NCCR MICROBIOMES

NCCR AntiResist

SIB Swiss Institute of Bioinformatics

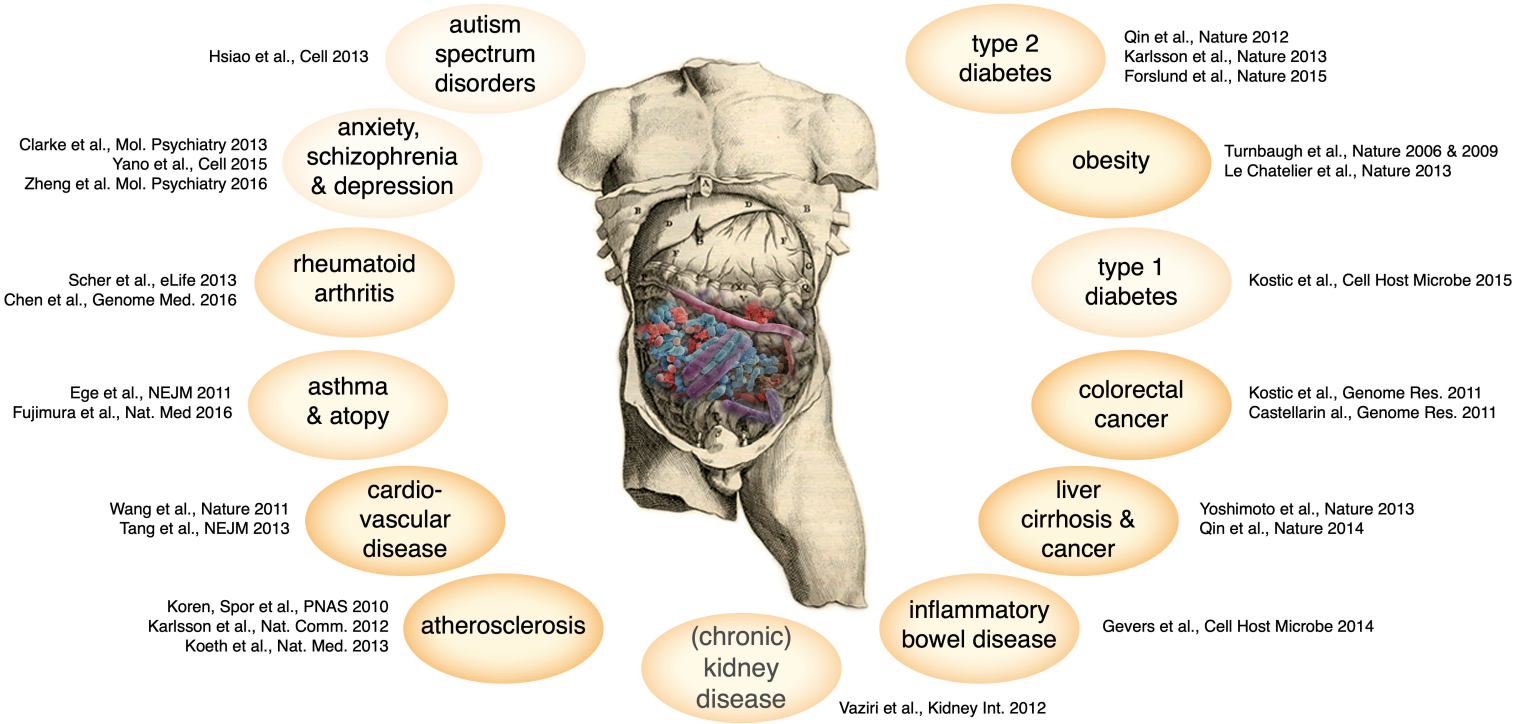
Univariate statistical tests for metagenomic data

Project 3

Spring School Bioinformatics and computational approaches in Microbiology

Alessio Milanese, Lukas Malfertheiner

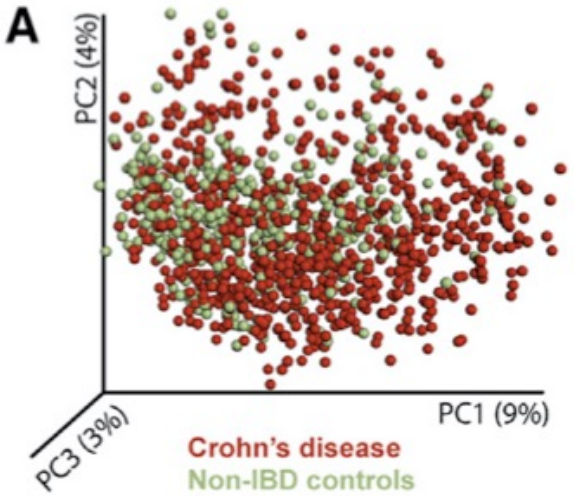
Comparing microbiome composition in case-control studies



Tools for microbial community comparison

Assessing difference in overall community structure

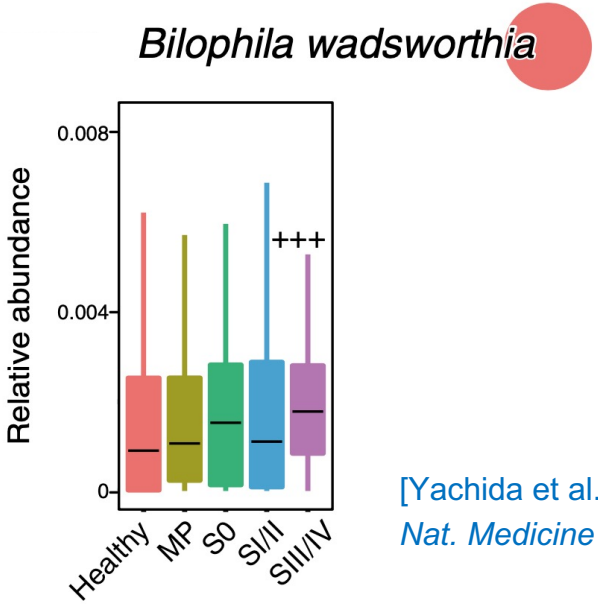
- Clustering
- Ordination



[Gevers et al. *Cell Host&Microbe* 2014]

Testing for changes in individual taxa

- Statistical testing

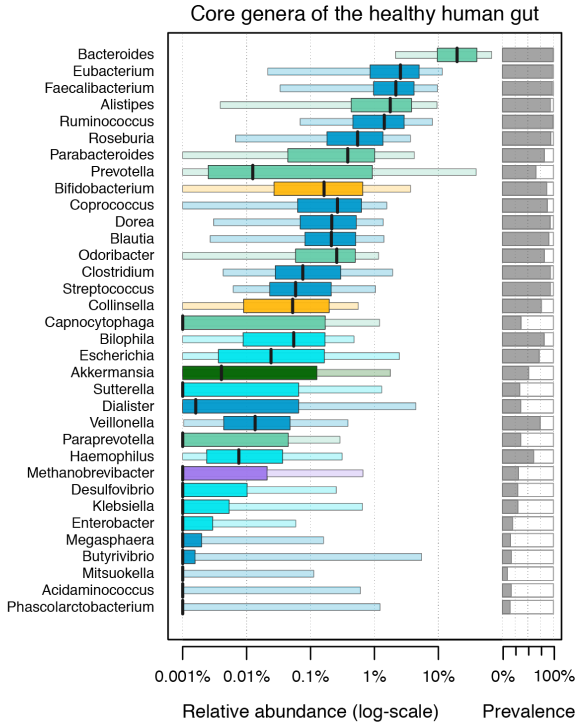


[Yachida et al. *Nat. Medicine* 2019]

Which statistical test is appropriate?

Some things to keep in mind:

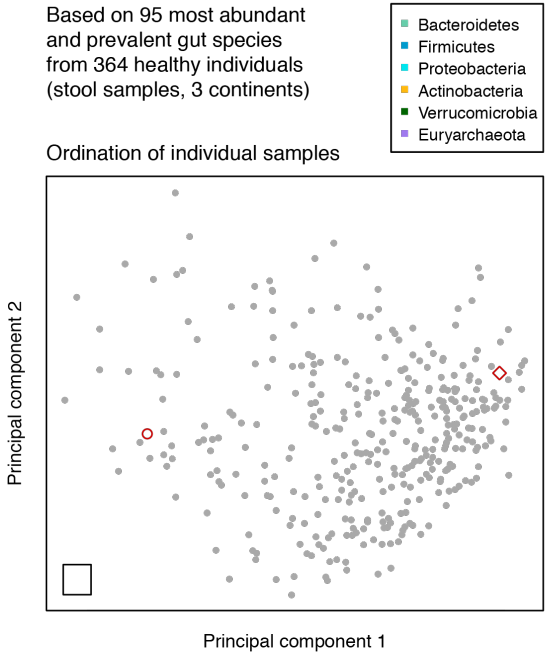
- Microbiome data show **zero-inflation**
- Microbiome data do **not** follow a **log-normal** distribution
- **Extreme variance** across individuals

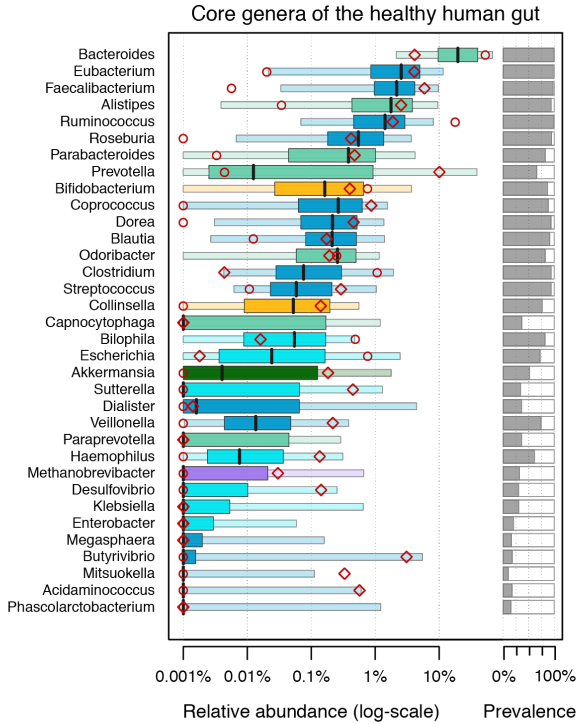


Which statistical test is appropriate?

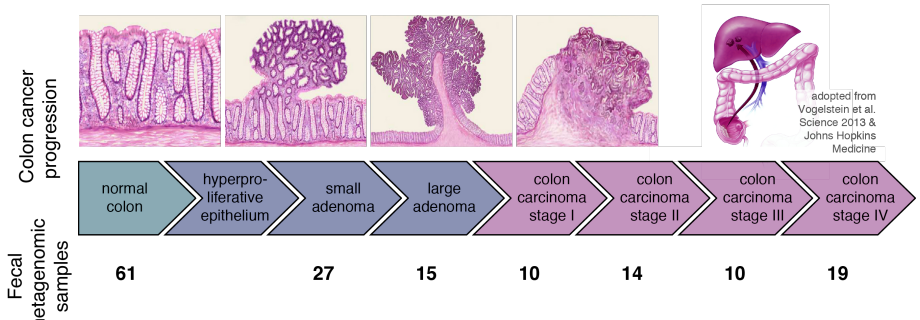
Some things to keep in mind:

- Microbiome data show **zero-inflation**
- Microbiome data do **not** follow a **log-normal** distribution
- **Extreme variance** across individuals





Colorectal cancer (CRC) as an introductory example



- Collected stool samples from 46 colorectal cancer (CRC) patients and 60 healthy controls
- Used metagenomic sequencing and profiled gut bacterial species
- Can microbiome differences be used for non-invasive detection of cancer?

[Zeller*, Tap*, Voigt* et al., *Mol. Syst. Biol.* 2014]

Statistically significant associations with CRC

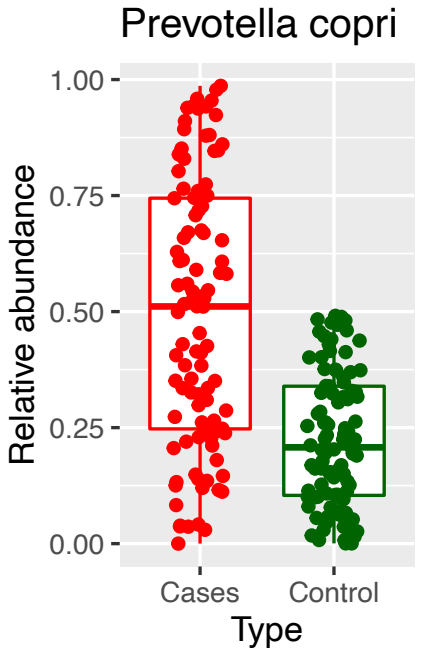
- How would you identify which species are associated to CRC?

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
B. vulgatus	0.2	0.5	0.5	0.1	0.2	0	0.1	0.2	0	0.3
P. copri	0.3	0.2	0.2	0.1	0.2	0.2	0.2	0.1	0.1	0
E. rectale	0.2	0	0	0.4	0	0	0	0	0.1	0.1
B. wexlerae	0	0.2	0.2	0	0.3	0.1	0.1	0.2	0.1	0
A. putredinis	0	0	0	0.3	0	0.3	0	0	0	0.3
E. coli	0	0	0	0	0	0.3	0.2	0.5	0.6	0.1
C. innocuum	0	0.1	0.1	0	0.2	0	0	0.1	0	0.2
R. intestinalis	0.3	0.1	0.1	0	0.1	0.1	0.3	0	0.1	0
A. finegoldii	0	0	0	0.1	0	0	0	0	0	0

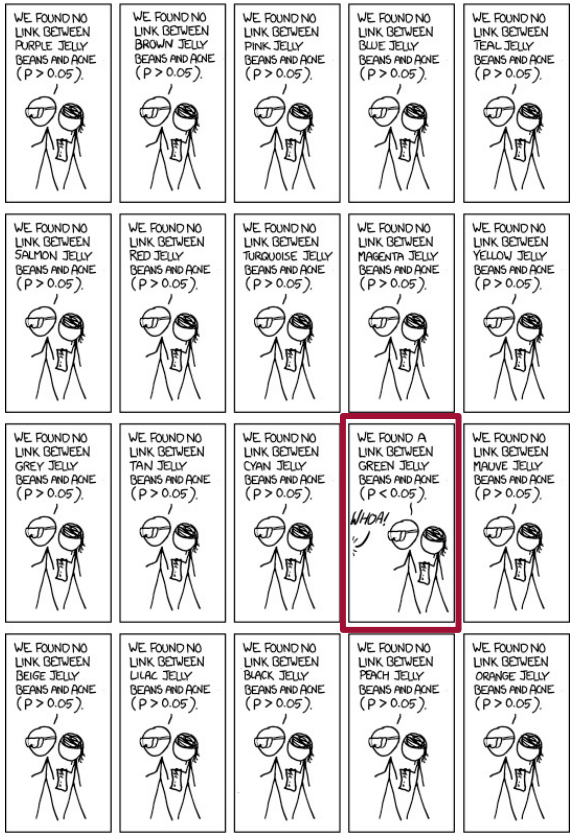
Statistically significant associations with CRC

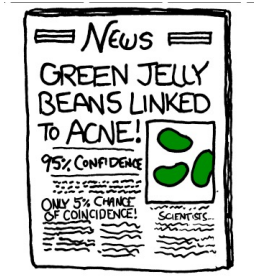
- How would you identify which species are associated to CRC?

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
B. vulgatus	0.2	0.5	0.5	0.1	0.2	0	0.1	0.2	0	0.3
P. copri	0.3	0.2	0.2	0.1	0.2	0.2	0.2	0.1	0.1	0
E. rectale	0.2	0	0	0.4	0	0	0	0	0.1	0.1
B. wexlerae	0	0.2	0.2	0	0.3	0.1	0.1	0.2	0.1	0
A. putredinis	0	0	0	0.3	0	0.3	0	0	0	0.3
E. coli	0	0	0	0	0	0.3	0.2	0.5	0.6	0.1
C. innocuum	0	0.1	0.1	0	0.2	0	0	0.1	0	0.2
R. intestinalis	0.3	0.1	0.1	0	0.1	0.1	0.3	0	0.1	0
A. finegoldii	0	0	0	0.1	0	0	0	0	0	0



Multiple testing correction

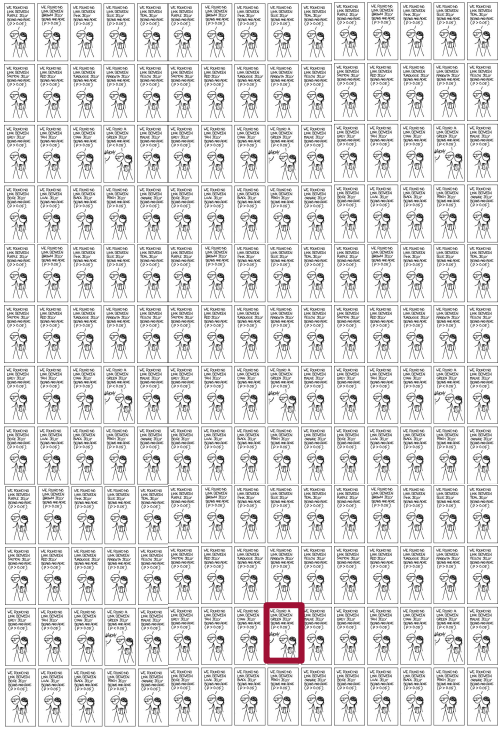


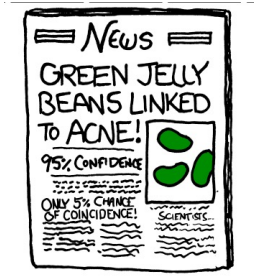


- Since we test several hundreds of taxa, some tests will be “significant” by chance
- It is thus crucial to perform a multiple testing correction, e.g.
  - The Benjamini-Hochberg procedure controls the false discovery rate (proportion of true positives among those for which the null hypothesis is rejected)
  - The Bonferroni procedure controls the family-wise error rate (probability of the significant set to contain any false positive)

https://www.xkcd.com/882/

Multiple testing correction





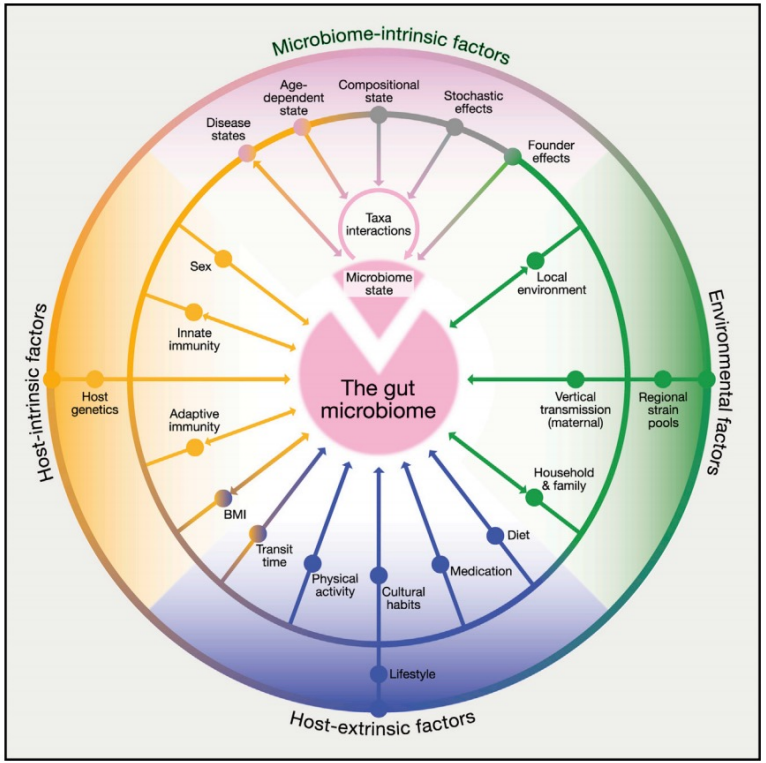
- Since we test several hundreds of taxa, some tests will be “significant” by chance
- It is thus crucial to perform a multiple testing correction, e.g.
  - The Benjamini-Hochberg procedure controls the false discovery rate (proportion of true positives among those for which the null hypothesis is rejected)
  - The Bonferroni procedure controls the family-wise error rate (probability of the significant set to contain any false positive)

https://www.xkcd.com/882/

Technical and biological effects on community composition can be challenging to deconvolute

- Technical factors can strongly affect microbial community profiles (**batch effects**), e.g. DNA extraction protocols, sequencing approach (16S primers), bioinformatic profiling

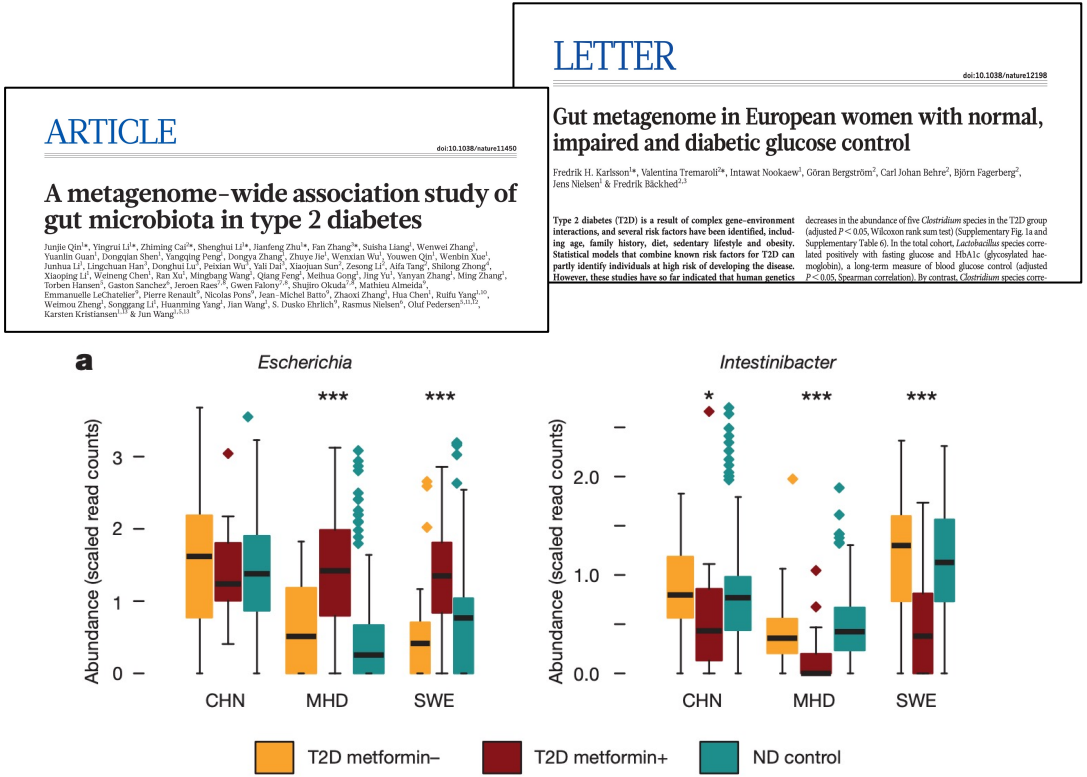
- Biological factors other than that of interest can affect profiles (**confounders**), e.g. medication, lifestyle, host demographics



[Schmidt et al. *Cell* 2018]

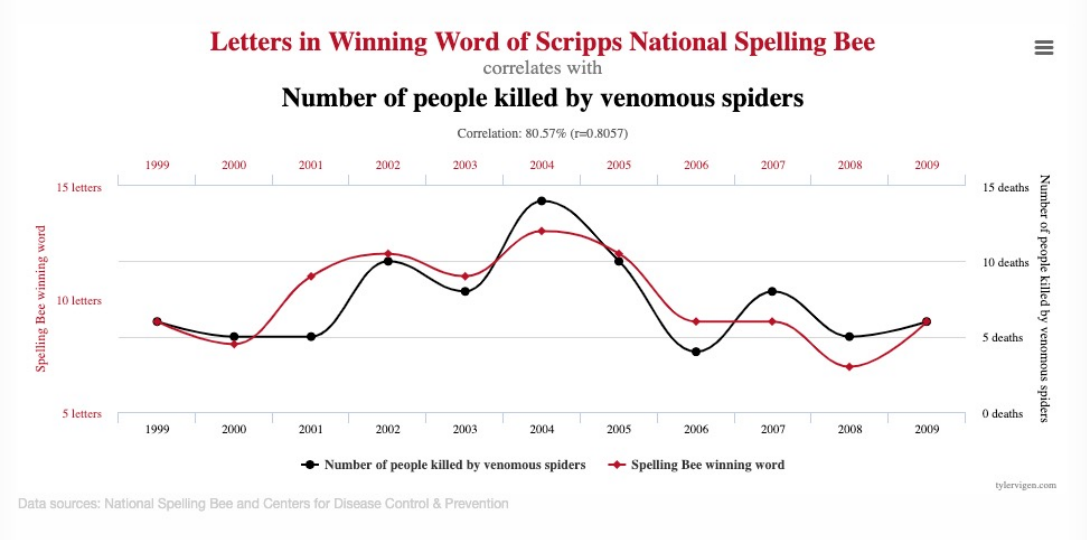
Caveat: confounding (here due to metformin)

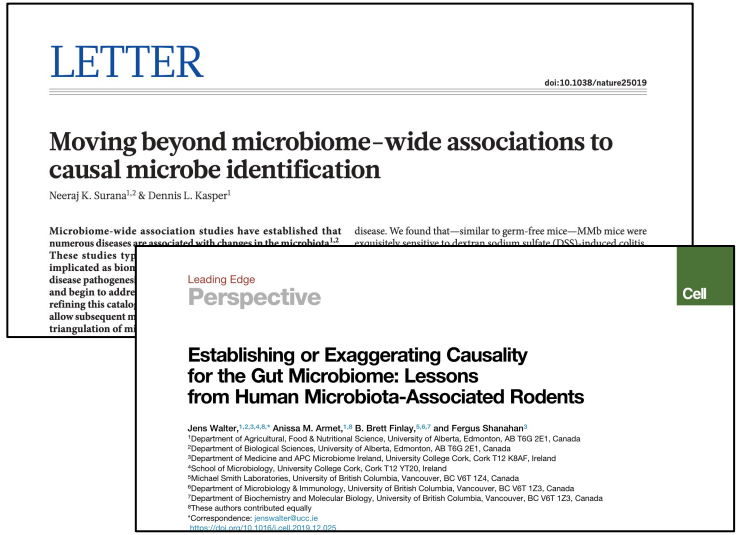
- Two studies reported associations between the gut microbiome and type 2 diabetes
  - However, there was little overlap in the set of associated taxa
- Metformin is a common medication for treatment of type 2 diabetes
- Metformin alters the composition of the gut microbiome



[Forslund et al. *Nature* 2015]

Caveat: association does not imply causation





LETTER

doi:10.1038/nature25019

Moving beyond microbiome-wide associations to causal microbe identification

Neeraj K. Surana[1,2] & Dennis L. Kasper[1]

Microbiome-wide association studies have established that numerous diseases are associated with changes in the microbiota[1,2] These studies typ implicated as bion disease pathogenes and begin to addre refining this catalo allow subsequent m triangulation of mi

disease. We found that—similar to germ-free mice—MMb mice were exquisitely sensitive to dextran sodium sulfate (DSS)-induced colitis

Leading Edge

Perspective

Cell

Establishing or Exaggerating Causality for the Gut Microbiome: Lessons from Human Microbiota-Associated Rodents

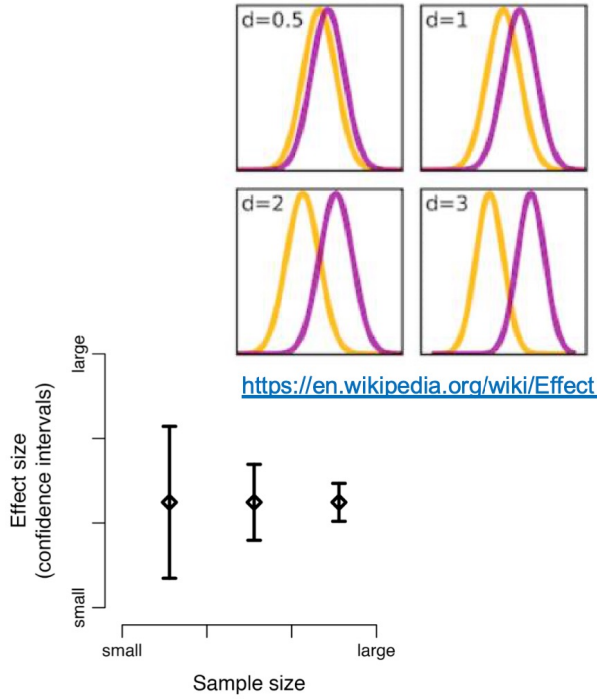
Jens Walter,[1,2,3,4,8,*] Anissa M. Armet,[1,8] B. Brett Finlay,[5,6,7] and Fergus Shanahan[3]
[1]Department of Agricultural, Food & Nutritional Science, University of Alberta, Edmonton, AB T6G 2E1, Canada
[2]Department of Biological Sciences, University of Alberta, Edmonton, AB T6G 2E1, Canada
[3]Department of Medicine and APC Microbiome Ireland, University College Cork, Cork T12 K8AF, Ireland
[4]School of Microbiology, University College Cork, Cork T12 YT20, Ireland
[5]Michael Smith Laboratories, University of British Columbia, Vancouver, BC V6T 1Z4, Canada
[6]Department of Microbiology & Immunology, University of British Columbia, Vancouver, BC V6T 1Z3, Canada
[7]Department of Biochemistry and Molecular Biology, University of British Columbia, Vancouver, BC V6T 1Z3, Canada
[8]These authors contributed equally
*Correspondence: jenswalter@ucc.ie
https://doi.org/10.1016/j.cell.2019.12.025

Caveat: significance not to be confused with effect size

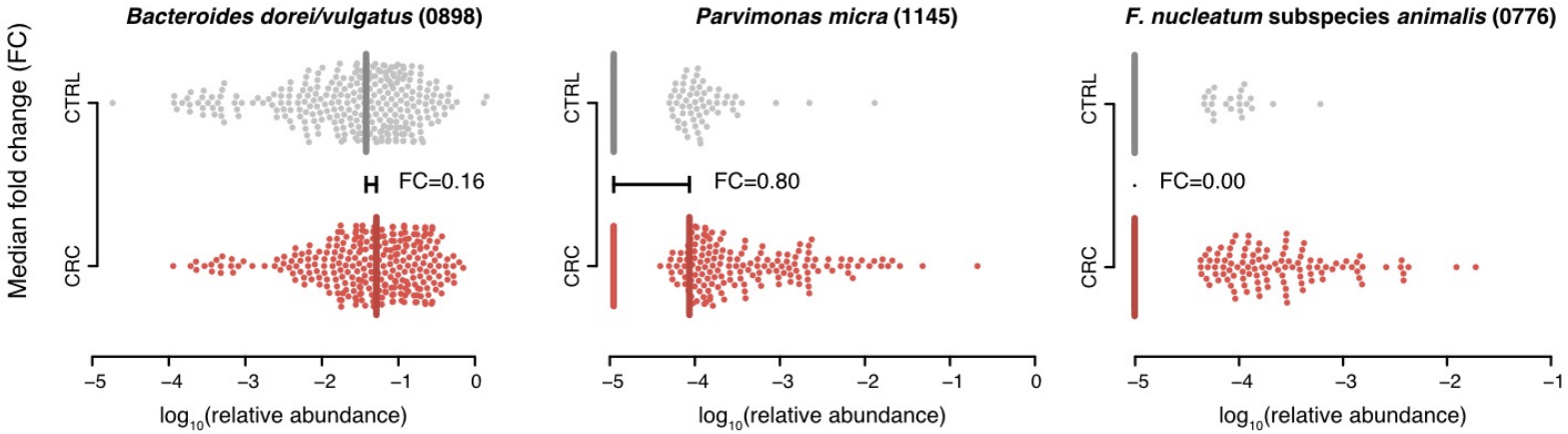
- Statistical significance does not mean that the difference is big, important or biologically significant.
  It simply means you can be confident that there is a difference.

- Any (even a tiny) difference can create a significant results if the sample size is large enough

- What is a good effect size measure for microbiome data?

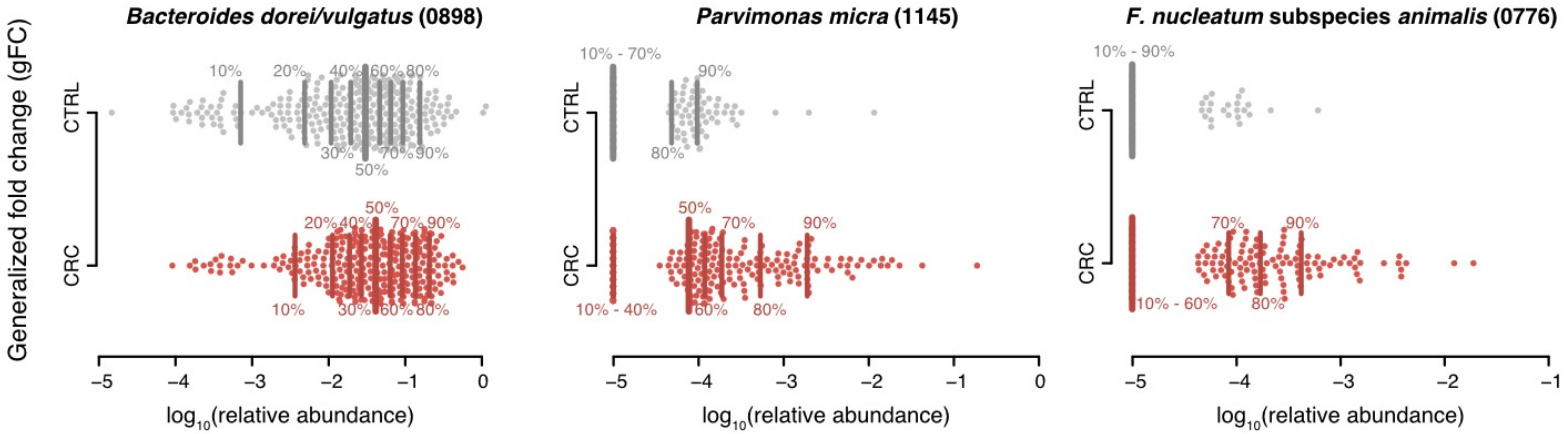


Generalized fold change as measure for effect size



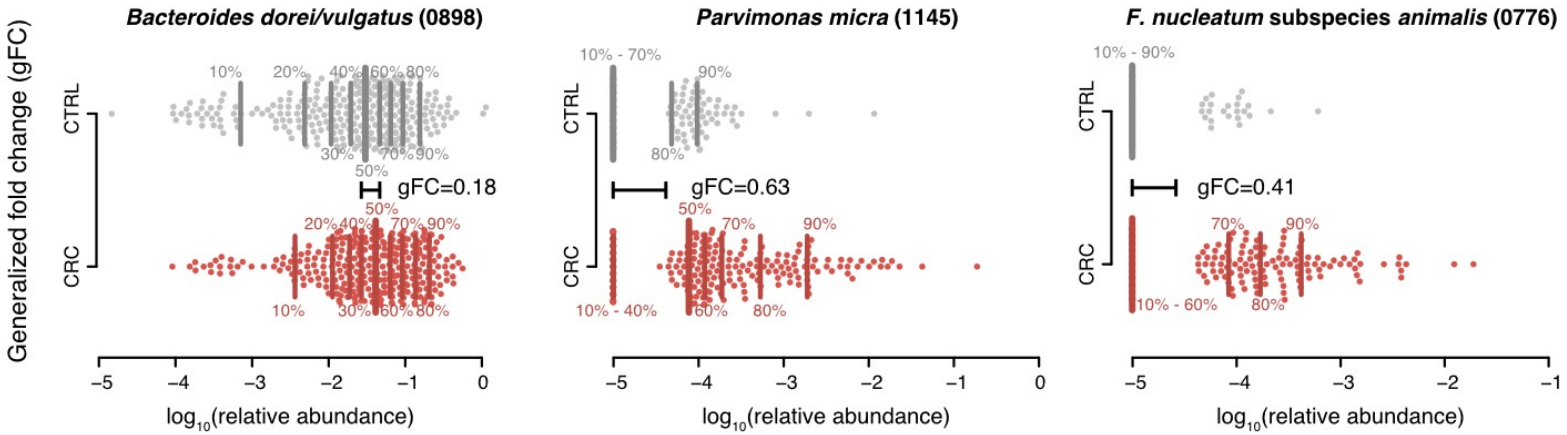
[Wirbel et al., *Genome biology* 2021]

Generalized fold change as measure for effect size



[Wirbel et al., *Genome biology* 2021]

Generalized fold change as measure for effect size



[Wirbel et al., *Genome biology* 2021]

Exercises

- Download the provided dataset with healthy and CRC samples profiled with mOTUs
- Try to identify which mOTUs are enriched or depleted in colorectal cancer patients
- Use SIAMCAT association testing on the samples you downloaded