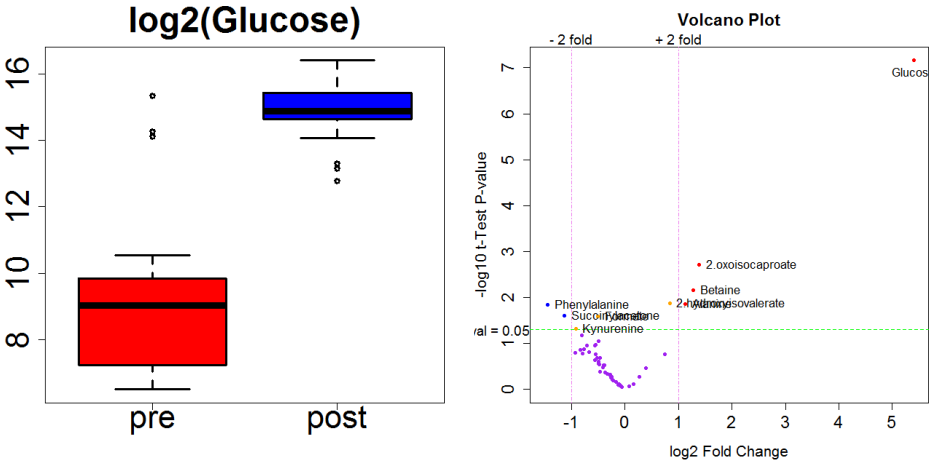
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

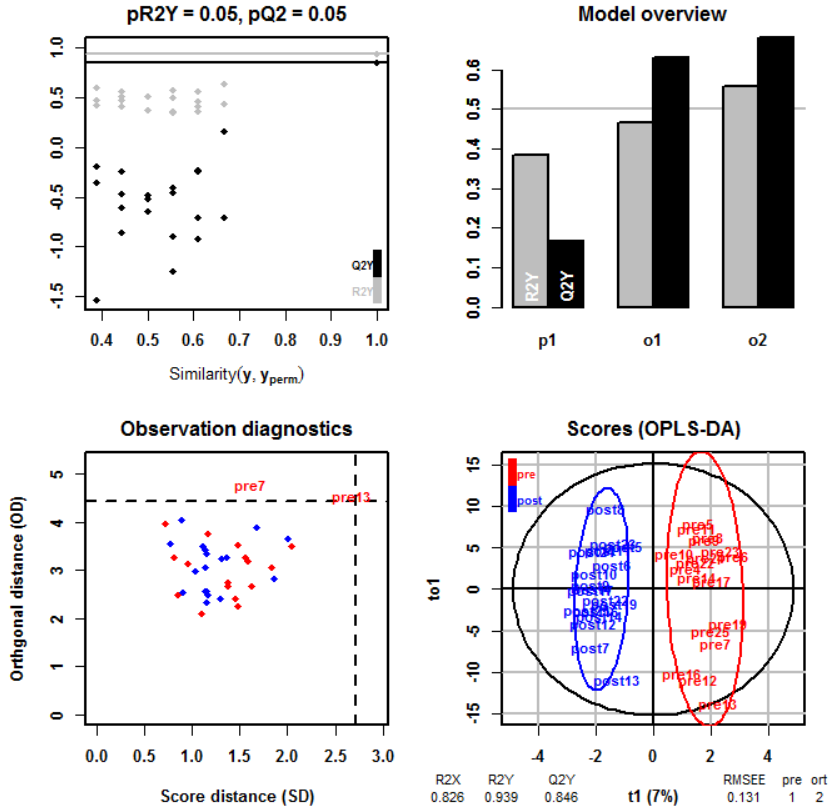
Eidgenössisches Departement für Wirtschaft,
Bildung und Forschung WBF

Agroscope

Vergleich einer univariaten und einer multivariaten «OMICS» - Datenanalyse; zB. für Genomics, Transcriptomics, Proteomics, *Metabolomics*, Foodomics, …

Dominik Guggisberg, Agroscope



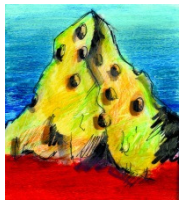


CAS/DAS Angewandte Statistik,

04.05.2018

www.agroscope.ch I gutes Essen, gesunde Umwelt

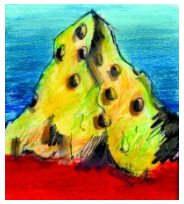
Agenda



1. **Einleitung / Ziele**
2. **Univariat;** traditionelle Tests, parametrisch, nicht-parametrisch, FDR (false discovery rate)
   Biologische Effekte («Volcano Plot» ≅ Scatter-plot)
   R: library(metabolomics) (cran.r-project.org/web/packages)
3. **Multivariat;** PCA, PLS-DA, OPLS-DA
   «Feature selection», VIP (Variable Importance in Projection)
   R: library(ropls) (bioconductor.org)
4. **Zusammenfassung/Ausblick**

Agroscope

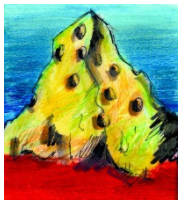
1. Einleitung / Ziele:

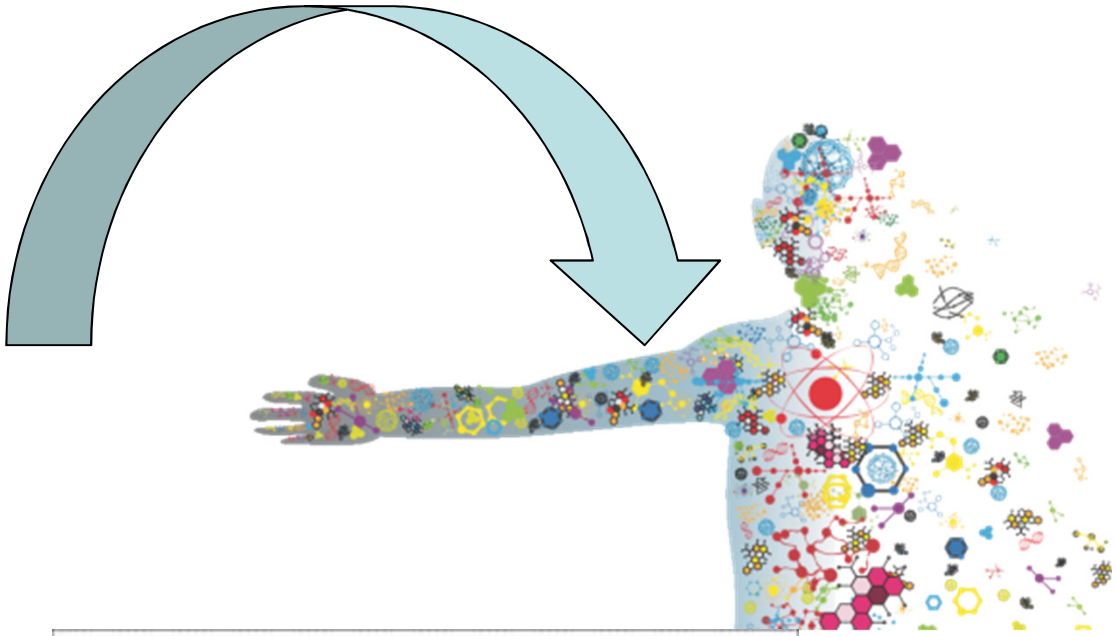


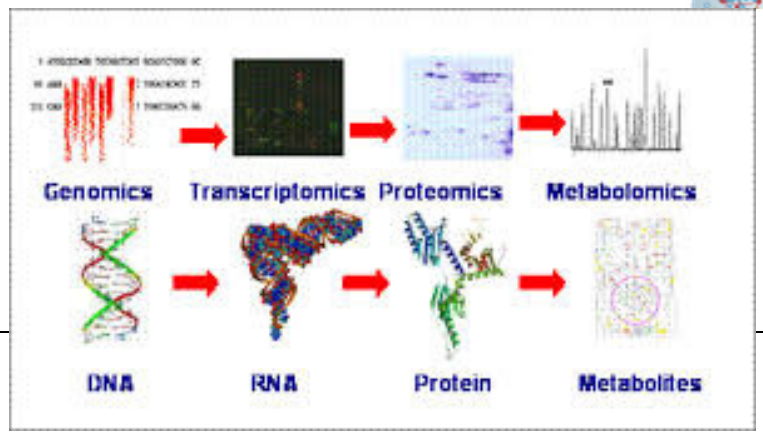
Die Erforschung des Metaboloms wird als **Metabolomik** bezeichnet (im englischen: *metabolomics*). Diese umfasst die Wechselwirkung der darin enthaltenen Metaboliten, deren Identifizierung und Quantifizierung. (=> Einfluss auf den Stoffwechsel verstehen.)

Agroscope

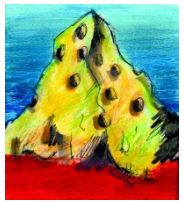
zB. Metabolomics







Datensatz: library(metabolomics)*; data(treated)



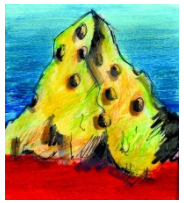
S	Group	met1	met2	met3	. . . .	metN
S1	A	0.6358	0.0851	0.3665	. . . .	1.0024
S2	A	0.5871	0.0935	0.3421	. . . .	1.0329
.....	...	....	....	....		....
S19	B	0.6650	1.0705	0.6710	. . . .	0.7319
S20	B	0.6907	1.0341	0.6858	. . . .	0.7376 ...

A metabolomics study with ***paired*** observations.
Description: A data frame with data collected from several subjects ***before*** and ***after*** a specific treatment.
(=> Diabetes-Forschung, 2012)

n<- ncol(treated) # n= 54 (53 Metaboliten)
m<- nrow(treated) # m= 36 (= 2x18 Subjekte)

* cran.r-project.org/web/packages

2. Univariate Analyse

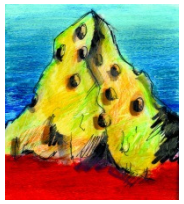


- Parametrisch: T-Test, ANOVA
- Nicht-parametrisch: Wilcoxon-Test (Mann-Whitney), Kruskal-Wallis
- Probleme: “Alpha-Fehler - Kumulierung” (=> multiples Testen)
- => False Discovery Rate (FDR): Korrektur mit Benjamini-Hochberg[1]
- => Filterung der Daten, damit weniger “Features” vorhanden sind: Nachteil; ev. “verlieren” wir wichtige “Features”?
- Grafik: “Volcano Plot”

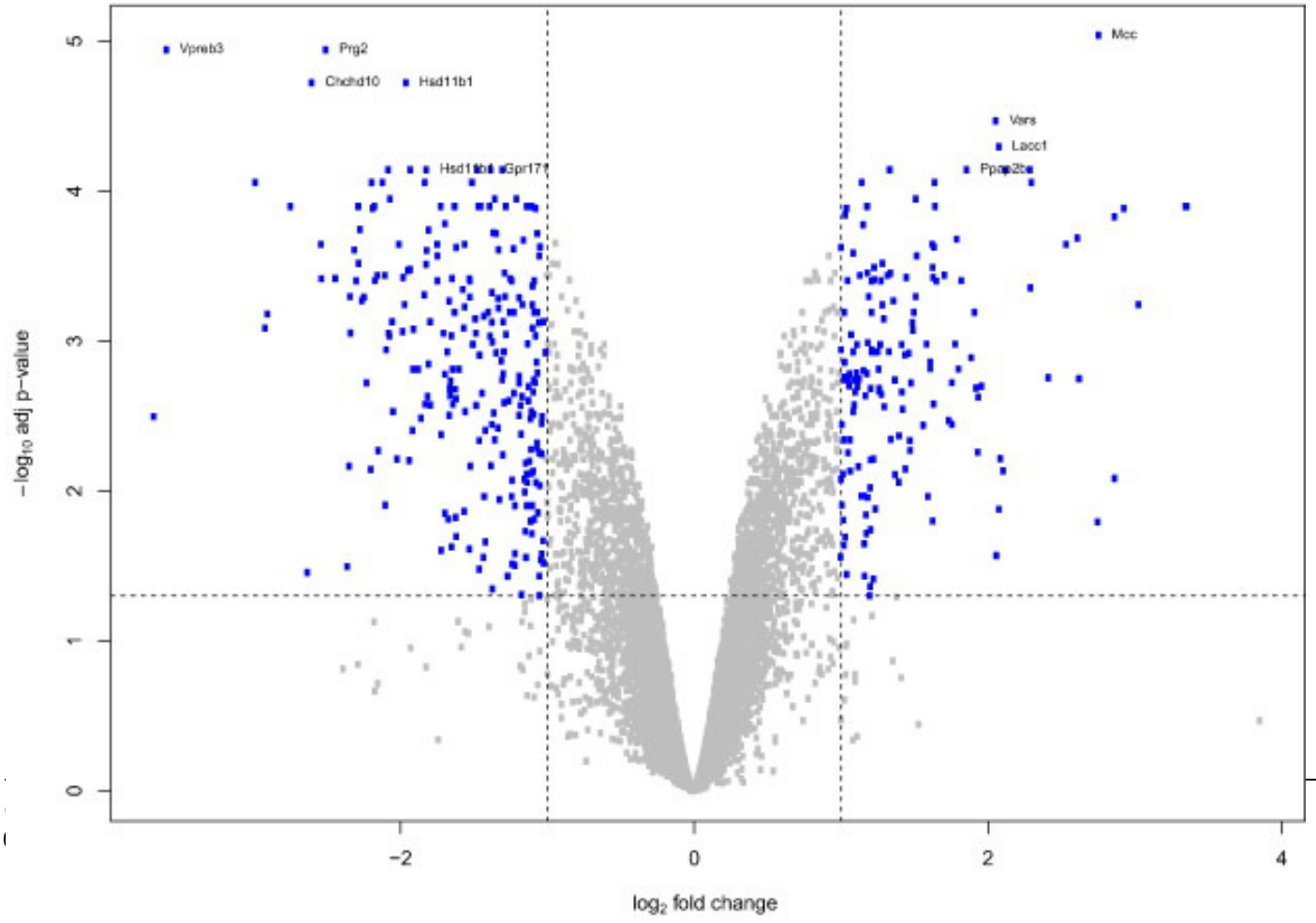
Der Begriff wurde erstmals 1995 von Yoav Benjamini und Yosi Hochberg definiert.[1]

Benjamini, Yoav; Hochberg, Yosef: "Controlling the false discovery rate: a practical and powerful approach to multiple testing" In: Journal of the Royal Statistical Society, Series B Nr. 57, 1995, S. 289–300

Volcano Plot (Typ: Scatterplot)

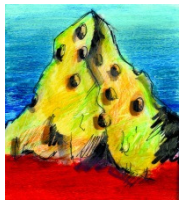


- x-Achse: log2 fold change = log 2 (mean post/mean pre)
- y-Achse: - log10 (p-values)



Agroscope

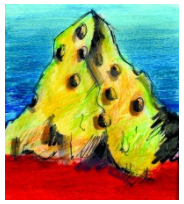
Volcano plot für data(treated) (library(metabolomics))



- ### Volcano plot
- data(treated)
- treated.log <- LogTransform(treated, base = 2)$output
- results <- TwoGroup(treated.log, ***paired = TRUE***)$output
- (Description TwoGroup: This function computes t- statistics, p-values, adjusted p-values, fold changes and standard errors for each metabolite given a series of replicates and two biological conditions).

```
> results
                      t-statistic      p-value BH Adjusted p-value fold change standard error
2.hydroxybutyrate       -1.8408367 8.316721e-02        1.377457e-01 -0.28430458      0.1544431
2.hydroxyisobutyrate     0.3673505 7.178907e-01        7.460433e-01  0.07671834      0.2088423
2.hydroxyisovalerate     4.1505989 6.693494e-04        5.912587e-03  0.83338203      0.2007860
2.hydroxyvalerate       -2.5875902 1.916541e-02        4.416378e-02 -0.40917314      0.1581290
2.methylglutarate       -1.7963295 9.023429e-02        1.449217e-01 -0.33423740      0.1860669
2.oxoisocaproate         8.0900778 3.132787e-07        8.301887e-06  1.38511010      0.1712110
```

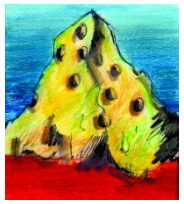
Volcano plot für data(treated) (library(metabolomics))



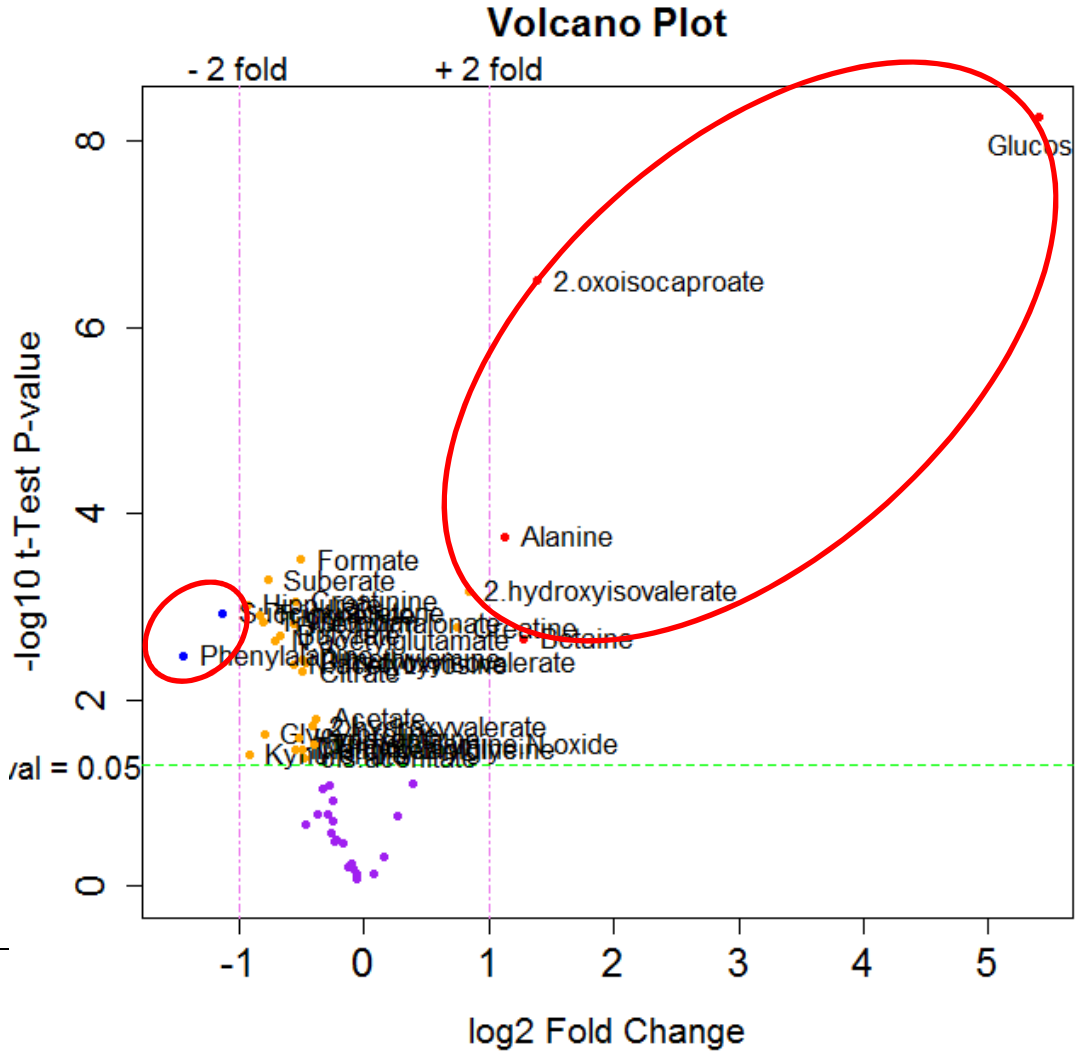
- ### Volcano plot
- data(treated)
- treated.log <- LogTransform(treated, base = 2)$output
- results <- TwoGroup(treated.log, ***paired = TRUE***)$output
- (Description TwoGroup: This function computes t- statistics, p-values, adjusted p-values, fold changes and standard errors for each metabolite given a series of replicates and two biological conditions).

```
> results
                      t-statistic      p-value BH Adjusted p-value fold change standard error
2.hydroxybutyrate       -1.8408367 8.316721e-02        1.377457e-01 -0.28430458      0.1544431
2.hydroxyisobutyrate     0.3673505 7.178907e-01        7.460433e-01  0.07671834      0.2088423
2.hydroxyisovalerate     4.1505989 6.693494e-04        5.912587e-03  0.83338203      0.2007860
2.hydroxyvalerate       -2.5875902 1.916541e-02        4.416378e-02 -0.40917314      0.1581290
2.methylglutarate       -1.7963295 9.023429e-02        1.449217e-01 -0.33423740      0.1860669
2.oxoisocaproate         8.0900778 3.132787e-07        8.301887e-06  1.38511010      0.1712110
```

Fortsetzung: Volcano Plot

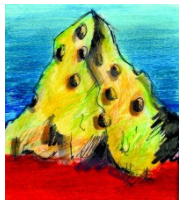


- pval <- results[, 2]
- fc <- results[, 4]
- VolcanoPlot(fc, pval, lwd=1, cexlab = 1.1, cex=1.2, cex.axis=1.5, cex.main=1.4, cex.lab=1.3)



Agroscope

3. Multivariate Analyse:

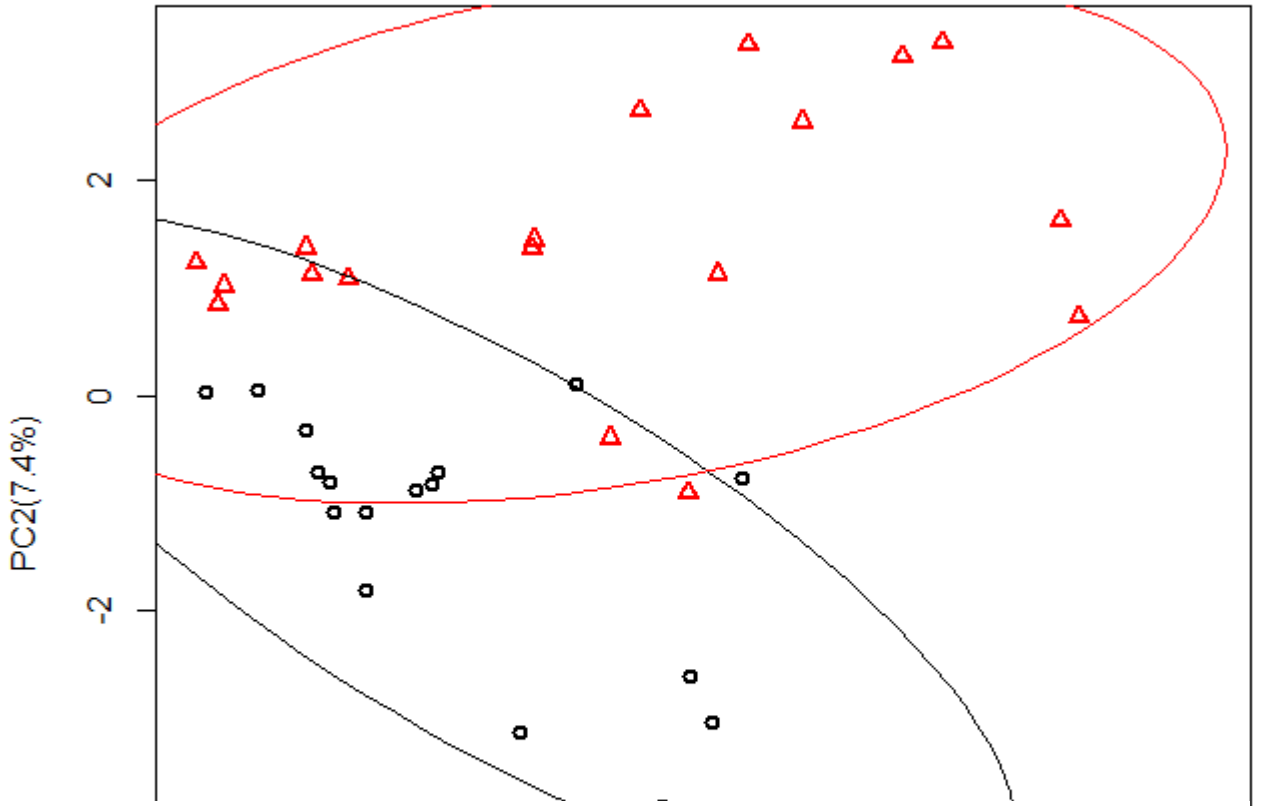


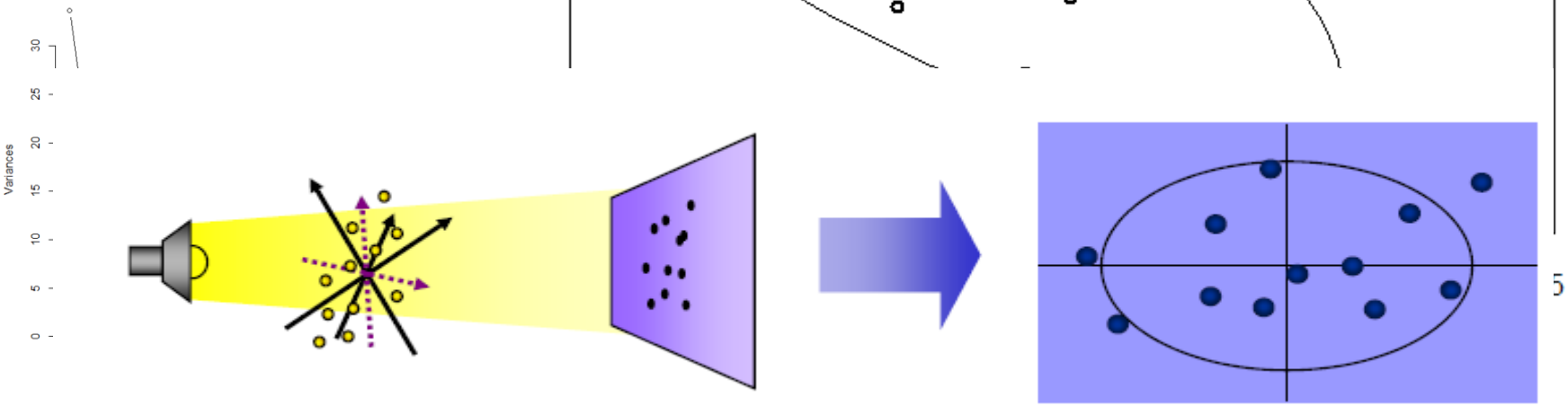
- The ropls R package implements the **PCA**, **PLS(-DA)** and **OPLS(-DA)** approaches with the original, **NIPALS**-based, versions of the algorithms. It includes the **R2** and **Q2** quality metrics, the permutation **diagnostics**, the computation of the **VIP** values, the score and orthogonal distances to detect **outliers**, as well as many **graphics** (scores, loadings, predictions, diagnostics, outliers, etc).

library(ropls), www.bioconductor.org

Hauptkomponentenanalyse

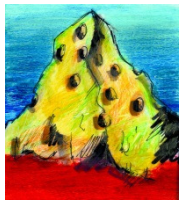
- **PCA center=T, scale=T, (nicht-log)**

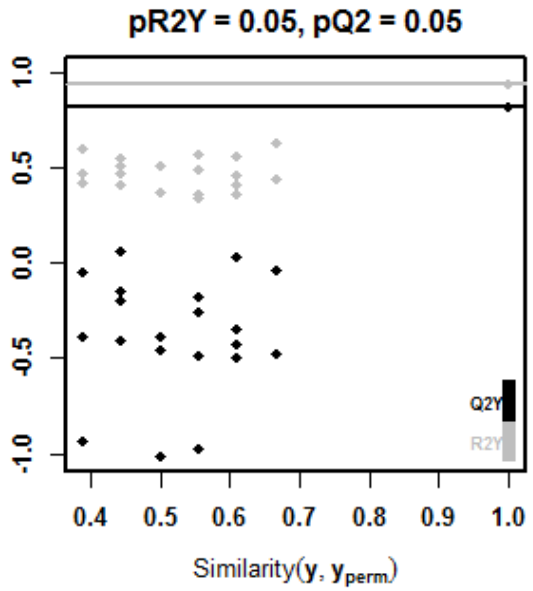


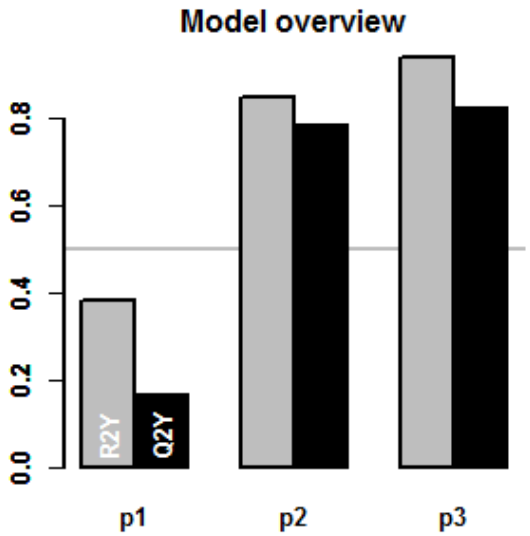


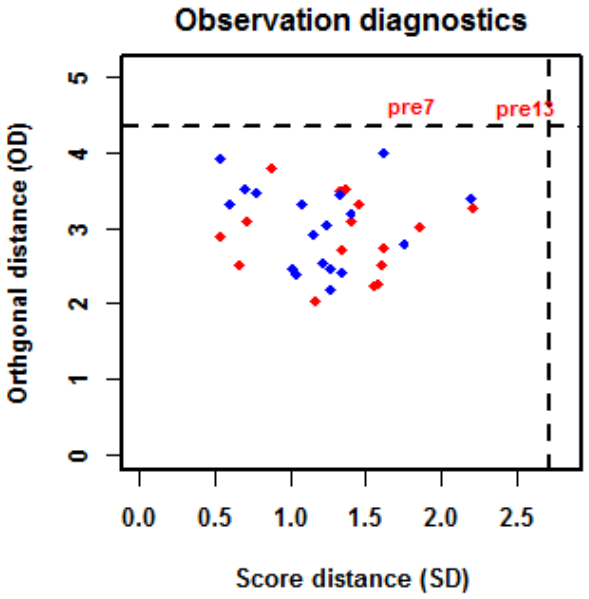
Agroscope

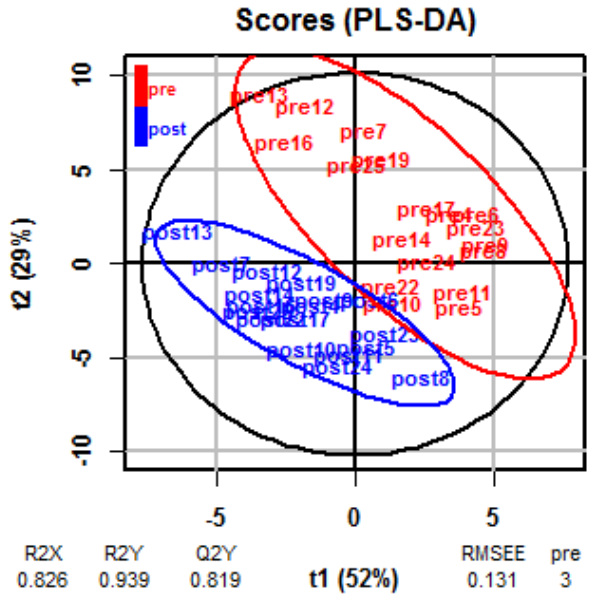
PLS-DA (log-transformierte Daten)





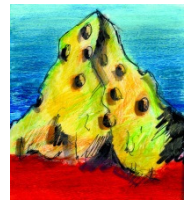


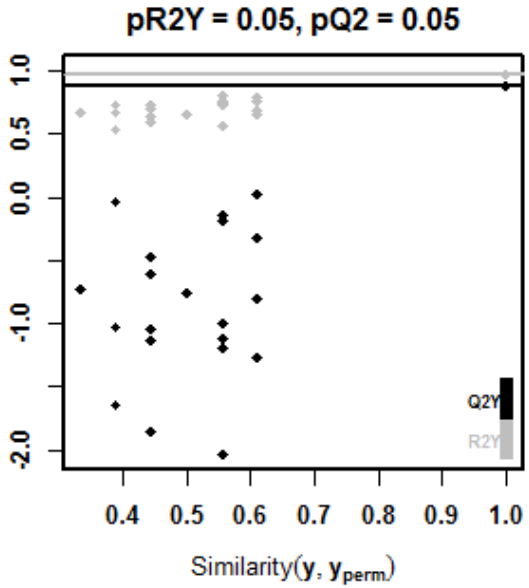


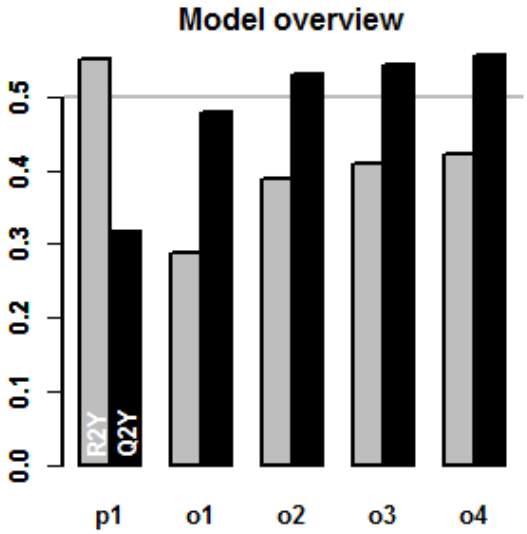


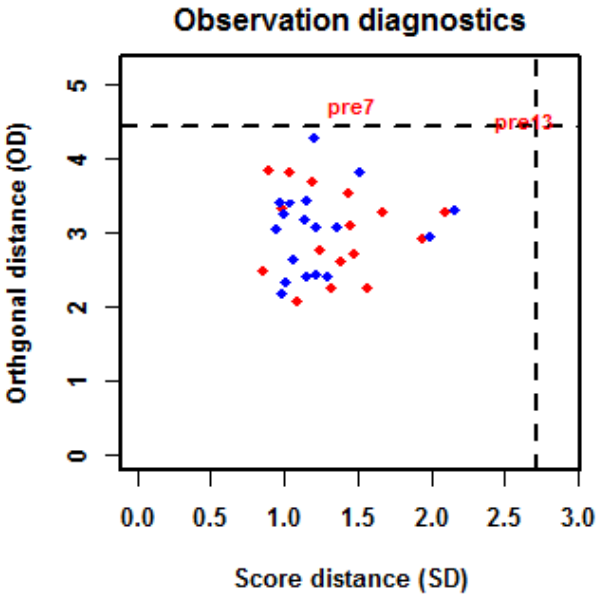
Latent variable regression based on covariance between X (predictors) and Y (response). (ideal for multi-collinear predictors!)

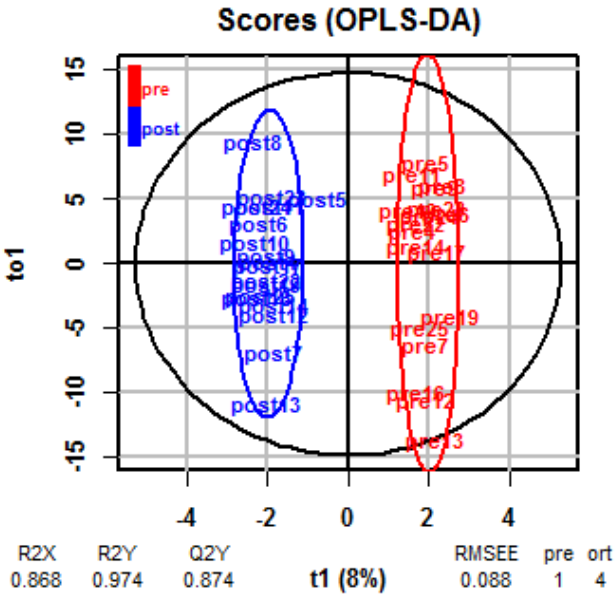
OPLS-DA (log-transformierte Daten)







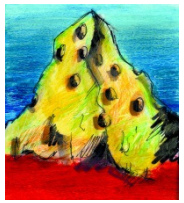




Model separately the variations of X (predictors) correlated and orthogonal to Y (response). Q2Y metrics and permutation testing to avoid overfitting. VIP (loading weights and variability of the response explained by this component: feature selection!)

Agroscope

VIP (Variable Importance in Projection)

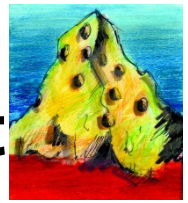


- head(newdata1, 10), log-transformierte Daten

Glucose	2-oxoisocaproate	2-hydroxyisovalerate	Betaine
3.178962	1.906171	1.664372	1.657276
Alanine	**Phenylalanine**	**Formate**	**Succinylacetone**
1.657078	1.603227	1.525911	1.499841
Kynurenine	**Tryptophan**		
1.321389	1.301598		

(sortiert!)

4. Zusammenfassung: Univariat/Multivariat

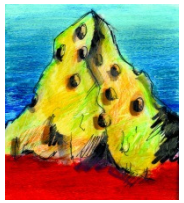


- **Univariat (Volcano-Plot)**
- Glucose
- Oxoisocaproate
- Alanine
- Betaine
- Succinylacetone
- Phenylalanine

- **Multivariat (OPLS-DA, log-Daten)**
- Glucose
- Oxoisocaproate
- **Hydroxyisovalerate**
- Betaine
- Alanine
- Phenylalanine
- **Formate**
- Succinylacetone
- …

Agroscope

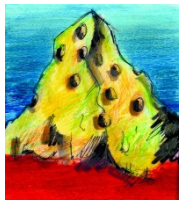
Skalierung: Unit variance oder Pareto



- Unit variance scaling: jede Variable wird durch die Standardabweichung dividiert.
- Pareto scaling: jede Variable wird durch die Wurzel der Standardabweichung dividiert (häufig in Metabolomics-Studien verwendet).
- In der Literatur wird zB. auch empfohlen:
  Range scaling: $\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{(xi_{max} - xi_{min})}$

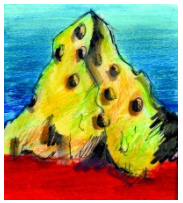
- Robert A van den Berg,[1] Huub CJ Hoefsloot,[2] Johan A Westerhuis,[2] Age K Smilde,[1,2] and Mariët J van der Werf[1], «Centering, scaling, and transformations: improving the biological information content of metabolomics data", BMC Genomics, 2006.

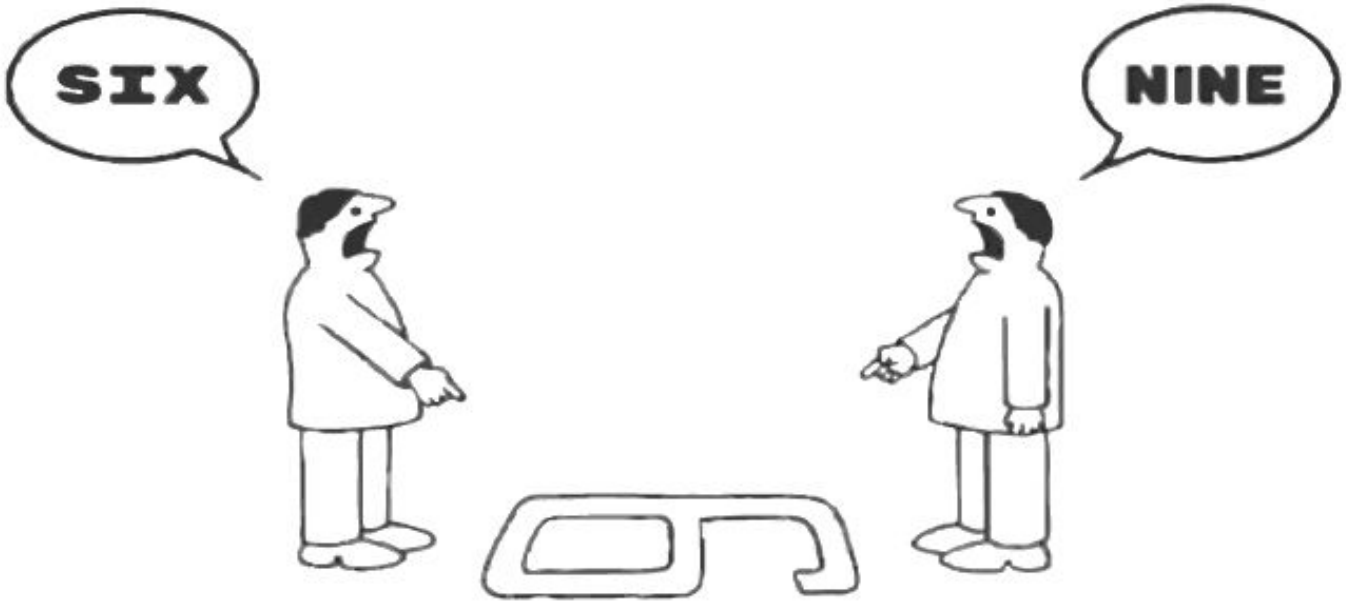
4. Ausblick



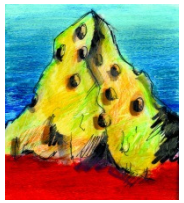
- Univariat oder multivariat?
- Unit variance, Pareto oder Range scaling?
- Log-Transformation?
- Plausibilitätsprüfung , „Know-how“ aus der Praxis!

Agroscope





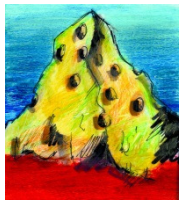
Just because you are right,
does not mean, I am wrong.
You just haven't seen life
from my side.

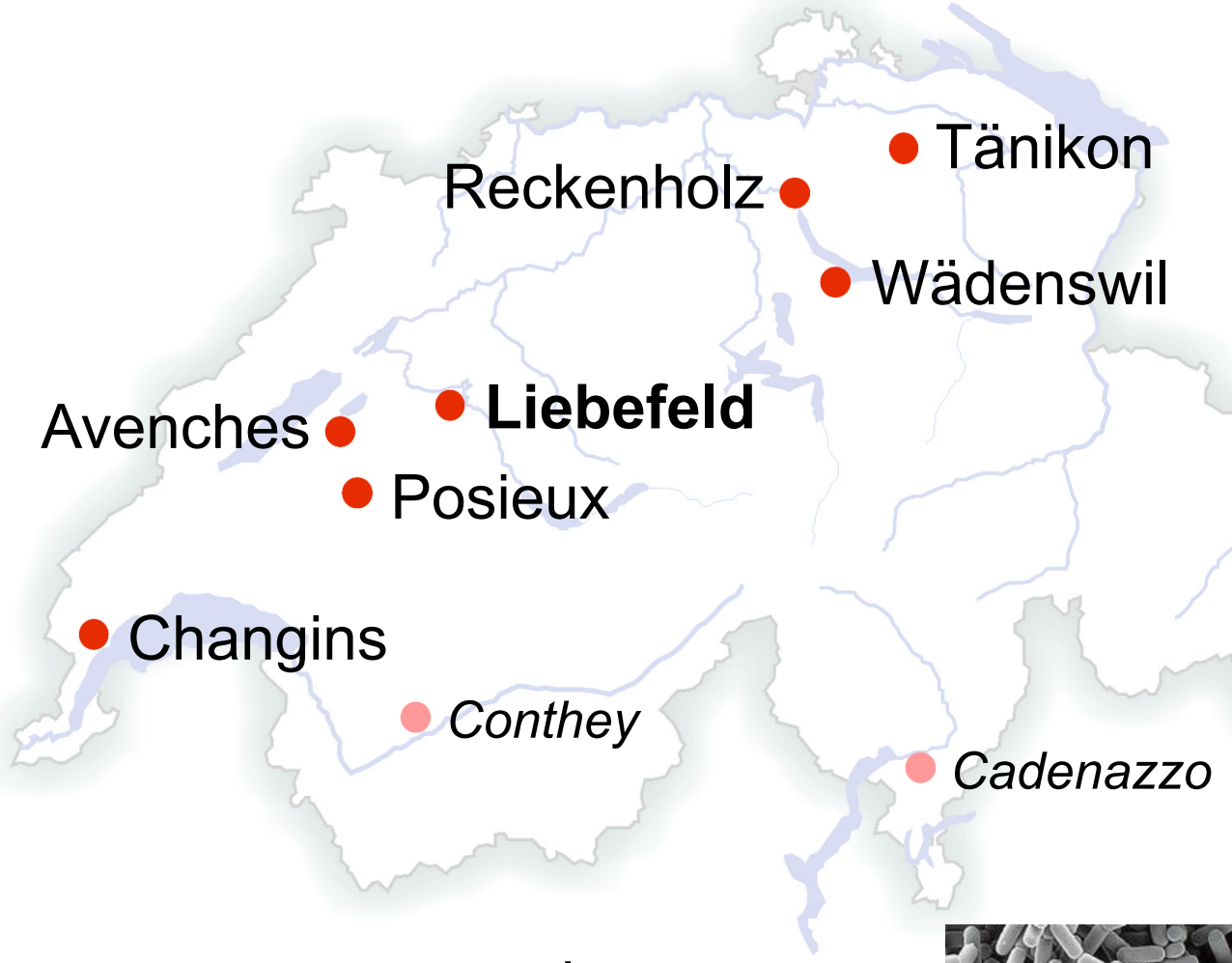


Fragen?

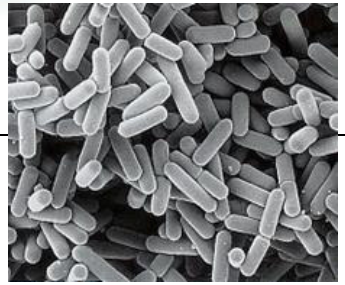
Besten Dank für eure Aufmerksamkeit

Agroscope

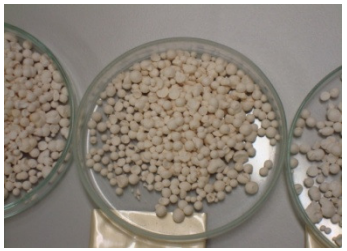


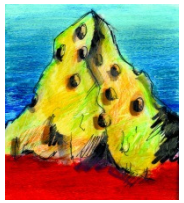


www.agroscope.ch



Agroscope: Liebefeld Milchsäurebakterien





Agroscope