

Accounting for asymmetry and batch effects in meta- transcriptomics



Western

Greg Gloor

U. Western Ontario

Department of Biochemistry

gslot@uwo.ca : @gbgloor

github.com/ggloor/compositions/presentations



Schulich
MEDICINE & DENTISTRY

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

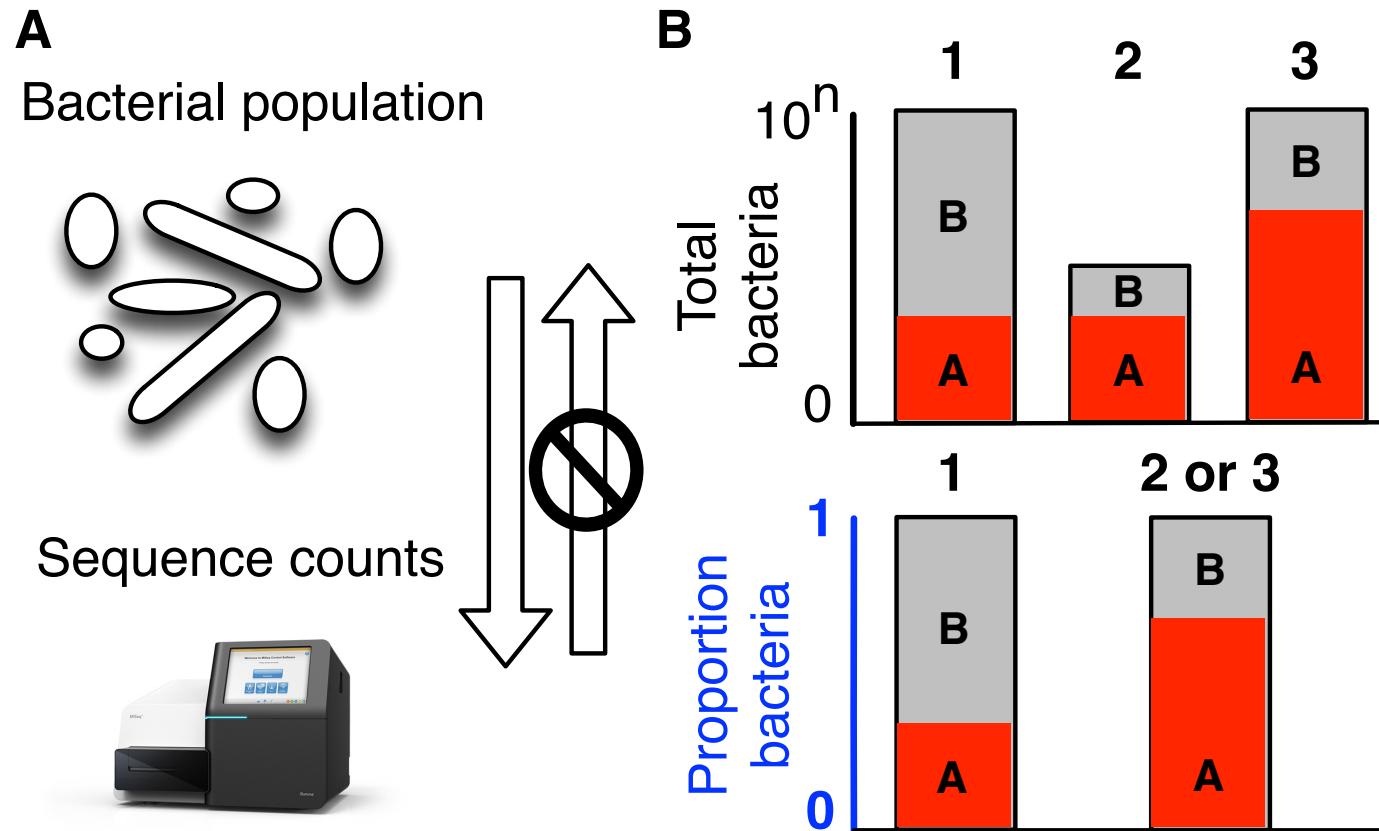


Jia R. Wu
PhD student
U. Waterloo



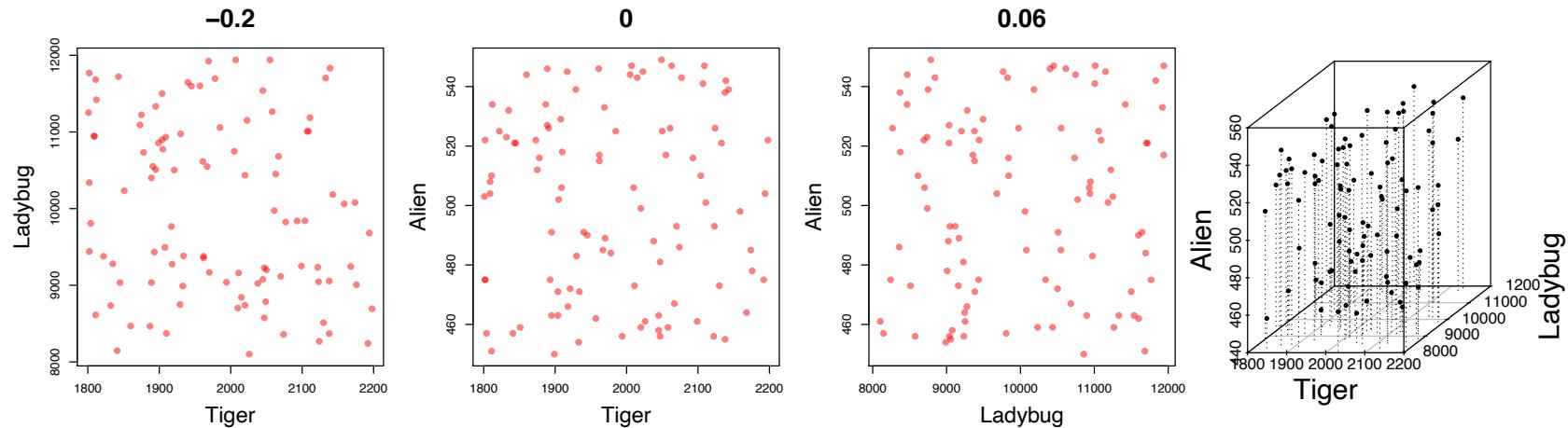
Thom Quinn
PhD student
Deakin U.

The sequencing meat grinder



Gloor et al Front. Micro. 2017

The machine delivers ‘counts’

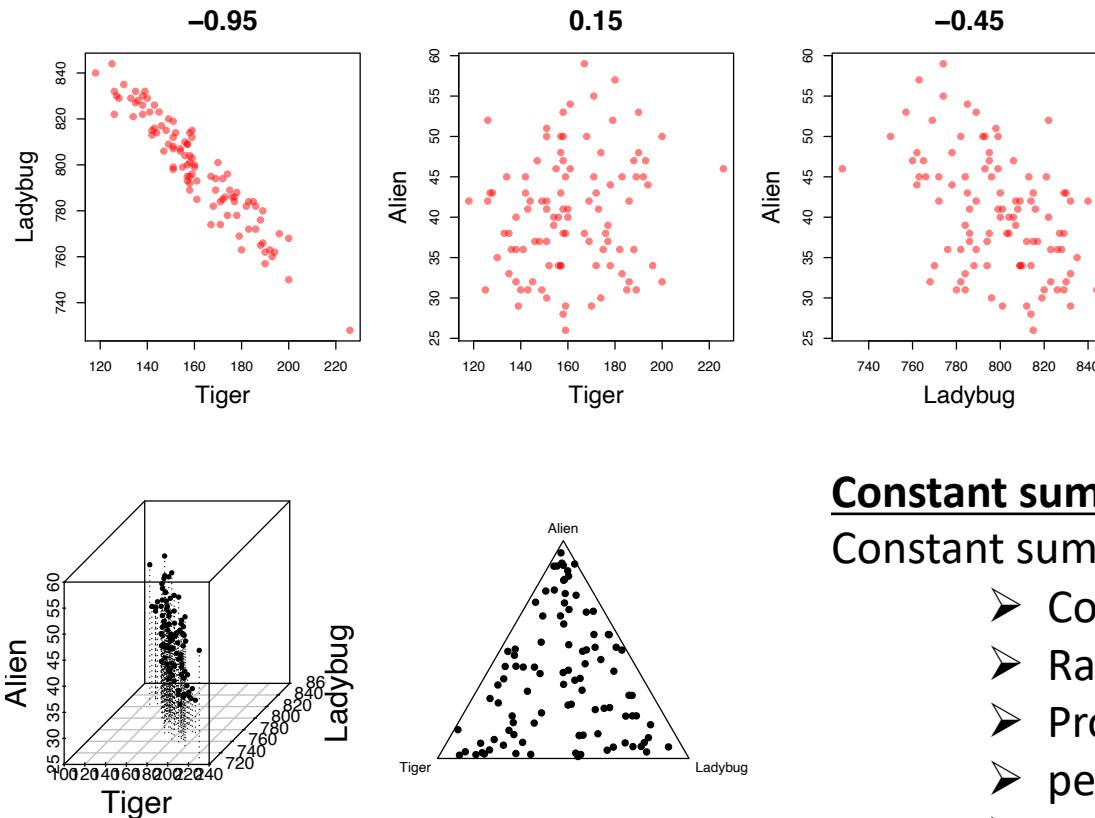


Example 100 random sample sets

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

Gloor et al Front. Micro. 2017

'Counts' have a maximum sum



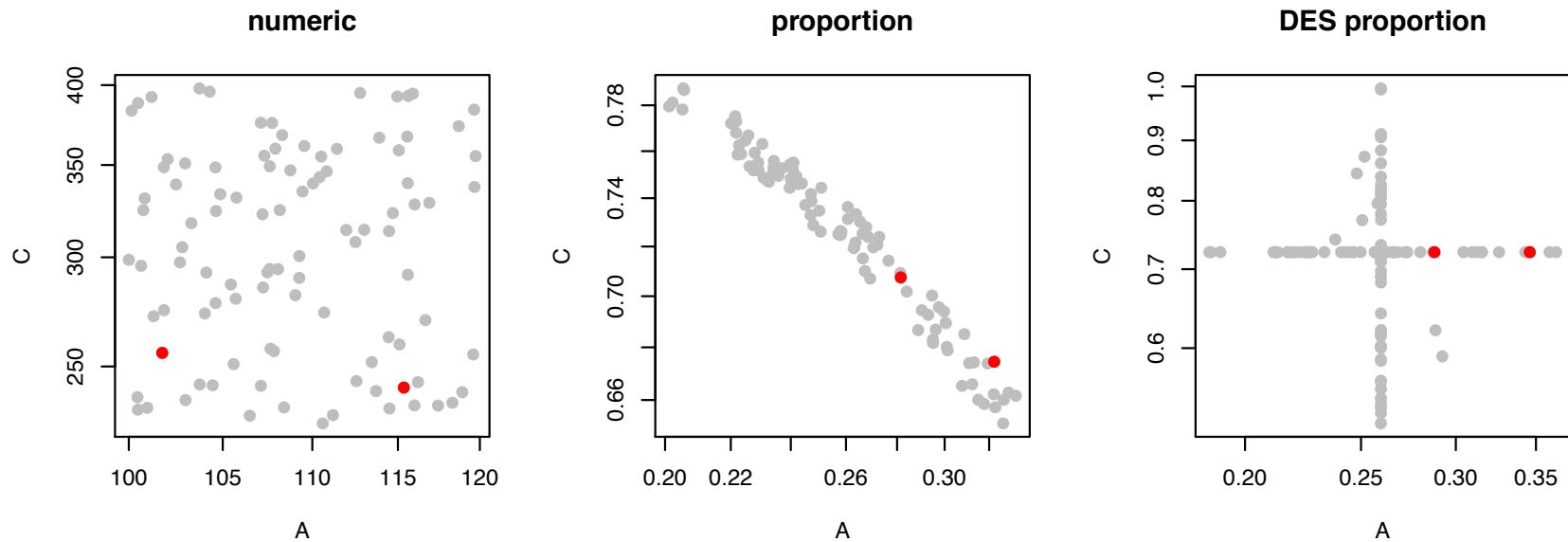
Constant sum of 1000

Constant sum operations:

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

Gloor et al Front. Micro. 2017

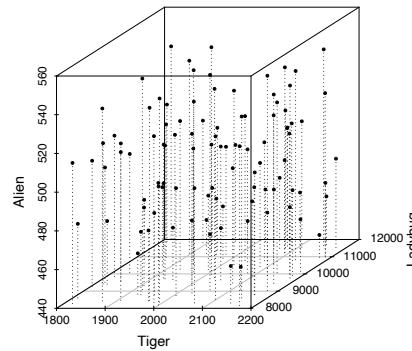
Normalization is no help at all



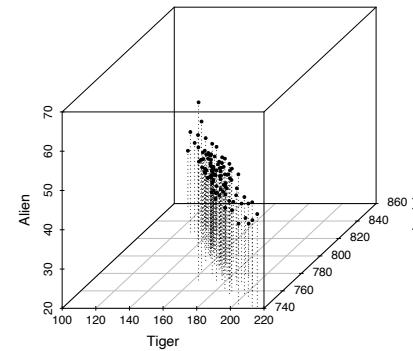
Once data are on the simplex, they cannot be removed without additional information

Log-ratios for CoDa

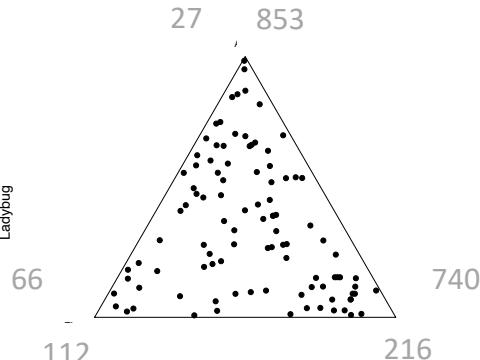
Count



Sequencing



Simplex



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

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

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

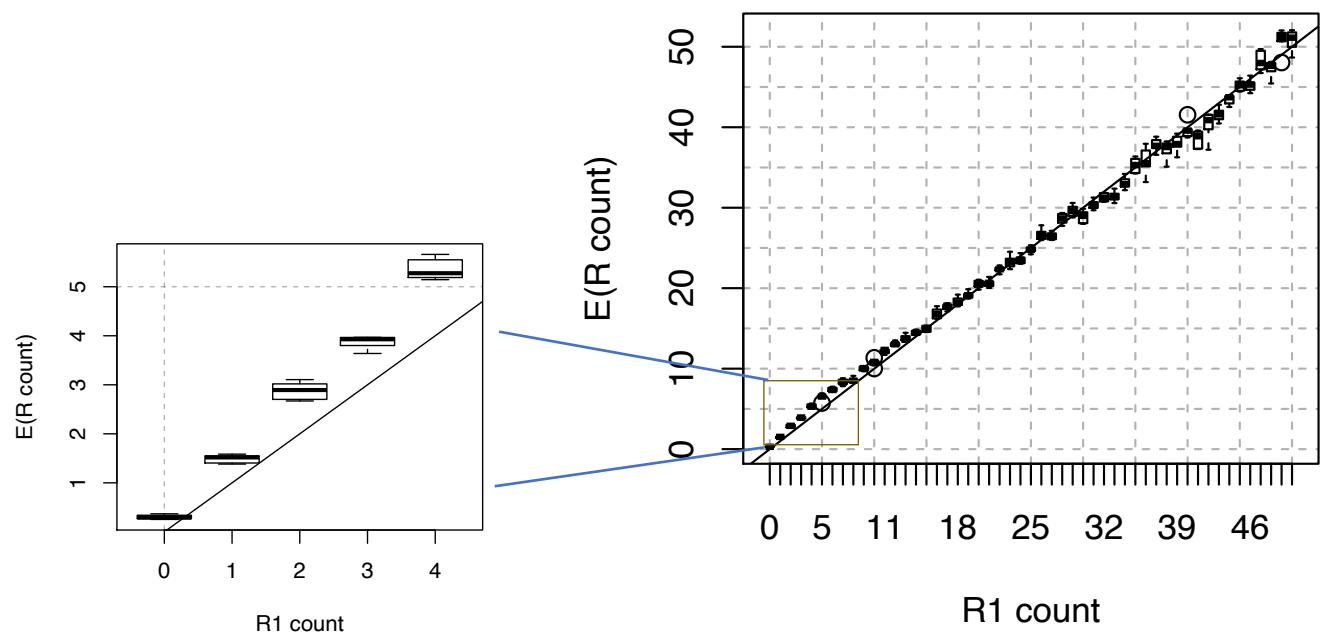
Aitchison 1986. Stat. Anal. Comp Data
Pawlowsky-Glahn, 2015. Mod. Anal. CoDa

The 0 in the room

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

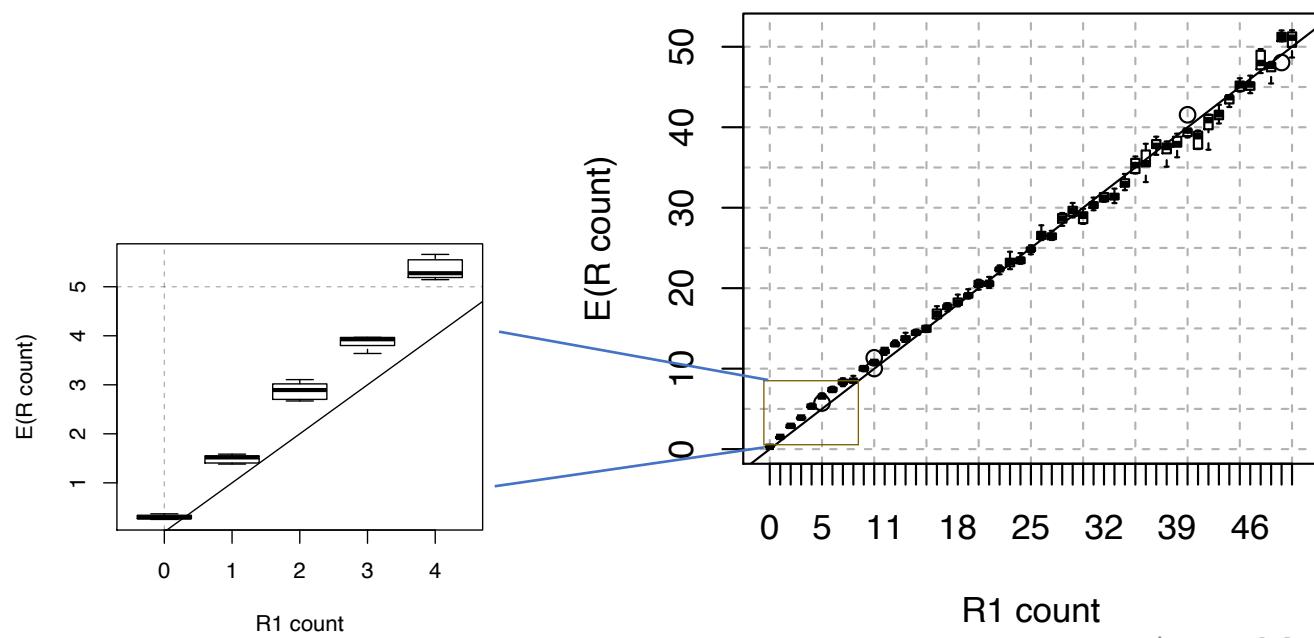
$$X = [x_1, x_2, x_i \dots x_D] = [p_1, p_2, p_i \dots p_D]\alpha ; \alpha = \sum X; \sum P = 1$$

$$x_i = (P_i | f_i)\alpha$$

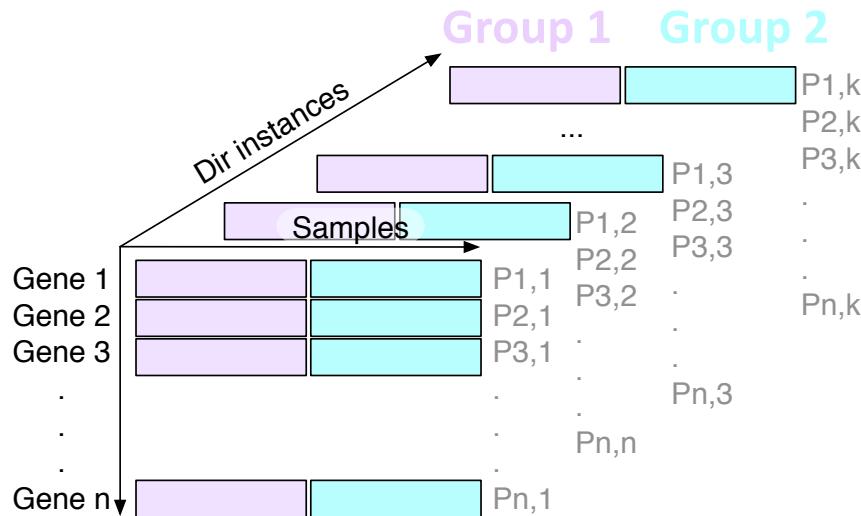


ALDEx2

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

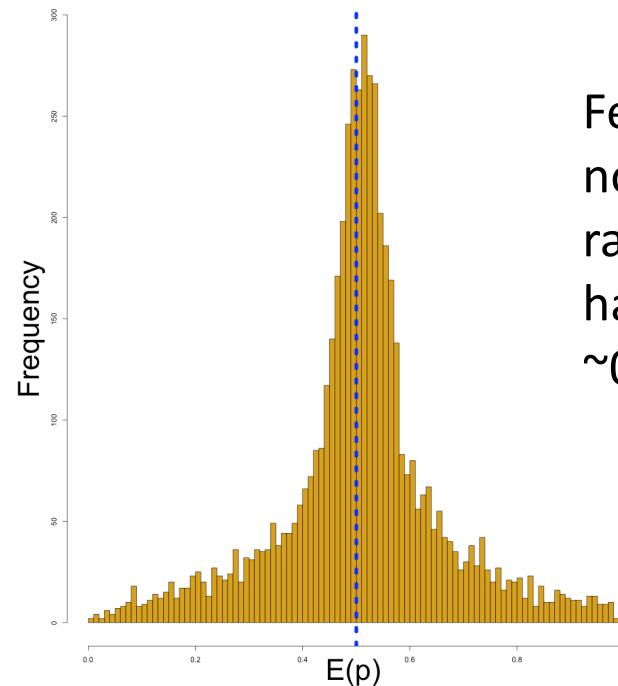
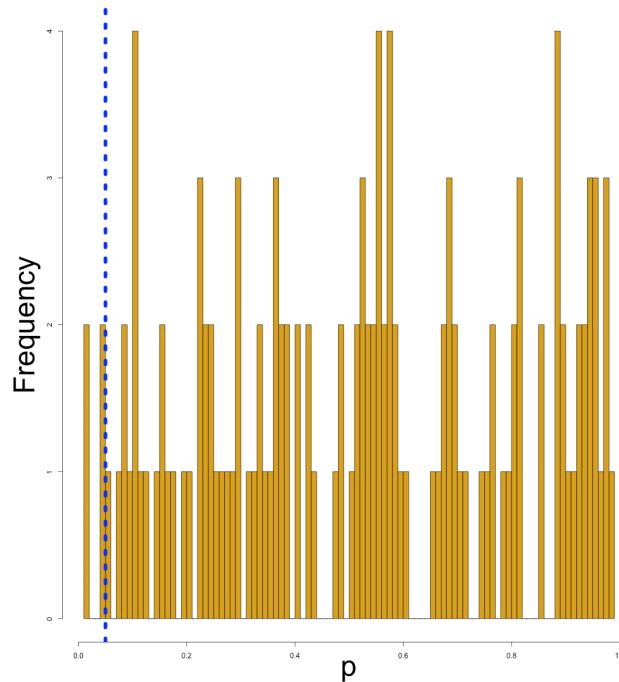


ALDEx2: probabilistic CoDa



- 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

ALDEx2 Expected value



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

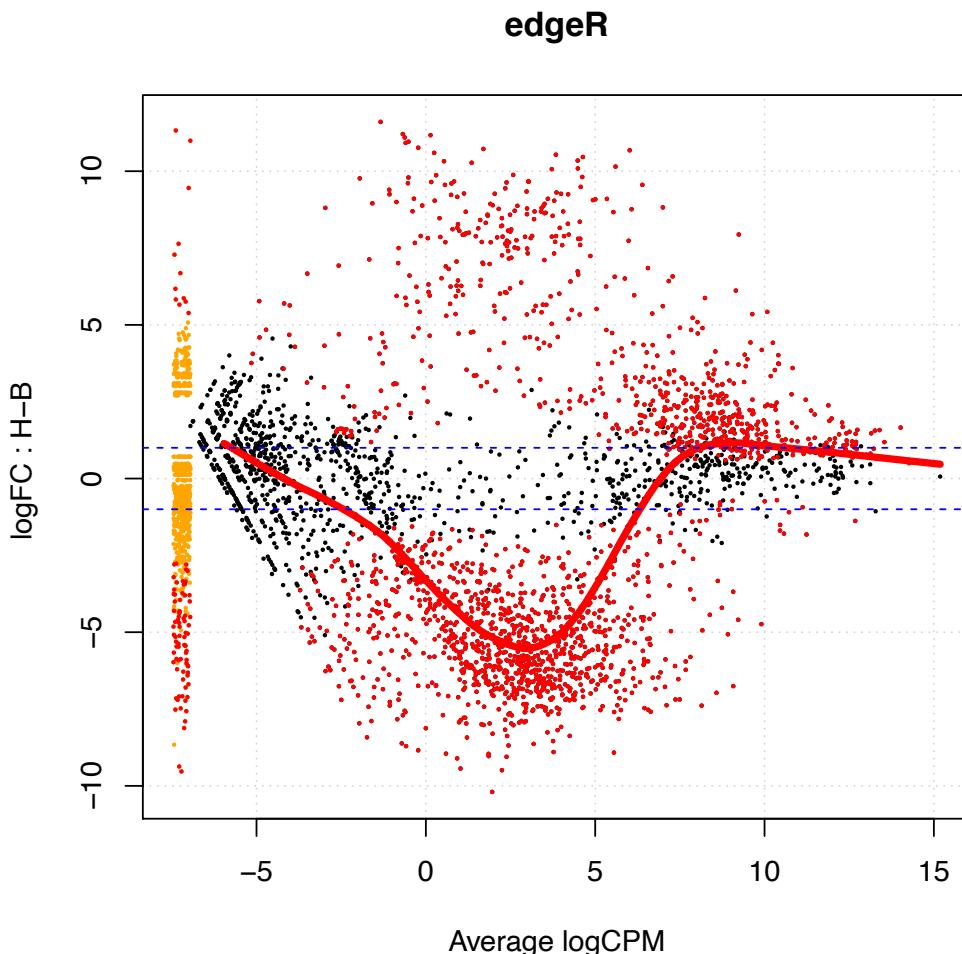
Meta-transcriptomics of an ecosystem

- 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

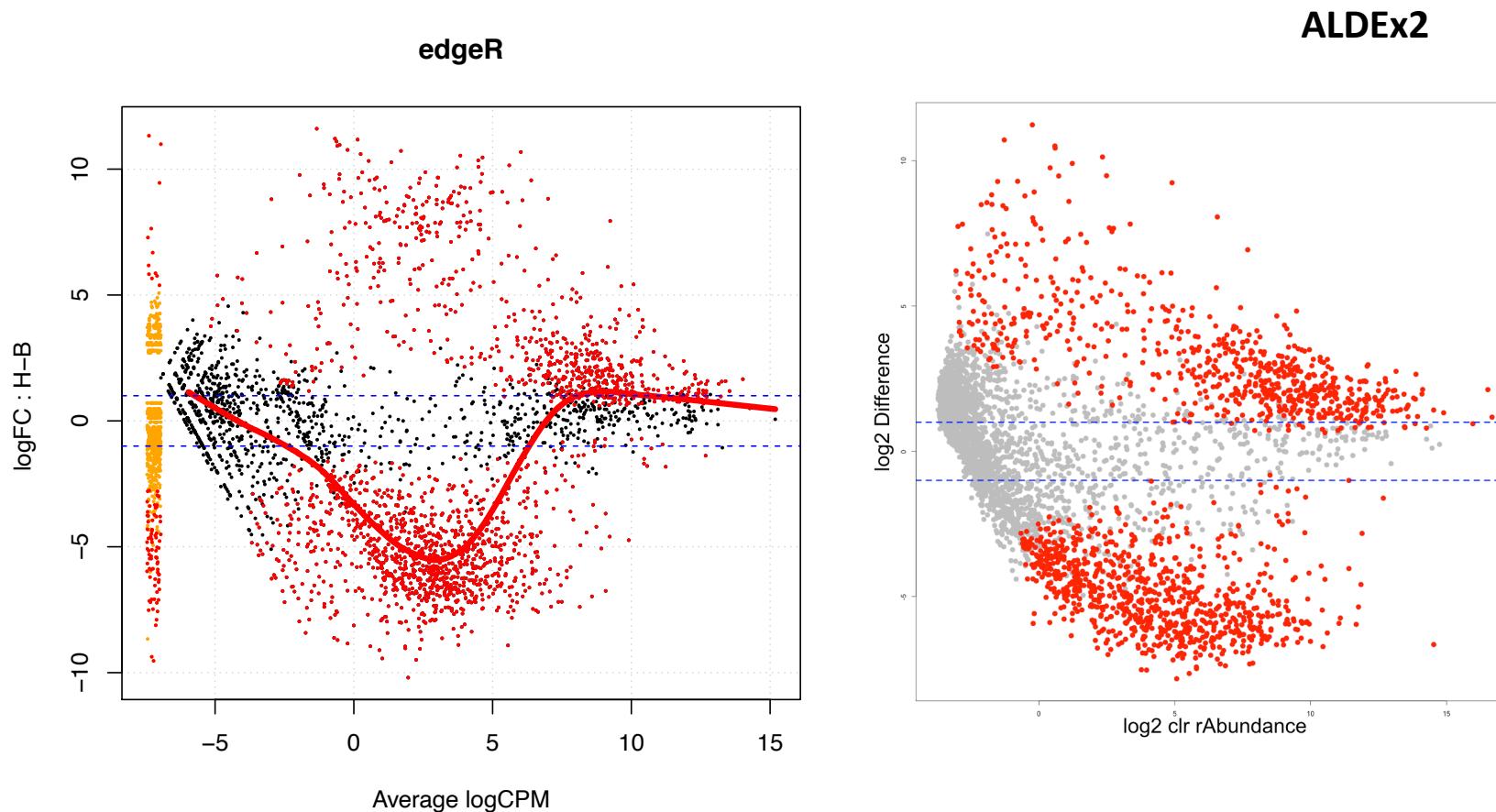
- Each function is a point
- Red are 'DE'

Obviously unbalanced dataset

What assumptions do not fit?

- Majority invariant?
- Dispersion \sim abundance?
- Counts?

So let's try probabilistic CoDa



Strategies to center the data

$$\mathbf{x}_{clr} = \log\left(\frac{x_i}{G(\mathbf{x})}\right)_{i=1\dots D}$$

All

$$\mathbf{x}_{alr} = \log\left(\frac{x_i}{x_D}\right)_{i=1\dots D-1}$$

One

$$\mathbf{x}_{i,IQLR} = \log\left(\frac{\mathbf{x}_{i,j=1\dots D}}{\mathbf{G}(IQVF)}\right)$$

Robust midpoint variance

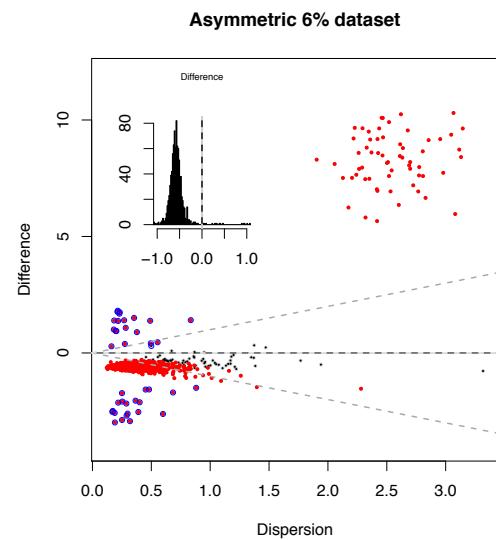
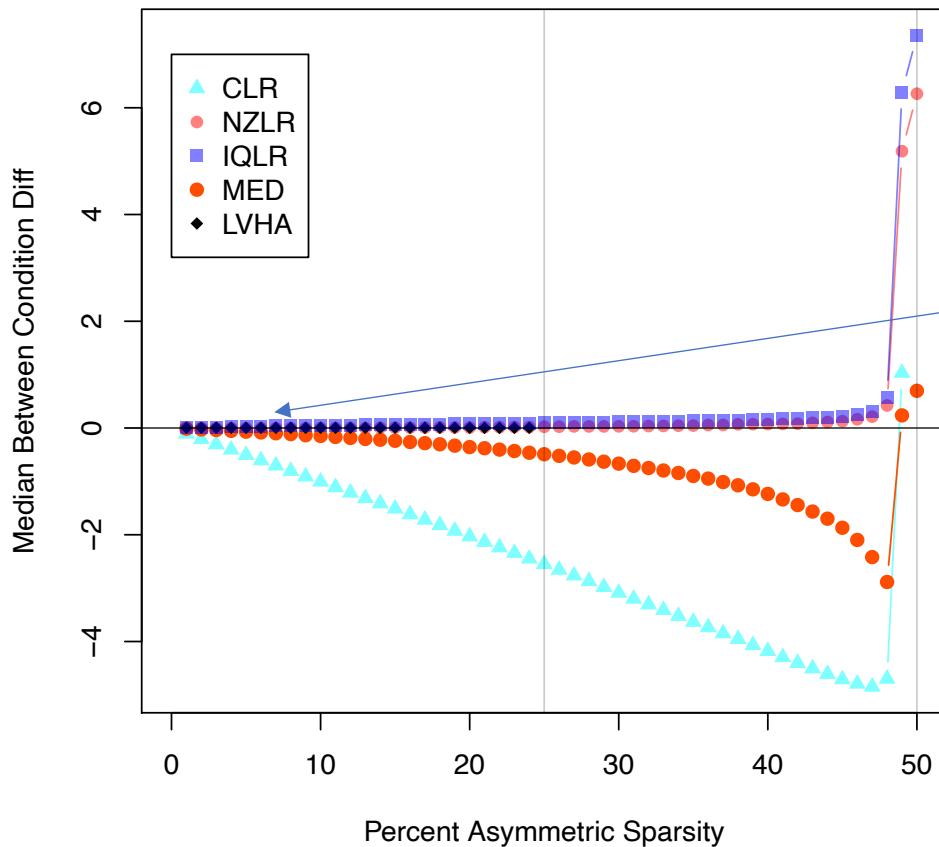
$$\mathbf{x}_{i,LVHA} = \log\left(\frac{\mathbf{x}_{i,j=1\dots D}}{\mathbf{G}(LVHA)}\right)$$

Low variance, high abundance

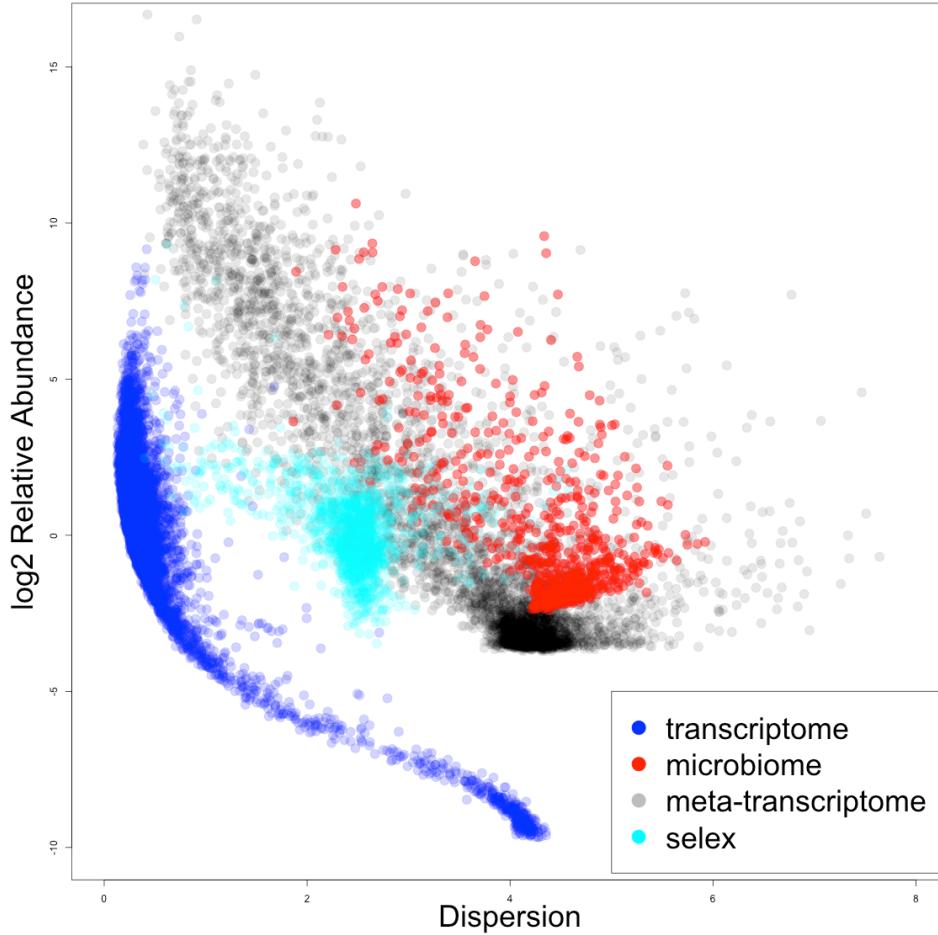
$$\mathbf{x}_{i,MED} = \log(\mathbf{x}_{i,j=1\dots D}) - \text{MED}(\log(\mathbf{x}_i))$$

Median

Properties in simulated data



Dispersion v. rAbundance



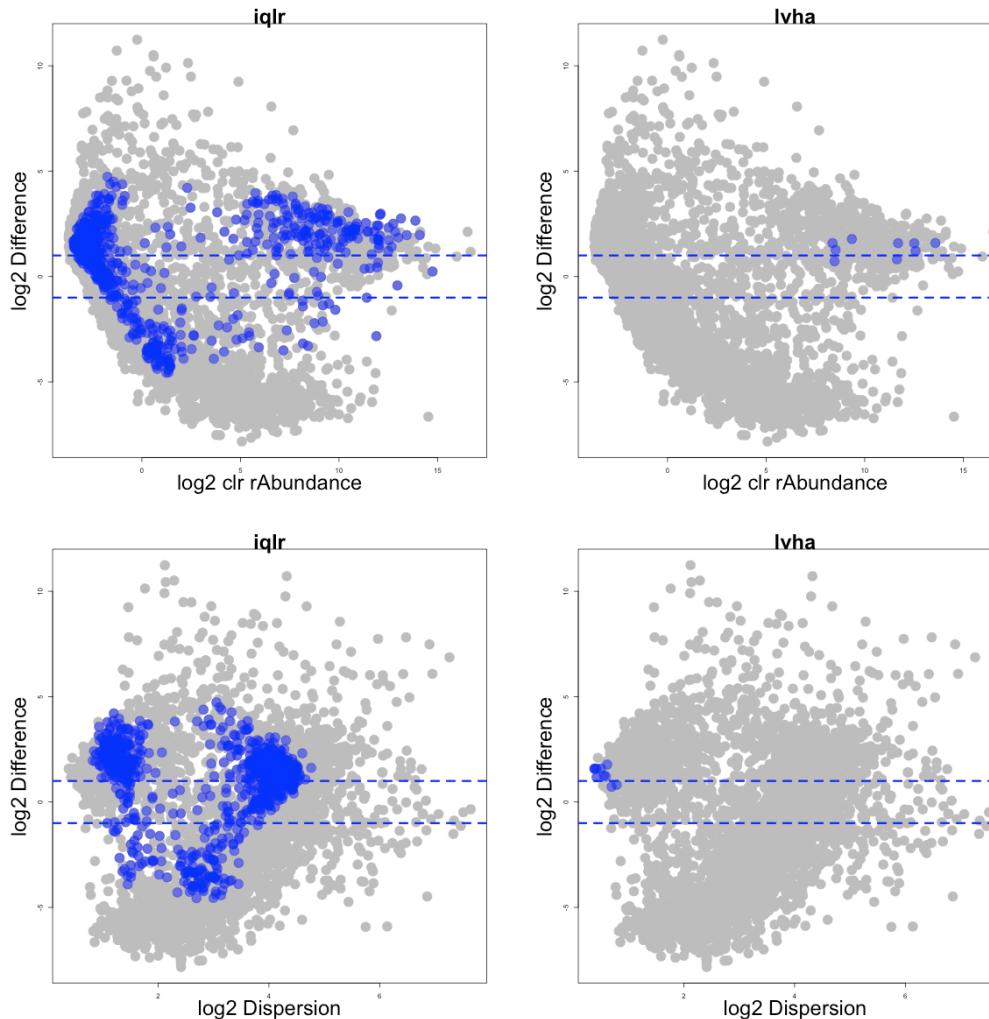
iqlr high variance datasets and low reproducibility

- tag-sequencing

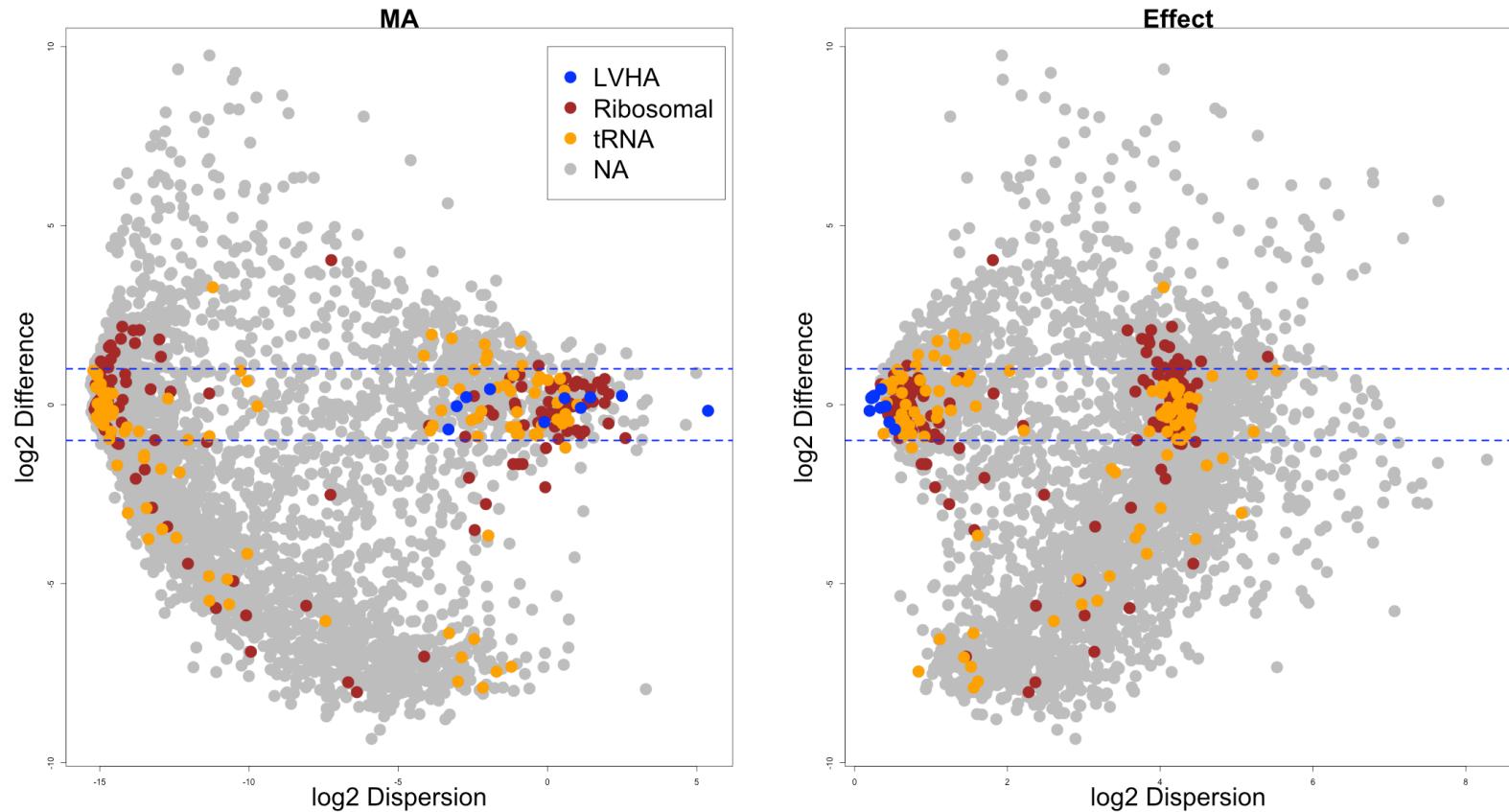
lvha low variance datasets and high reproducibility

- transcriptomes

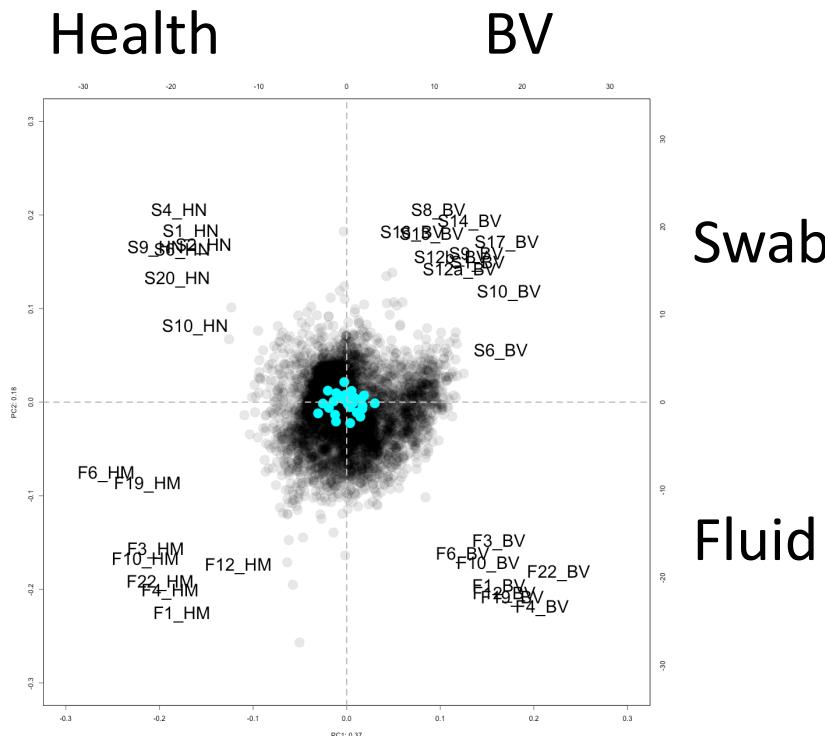
In meta-transcriptome data



'Housekeeping' genes are centered



Two disparate datasets

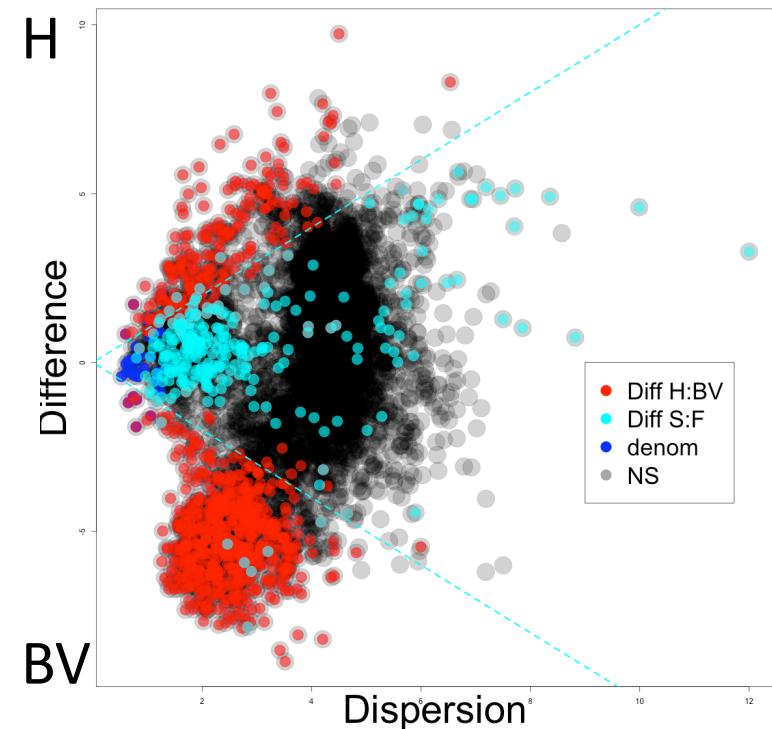
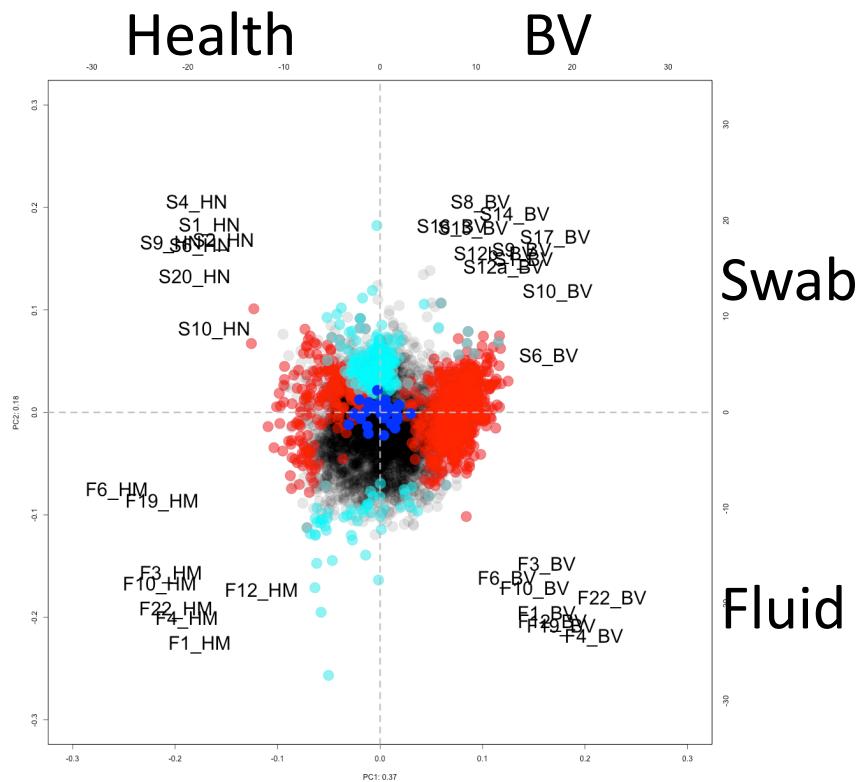


Swabs collected in Ontario
PRJEB31833

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

Gregor Reid
Jeremy Burton
Jia R. Wu
Daniel Giguere



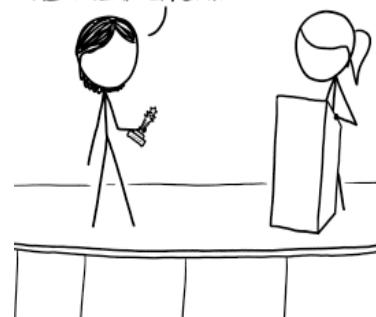
THE W. GARFIELD WESTON
FOUNDATION

NIH National Institutes of Health
Turning Discovery Into Health

People. Discovery. Innovation.

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.



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

Ontario Genomics

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

Ontario Centres of
Excellence
Where Next Happens



CIHR IRSC

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



Vera Pawlowsky-Glahn
Juan Jose Egozcue

Justin Silverman
philr
Tom Quinn
propr