UNIVERSITÄT LEIPZIG

# Quantitative Analysis of Proteomics Mass Spectrometry Data

**Sebastian Gibb and Korbinian Strimmer**
IMISE, University of Leipzig

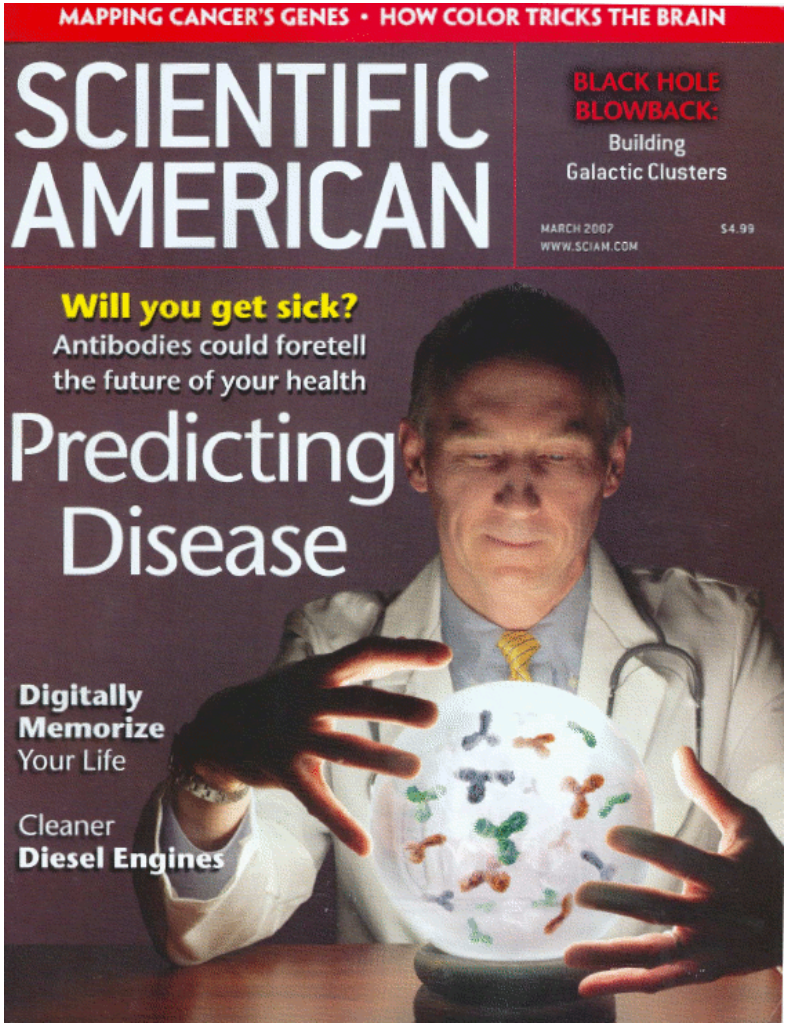
EMS 2013
Budapest
21 July 2013

# Inhalt


1. Introduction
   - Personalized Medicine
   - Why Proteomics?
   - Statistical Challenges
2. Preprocessing Mass Spectrometry Data
   - MALDIquant Software
   - Workflow
     - Data import, smoothing, baseline correction, calibration, peak detection, peak alignment, peak binning, intensity matrix
3. High-Level Analysis
   - Dichotomization
   - Classification and Ranking with Binary Predictors
   - Example from Cancer Proteomics
4. Conclusion and Outlook

# I. Introduction

# Predicting Disease



# Personalized Medicine

Medication and tests approved for personalized medicine in Germany:

**TABELLE**

**In Deutschland zugelassene Medikamente aus dem Bereich der personalisierten Medizin (Quelle: Verband der forschenden Arzneimittelhersteller)**

| Wirkstoffe | Krankheitsgebiet | Test auf | Testbeschreibung | Konsequenz aus dem Test | was wird getestet | Status | Quelle | Bemerkung |
|---|---|---|---|---|---|---|---|---|
| Abacavir | HIV/ Aids | Nebenwirkungen | Test auf Vorhandensein des HLA-B*5701-Allels (erhöht Risiko für Überempfindlichkeit) | keine Anwendung bei positivem Test | Blut | Pflichttest seit Feb 2008 | European Public Assessment Report (EPAR) | positives Testgergebnis bei ca. 5 % aller Patienten; bei 48-61 % dieser Patienten Überempfindlichkeitsreaktion; vor Testpflicht Hinweis auf mögliche schwere Nebenwirkungen |
| Anastrozol | Onkologie/bestimmte Formen von Brustkrebs | Wirksamkeit | Test auf Hormonrezeptor-positive Brustkrebszellen | Anwendung nur bei positivem Test | Krebszellen | Pflichttest seit Juni 1996 | Fachinformation | bei metastasierten Brustkrebs Zulassung auch ohne Vortest |
| Arsentrioxid | Onkologie/Akute Promyelozyten-Leukämie | Wirksamkeit | Test auf Vorhandensein des Promyelozytenleukämie/Retinsäurerezeptor-alpha (PML/RAR-alpha) Gens | Anwendung nur bei positivem Test | Gewebeprobe (Knochenmark) | Pflichttest seit März 2002 | European Public Assessment Report (EPAR) | keine |
| Azathioprin | Immunsuppressivum | Nebenwirkungen | Test auf Thiopurin-Methyltransferase (TPMT)-Mangel durch Gen- oder Enzymtest wegen möglicher extremer myelosuppressiver Wirkung | keine Anwendung bei positivem Test | Blut | von der Fachinformation empfohlener Test | Fachinformation | positives Testgergebnis bei ca. 0,3 % der Patienten, 10 % mit mittlerem Risiko; Test wird nur für den Fall empfohlen, dass sofort volle Dosis gegeben werden muss |
| Carbamazepin | Epilepsie | Nebenwirkungen | Test auf Vorhandensein des HLA-B*1502-Allels (erhöht Risiko für schwere Hautreaktionen) | keine Anwendung bei positivem Test | Blut | Empfohlener Test | Fachinformation | keine |
| Cetuximab | Onkologie / Darmkrebs | Wirksamkeit | Test auf nicht-mutiertes (Wildtyp) KRAS-Gen | Anwendung nur bei nicht-mutierter KRAS-Variante | Gewebeprobe | Pflichttest seit Juli 2008 | European Public Assessment Report (EPAR) | nicht-mutierte KRAS-Variante bei ca. 60 % der Patienten |
| Dasatinib | Onkologie / akute lymphatische Leukämie | Wirksamkeit | Test auf Philadelphia Chromosom; per FISH oder PCR (Polymerase Kettenreaktion) | Anwendung nur bei positivem Test | Blut | Pflichttest seit Nov 2006 | European Public Assessment Report (EPAR) | positives Testergebnis bei ca. 30 % der ALL-Patienten |

from: Nicola Siegmund-Schultze. 2011. Deutsches Ärzteblatt 108:1904-1909

# Biomarker Discovery

General aim of personalized medicine:

"the right drug for the right person"

Biomarkers are essential for:

- Classifying heterogeneous subtypes of disease
- Pinpoint precise diagnosis
- Targeted therapy
- Understanding the provenance of disease

# Data from High-Throughput Platforms

Major techologies to collect genomic high-througput data:
Microarrays and


Next-Generation Sequencing

# Proteomics

Microarrays and NGS allow to measure concentration of RNA expressions → Transcriptomics

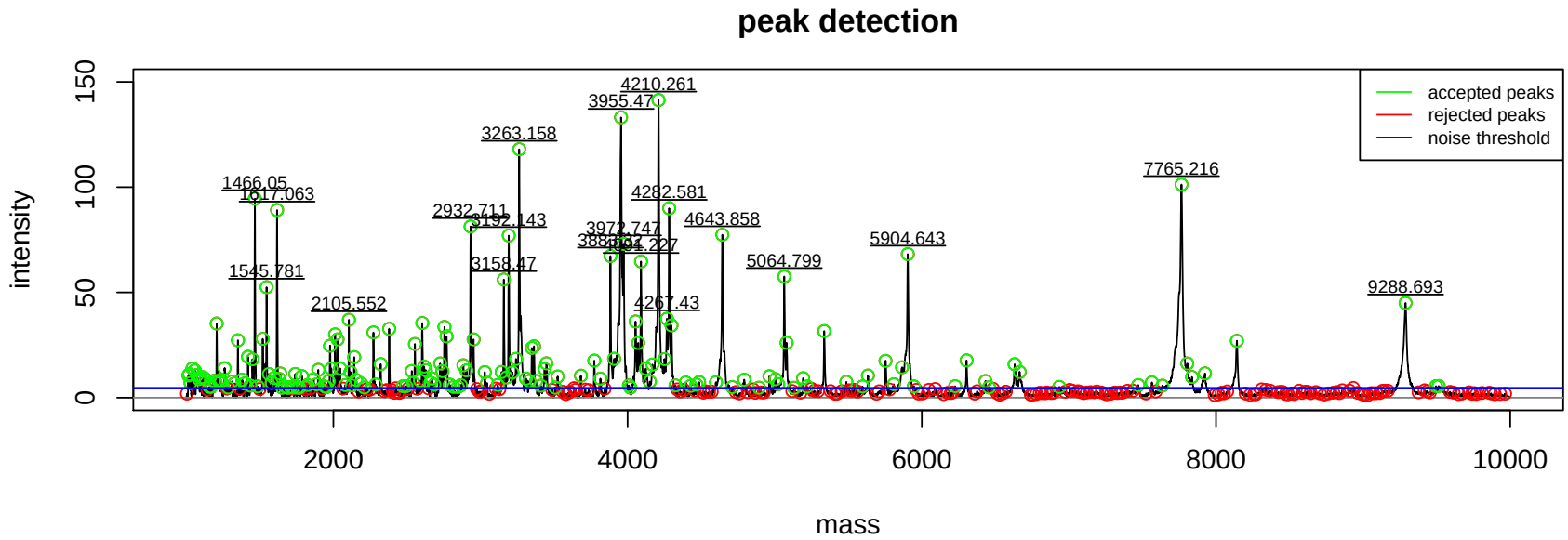
However, in many clinical settings (e.g. clinical diagnostics) one is often interested in expression of proteins, peptides and amino acids.

Advantage: direct biological and medical interpretation!

Systematic study of proteins and related products → Proteomics

# Proteomic Mass Spectrometry

High-throughput technology for measuring proteins:
Mass Spectrometry



# MALDI Technology


© Bruker Daltonics

MALDI: Matrix-Assisted Laser Desorption/Ionization

Allows the analysis of large organic molecules such as proteins.
(note “matrix” is biochemical jargon referring to carrier material that helps the protein ionization).

# MALDI-TOF Principle

| Ion Source: MALDI | Mass Analyzer: TOF | Detector |
|---|---|---|
| Matrix-Assisted Laser Desorption/Ionization | Time Of Flight ($t \propto \sqrt{\frac{m}{q}}$) | Quantity Measurement |

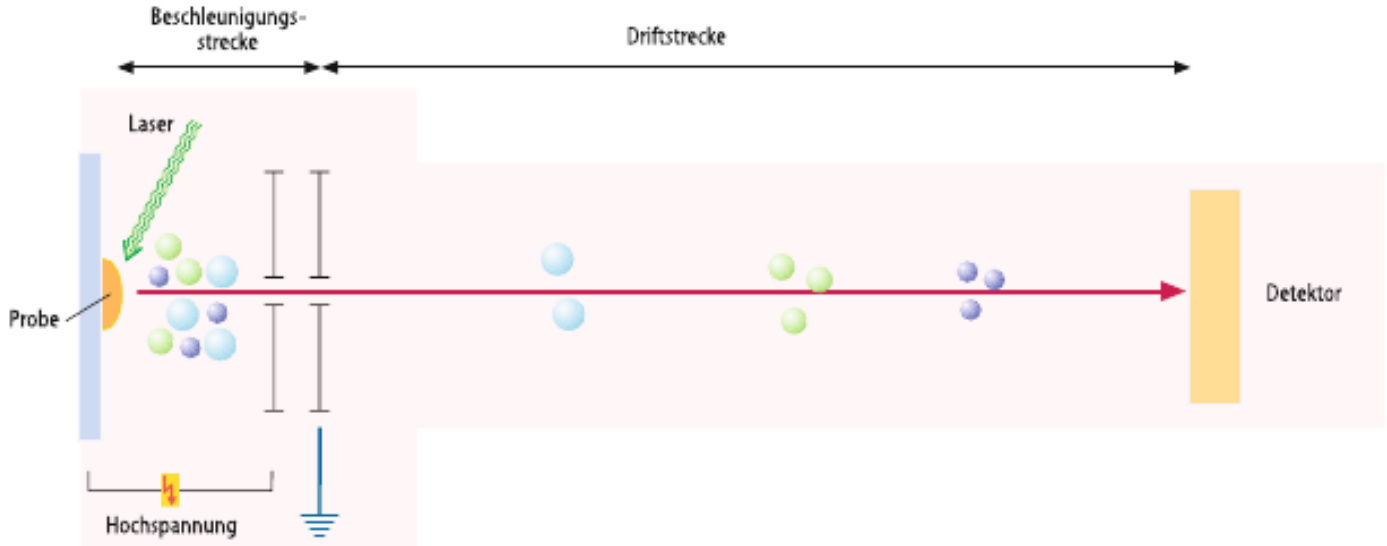



Abb. 3.14.; S. 67; "Biochemie & Pathobiochemie", Löffler G., 8. Auflage (2007), Springer Medizin Verlag

# MALDI Benefits

Advantages:

- Reliable well established technology (many variants exists)
- Often combined with other experimental techniques (gels, imaging etc).
- Very cheap to run compared with other high-throughput technologies (i.e. many replications possible)

# Statistical Challenges

Analysis of proteomics mass spectrometric data is more complicated than that of gene expression!

Overview of the major statistical challenges summarized by Morris et al (2007):

Preprocessing:

- Removal of systematic biases in the data
- Peak identification
- Peak alignment across spectra
- Quantification and calibration of relative peak intensities

Develop suitable methods for multivariate analysis:

- Differential protein expression
- Classification and prediction
- Feature selection

In addition, many considerations from transcriptomics also apply here (such high dimensionality, sparse models etc).

# II. Preprocessing Mass Spectrometry Data

# MALDIquant Software

In the form of `MALDIquant` we have developed a complete processing pipeline for proteomics data in R.

Motivation:

- Only relatively few open source software solutions available and very few for the R platform.
  http://strimmerlab.org/notes/mass-spectrometry.html
- No MALDI-TOF package fitting our needs for clinical diagnostics.
- Necessity of handling both technical and biological replicates.
- Nonlinear Peak alignment necessary for many spectra.
- Modular and easy to customize analysis routines.
- Testbed for studying new methodologies (e.g. for quantification).

# MALDIquant Family

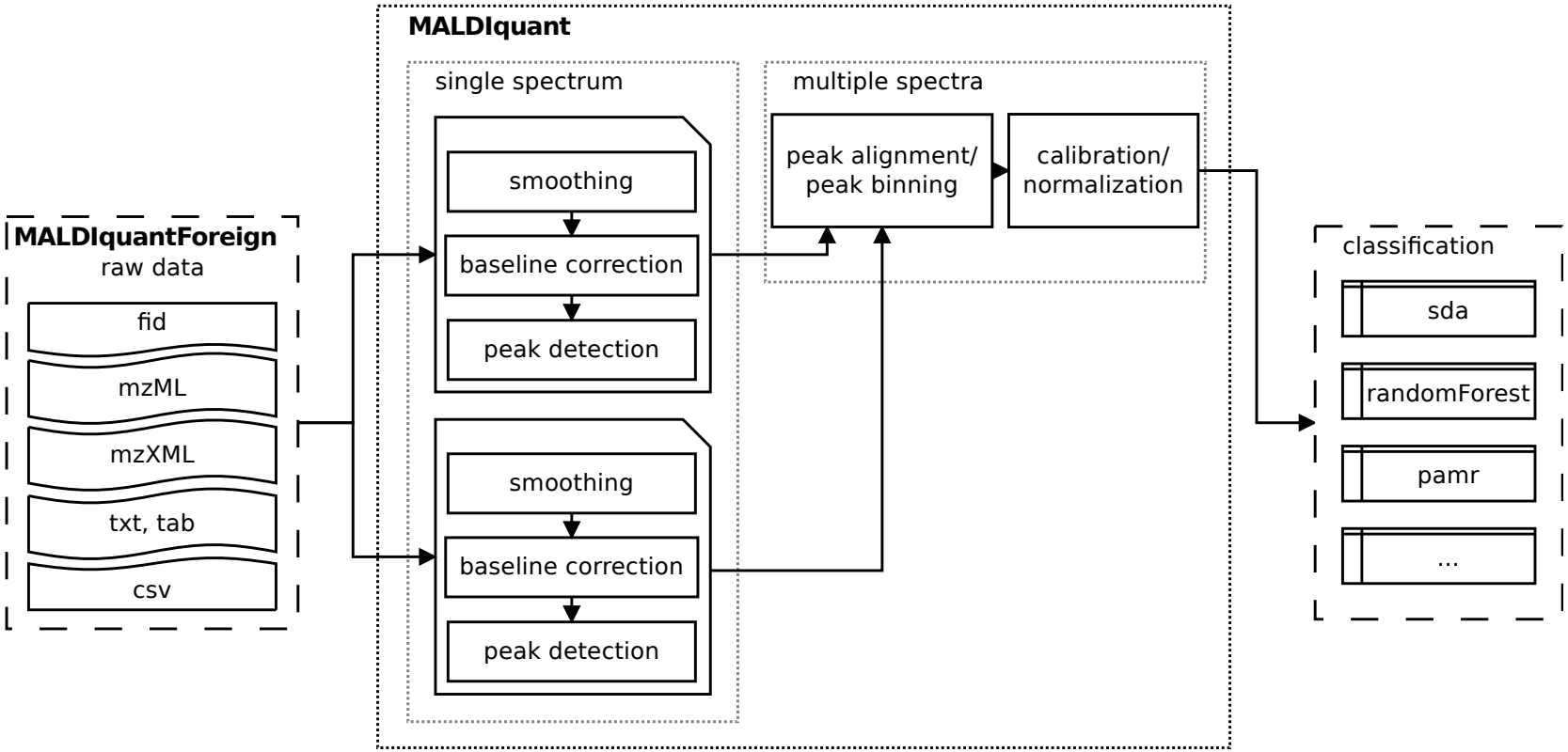
## MALDIquant

Complete analysis pipeline for MALDI-TOF and other 2D mass spectrometry data.

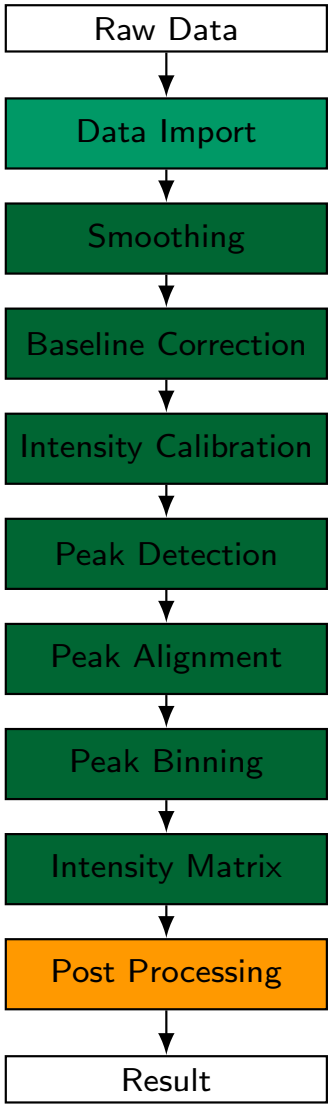
## MALDIquantForeign

- Import raw data (txt, tab, csv, Bruker Daltonics fid, Ciphergen XML, mzXML, mzML).
- Export into common formats (txt, tab, csv, msd, mzML).

# MALDIquant Family



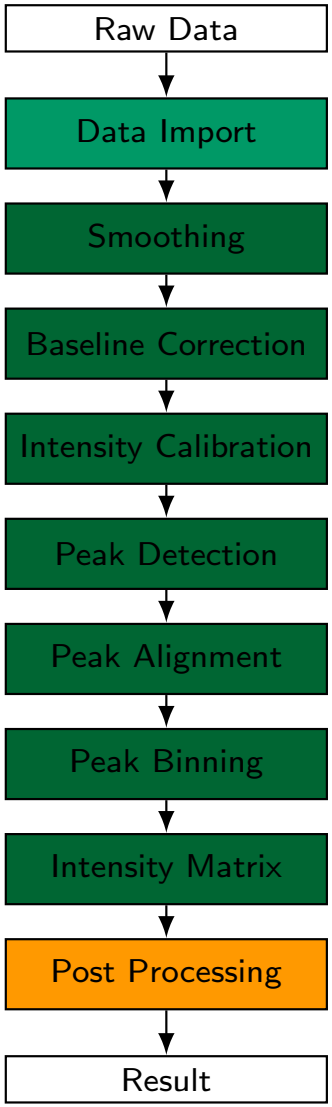
# Data Import




MALDIquant automatically recognizes many mass spectrometry data formats (including native Bruker files).

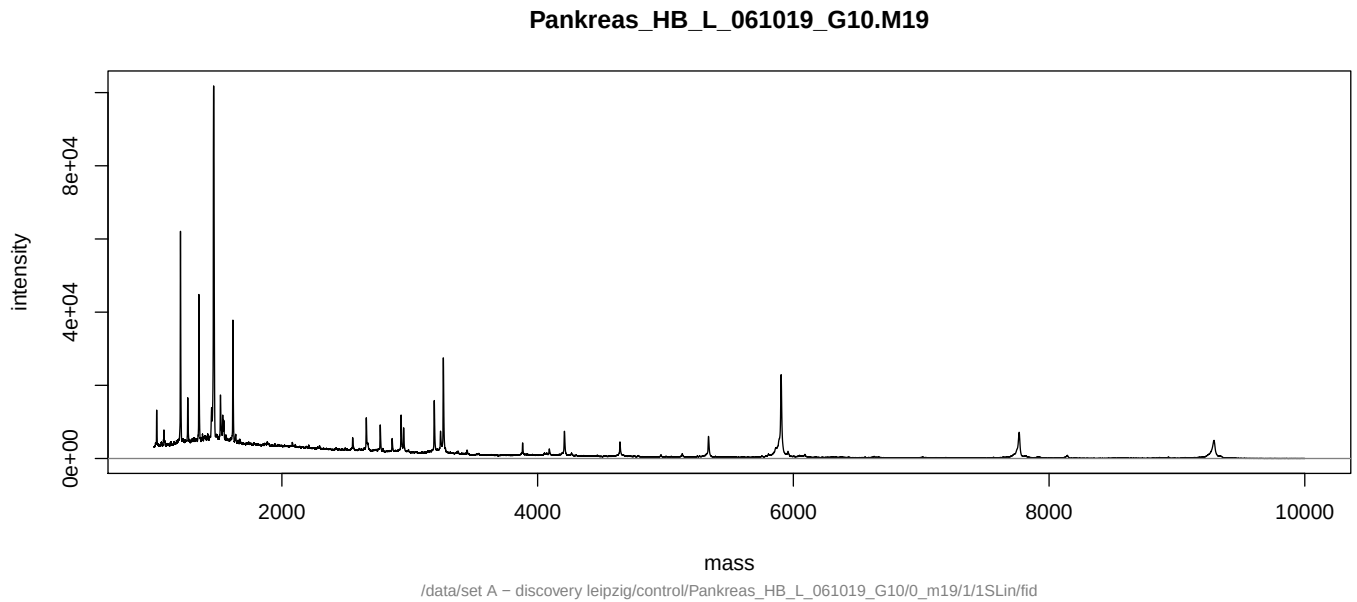
```
## load MALDIquant
library("MALDIquant")
## load MALDIquantForeign
library("MALDIquantForeign")

spectra <- import("/data/ms/raw")
```

# Data Import

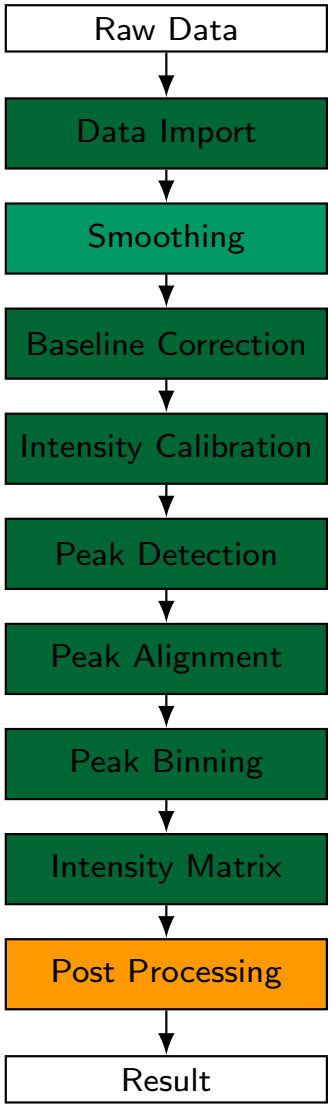



Here is a raw example spectrum from the Fiedler et al. (2009) cancer data set.

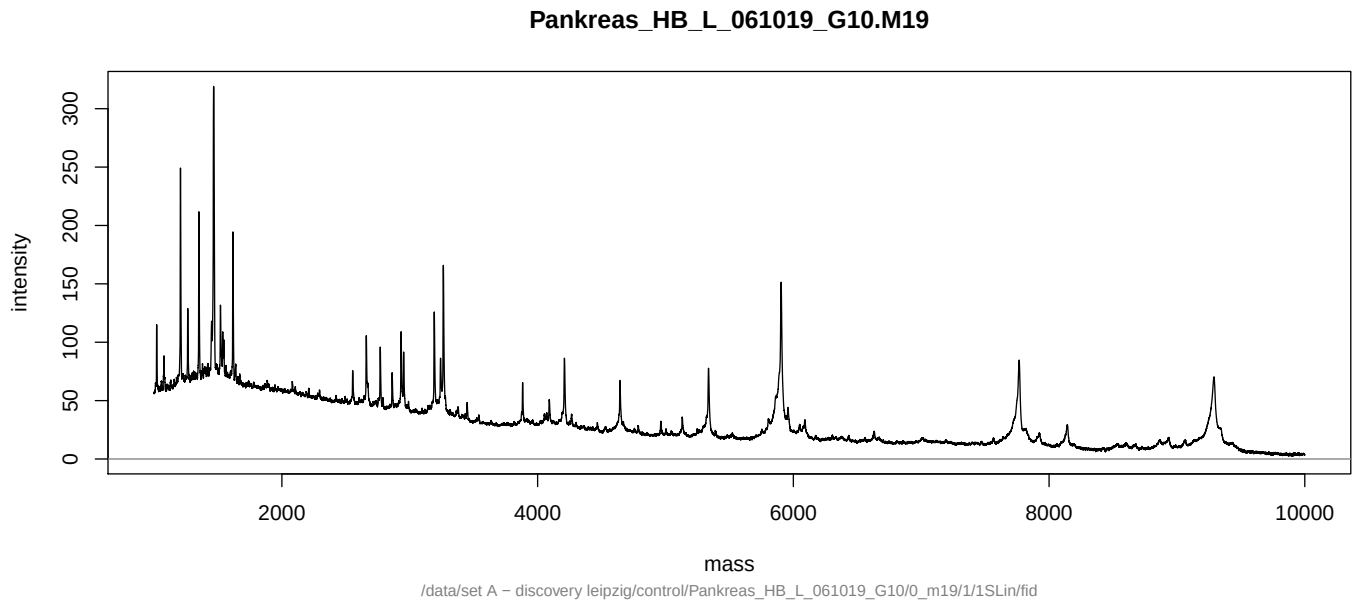
```
plot(spectra[[1]])
```



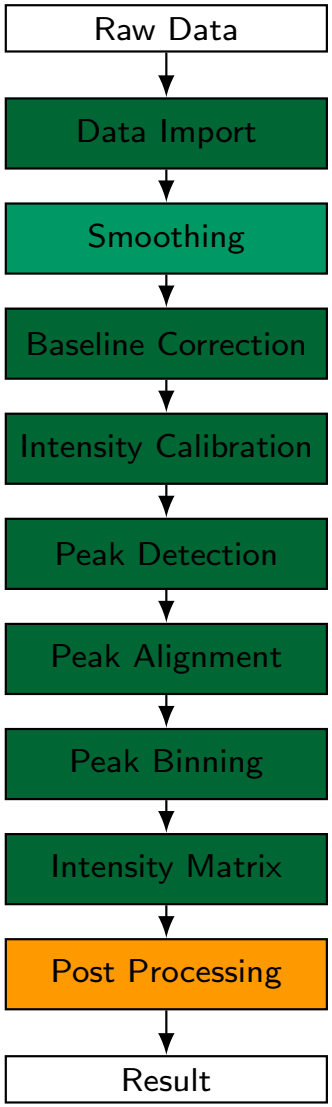
# Variance Stabilizing/Smoothing



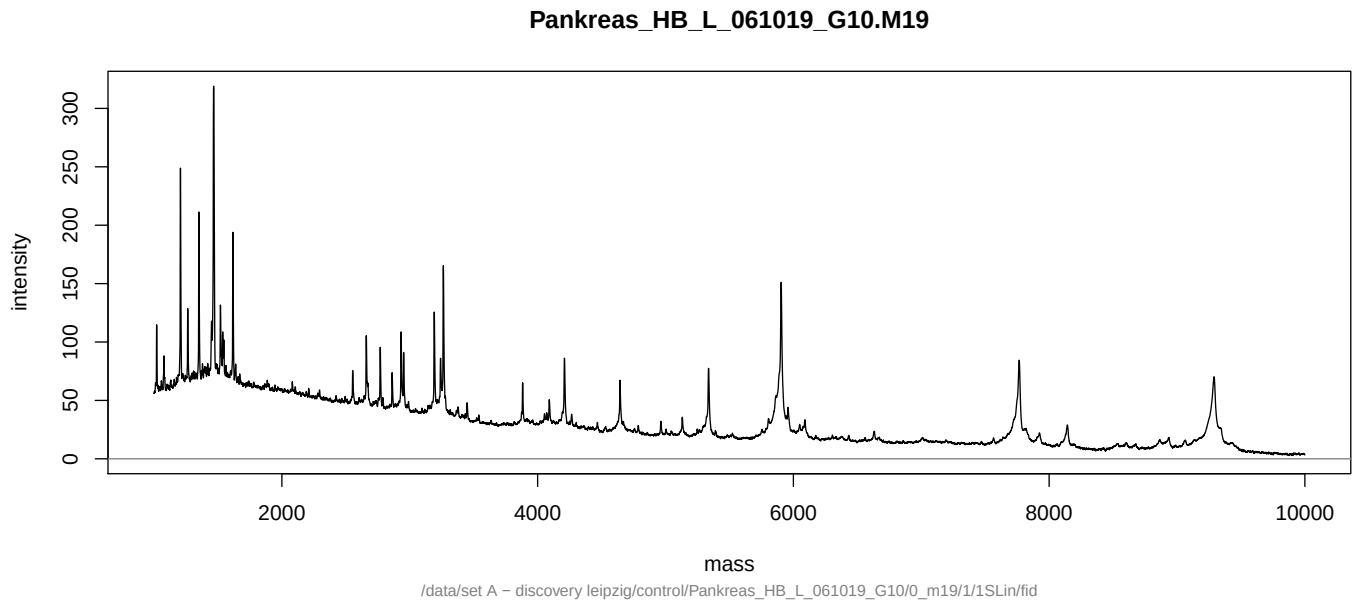

```
spectra <- transformIntensity(spectra, sqrt)
plot(spectra[[1]])
```



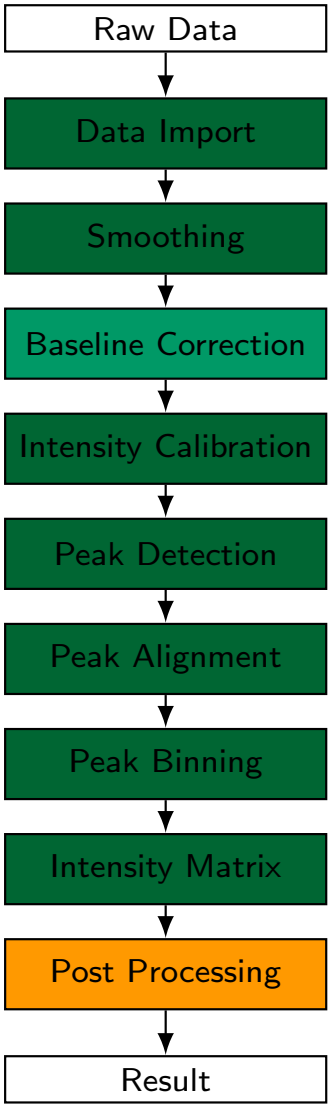
# Variance Stabilizing/Smoothing




```
movingAverage <- function(y) {
    return(filter(y, rep(1, 5)/5, sides = 2))
}
spectra <- transformIntensity(spectra, movingAverage)
```



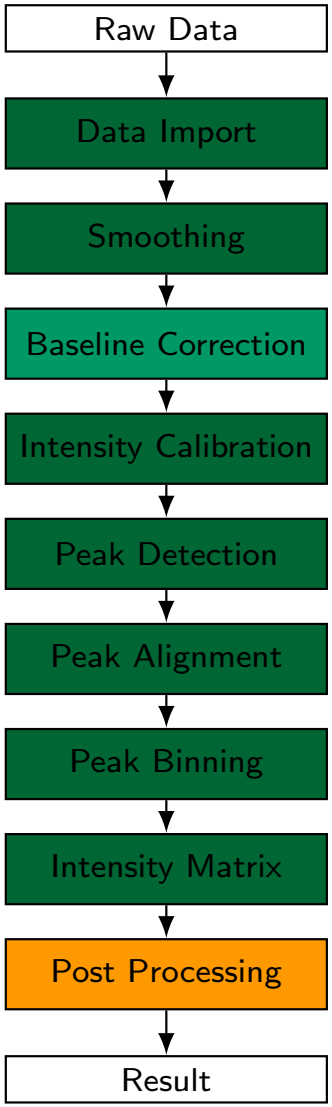
# Baseline Correction

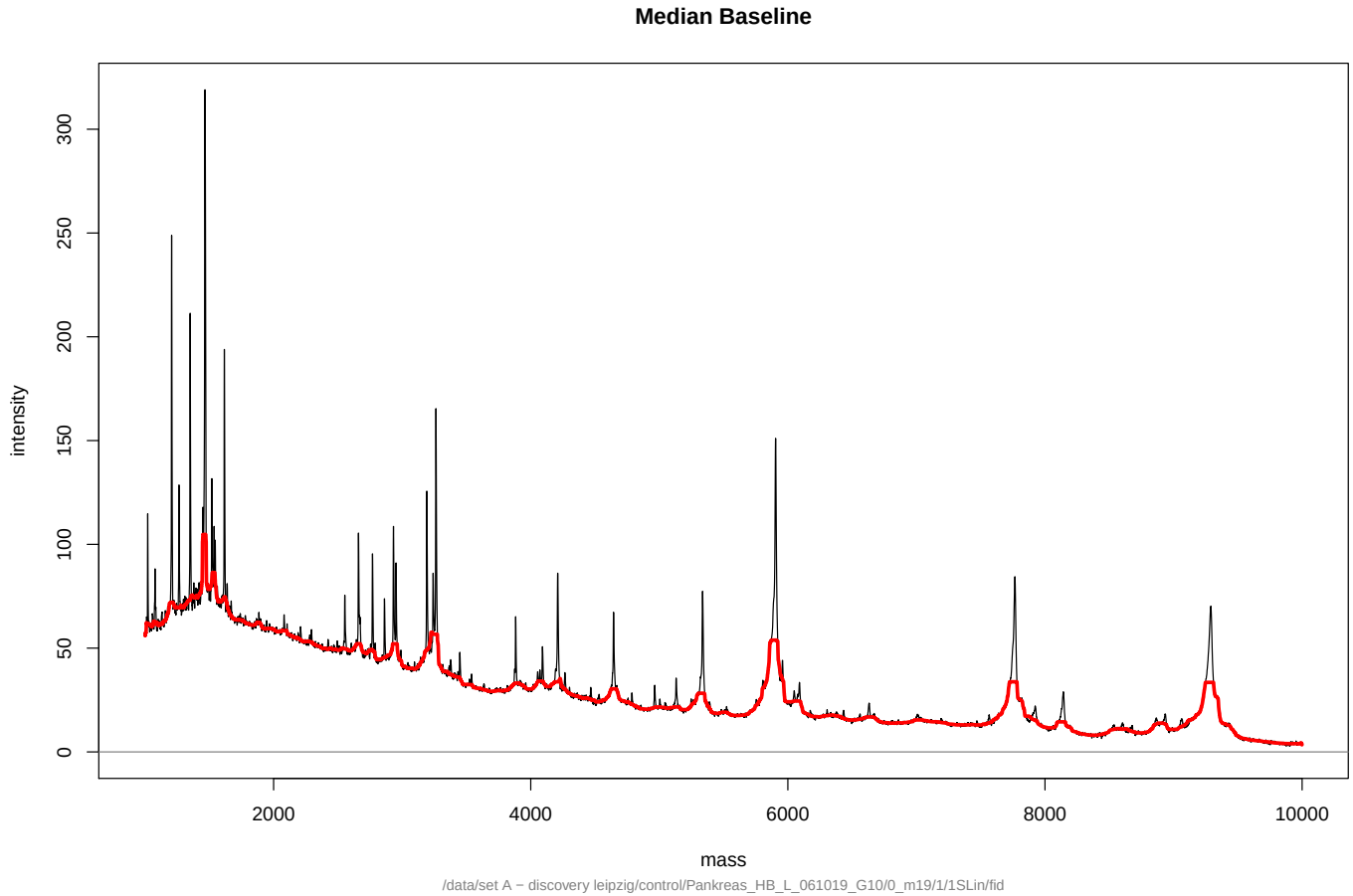



For correct quantification of peak intensities it is necessary to conduct baseline correction. This accounts for systematic bias, such as matrix effects.

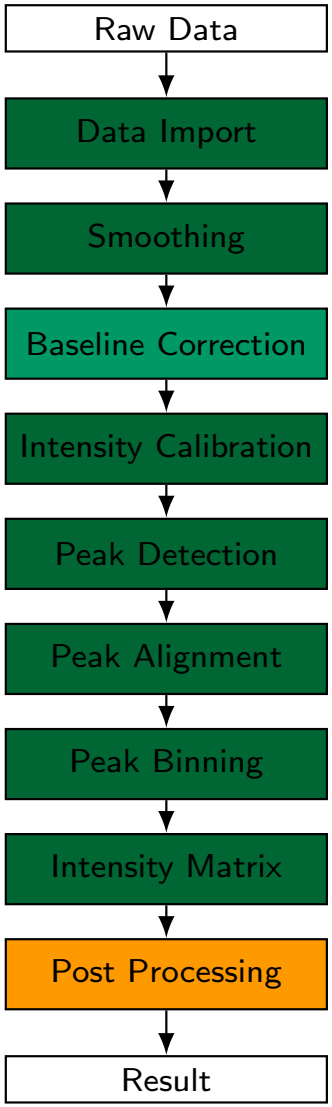
MALDIquant implements a number of correction algorithms, including the SNIP approach by Ryan et al (1988) and the TopHat filter.

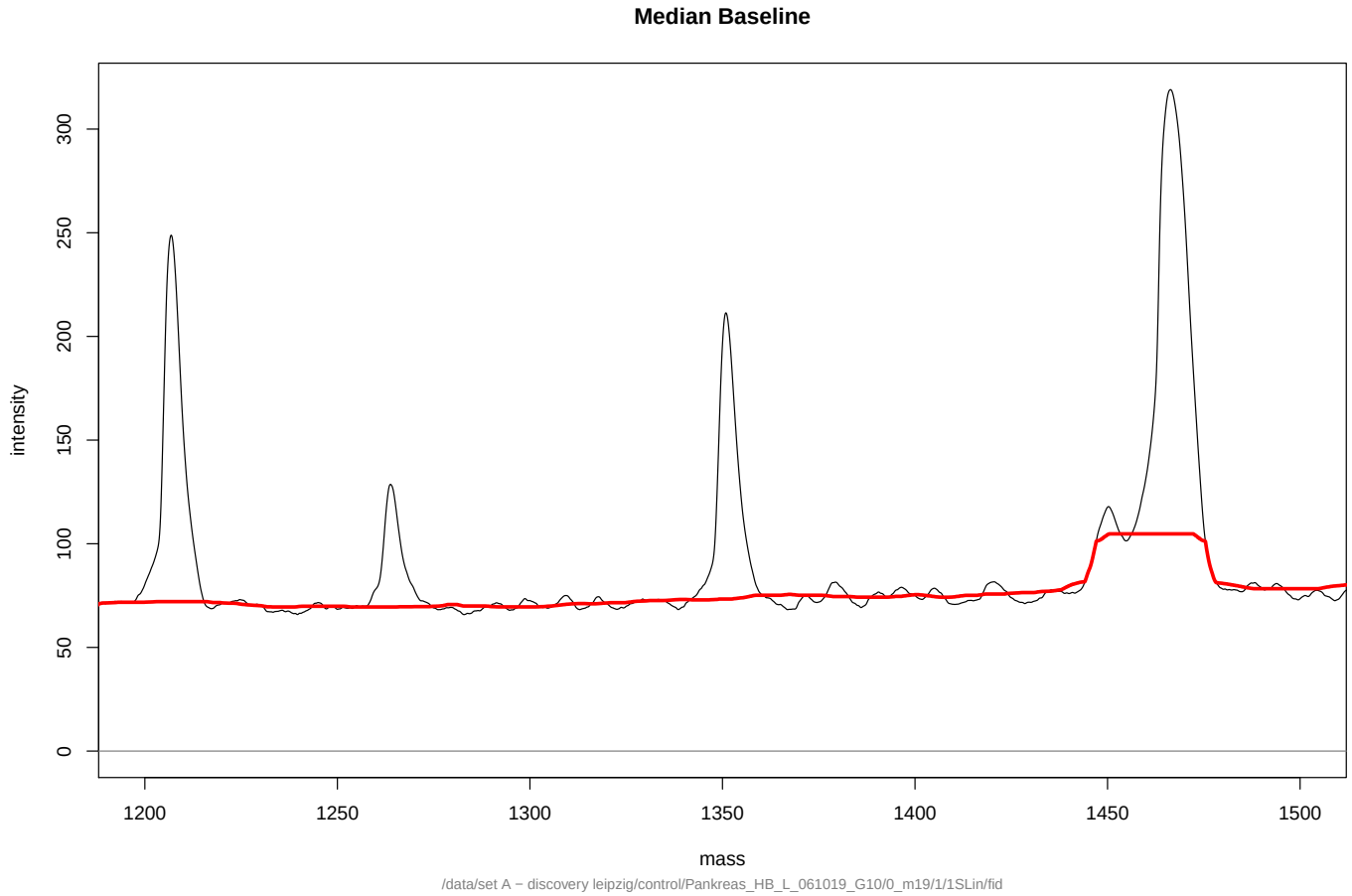
# Baseline Correction



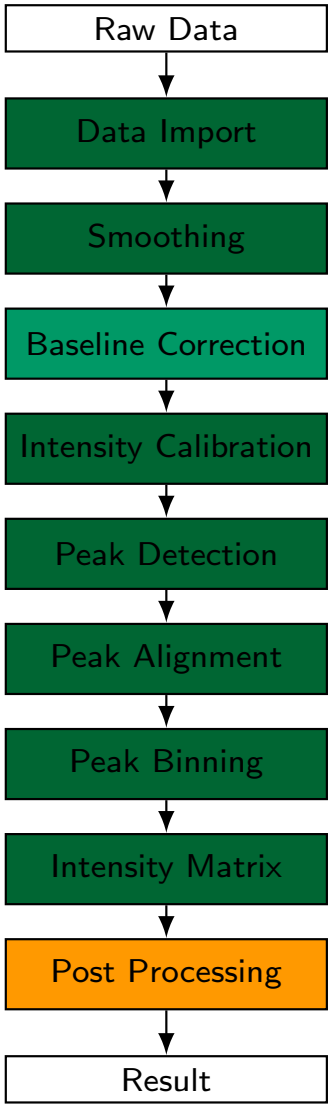



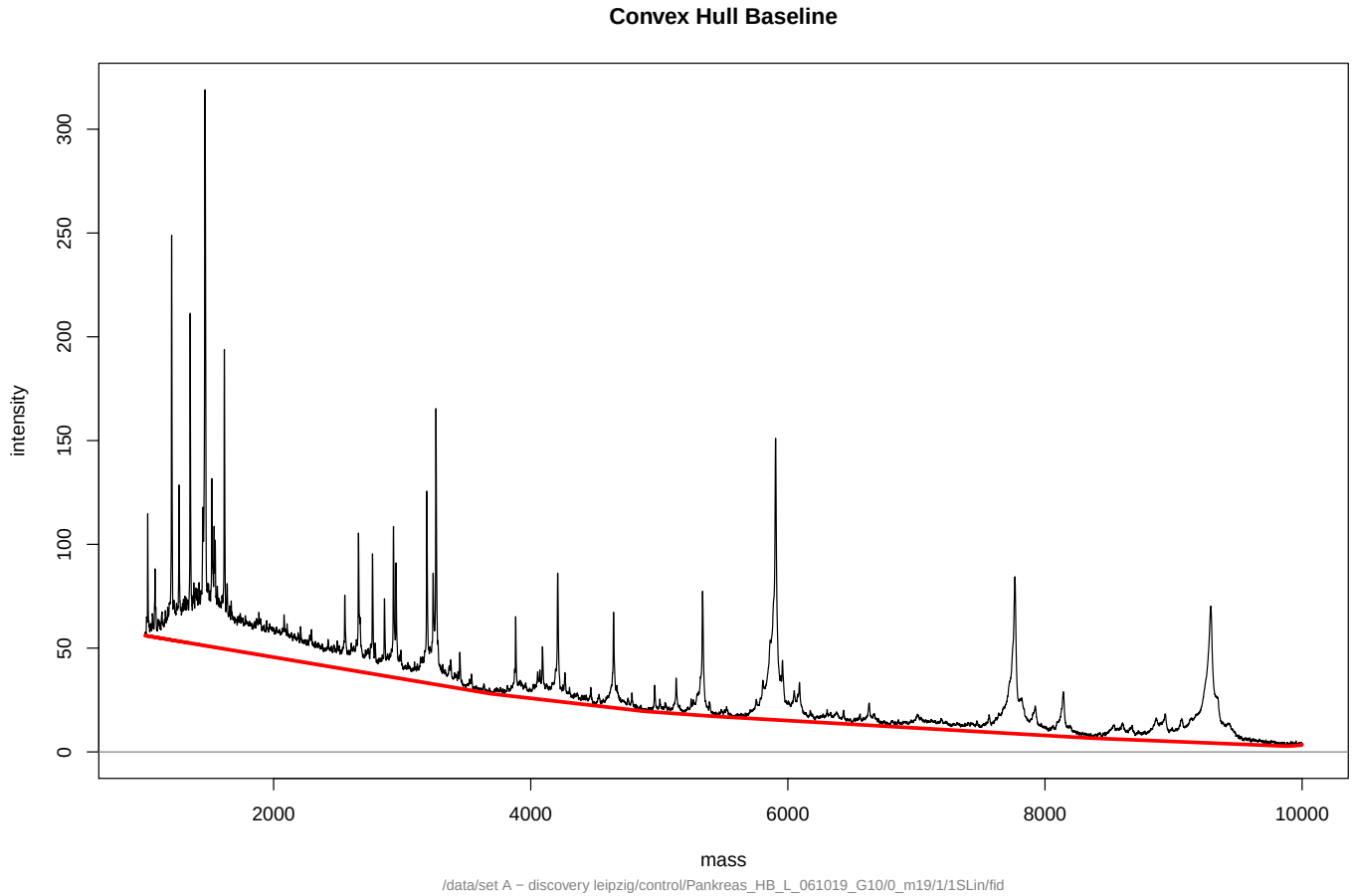
# Baseline Correction



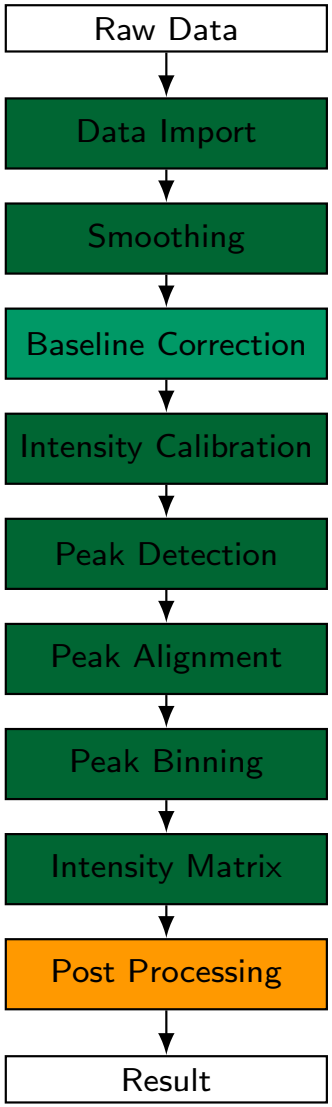



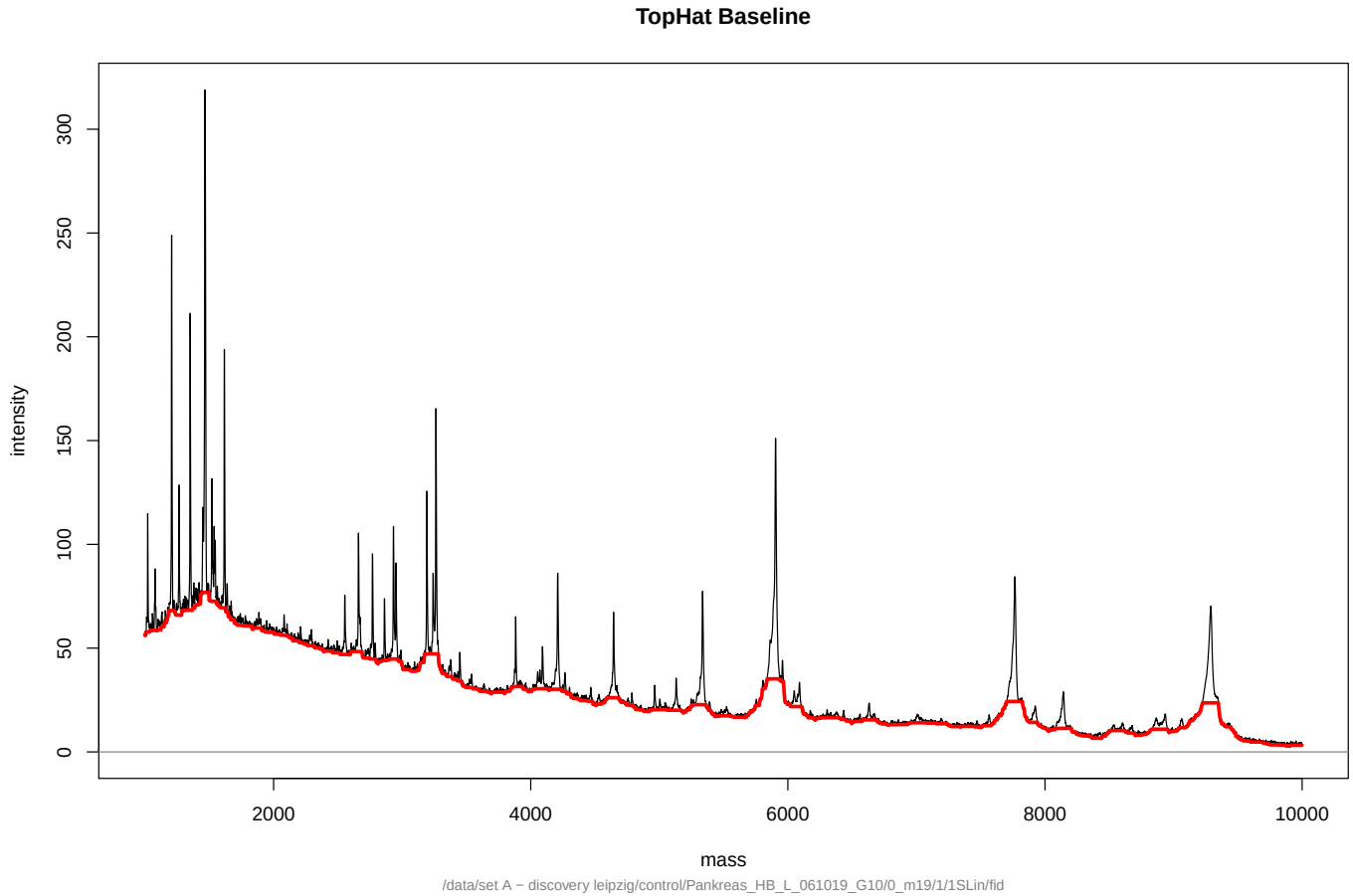
# Baseline Correction



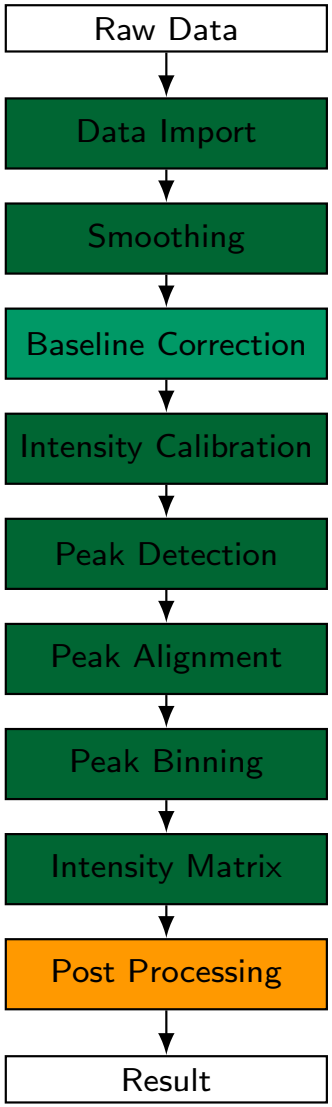



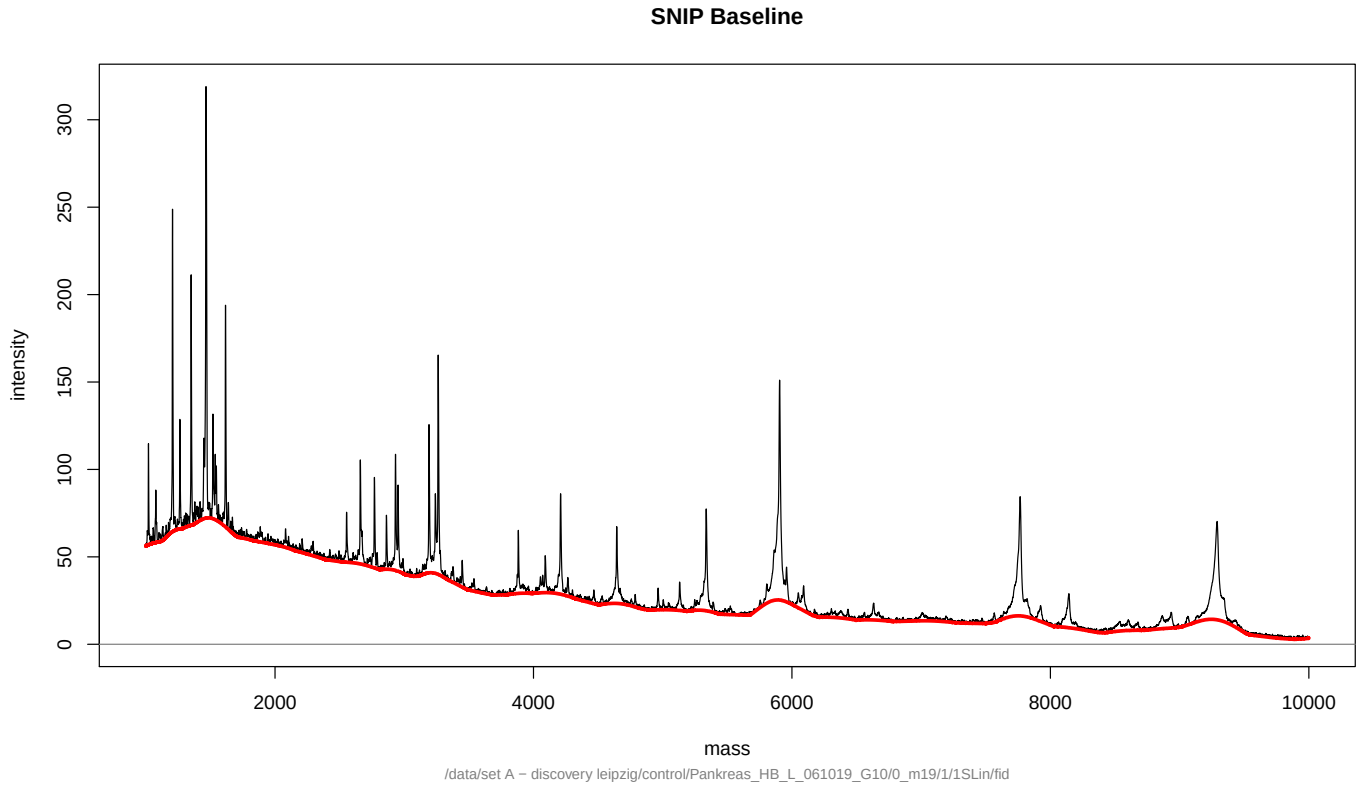
# Baseline Correction



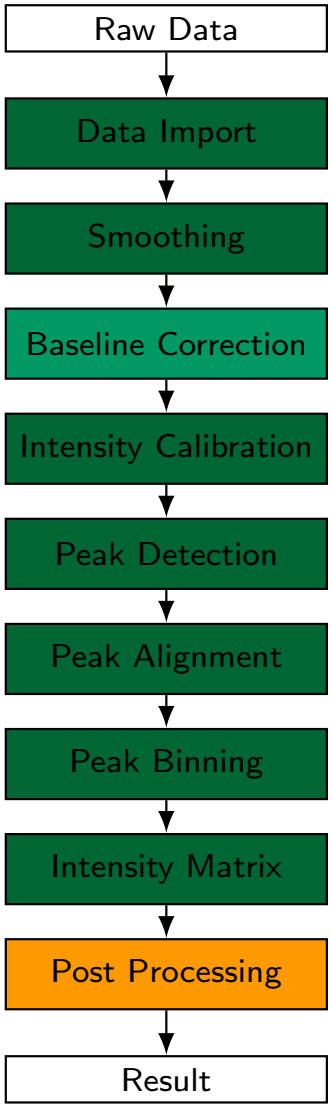
# Baseline Correction



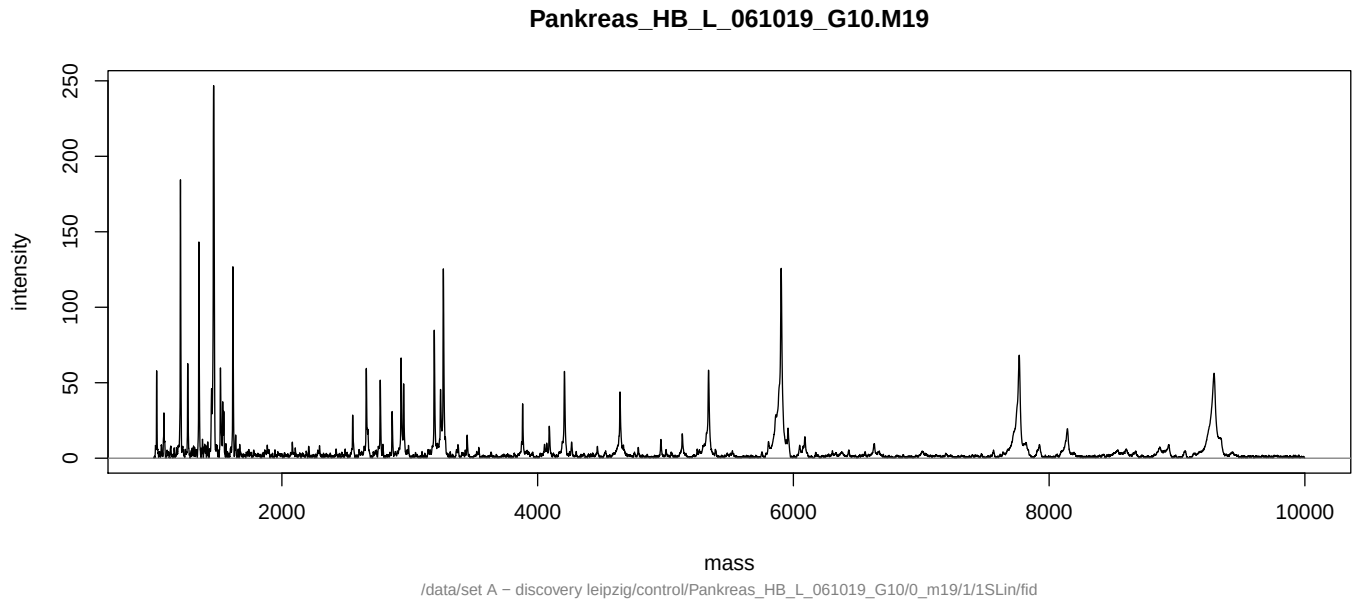




C. G. Ryan, E. Clayton, W. L. Griffin, S. H. Sie, and D. R. Cousens. SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications. *Nucl. Instrument. Meth. B*, 34: 396–402, 1988

# Baseline Correction

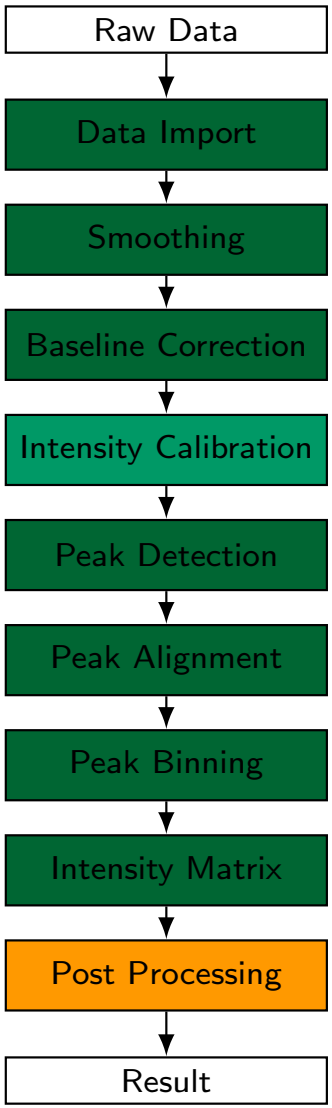



```
spectra <- removeBaseline(spectra, method = "SNIP")
plot(spectra[[1]])
```



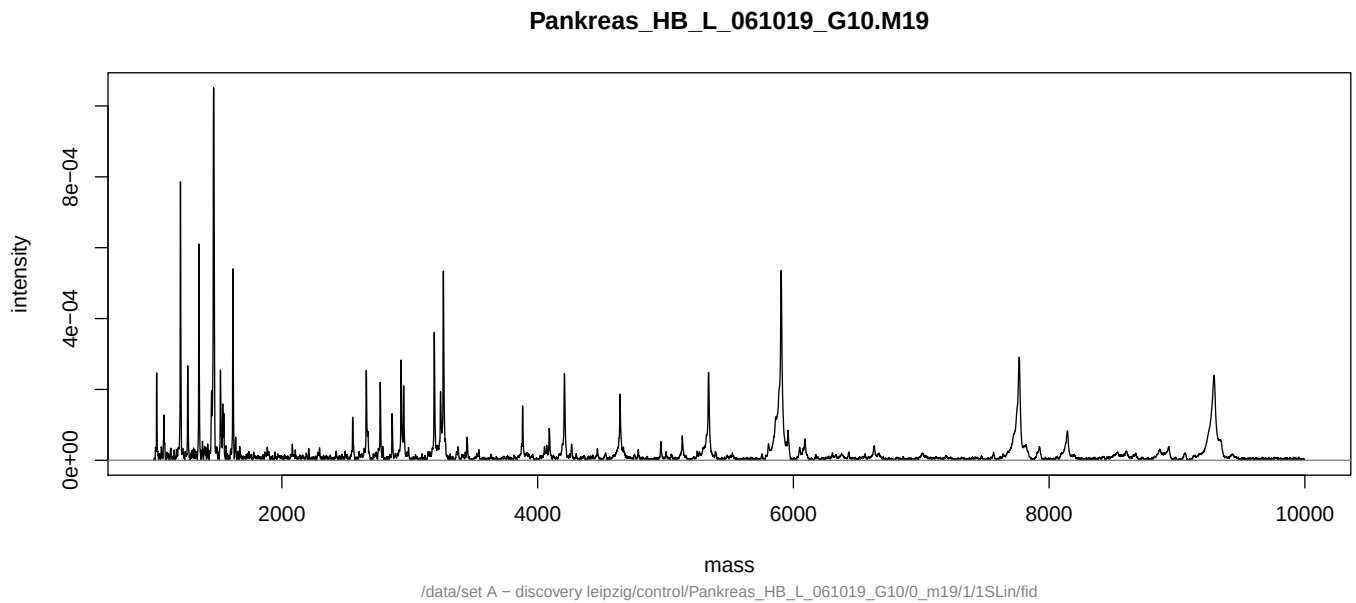

C. G. Ryan, E. Clayton, W. L. Griffin, S. H. Sie, and D. R. Cousens. SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications. *Nucl. Instrument. Meth. B*, 34: 396–402, 1988

# Intensity Calibration

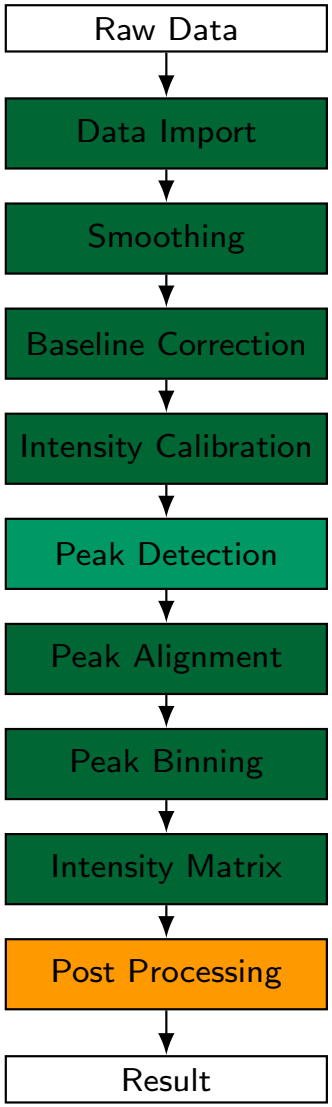


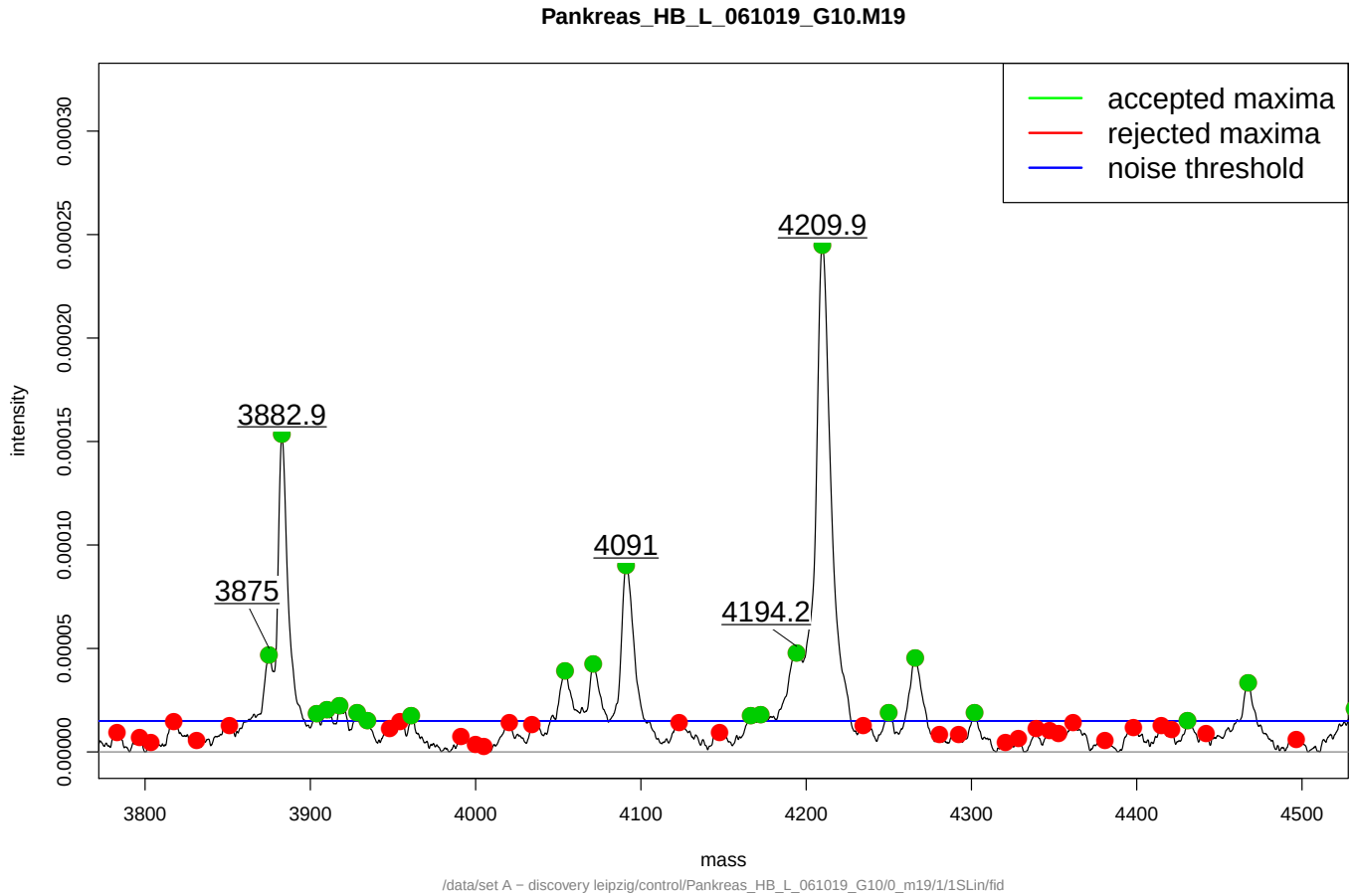

As with gene expression data, mass spectrometry data need to be calibrated. Several methods are available (TIC, Median, probabilistic quotient normalization).

```
spectra <- standardizeTotalIonCurrent(spectra)
plot(spectra[[1]])
```

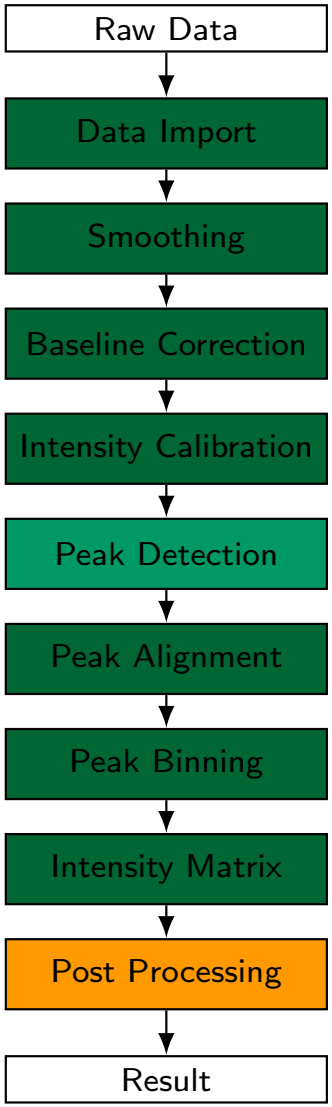


# Peak Detection

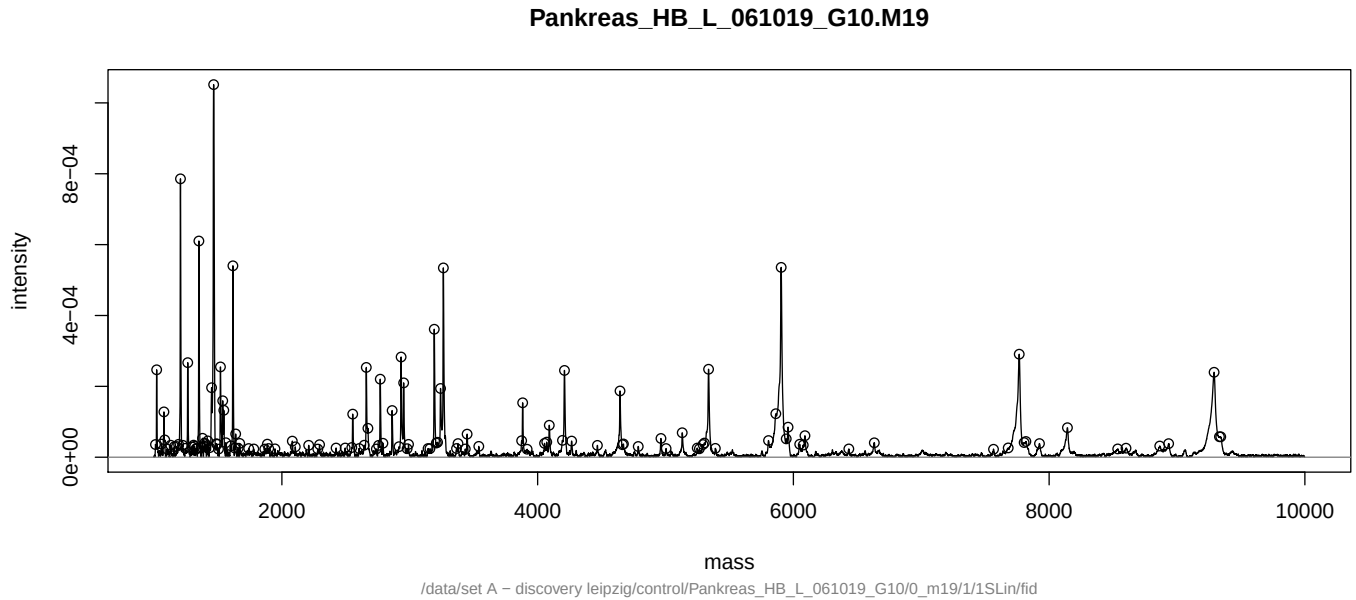


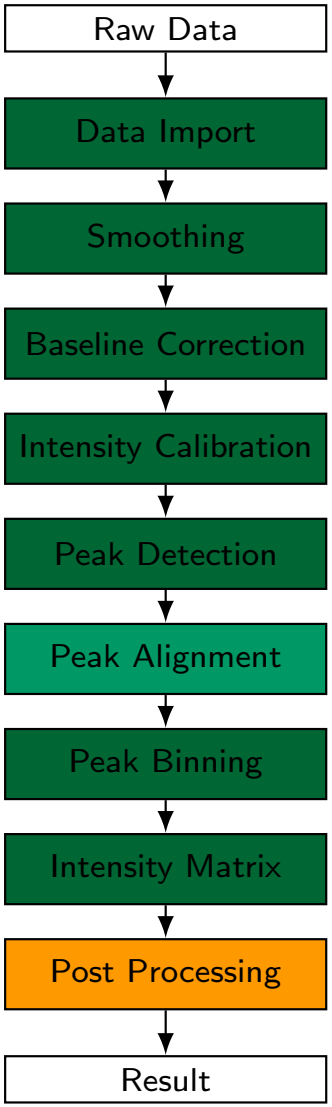
# Peak Detection




```
peaks <- detectPeaks(spectra, SNR = 3)
plot(spectra[[1]])
points(peaks[[1]])
```



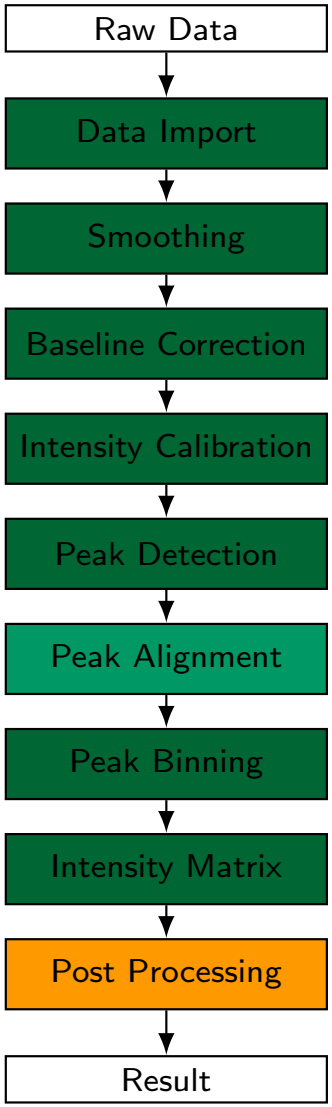
# Peak Alignment

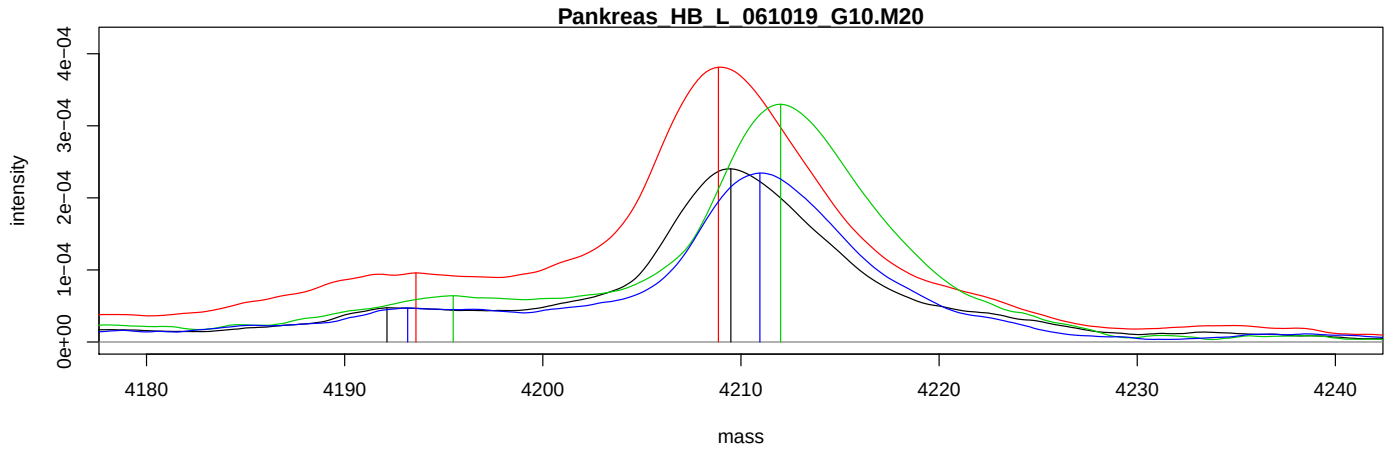


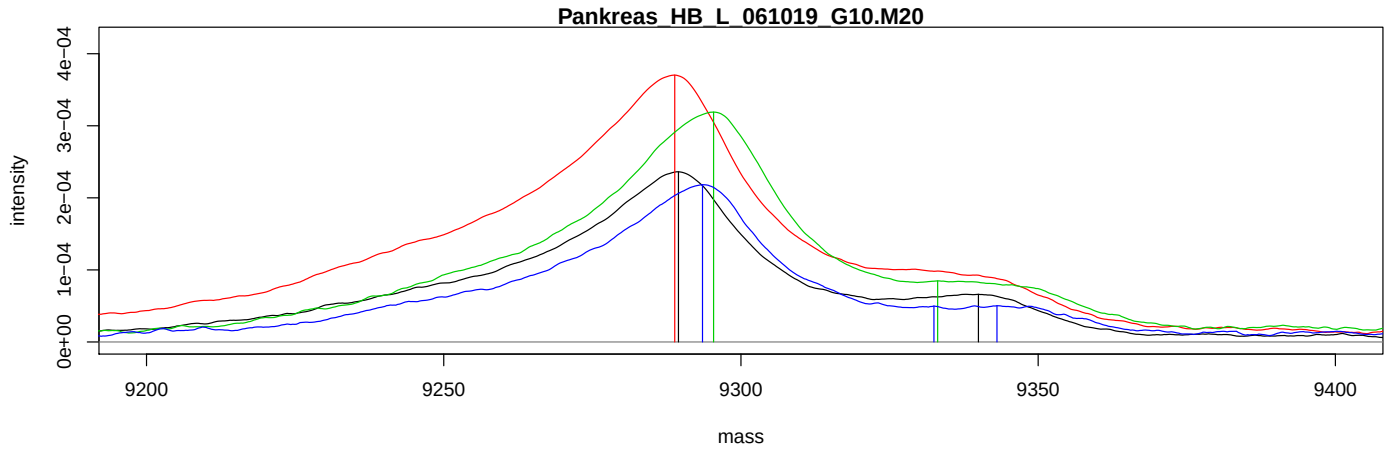

Peak alignment accross multiple spectra is a very challenging issue in mass spectrometry analysis:

- there are many peaks that occur only in few spectra
- non-linear mass shifts requires calibration along $x$-axis
- peak mapping needed for comparison and identification of markers

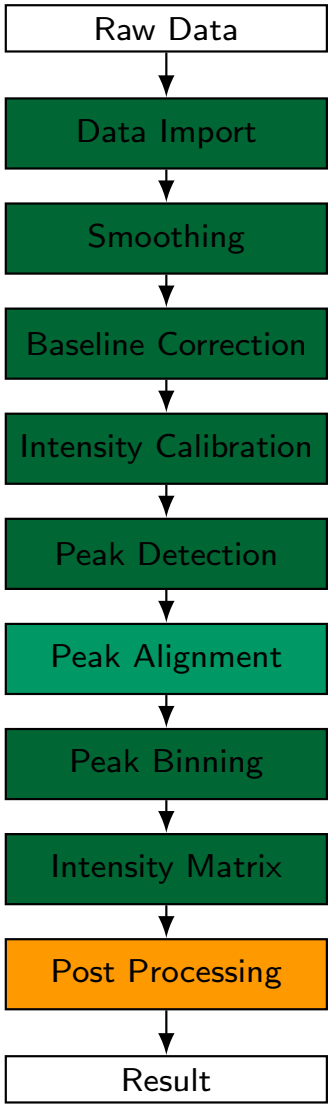
# Peak Alignment/The Problem



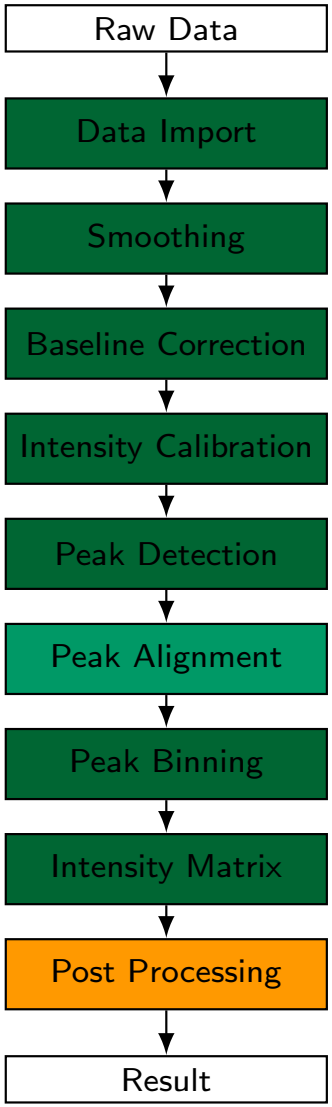
# Peak Alignment/Possible Solutions




| | |
|---|---|
| **DTW** [Clifford et al. 2009] | very slow & high memory usage $m_1 * m_2$: $((5 * 10^4)^2 * 48 Bytes \approx 120 GB)$ |
| **COW** [Veselkov et al. 2009] | slow, must run segment-wise to overcome nonlinear shift |
| **PTW** [Bloemberg et al. 2010] | fast ($\approx$ 3 sec/sample on an i5@2.5 GHz, 4 GB RAM) |

...

# Peak Alignment/Our Approach




## inspired by [Wang et al. 2010]

reference peaks: landmark peaks $\Rightarrow$ occur in most spectra

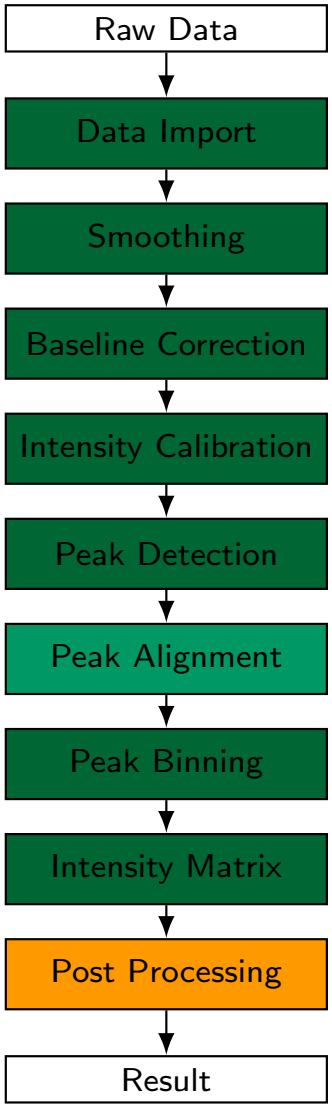
## inspired by [He et al. 2011]

- peak matching (choose highest peak in range)
- mass vs diff plot $\rightarrow$ estimate warping function

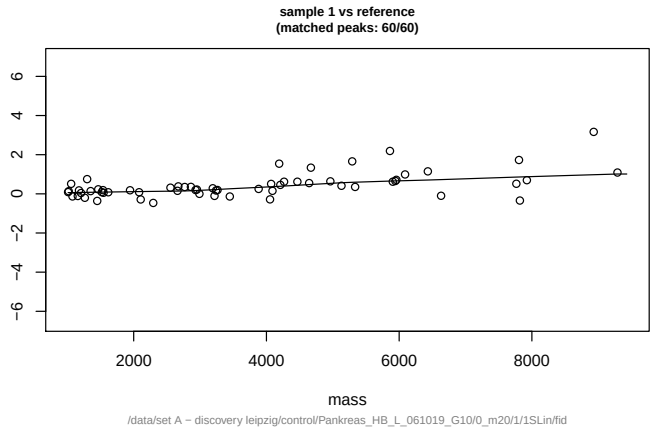
Bing Wang, Aiqin Fang, John Heim, Bogdan Bogdanov, Scott Pugh, Mark Libardoni, and Xiang Zhang. Disco: Distance and spectrum correlation optimization alignment for two-dimensional gas chromatography time-of-flight mass spectrometry-based metabolomics. *Analytical Chemistry*, 82(12):5069–5081, 2010

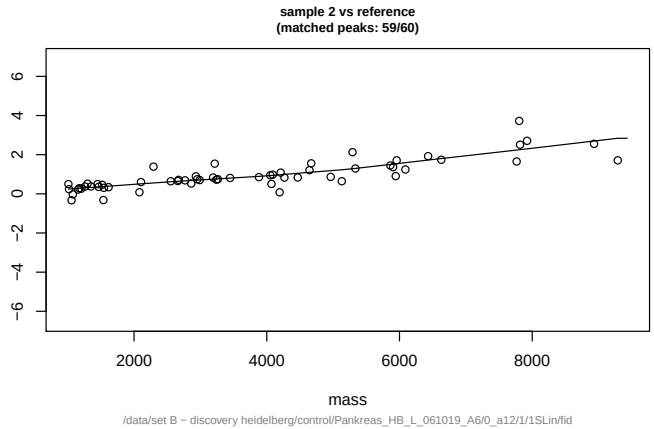
Q P He, J Wang, J A Mobley, J Richman, and W E Grizzle. Self-calibrated warping for mass spectra alignment. *Cancer Inform*, 10:65–82, 2011

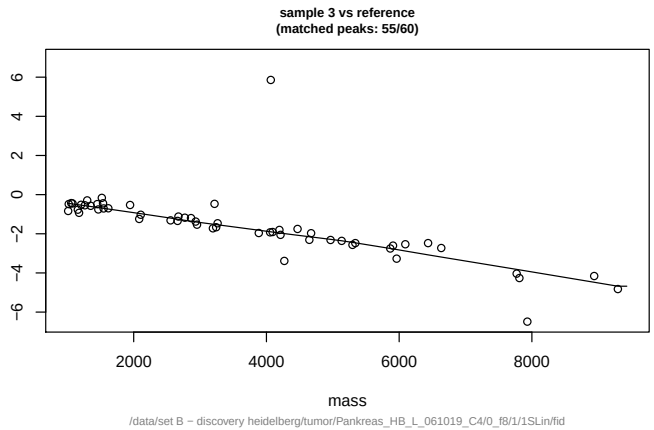
# Peak Alignment/Warping Functions

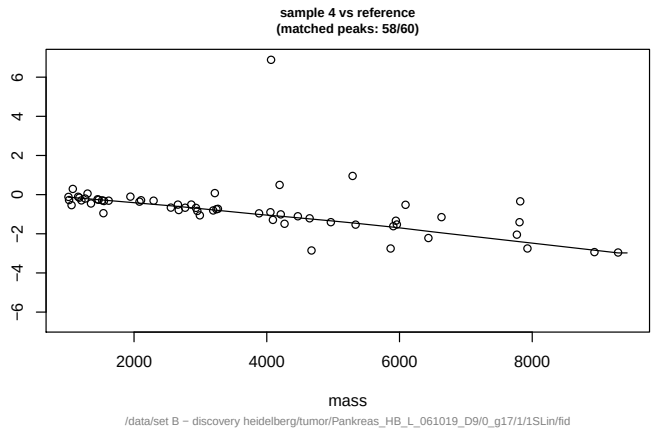



```
warpingFunctions <- determineWarpingFunctions(peaks)
```

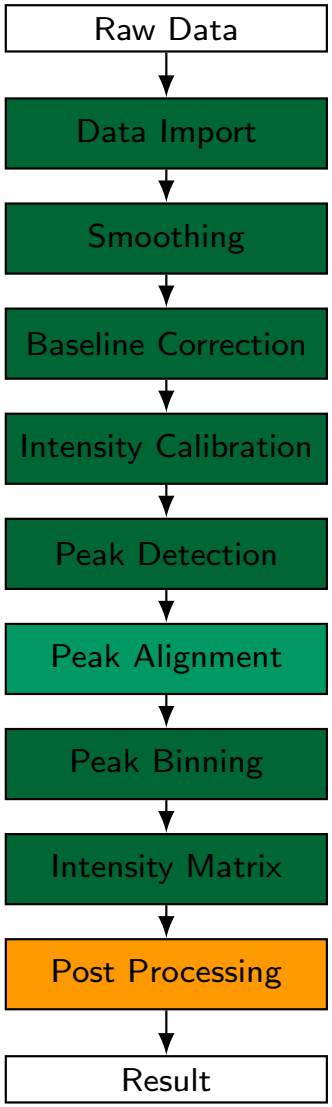


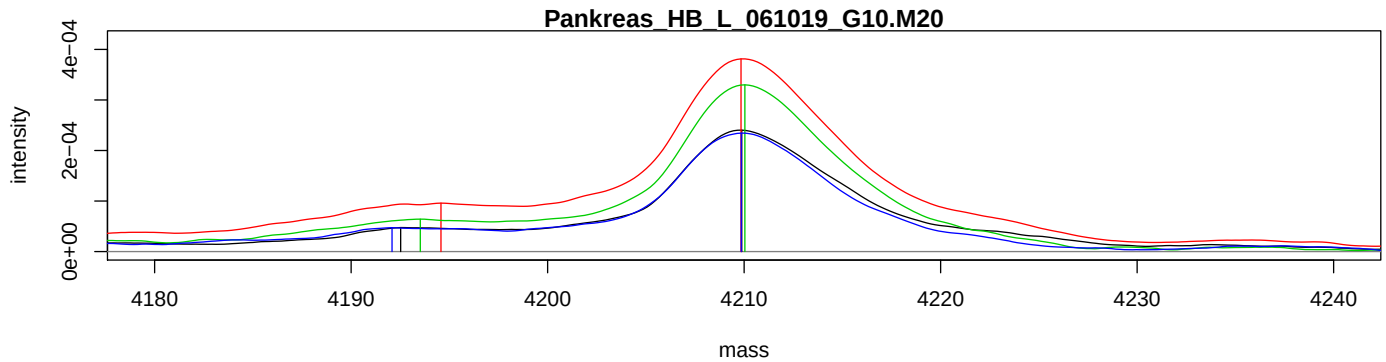



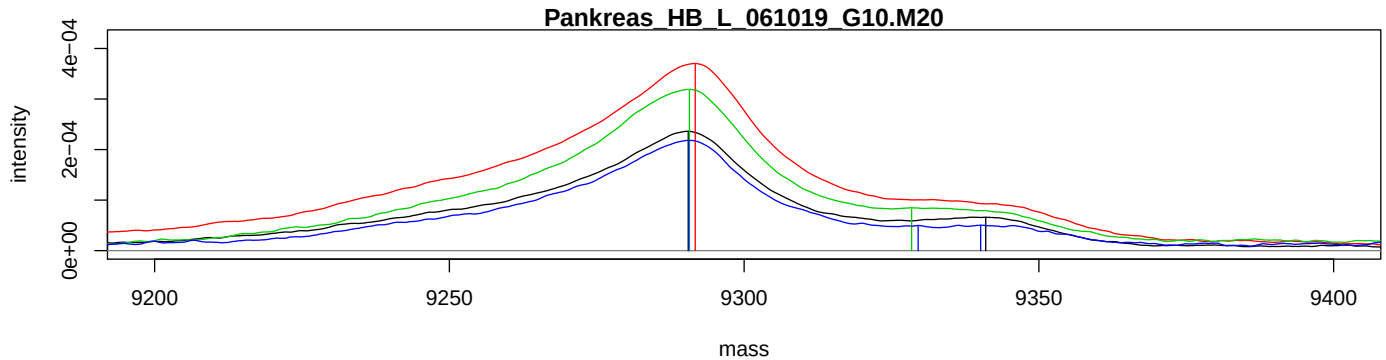



# Peak Alignment/Warping Functions

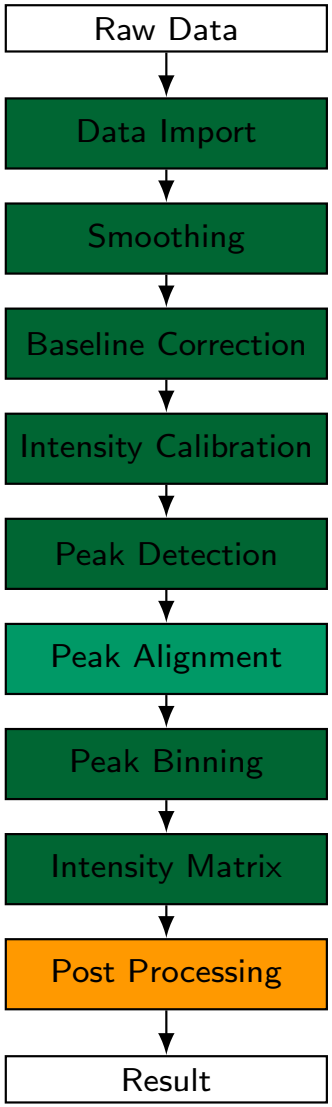


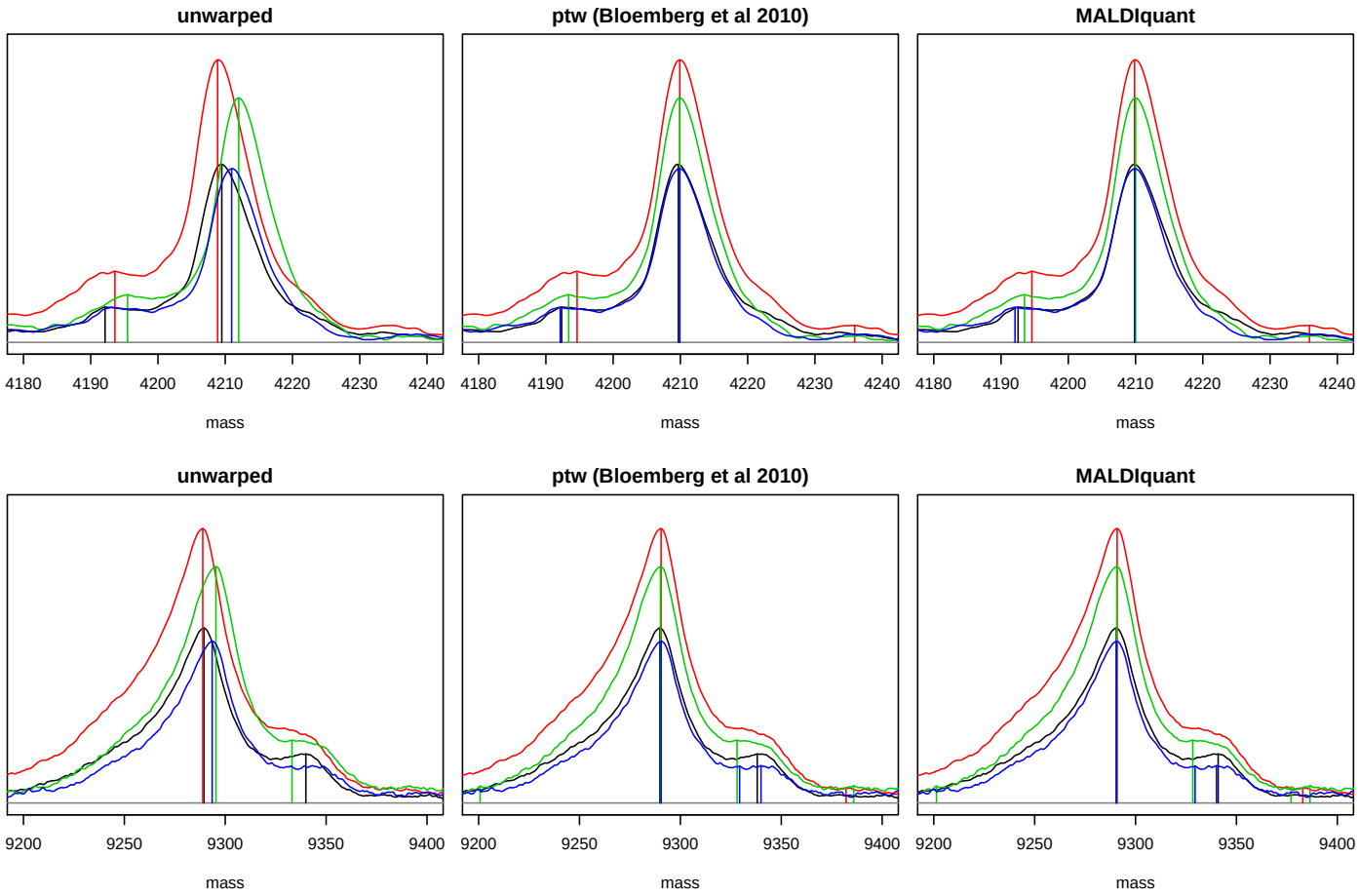

```
spectra <- warpMassSpectra(spectra, warpingFunctions)
peaks <- warpMassPeaks(peaks, warpingFunctions)
```



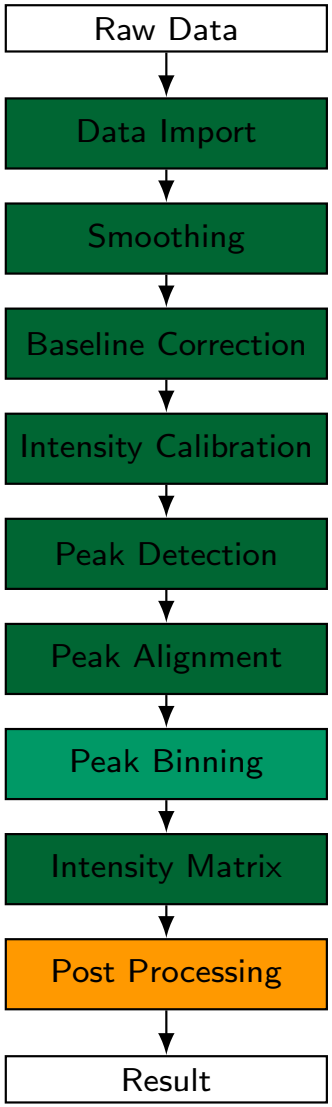



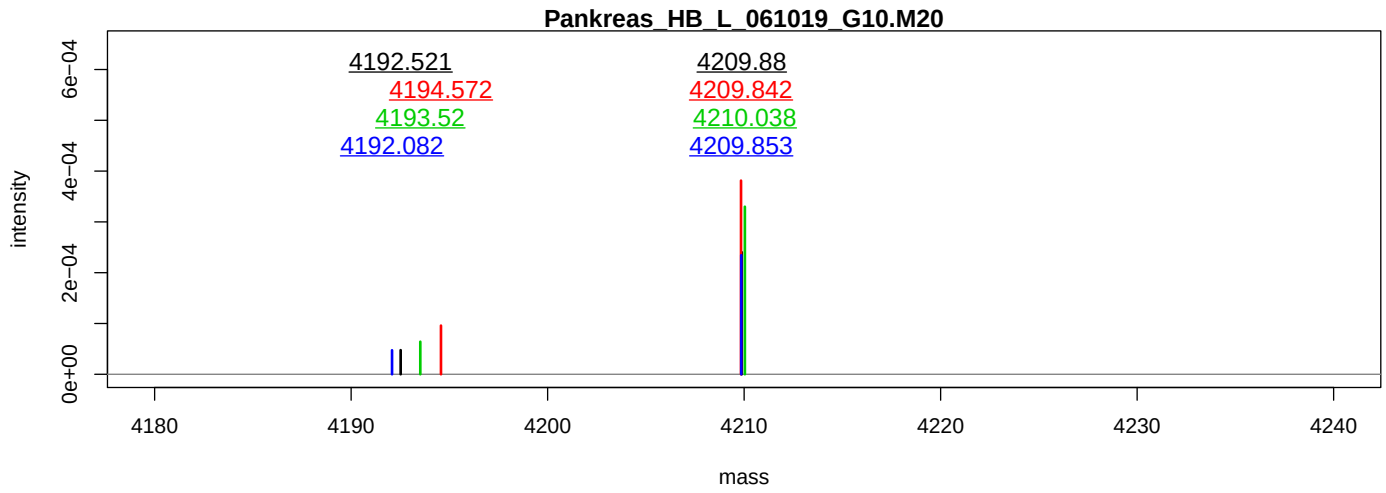
# Peak Alignment/Comparison

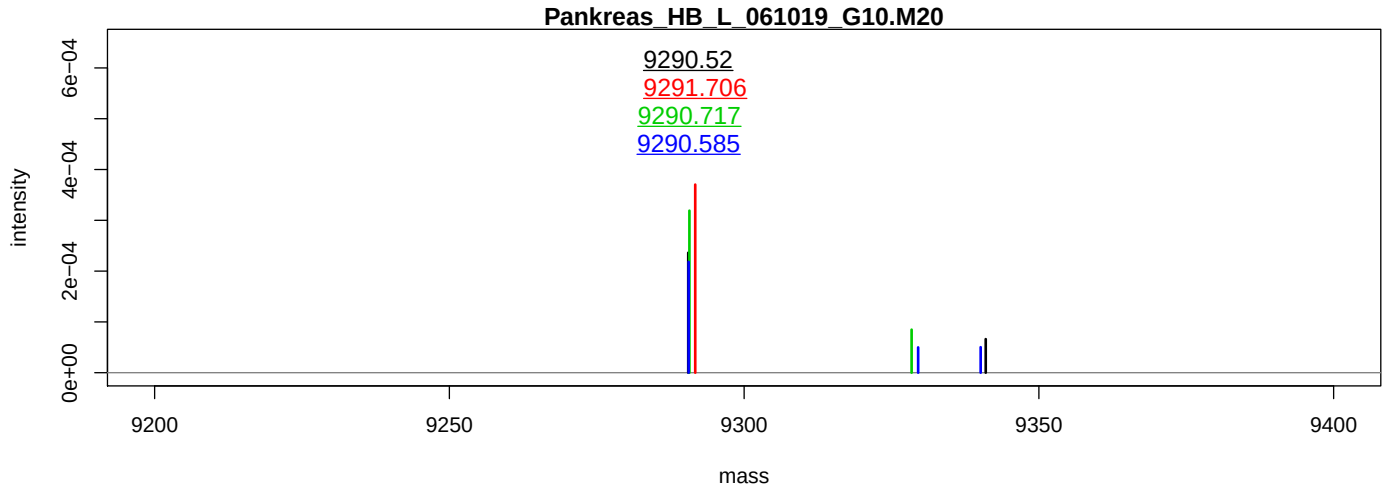


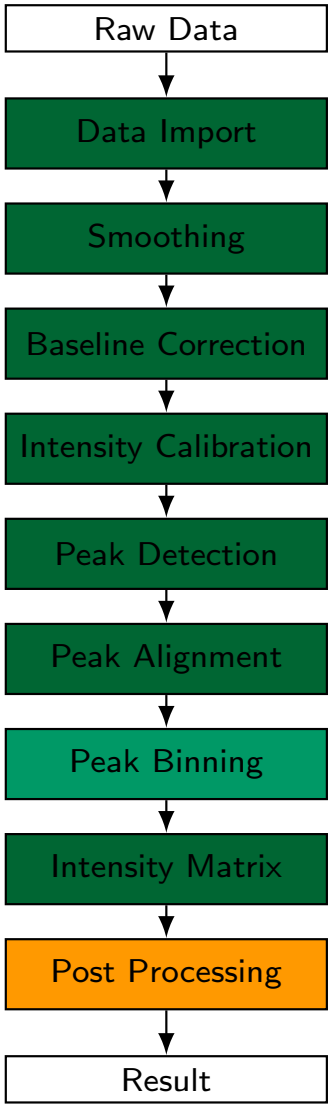
# Peak Binning



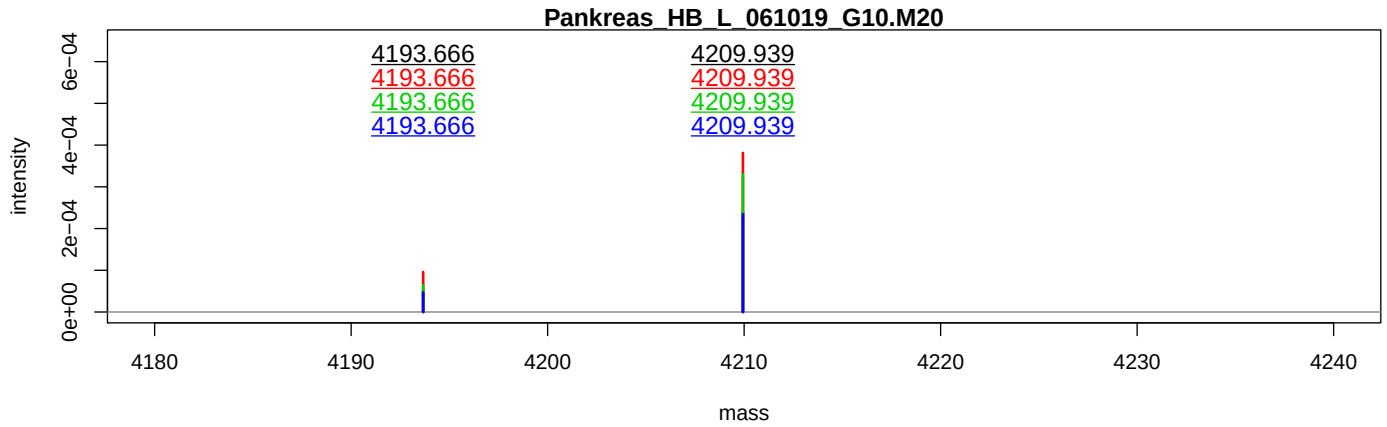



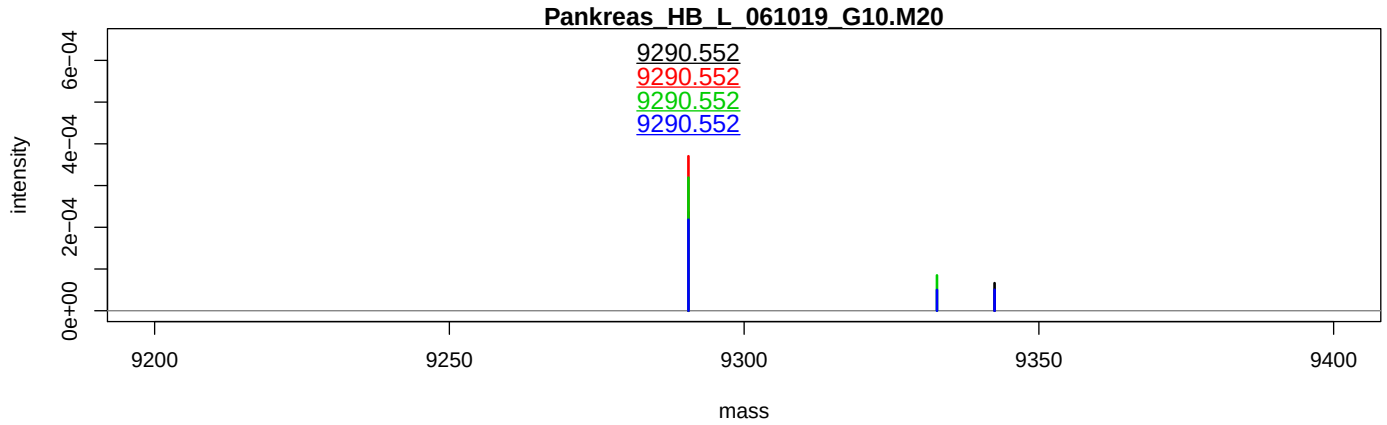



# Peak Binning

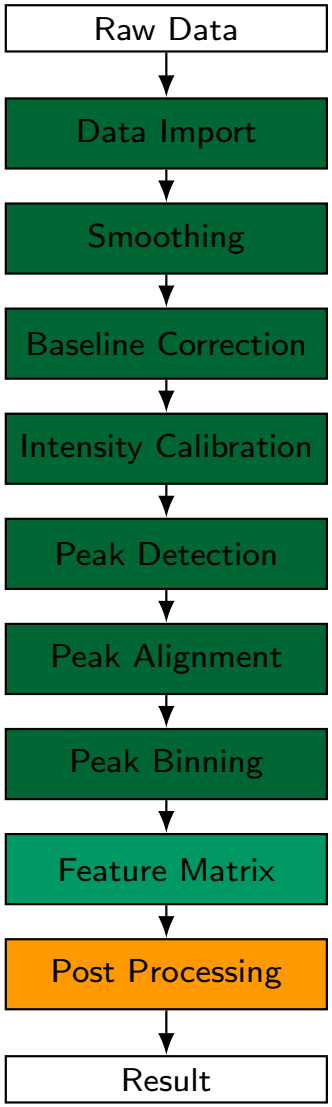



```
peaks <- binPeaks(peaks)
```



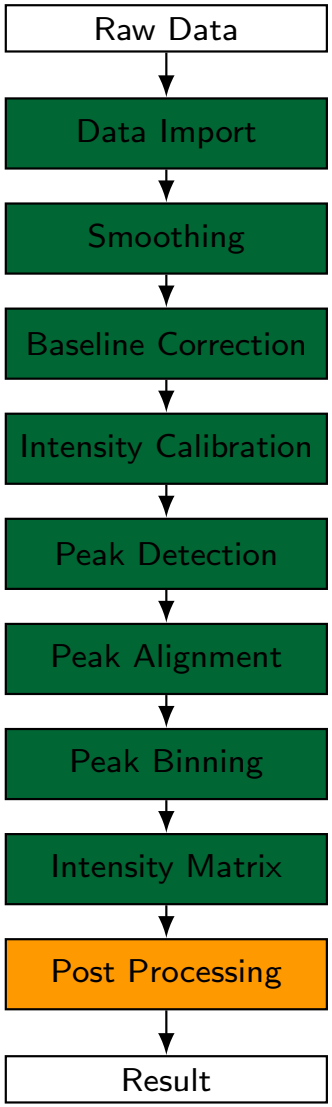



# Feature Matrix



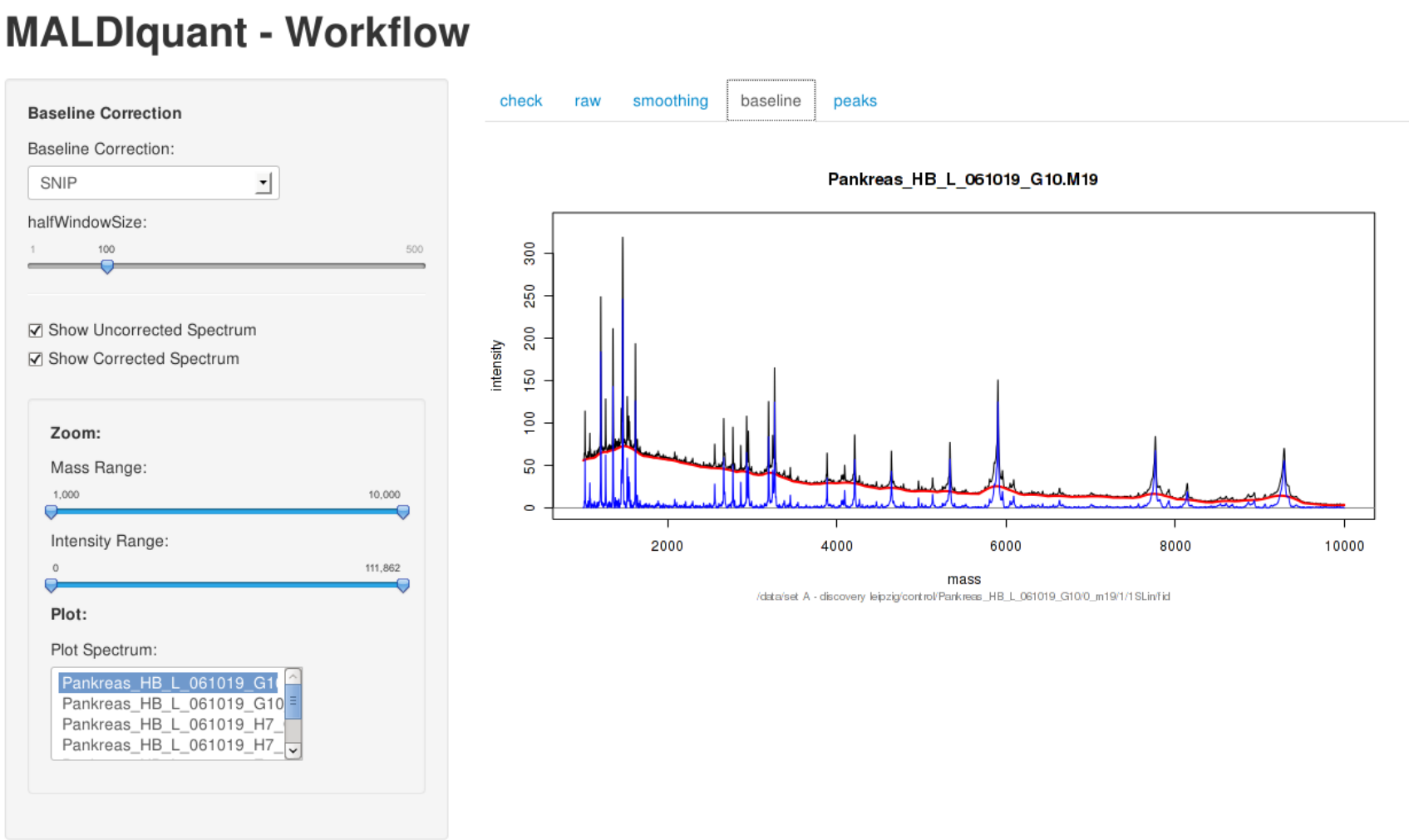

```
featureMatrix <- intensityMatrix(peaks)
featureMatrix[1:12, 1:2]

##       1011.67823313295 1020.66971843048
##  [1,]     3.587257e-05     0.0002467926
##  [2,]     3.170674e-05     0.0002550081
##  [3,]     3.517940e-05     0.0002428846
##  [4,]     3.430174e-05     0.0002563571
##  [5,]     3.122426e-05     0.0004472052
##  [6,]               NA     0.0004505502
##  [7,]     2.680547e-05     0.0004240763
##  [8,]     3.484823e-05     0.0003559640
##  [9,]     4.327525e-05     0.0001619205
## [10,]     3.357397e-05     0.0001527801
## [11,]     4.160095e-05     0.0002912183
## [12,]     3.848561e-05     0.0002911327
```

# Workflow Summary




```
## load libraries
library("MALDIquant")
library("MALDIquantForeign")

## load data
spectra <- import("/data/ms/raw")

## run spectrum based workflow
spectra <- transformIntensity(spectra, sqrt)
spectra <- transformIntensity(spectra, movingAverage)
spectra <- removeBaseline(spectra)
spectra <- standardizeTotalIonCurrent(spectra)
peaks <- detectPeaks(spectra)

## run peak based workflow
warpingFunctions <- determineWarpingFunctions(peaks)
peaks <- warpMassPeaks(peaks)
peaks <- binPeaks(peaks)
featureMatrix <- intensityMatrix(peaks)

## use featureMatrix for further analysis
```

# MALDIquant GUI



# III. High-Level Analysis

# Multivariate Analysis of Omics Data

In the last decade a multitude of statistical methods have been developed for high-dimensional gene expression data. For example:

- regularized $t$-scores (moderated $t$, shrinkage $t$) for gene ranking,
- regularized regression and classification (e.g. shrinkage discriminant analysis),
- sparse models, structured models, latent class models, and
- high-dimensional testing (e.g. Higher Criticism or FDR)

Can we use these techniques also for mass spectrometry data?

# Peaks as Biomarkers

Mass spectrometric peaks contain both binary and continuous information:

1. peak may be present or absent (NA in intensity matrix)
2. intensity of peak may be up or down regulated

Both properties need to be taken into account!
Thus, usual methods from gene expression data are generally not appropriate! (But nonetheless used in practice.)

# Solution: Dichotomization

Local peak thresholding (Tibshirani et al 2004) uses a peak-specific thresholding rule $I_{\text{thresh}}(m)$ to dichotomize spectral data:

1. peak intensity $I(m) > I_{\text{thresh}}(m) \rightarrow 1$
2. peak intensity $I(m) \leq I_{\text{thresh}}(m) \rightarrow 0$

This allows to take account both of absent/present as well as up/down regulated peaks.
The thresholding rule is estimated from the data by maximizing the separation between the groups (as measured by a variable importance criterion).

$\rightarrow$ for mass spectrometry data we need categorical data analysis!

# Classification and Ranking with Binary Predictors

Large-scale analysis of binary data is routine in machine learning, especially in text mining.

We suggest using similar techniques as in text mining for the analysis of mass spectrometry data, in particular:

- multivariate Bernoulli and related models for classification,
- corresponding variable importance measures (typically based on KL entropy) for peak ranking, and
- regularized inference for application in high-dimensional settings.

# Multivariate Bernoulli Independence Rule

One of the most popular approaches and highly effective approach for classification of gene expression data is PAM (Tibshirani et al 2003), a variant of shrinkage diagonal discriminant analysis.

We use the same idea for mass spectrometry data:

- assume "diagonal" multivariate Bernoulli distributions for each group, $\mathbf{X}_k = (X_1, \ldots, X_d) \sim B_d(\boldsymbol{\mu}_k)$ with $\Pr(\mathbf{x}|\boldsymbol{\mu}_k) = \prod_{i=1}^{d} \Pr(x_i|\mu_{i,k})$
- Bayes rule gives discriminant function $d_k \propto \log \Pr(k|\mathbf{x})$.
- Regularized training of discriminant rule in situations with $d \geq n$.

This MVB independence rule, while very simply, has shown to be highly effective (e.g. Park 2009).

# Pancreatic Cancer Proteomics Study

G.M. Fiedler et al. 2009. *Serum peptidome profiling revealed platelet factor 4 as a potential discriminating peptide associated with pancreatic cancer*. Clinical Cancer Research **11**: 3812–3819.

- Large study conducted in Leipzig and Heidelberg
- 120 participants (60 healthy vs. 60 cancer)
- 4 technical replicates per sample
- 480 MALDI spectra

# Preprocessing and Peak Filtering

- Number of detected peaks per spectrum (of 480): between 134 and 257
- After merging of technical replicates, keeping only peaks that occur in all 4 measured spectra: between 52 and 146
- Keeping only peaks that occur with 75 % frequency in each group (cancer and control): between 23 and 56

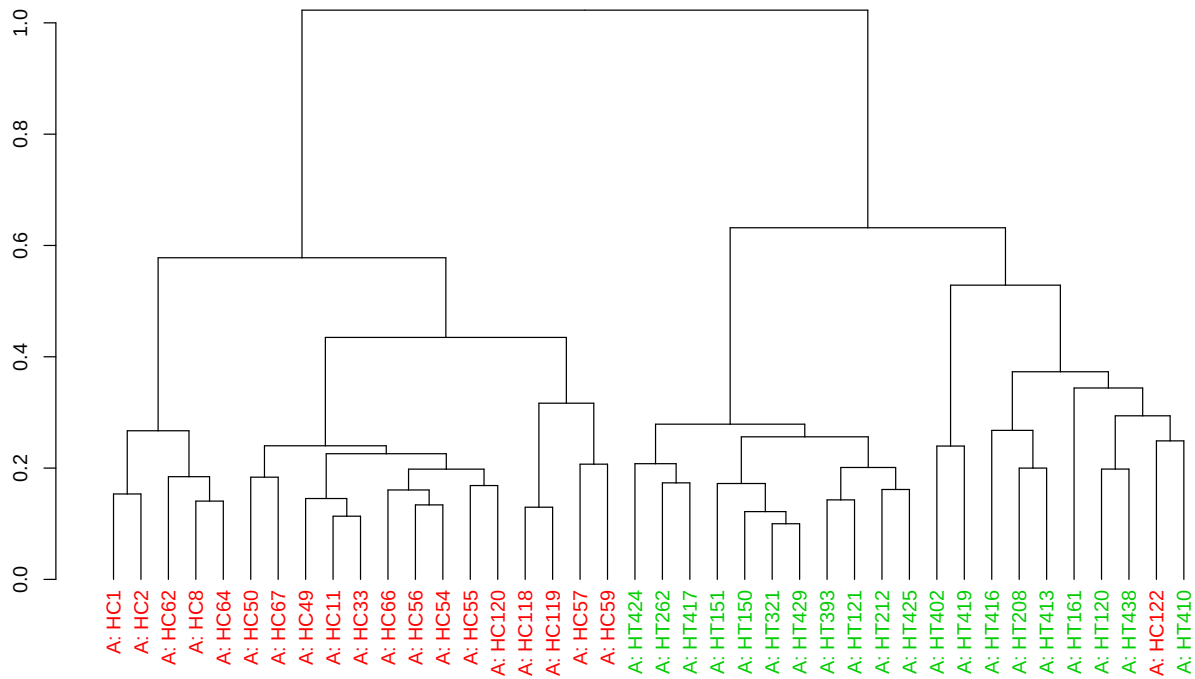
# Comparison of Rankings

Shrinkage $t$ vs. categorical ranking (top 10 peaks):

| Ranking | metric | discrete |
|---|---|---|
| 1 | **4494.71** | **4494.71** |
| 2 | **2755.41** | 8937.02 |
| 3 | **4250.76** | 4467.81 |
| 4 | 8131.39 | **5945.42** |
| 5 | **2022.73** | **2022.73** |
| 6 | 1627.86 | 1866.07 |
| 7 | 3920.23 | **4250.76** |
| 8 | 8144.25 | **2755.41** |
| 9 | **5945.42** | 5906.06 |
| 10 | 5266.03 | 2953.13 |

# Clustering of Dichotomized Data

Dichotomized data exhibits near perfect split in healthy and cancer patients!



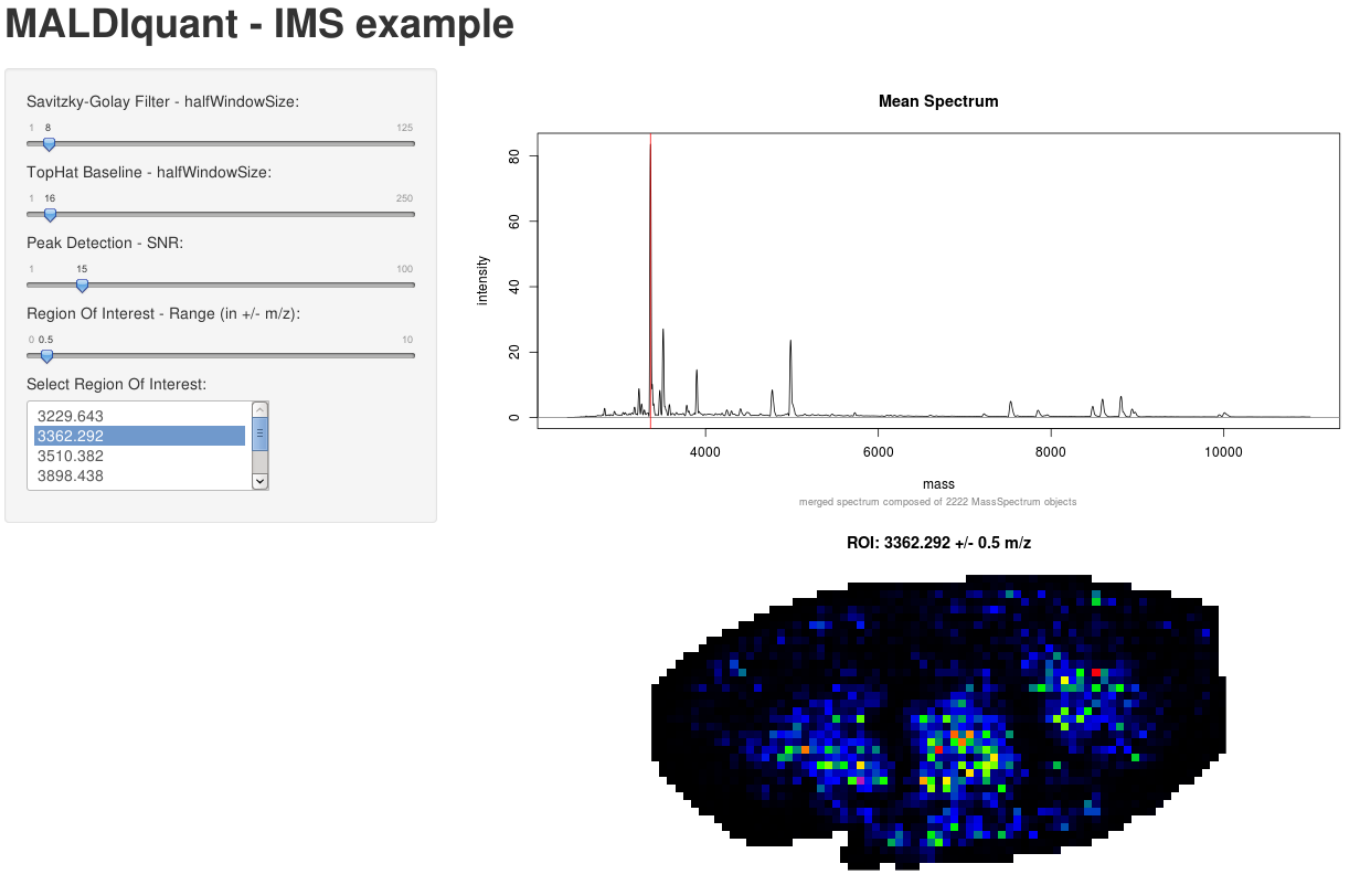
# IV. Conclusion and Outlook

# Summary

- We have developed MALDIquant, a comprehensive R package for analysis of mass spectrometry data, with focus on clinical diagnostics.
- MALDIquant is easy to use yet incorporates state-of-the art methods for preprocessing, calibration, quantification and visualization
- Multivariate analysis is best done on dichotomized peak data, using regularized categorical data analysis.

# Outlook: IMS

Imaging Mass Spectrometry (IMS):
combining MS data with spatial information



Many Thanks for Your Interest!

# Availability

http://strimmerlab.org/software/maldiquant/

```
## get newest MALDIquant/MALDIquantForeign directly from CRAN
install.packages(c("MALDIquant", "MALDIquantForeign"))
```

S. Gibb and K. Strimmer. 2012. MALDIquant: a versatile R package for the analysis of mass spectrometric data. *Bioinformatics* **28**:2270-2271.