# Tools for comparative metagenomics

EBI Metagenomics Bioinformatics Course
November 2020

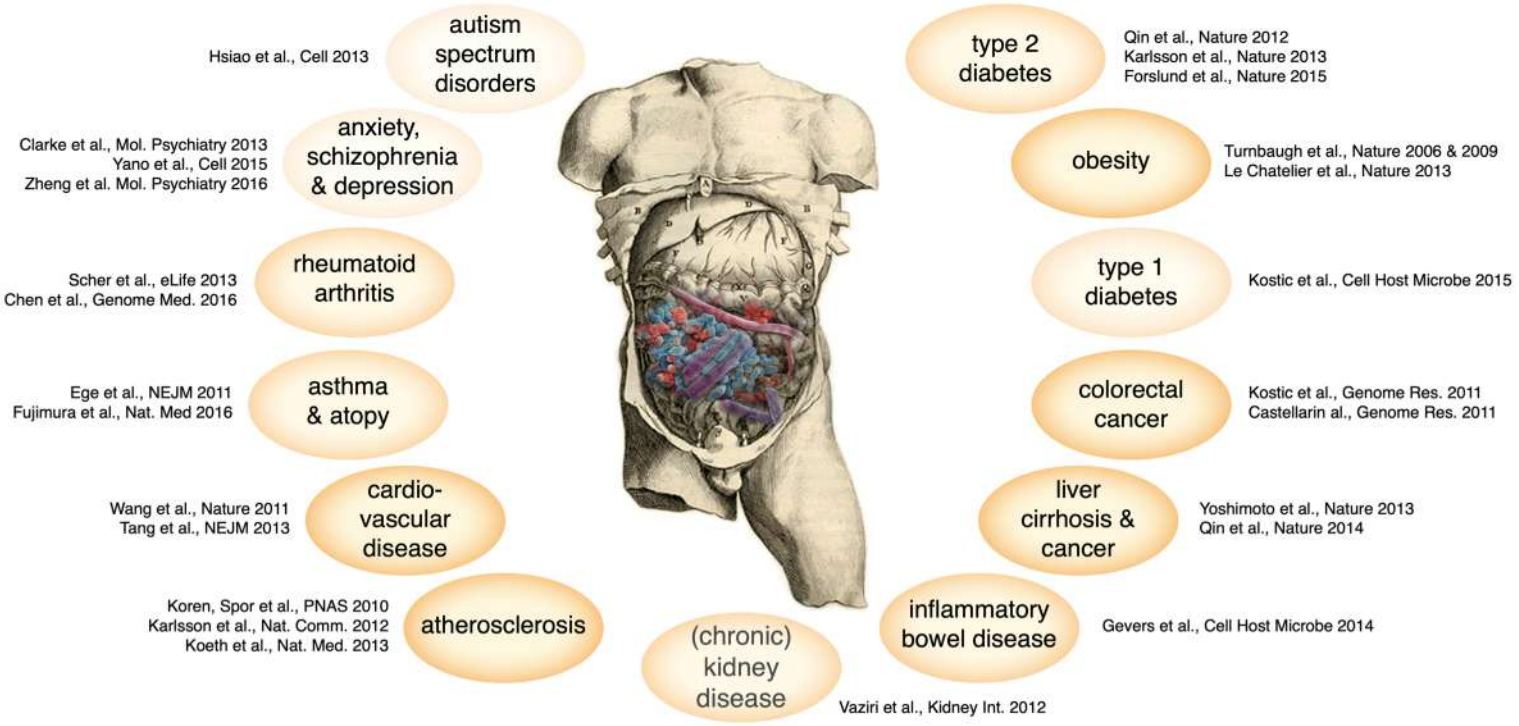
Georg Zeller & Jakob Wirbel

de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

EMBL

# Univariate statistical tests for metagenomic data

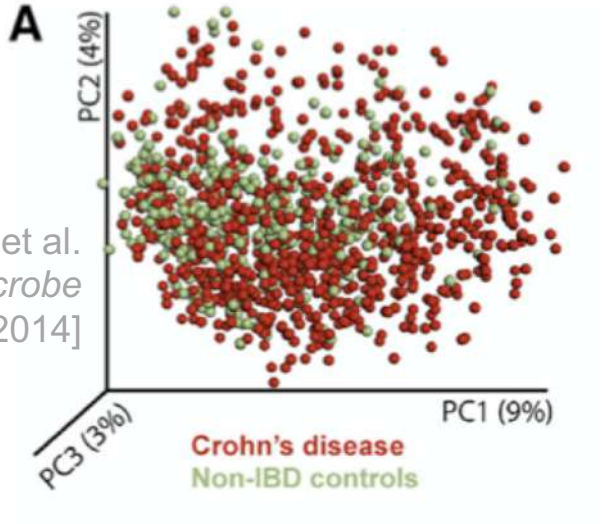
# Comparing microbiome composition in case-control studies



# Tools for microbial community comparison

**Assessing difference in overall community structure**

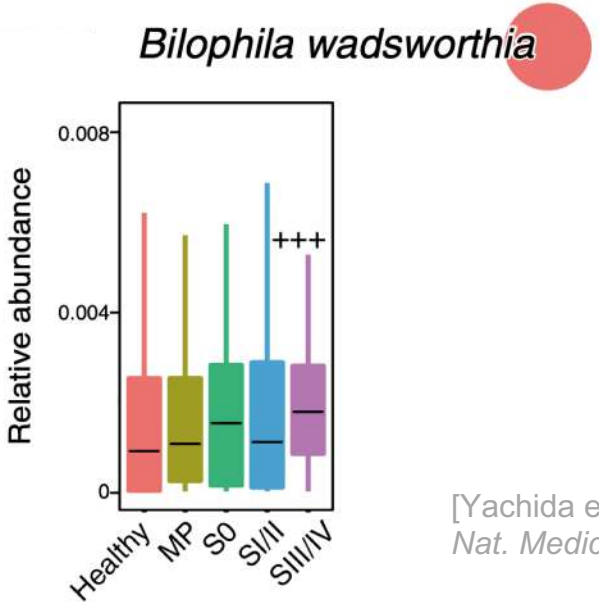
- Clustering
- Ordination




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

**Testing for changes in individual taxa**

- Statistical testing

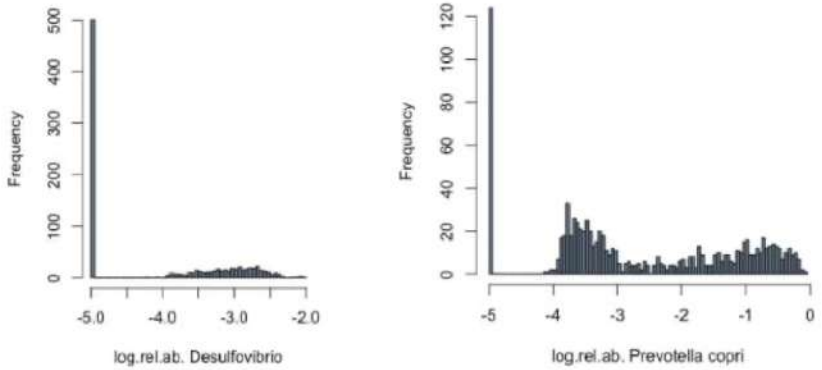


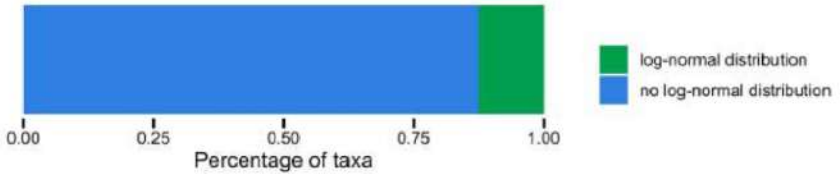

[Yachida et al. *Nat. Medicine* 2019]

# Which statistical test is appropriate?

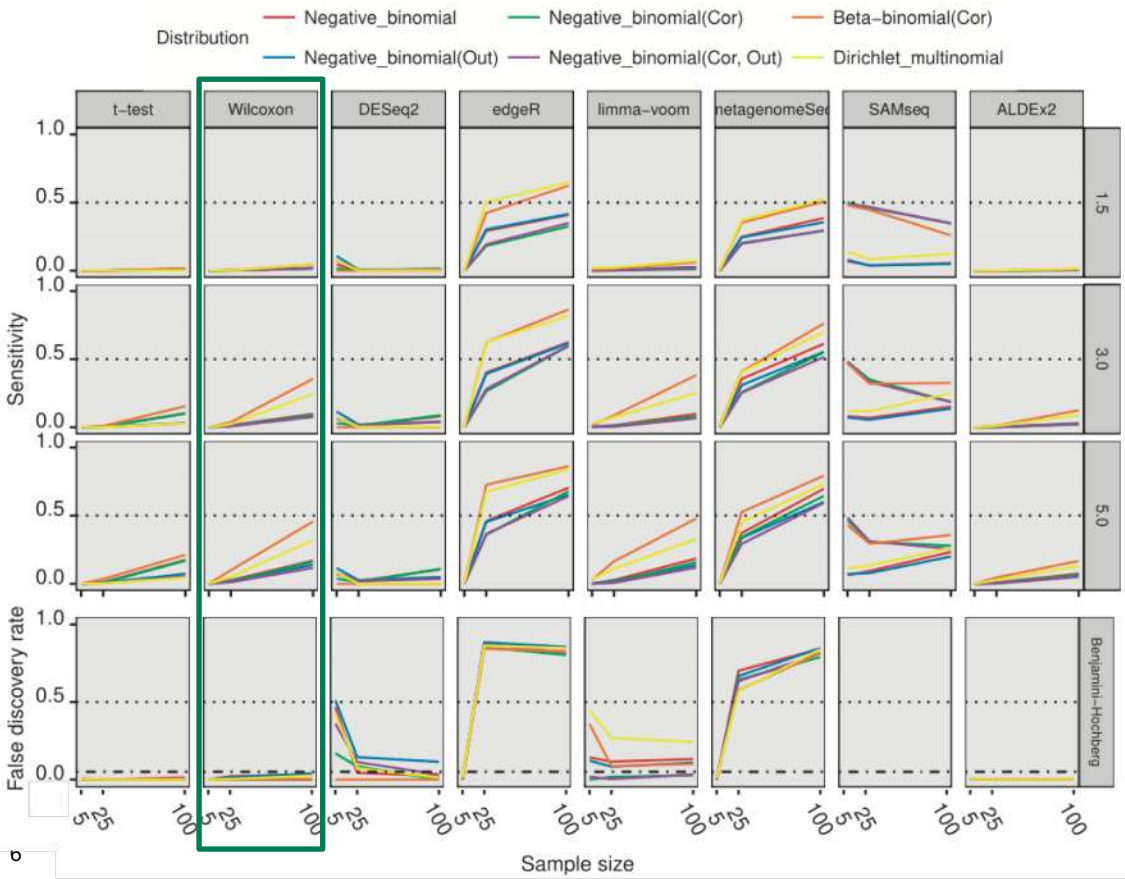
Some things to keep in mind:

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



# Nonparametric Wilcoxon test suitable for metagenomics

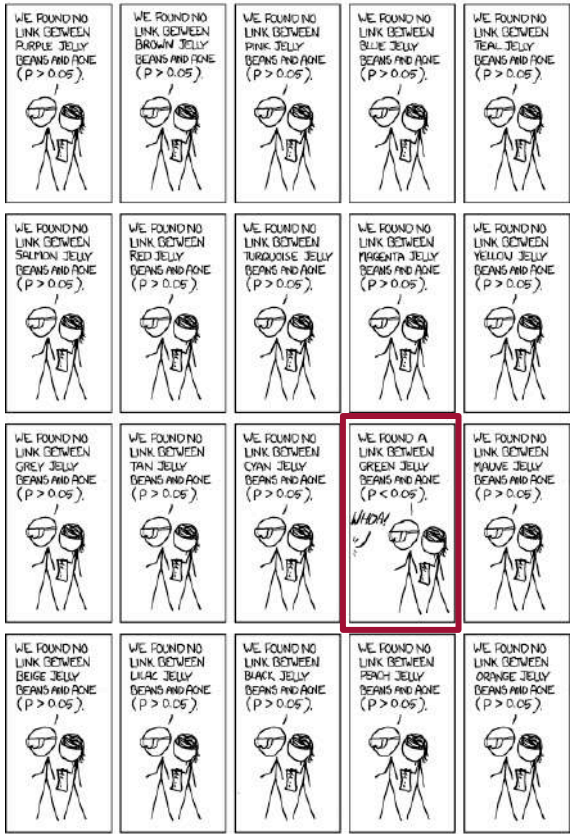


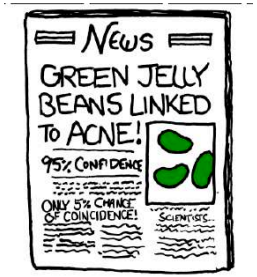

Simulations suggest the Wilcoxon test to…

- maintain false-discovery-rate control,
- have reasonable sensitivity (stat. power), which increases with sample size.

[Hawinkel et al. *Brief. Bioinform.* 2017]

# Multiple testing correction

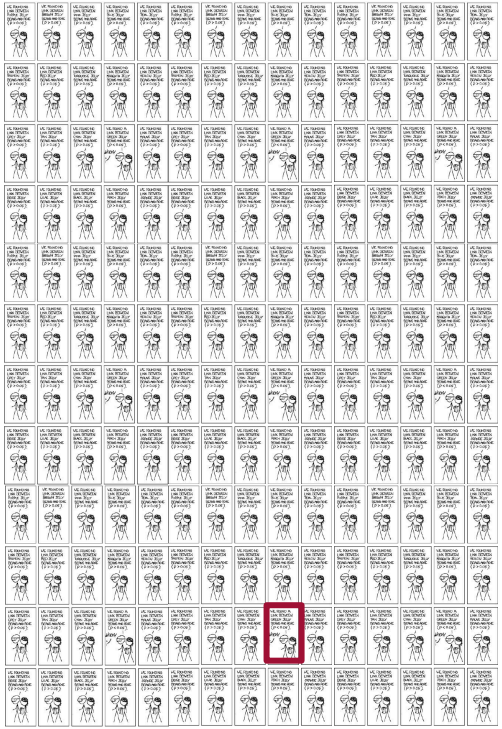


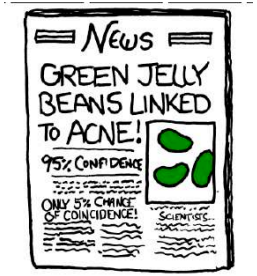




- Since we test several hundreds of taxa, some tests will be “significant” by chance
- It is thus crucial to perform a multiple testing correction, e.g.
  - The Benjamini-Hochberg procedure controls the false discovery rate (proportion of true positives among those for which the null hypothesis is rejected)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



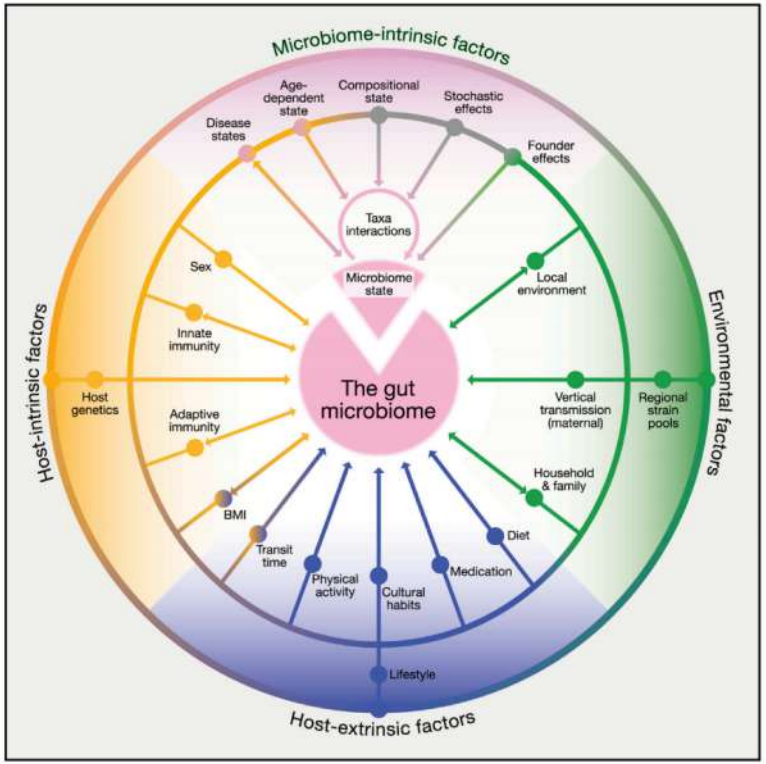



- Since we test several hundreds of taxa, some tests will be “significant” by chance
- It is thus crucial to perform a multiple testing correction, e.g.
  - The Benjamini-Hochberg procedure controls the false discovery rate (proportion of true positives among those for which the null hypothesis is rejected)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

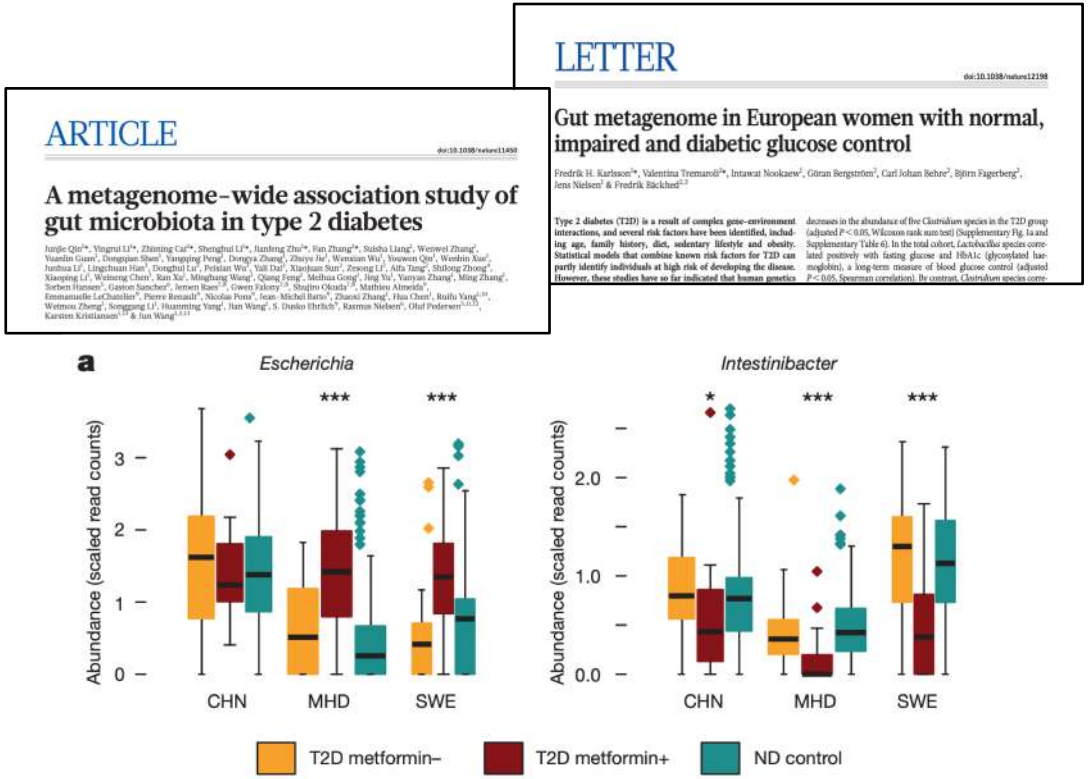
- Technical factors can strongly affect microbial community profiles (**batch effects**), e.g. DNA extraction protocols, sequencing approach (16S primers), bioinformatic profiling
- Biological factors other than that of interest can affect profiles (**confounders**), e.g. medication, lifestyle, host demographics




[Schmidt et al. *Cell* 2018]

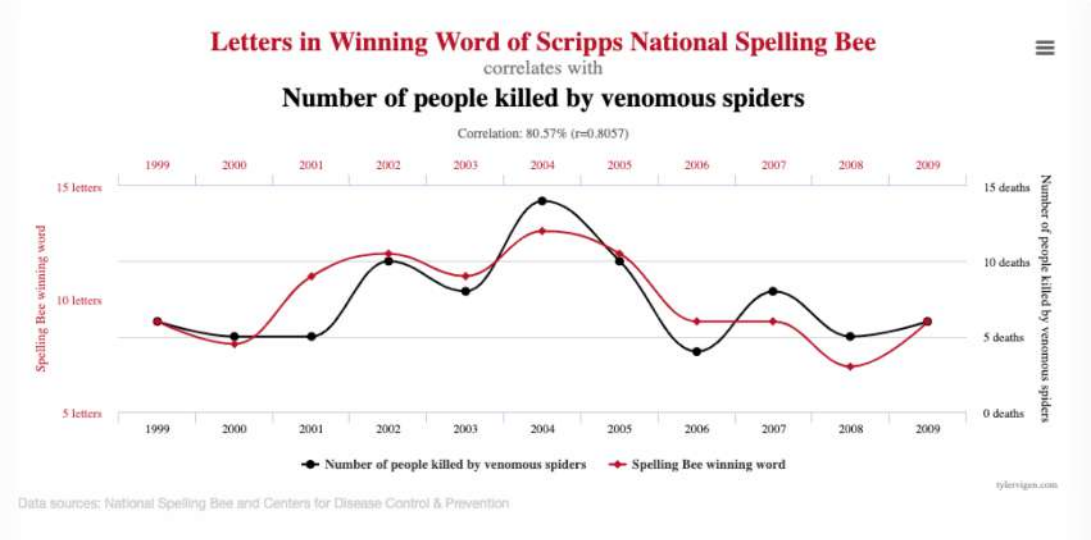
# Caveat: confounding (here due to metformin)

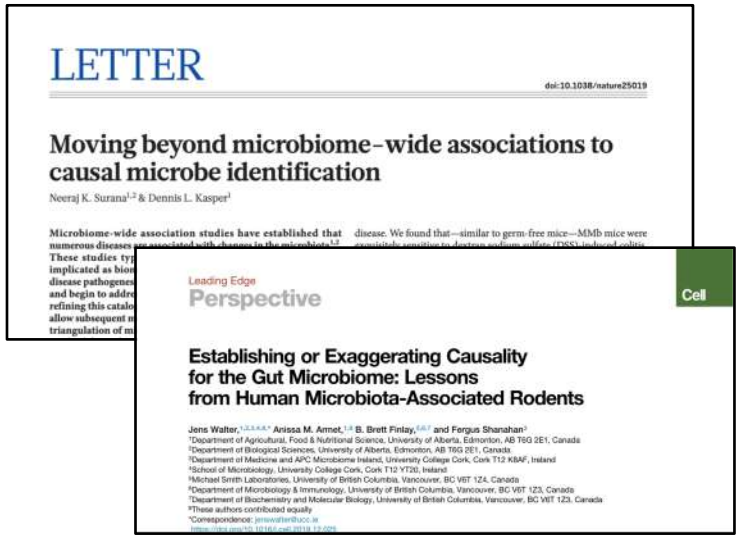
- Two studies reported associations between the gut microbiome and type 2 diabetes
  - However, there was little overlap in the set of associated taxa
- Metformin is a common medication for treatment of type 2 diabetes
- Metformin alters the composition of the gut microbiome




[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 n triangulation of m

disease. We found that—similar to germ-free mice—MMb mice were

Leading Edge

Perspective

Cell

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

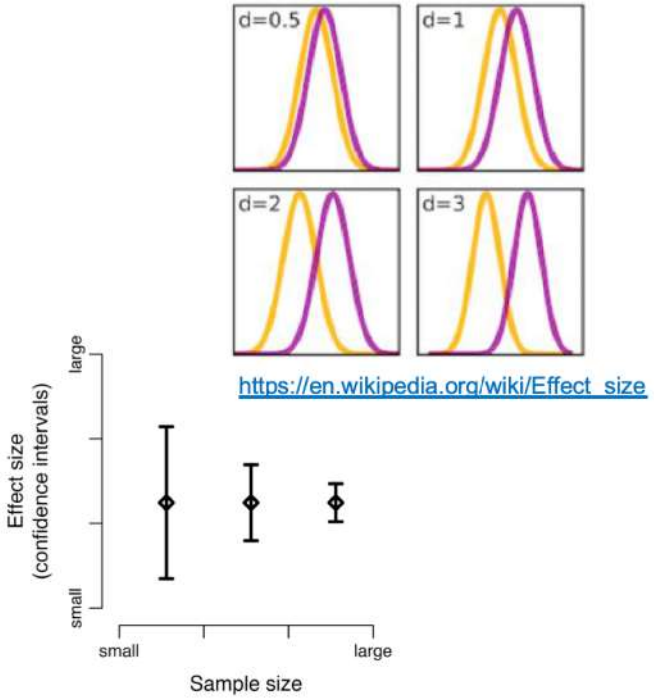
Jens Walter,[1,2,3,4,8,*] Anissa M. Armet,[1,8] B. Brett Finlay,[5,6,7] and Fergus Shanahan[3]
[1]Department of Agricultural, Food & Nutritional Science, University of Alberta, Edmonton, AB T6G 2E1, Canada
[2]Department of Biological Sciences, University of Alberta, Edmonton, AB T6G 2E1, Canada
[3]Department 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.2019.12.029

# Caveat: significance not to be confused with effect size

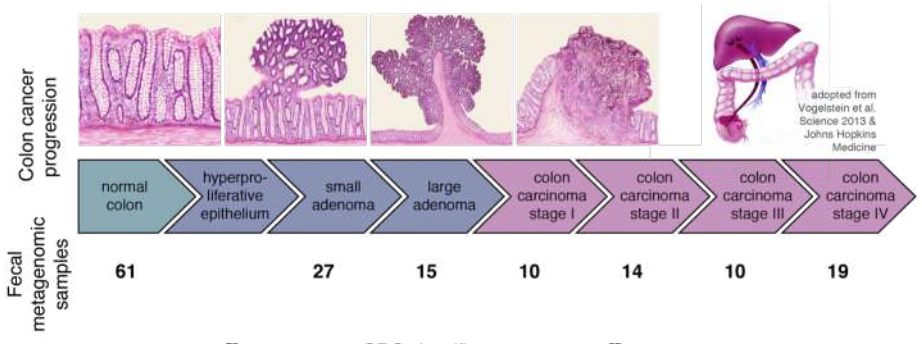
- Statistical significance does not mean that the difference is big, important or biologically significant.
  It simply means you can be confident that there is a difference.

- 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?



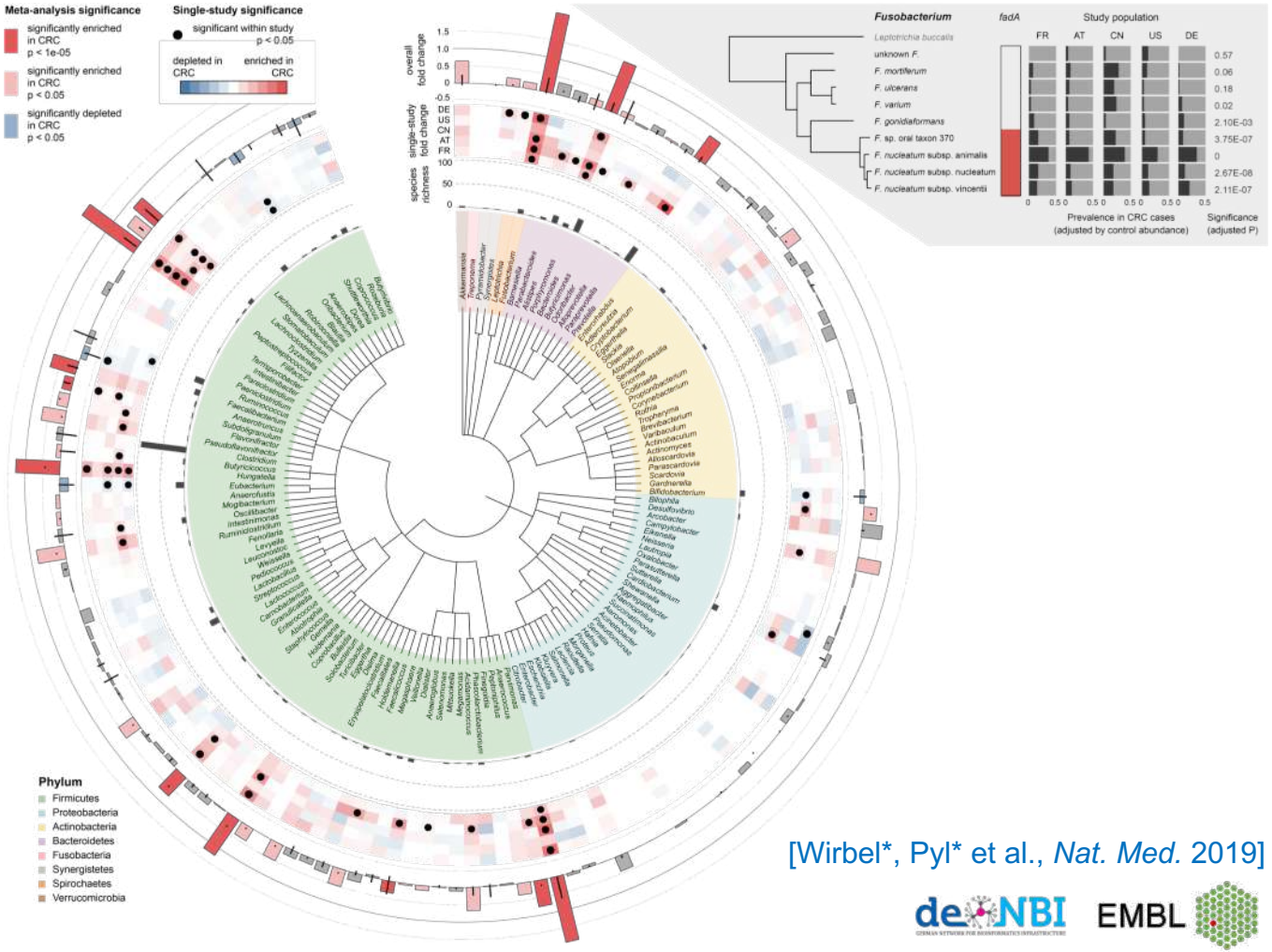
# Colorectal cancer (CRC) as an introductory example




- Collected stool samples from 53 colorectal cancer (CRC) patients and 88 healthy controls
- Used metagenomic sequencing and profiled gut bacterial species
- Can microbiome differences be used for non-invasive detection of cancer?

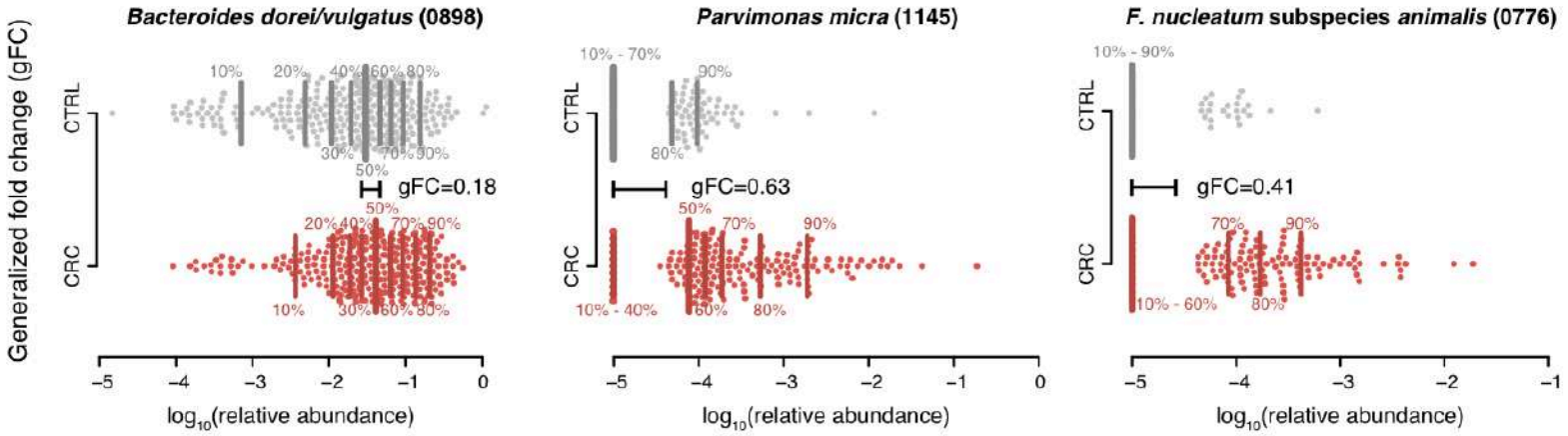
[Zeller*, Tap*, Voigt* et al., *Mol. Syst. Biol.* 2014]
[Wirbel*, Pyl*, et al., *Nat. Med.* 2019]

# Statistically significant associations with CRC across five studies



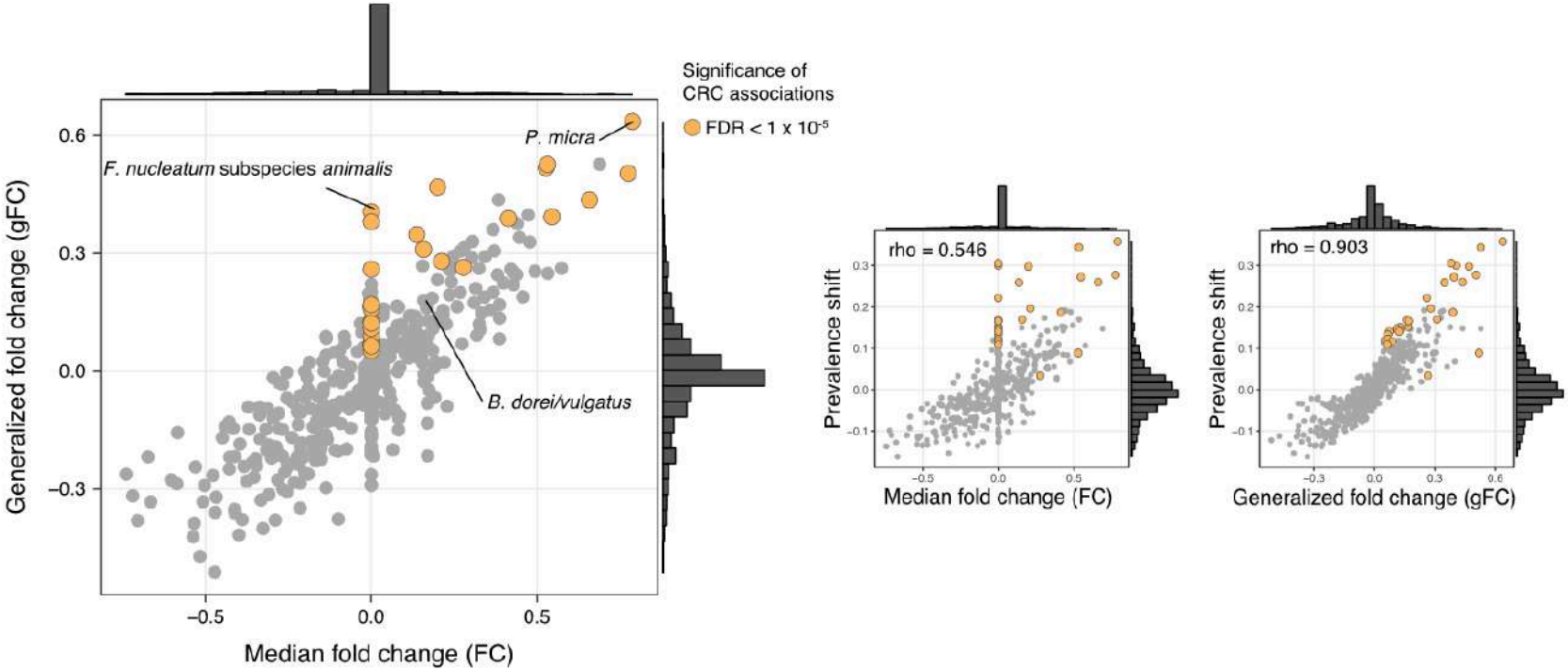

[Wirbel*, Pyl* et al., *Nat. Med.* 2019]

# Generalized fold change as measure for effect size




[Wirbel*, Pyl* et al., *Nat. Med.* 2019]

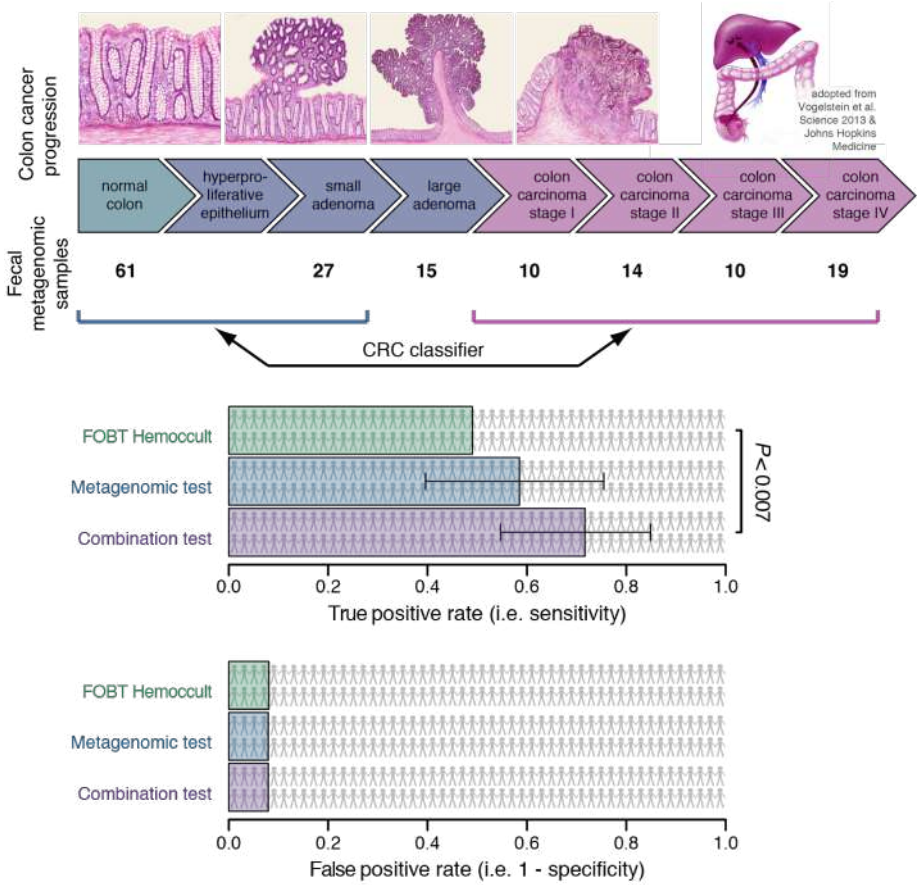
# Generalized fold change as measure for effect size




[Wirbel*, Pyl* et al., *Nat. Med.* 2019]

# Machine learning / statistical modelling of metagenomic data

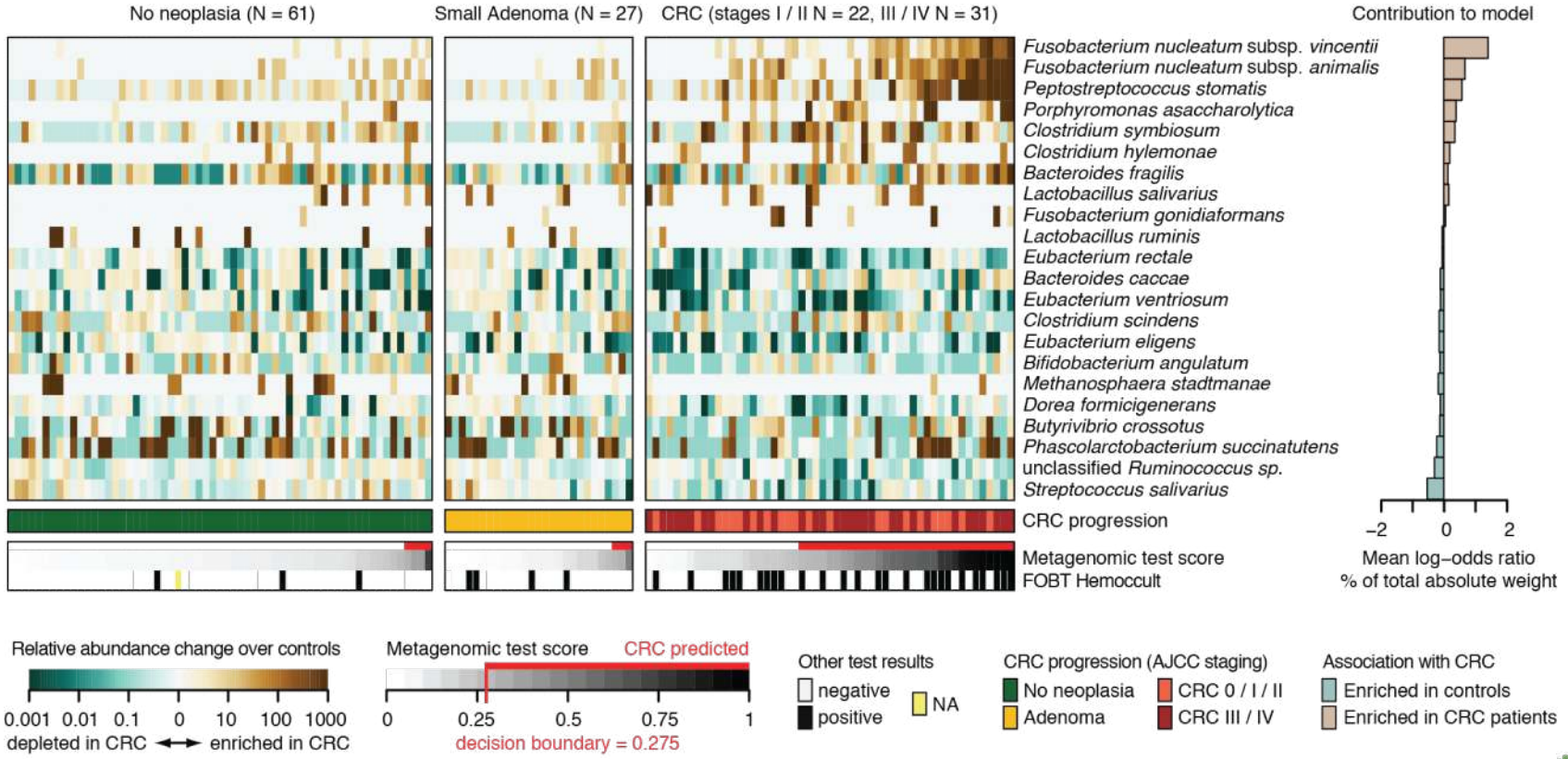
# Colorectal cancer example (continued)




- Collected stool samples from 53 colorectal cancer (CRC) patients and 88 healthy controls
- Used metagenomic sequencing and profiled gut bacterial species
- Can microbiome differences be used for non-invasive detection of cancer?
- How does metagenomic detection compare to standard noninvasive diagnostic test (FOBT)?

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

# A microbiome “signature” of colorecatal cancer

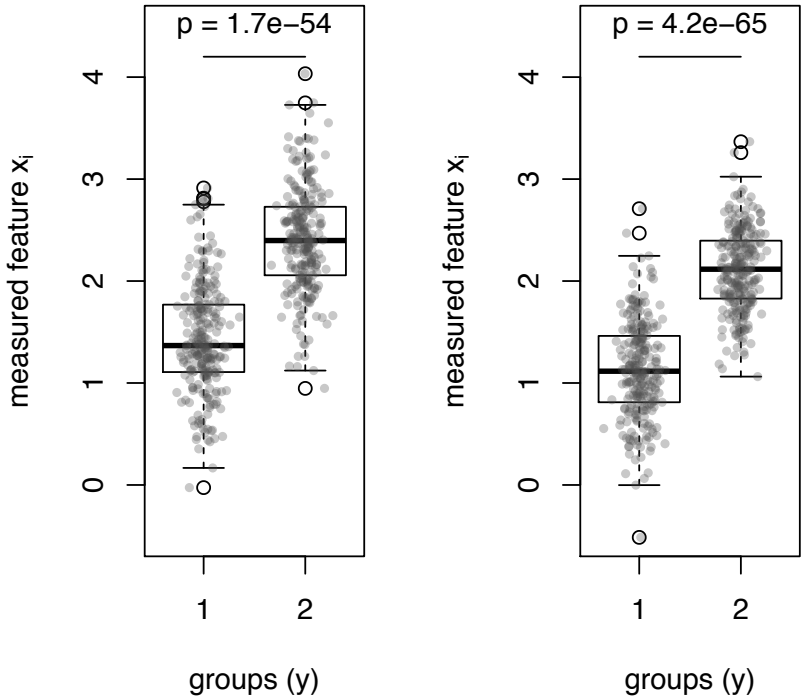





[Zeller*, Tap*, Voigt* et al., *Mol. Syst. Biol.* 2014]

# Descriptive statistics versus statistical modeling

- **Hypothesis testing:**
  Could the observed difference also be observed by chance?
- **Modeling:**
  Given only the measurement, can we tell which group the measurement corresponds to?
- Recall that ***P*-values depend on both effect size and sample size!**



# Why statistical modelling / machine learning?

- Modeling ideally **extracts the essence** of a biological phenomenon
- Model needed to **make predictions on new data**
  (necessary e.g. for microbiome-based diagnostics)
- **Prediction accuracy** is often a more **meaningful measure of association** than statistical significance of differences
- Suitable methods can **select predictive taxa** (and ignore others)
- **Sparse statistical models** are based on only „few“ taxa,
  therefore useful for microbiome **biomarker / signature extraction**

$$y_i = f(\mathbf{x}_i) + \varepsilon$$

For $i$ samples / patients
$y_i$ – label (e.g. disease or control), always binary herein
$x_i$ – features (e.g. species abundance profile, a vector)
$f$ – our model
$\epsilon$ – modeling error

# Introduction to notation and input data format

- **Feature** data **X** (also observations, predictors):
  n x p matrix $x_{ij}$
  **species/gene abundances** in rows (i),
  **samples/patients** in columns (j)

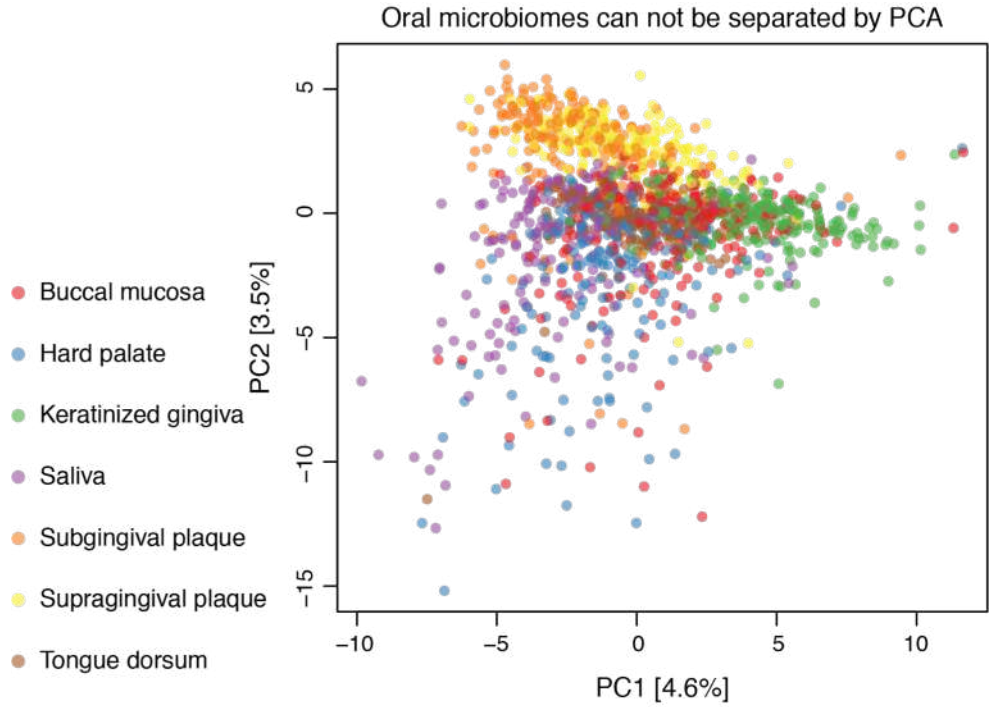
  observations based on which we wish to make predictions
  $\mathbf{x}_i$ denotes the feature vector, i.e. abundance profile, for the i-th sample

- **Label** data **y** (also dependent variable, response):
  vector of length n, containing binary values in our cases

  the phenomenon which we wish to predict:
  **disease vs. healthy, response vs. non-response** etc.

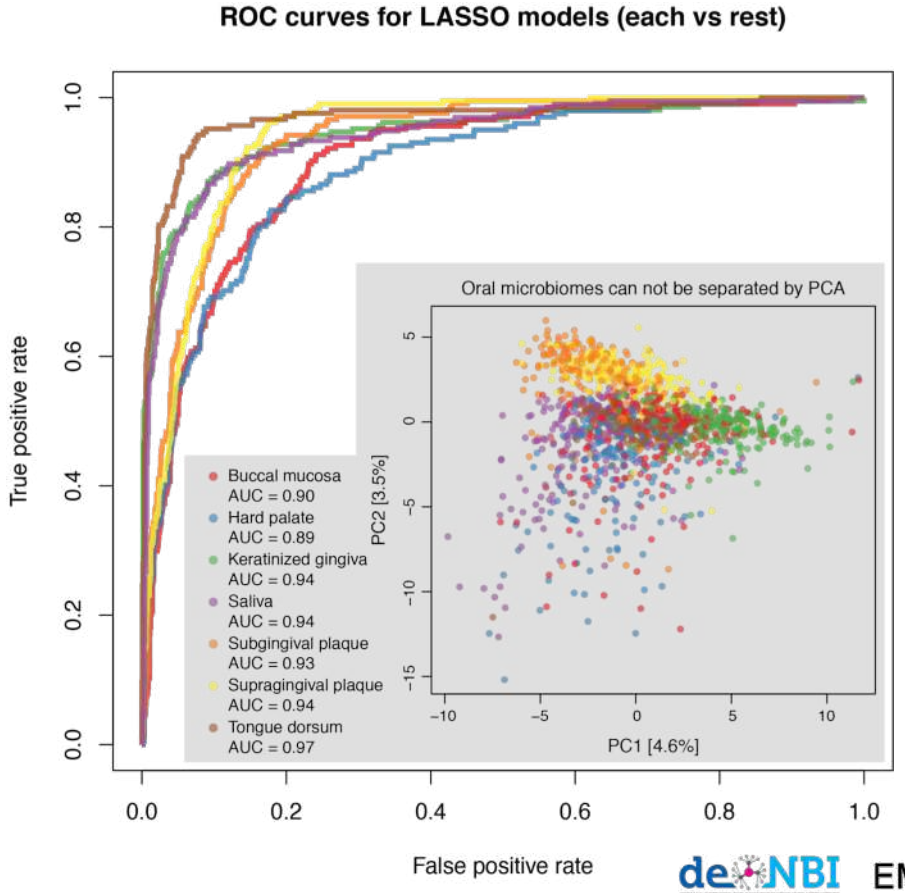
# Ordination versus modelling (I)

- Using PCoA (with various dissimilarity measures), it is difficult to resolve for each oral microbiome sample the precise sampling site.

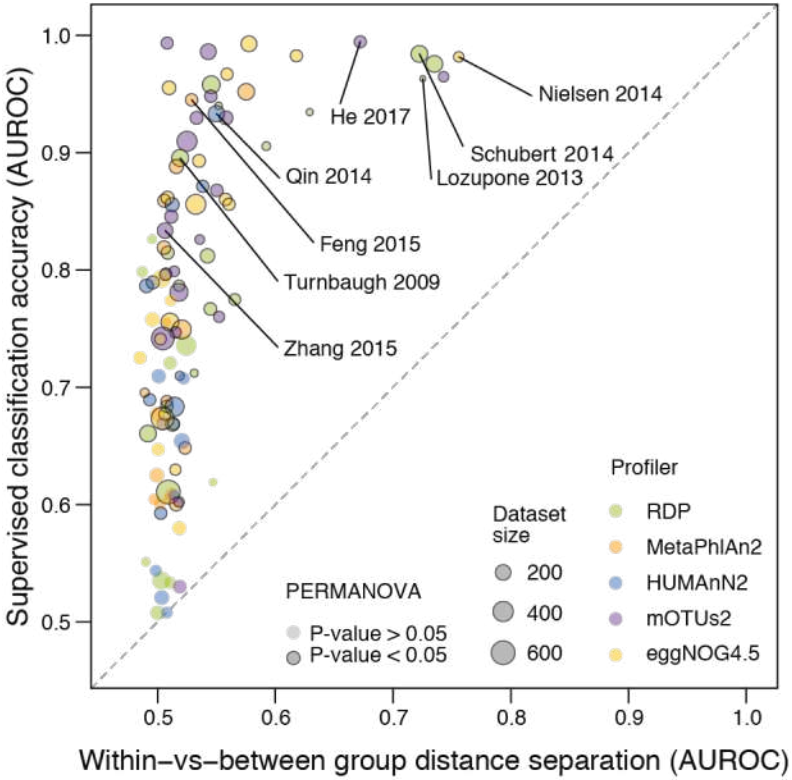


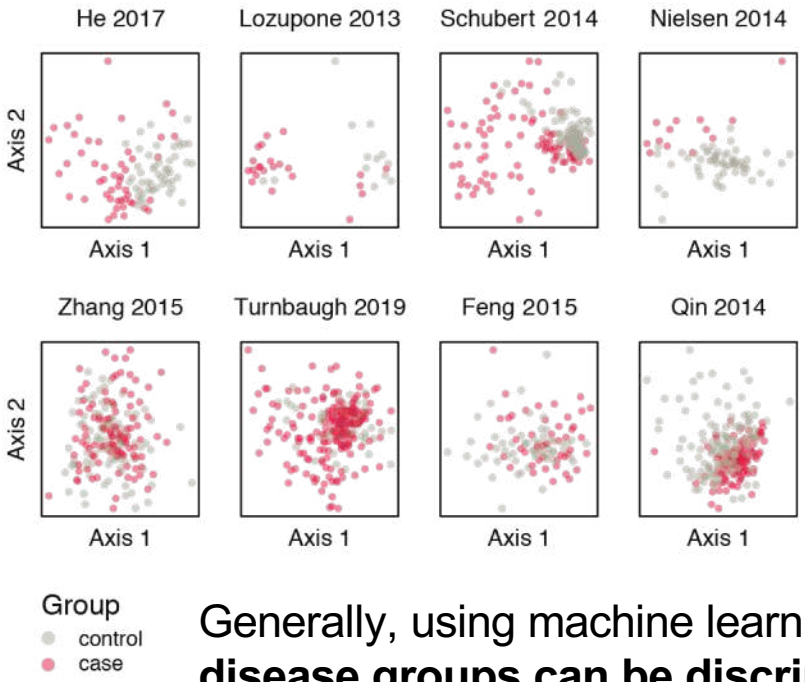
# Ordination versus modelling (I)

- Using PCoA (with various dissimilarity measures), it is difficult to resolve for each oral microbiome sample the precise sampling site.
- Statistical models, in contrast, can very accurately recognize sample origin.



# Ordination versus modelling (II)




Generally, using machine learning, **disease groups can be discriminated from controls more accurately** than based on ecological distances (beta-diversity analysis).

# A typical machine learning workflow


[Wirbel et al., BioRxiv 2020]

**siamcat.embl.de**

SIAMCAT

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

phyloseq

mlR

docker

COMMON WORKFLOW LANGUAGE

## Starting with SIAMCAT

```
> source("https://bioconductor.org/biocLite.R")
> biocLite("SIAMCAT")
> browseVignettes("SIAMCAT")
```

File formats supported:
- phyloseq
- BIOM
- LEfSe
- MaAsLin
- metagenomeSeq

EMBL MICROBIOME TOOLS

microbiome-tools.embl.de

This workflow is implemented in the SIAMCAT Bioconductor package, which we will explore in detail in the practical.



de.NBI EMBL

# What to use as input (features)?

- Use your **domain expertise** to engineer features that are likely predictive of the phenomenon of interest – microbiome examples:
  - Species abundances (or higher / lower resolution taxonomic profiles)
  - Metabolic pathway abundance (e.g. KEGG / CAZy maps)
  - Functional gene annotations (GO terms, domains, …)
  - Orthologous gene families (COGs, eggNOG families, …)
  - Toxins, virulence factors, ABX resistance genes, …

- Consider **interpretability** –
  predictive species/metabolic pathways may be preferred over k-mers or log-ratios

- Importantly, do **NOT use the label** information for selecting features for modeling (more on this later)

# Model evaluation (classification)

In many applications, classes aren't equal – neither are errors!

| | | True condition | |
|---|---|---|---|
| | | positive ("cancer") | negative ("healthy") |
| **Predicted condition** | positive ("predicted to have cancer") | True positives<br>TP | False positives<br>FP<br>(Type I errors) |
| | negative ("predicted not to have cancer") | False negatives<br>FN<br>(Type II errors) | True negatives<br>TN |

True positive rate (TPR, **sensitivity**, **recall**)
True negative rate (TNR, **specificity**)
False positive rate (FPR, 1 – specificity )

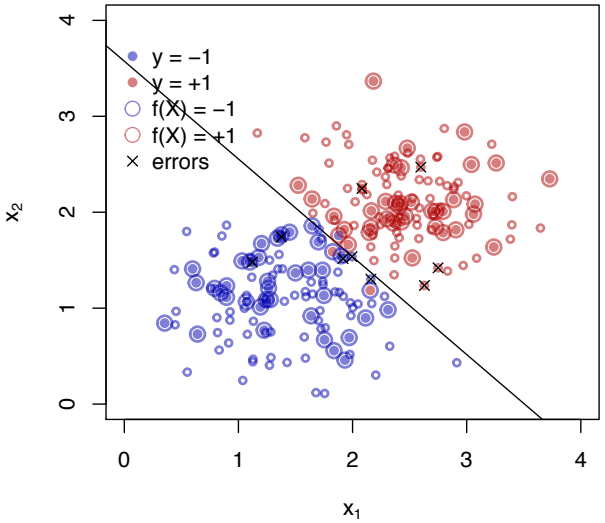
- are all **independent of prevalence** (fraction of positives in the population)

Precision (positive pred. value, PPV)
False discovery rate (FDR, 1 – precision)

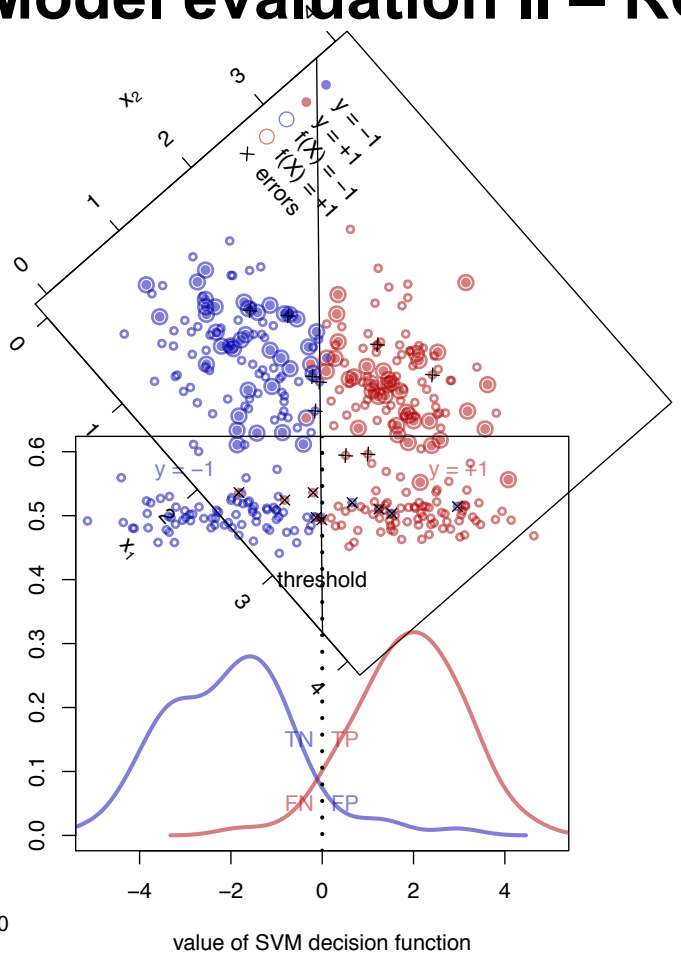
- are both **dependent on prevalence** (fraction of positives in the population)

[these and more measures on en.wikipedia.org/wiki/Evaluation_of_binary_classifiers]

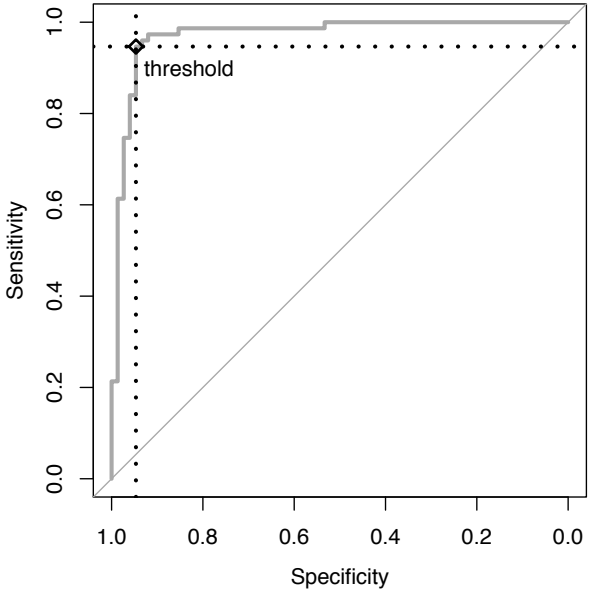
# Model evaluation II – ROC curves

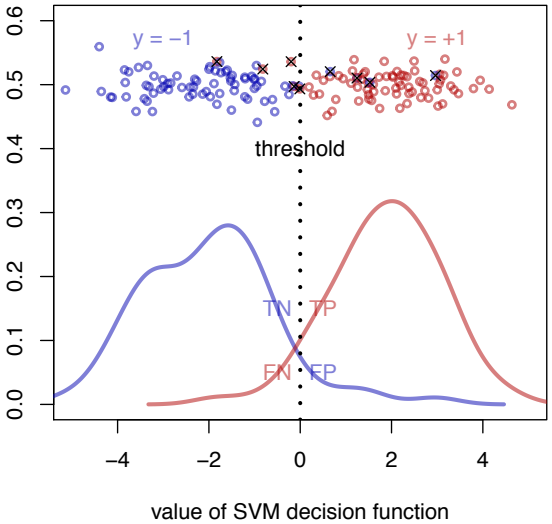


# Model evaluation II – ROC curves



# Model evaluation II – ROC curves

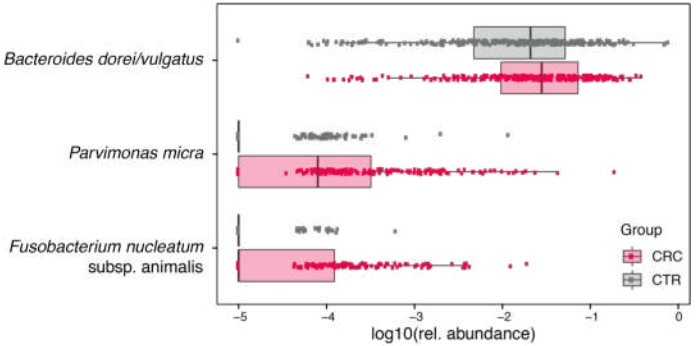


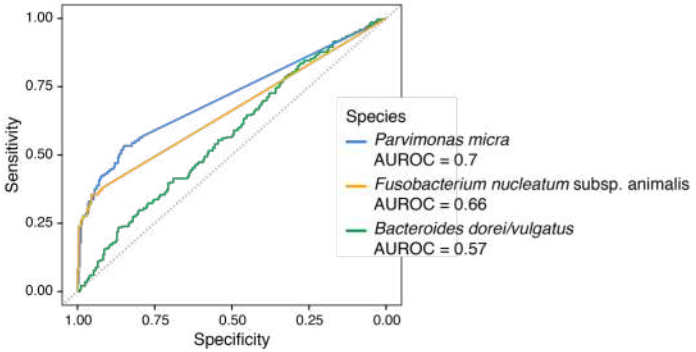




- Change decision threshold to obtain other **trade-offs between sensitivity and specificity**
- Receiver operating characteristic (ROC) curve plots all of them
- **Area under the ROC curve** as a summary statistic

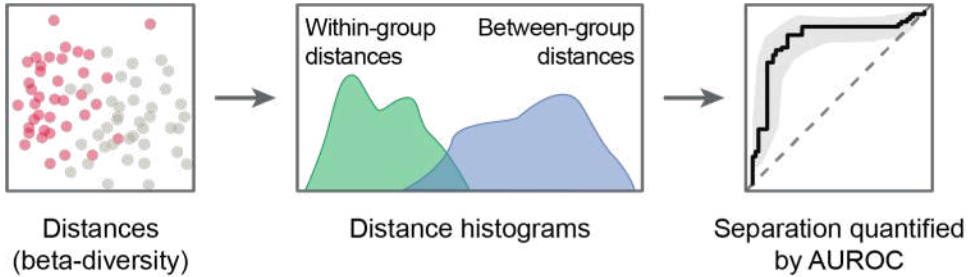
# ROC curves from single features / distances

- Enrichment of a species in disease group can be directly quantified using ROC curves (disease biomarker).

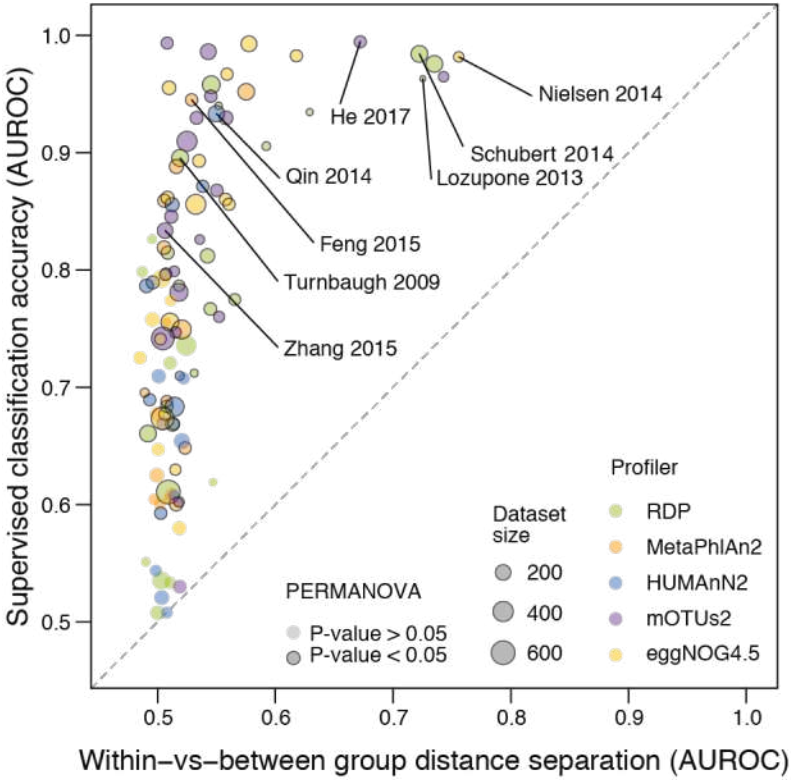


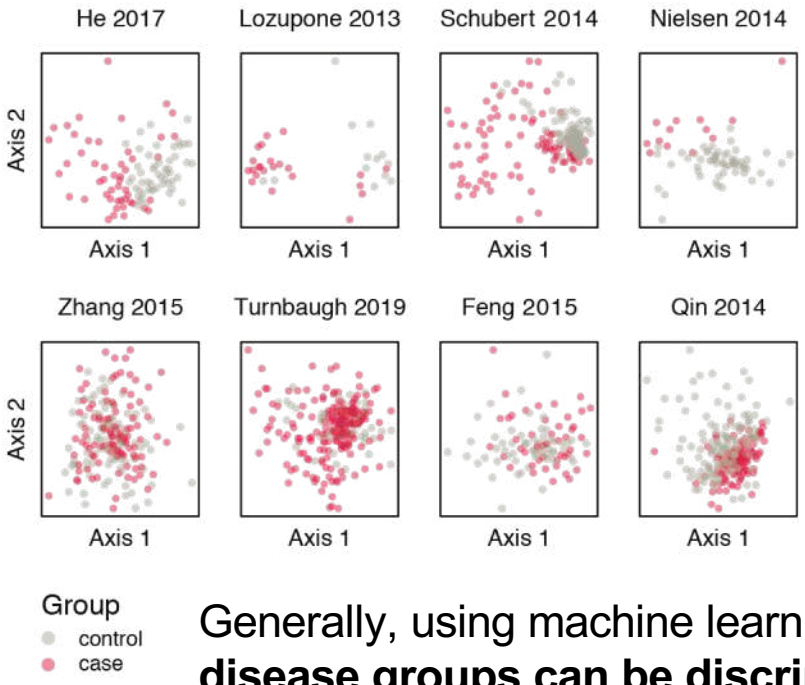




- Separation between groups in terms of pairwise dissimilarities can also be assessed using ROC curves.



# Ordination versus modelling (II) - revisited

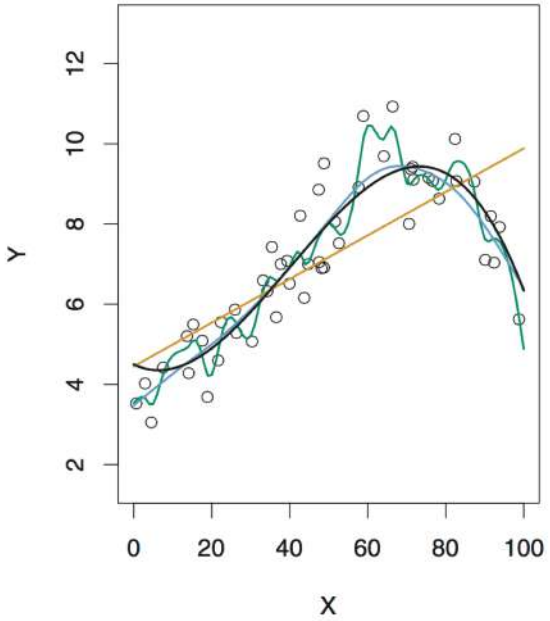


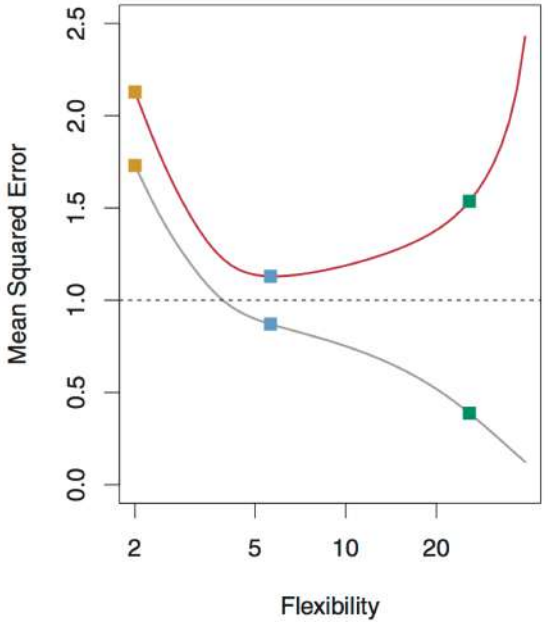




Generally, using machine learning, **disease groups can be discriminated from controls more accurately** than based on ecological distances (beta-diversity analysis).

# Model evaluation III – assessing generalization

- What might seem a good idea at first: Minimizing the **training error**…
  But with increasing flexibility, models will fit the training data better and better.
- Better: maximize **generalization** to new data sets…
  Since **overfitting** the training data will result in poor generalization (i.e. large **test error**)




Here for illustration, smoothing splines are used where model flexibility / complexity increases with the degree of the polynomials.

[James, Witten, Hastie & Tibshirani, *Springer* 2013]

# Resampling data for external validation or cross validation

Some data needs to be reserved for model evaluation….

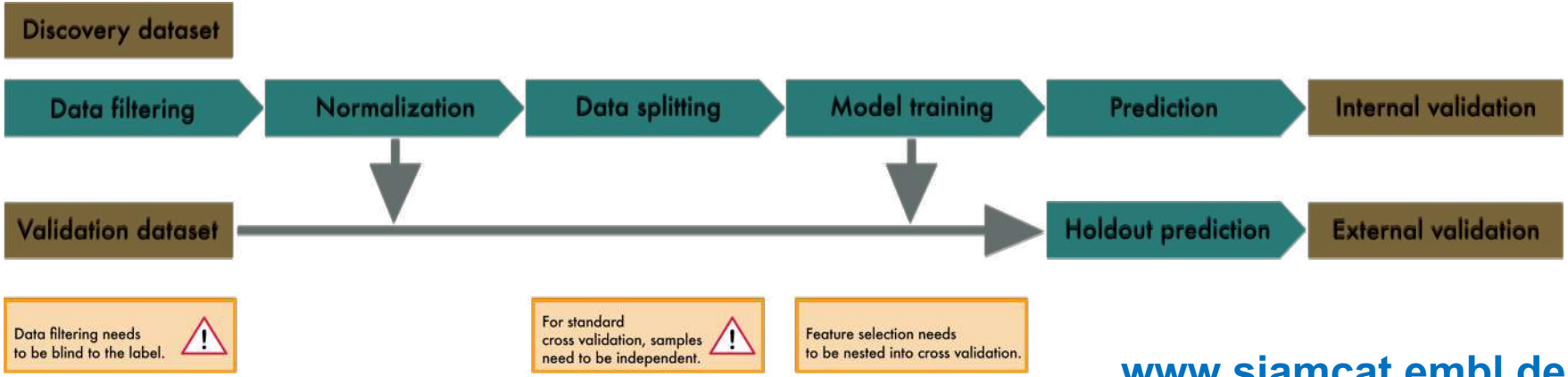
- Validation on external data


- Train model on training set
- Test on test set
- Assess error on test predictions

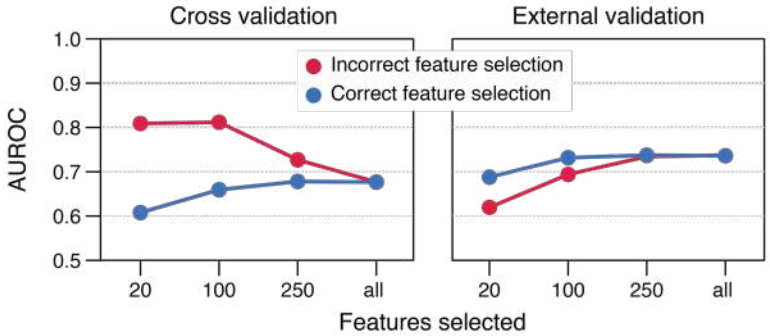
- Cross-validation (CV)


- For each CV fold:
  - Train a model on training set
  - Predict on the test set
- Either concatenate or average predictions from (all) test sets to estimate error
- More efficient use of (training) data

# Cross-validation pitfalls I – illustration




**www.siamcat.embl.de**




Importantly, all components of model building, **including feature selection,** need to be **nested into cross validation** and/or externally validation to **avoid overfitting**!

[Wirbel et al., *bioRxiv* 2020]

# Cross-validation pitfalls II

- **Cross validation works under the i.i.d. assumption** (observations have the same probability distribution and are mutually independent)
  - E.g. a series of (fair or unfair) coin flips is i.i.d. as the next flip doesn't depend on the previous ones.

- However, biological samples are **rarely completely independent**:
  - Multiple time-point measurements from the same subject or related subjects
  - Spatial structure / dependencies between measurements

- Data (sets) are **not always identically distributed**
  - Batch effects: e.g. experiments or diagnostic tests performed in different labs (by different technicians, at different times, using different reagent lots, …) may exhibit (subtle) distributional shifts

# Take home messages

- **Model fitting is easy, model evaluation is not at all**!
  Understand the generalization assessed – consult experts!

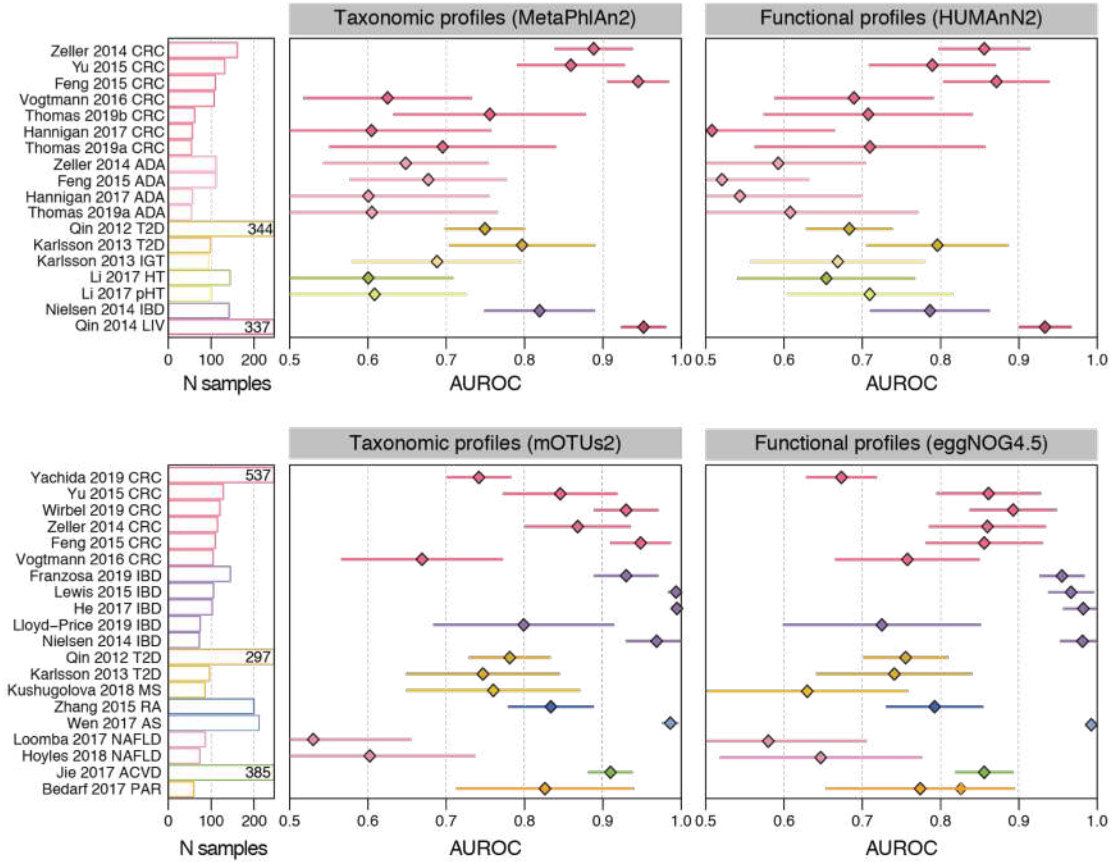
- Beware of **overfitting** – especially on small data sets, especially with complex algorithms!
  Typically N > 50, better > 100 per group is a requirement; start with simple algorithms first

- **Trade off interpretability** (white-box models) **and** maximal prediction **accuracy** wisely!

- **Models can be confounded too!**
  [see e.g. Forslund et al., *Nature* 2015 or Vujkovic-Cvijin et al., *Nature* 2020]

- Diagnostic application is relatively straightforward, but underlying **mechanisms are generally difficult to glean** from models (predictability does NOT imply causality!)

# Outlook – disease classification using SIAMCAT




**www.siamcat.embl.de**

[Wirbel et al., *bioRxiv* 2020]