Swiss Institute of Bioinformatics

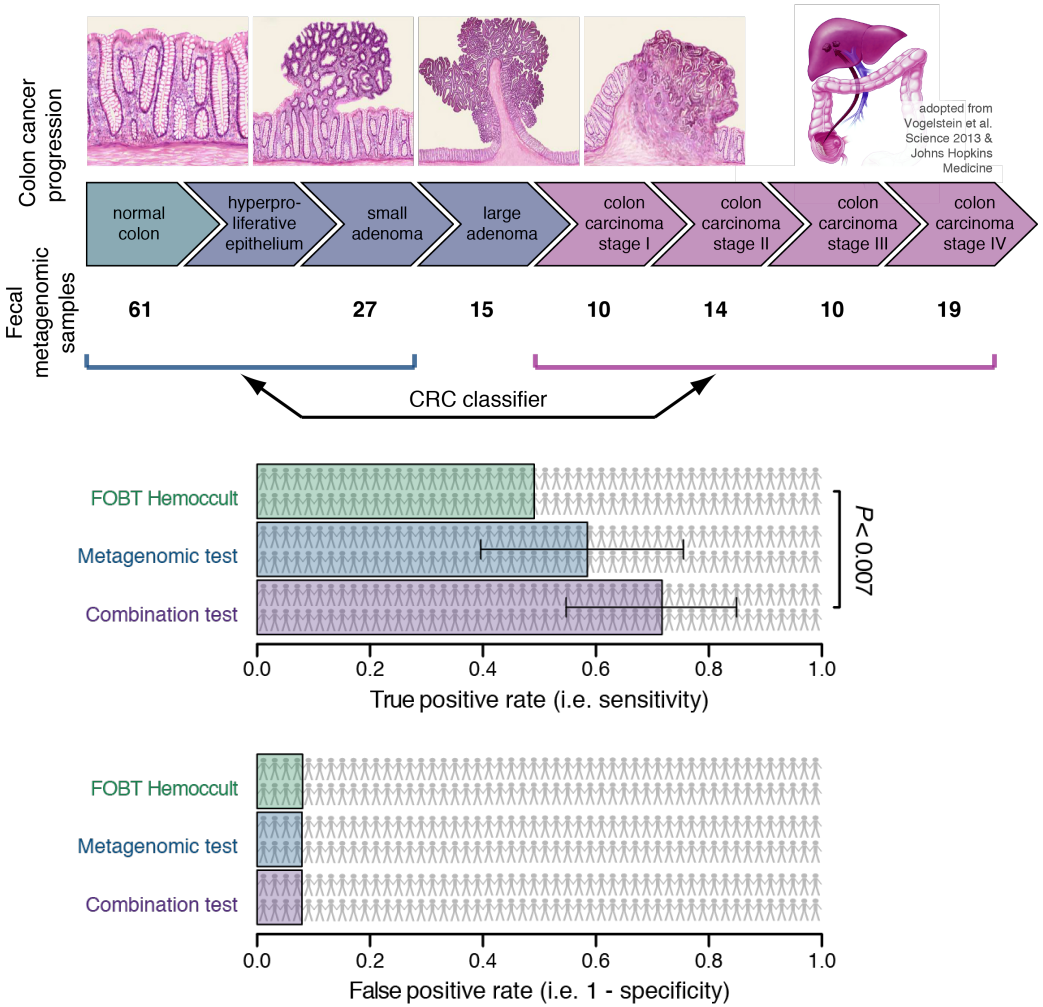
Machine learning / statistical modelling of metagenomic data

Alessio Milanese, Lukas Malfertheiner

Project 3

Spring School Bioinformatics and computational approaches in Microbiology

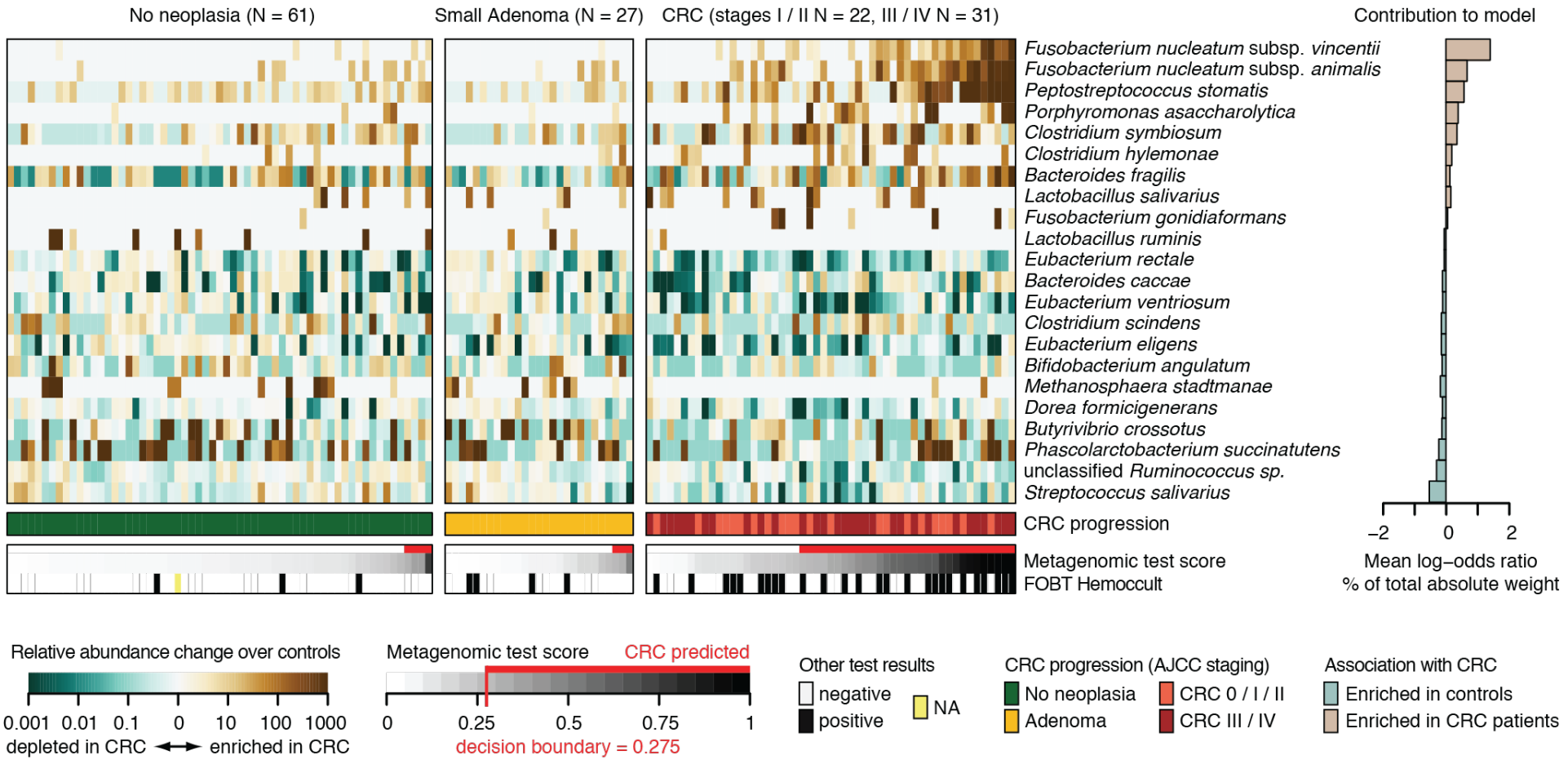
Colorectal cancer example (continued)



- 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?
- How does metagenomic detection compare to standard noninvasive diagnostic test (FOBT)?

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

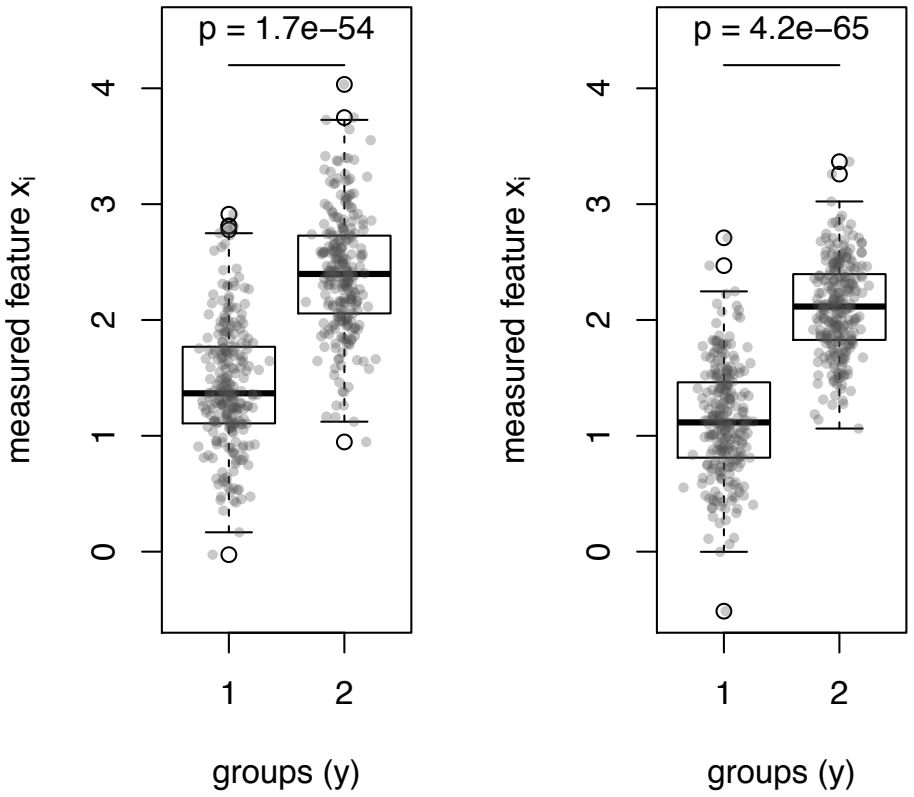
A microbiome “signature” of colorecatal cancer



[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
ϵ – 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)

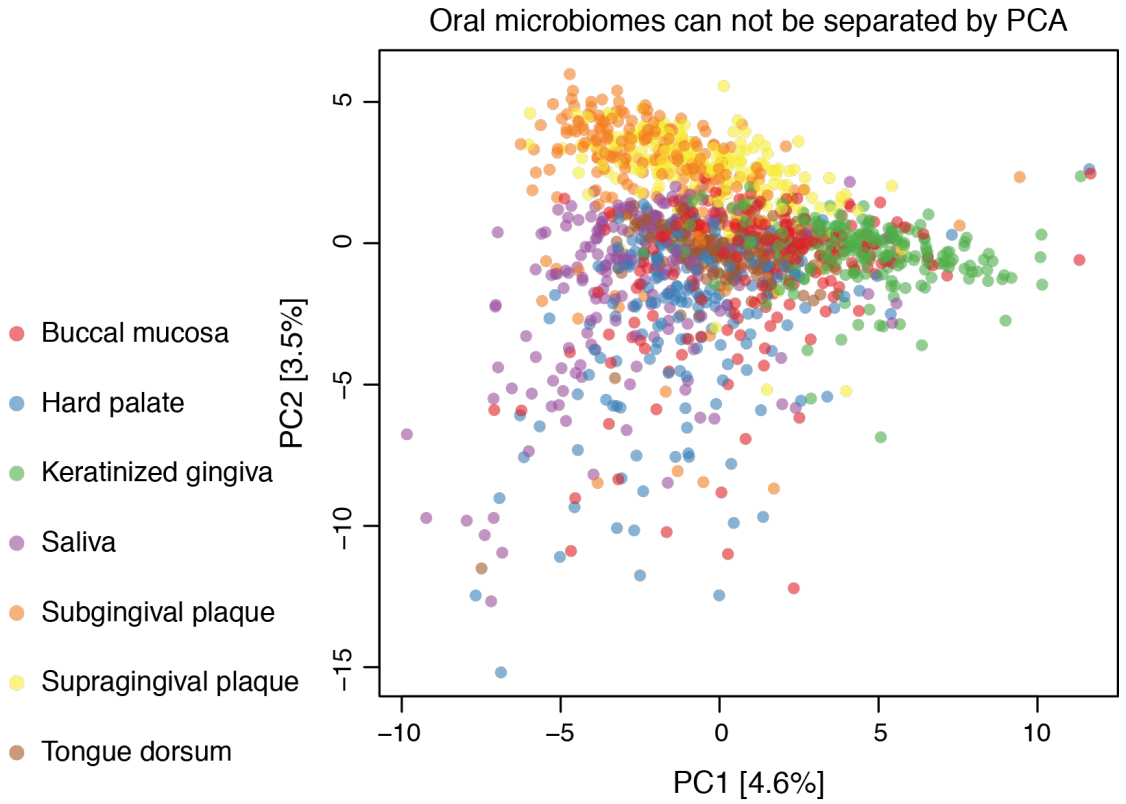
  observations based on which we wish to make predictions
  $\mathbf{x}_i$ denotes the feature vector, i.e. abundance profile, for the i-th sample

- **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.

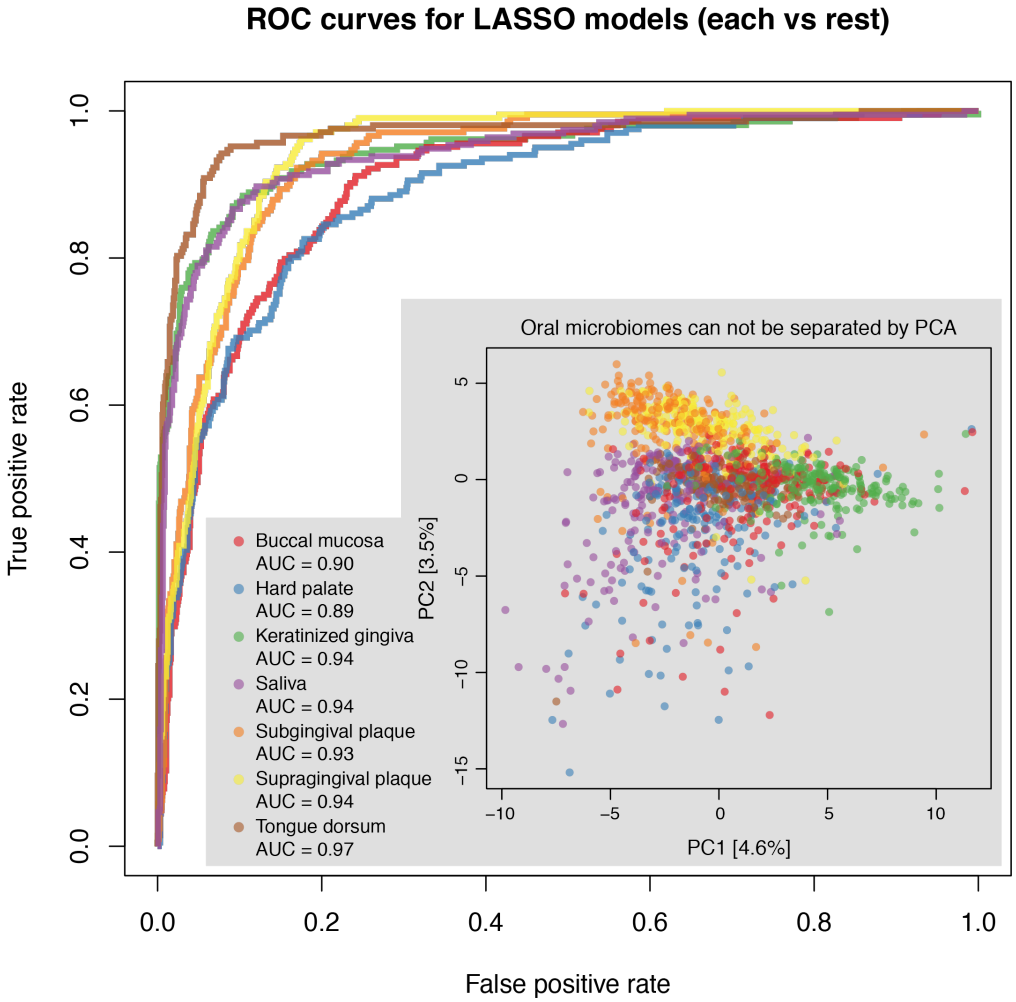
Ordination versus modelling (I)

- 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.



A typical machine learning workflow

[Wirbel et al., *Genome Biol.* 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.

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

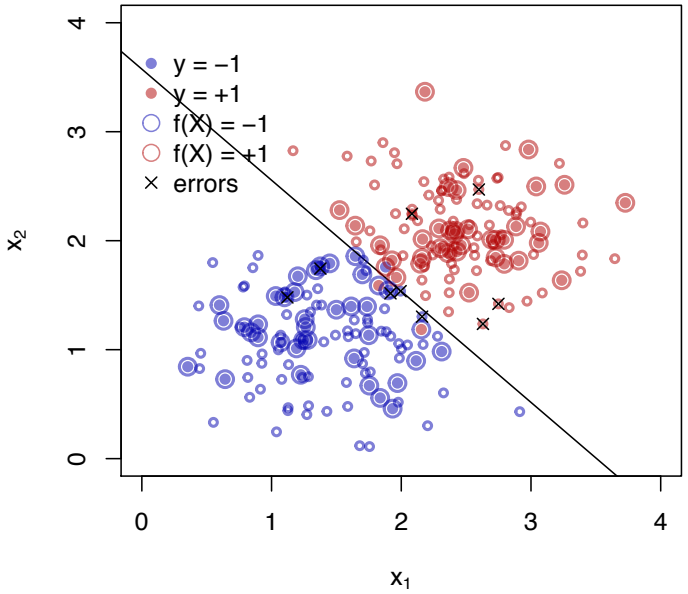
- are all **independent of prevalence** (fraction of positives in the population)

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

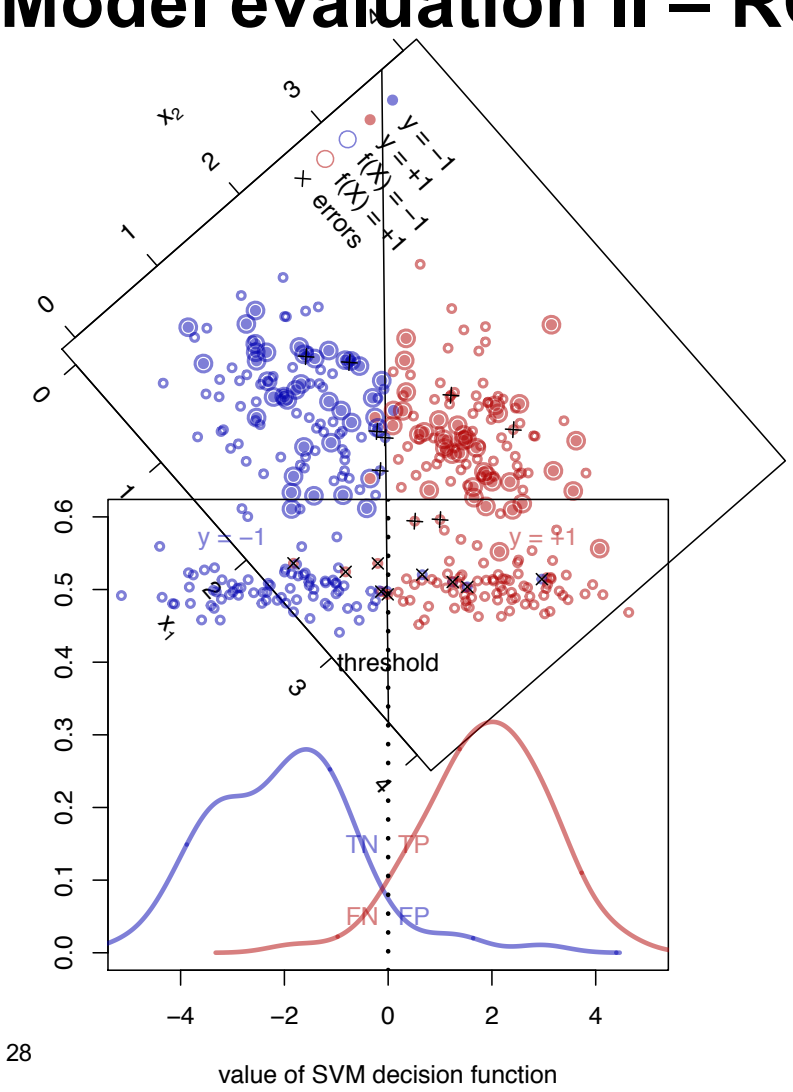
- are both **dependent on prevalence** (fraction of positives in the population)

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

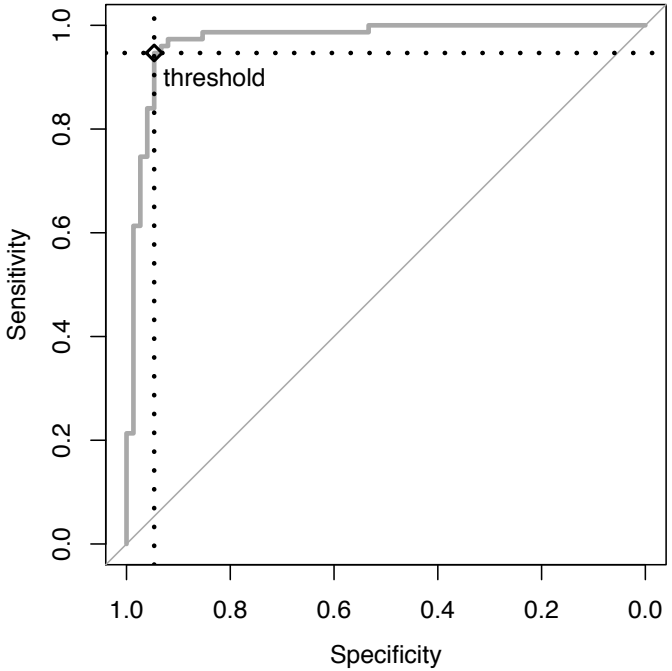
Model evaluation II – ROC curves

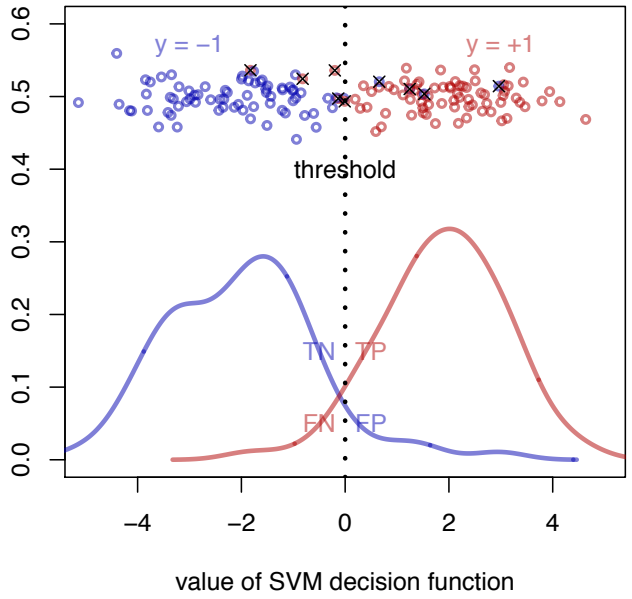


Model evaluation II – ROC curves



Model evaluation II – ROC curves

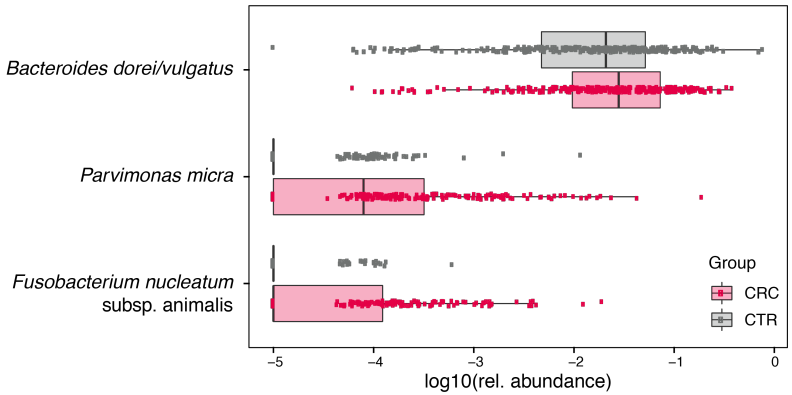


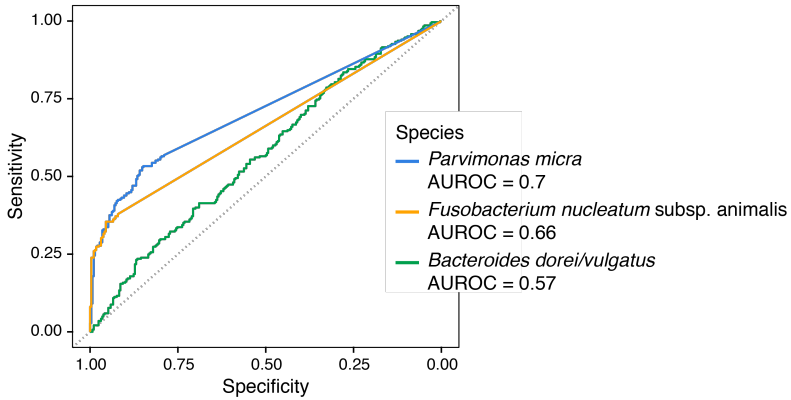


- Change decision threshold to obtain other **trade-offs between sensitivity and specificity**
- Receiver operating characteristic (ROC) curve plots all of them
- **Area under the ROC curve** as a summary statistic

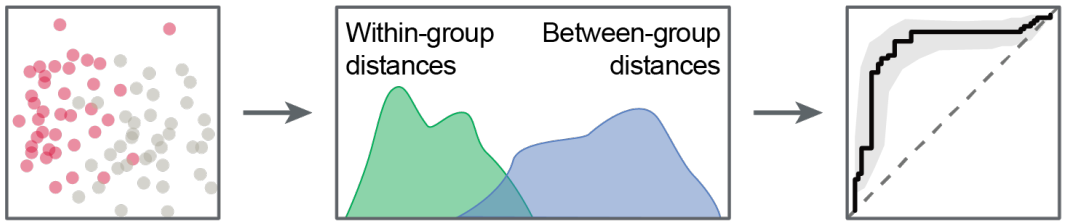
ROC curves from single features / distances

- 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.



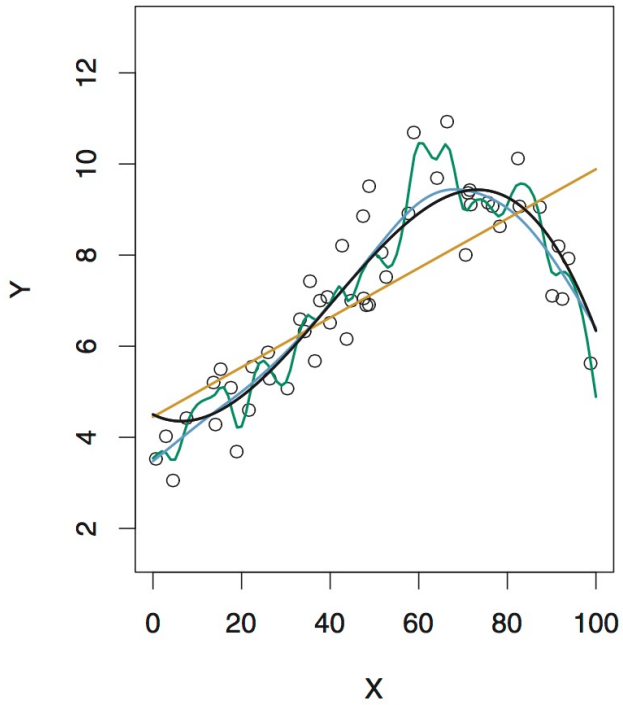
Distances (beta-diversity)

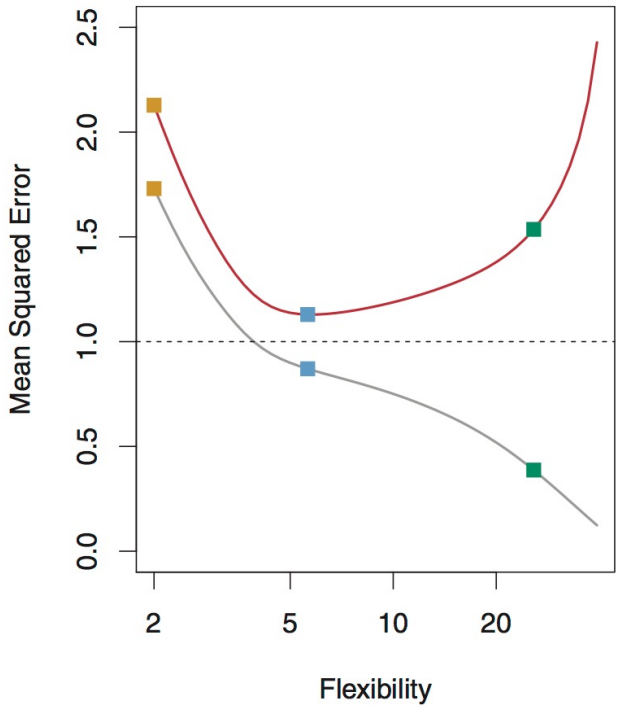
Distance histograms

Separation quantified by AUROC

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]

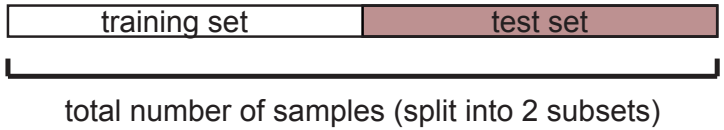
Resampling data for external validation or cross validation

Some data needs to be reserved for model evaluation….

Resampling data for external validation or cross validation

Some data **– always! –** 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 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

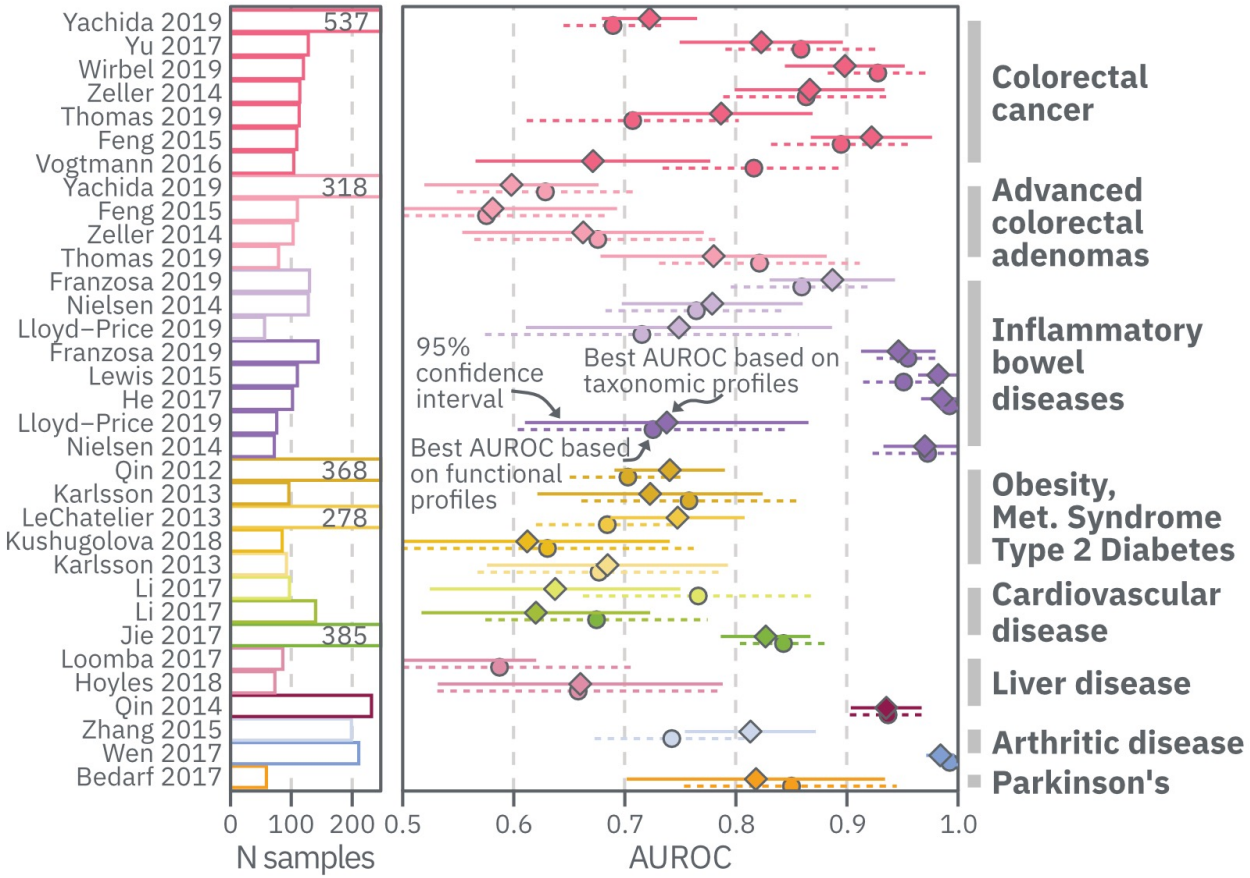
- **Model fitting is easy, model evaluation is not at all**!
  Understand the generalization assessed – consult experts!

- 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!

- 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., *Genome Biol.* 2020]