Metabolomic and transcriptomic data analysis of Bioplastic-producing *Arabidopsis* **using R, exploRase and GGobi**

Iowa State University

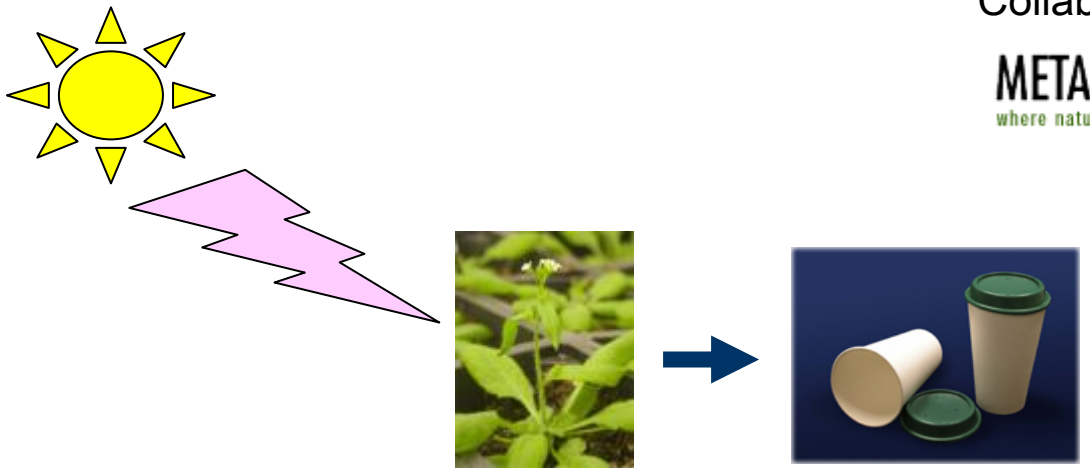
Suh-Yeon Choi, Michael Lawrence, Dianne Cook, Heike Hofmann, Lauralynn Kourtz, Kristi Snell, Basil J. Nikolau and Eve Syrkin Wurtele

Outline

- Introduction to bioplastic-producing plants
- Challenges in metabolomic data analysis
- Development of R based preprocessing tool for metabolomic data analysis
- Omics data analysis using exploRase

Goal

Collaboration with METABOLIX where nature performs™



Bioplastic-producing *Arabidopsis*

Metabolomics

Transcriptomics

Bioinformatics:
What limits bioplastic production in plants?

Optimize the production of bioplastic in plants

Procedure

Control plants

Bioplastic Producing plants

Extract Metabolites – metabolomic analysis

Extract RNA – transcriptomic analysis

Metabolomics data acquisition

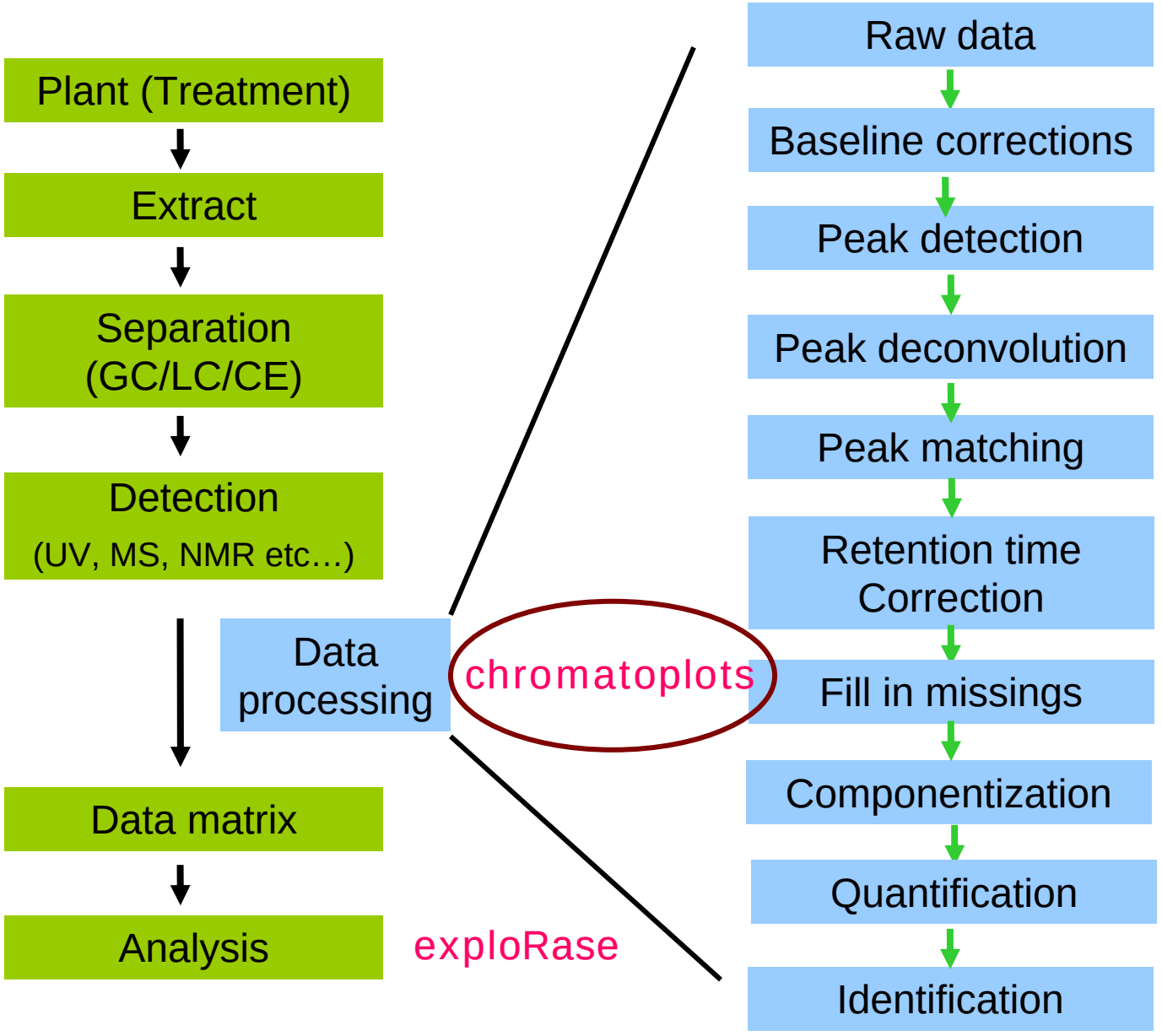


Image of the Raw Data

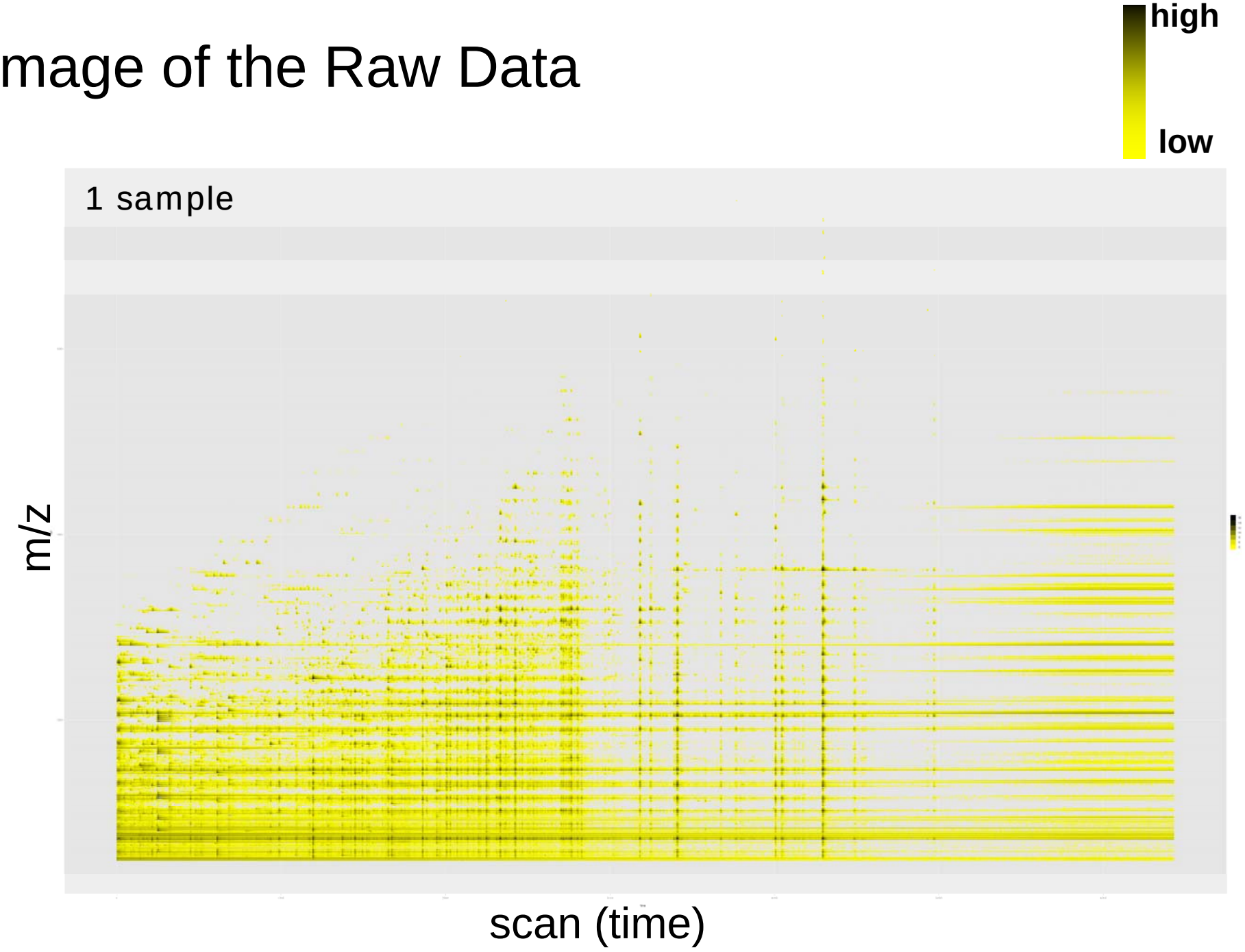
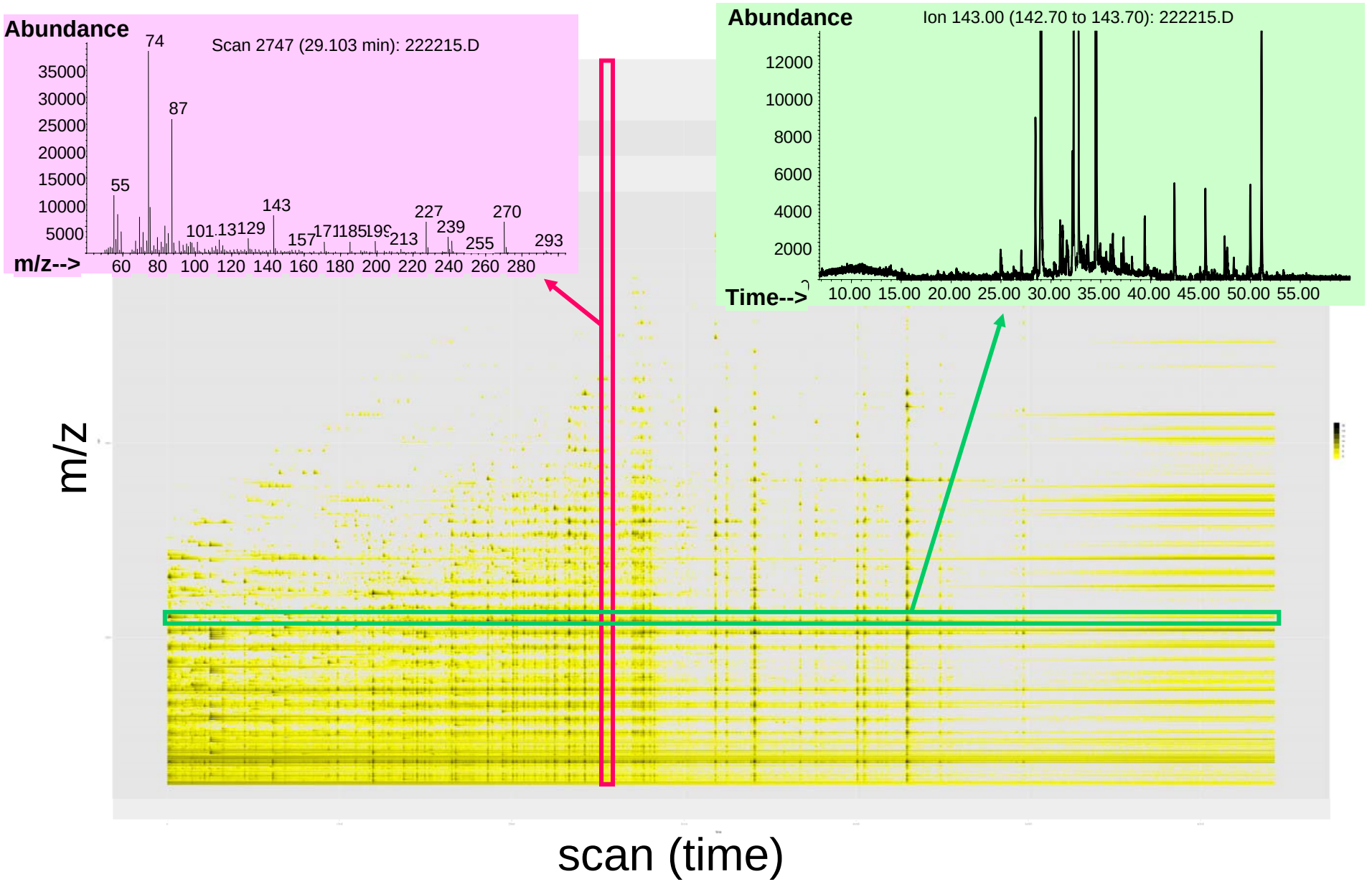
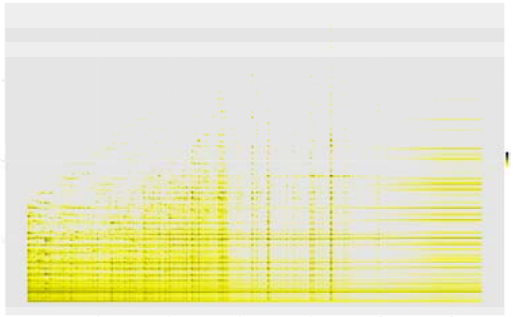


Image of the Raw Data



Goal of preprocessing of metabolomics data

- Identify components from peaks in intensity
- Label the components as specific metabolites



Data matrix

Metabolites	WT plant1	PHB plant1	WT plant2	...
malate	100	200	110	
citrate	50	25	60	

⋮

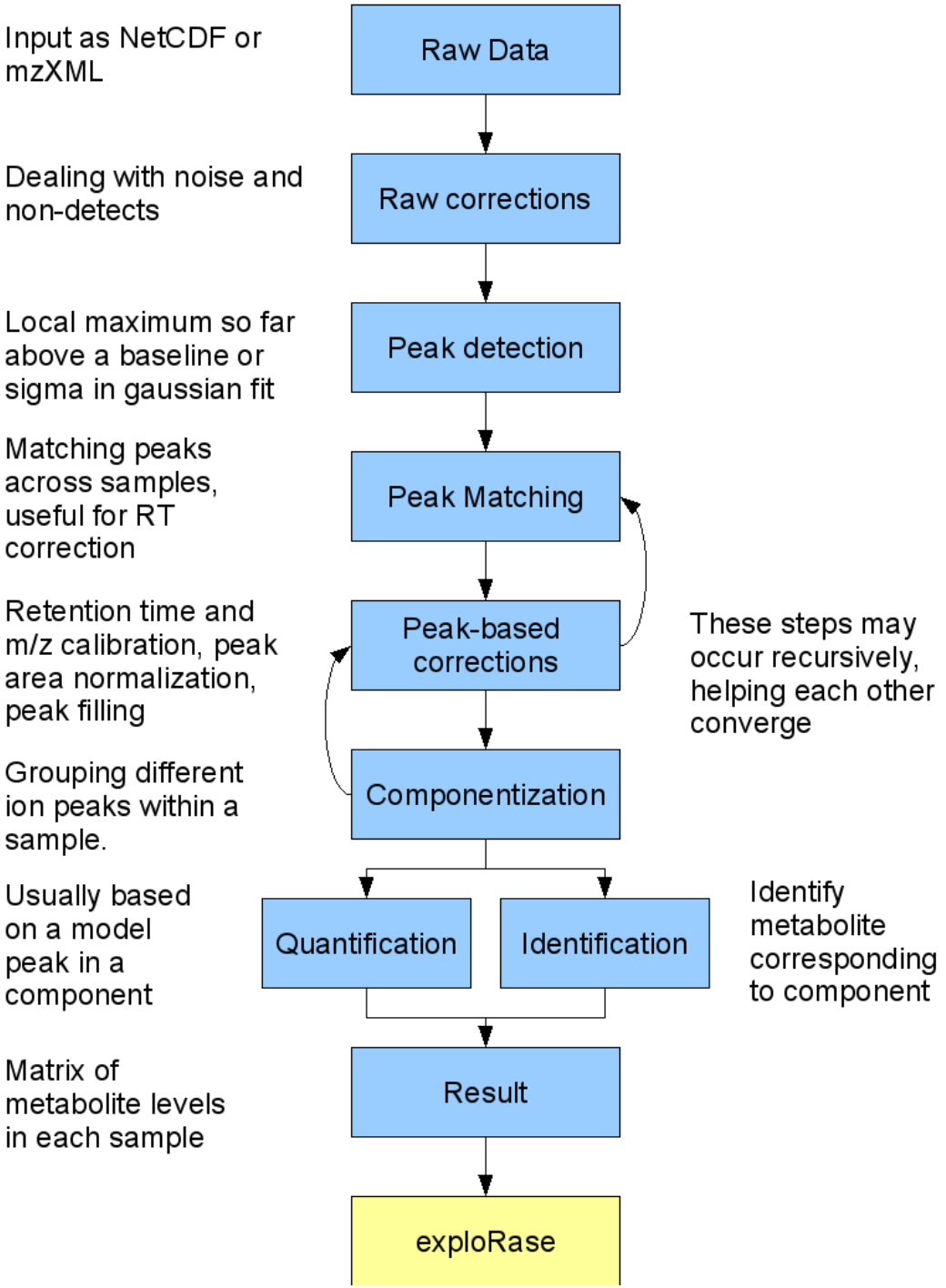
Limitation of existing tools

- Larger number of samples used with underlying experimental design
  - Most software analyze the data one by one
- Larger number of peaks of interest
  - More than ~300 metabolites detected per run
- No unified method
  - Each software uses their own algorithms
  - No comprehensive software
  - Commercial software ; cannot be shared by biologists
- Some bioinformatic tools have been developed (AMDIS, XCMS, MZMine, etc), but they are lacking
  - Limited diagnostics, especially interactive visualizations
  - Do not leverage experimental design

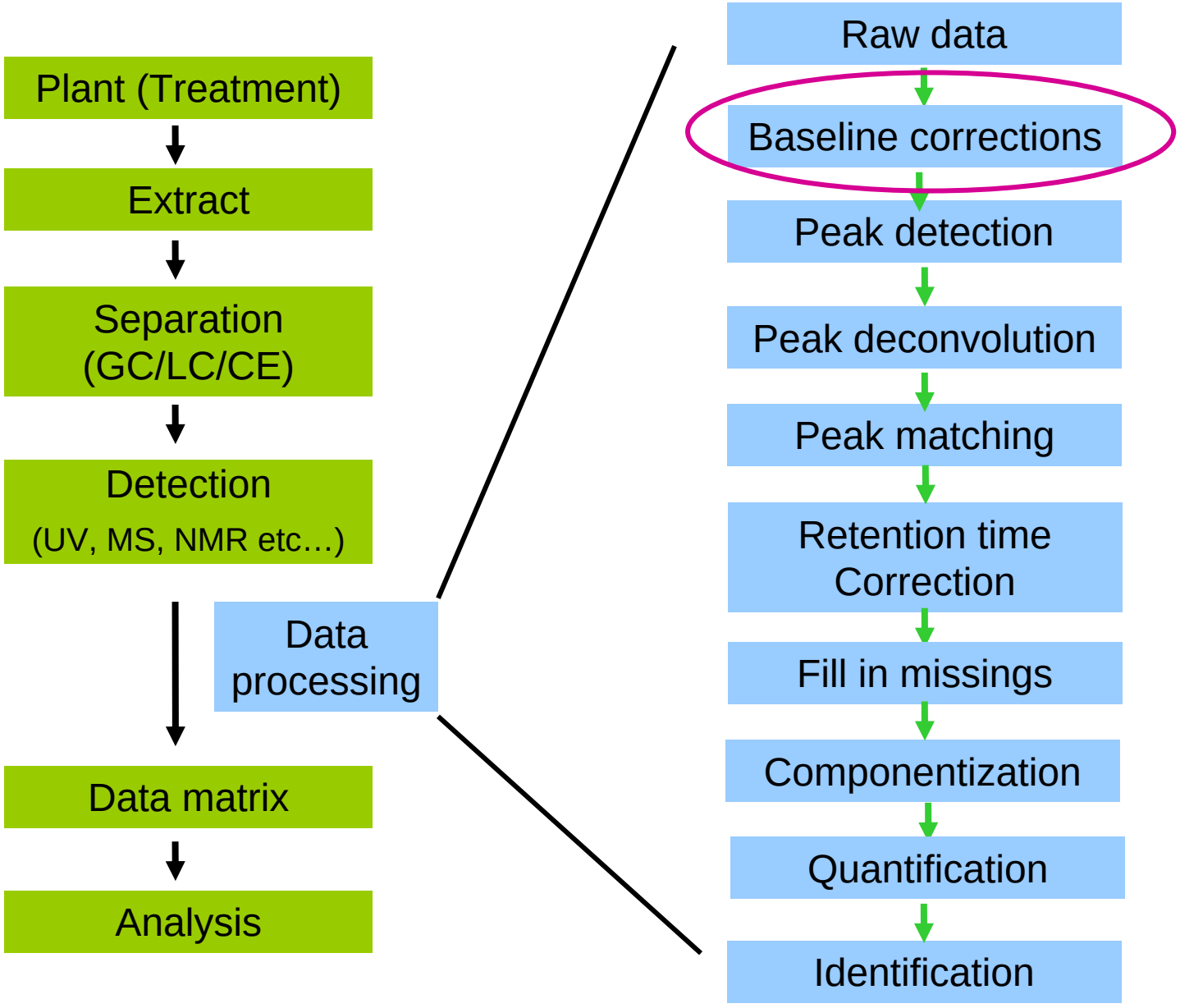
Features/goals of new tools

1. Automated data processing tool for large set of data (over hundreds samples..)
2. Have experimental design information in data processing
3. User inspection feature during processing (over replicates, etc…)
4. User friendly GUI wizard

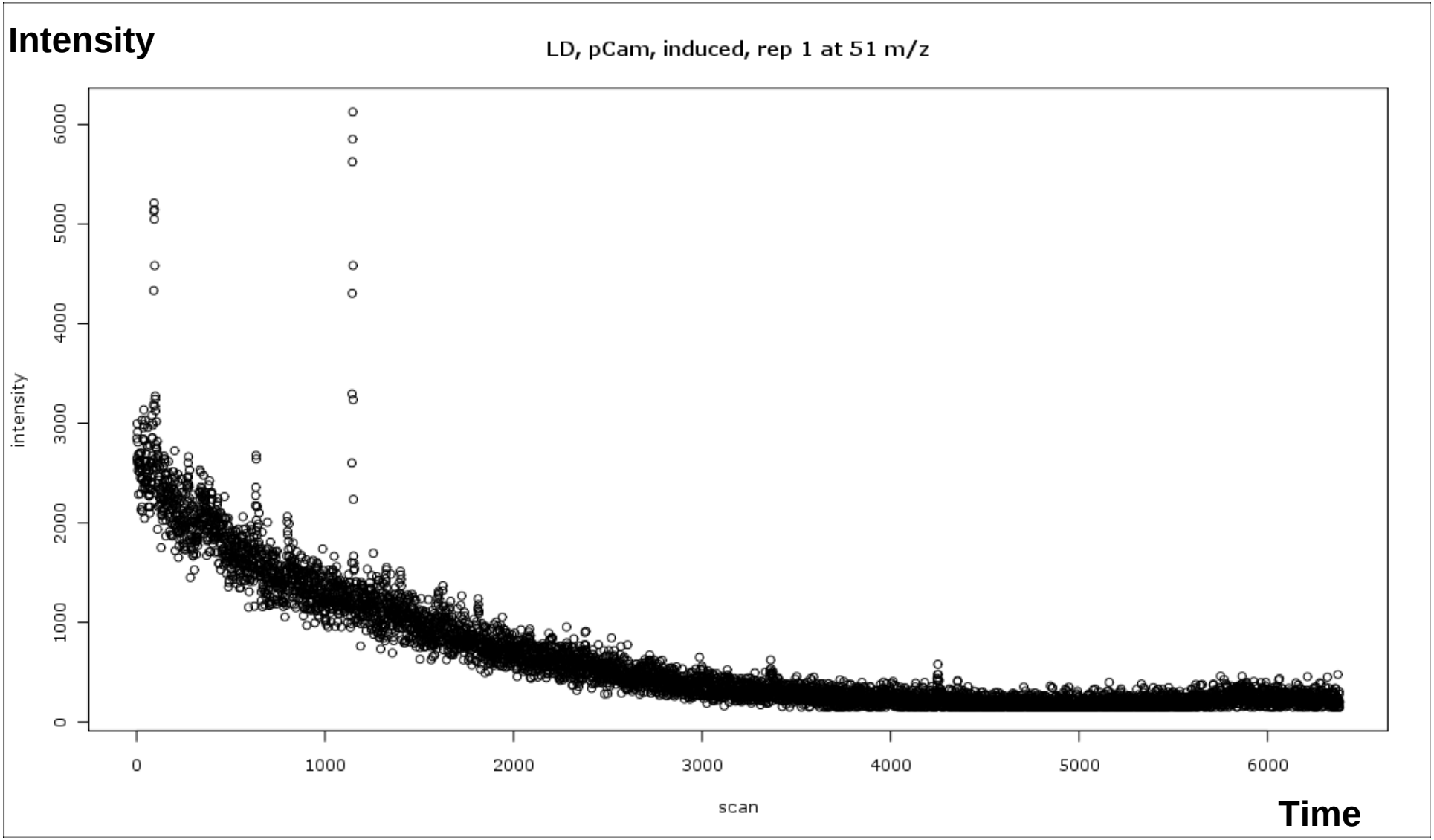
Proposed pipeline



Metabolomics data acquisition



Where is the Baseline?



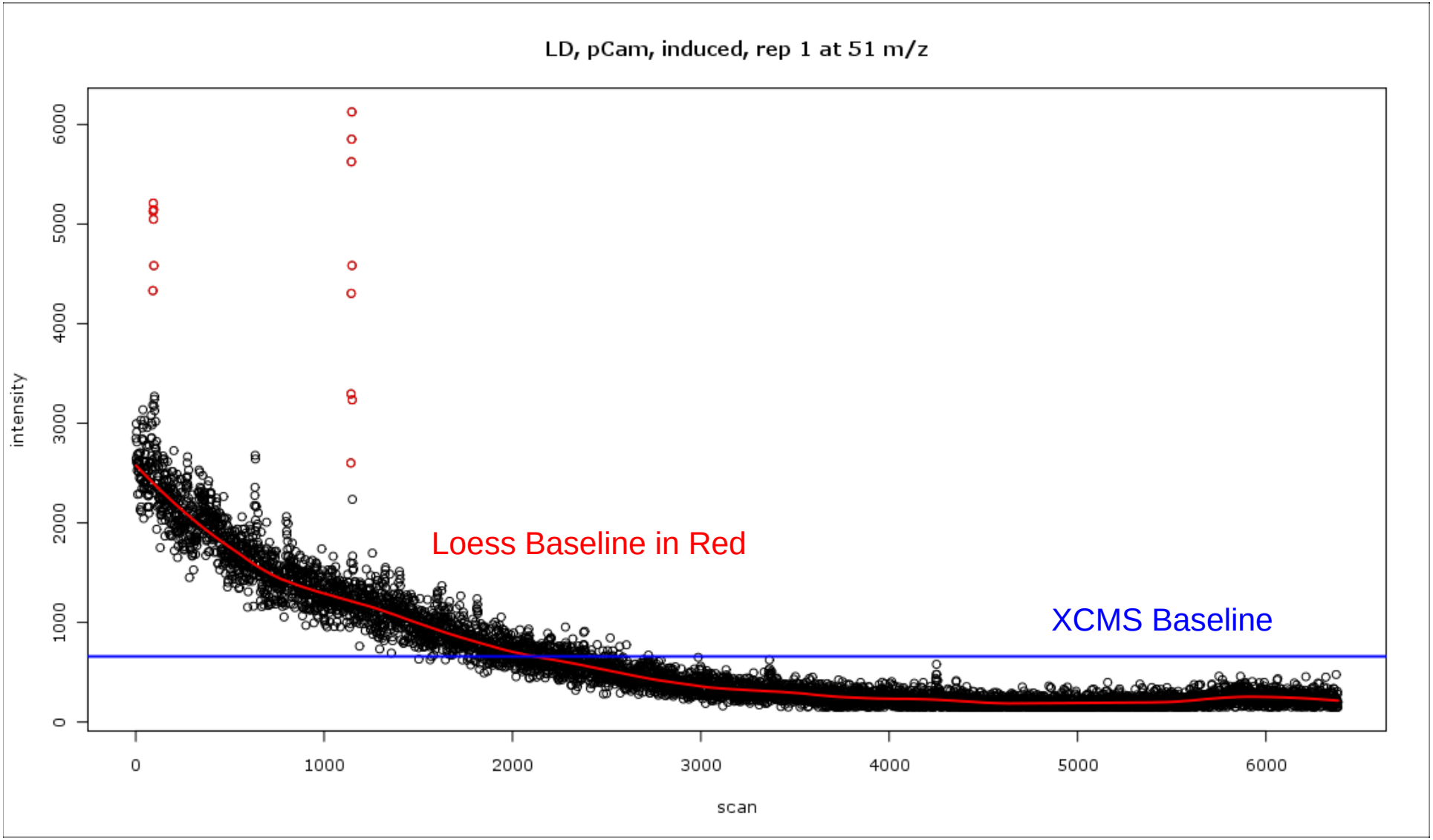
Background correction – existing solution

- AMDIS
  - baseline from a linear regression on all points below the median in the fitting region
  - not robust to high signal
- XCMS
  - Baseline from the second derivative of the filter translates the signal to curvature
  - subtracting linear background
- MathDAMP
  - RBE (Robust Baseline Estimation), a loess smoother that is weighted (Tukey biweight function)
  - robust to outliers (peaks)

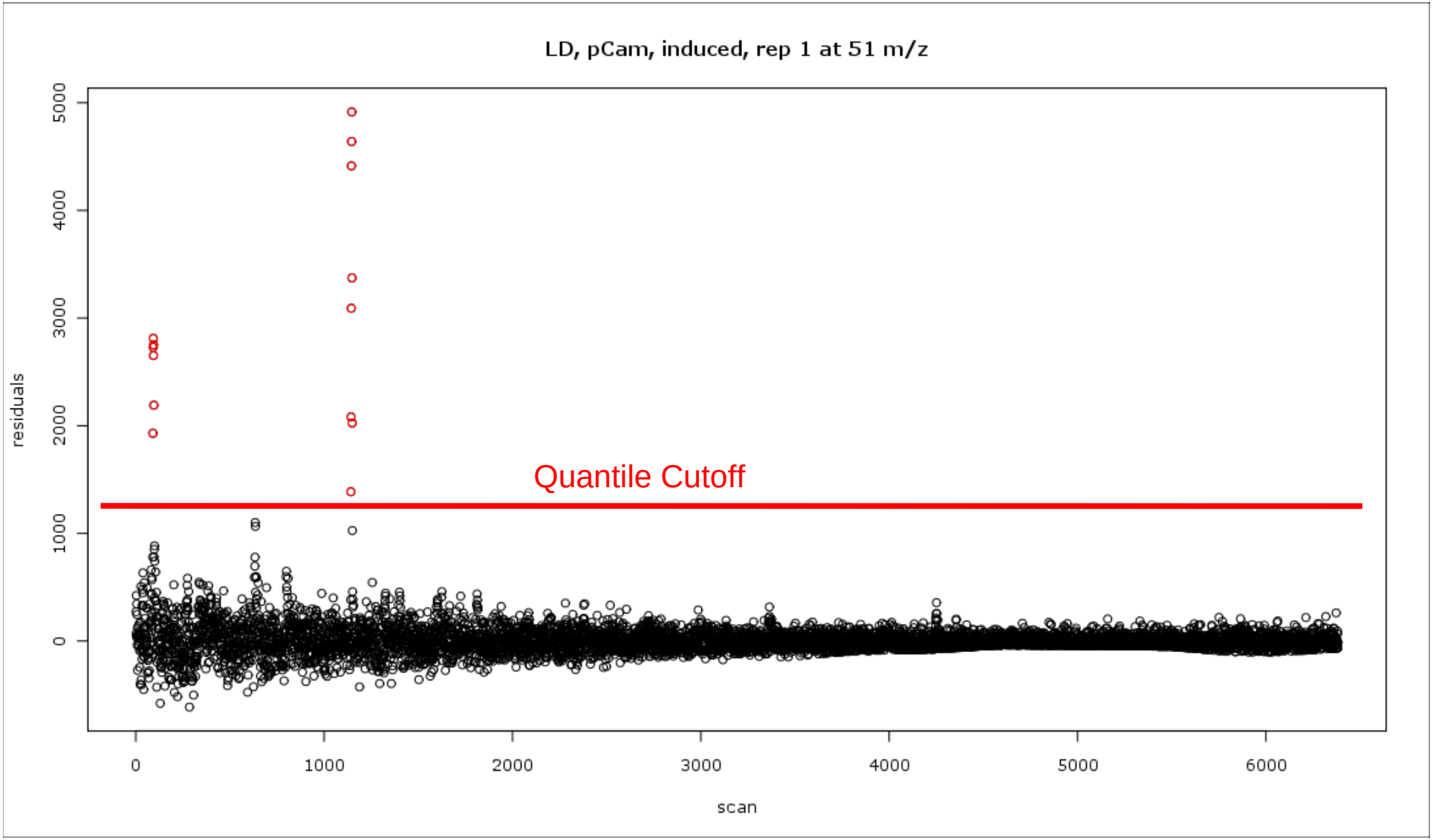
Background correction - Loess Baseline Subtraction

- Approach used in MathDAMP
- Fit loess model to the raw profile.
- Needs to be robust to avoid fitting the peaks.
- Iterate loess fits, weighting cases with positive residuals by the Tukey biweight function (Ruckstuhl et al., 2001).

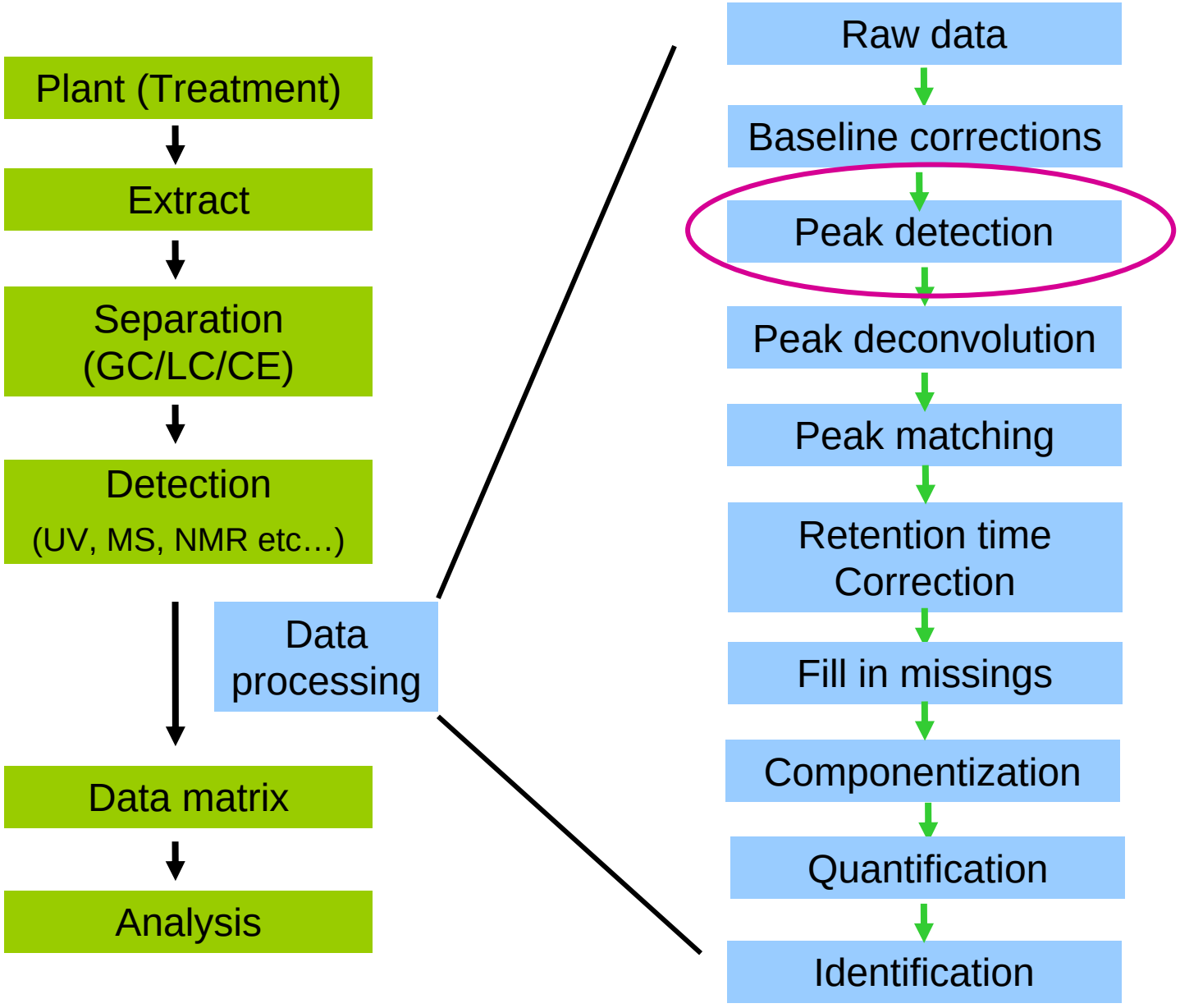
Loess Baseline Fit



After Baseline Subtraction

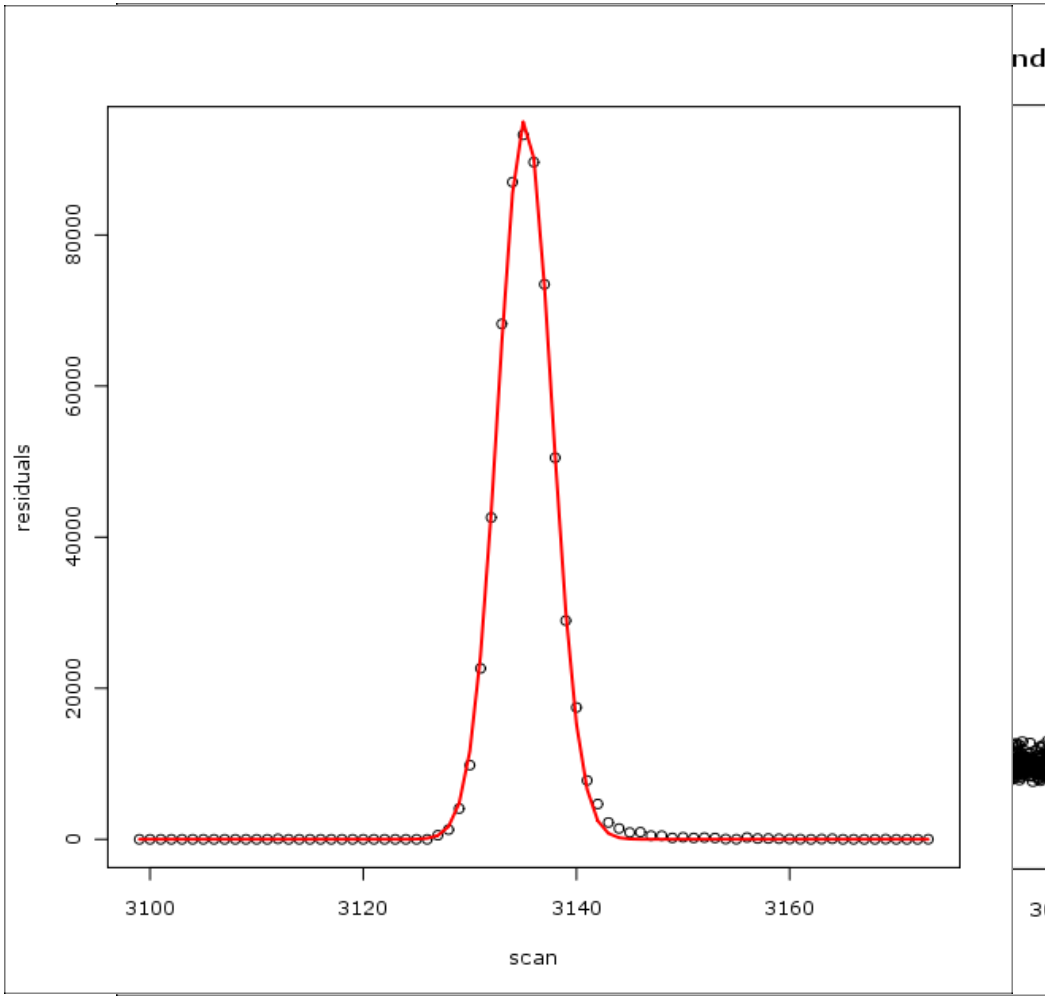


Metabolomics data acquisition



Peak Detection

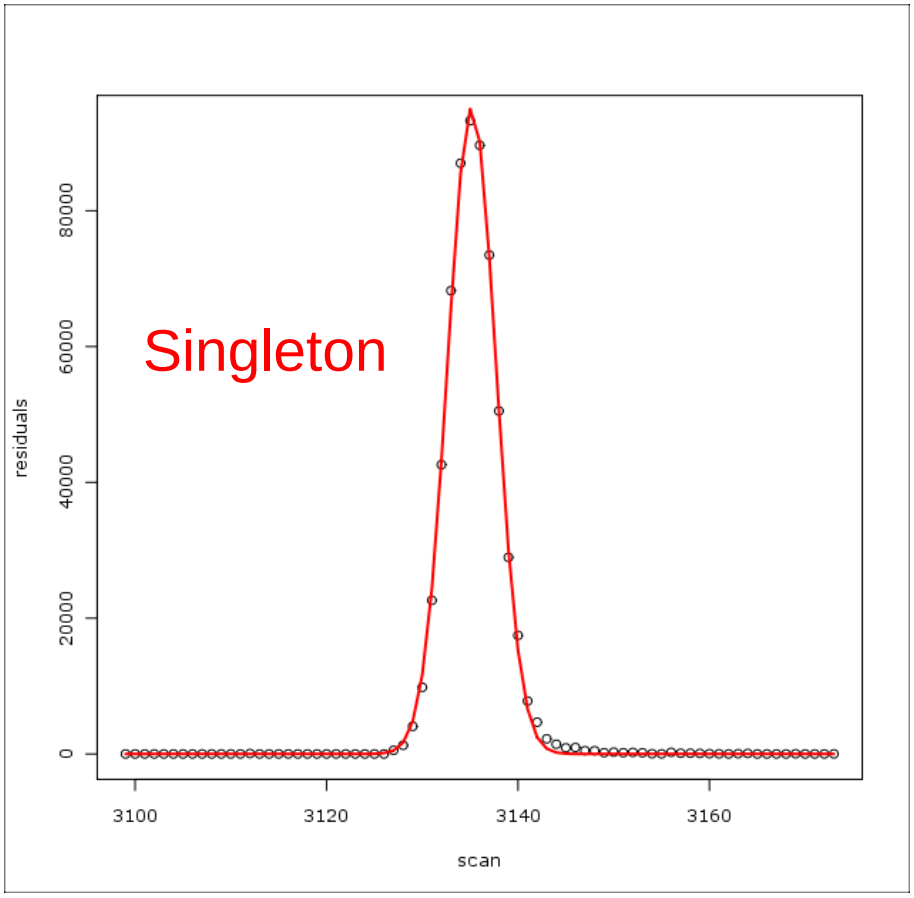
- Peaks are local maxima above some cutoff and exceeding adjacent minima by some threshold.
- Cutoff is a global quantile of the residuals.
- The threshold is a multiple of the standard deviation of the (residual) intensities.
- Similar approach to AMDIS.

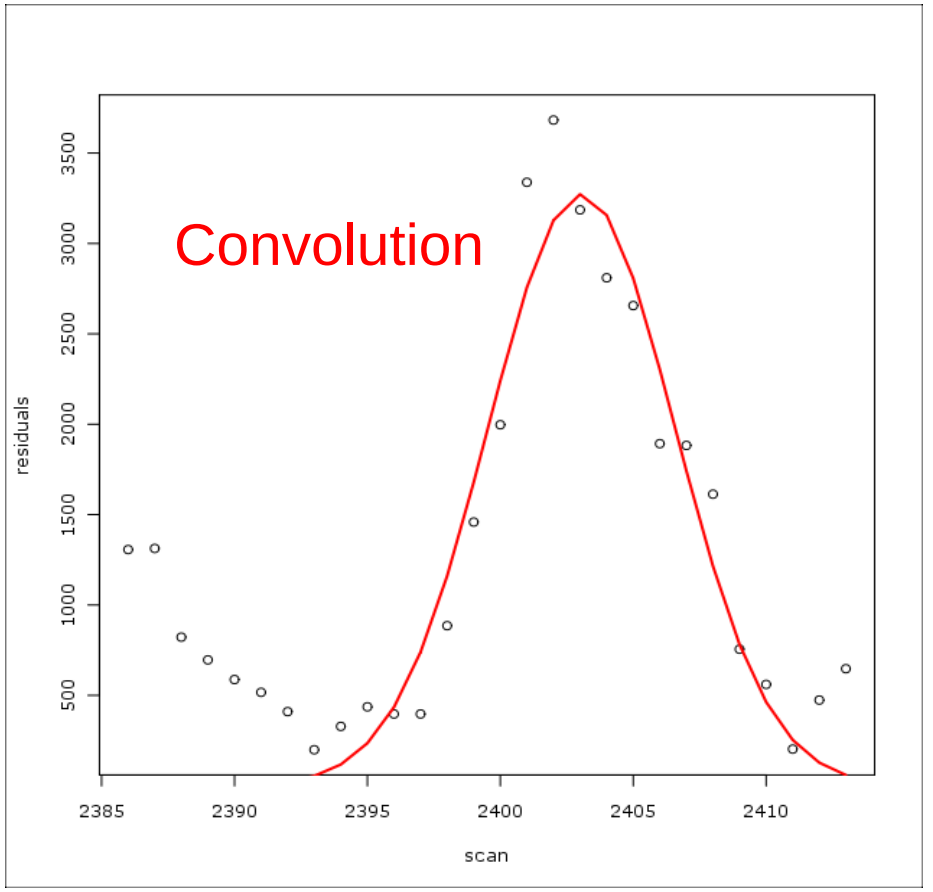


Considering the Peak Shape

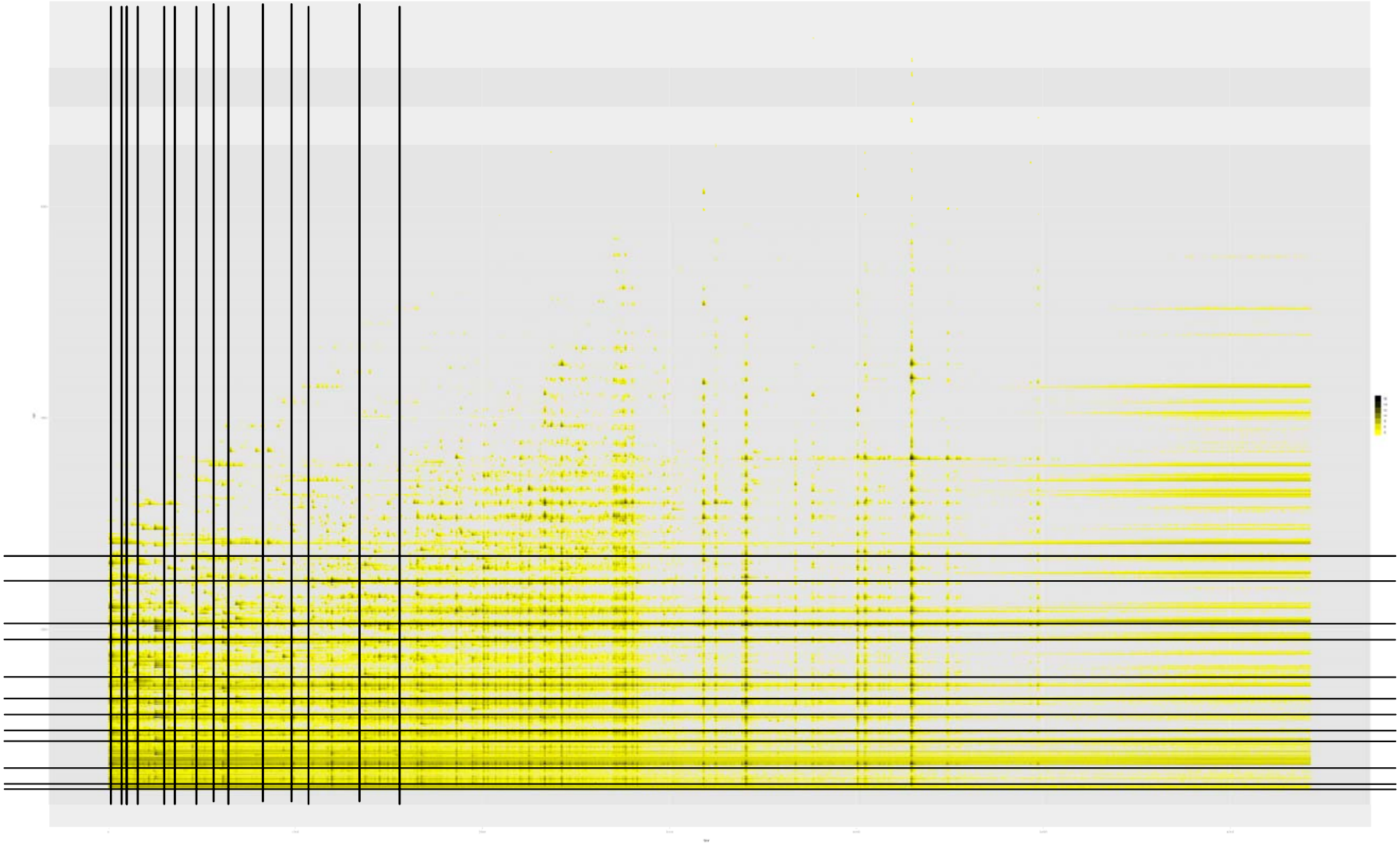
- We expect a peak to have a gaussian shape, so we fit a gaussian function to the neighborhood around each maxima.
- Neighborhoods are not allowed to overlap.
- Fits with extremely large sigma are discarded.
- About 4000 peaks detected per sample.

Example Peak Fits

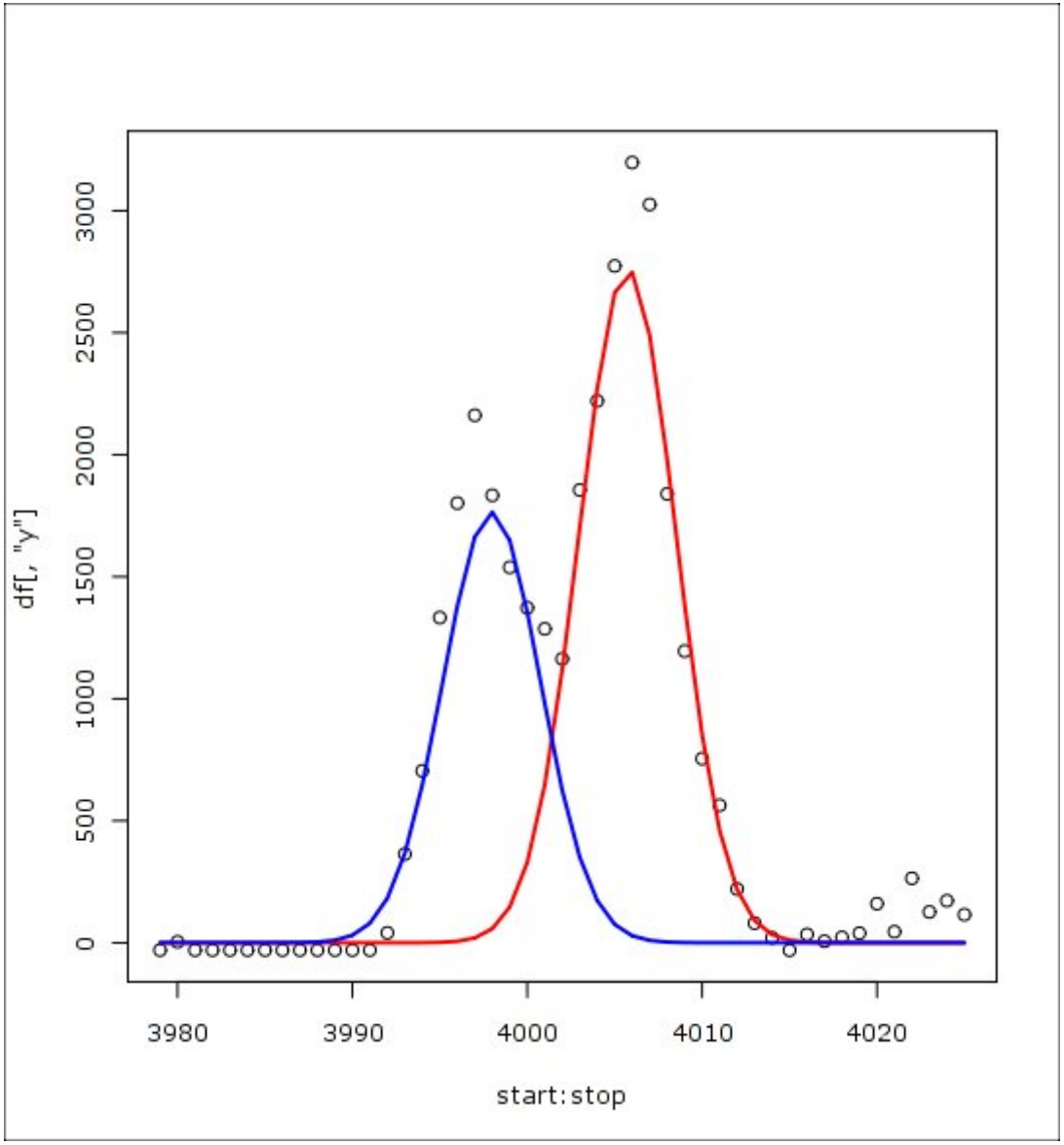




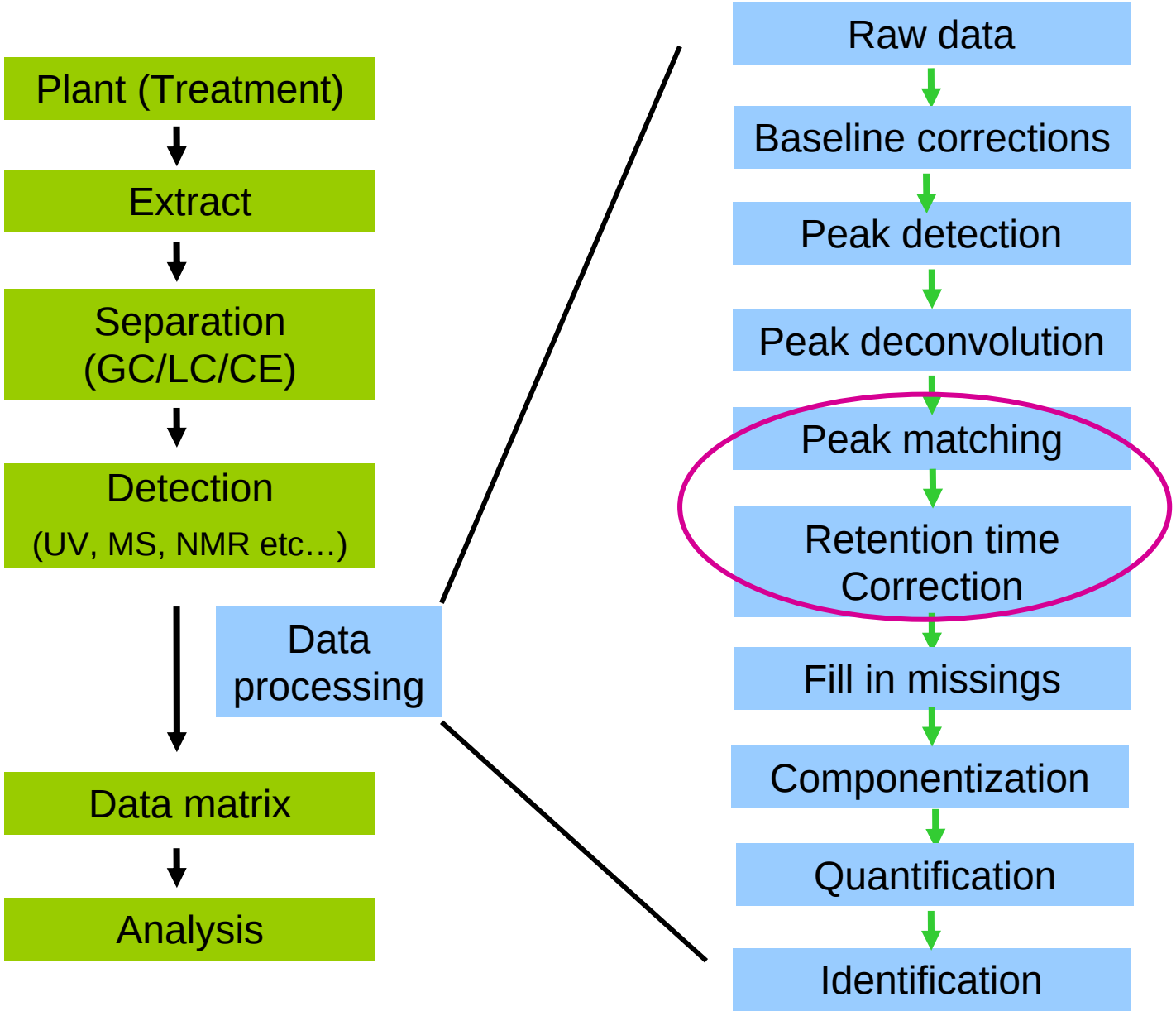
Slicing and Dicing for the Peaks



Convoluted peak detection

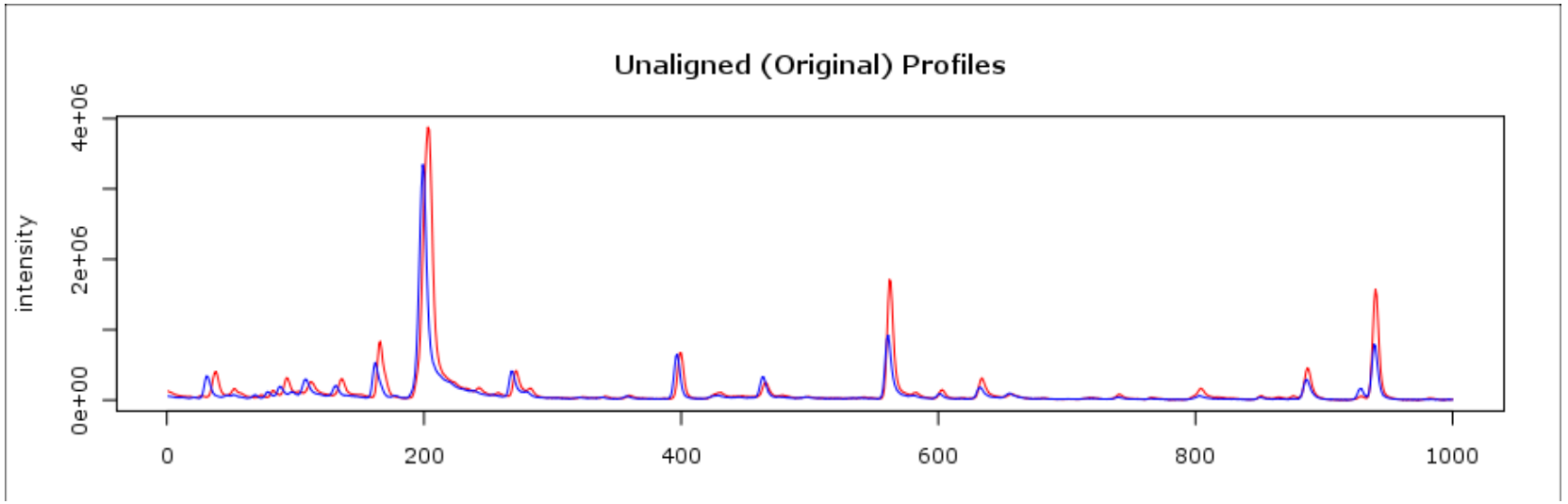


Metabolomics data acquisition



Comparing Samples

- To compare, they need to be aligned.
  - The m/z is assumed to be relatively stable.
  - Retention time likely requires correction, due to instability of the column across runs.
- Peaks between replicates should be consistent.



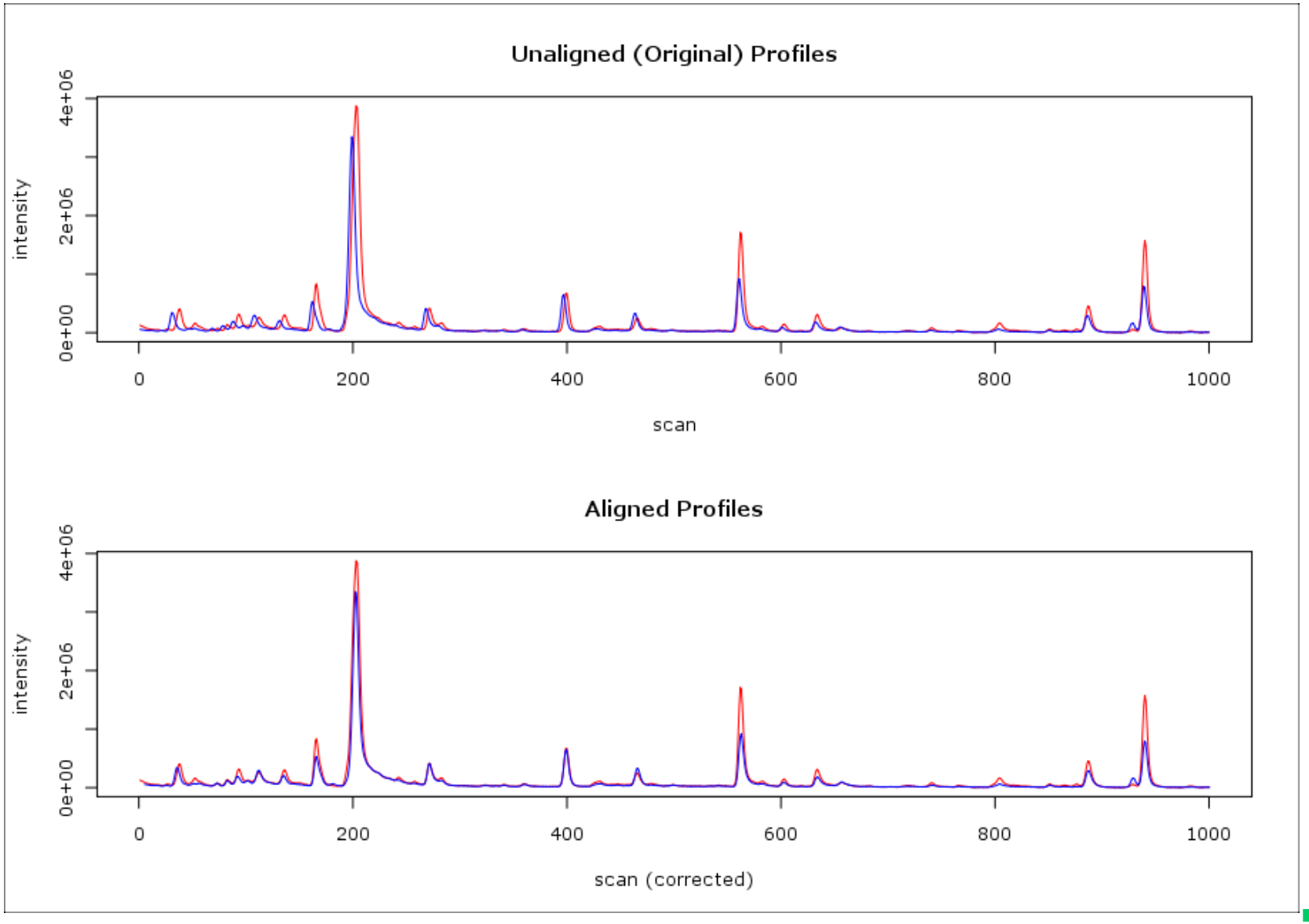
Retention time correction – existing solutions

- AMDIS – RI based (not precise)
- METIDEA – AMDIS + selective ion matching
- MetAlign – selective ion matching + back and forth..
- XCMS – fitting by Gaussian density estimation function

Retention time correction

- Consider the peaks in the TIC (Total Ion Count) profile, the sum over m/z (Krebs et al., 2006).
- Greedily match by the pairwise correlation between spectral intensity vectors
- Fit robust loess to ignore outliers (mismatches).
- Visually explore results using rggobi.

RT Correction Results

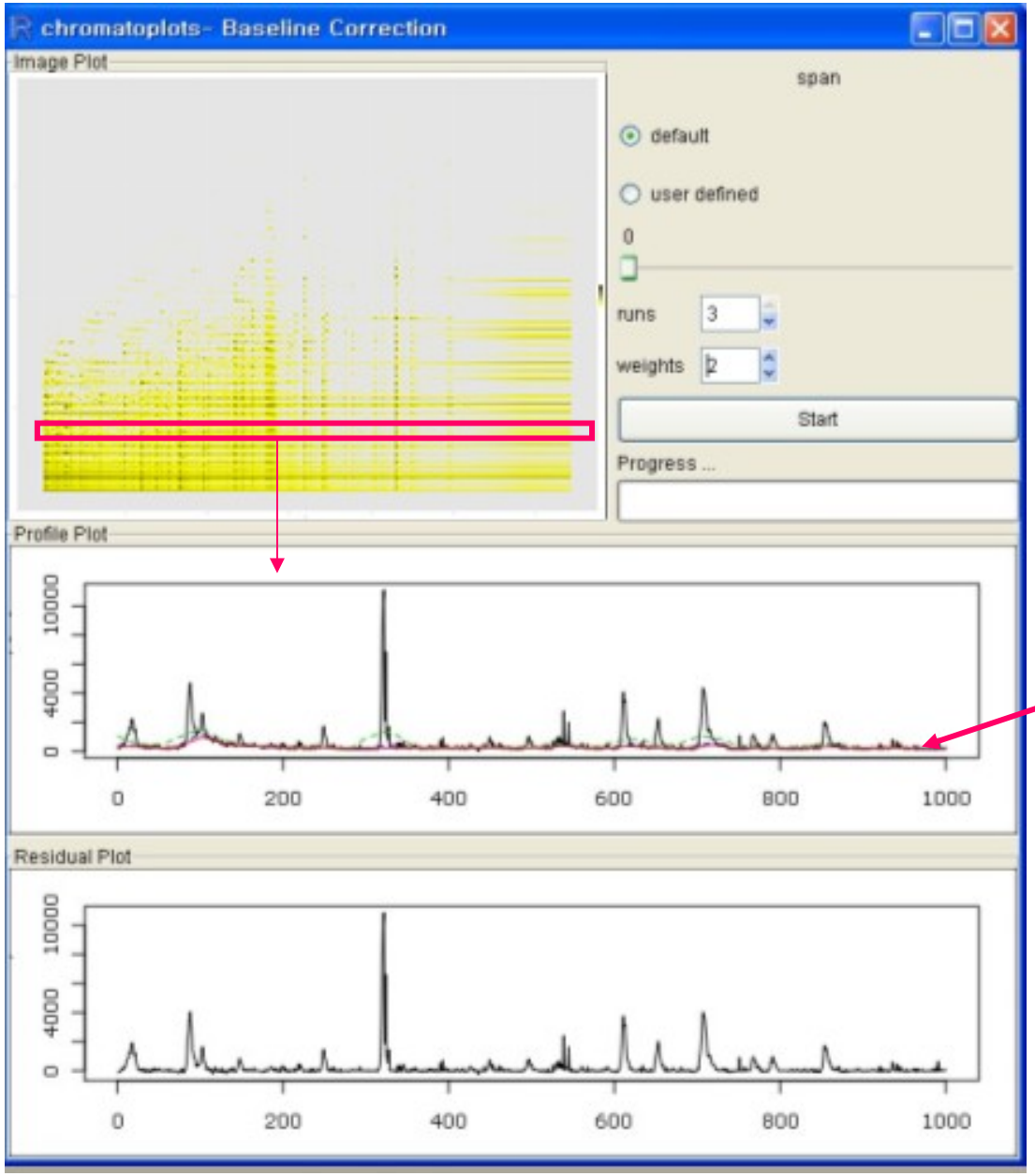


GUI : chromatoplot (baseline correction)

Raw image

Profile plot

Residual plot



Option windows ; User can select

Baseline from loess fit

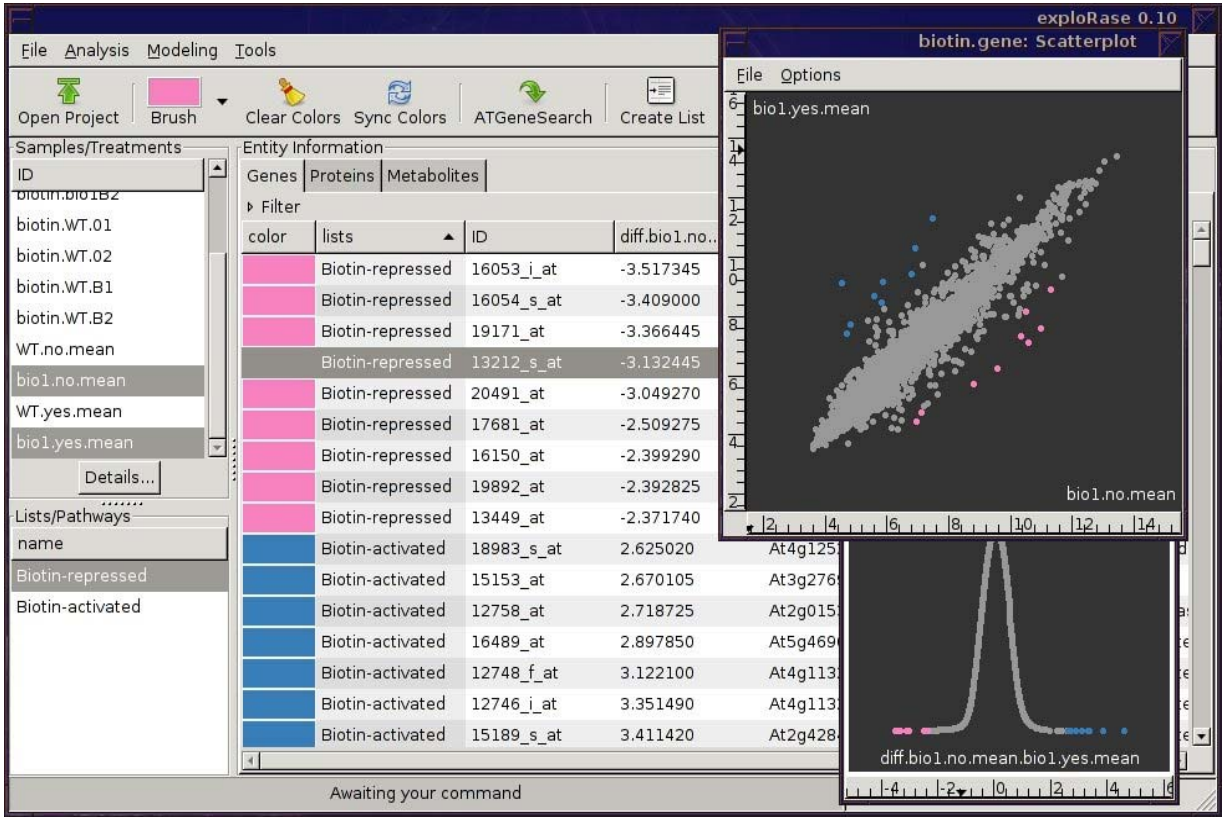
Baseline corrected

Next Steps

- Deconvolution of the peaks
- Matching the peaks across data set
- Identify and quantify the metabolites
  - A scriptable implementation of the methods
  - A biologist-accessible GUI
  - Plenty of interactive graphics for diagnostics
  - Integration with Bioconductor (xcms, MassSpecWavelet)

Metabolites	WT plant1	PHB plant1	WT plant2	...
malate	100	200	110	
citrate	50	25	60	

exploRase : Omics data analysis tool



Metabolic Network

- R: http://www.r-project.org/
- RGtk2: http://www.ggobi.org/RGtk2/
- rggobi: http://www.ggobi.org/rggobi/
- ggobi: http://www.ggobi.org/
- exploRase : http://www.metnetdb.org/MetNet_exploRase.htm
- chromatoplots : not available yet

Acknowledgement

- **Department of Statistics**
  - Prof. Dianne Cook
  - Prof. Heike Hofmann
  - Michael Lawrence
  - Dr. Eun-Kyung Lee
- **Metabolix. Inc.**
  - Dr. Lauralynn Kourtz
  - Dr. Kristi Snell
- **Department of Genetics, Developmental and Cell Biology**
  - Prof. Eve Wurtele
  - Suh-Yeon Choi
- **Department of Biochemistry, Biophysics and Molecular Biology**
  - Prof. Basil Nikolau
  - Dr. Wenxu Zhou
- **W.M. Keck Metabolomics Research Lab**
  - Dr. Ann Perera