Tools for comparative metagenomics

Georg Zeller
Team Leader, SCB, EMBL Heidelberg

All my course material at:
https://github.com/gezel/ebi_metagenomics_2018

de NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

EMBL

Biological diversity of microbial ecosystems




Meta HIT

HMP

An FP7 financed Project
RUMINOMICS

earth microbiome project

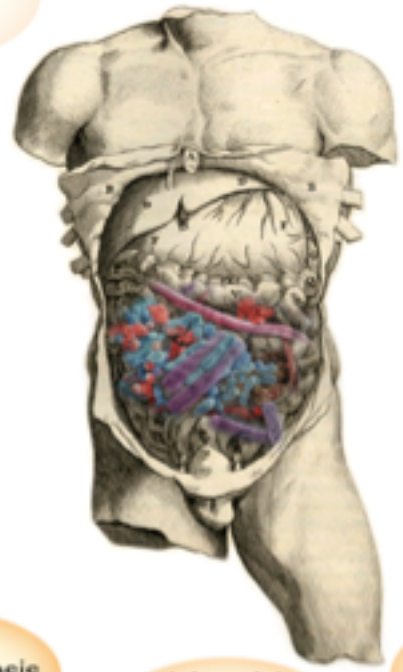
TARA OCEANS



Comparative metagenomics – Exploring biological factors affecting microbiome composition

- How do microbial communities **differ between habitats**?
  - E.g.: mouse vs. human gut, farm soil vs. grassland
- What **external factors are associated** with community composition?
  - E.g. ocean temperature & depth, soil pH, human diseases
- Is a microbial community **stable over time, how does it cope with environmental fluctuations**?
  - E.g. seasonality, host age, global warming, changing diets, co-evolution with host
- How do microbial communities **respond to perturbations**?
  - E.g. antibiotics and other drug treatments, fertilization, pollution, pathogen invasion

Comparing microbiome composition in case-control studies (disease associations)



autism spectrum disorders — Hsiao et al., Cell 2013

anxiety, schizophrenia & depression — Clarke et al., Mol. Psychiatry 2013; Yano et al., Cell 2015; Zheng et al. Mol. Psychiatry 2016

rheumatoid arthritis — Scher et al., eLife 2013; Chen et al., Genome Med. 2016

asthma & atopy — Ege et al., NEJM 2011; Fujimura et al., Nat. Med 2016

cardio-vascular disease — Wang et al., Nature 2011; Tang et al., NEJM 2013

atherosclerosis — Koren, Spor et al., PNAS 2010; Karlsson et al., Nat. Comm. 2012; Koeth et al., Nat. Med. 2013

(chronic) kidney disease — Vaziri et al., Kidney Int. 2012

inflammatory bowel disease — Gevers et al., Cell Host Microbe 2014

liver cirrhosis & cancer — Yoshimoto et al., Nature 2013; Qin et al., Nature 2014

colorectal cancer — Kostic et al., Genome Res. 2011; Castellarin al., Genome Res. 2011

type 1 diabetes — Kostic et al., Cell Host Microbe 2015

obesity — Turnbaugh et al., Nature 2006 & 2009; Le Chatelier et al., Nature 2013

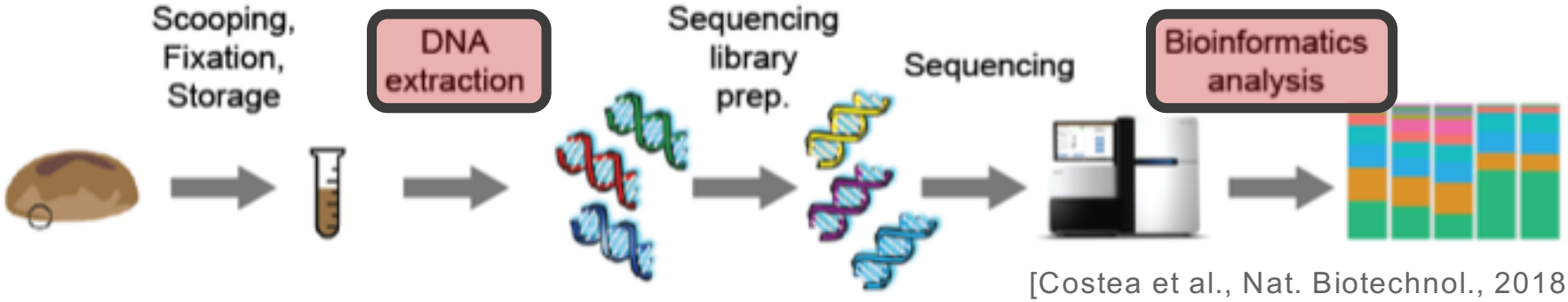
type 2 diabetes — Qin et al., Nature 2012; Karlsson et al., Nature 2013; Forslund et al., Nature 2015

Compare the microbiome of a patients (cases) to appropriate unaffected volunteers (controls) and identify changes in microbiome composition between groups

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), bioinformatics profiling
- Biological factors other than that of interest can affect profiles (**confounders**): E.g Medication, lifestyle, host demographics

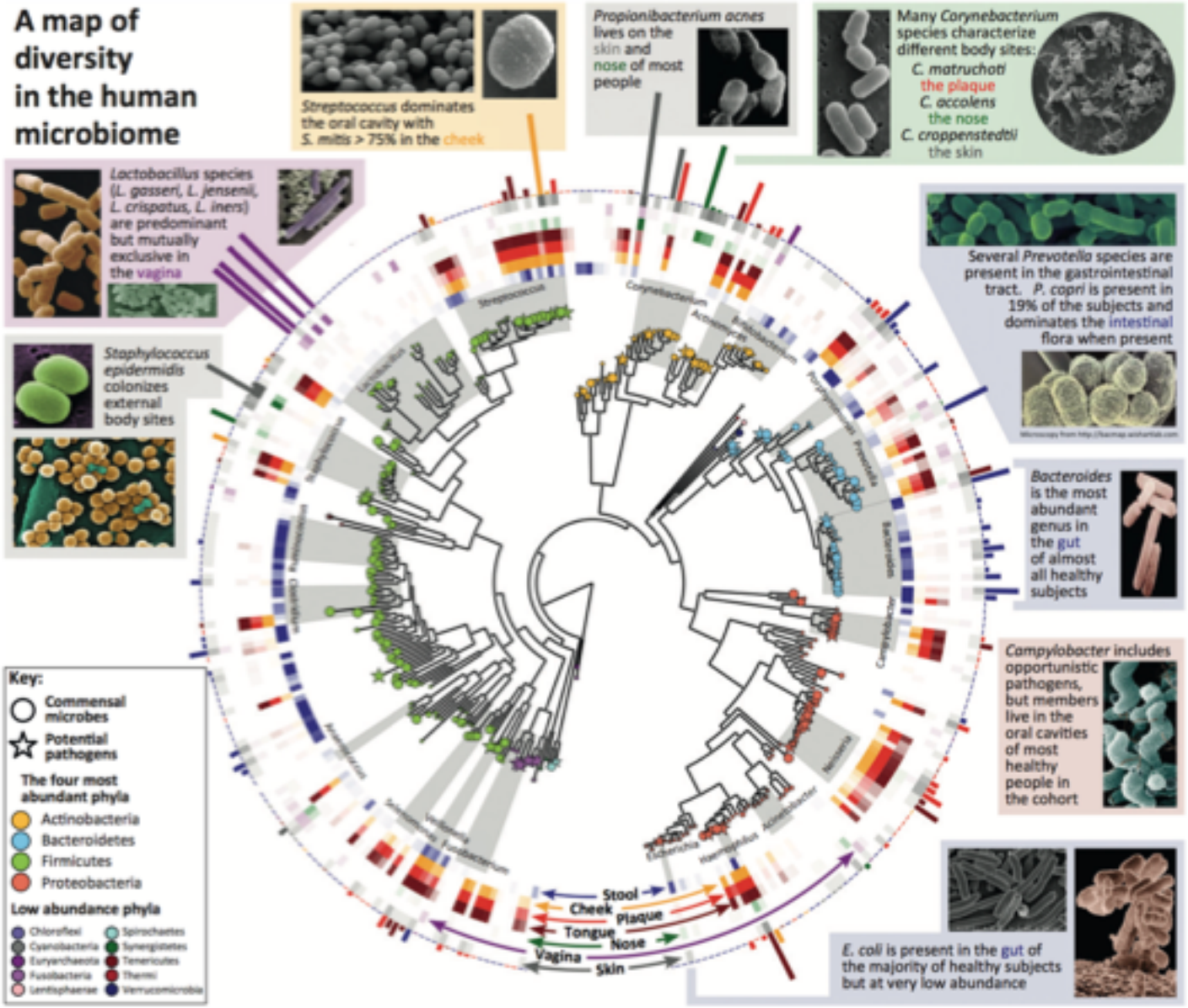
[Schmidt et al., Cell, 2018]



[Costea et al., Nat. Biotechnol., 2018]

PART I:
EXPLORATION, VISUALIZATION & TESTING

Let's start with visualization…



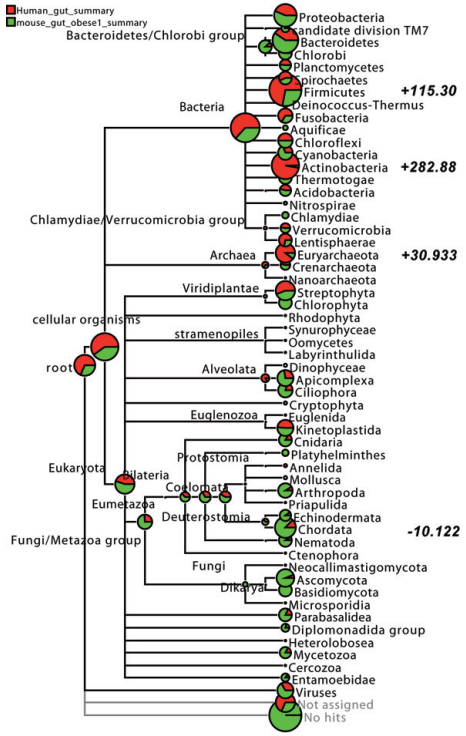
[Fig. from Morgan et al., Trends Genet. 2013]

Visual community comparison

Relative abundance of taxa as

- Heatmap (taxa vs environment/conditions)
- Bar or pie charts
- Alongside a phylogenetic or taxonomic tree

➢ Important limitation:
variance between samples of the
same group is concealed by “averaging”

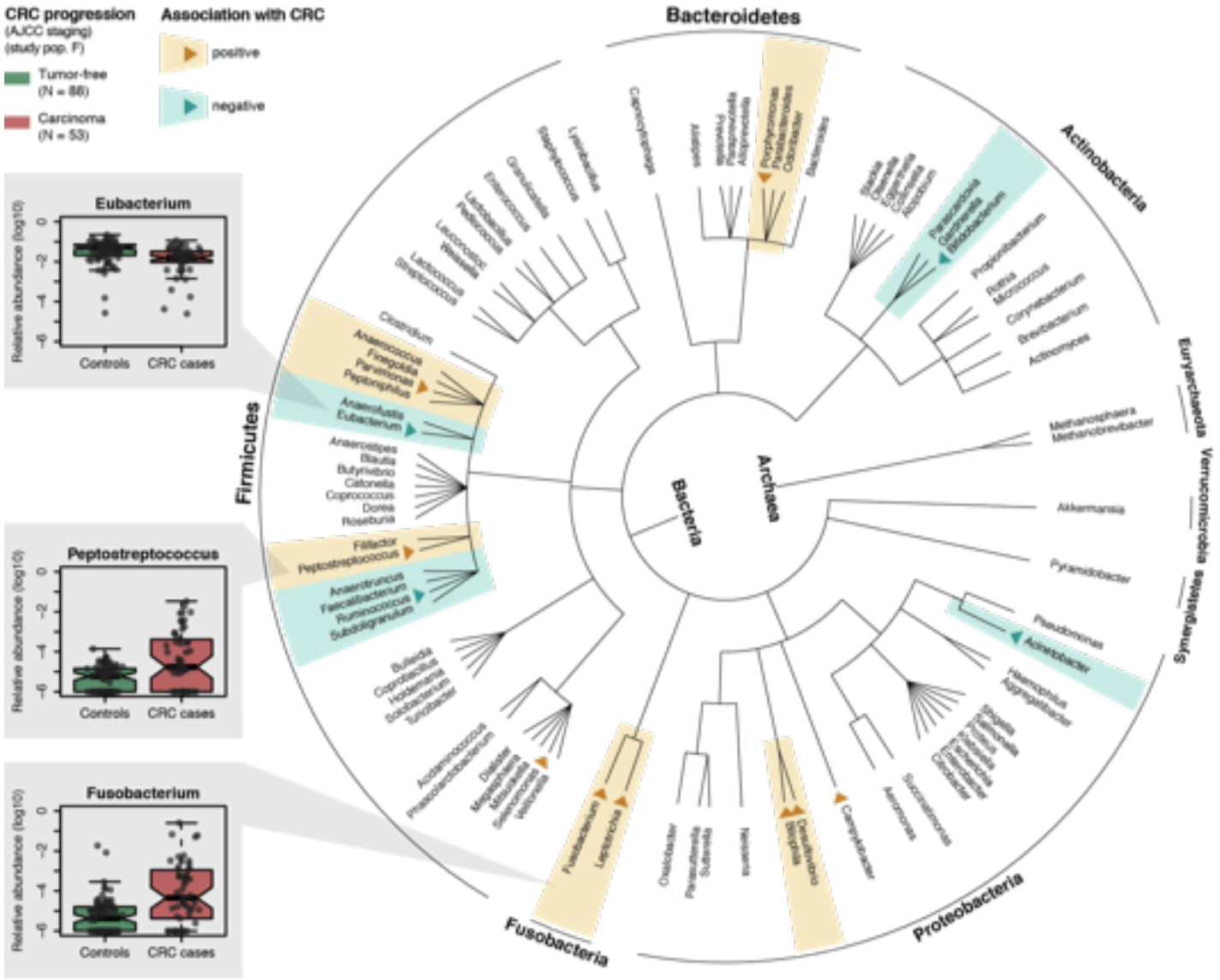


[Fig. from Huson et al. BMC Bioinform 2009; MEGAN]

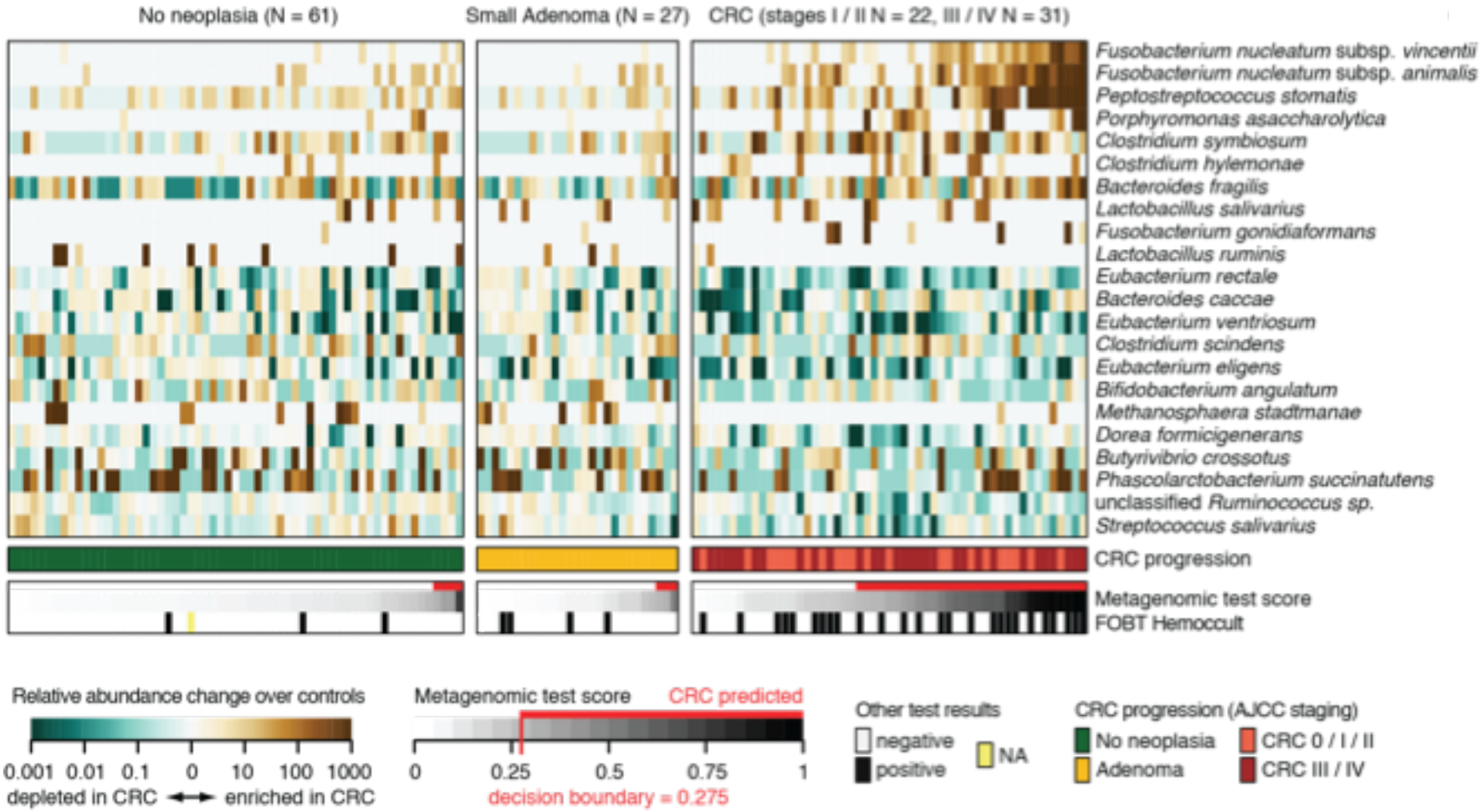
How to visualize sample-to-sample variance?

- Boxplots (visual account of variance within groups)
- Heatmaps (taxa vs samples, grouped by condition)
- Ordination (embedding high-dimensional space in 2 or 3D)

Example I: Colorectal-cancer associated gut microbiome alterations visualized on taxonomic tree



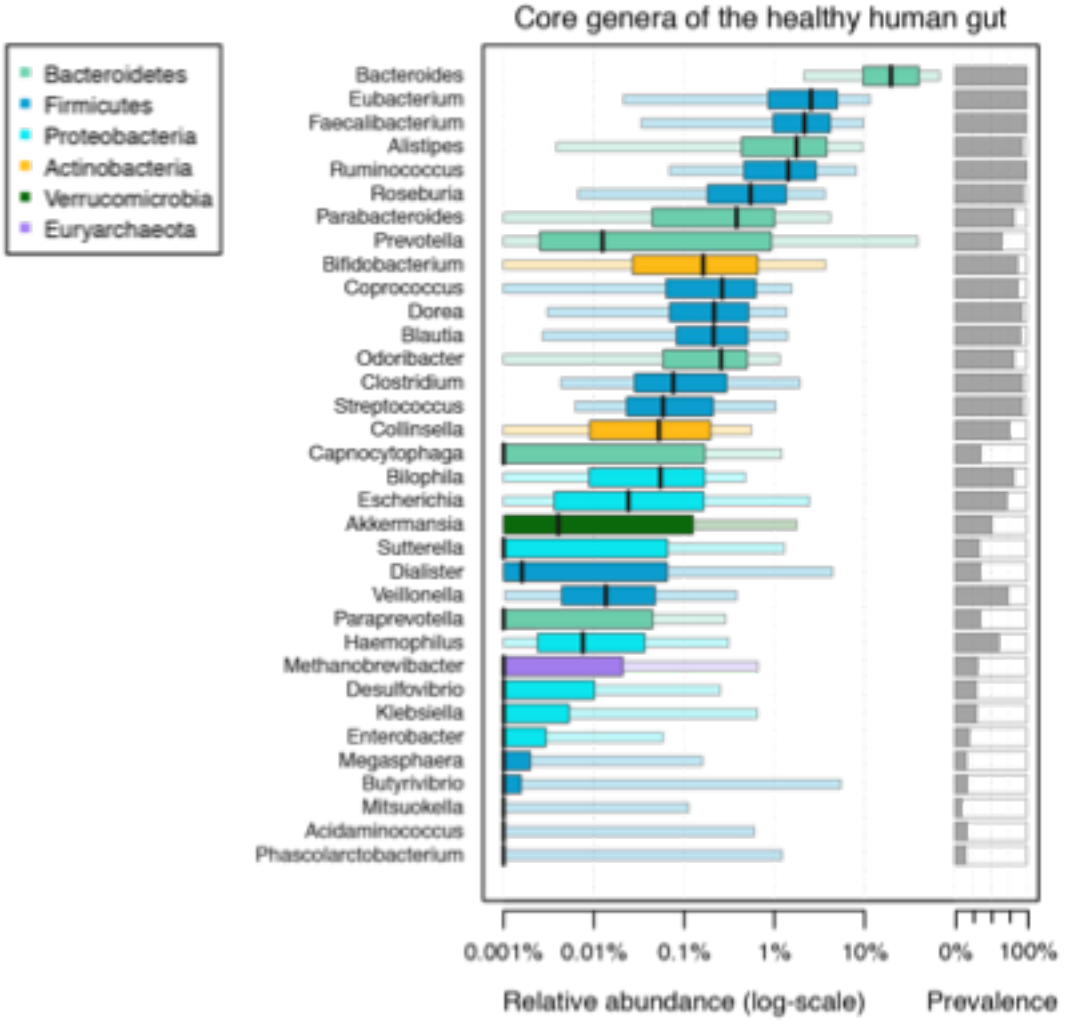
Example II: Heatmap of gut microbiome changes in colorectal cancer



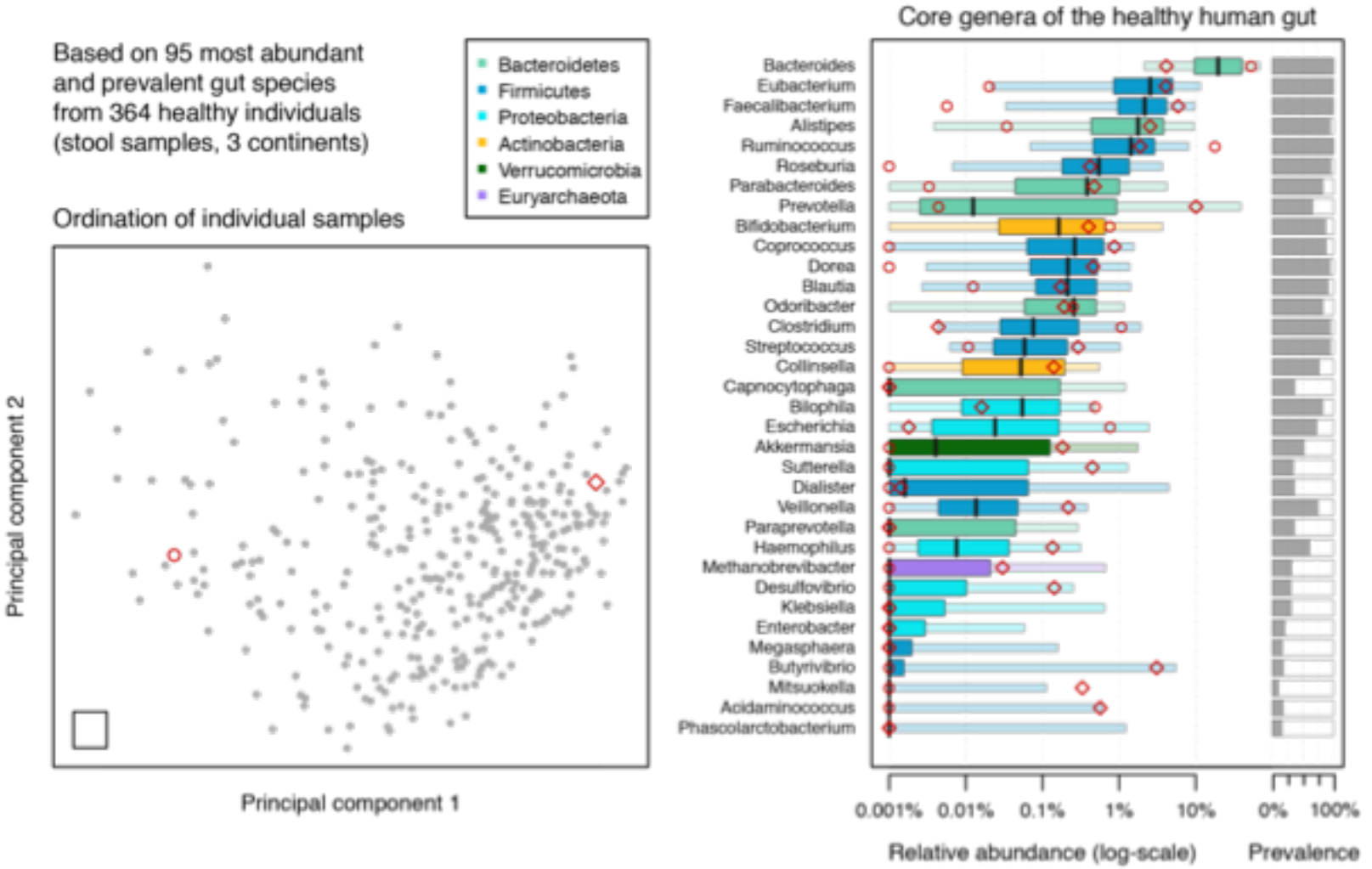
[Zeller*, Tap*, Voigt* et al., Mol. Syst. Biol. 2014; SIAMCAT]

Example III: Individuality of the gut microbiome

Based on 95 most abundant and prevalent gut species from 364 healthy individuals (stool samples, 3 continents)



Example III: Individuality of the gut microbiome



Extreme within-group variance is a key characteristic of many microbial communities

Tools for microbial community comparison

Assessing differences in overall community structure

- **Community (dis-)similarity / measures:**
  Summarize dissimilarity between two communities as a single number
  - Clustering
  - Comparing dissimilarity within and between groups

- **Ordination:**
  Reduce high-dimensional data to 2 to 3 dimensions while preserving dissimilarity relation between samples as well as possible
  - Use with custom dissimilarity
  - Visualization by projection
  - Analyze major sources of variance

Testing for changes in individual taxa

- Apply a **statistical test** to each taxon to ask whether its abundance significantly differs between groups

- Robust, **nonparametric tests** (Wilcoxon or Kruskal Wallis tests) are recommended

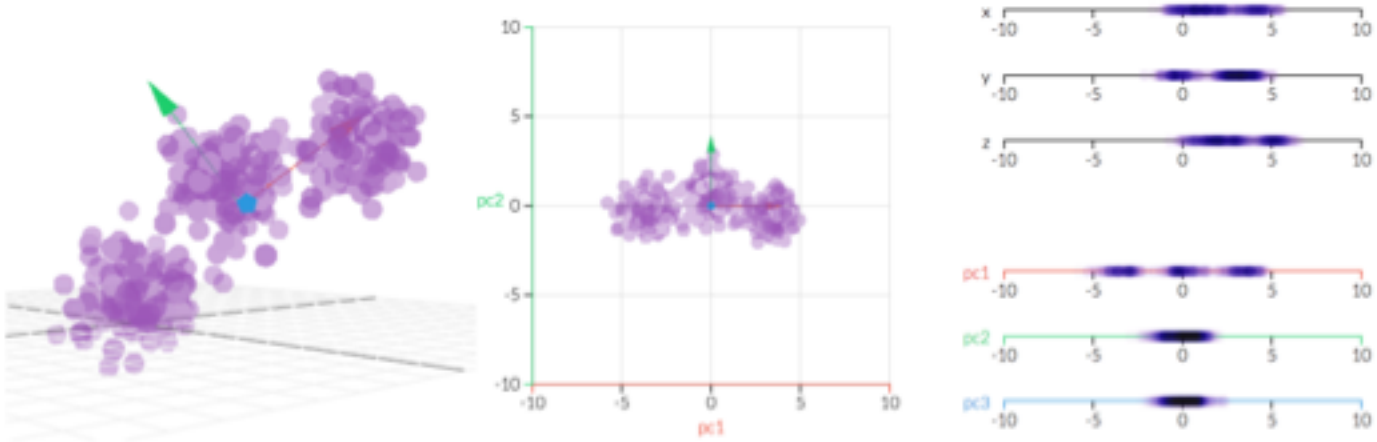
- **Correct for multiple testing** to e.g. control the false discovery rate

- Individual taxa affected by treatment, useful biomarkers, ecological indicator species, …
- Often more sensitive and interpretable than ordination!

Dissimilarity measures for microbial community comparison

Euclidean	$d_{kl} = \sqrt{\sum_{i=1}^{n}(p_{ik} - p_{il})^2}$	p_{ij} is the (relative) abundance of taxon i in sample j
Manhattan	$d_{kl} = \sum_{i=1}^{n} \lvert p_{ik} - p_{il} \rvert$	n equals the number of taxa
Gower	$d_{kl} = \frac{1}{n} \sum_{i=1}^{n} \frac{\lvert p_{ik} - p_{il} \rvert}{max(p_i) - min(p_i)}$	
Canberra	$d_{kl} = \frac{1}{\hat{n}} \sum_{i=1}^{n} \frac{\lvert p_{ik} - p_{il} \rvert}{p_{ik} + p_{il}}$	$\hat{n}$ equals the number of taxa with nonzero abundance in at least one sample
Bray-Curtis	$d_{kl} = \frac{\sum_{i=1}^{n} \lvert p_{ik} - p_{il} \rvert}{\sum_{i=1}^{n} (p_{ik} + p_{il})}$	
Jaccard	$d_{kl} = \frac{b_{kl}}{(1 + b_{kl})}$	b_{kl} is the Bray-Curtis dissimilarity between sample k and sample l.
unweighted Unifrac	$\frac{\sum_{i=1}^{n} b_i \lvert q_{ik} - q_{il} \rvert}{\sum_{i=1}^{n} b_i \max(q_{ik}, q_{il})}$	q_{ik} is 1 if taxon i is present in sample k and 0 otherwise.
weighted Unifrac	$\sum_{i=1}^{n} b_i \lvert \frac{p_{ik}}{\sum_{i=1}^{n} p_{ik}} - \frac{p_{il}}{\sum_{i=1}^{n} p_{il}} \rvert$	b_i refers to branch i. $\sum_{i=1}^{n} p_{ik}$ is equal to the sum of abundances in sample k.

Ordination: principal component analysis and generalizations

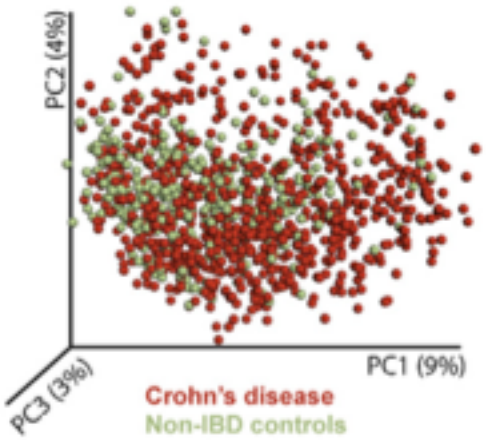


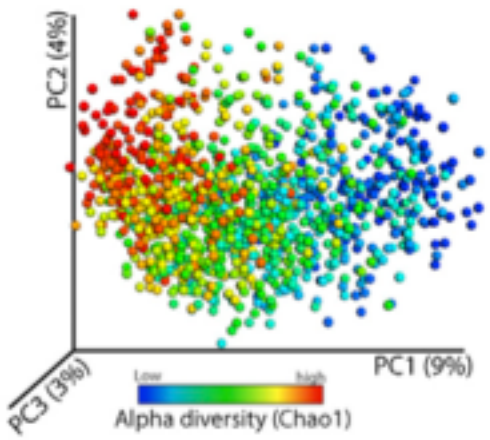
[Fig source: http://setosa.io/ev/principal-component-analysis/]

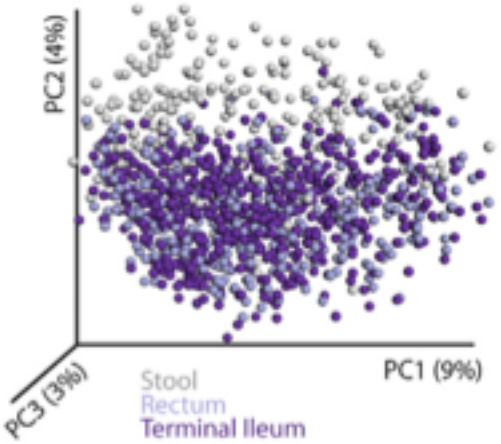
- PCA: transformation into an orthogonal basis such that first principal component is aligned with most variance and so on. Project to 2D (or 3D) for visual exploration
  - ➢ Implicitly works with Euclidean distance

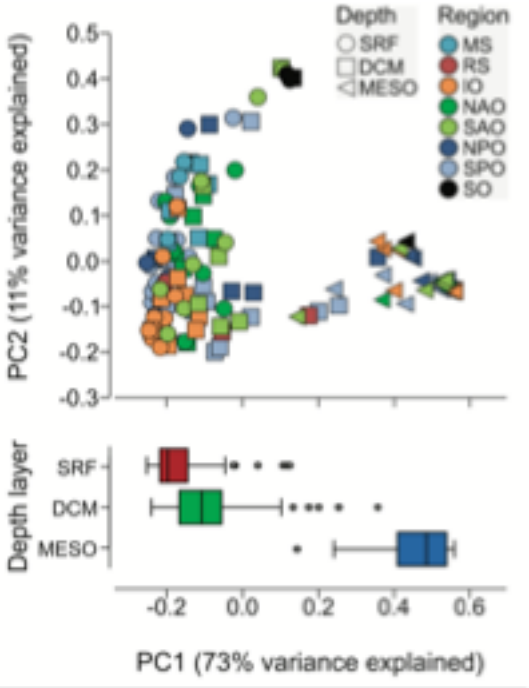
- PCoA: Generalization that finds a low-dimensional embedding that preserves a pre-specified distance matrix as well as possible.
  - ➢ Can be used with any user-supplied dissimilarity matrix

Ordination as a tool to reveal ecological differences between microbial community samples









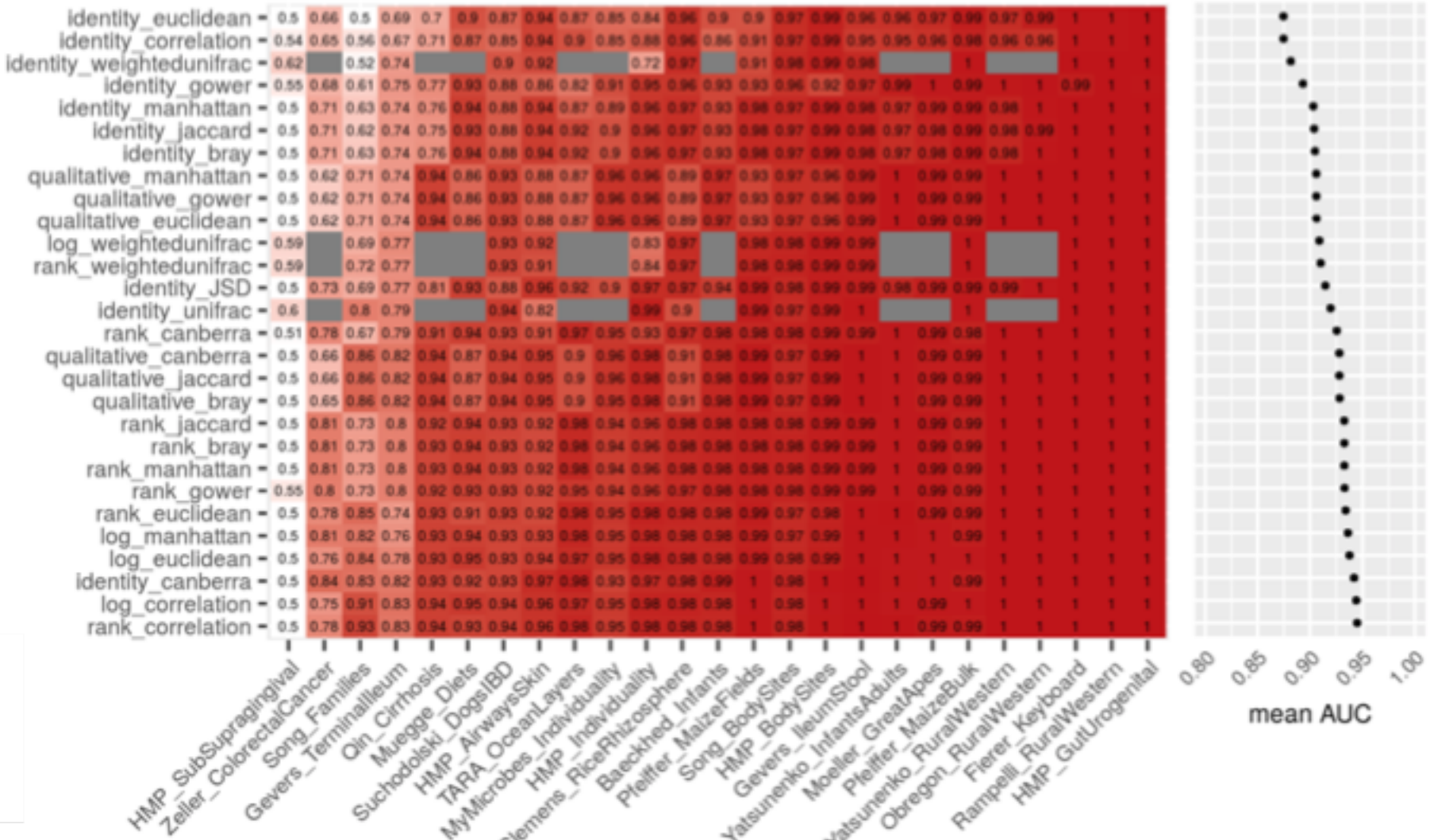
Top: large cohort of inflammatory bowel disease patients and controls shows alpha diversity gradient and clustering by sampling location

[Gevers et al., Cell Host Microbe 2014]

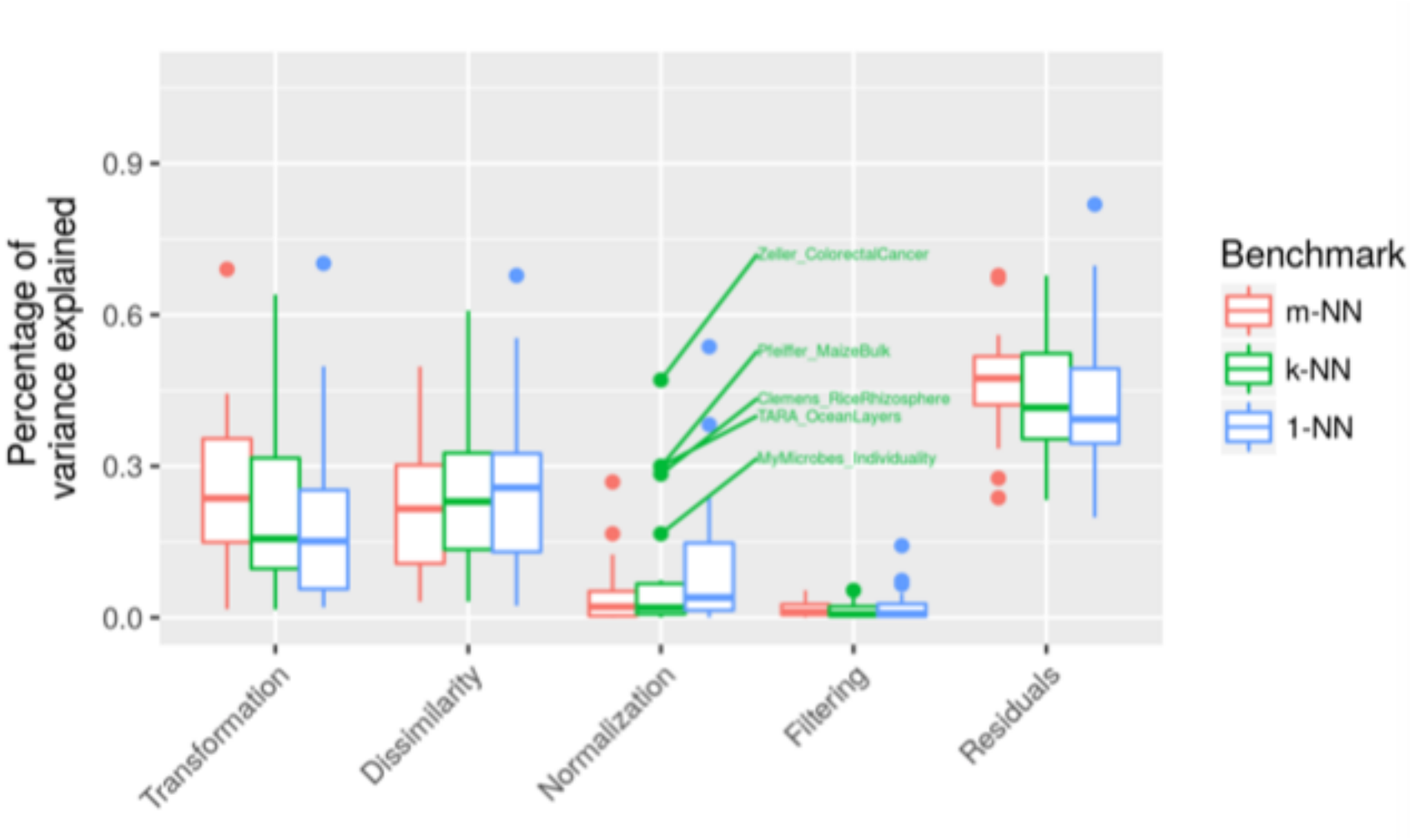
Left: TARA Oceans data reveals clustering by sampling depth

[Sunagawa et al., Science 2015]

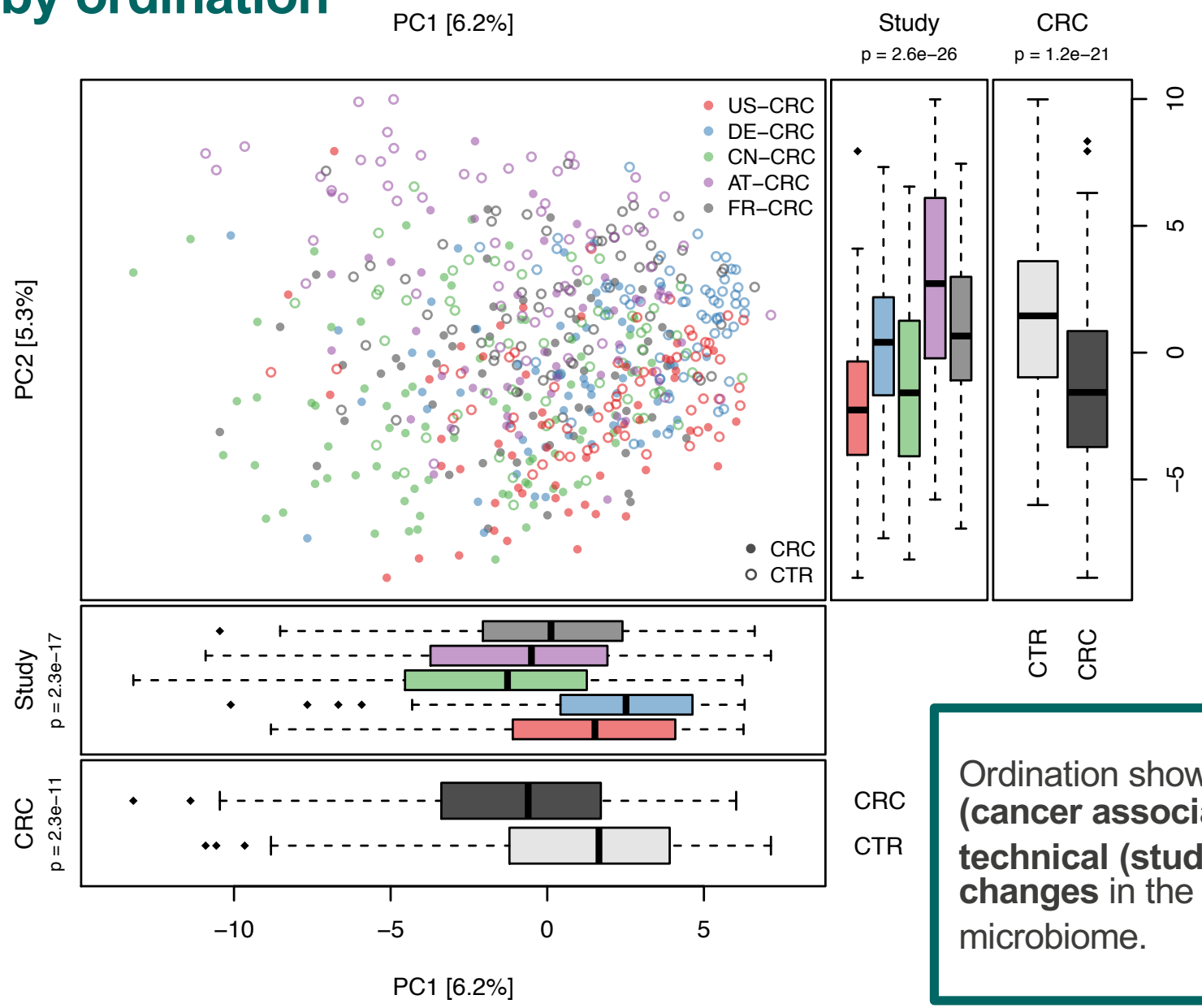
Comparison of dissimilarity measures



Factors influencing dissimilarity analysis

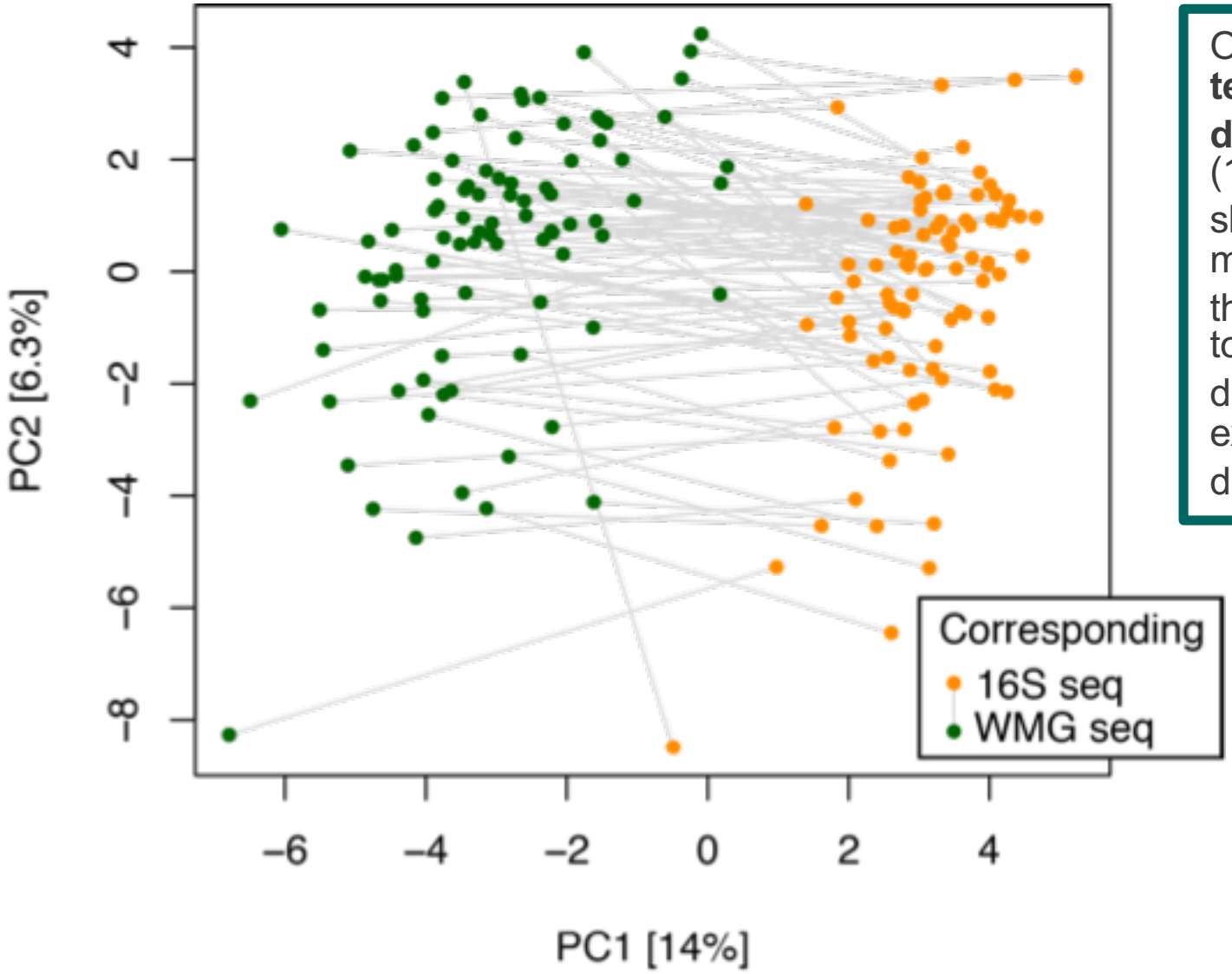


Joint visualization of multiple influences by ordination



Ordination shows **biological (cancer associated) and technical (study associated) changes** in the gut microbiome.

Example of technical differences due to sequencing approach



Ordination shows **technical differences** (16S versus shotgun metagenomics of the same samples) to be more dramatic than any expected biological difference.

Tools for microbial community comparison

Assessing differences in overall community structure

- **Community (dis-)similarity / measures:**
  Summarize dissimilarity between two communities as a single number
  - ➢ Clustering
  - ➢ Comparing dissimilarity within and between groups

- **Ordination:**
  Reduce high-dimensional data to 2 to 3 dimensions while preserving dissimilarity relation between samples as well as possible
  - ➢ Use with custom dissimilarity
  - ➢ Visualization by projection
  - ➢ Analyze major sources of variance

Testing for changes in individual taxa

- Apply a **statistical test** to each taxon to ask whether its abundance significantly differs between groups

- Robust, **nonparametric tests** (Wilcoxon or Kruskal Wallis tests) are recommended

- **Correct for multiple testing** to e.g. control the false discovery rate

- ➢ Individual taxa affected by treatment, useful biomarkers, ecological indicator species, …
- ➢ Often more sensitive and interpretable than ordination!

Summary of caveats

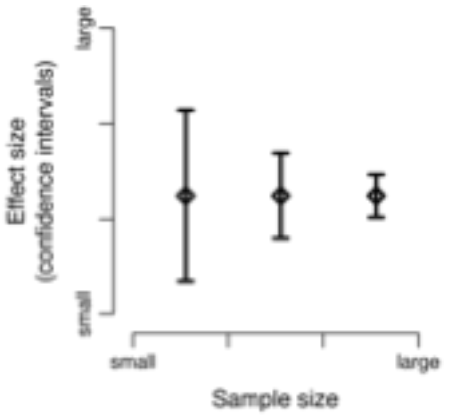
- Beware of technicalities such as **inappropriate data preprocessing**, **strange dissimilarity measures** and **projection** artefacts!
  [see e.g. Morton et al. mSystems 2017]

- **Disentangling biological and technical effects** can be difficult!
  [e.g. geography and DNA extraction protocol effects, as these are rarely uncorrelated]

- **Confounding** with biological effects from uncontrolled sources is difficult to detect and correct for!
  [see e.g. Forslund et al. Nature, 2015, Qin et al. Nature 2012, Karlsson et al. Nature 2013]

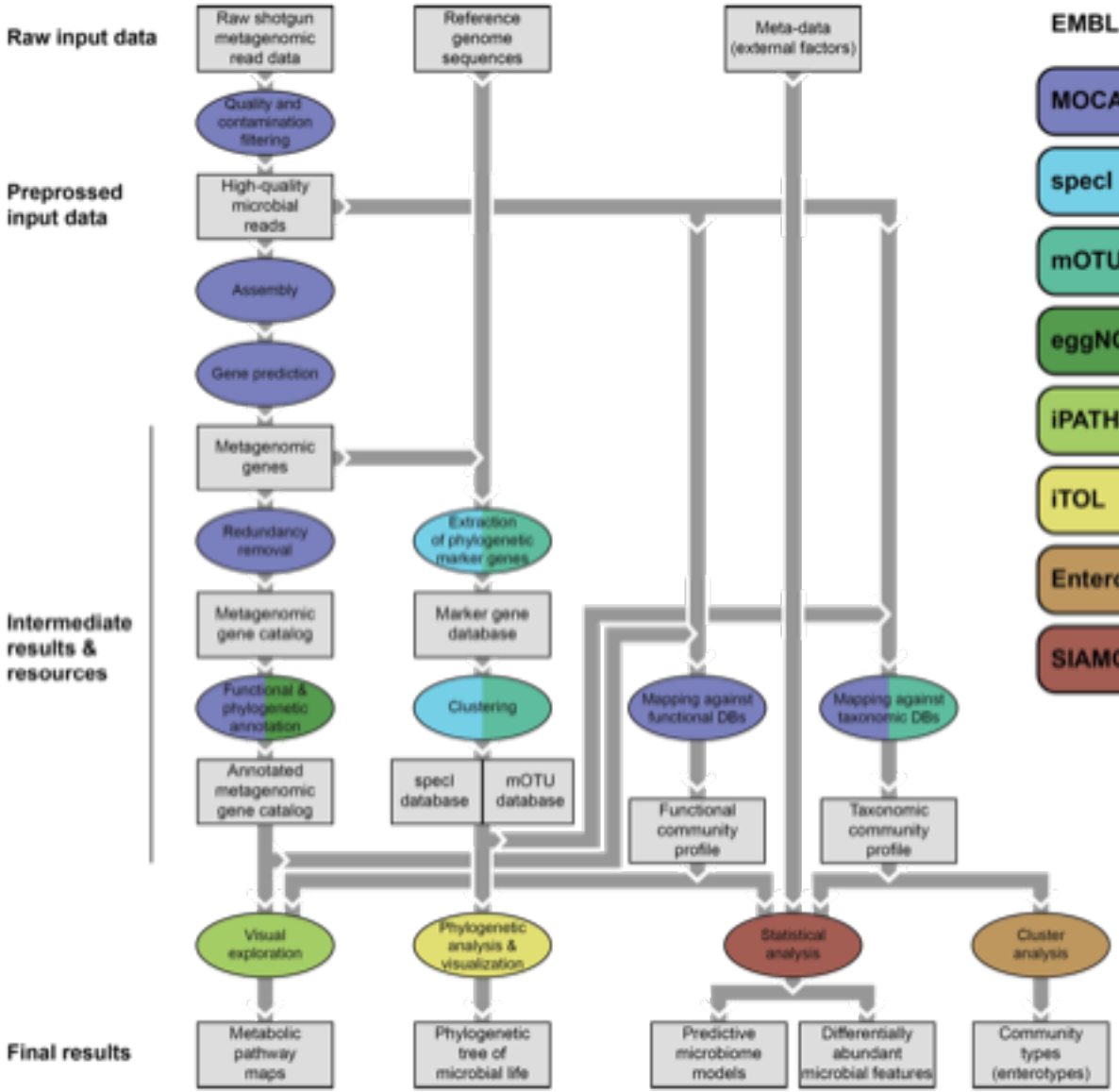
- Correlation does not imply **causation**!
  [see e.g. http://www.tylervigen.com/spurious-correlations]

- Statistical significance should not be confused with **effect size**, nor with **biological significance**!



EMBL microbiome analysis toolkit

www.microbiome-tools.embl.de



EMBL Microbiome Tools

- **MOCAT** — Metagenomics profiling pipeline
- **specI** — Prokaryotic species identification
- **mOTUs** — Metagenomic species profiling
- **eggNOG** — Resource for orthologous genes
- **iPATH** — Cellular pathway mapping tool
- **iTOL** — Interactive tree visualization
- **Enterotyping** — Gut microbiome community typing
- **SIAMCAT** — Statistical analysis of microbiome data

de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

https://github.com/gezel/ebi_metagenomics_2018

PART II:
STATISTICAL MODELLING

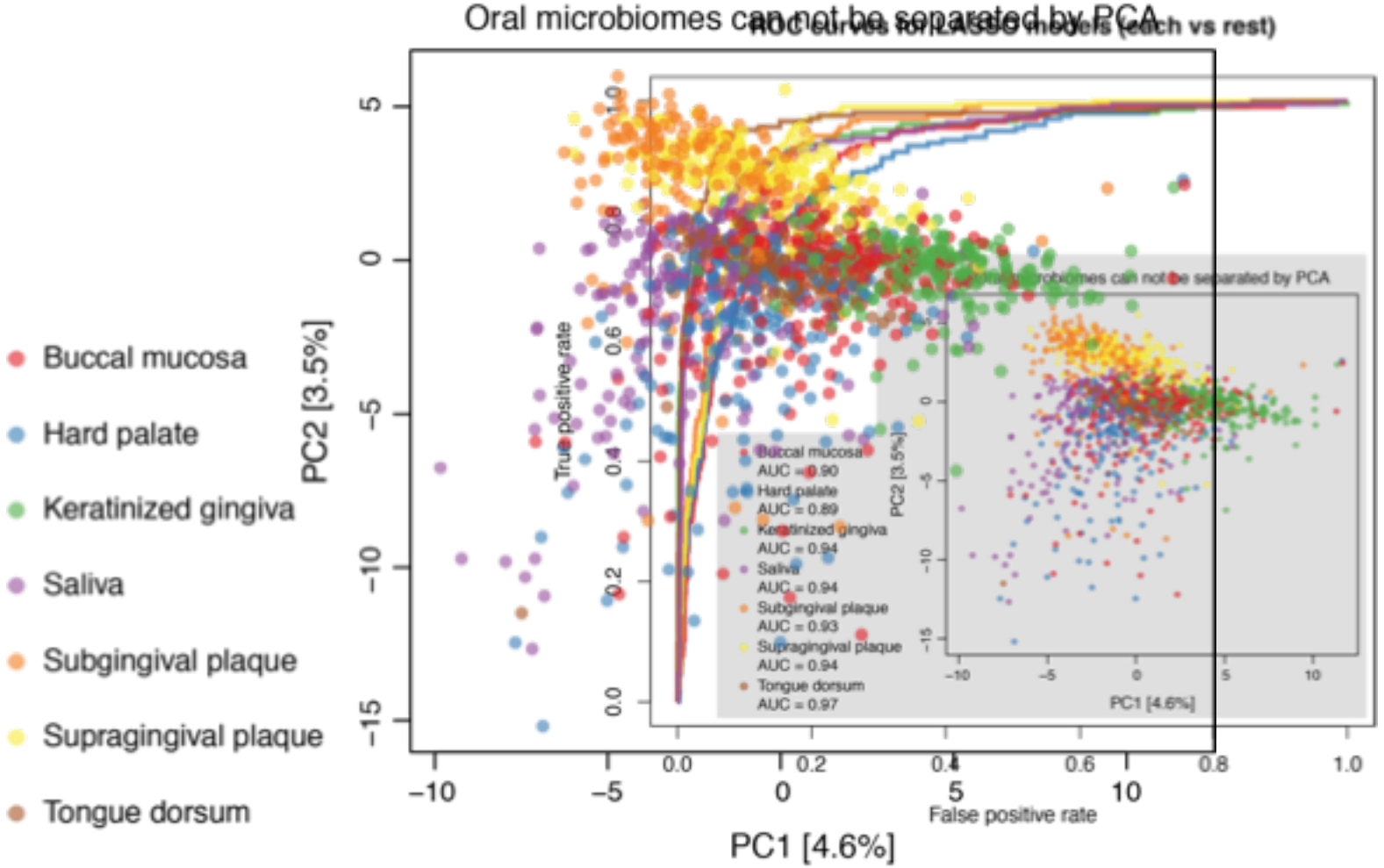
Why statistical modeling / machine learning?

- Modeling ideally **extracts the essence** of a biological phenomenon
- Good models make **accurate 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
- **Sparse statistical models** are „in between“ univariate and multivariate testing, e.g. based on „a few “ taxa (**biomarkers**)
- Suitable methods can **select predictive taxa** (and ignore others)

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

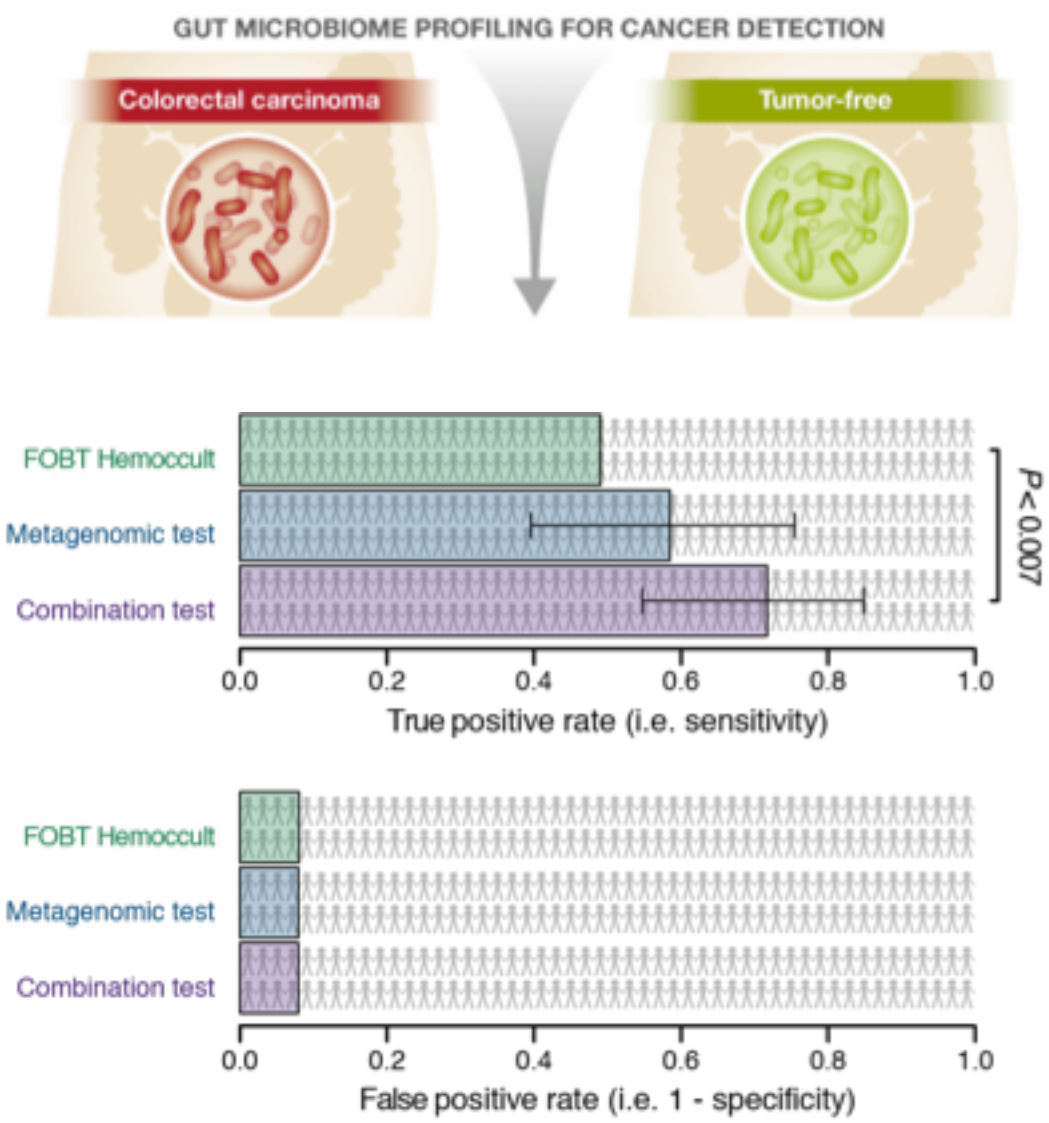
For i samples / patients
y_i – label (e.g. disease or control)
x_i – features (e.g. species abundance profile, a vector)
f – our model
ϵ – modeling error

PCoA vs modeling



While it can be difficult to resolve for each oral microbiome sample the precise sampling site using PCoA (with various dissimilarity measures), one can train statistical models that can very accurately recognize sample origin.

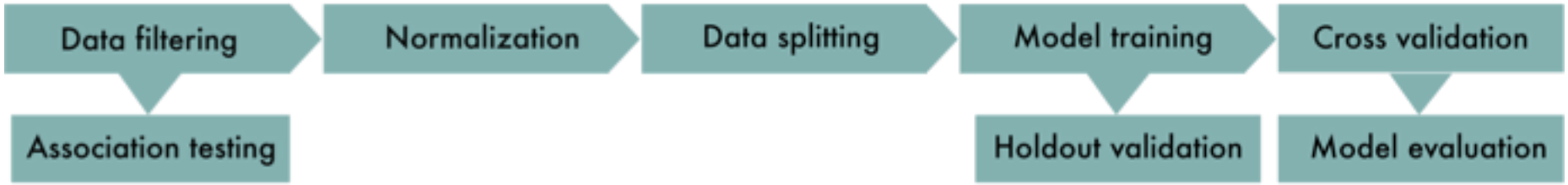
Detecting colorectal cancer (CRC) from fecal microbiome profiles



- 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 clinical tests (FOBT)?

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

A statistical modeling workflow



SIAMCAT Bioconductor 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

The above workflow is specifically implemented in the SIAMCAT Bioconductor package, but still fairly general for any statistical modeling task.

Feature preparation

- Use your **domain expertise** to engineer features that are likely predictive of the phenomenon of interest – microbiome examples:
  - Species abundances (or higher/lower resolution 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
- Importantly, do **NOT use the label** information for feature filtering / preprocessing!

Model evaluation I – classification errors

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 TP	False positives FP (Type I errors)
	negative ("predicted not to have cancer")	False negatives FN (Type II errors)	True negatives TN

True positive rate (TPR, **sensitivity**, **recall**) = TP / (TP + FN)
True negative rate (TNR, **specificity**) = TN / (TN + FP)
False positive rate (FPR, 1 – specificity) = FP / (TN + FP)

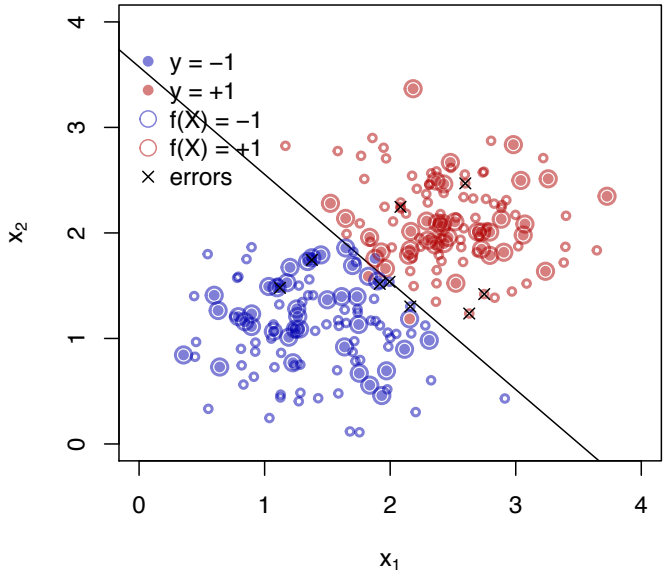
- are all **independent on prevalence** (of positives in the total population)

Precision (positive pred. value, PPV) = TP / (TP + FP)
False discovery rate (FDR, 1 – precision) = FP / (FP + TP)

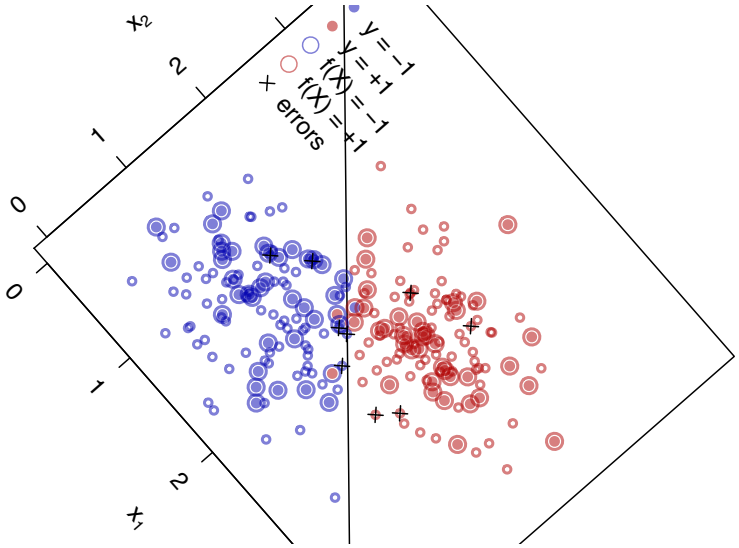
- are both **dependent on prevalence** (of positives in the total population)

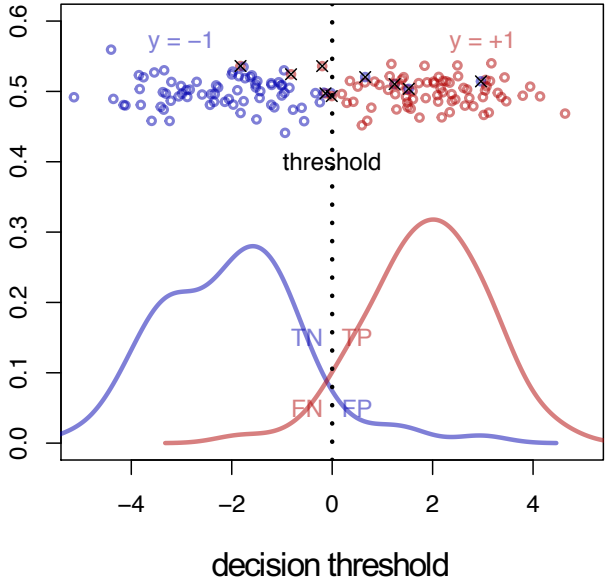
[these and more measures on en.wikipedia.org/wiki/Evaluation_of_binary_classifiers]

Model evaluation II – ROC curves

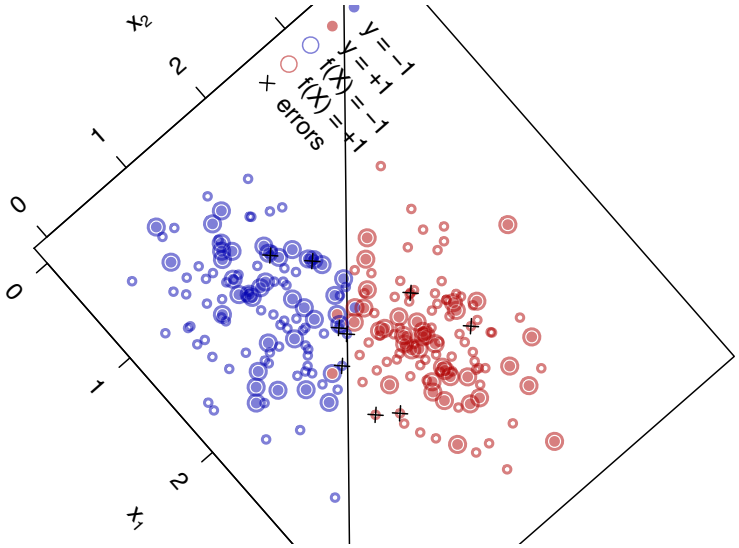


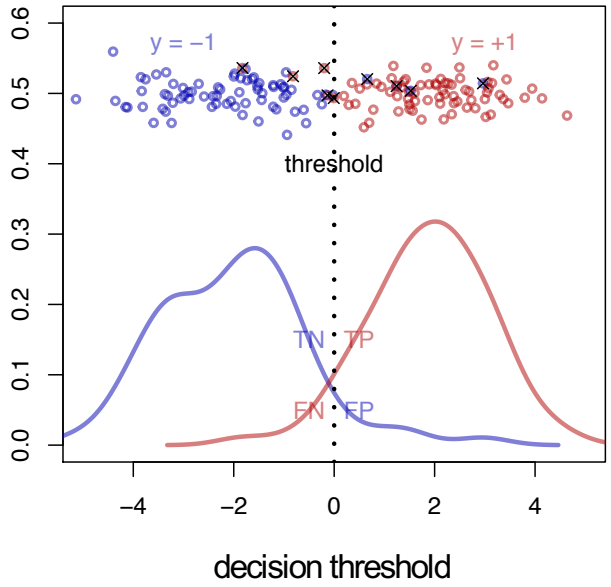
Model evaluation II – ROC curves

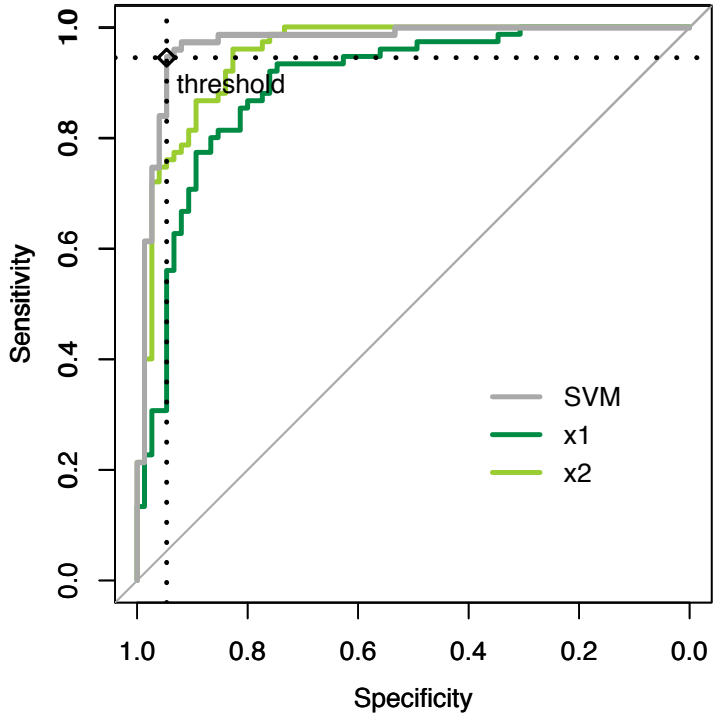




Model evaluation II – ROC curves



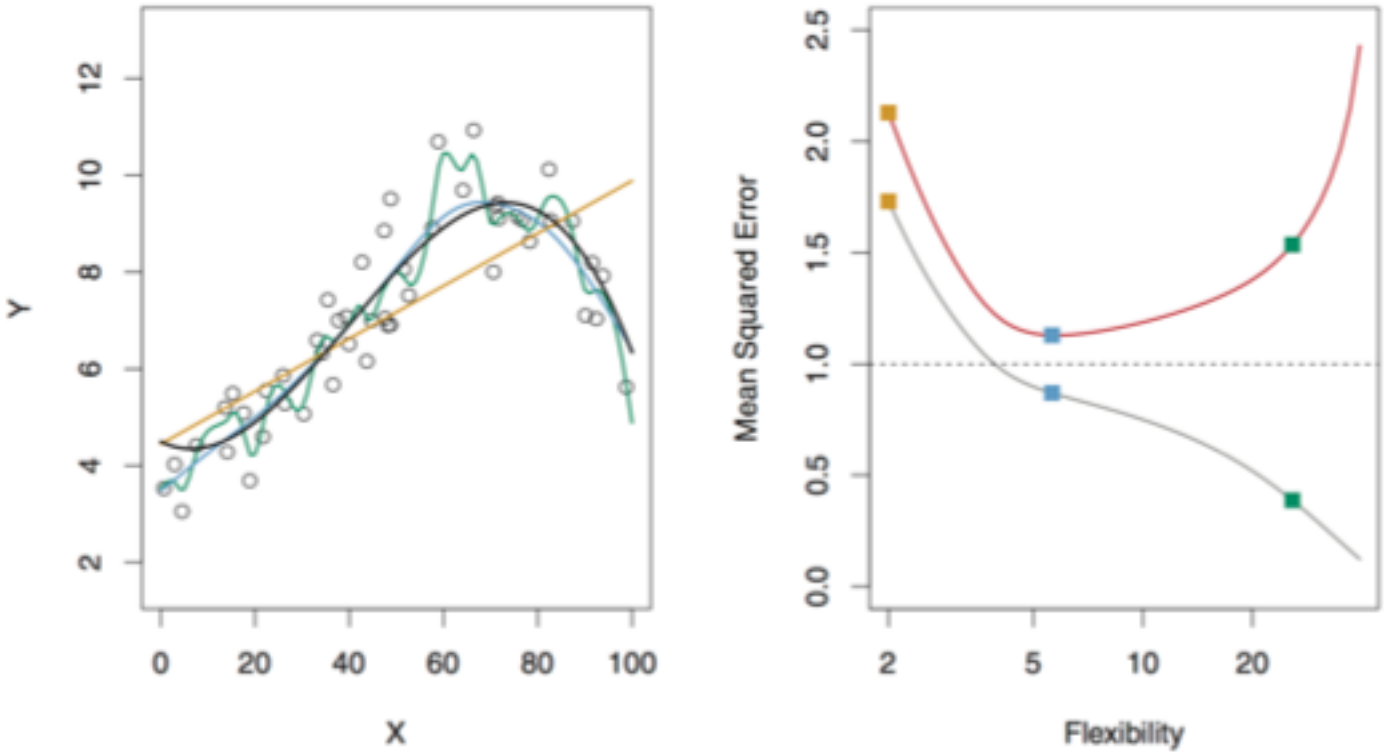




- Change decision threshold to obtain other trade-offs between sensitivity and specificity
- Receiver operating characteristic (ROC) curve plots all of them
- Area under the ROC curve as a summary

Model evaluation III – assessing generalization

- Minimizing the training error
  With increasing flexibility, models will fit the training data better and better
- Generalization to new data sets
  Overfitting the training data will result in poor generalization



Here smoothing splines are used; model flexibility / complexity increases with the degree of the polynomials.

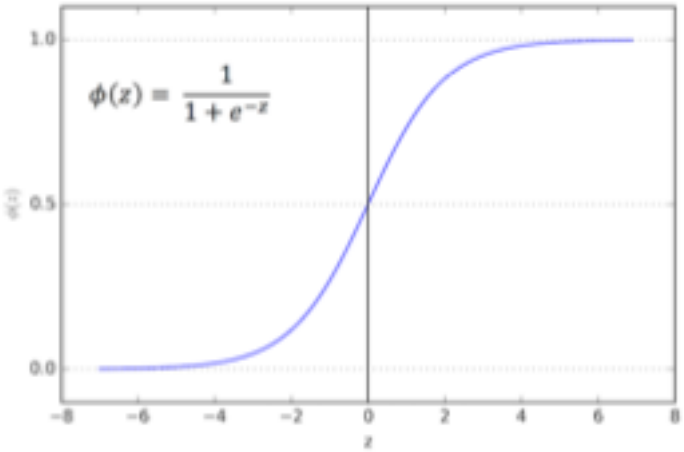
[James, Witten, Hastie & Tibshirani, Springer 2013]

Training and prediction (with the LASSO)

The (LASSO logistic regression) **model**:

$$f(\mathbf{x_i}) = \phi(\mathbf{w}^T\mathbf{x_i}) = \phi(\sum_{j=1}^{p} w_j x_{ji})$$

The logit function ϕ:



Fitting the model (for the simple LASSO without logit):

$$\min_{\mathbf{w}} \sum_{i}^{n} (y_i - \mathbf{w}^T\mathbf{x_i})^2 + \lambda \sum_{j}^{p} |w_j|$$

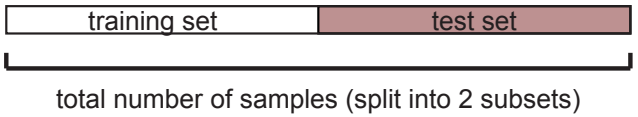
Applying the trained model to new data:

$$\hat{y}_i = f(\hat{\mathbf{x}}_\mathbf{i}) = \phi(\mathbf{w}^T\hat{\mathbf{x}}_\mathbf{i})$$

Splitting 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

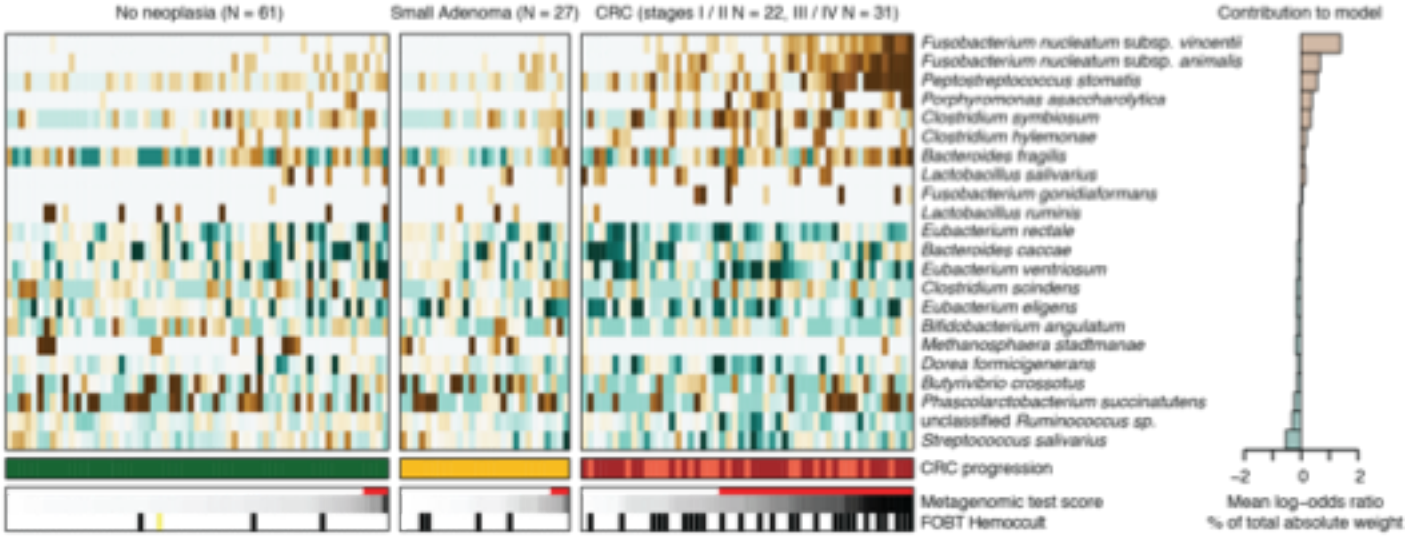
Cross-validation pitfalls

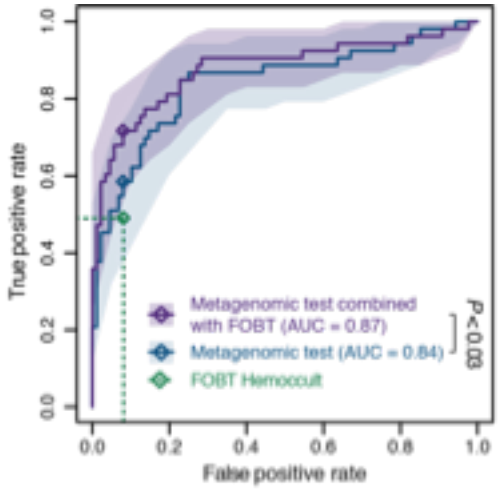
- **Cross validation works under the i.i.d. assumption** (observations have the same probability distribution and are mutually independent)
  - E.g. a series of (fair or unfair) coin flips is i.i.d. as the next flip doesn't depend on the previous ones.

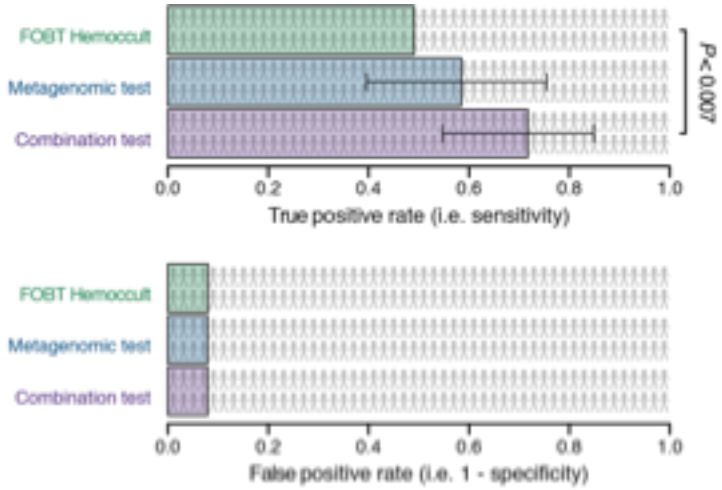
- However, biological samples are often **not independent**:
  - Multiple time-point measurements from the same subject or related subjects
  - Spatial structure / dependencies between measurements

- 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

Model evaluation & interpretation







Fecal metagenomic profiling as a novel method for non-invasive CRC detection.

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

Summary of caveats

- **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 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, Qin et al. Nature 2012, Karlsson et al. Nature 2013]

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

https://github.com/gezel/ebi_metagenomics_2018