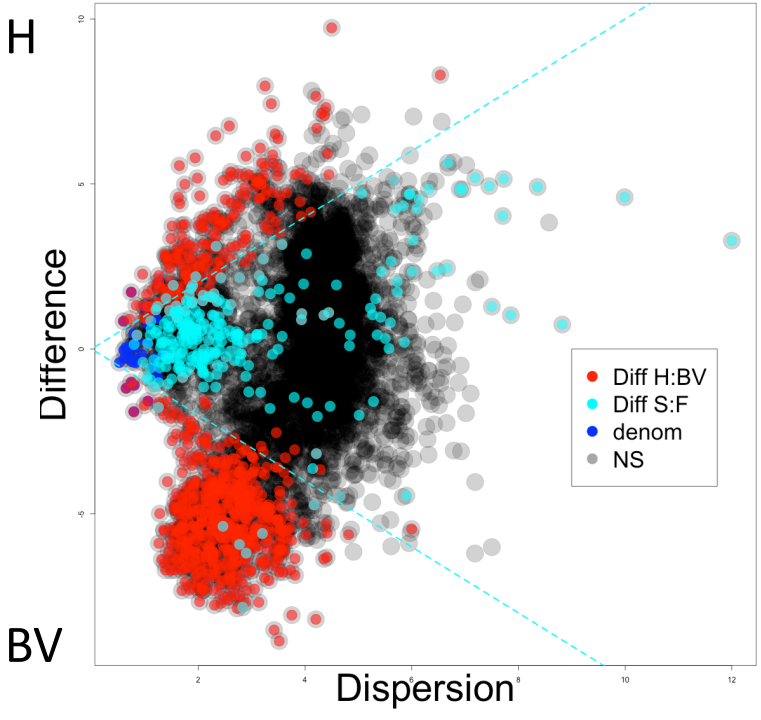
Swabs collected in Ontario PRJEB31833

Fluid collected in Germany PRJEB21446

Data centred with LVHA

Output from aldex.glm

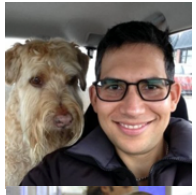




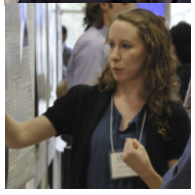
Conclusions

- Asymmetry in datasets is surprisingly common
- Proper centering is needed to prevent non-sensical answers
  - LVHA and IQLR work in a compositional setting
- Choosing the proper centering method depends on the var-abun relationship
- Meta-transcriptomes are hard
  - Can at least get reproducible answers across datasets
  - Standard approaches work with some effort

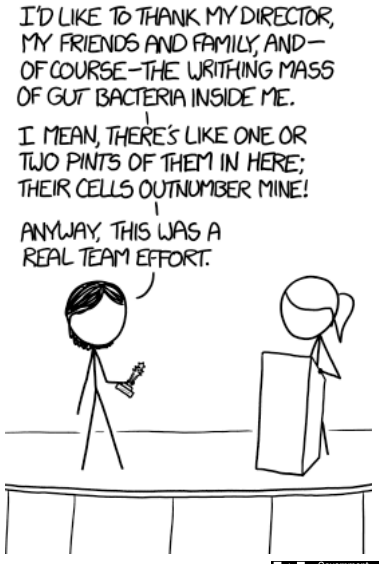
Canadian Centre for Human Microbiome and Probiotic Research



Andrew Fernandes



Jean Macklaim



CoDa

Vera Pawlowsky-Glahn
Juan Jose Egozcue

Justin Silverman
philr
Tom Quinn
propr

Gregor Reid
Jeremy Burton
Jia R. Wu
Daniel Giguere

THE W. GARFIELD WESTON FOUNDATION

NIH National Institutes of Health
Turning Discovery Into Health

NSERC CRSNG
People. Discovery. Innovation.

Government of Canada / Gouvernement du Canada
Agriculture and Agri-Food Canada

Ontario Genomics

Ontario Centres of Excellence
Where Next Happens

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