May Institute 2017

Computation and statistics for mass spectrometry and proteomics

Fundamentals of non-targeted proteomics and metabolomics

Oliver Kohlbacher

University of Tübingen and

MPI for Developmental Biology

KohlbacherLab.org | @okohlbacher

MAX-PLANCK-GESELLSCHAFT

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Today's Schedule

Tuesday 5/2/2017	
8:00 AM	Bring your own data or Skyjam
9:00 AM	Lecture: Label-free quantitative proteomics.
10:30 AM	**Refreshments**
11:00 AM	Hands-on: Label-free quantification workflows
12:30 PM	**Lunch Break**
1:30 PM	Lecture: Introduction to non-targeted metabolomics.
2:30 PM	Hands-on: Metabolite profiling workflow.
3:00 PM	**Refreshments**
3:30 PM	Hands-on: Differential quantification of metabolites, visualization, report generation
5:00 PM	Questions and practice with own data
6:00 PM	**Adjourn**

Metabolome vs. Proteome

- **Size and complexity** of the metabolome still largely unknown
- Similar to protein sequence databases, there are also **metabolite databases** listing all known metabolites (usually contains **tens of thousands** of metabolites)
- Differences between **proteome and metabolome**
  - Metabolites belong to wider ranger of chemical compound classes (lipids, sugars, amino acids)
  - Proteins have a more homogenous chemistry (20 proteinogenic amino acids)
  - Metabolites can have complex structures that require a structural formula for a comprehensive description
  - Proteins have a simple, linear structure that can be represented by a sequence
  - Metabolites are **light**: average metabolite mass a 100-300 Da
  - Proteins are **heavy**: median protein length around 300-500 aa, about 40,000 Da molecular weight

Metabolites

- Metabolites comprise a heterogeneous set of biomolecules: all small molecules in a system excepting salts and macromolecules (proteins, long peptides, RNA, DNA)
- Lipids and sugars are metabolites as well
- There are separate fields dealing with lipids and sugars (lipidomics, glycomics), techniques are very similar

Examples:

Metabolite	mol l^{-1}	Metabolite	mol l^{-1}	Metabolite	mol l^{-1}
Glutamate	9.6×10^{-2}	UDP-glucuronate (51)	5.7×10^{-4}	*N*-Acetyl-ornithine (79)	4.3×10^{-5}
Glutathione	1.7×10^{-2}	ADP	5.6×10^{-4}	Gluconate (80)	4.2×10^{-5}
Fructose-1,6-bisphosphate	1.5×10^{-2}	Asparagine (52)	5.1×10^{-4}	Malonyl-CoA (81)	3.5×10^{-5}
ATP	9.6×10^{-3}	α-Ketoglutarate	4.4×10^{-4}	Cyclic AMP (82)	3.5×10^{-5}

Extracted from Bennett et al.: some of the most abundant small molecules in *E. coli*

Bennett et al. *Nature Chemical Biology*.
Bennett et al. *Nature Chemical Biology*. 2009

Metabolomics Techniques

- Fundamentally two types of approaches
  - **Targeted metabolomics**
    - Identify only a well-defined subset of metabolites, but those with higher accuracy (hundreds?)
    - All of these metabolites can then be identified
  - **Non-targeted metabolomics (metabolic profiling)**
    - Try to see as much of the metabolome as possible (thousands and more)
    - Majority of metabolites can be seen
    - Only a small fraction will be identified
- Similarly, there is also targeted and non-targeted proteomics
- In proteomics, the identification problem is less difficult, though, which is why this distinction is more relevant in metabolomics (where identification is much harder)

Metabolite Quantification

- **Label-free proteomics** is similar to **non-targeted metabolomics**
- Overall workflow is identical
  - Feature finding
  - Map alignment
  - Feature linking
- Feature-finding approaches are algorithmically **similar** to those used in proteomics
  - Mass traces usually at the heart of the algorithm
  - Assembly into features can be done similarly
- However, there are some **differences**
  - Isotopic patterns differ from proteomics (no averagine!)
  - Mass range and charge states are different

Feature Finding – Terms

Map:

Two-dimensional data set (RT, m/z) containing the MS signal from one LC-MS run.

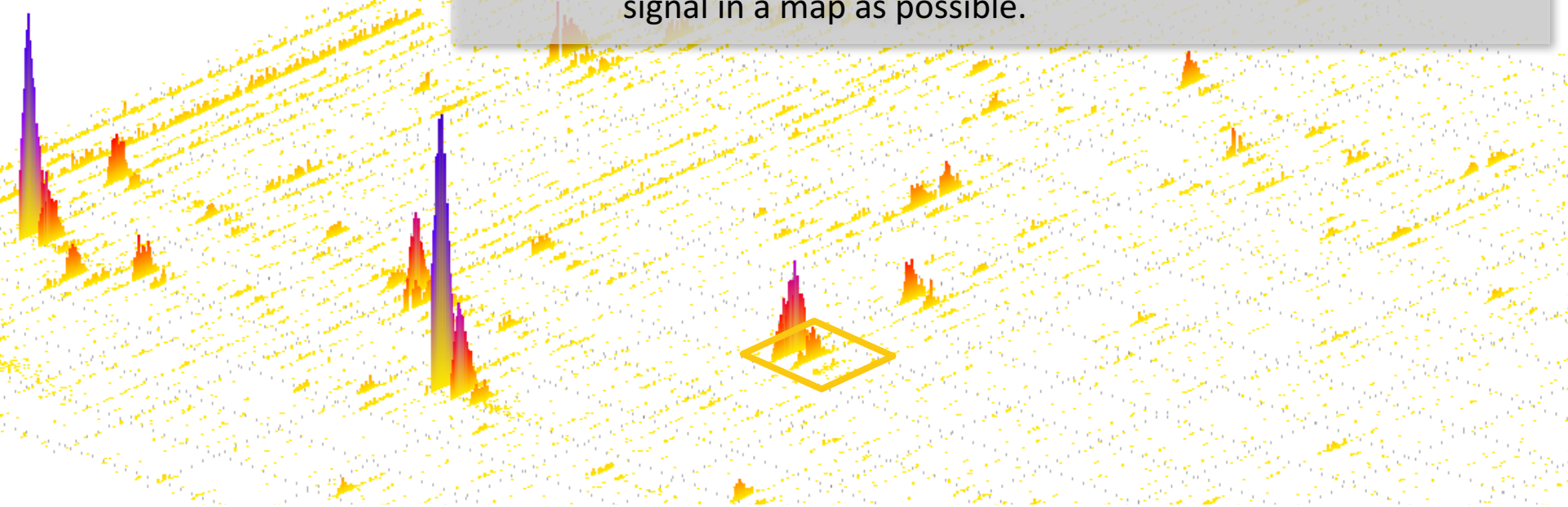
Feature:

The sum of all the MS signals caused by the same analyte in a specific charge state.

Different adducts will result in distinct features. Primarily characterized by RT, m/z, charge, intensity.

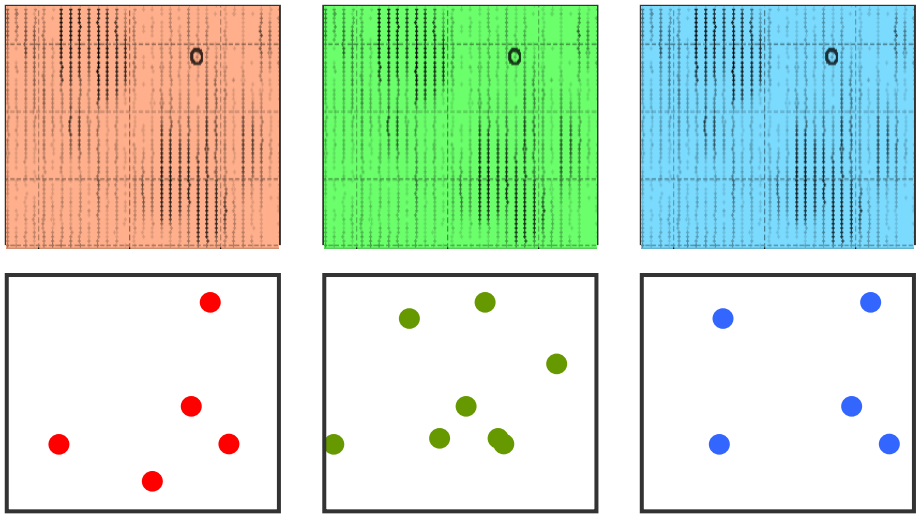
Feature finding:

Finding the set of features explaining as much of the signal in a map as possible.



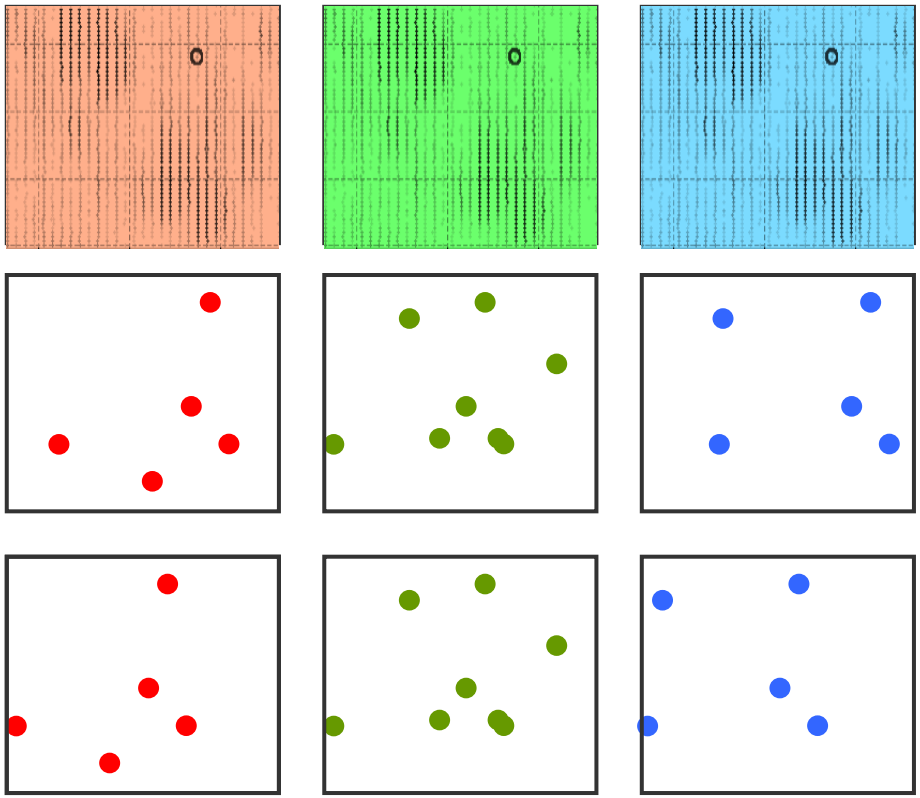
Metabolic Profiling

1. **Find** features in all maps



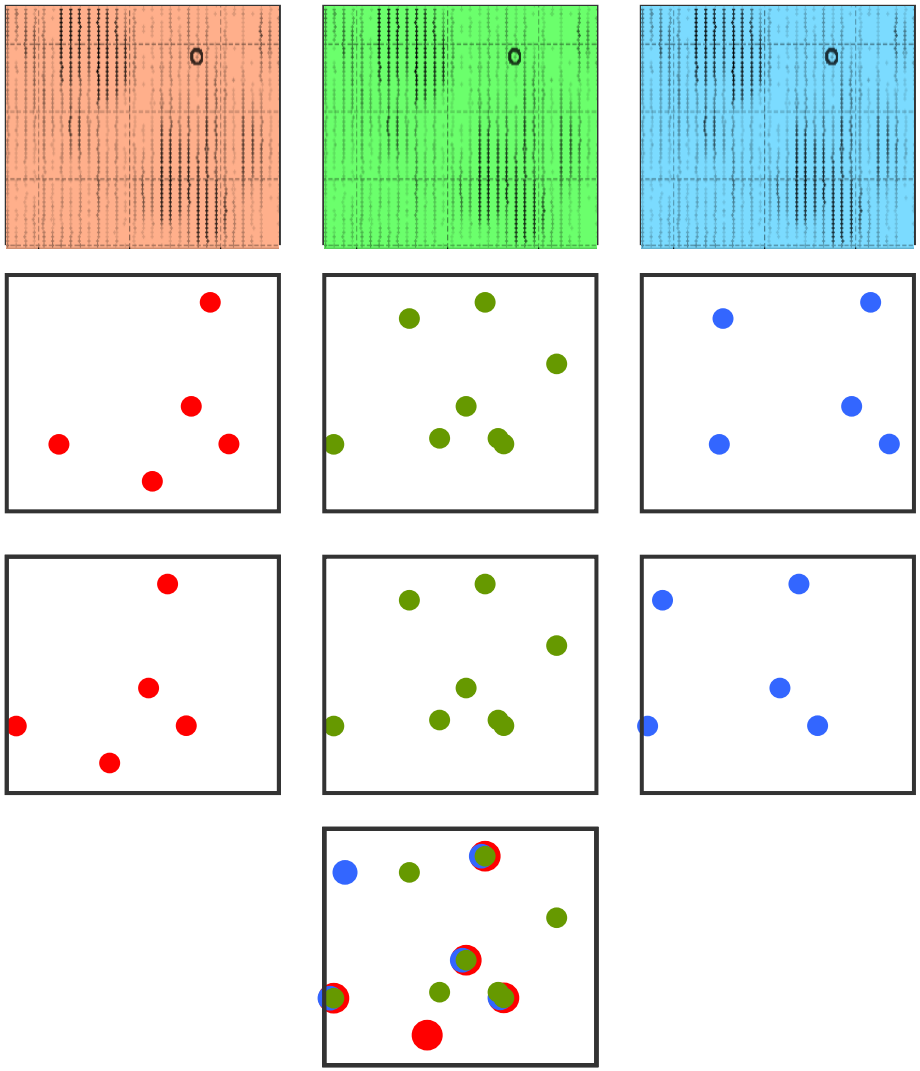
Metabolic Profiling

1. **Find** features in all maps
2. **Align** maps



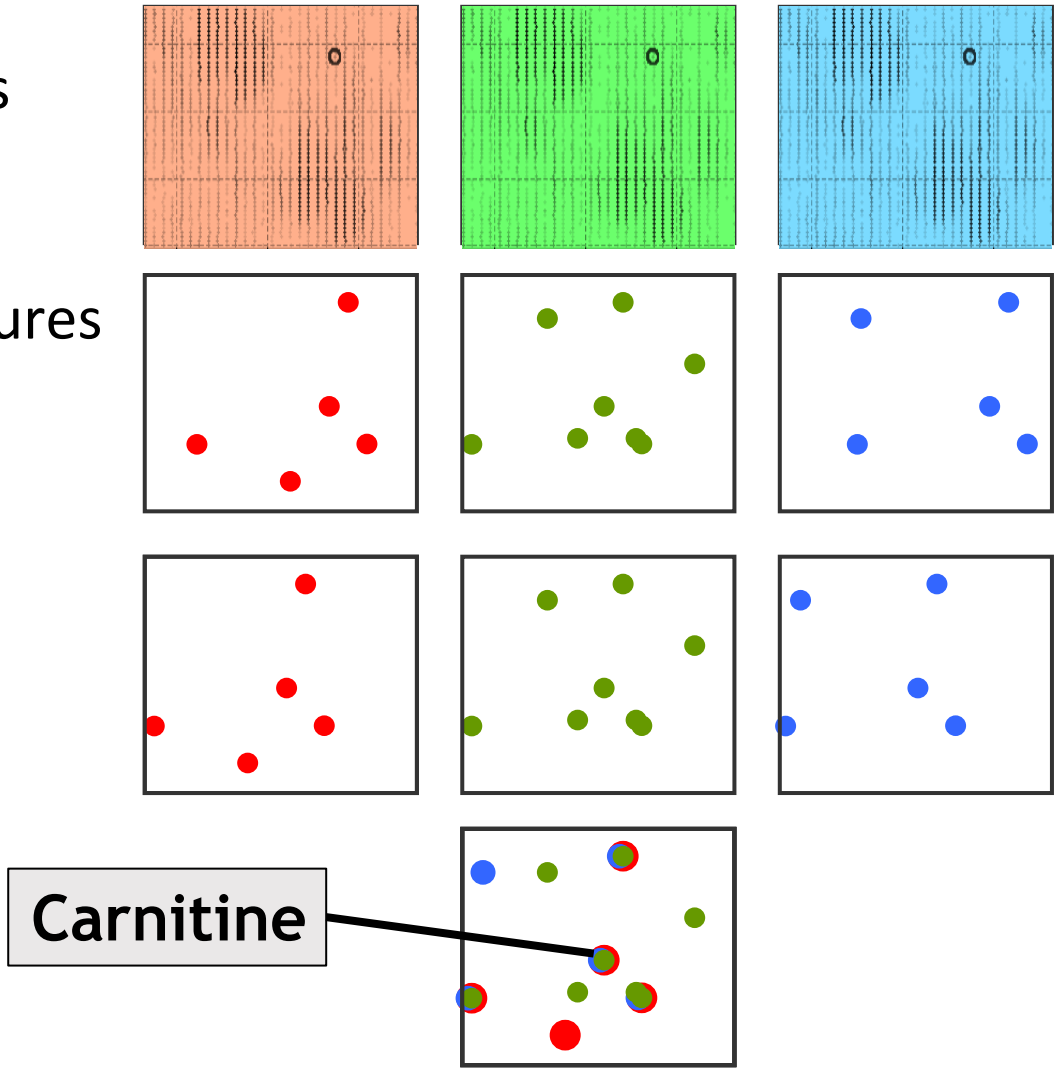
Metabolic Profiling

1. **Find** features in all maps
2. **Align** maps
3. **Link** corresponding features



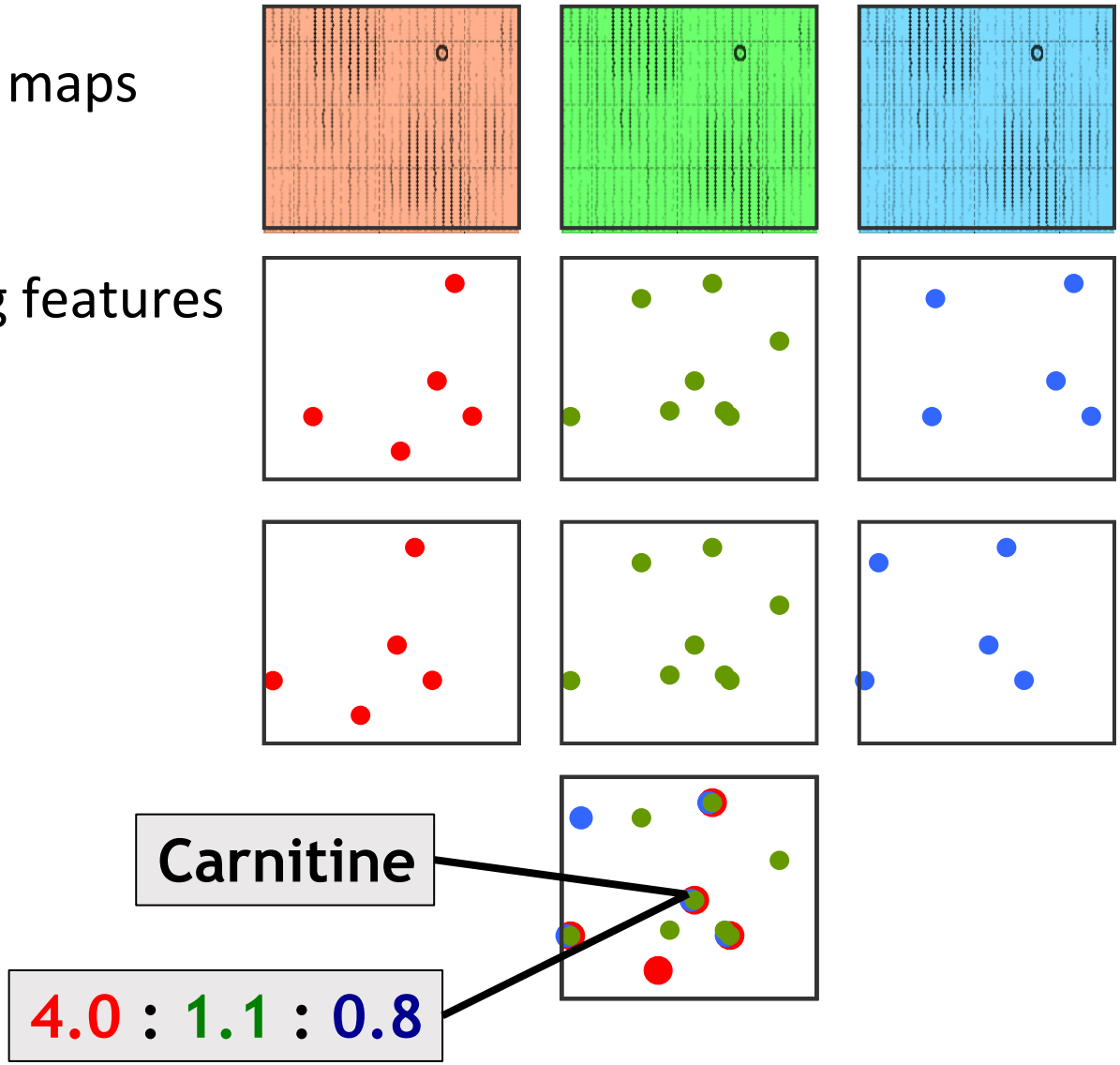
Metabolic Profiling

1. **Find** features in all maps
2. **Align** maps
3. **Link** corresponding features
4. **Identify** features



Metabolic Profiling

1. **Find** features in all maps
2. **Align** maps
3. **Link** corresponding features
4. **Identify** features
5. **Quantify**



Feature Finding in MTX – Issues

- Proteomics feature finding algorithms make extensive use of the **averagine** hypothesis: peptides have a well-defined average composition
- Metabolites are chemically much more diverse than peptides

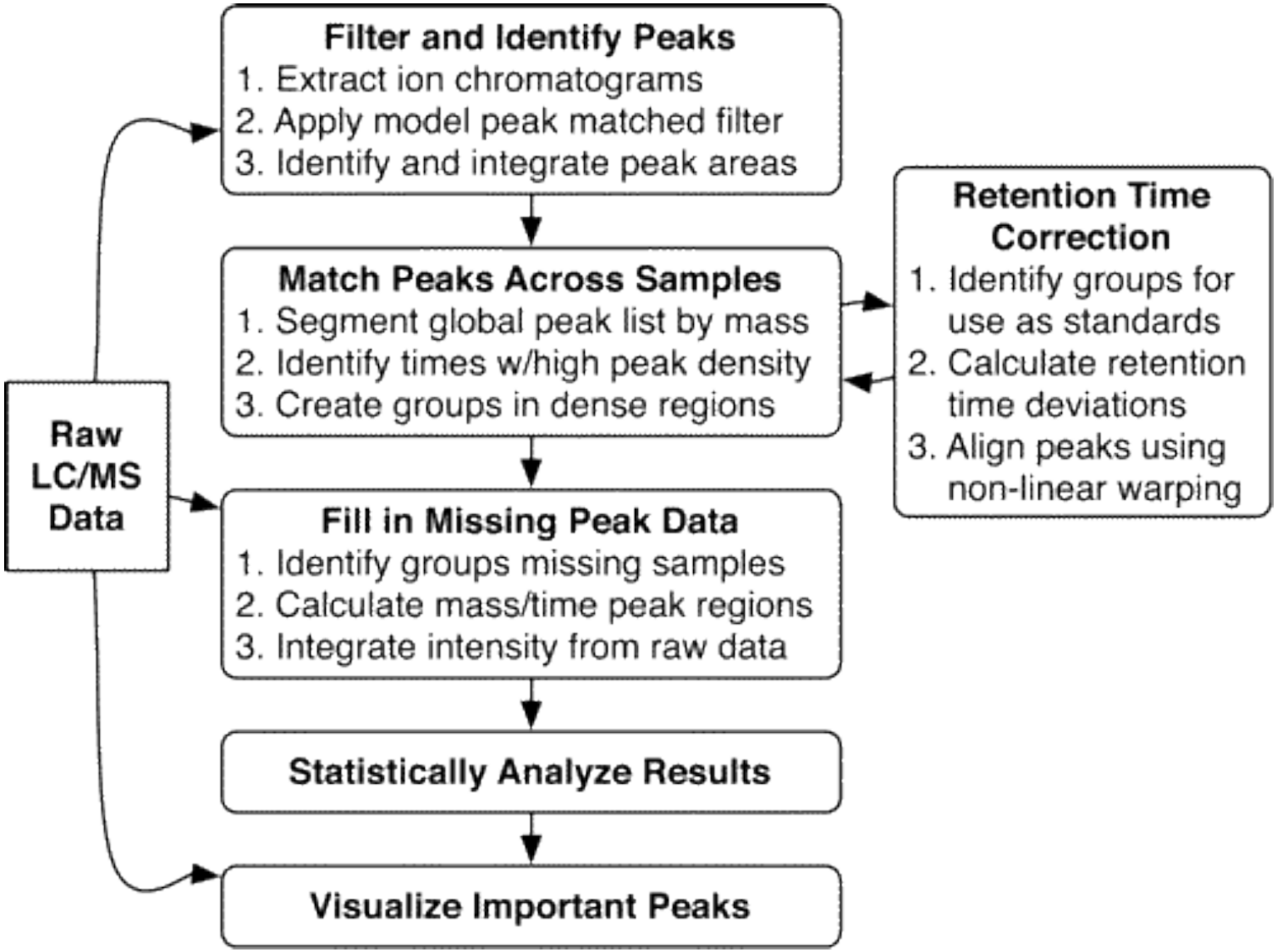
- Feature finding algorithms are often very sensitive to the choice of **parameters**
- Tuning these parameters can be a challenge

- **Sensitivity** is often an issue in feature finding: distinguishing signal from noise can be a challenge
- Lack of sensitivity is often a problem for large-scale studies – missing values

XCMS

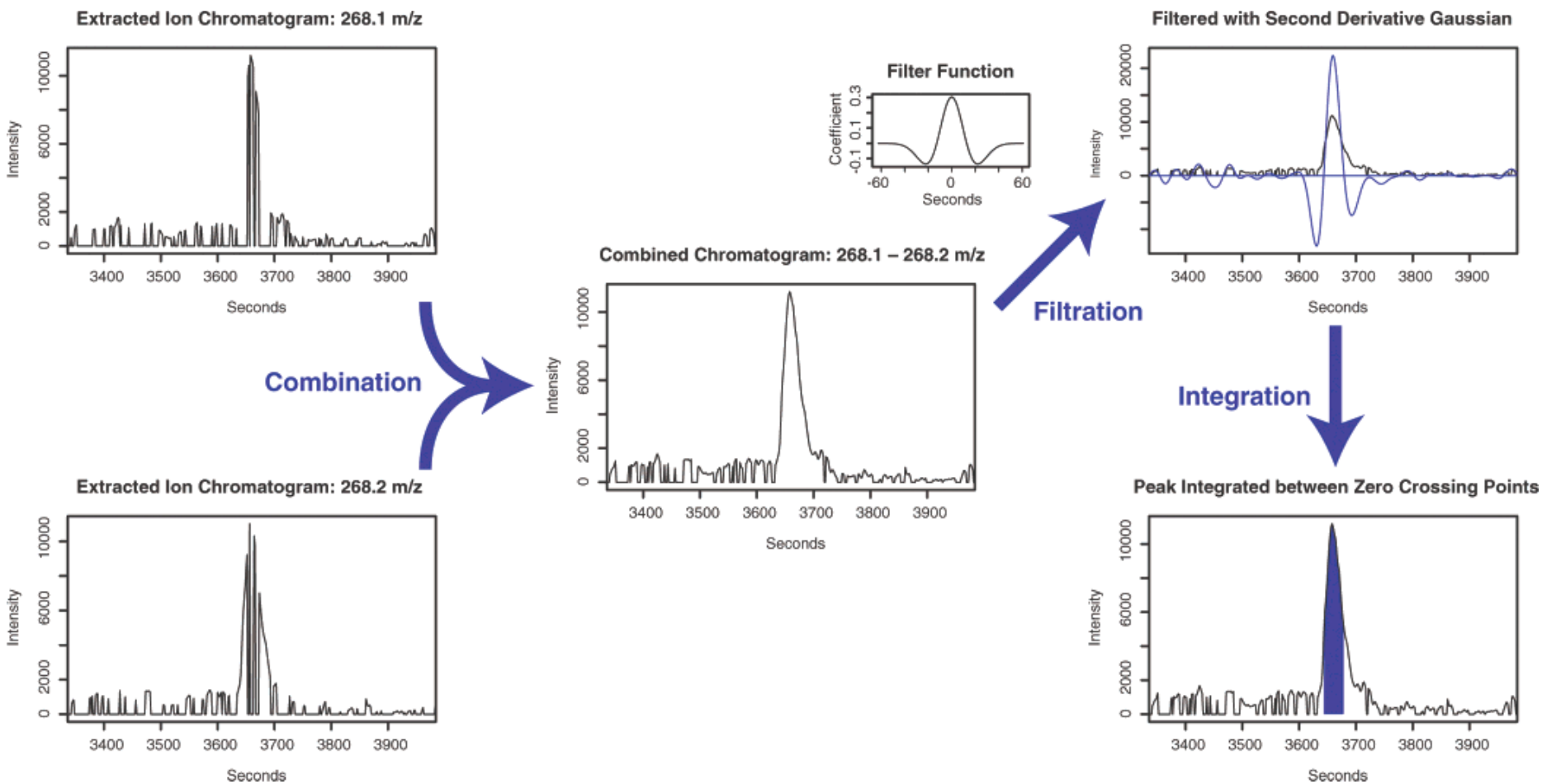
- XCMS is a Bioconductor package, written in R
- **Key ideas**
  - Extract mass traces by binning peaks w.r.t. m/z
  - Treat mass bins as distinct mass traces
  - Detect peaks in these mass traces using standard methods from signal processing
  - Align detected mass traces in the RT dimension across maps using nonlinear de-warping

XCMS



Smith et al., Anal. Chem., 2006, 78, 779-787

XCMS

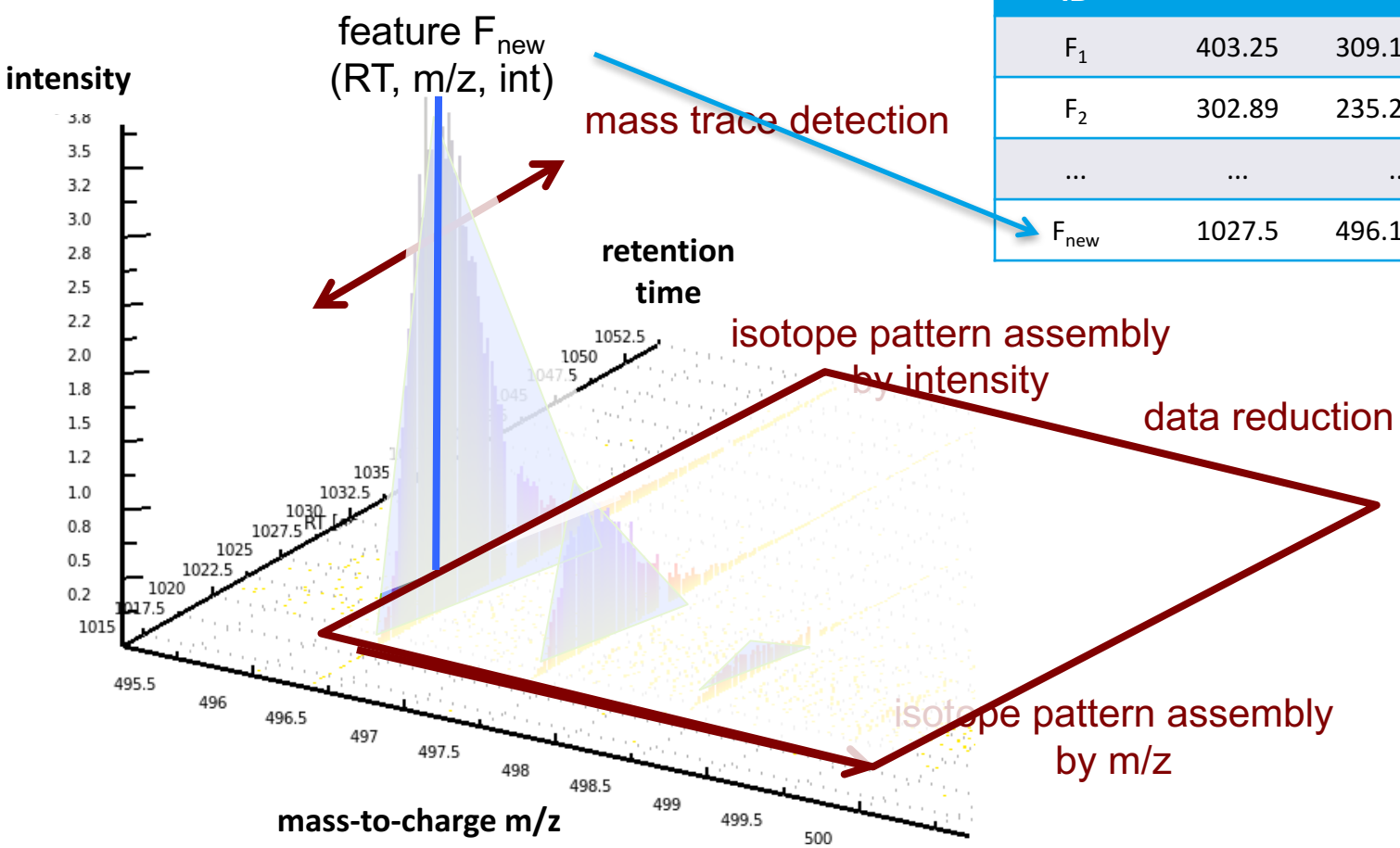


Smith et al., Anal. Chem., 2006, 78, 779-787

XCMS

- XCMS has become the quasi standard for LC-MS metabolomics data analysis
- Recent versions include more advanced methods, including wavelet peak detection
- For many tasks (e.g., biomarker detection), the identification of differential mass traces is sufficient (lower complexity of metabolomics data sets)
- Other software packages also assemble mass traces back to features (e.g., OpenMS FeatureFinderMetabo)
- Advantages here:
  - Profit from additional information, increase specificity
  - Reduced number of signals (multiple mass traces per feature)

OpenMS - Metabolite Feature Finding

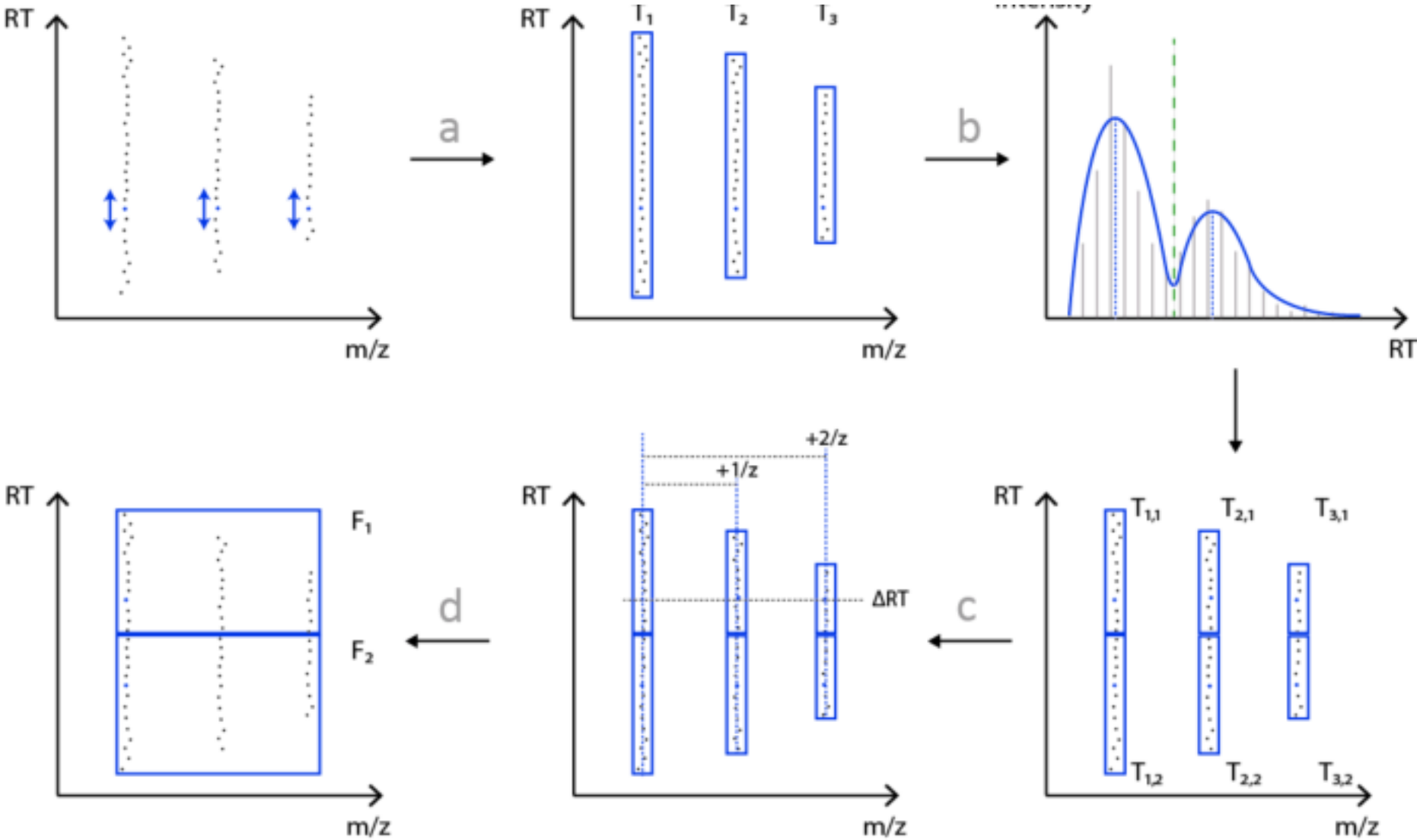


MS data condensed to feature list:

Feature ID	RT	m/z	intensity
F_1	403.25	309.13455	345923.1
F_2	302.89	235.20503	8109.5
...	...	...	...
F_{new}	1027.5	496.11304	45209.8

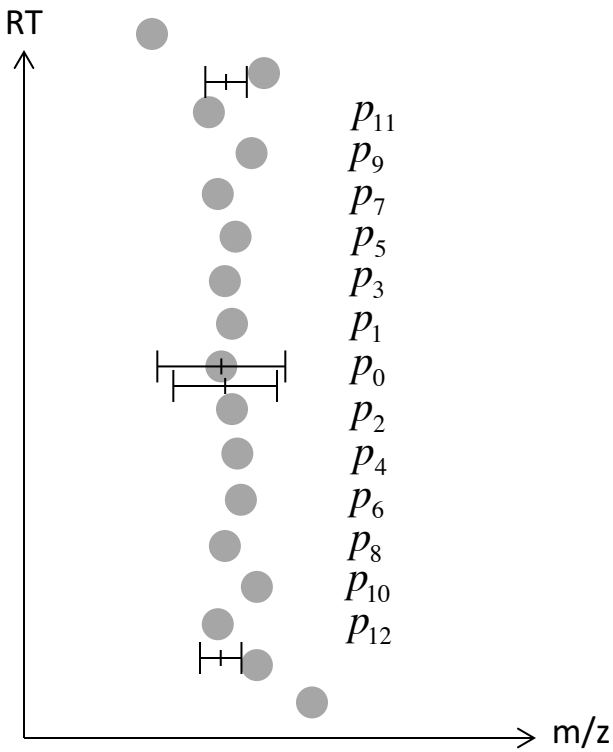
Kenar et al., Mol. Cell. Prot., 2014, 13(1):348-59. doi: 10.1074/mcp.M113.031278

Algorithmic Overview



Kenar et al., Mol. Cell. Prot., 2014, 13(1):348-59. doi: 10.1074/mcp.M113.031278

Mass Trace Detection



$\mu_{02} = 521.4238571$ m

$\sigma^2_{02} \approx 0.006371$ mass_error_param

$T = (p_{02}, p_{01}, p_{14}, p_2, p_0, p_1, p_3, \ldots, p_{11})$

- A mass spectrometric peak *p* is given by

$$p = (t, m, i)$$

t : retention time, m : mass-to-charge ratio, i : intensity

- A mass trace *T* is a list of peaks:

$$T = \left(p_1, p_2, \ldots, p_k, p_l, \ldots, p_n\right) \qquad t_k < t_l \, \forall k < l$$

- m/z error model is adaptive
- Online Gaussian density estimation

$$\mu_{n+1} = \frac{w_n \cdot \mu_n + i_{n+1} \cdot m_{n+1}}{w_n + i_{n+1}}$$

centroid m/z

$$\sigma^2_{n+1} = \frac{w_n \cdot \sigma^2_n + i_{n+1} \cdot \left(m_{n+1} - \mu_{n+1}\right)^2}{w_n + i_{n+1}}$$

m/z error

$$w_n = \sum_k^n i_k$$

weight

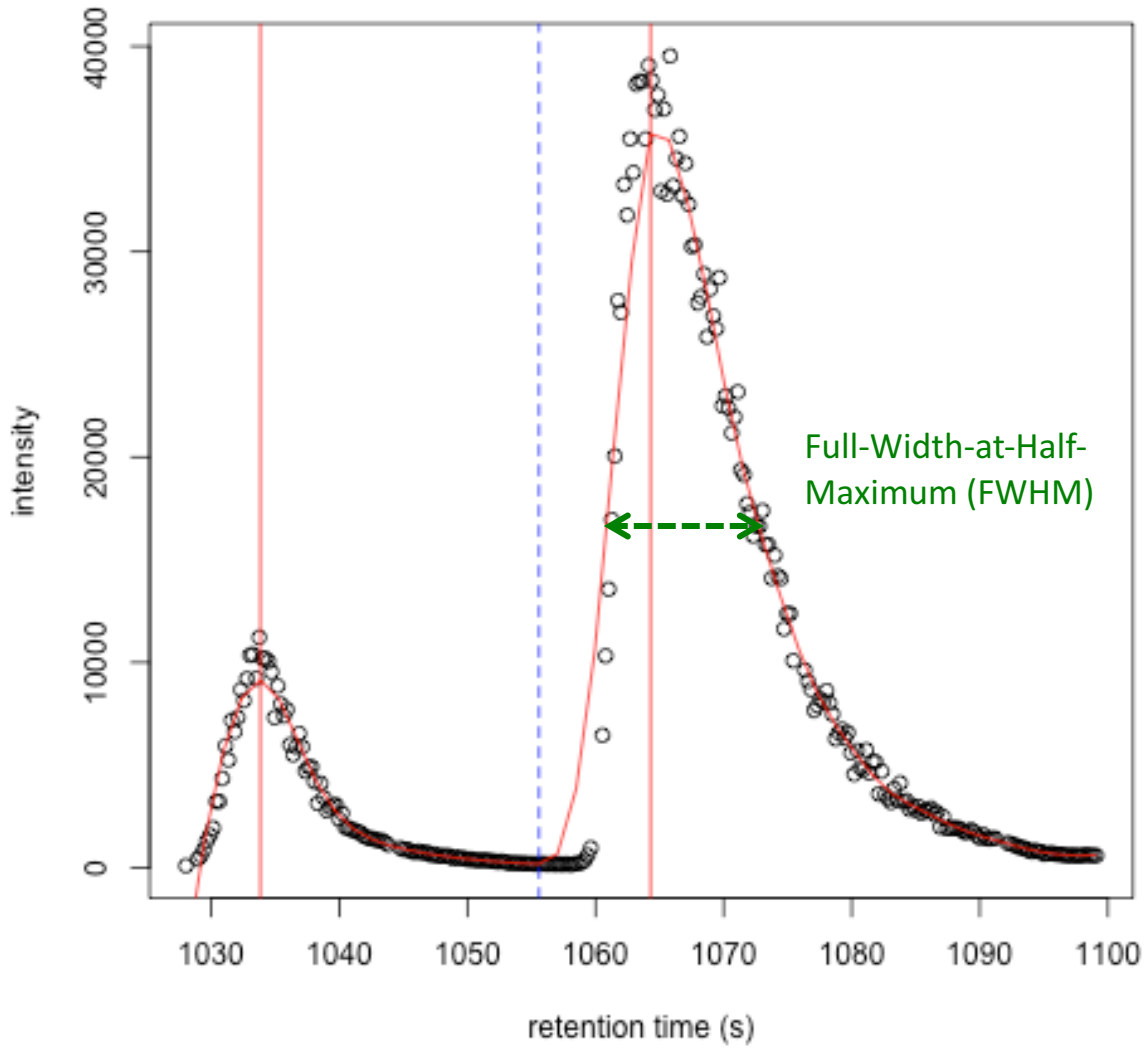
$$\mu_n - 3 \cdot \sigma_n \le m_{n+1} \le \mu_n + 3 \cdot \sigma_n$$

m/z constraint

Kenar et al., Mol. Cell. Prot., 2014, 13(1):348-59. doi: 10.1074/mcp.M113.031278

Peak Separation

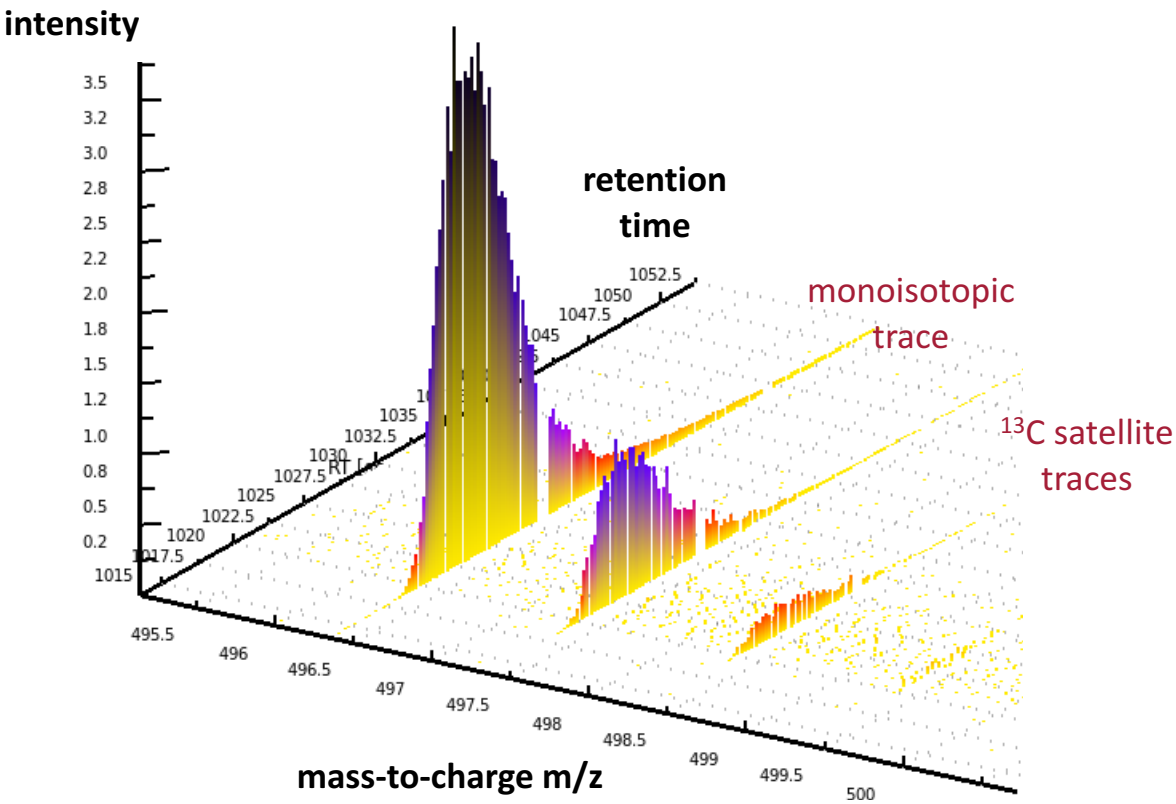
- Split chromatographic peaks overlapping in retention time



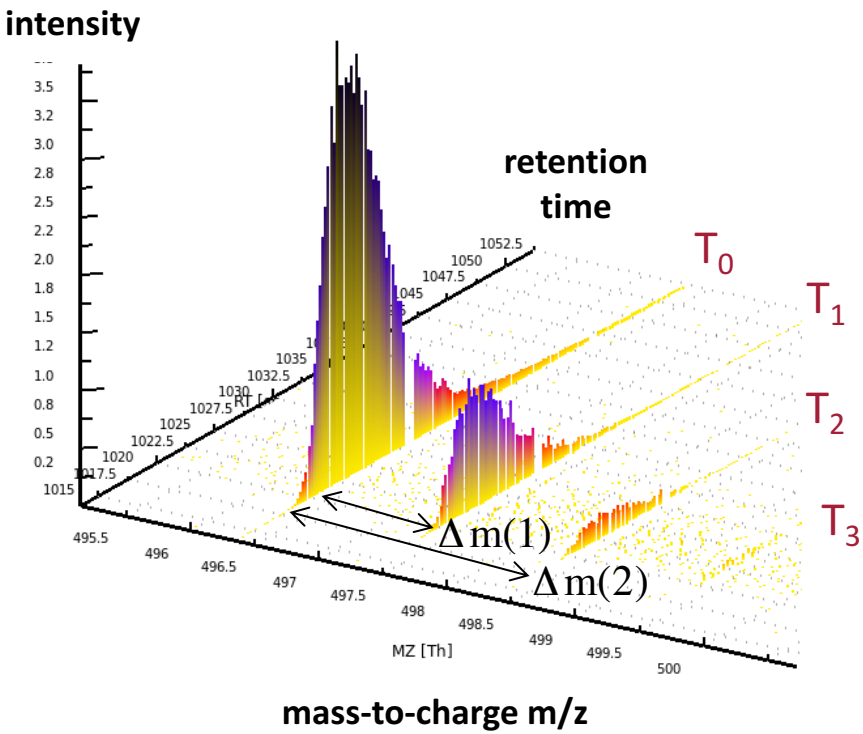
Kenar et al., Mol. Cell. Prot., 2014, 13(1):348-59. doi: 10.1074/mcp.M113.031278

Feature Assembly

- Identify mass traces belonging to the same feature
- Multiple explanations are possible
- Create all potential hypotheses and score them



Feature Scoring – m/z



- m/z distances T_0 and T_j:

$$\Delta m(j) = \left| \bar{m}_0 - \bar{m}_j \right|$$

- Theoretical m/z distances:

$$\mu(j) = 1.0033\ \mathrm{Da} \cdot \frac{j}{z}$$

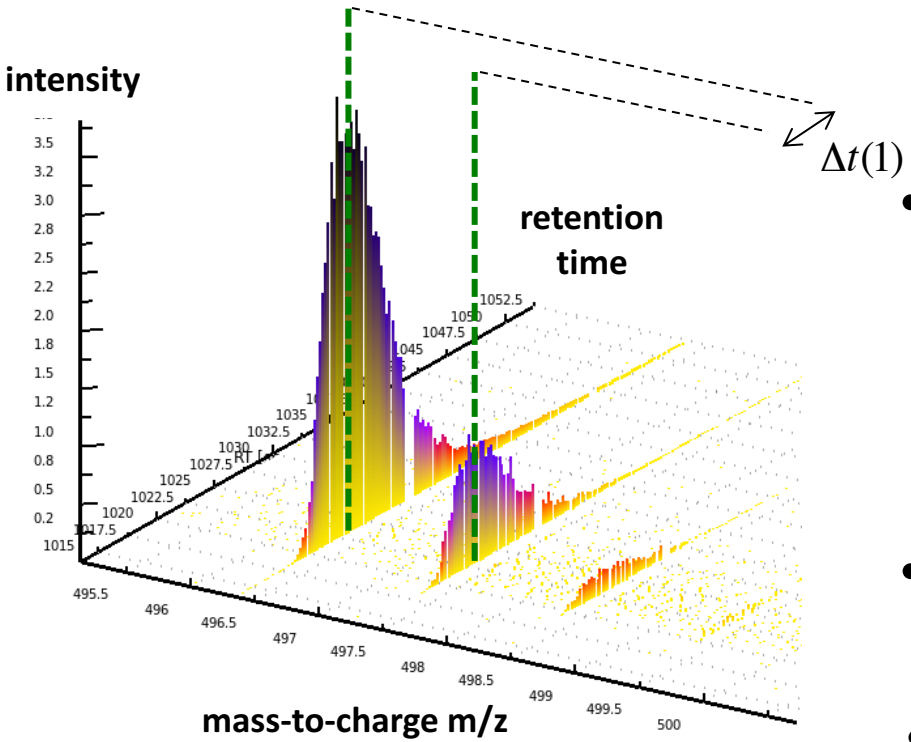
- Mass errors for T_0 and T_j:

$$\sigma^2(j) = \sigma_0^2 + \sigma_j^2$$

- Pairwise scoring function:

$$S_{\Delta m}(j) = \begin{cases} e^{-\frac{(\Delta m(j) - \mu(j))^2}{2\sigma^2(j)}}, & \text{if } \mu(j) - 3 \cdot \sigma(j) \le \Delta m(j) \le \mu(j) + 3 \cdot \sigma(j) \\ 0 & \text{else.} \end{cases}$$

Feature Scoring – RT



- RT shifts between T_0 and T_j:

$$\Delta t(j) = \left| \bar{t}_0 - \bar{t}_j \right|$$

- Gaussian error model with

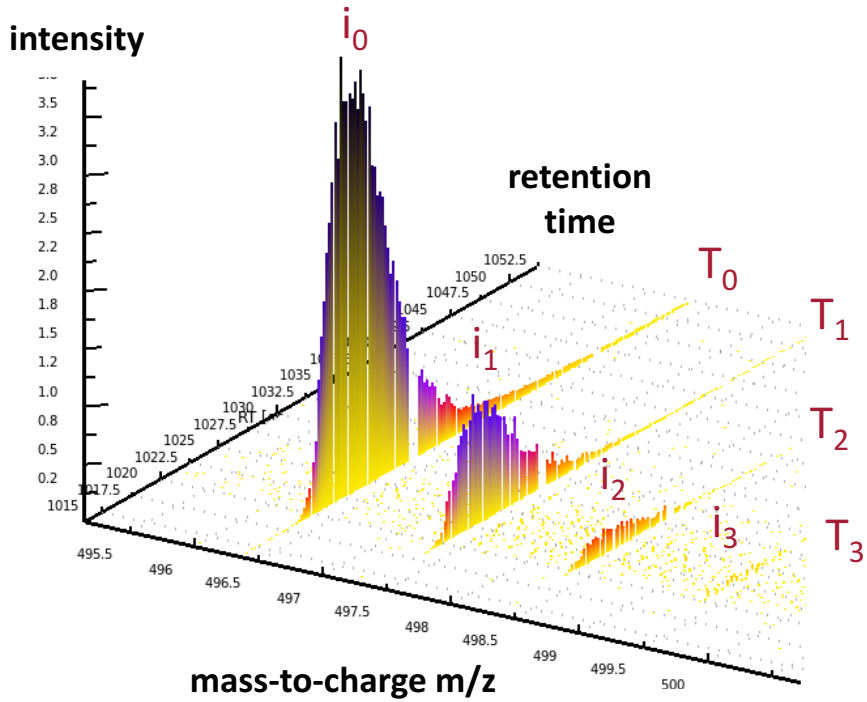
$$\mu_{\Delta RT} = 0 \qquad \sigma^2_{\Delta RT} = \left(\frac{\Delta t_{0.5}}{2\sqrt{2\ln 2}} \right)^2$$

- Pairwise scoring function:

$$S_{\Delta RT}(j) = \begin{cases} e^{-\frac{(\Delta t(j))^2}{2\sigma^2_{\Delta t}}}, & \text{if } -3\cdot\sigma_{\Delta t} \le \Delta t(j) \le 3\cdot\sigma_{\Delta t} \\ 0 & \text{else.} \end{cases}$$

Feature Scoring – Intensity

- **Problem:** There is no 'averagine' for metabolites
- **Idea**
  - Enumerate metabolite compositions and learn intensities
  - 'Golden rules' describe likely chemistry *(Kind & Fiehn, BMC Bioinfo, 2007)*
  - Generate all compositions, remove unlikely ones based on heuristics



24 mio. compositions

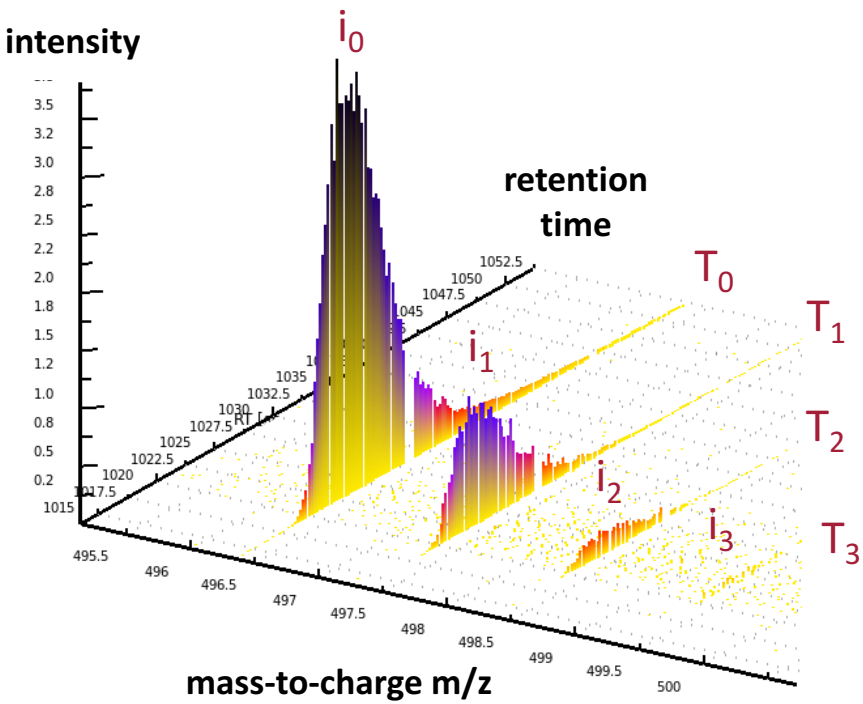
↓

Random subsample: 115 k compositions

↓

SVM (RBF kernel)

Feature Scoring – Intensity



- Intensity ratio of T_0 and T_j:

$$r(j) = \frac{i_j}{i_0}$$

- Assess if valid isotope ratios:

$$m(T_0), r(0), r(1), \ldots, r(5)$$

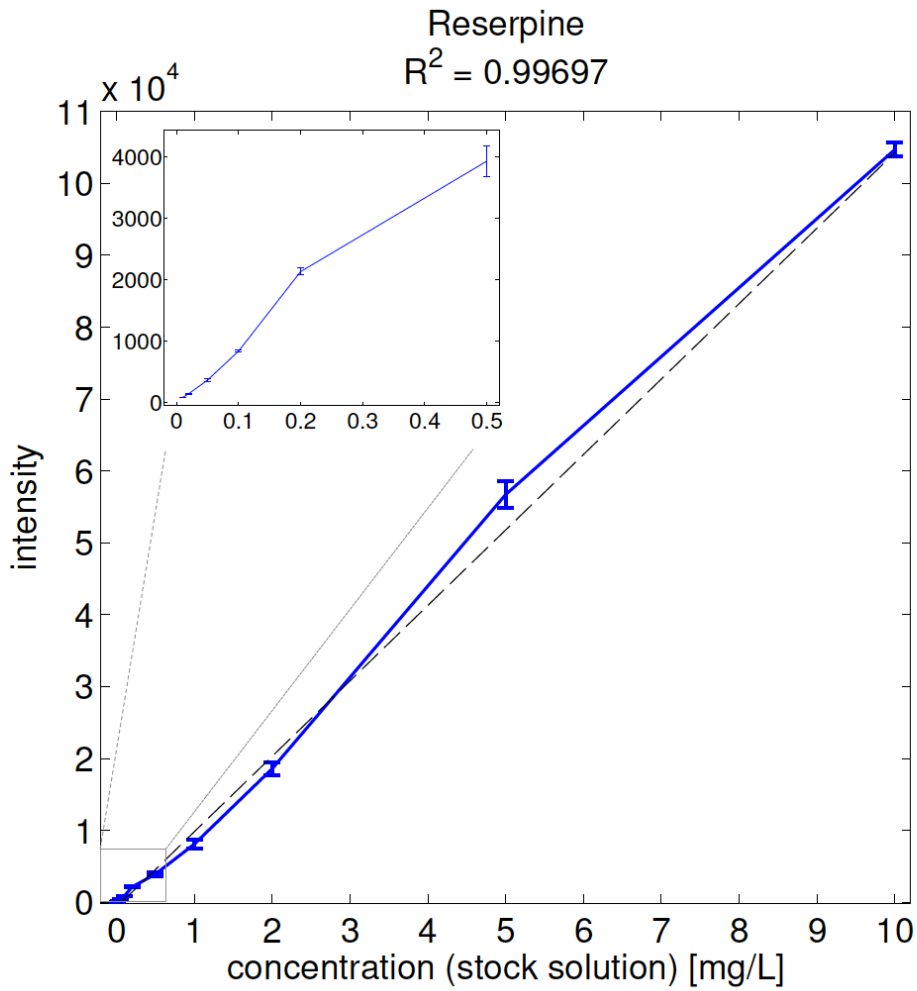
↓

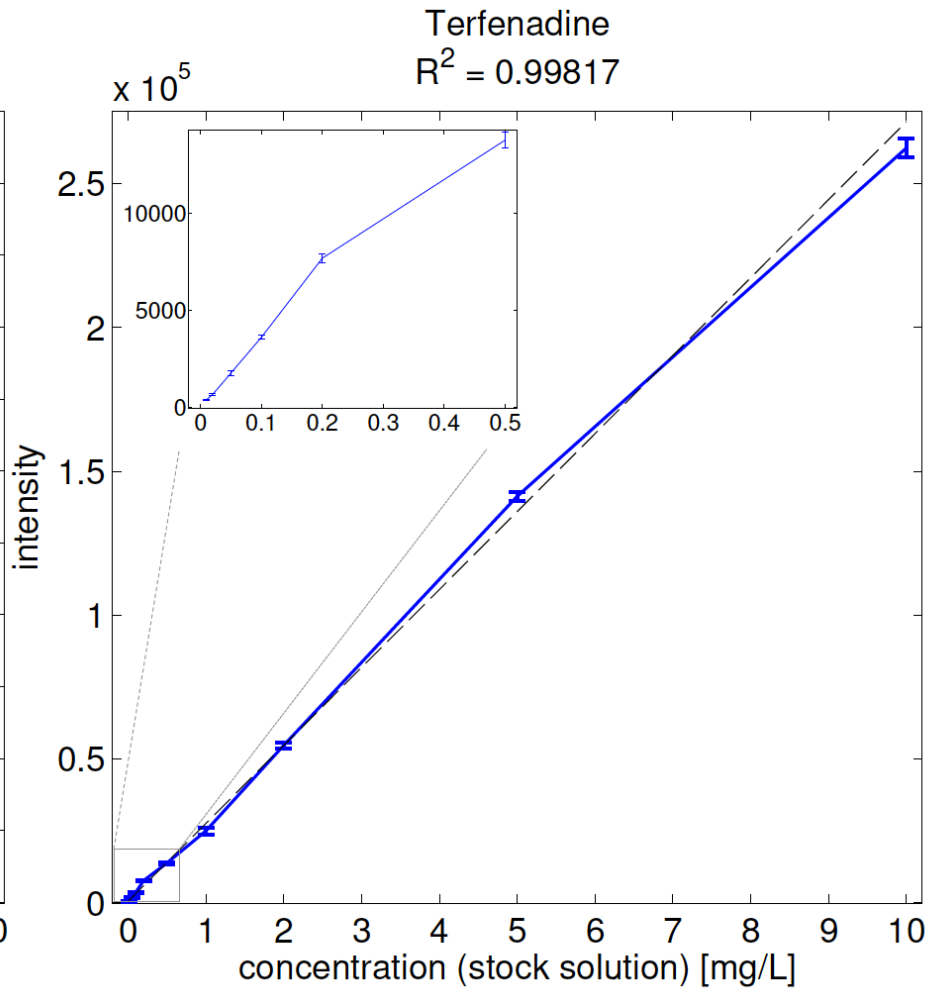
SVM
RBF kernel

↓

Yes, it is a legal isotope pattern, ***keep it***
Or
No, it is not a legal isotope pattern,
discard it

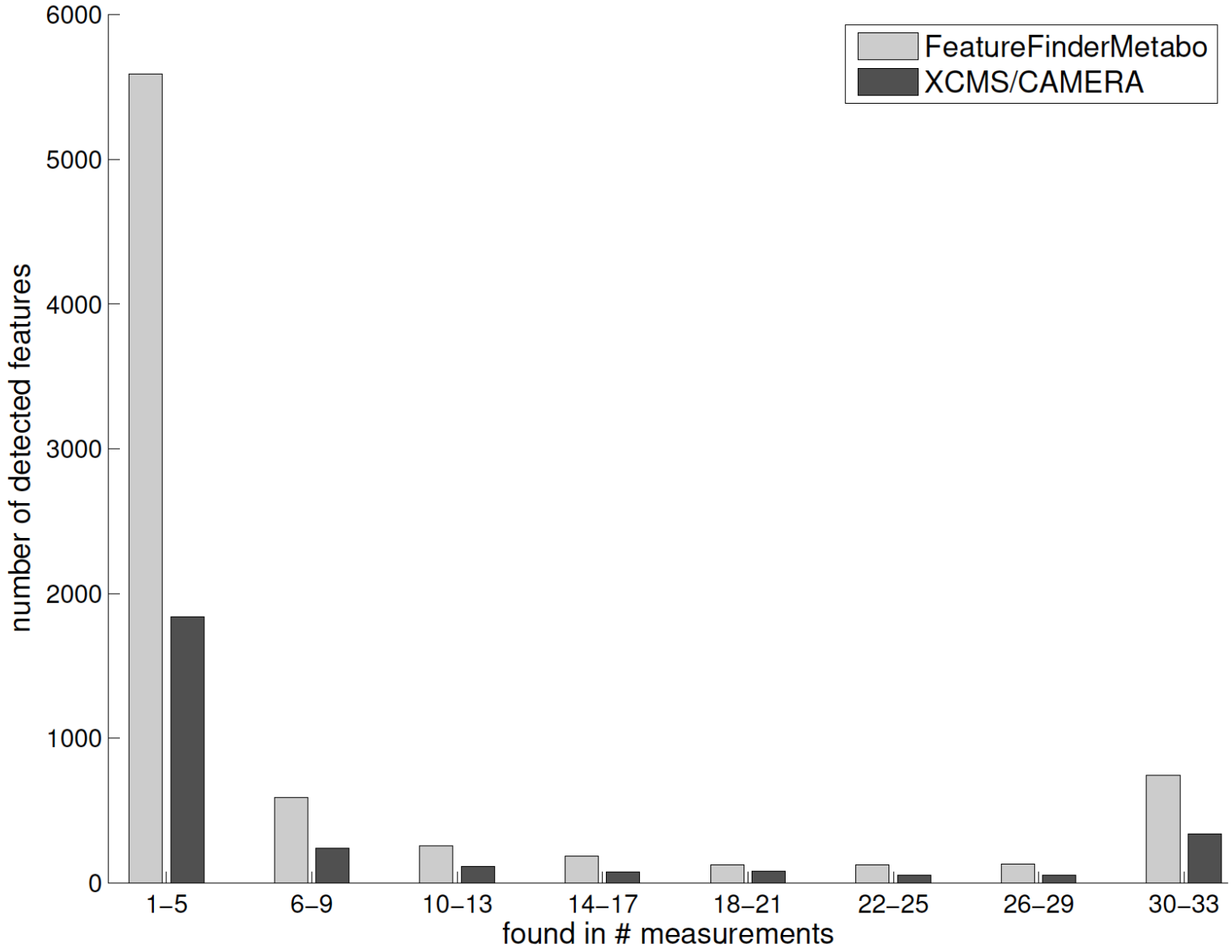
Quantification Linearity – Spike-In





Kenar et al., Mol. Cell. Prot., 2014, 13(1):348-59. doi: 10.1074/mcp.M113.031278

Sensitivity – Human Plasma



Kenar et al., Mol. Cell. Prot., 2014, 13(1):348-59. doi: 10.1074/mcp.M113.031278

Specificity – Synthetic Data

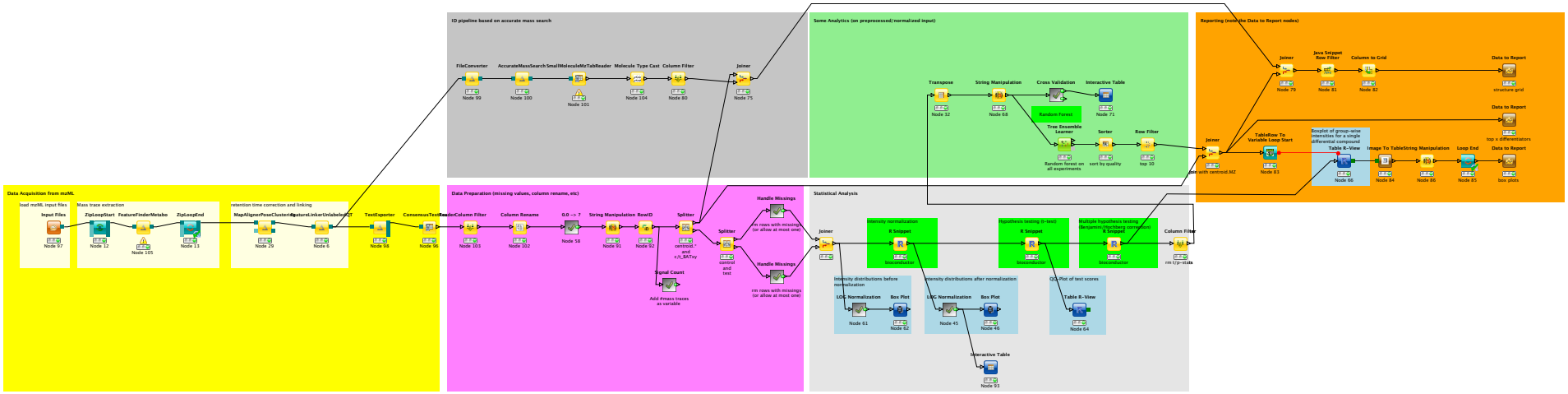
- **Benchmarking feature detection algorithms is HARD**
  - Multiple metrics are required: linearity, sensitivity, specificity
  - Sensitivity needs to be balanced with specificity
  - Experimental data does not come with a well-defined ground truth
- **Idea**
  - Simulated LC-MS data with known composition
  - Take a well-defined experimental dataset (identification lists from a metabolomics study, plant metabolites)
  - OpenMS LC-MS simulator was expanded to generate metabolite data

Method	Recall	Precision	F-score
OpenMS	96%	97%	0.97
XCMS/Camera	88%	37%	0.52

NON-TARGETD METABOLOMICS WITH OPENMS

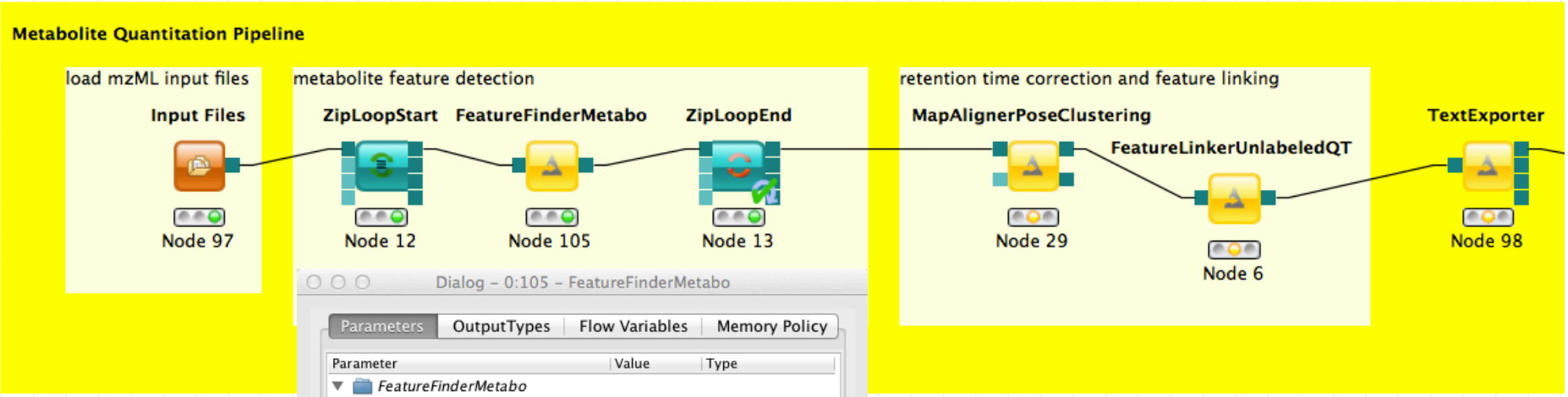
- Workflows for non-targeted metabolomics
- Metabolomics workflows with OpenMS in KNIME
- Integration into Compound Discoverer

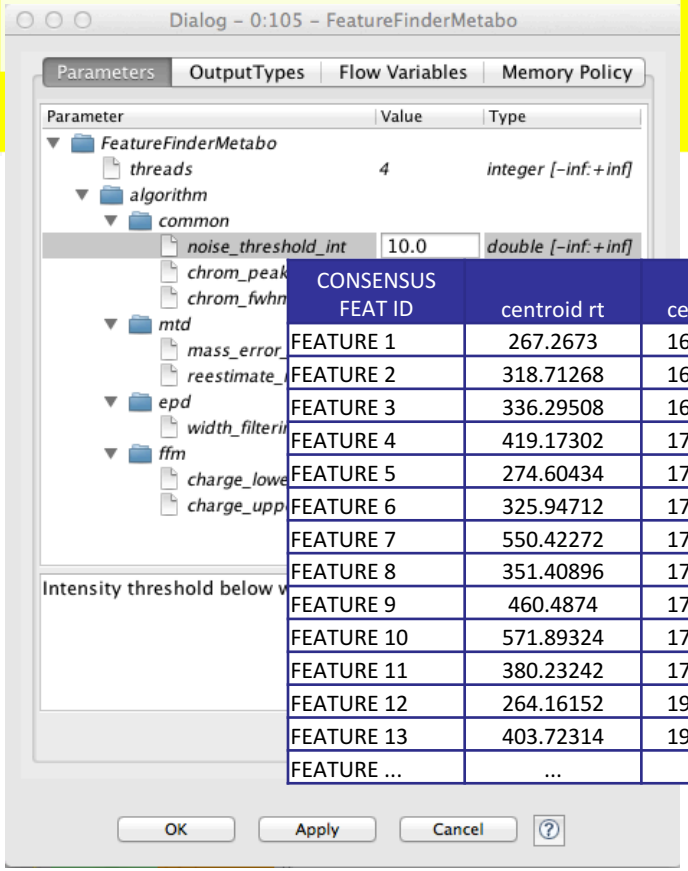
Metabolomics – Biomarker ID



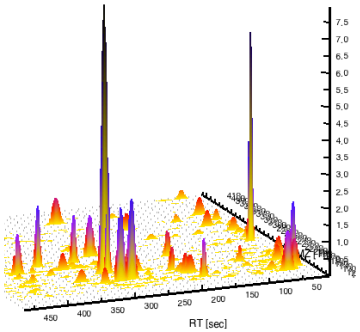
- Complex workflow analyzing a diabetes-related metabolomics biomarker study
  - Data preprocessing (yellow)
  - Quantification (purple)
  - Identification based on accurate mass/HMDB (gray)
  - Detection of distinctive features, statistics (green/gray)
  - Reporting of differential features and their structures (orange)

Metabolite Quantitation





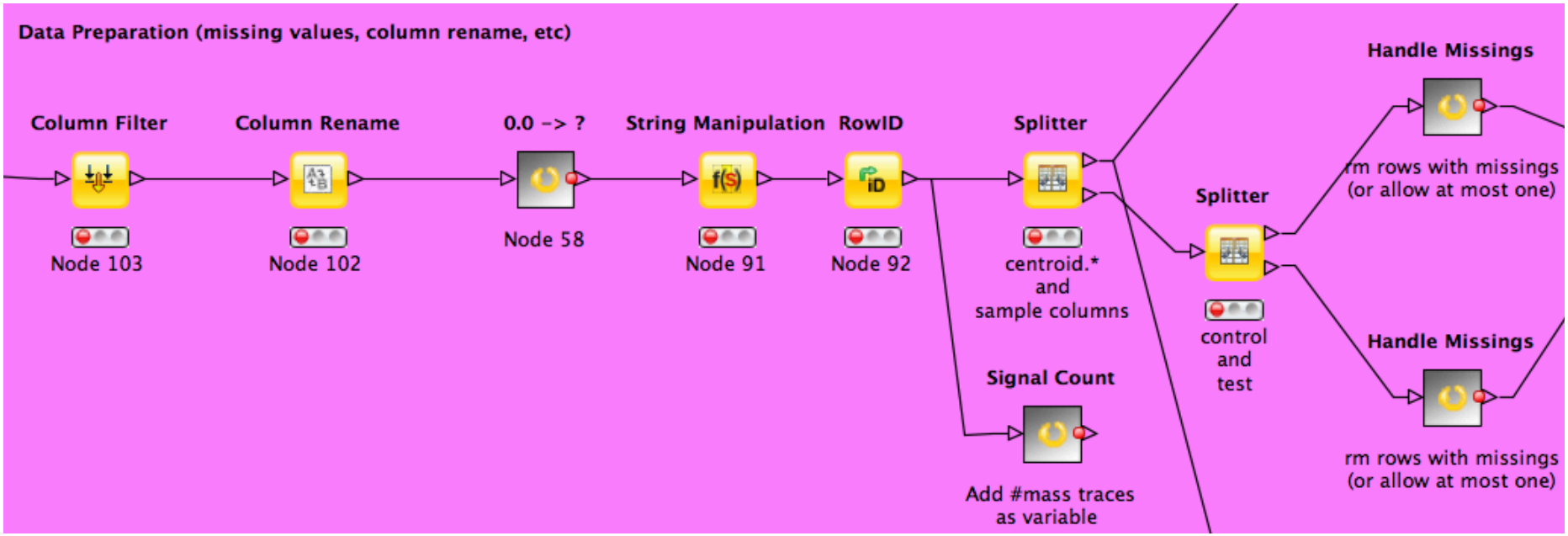
MSConvert



5 controls vs.
5 samples

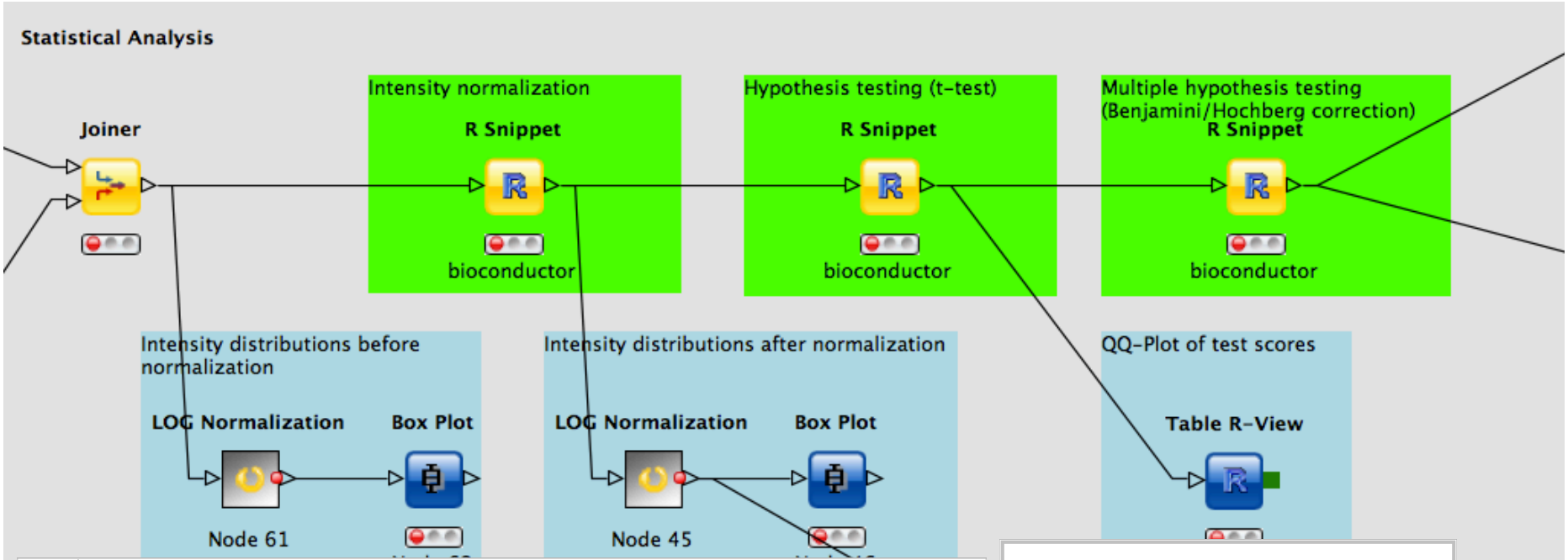
CONSENSUS FEAT ID	centroid rt	centroid m/z	...	charge	sample 1 intensity	sample 2 intensity
FEATURE 1	267.2673	163.0753568	...	1	5288099840	50020923440
FEATURE 2	318.71268	163.0753568	...	1	18835900	17835200
FEATURE 3	336.29508	163.0753568	...	1	7285210	6285210
FEATURE 4	419.17302	179.0702718	...	1	175022000	105022000
FEATURE 5	274.60434	179.0702718	...	1	44317400	33317400
FEATURE 6	325.94712	179.0702718	...	1	11875200	12879200
FEATURE 7	550.42272	179.0702718	...	1	4871360	5071360
FEATURE 8	351.40896	179.0702718	...	1	2919350	1019350
FEATURE 9	460.4874	179.0702718	...	1	2021340	3221340
FEATURE 10	571.89324	179.0702718	...	2	1546820	1446820
FEATURE 11	380.23242	179.0702718	...	2	1993120	1893120
FEATURE 12	264.16152	195.0651868	...	2	269592992	279592532
FEATURE 13	403.72314	195.0651868	...	2	21862600	20342600
FEATURE ...	...	...	...	...	...	...

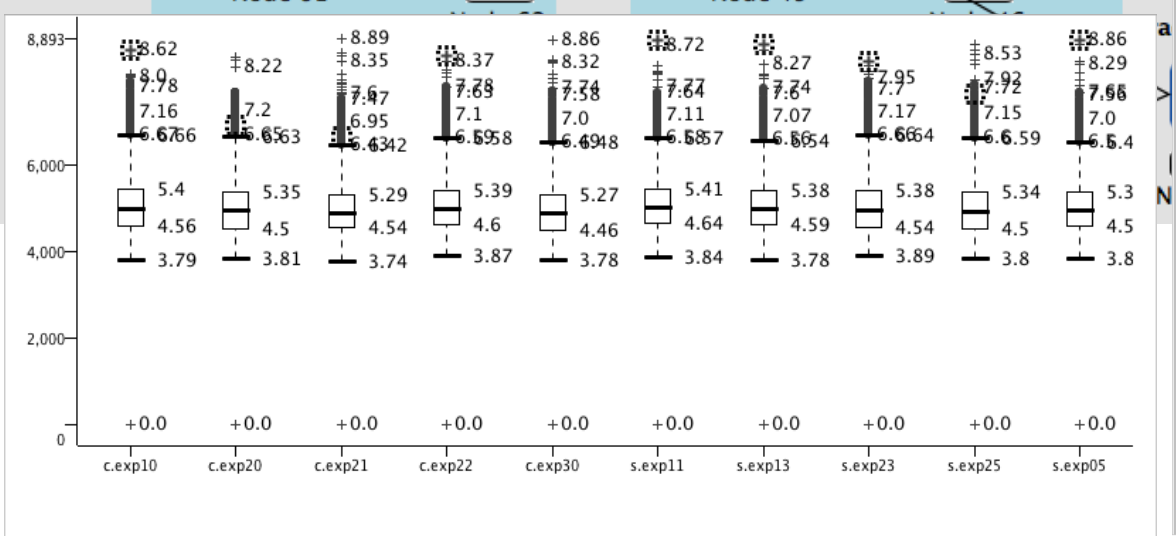
Data Table Magic

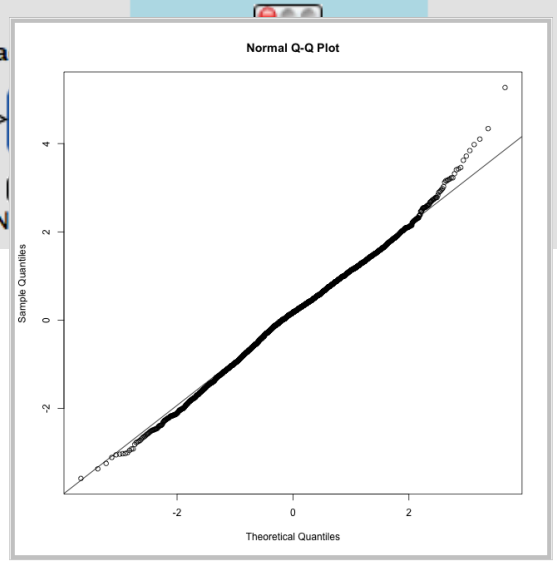


Row ID	centroid.rt	centroid.mz	...	charge	Control_1	Control_2	Sample_2	Sample_2
FEATURE 1	267.2673	163.0753568	...	1	5288099840	50020923440	5288099840	50020923440
FEATURE 2	318.71268	163.0753568	...	1	18835900	17835200	18835900	17835200
FEATURE 3	336.29508	163.0753568	...	1	7285210	6285210	7285210	6285210
FEATURE 4	419.17302	179.0702718	...	1	175022000	105022000	175022000	105022000
FEATURE 5	274.60434	179.0702718	...	1	44317400	33317400	44317400	33317400
FEATURE 6	325.94712	179.0702718	...	1	11875200	12879200	11875200	12879200
FEATURE 7	550.42272	179.0702718	...	1	4871360	5071360	4871360	5071360
FEATURE 8	351.40896	179.0702718	...	1	2919350	1019350	2919350	1019350
FEATURE 9	460.4874	179.0702718	...	1	2021340	3221340	2021340	3221340
FEATURE 10	571.89324	179.0702718	...	2	1546820	1446820	1546820	1446820
FEATURE 11	380.23242	179.0702718	...	2	1993120	1893120	1993120	1893120
FEATURE 12	264.16152	195.0651868	...	2	269592992	279592532	269592992	279592532
FEATURE 13	403.72314	195.0651868	...	2	21862600	20342600	21862600	20342600
FEATURE ...	...	...	...	...	...	...	...	...

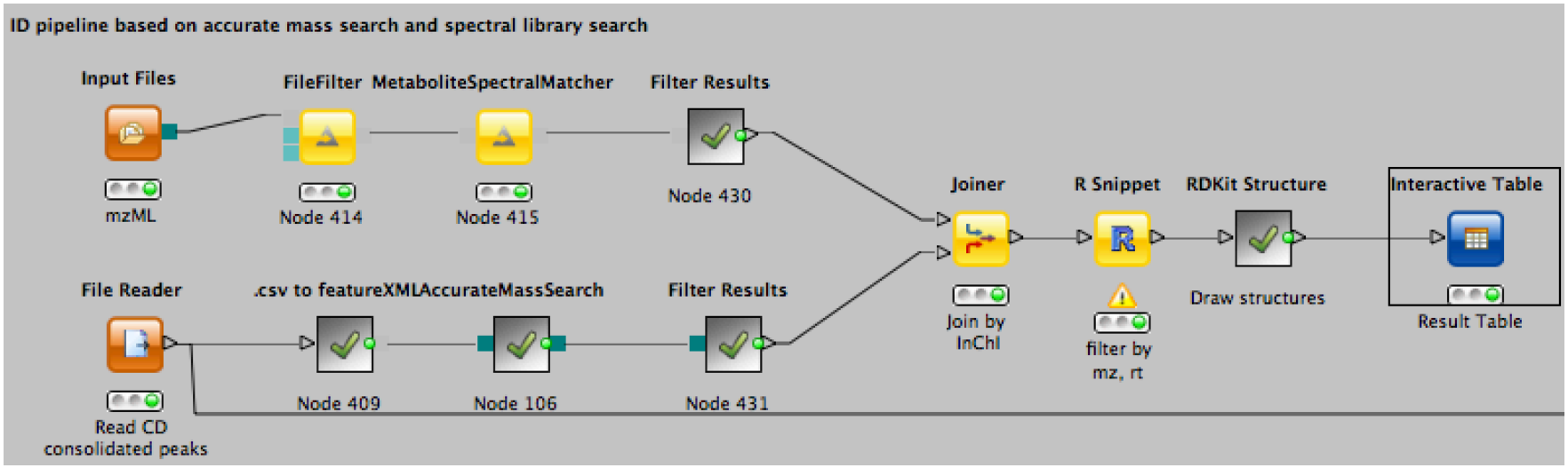
Multiple Hypothesis Testing







Metabolite ID

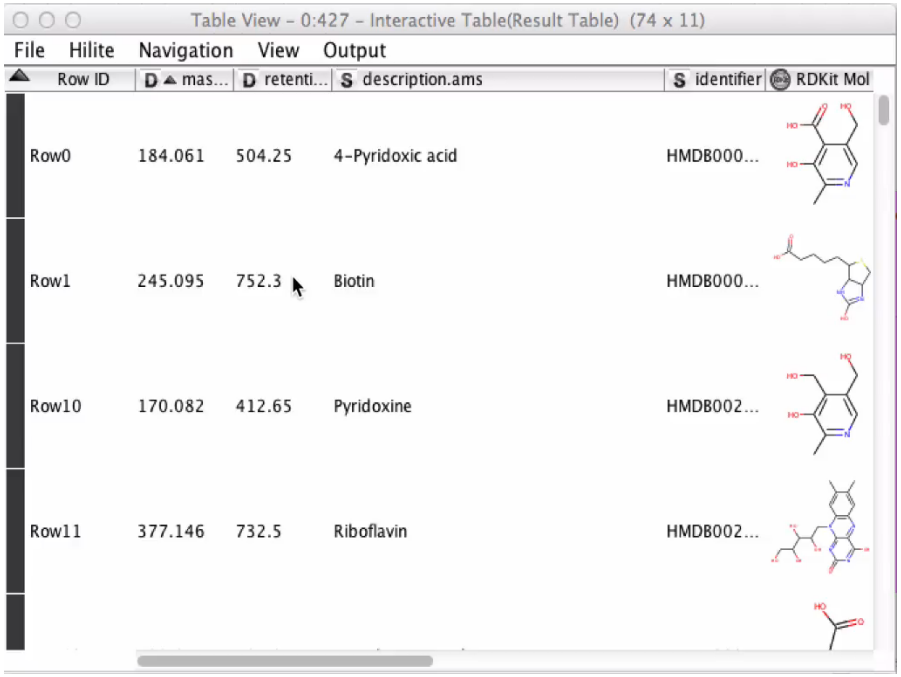


Multiple ID strategies

- Accurate mass
- Retention time database
- Retention time prediction
- Spectral matching

KNIME provides

- Online access to structure databases
- Structure visualization
- Cheminformatics
  - Metabolization
  - Substructure search



References

- **XCMS**
  - C.A. Smith, E.J. Want, G.C. Tong, R. Abagyan, and G. Siuzdak. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. Anal. Chem., 2006,
- **FeatureFinderMetabo**
  - Kenar, E, Franken, H, Forcisi, S, Wörmann, K, Häring, H, Lehmann, R, Schmitt-Kopplin, P, Zell, A, and Kohlbacher, O (2014). Automated Label-Free Quantification of Metabolites from LC-MS Data. Mol. Cell. Prot., 13(1):348-59. http://dx.doi.org/10.1074/mcp.M113.031278