Using metabolomics and robust computational methods to predict donor traits in forensics and other fields of research

Progress Report by Johan Lassen

Supervisor: Palle Villesen
Co-supervisor: Kirstine L. Nielsen

Table of Contents

Introduction

Metabolomics is the study of small molecules used in a variety of fields such as clinical, environmental, food, and pharmacological sciences. The small molecules reflect the current state of a sample and is useful when describing phenotypes uncorrelated to traditional omics such as genomics and transcriptomics.

The metabolomics data are acquired through mass spectrometry-based methods (MS) or nuclear magnetic resonance (NMR). NMR detects several hundred compounds, and mass spectrometry methods may detect tens of thousands of molecules and are thus highly used for metabolomics. The mass spectrometry methods are coupled to gas chromatography (GC) or liquid chromatography (LC). In practice the methods are abbreviated GCMS and LCMS, of which the latter may be preferred for metabolomics due to its high sensitivity when detecting molecules.

LCMS methods include targeted and untargeted analyses: Targeted methods measure specific compounds of interest, while untargeted methods measure everything detectable. In practice targeted methods yields high quality data of known compounds, in contrast to untargeted which yield thousands of features that only represent *potential* molecules, due to a risk of experimental artefacts.

The untargeted methods are often deployed to narrow down potential biomarkers, followed by identification and validation of the potential compounds by targeted methods. This is because the targeted LCMS method ignores a great fraction of the global pool of metabolites, thus biasing the results. As we cannot be sure that a potential compound of untargeted data represents a true compound, the identification of the biomarkers relies on an extensive statistical pipeline of transformations and denoising.

The denoising of the data is critical to obtain reliable results from the untargeted data, and current computational methods provide improvements to data quality. However, several processing steps are under constant development because they perform inconsistently or might be further improved. These steps include peak calling, batch normalization and biomarker identification by multivariate modeling (figure 1).

Essential processing steps that ensure high data quality

The raw data from the untargeted methods are 3-dimensional chromatograms: Compound intensities measured at given retention times (rt) and mass to charge ratios (m/z). The retention times and the mass charge ratios represent physiochemical properties of the compounds, making it possible to distinguish the thousands of chromatogram peaks from each other.

The peak calling deconvolves the peaks into a feature table, where each feature represents the same peak between all samples and that the intensity (value) represents the area under the curve. XCMS is a popular tool for peak calling, but other tools exist including MZmine, OpenMS, and SLAW. The first step of XCMS is sample-wise peak identification for which the parameters *min/max peak width*, *ppm*, and *signal to noise ratio* must fit the data. Following this peak grouping, peak alignment and peak integration is carried out. All the steps require specific parameter settings to work properly and yield quality data.

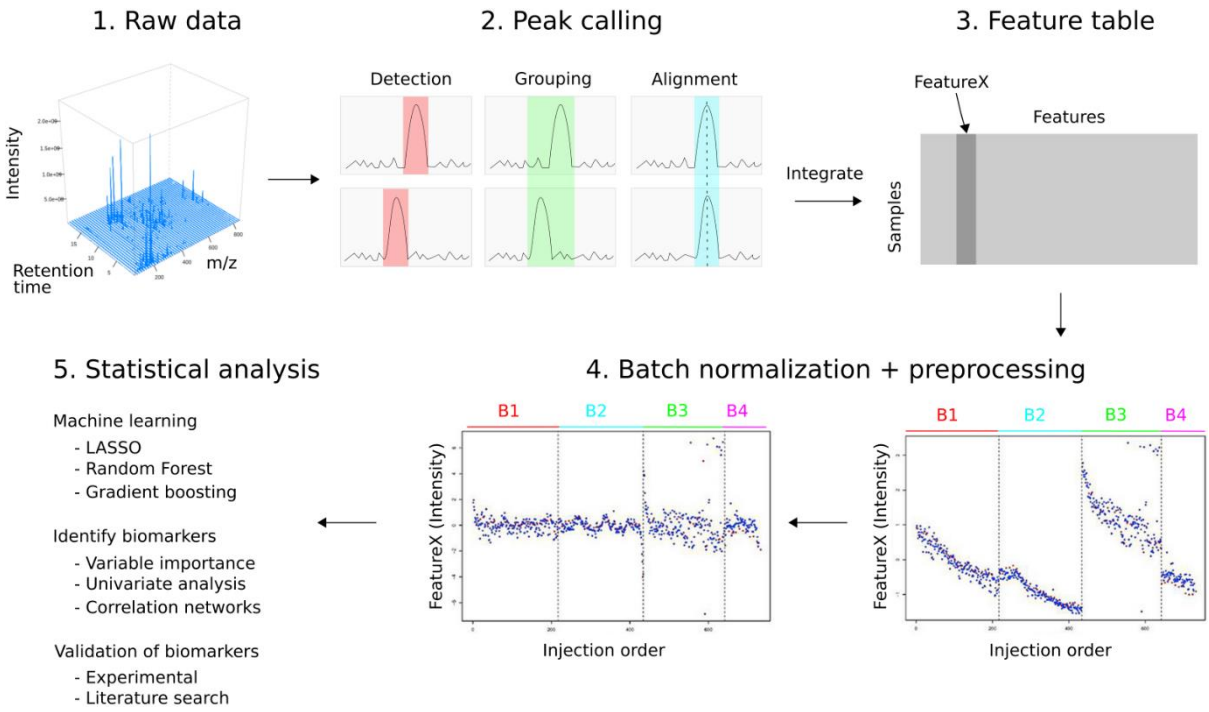

Figure 1 | Workflow: From chromatogram to biomarker. (1) Raw LCMS 3D data consisting of peaks with intensities at given retention times and mass charge ratios. (2) The peaks are called by XCMS to quantify molecule intensities and make molecules comparable between samples (alignment). (3) The peak calling returns a feature table feature (potential molecules) intensities for each sample. (4) Batch normalization and statistical processing remove variance depending on batch and injection order. (5) Analyze the cleaned data using robust multivariate methods. Step 4 figure is from
https://www.sciencedirect.com/science/article/abs/pii/S0003267019301849?via%3Dihub

The feature table undergoes normalization to adjust for potential differences in e.g., sample quality and batches. For instance, sample quality might contribute to more variance than biological causation. This is a simple example that may be adjusted by row normalization, that scales all samples to the same intensity. Trends caused by batch effects, date of sampling, maintenance cycles of LC-MS equipment, sample quality, and small deviances in standard operating procedures, might have non-linear relationships that are harder to correct for – several tools attempt to adjust this, including WaveICA, NormAE, and Combat.

After normalization multivariate statistical analysis reveal any molecular patterns correlated to the outcome variable. This requires robust methods as the features represent a mix of true signals and experimental artifacts; some features are false positives, due to poor peak identification; some features are isotopes or fragments of mother ions contributing to notable covariance; and some features may exist due to carry-over between samples when running the LC-MS.

Robust machine learning methods are especially suitable to meet the noisy requirements, as they penalize (e.g., LASSO, Neural nets, gradient boosting, random forest) or factorize (PLS-DA, OPLS-DA) to disregard the redundant features. The robustness ensures that machine learning based feature selection contains the most predictive features i.e., the potential biomarkers. The further identification of the biomarkers is based on database matching (e.g., HMDB or KEGG) followed by

experimental validation, to ensure that the feature selected compounds are not experimental artefacts.

In conclusion the data quality in metabolomics is highly affected by the data acquisition (lab methods) and the statistical analysis. Often researchers only have domain knowledge in either analytical chemistry or data science. Resultingly, tools as XCMS online, OPLS-DA by SIMCA©, and MetaboAnalyst provide bioinformatics tools that empowers non-expert users to perform statistical analyses.

As the bioinformatics tools aim at the generalist, there is a gap between the gold-standard and the cutting-edge of newly developed methods. OPLS-DA (machine learning) might for instance be outperformed by Random Forest, and XCMS R-based peak calling with automatic parameter tuning might outperform XCMS-online. By use of the cutting-edge methods the denoising and analysis of untargeted data might improve and consequently yield more/better biomarkers for the targeted methods.

Aim of project

My aim is to use metabolomics to characterize individual traits such as age or diseases. The traits of interest are useful for forensics screenings of perpetrators and clinical screenings for early detection of diseases. To achieve reliable results, I will use and develop state of the art statistical methods for metabolomics data. This includes improving parts of the workflow described, but also using it in practice for biological minded studies (see *planned studies*). I hope to contribute with new methods and ideas for experts while using strong statistical practices in applied studies to push the common base of metabolomics data analysis.