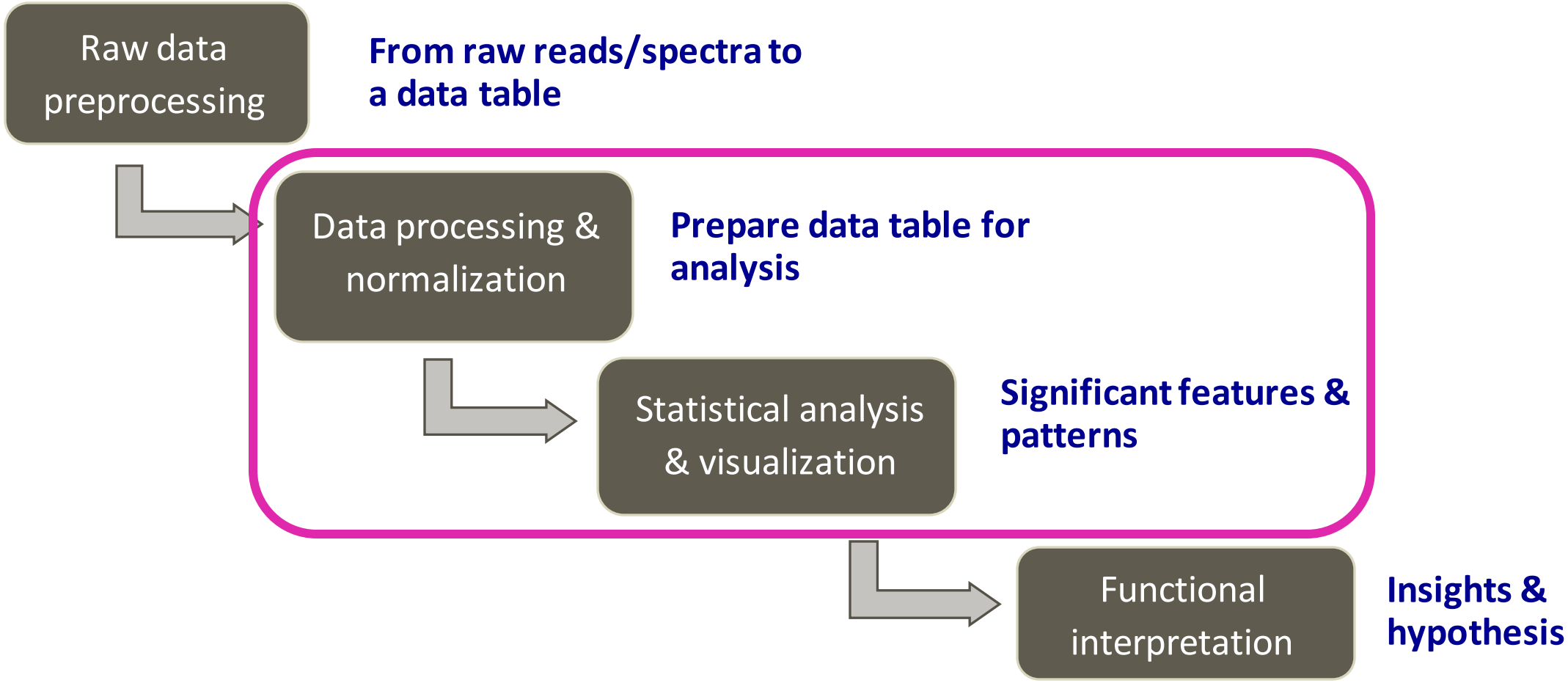
19th Annual Conference of the Metabolomics Society
METABOLOMICS 2023
NIAGARA FALLS, CANADA JUNE 18-22

Statistical analysis (I)
-- simple experimental designs

Jessica Ewald, Postdoctoral Fellow
jessica.ewald@mcgill.ca
McGill University, Montreal, QC Canada

Omics Data Analysis (in a nutshell)



DATA PROCESSING

-- prepare data for main analysis

Data processing ….

General steps

1. (Samples) Quality checking
2. (Features) Missing value imputation
3. (Features) Data filtering
4. (Both) Normalization

Sample Space

Feature Space

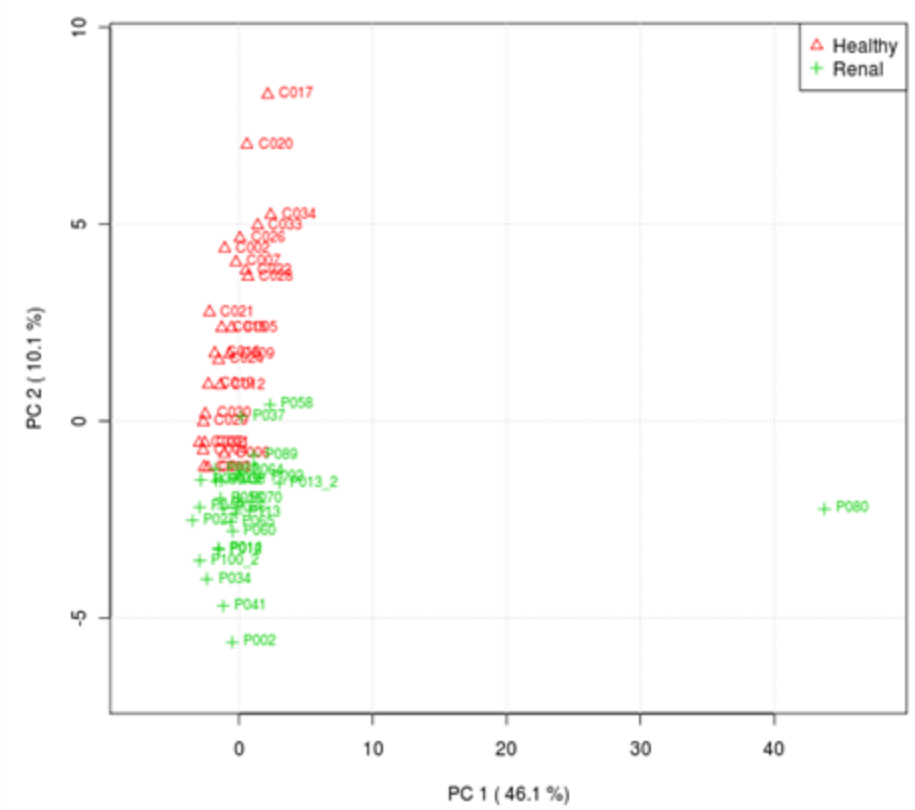
	PIF_178	PIF_087	PIF_090	NETL_005_V1
1,6-Anhydro-beta-D-glucose	40.85	62.18	270.43	154.47
1-Methylnicotinamide	65.37	340.36	64.72	52.98
2-Aminobutyrate	18.73	24.29	12.18	172.43
2-Hydroxyisobutyrate	26.05	41.68	65.37	74.44
2-Oxoglutarate	71.52	67.36	23.81	1199.91
3-Aminoisobutyrate	1480.3	116.75	14.3	555.57
3-Hydroxybutyrate	56.83	43.82	5.64	175.91
3-Hydroxyisovalerate	10.07	79.84	23.34	25.03
3-Indoxylsulfate	566.8	368.71	665.14	411.58
4-Hydroxyphenylacetate	120.3	432.68	292.95	214.86
Acetate	126.47	212.72	314.19	37.34
Acetone	9.49	11.82	4.44	206.44
Adipate	38.09	327.01	131.63	144.03
Alanine	314.19	871.31	464.05	589.93
Asparagine	159.17	157.59	89.12	273.14
Betaine	109.95	244.69	116.75	278.66
Carnitine	265.07	120.3	25.03	200.34
Citrate	3714.5	2617.57	862.64	13629.61
Creatine	196.37	212.72	221.41	85.63
Creatinine	16481.6	15835.35	24587.66	20952.22
Dimethylamine	632.7	607.89	735.1	1064.22
Ethanolamine	645.48	487.85	407.48	820.57
Formate	441.42	252.14	249.64	468.72
Fucose	336.97	198.34	186.79	407.48
Fumarate	7.69	18.92	7.1	96.54
Glucose	395.44	8690.62	1352.89	862.64
Glutamine	871.31	601.85	301.87	1685.81
Glycine	2038.56	1107.65	620.17	5064.45
Glycolate	685.4	651.97	141.17	70.81

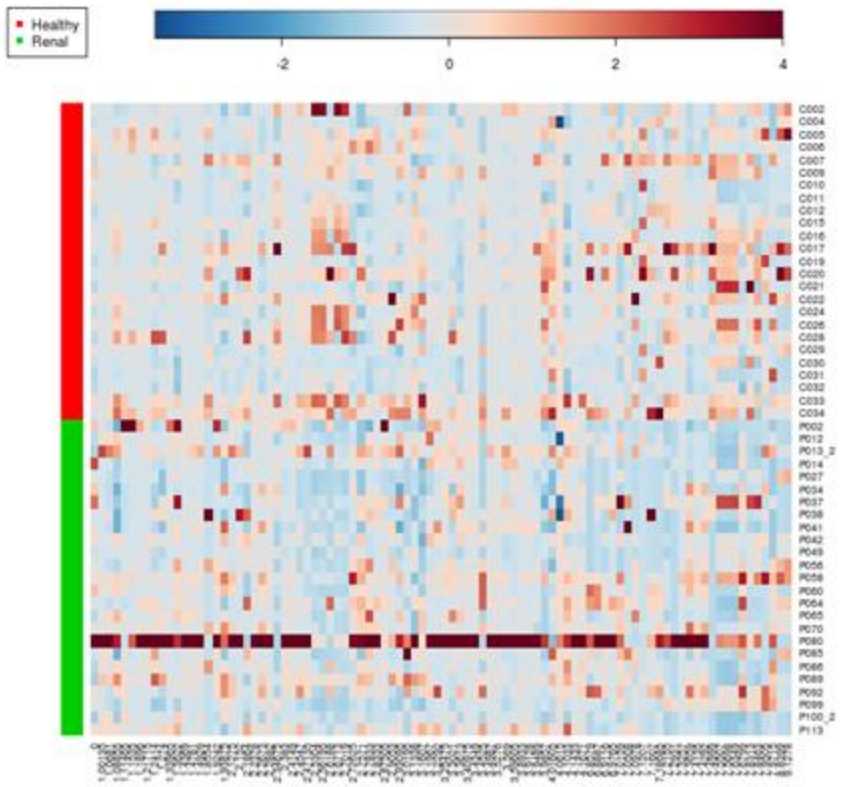
Quality checking

- The first & most critical step before analysis
  - ✓Garbage in and garbage out
- Depending on
  - ✓Good experimental design
  - ✓Good laboratory practice
- Pay attention to
  - ✓Outliers
  - ✓Batch effects

Outliers (I)

- Relative to the majority

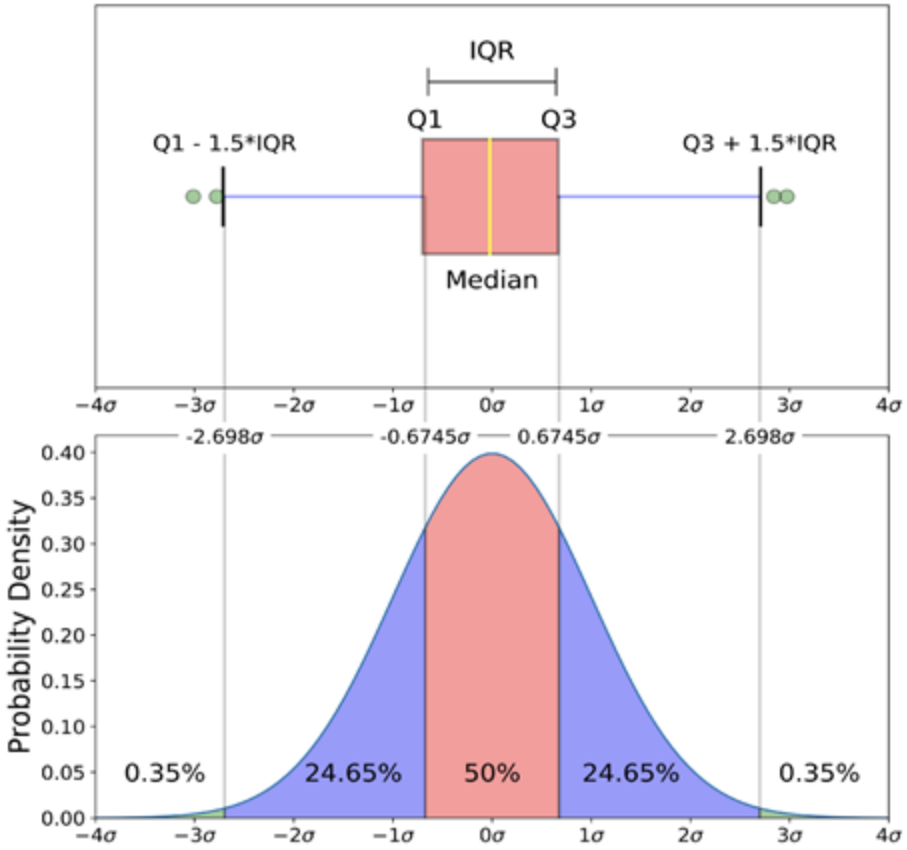




Outliers (II)

Mainly concerns **sample** outliers, not on feature space

- Interested in large, systemic outliers: measurements impacted for whole sample
- Common & normal to have feature outliers
- Statistical feature outliers could be our target of interest

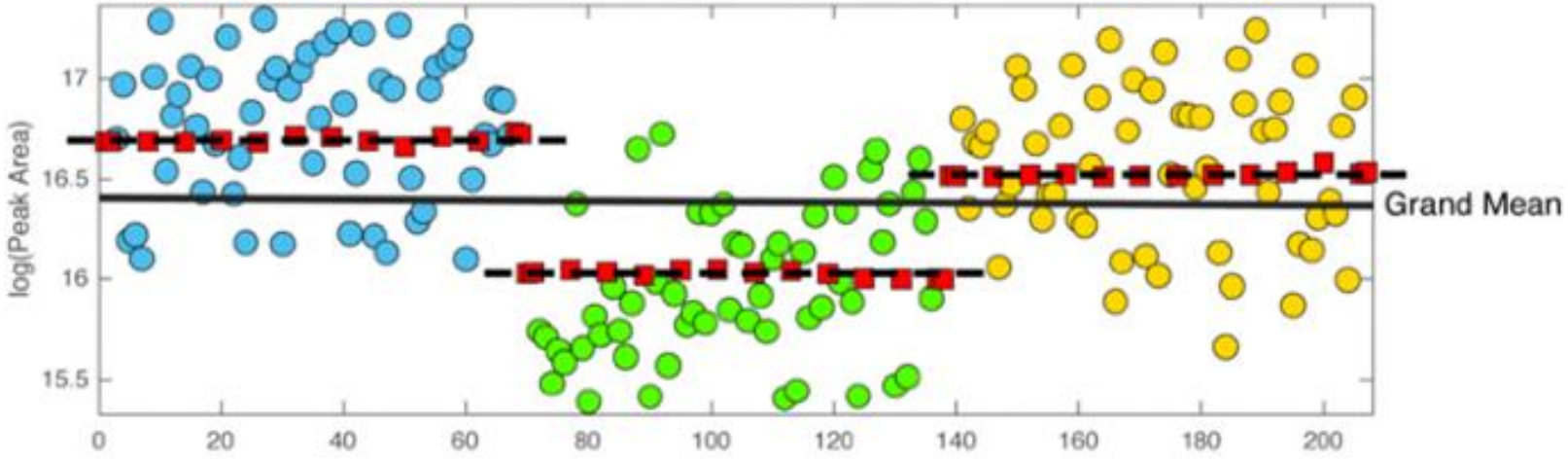


Source: towardsdatascience

Batch effects (I)

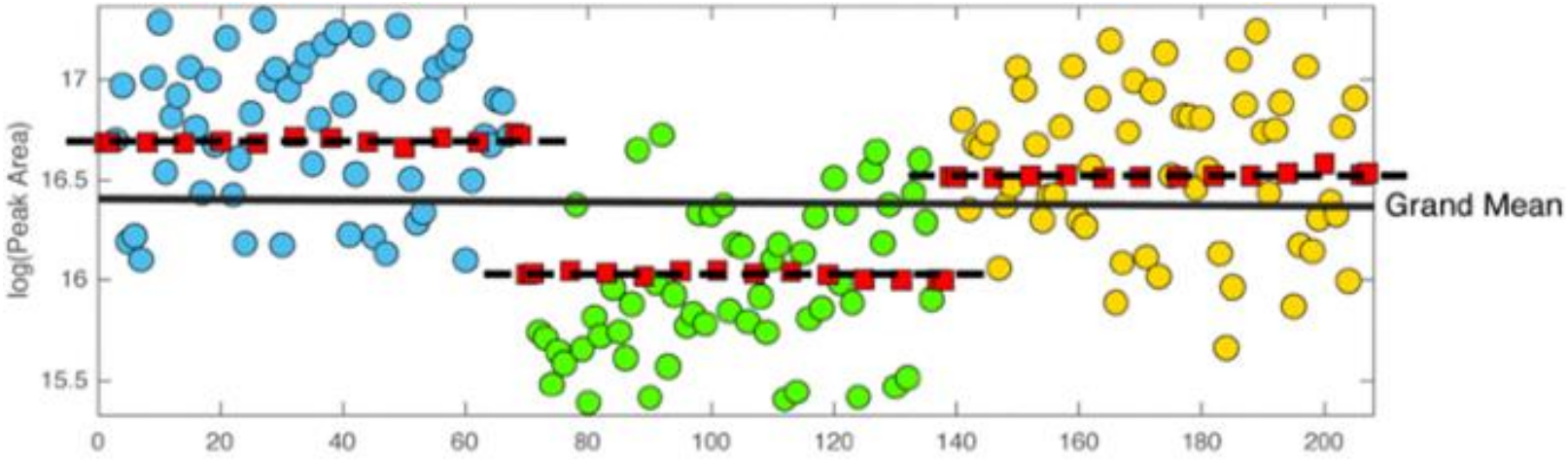
Omics data can have batch effects

- Display overall or systematic differences
- Technical (not biological) reasons
  - Sample preparation, machine run, technician, time before running samples



Metabolomics (2018) 14:72

Batch effects (II)



Control	Drug A	Drug B

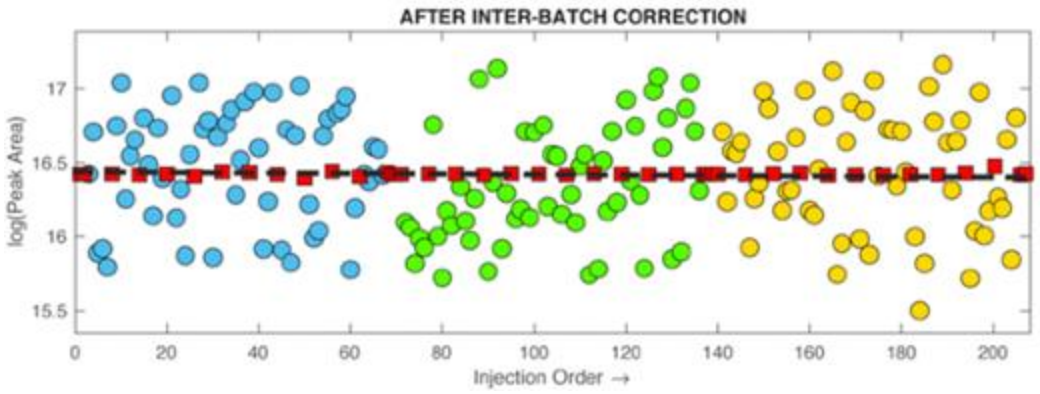
1/3 Control 1/3 Drug A 1/3 Drug B	1/3 Control 1/3 Drug A 1/3 Drug B	1/3 Control 1/3 Drug A 1/3 Drug B

Very bad!!! We cannot use statistics to correct for anything

Good experimental design

Metabolomics (2018) 14:72

Batch effects (III)



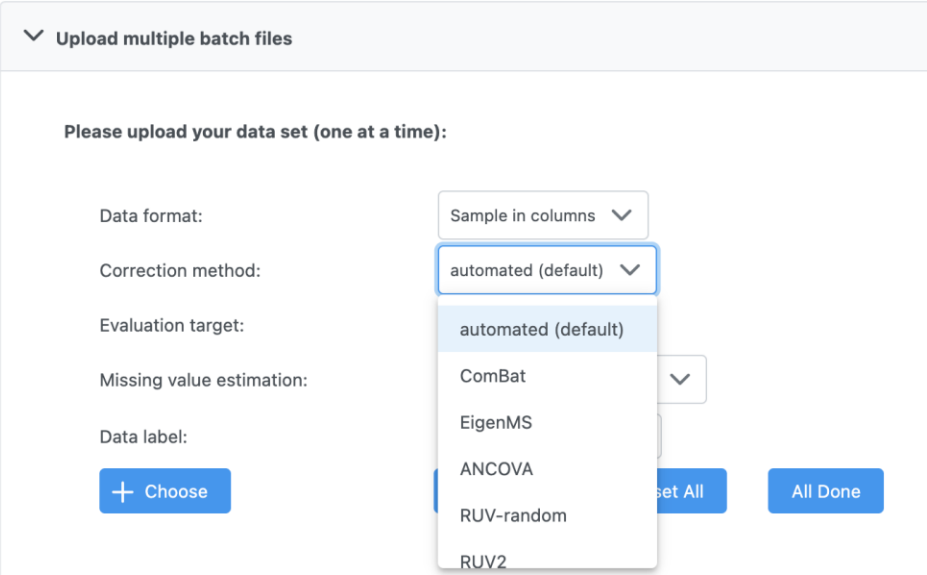
Metabolomics (2018) 14:72

Batch effect correction methods perform batch effect correction prior to statistical analysis;

- Internal standards
- Computational estimation

Implementation in MetaboAnalyst

- MetaboAnalyst currently supports nine well-established methods (ComBat, EigenMS, QC-RLSC, ANCOVA, RUV-random, RUV2, RUVseq, NOMIS and CCMN) for batch effect correction.
- The automated (default) will return the results with least distance among batches.



Some methods can include batch variables within the model for statistical analysis, such that differences associated with batch are accounted for during analysis. This is the concept we will use for meta-data or complex design

Missing values (I)

- Common in omics data. Can be introduced during data collection, or by algorithms during raw data pre-processing (i.e. peak picking)
- Most algorithms will complain if input contains missing values

1.4781	2	1.05	1.84	0.89	1.33	1.94	1.43	0.85	1.52	1.48	2.52	2.22
1.4929	2.03	1.06	1.86	0.88	1.33	1.13	1.46	0.88	0.97	1.47	1.88	2.19
1.8554	NA	0.47	0.83	NA	1.31	NA	NA	NA	NA	0.6	NA	0.79
1.9242	1.82	1.59	1.45	1.73	3.13	1.91	1.79	1.58	1.77	1.99	3.26	3.35
1.93875	NA	0.76	1.1	0.83	2.62	0.8	1.53	1.45	0.94	1.8	NA	3.36
2.1275	1.19	0.72	NA	NA	2.88	NA	1.68	NA	0.94	1.1	NA	3.16
2.152	NA	2.25	2.9	1.25	8.75	5.02	1.09	1.91	1.33	3.13	2.84	4.18
2.1864	NA	1.3	2.7	0.8	3.47	2.84	NA	0.9	NA	1.98	2.05	2.35
2.2378	1.58	0.61	1.03	1.75	0.84	1.51	0.72	0.77	1.15	0.85	0.79	0.96

Missing value (II)

- Goal: "guess" reasonable values
  - Must understand why the data are missing
  - Choose the appropriate imputation strategy

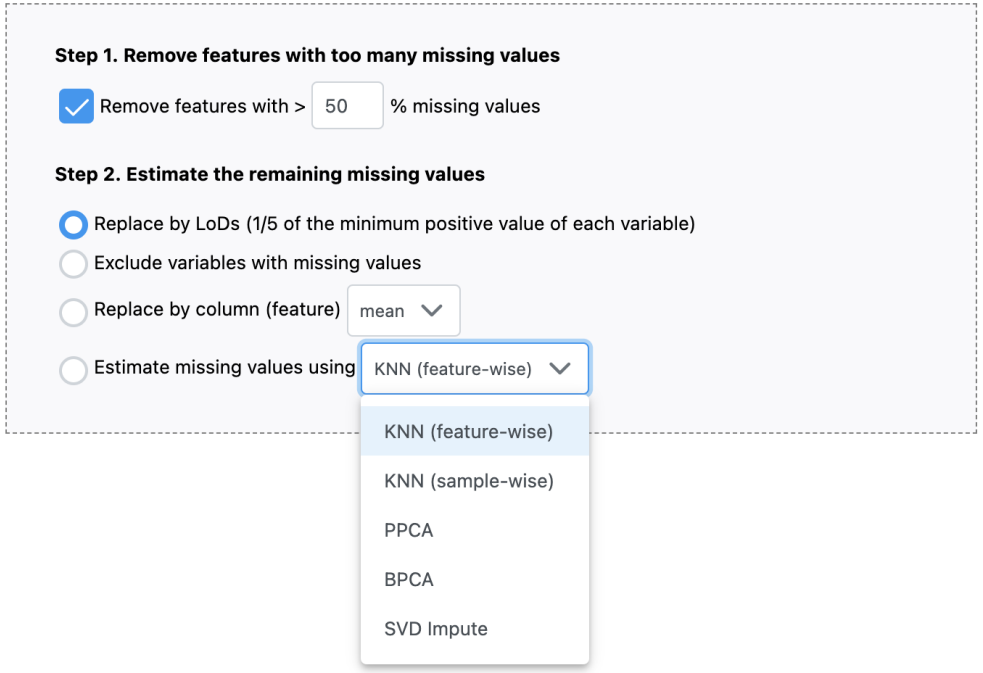
- Options:
  - Missing completely at random:
    - Ie. Machine fails randomly
  - Missing at random:
    - Depends on sample characteristics (age, sex, etc.)
    - Within group of 'similar' samples, missingness is random
  - Missing not at random:
    - Depends on the true value of the measured variable
    - Ie. Missing because metabolite is below the detection limit

Int J Epidemiol. 2014 Aug; 43(4): 1336–1339. PMCID: PMC4121561
Published online 2014 Apr 4. doi: 10.1093/ije/dyu080 PMID: 24706730

What is the difference between missing completely at random and missing at random?

Krishnan Bhaskaran* and Liam Smeeth

Great explanation!



Feature filtering (I)

- Not all features are informative
- There are redundancies in omics data for most features
- Filtering non-informative features before statistical analysis can often <u>significantly</u> improve the power

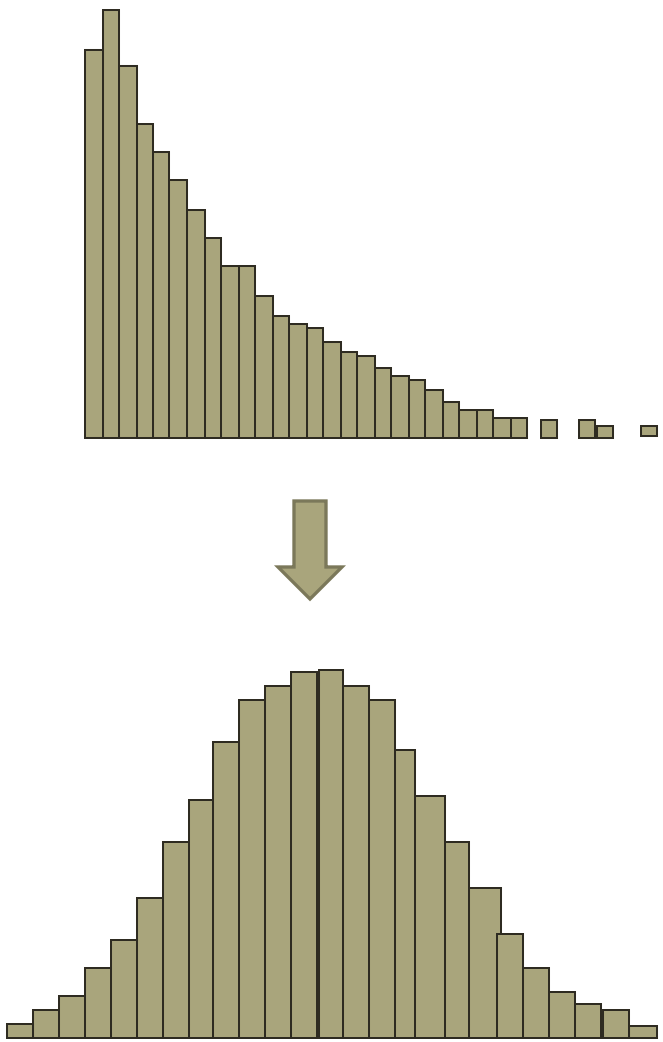
Feature filtering (II)

- **Low quality**
  - Too many missing values
  - Hard to measure: low repeatability based on QC
- **Low abundance**
  - Variables of very small values (close to baseline or detection limit).
- **Low variance**
  - Variables that are near-constant values throughout the experiment conditions (housekeeping or homeostasis)

DO NOT filter features based on their p-values or fold changes

Normalization (I)

- Most statistical methods work best when variables are normally distributed
  - Biological measurements are often right skew
- Variable abundance levels can vary across several magnitudes
  - Inconvenient for visualization
- Adjust other effects:
  - Dilutions, tissue volumes, etc

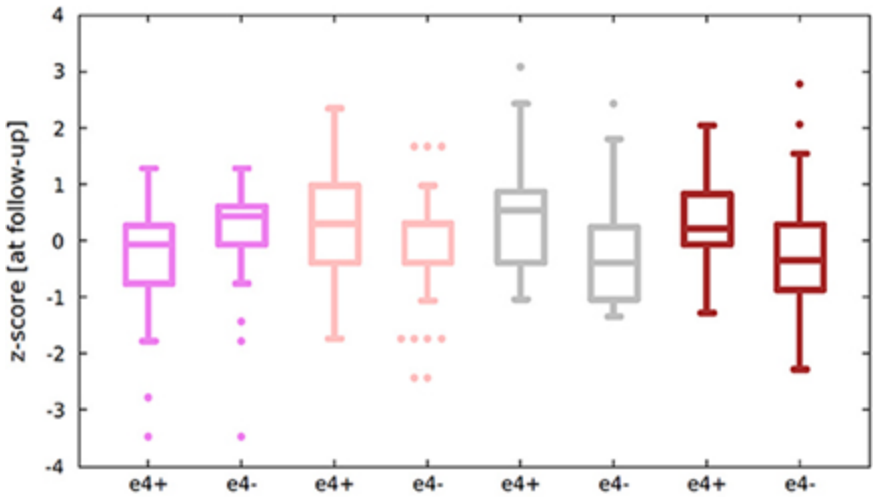


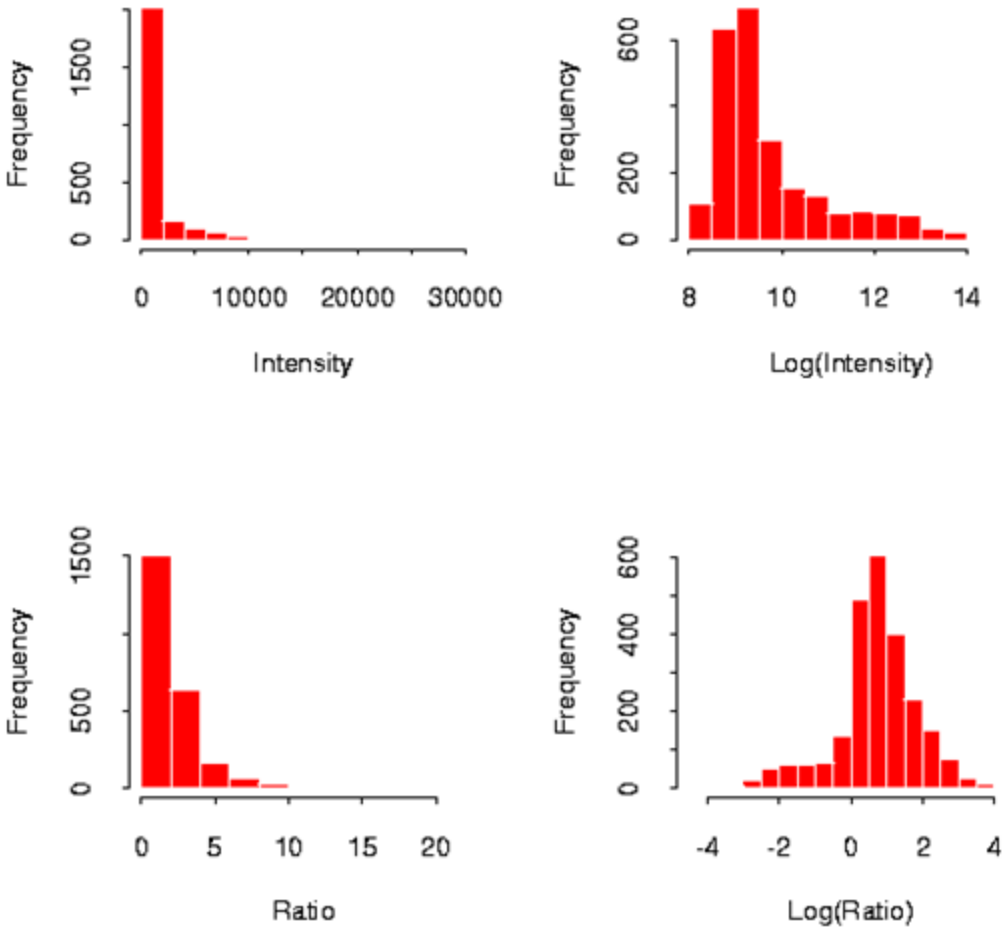
Normalization (II)

But ….. normalization often makes data difficult to interpret

Try simple methods first

- Most physiological measures are log-normal
- Auto-scale (unit transformation, or Z-score)





Normalization (III)

Many methods are available

✓Centering

✓Scaling

✓Transformation

There is NO guarantee of global normal distribution in omics data

Method	Formula	Unit	Goal	Advantages	Disadvantages
Centering	$\tilde{x}_{ij} = x_{ij} - \bar{x}_i$	0	Focus on the differences and not the similarities in the data	Remove the offset from the data	When data is heteroscedastic, the effect of this pretreatment method is not always sufficient
Autoscaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$	(-)	Compare metabolites based on correlations	All metabolites become equally important	Inflation of the measurement errors
Range scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\left(x_{i_{\max}} - x_{i_{\min}}\right)}$	(-)	Compare metabolites relative to the biological response range	All metabolites become equally important. Scaling is related to biology	Inflation of the measurement errors and sensitive to outliers
Pareto scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$	0	Reduce the relative importance of large values, but keep data structure partially intact	Stays closer to the original measurement than autoscaling	Sensitive to large fold changes
Vast scaling	$\tilde{x}_{ij} = \frac{\left(x_{ij} - \bar{x}_i\right)}{s_i} \cdot \frac{\bar{x}_i}{s_i}$	(-)	Focus on the metabolites that show small fluctuations	Aims for robustness, can use prior group knowledge	Not suited for large induced variation without group structure
Level scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\bar{x}_i}$	(-)	Focus on relative response	Suited for identification of e.g. biomarkers	Inflation of the measurement errors
Log transformation	$\tilde{x}_{ij} = {}^{10}\log\left(x_{ij}\right)$ $\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	Log 0	Correct for heteroscedasticity, pseudo scaling. Make multiplicative models additive	Reduce heteroscedasticity, multiplicative effects become additive	Difficulties with values with large relative standard deviation and zeros
Power transformation	$\tilde{x}_{ij} = \sqrt{\left(x_{ij}\right)}$ $\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	√0	Correct for heteroscedasticity, pseudo scaling	Reduce heteroscedasticity, no problems with small values	Choice for square root is arbitrary.

BMC Genomics. 2006; 7: 142

STATISTICAL ANALYSIS

-- identify significant features & patterns

Objective

- Data are 'cleaned' and ready to analyze
  - No outliers
  - No missing values
  - Filtered out low quality or uninformative values
  - Normalized/transformed

- We want to identify features that are interesting in our research context
  - Metabolites with different abundance between "Control" and "Treated" samples

Univariate analysis

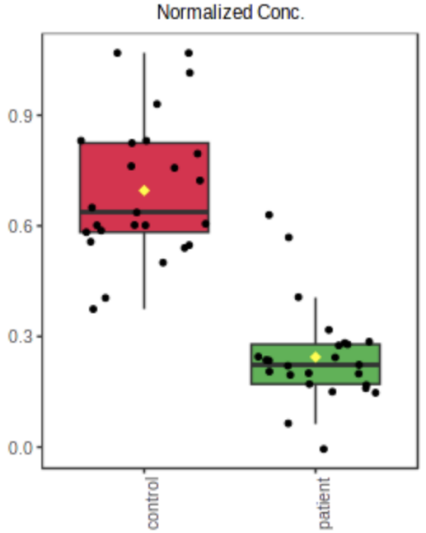
Test each feature independently (ignore their relationships to each other)

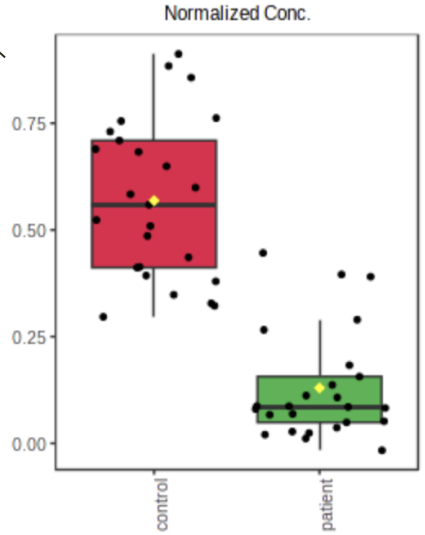
1. **T-tests**
   - **Compare the means between 2 conditions**
2. **ANOVA & post-hoc analysis**
   - **One factor with more than 2 levels (One-way ANOVA)**
   - **Two factors (Two-way ANOVA)**
3. Linear modeling (i.e., limma): more flexible analysis
   - Multiple factors
   - Time series
   - Covariates analysis

All these approaches are now available in MetaboAnalyst 5.0

Example Results

Name	t.stat	p.value	-log10(p)	FDR
Bin.2.54	9.7236	6.2898E-13	12.201	7.1299E-11
Bin.2.70	9.6702	7.5052E-13	12.125	7.1299E-11
Bin.0.82	-9.4206	1.7221E-12	11.764	1.0907E-10
Bin.0.94	-8.7213	1.8299E-11	10.738	8.6921E-10
Bin.2.66	7.4497	1.5087E-9	8.8214	5.1615E-8
Bin.0.78	-7.4277	1.6299E-9	8.7878	5.1615E-8
Bin.2.58	7.1607	4.1717E-9	8.3797	1.1323E-7
Bin.0.98	-6.592	3.1069E-8	7.5077	7.3788E-7
Bin.8.30	-6.3879	6.3908E-8	7.1944	1.2272E-6
Bin.1.02	-6.3848	6.4591E-8	7.1898	1.2272E-6
Bin.0.86	-6.1211	1.6376E-7	6.7858	2.8286E-6
Bin.1.70	-6.0919	1.8144E-7	6.7413	2.8729E-6
Bin.8.66	-5.9857	2.6369E-7	6.5789	3.854E-6
Bin.8.14	-5.8695	3.9655E-7	6.4017	5.3818E-6





P-value & multiple testing issue

1. P-value = probability of observed difference between groups if there is truly no effect (a.k.a the null hypothesis)
2. One "rejects the null hypothesis" when the p-value is less than the significance level α which is often 0.05 or 0.01
3. When the null hypothesis is rejected, the result is said to be statistically significant

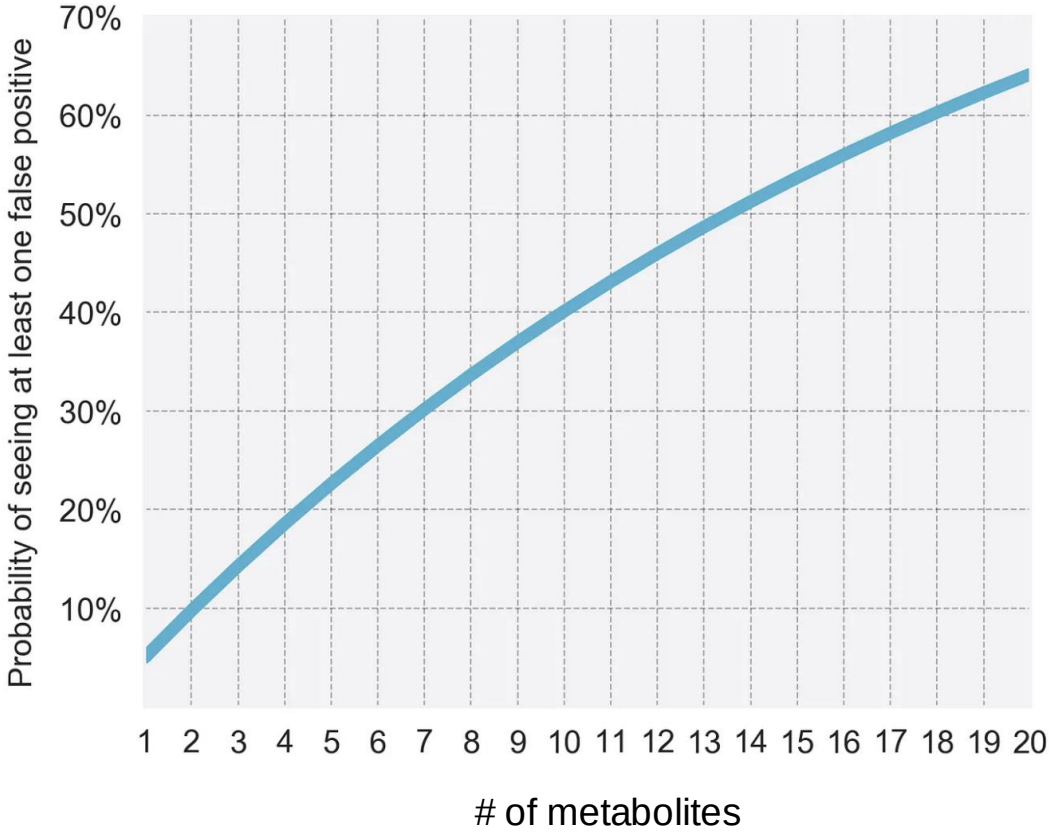
Performing T-tests on typical metabolomic data might result in performing ~10000 separate hypothesis tests. If we use a standard p value cut-off of 0.05, we would see 500 (10000*0.05) features to be deemed “significant” by chance!

Figure adapted from:

The third Ghost of Experimentation: Multiple comparisons

Skyscanner Engineering · Follow
Published in Towards Data Science · 6 min read · May 29, 2018

Adjusted p-values



Bonferroni (FWER)

- $\alpha = \alpha/n$
- Probability of ≥ 1 false positive = α
- Extremely strict

Benjamini-Hochberg (FDR)

- False discovery rate (ie. 0.05)
- False sig. metabs / Total sig. metabs
- Adjusted p-value or "q-value"

Principal Component Analysis (PCA)

Project high-dimensional data into lower dimensions that capture the most variance of the data

Assumption:

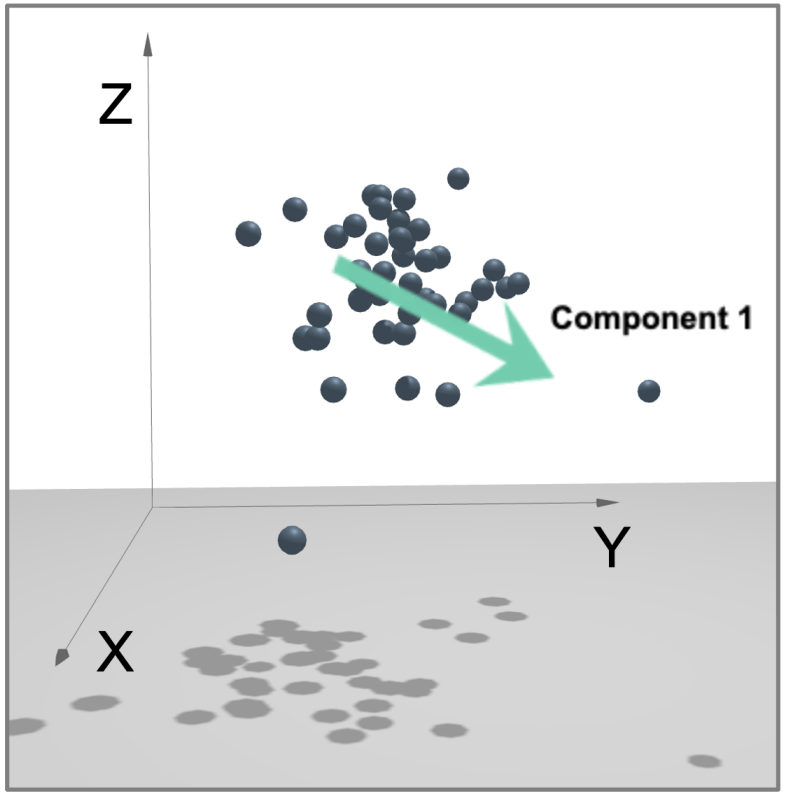
Main directions of variance

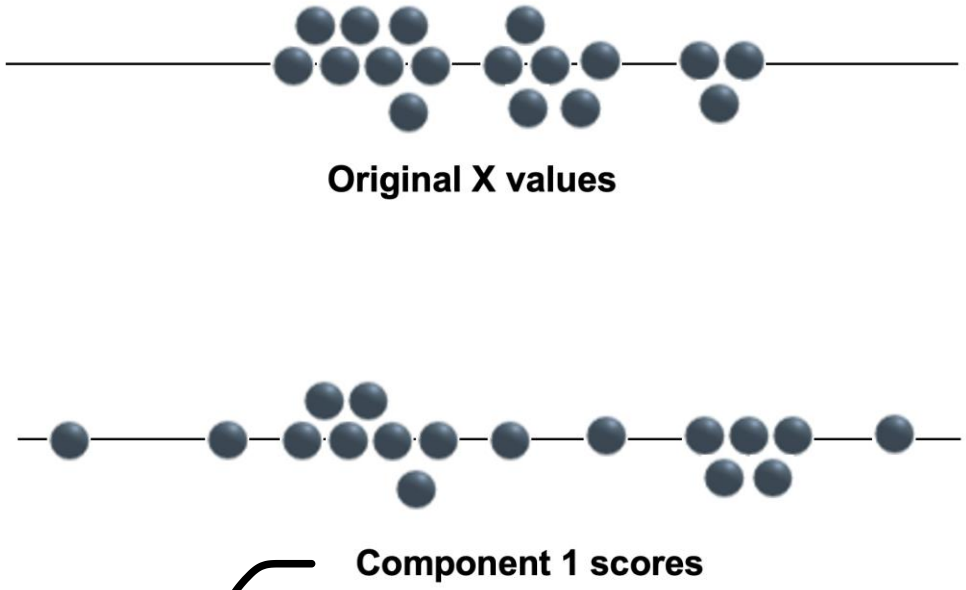
≈ major data characteristics

PCA Scores and Loadings

Component 1 Scores = Loadings x data

$PC1_1 = a*x_1 + b*y_1 + c*z_1$
$PC1_2 = a*x_2 + b*y_2 + c*z_2$





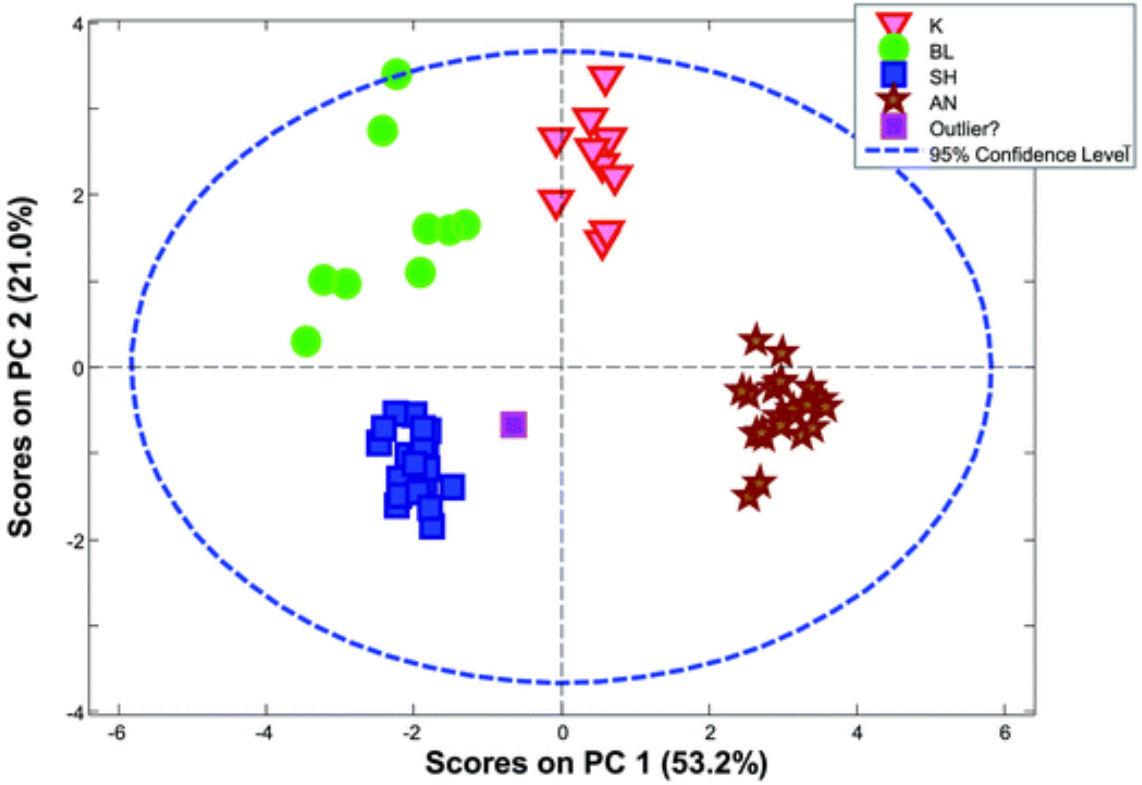
Integrates information from three variables – x, y, and z

Intuitive interpretation

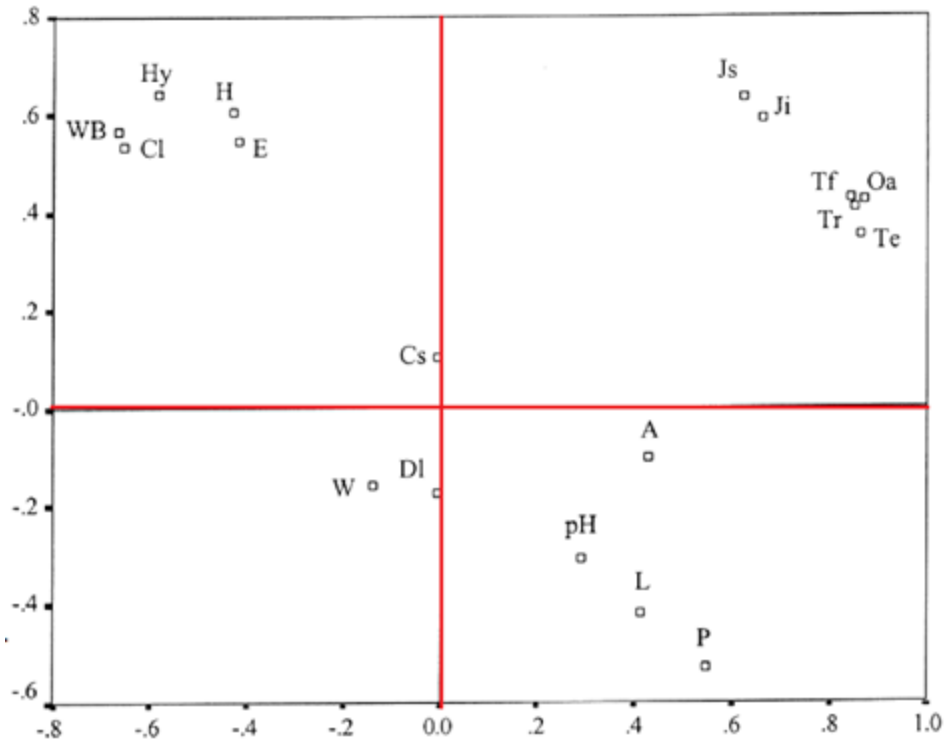
Scores = Loadings x data

$t_1 = p_1x_1 + p_2x_2 + p_3x_3 + \ldots + p_nx_n$

Scores



Loadings



Sample patterns (scores) are directly related to feature patterns (loadings)

PCA summary

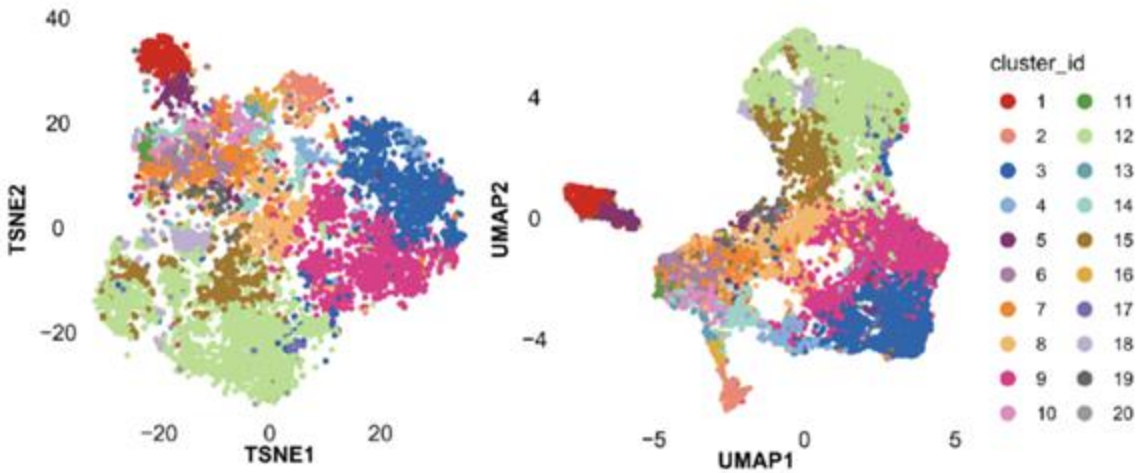
PCA integrates information from many variables into a few variables. It is widely used for:

- Data overview
- Outlier detection
- Find out relationships between variables

PCA is a linear method for dimension reduction.

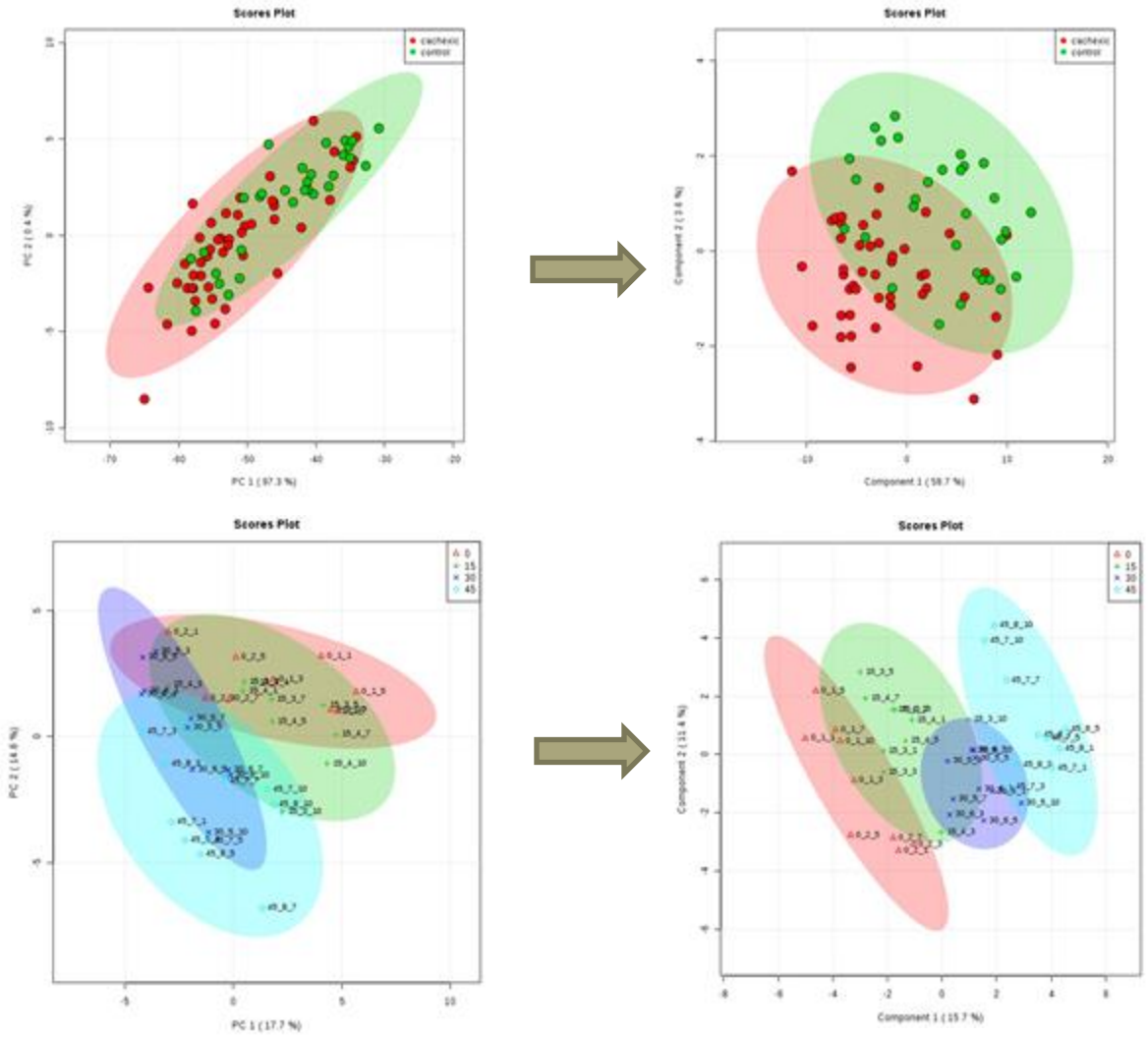
There are non-linear methods

- t-SNE, UMAP, etc



From unsupervised to supervised classification

PCA vs.
PLS-DA



PCA vs. PLS-DA

Scores = Loadings x data

$t_1 = p_1x_1 + p_2x_2 + p_3x_3 + \ldots + p_nx_n$

- In PCA we found loadings that computed scores that **maximized variance within the data**
  - Explain the main trends without considering metadata
- In PLS-DA, we find loadings that compute scores that **maximize variance between class groups**
  - Explain the main trends that separate the metadata

Caution! PLS-DA always produces some separation and is prone to overfitting

PLS-DA performance measures

PLS-DA is susceptible to over-fitting, and require more rigorous validation

1. **Cross validation** – whether the model can predict on new events
   - Sum of squares captured by the model (R^2)
   - Cross-validated R^2 (also known as Q^2)
   - Prediction accuracy
2. **Permutation tests** – whether the model captures real signals compared to null

Cross validations (CV)

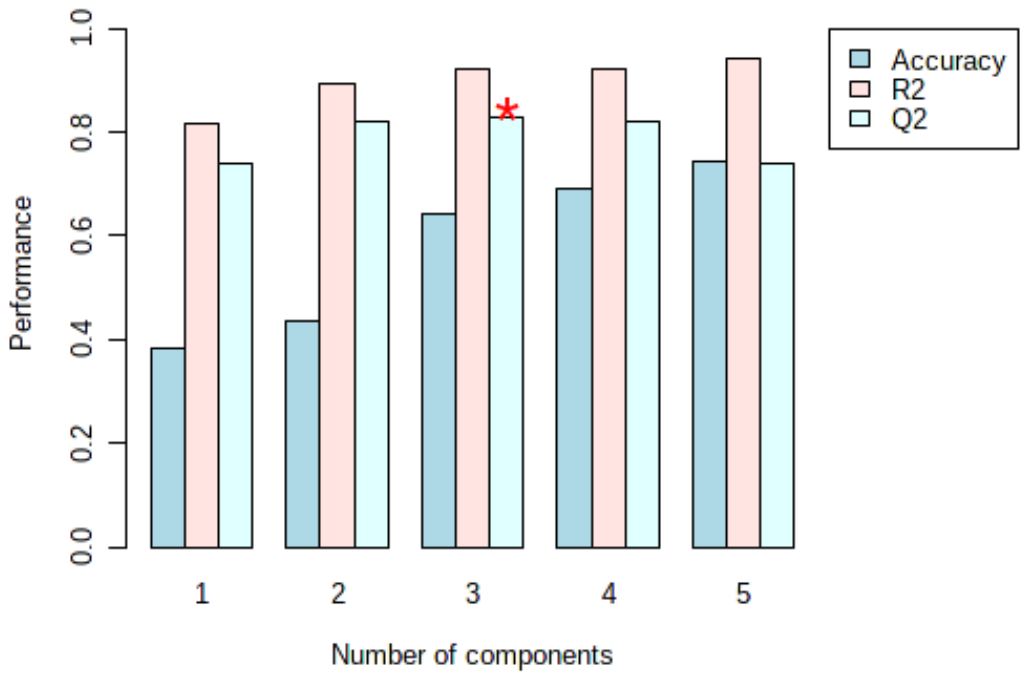
- Goal: test whether your model can predict class labels for new samples

PLS-DA (R^2 & Q^2)

Q2 is calculated via cross-validation to compute Predicted Residual Sum of Squares (PRESS).

For convenience, the PRESS is divided by the initial sum of squares and subtracted from 1 to resemble the scale of the R2.

Good predictions will have low PRESS or high Q2. Low or even **negative Q2** means that your model is not at all predictive or is overfitted.



Hands-On Demo