lipotype
Lipidomics for a better life

WHITE PAPER SERIES

White Paper: Big Data Lipidomics

Lipidomics is the large-scale study of lipids in biological systems. The analysis of large datasets, potentially containing up to thousands of lipidomes, is a challenging endeavour. We have established multiparametric statistical approaches, tailored to quantify lipid data. These methods are geared to identify lipid biomarkers. In this white paper a cohort of healthy subjects is compared with a cohort of diseased persons to identify lipid signatures that discriminate health from disease. Such signatures could potentially be useful for disease stratification or for diagnosis by means of predictive modelling (machine learning). In this white paper, we will guide you through the data analysis process aiming at the identification of lipid biomarkers and the evaluations of their performance.

Introduction

At Lipotype our expertise are lipids, lipid metabolism, lipid chemistry and lipidomics. Therefore, statistical analysis is always done with this knowledge as a basis. We treat lipids not as isolated entities but as a metabolic network with substructure and with their functions in mind. At Lipotype we are using the R programming language[1] for statistical computing with its rich environment of methods and plotting capabilities.

When working with lipidomic datasets consisting of hundreds of parameters (the lipids) in thousands of samples, a major challenge is to extract the relevant information. In the context of a biomarker identification study in which a cohort of healthy subjects is compared with a cohort of diseased subjects, the relevant information would be the lipids that discriminate health from disease. This information could then potentially be of use for disease stratification or diagnosis. In this white paper, we will guide you through the data analysis process aiming at the identification of biomarkers and the evaluation of their performance. The example is based on the analysis of 140 simulated human plasma lipidomes as acquired by Lipotype's mass spectrometry-based Shotgun Lipidomics technology.

[1] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Data preparation

Lipid amounts are generally provided in **moles (pmol)**, but often we prefer to standardize to the total amount of lipids within a sample, which we call **mol percentage (molp)**. This usually increases the robustness of the subsequent analysis.

When screening large high-throughput datasets, which are acquired in the course of several days or even weeks, we typically include reference samples in order to be able to assess analytical performance. Based on these reference samples, datasets are tested for measurement artifacts like **drift and batch effects**, and corrected if required.

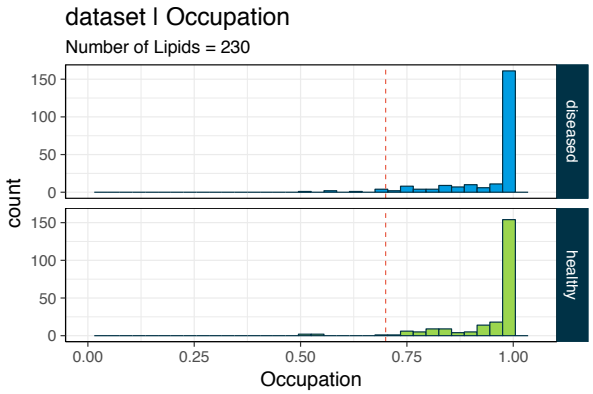

***Figure 1: Occupational Threshold**: Only lipids, which are present in a minimal fraction (here 70% of samples) are used for further analysis. In the present example, a total of 230 lipids pass the threshold.*

Additionally, we apply **occupational thresholds**, to only work with lipids that are present in a sufficiently large fraction, e.g. 70% of the study samples (Figure 1). Occupational thresholds may also be applied based on study groups.

We join all provided clinical and anthropometric data with the lipidomic dataset and check for overall consistency. To that end further measures can be applied, such as outlier filtering. Additionally, propensity matching can be performed in order to make sure that study cohorts are matching based on clinical and anthropometric parameters.

Data Analysis

Initially, we try to achieve a bird's-eye view of the dataset using **dimensionality reduction and clustering**.

With ***PCA*** (principal component analysis) we investigate the major variation within the sample set and look for segregation of cohorts (or experimental groups in general) within the principal components. If a segregation is found for subjects in the scores plot (Figure 2A), an analysis of the lipids or features in the PCA loadings (Figure 2B) can already be a rich source of information of the reasons for the segregation. In the present example, a segregation of the study cohorts in dimension 2 (PC2) can be observed, indicating potential differences in the lipid composition of the two cohorts. Note, however, that dimension 1 (PC1), which is the dimension with the largest variability in the dataset, does not allow for a separation of the two cohorts. This indicates that there is substantial variation in the dataset that is independent of the cohort assignments, or, in other words, does not reflect differences between healthy and diseased subjects. This is a typical situation for human plasma lipidomic datasets, which comprise significant inter-individual variability. To be able to cope with this intrinsic characteristic of human plasma, a sufficient (typically >100, depending on the effect size) number of samples needs to be provided in order to obtain satisfying statistical power.

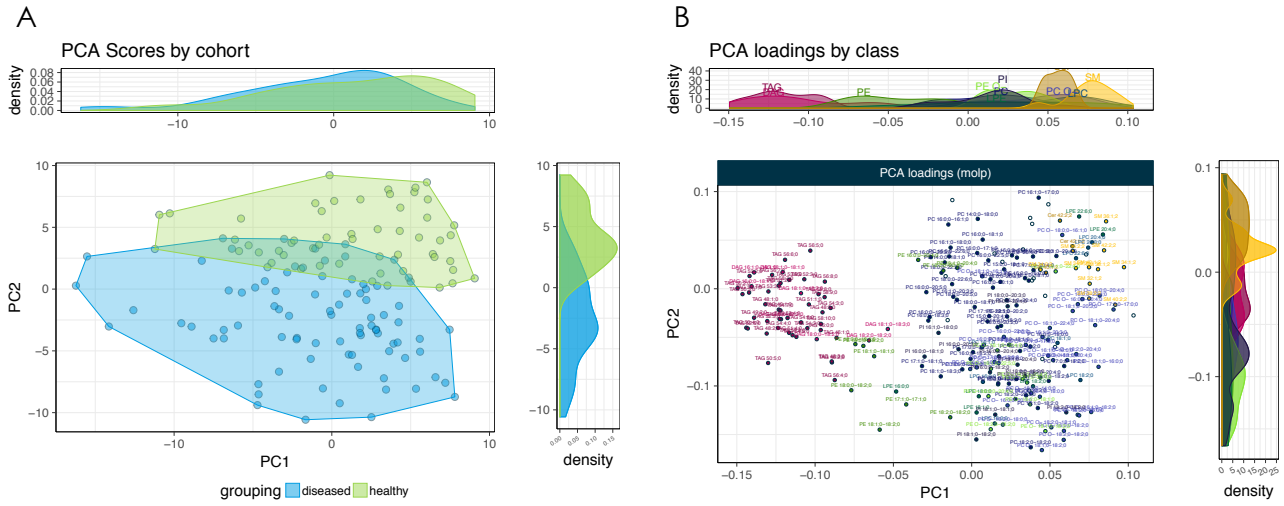

Figure 2: PCA Analysis *of a dataset. In the scores plot (A) a segregation of the samples can be observed in principle component 2. In the loading plot (B) individual lipid (sub-)species are show as the basis of the segregation.*

To identify the lipids that differ between cohorts (healthy vs diseased) we apply statistical hypothesis tests: the parametric **t-test** and its non-parametric alternative **Wilcoxon rank-sum test**, in their paired and unpaired variants. Covariates (such as age, gender or drug use) can be incorporated on the basis of linear models. *p*-values are adjusted for multiple comparisons. A volcano plot shown in Figure 3A readily visualizes the result. Lipids that are significantly different are highlighted (the "hits"). The fold-change (x axis) provides an idea about the effect size. Box plots (Figure 3B) help visualizing the distribution of the individual data points of the cohorts.

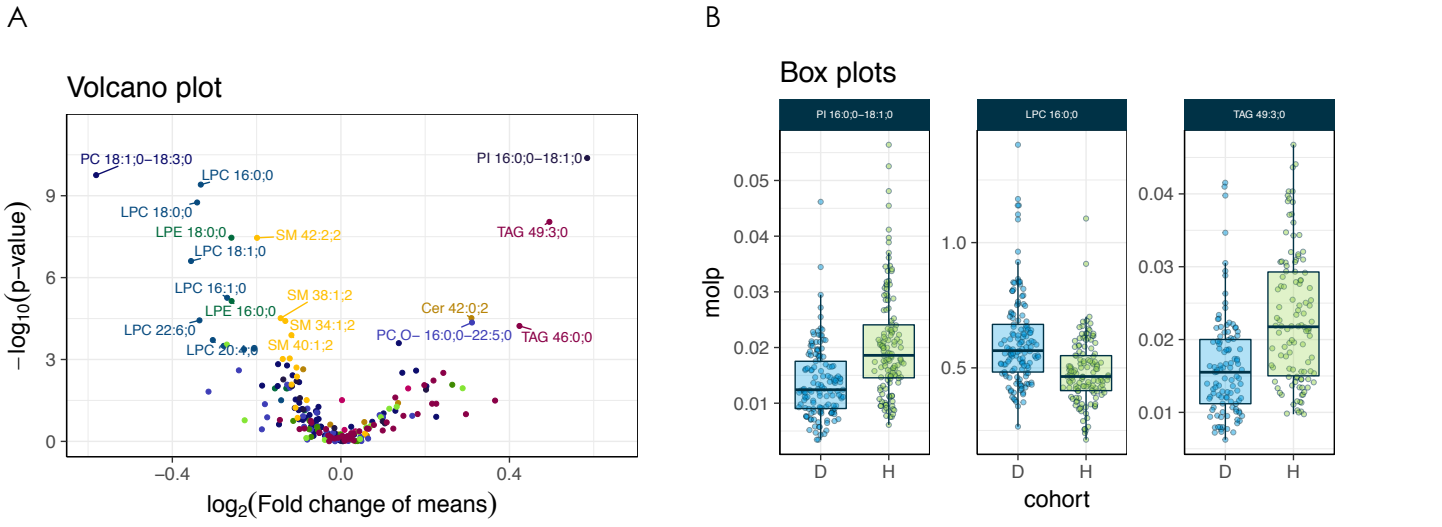

Figure 3: (A) Volcano plot *of lipid (sub-)species differences between two cohorts. Significant differences according to Wilcoxon rank-sum test and p-value adjustment (Bonferroni) are marked with their (sub-)species name.* ***(B) Box plot*** *of selected features from the Volcano plot.*

Enrichment Analysis / Pathway Analysis

Outputs of statistical analysis, as described above, are usually lists of individual lipid species. Enrichment analysis based on hypergeometric distribution or ranks, can provide **help with interpretation of these lists** by suggesting more general categories enriched within the results. Categories range from lipid classes, fatty acid composition, saturation profiles to pathway analysis and incorporate the full lipid knowledge of Lipotype. In the present example, we observe an enrichment of lysophosphatidylcholines (LPC) and lysophosphatidylethanolamines (LPE) among the hits (Figure 4), suggesting alterations in the activities of the respective phospholipid hydrolase as a cause for the observed differences.

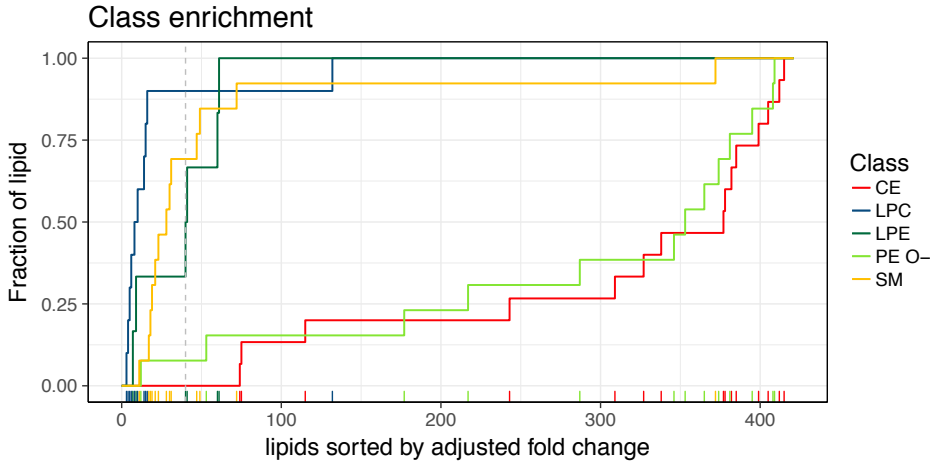

Figure 4: Enrichment analysis *of lipid classes over the data sorted according to the volcano plot above (Figure 3A). This is an example of a generalization from individual features to features that have more explanatory value.*

Predictive Modelling

With Predictive modelling we use a supervised algorithm to **predict class membership of future samples**. In the present example, the goal is to predict if a subject is healthy or diseased based on the plasma lipidome. It can also be applied to **estimate the performance** of a lipid signature to distinguish between study cohorts. For training a classification algorithm, we use 5× 10-fold cross validation on 80% of the data, while the remaining 20% are used as a hold-out test set. Data were centered, scaled and transformed to normal distribution. Missing values were median-imputed. All data pre-processing steps are performed within the cross-validation loop.

Classification

We trained several models (see Figure 5), of which the partial least squares discriminant analysis (pls) shows a good performance with an average cross validation classification accuracy of 95%. Cross validation sensitivity and specificity on the test set are 97% and 94%, respectively. Thus, 97% of the patients can be correctly identified as diseased based on the identified lipidomic signature. The most important predictors used for classification are shown in Figure 5B.

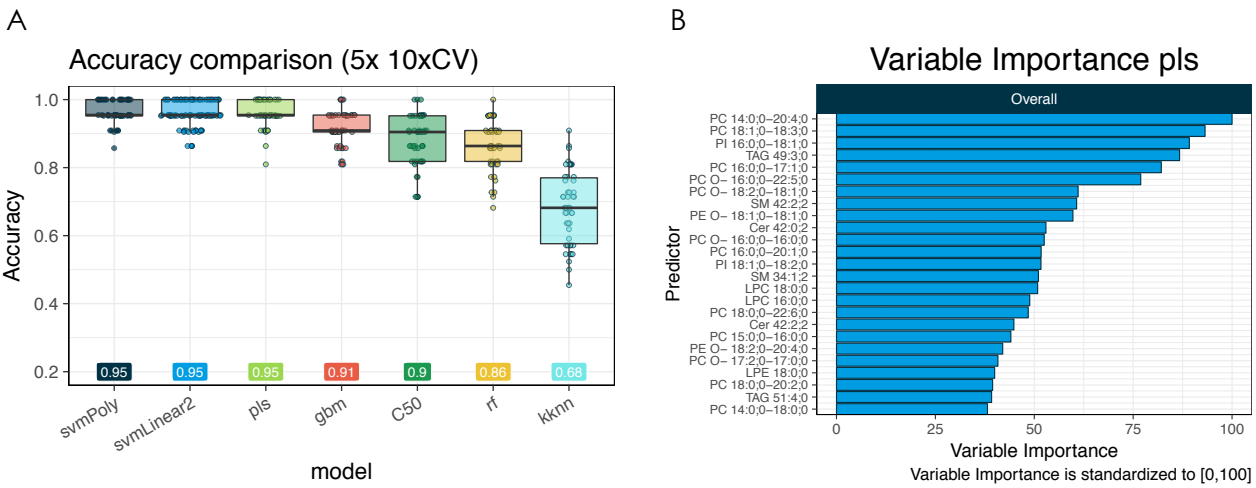

Figure 5: Example of Classification outputs: (A) *Overall performance of different models are compared.* ***(B)*** *Lipid (sub-)species of the partial least squares discriminant analysis (pls) are ranked according to their importance for the prediction.*

Continuous outcome variables

The examples above were dealing with categorical data, *i.e.* healthy vs. diseased. However, one might also want to relate lipidomics data to continuous variables such as BMI or blood glucose level. In that case, a correlation analysis or regression models would be appropriate methods.

Correlation analysis

Correlation analysis (Figure 6) is used to study the strength of the relationship between the amounts of individual lipids and a continuous clinical or anthropometric variable (e.g. blood glucose level). Significance estimates and covariates can be incorporated on the basis of linear models. In the present example, we could observe strong correlations between the amounts of individual lipids and blood glucose levels (Figure 6). This result suggests a complex interplay of glucose and lipid metabolism and might reveal novel therapeutic targets for the treatment of aberrant blood glucose levels or related metabolic disorders.

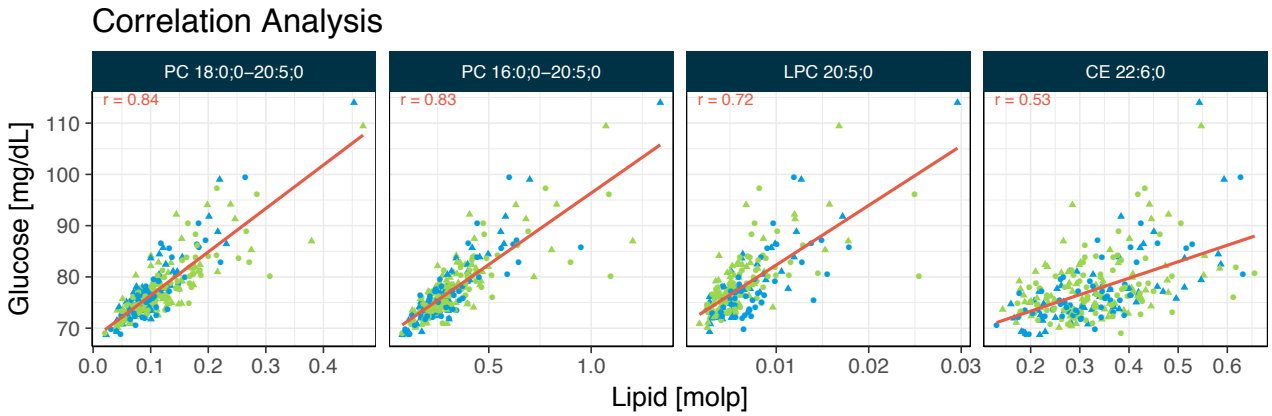

Figure 6: Correlations *of a continuous feature to individual lipid (sub-)species.*

Predictive Modelling

Instead of analyzing the correlation of individual lipid molecules with a continuous outcome variable, one can alternatively relate the entire lipidome with the outcome variable. In what is called a regression analysis, we typically train several models on root-mean-square error (RMSE) or coefficient of determination (R^2). *Models include*: Linear models, Partial Least Squares Regression (PLS-R), Lasso and Cubist.

What is delivered: EMPOWERING YOUR DATA ANALYSIS

Venturing into the analysis of large scale dataset is a challenging endeavor. Therefore, experts from Lipotype will interact with you closely starting from the first contact in order to understand your questions, needs and expectations. Based on the initial discussions, we will propose an analysis scheme that aims at getting the most out of your data. Upon agreement on an analysis goal, we will provide results in time-efficient and comprehendible manner. The customer will receive a report containing a summary the results and detailed descriptions of the statistical methods. Visualization of data is key to understanding results. Lipotype provides figures that can be used in presentations and data, which can be used in publications or reports. We will offer additional consultation after delivery of the report to make sure our customers get the most out of the results.

Applications for big data lipidomics

- Lipid biomarker identification and validation (pharmacodynamic, pharmacokinetic, CDx) for biotech and pharma industry as well as (pre-)clinical research
- Lipid biomarker identification and validation for clinical diagnostics (prognostic, diagnostic, patient stratification)
- Identification of novel, lipid-related therapeutic targets and mode-of-action studies in drug discovery phases
- Analysis of animal studies
- Analysis of clinical datasets in general
- Intervention studies for development of functional food/nutraceuticals
- Cosmetic claim support for (active) ingredients for cosmetic industry
- For cohort studies with limited sample size, we recommend LipotypeZoom as a quick and cost effective tool to interactively analyse your data.

References

Ciucci S. et al. "Enlightening discriminative network functional modules behind Principal Component Analysis separation in differential-omic science studies." Scientific Reports 2017, http://dx.doi.org/10.1038/srep43946

Frahnow T. et al. "Heritability and responses to high fat diet plasma lipidomics in a twin study" Scientific Reports 2017, http://dx.doi.org/10.1038/s41598-017-03965-6

Disclaimer: All data and results shown here are only for illustrative purposes. All data sets have been simulated and do not represent real values measured on actual samples.

Contact:

Dr. Oliver Uecke
T: +49 (0) 351 79653-45
sales@lipotype.com

Lipotype GmbH
Tatzberg 47, 01307 Dresden, Germany
www.lipotype.com