XCMS feature finding and retention time warping

David L. Tabb, Ph.D.

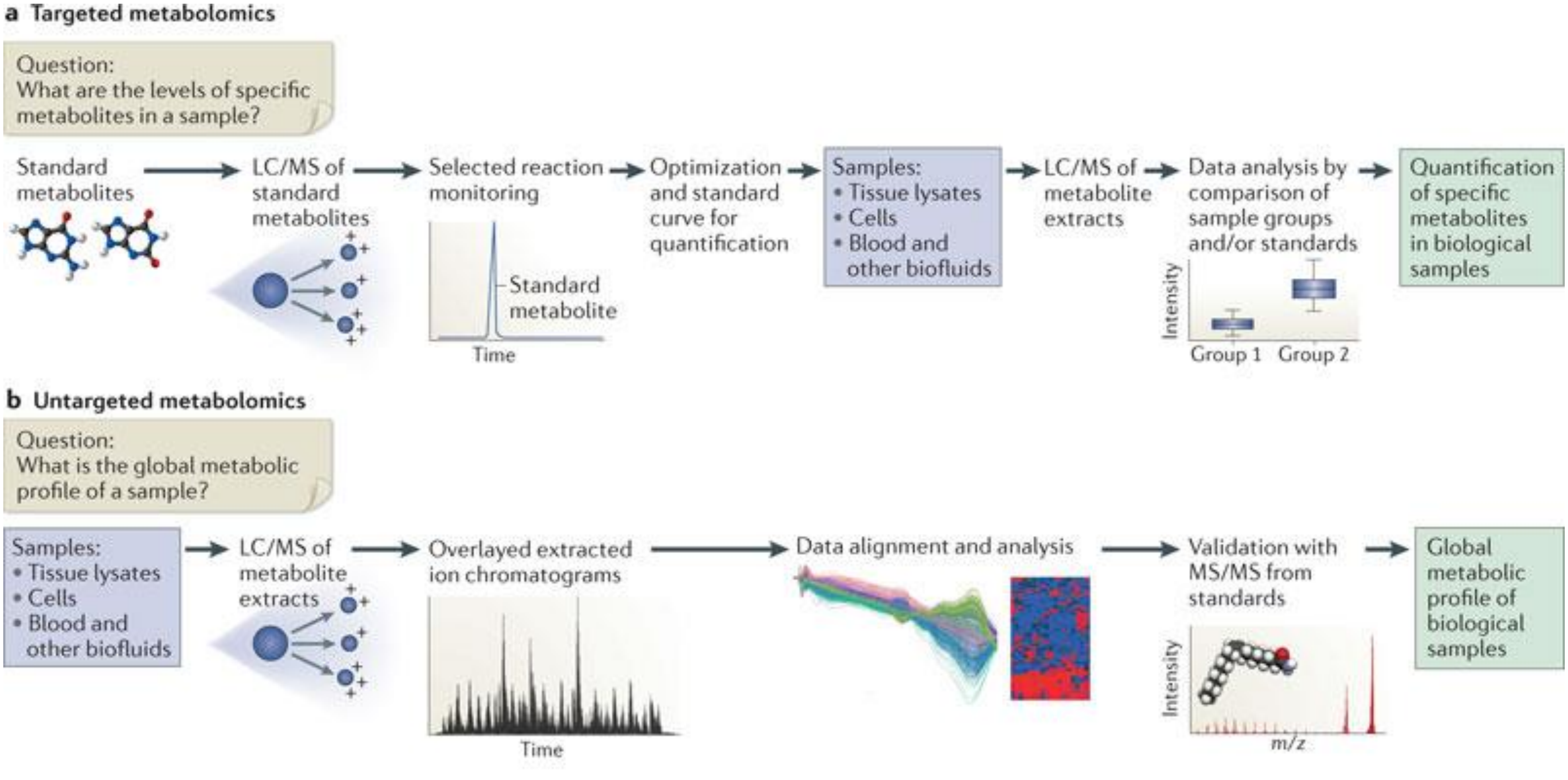
dtabb1973@gmail.com

With many valuable slide contributions from H. Paul Benton and Gary Siuzdak, The Scripps Research Institute

Overview

- Finding features by centWave
- Mapping major features across experiments
- Warping retention times in XCMS Obi-Warp

Metabolomics experiments may be targeted or untargeted



a Targeted metabolomics

Question:
What are the levels of specific metabolites in a sample?

Standard metabolites → LC/MS of standard metabolites → Selected reaction monitoring → Optimization and standard curve for quantification → Samples:
- Tissue lysates
- Cells
- Blood and other biofluids

→ LC/MS of metabolite extracts → Data analysis by comparison of sample groups and/or standards → Quantification of specific metabolites in biological samples

b Untargeted metabolomics

Question:
What is the global metabolic profile of a sample?

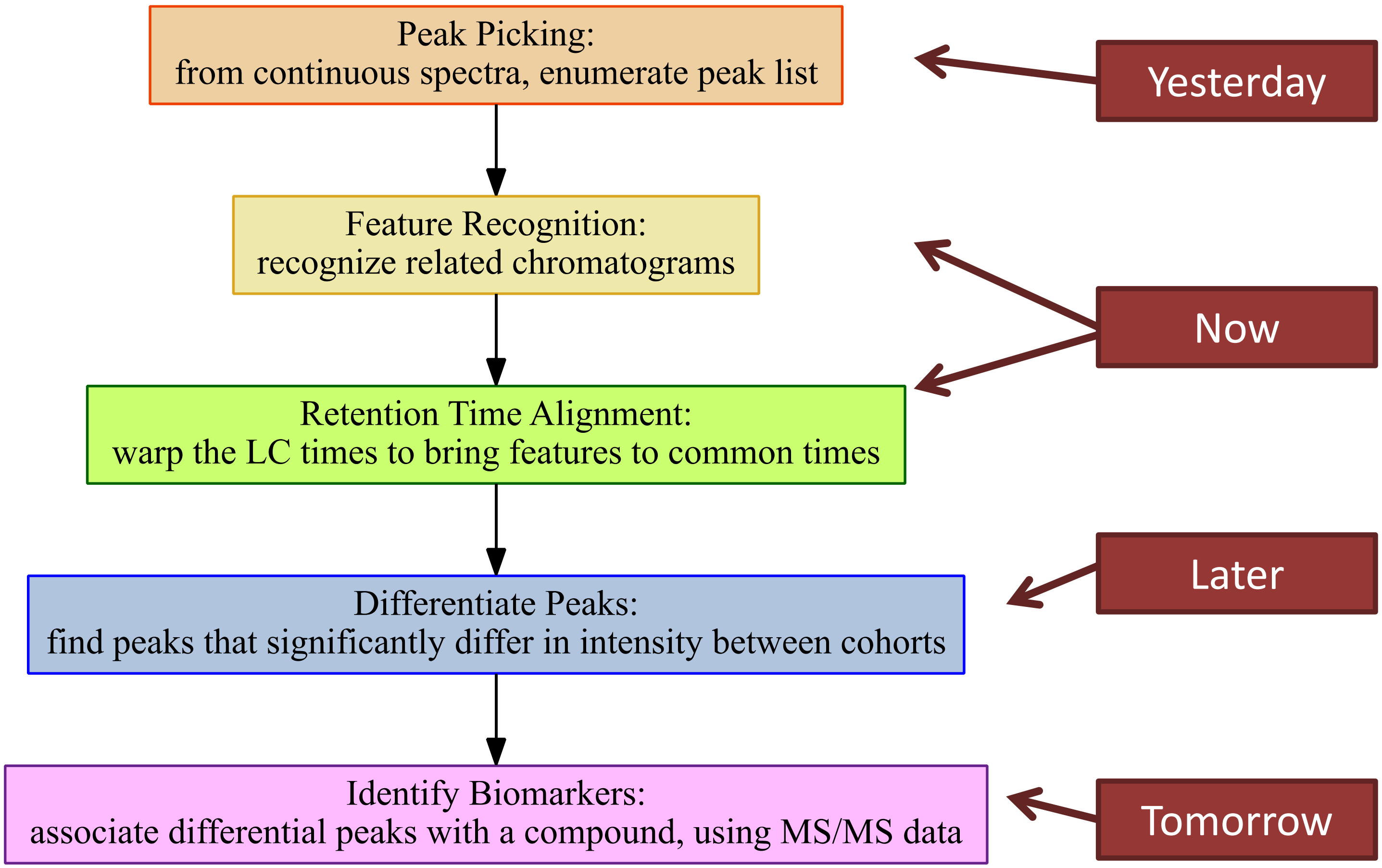
Samples:
- Tissue lysates
- Cells
- Blood and other biofluids

→ LC/MS of metabolite extracts → Overlayed extracted ion chromatograms → Data alignment and analysis → Validation with MS/MS from standards → Global metabolic profile of biological samples

Nature Reviews | Molecular Cell Biology

Patti et al, *Nature Rev Mol Cell Biology* (2012) 13:263

Untargeted metabolome informatics



Peak Picking:
from continuous spectra, enumerate peak list

Feature Recognition:
recognize related chromatograms

Retention Time Alignment:
warp the LC times to bring features to common times

Differentiate Peaks:
find peaks that significantly differ in intensity between cohorts

Identify Biomarkers:
associate differential peaks with a compound, using MS/MS data

Yesterday

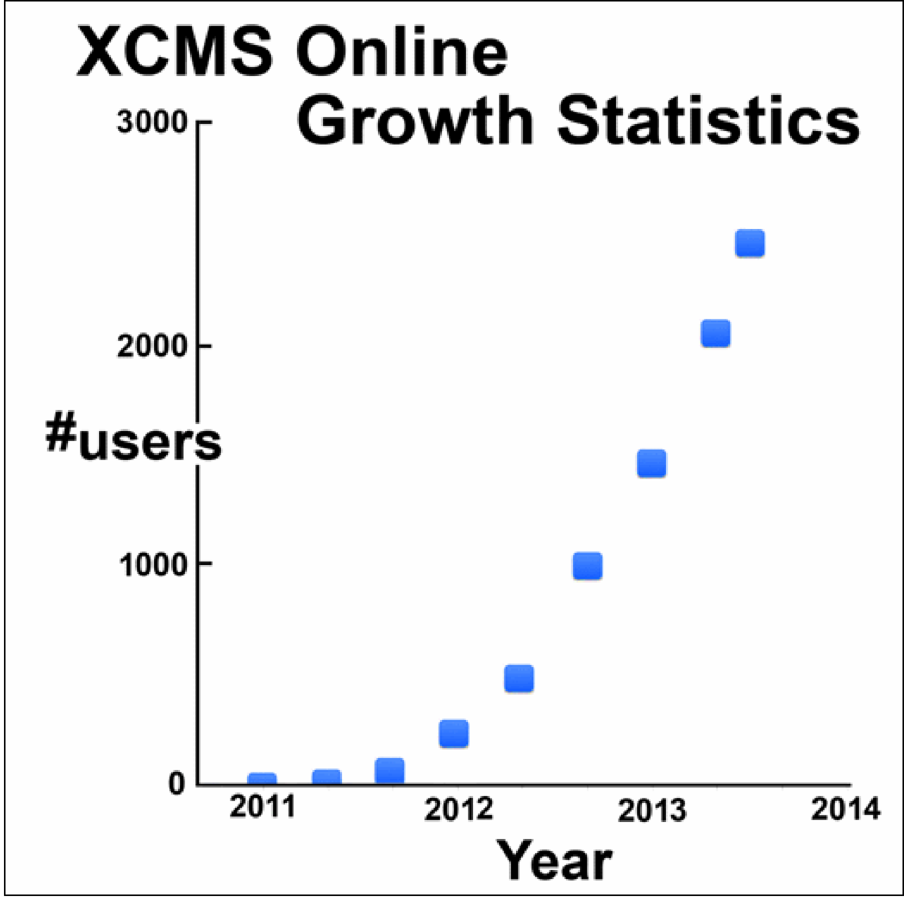
Now

Later

Tomorrow

XCMS
Online

Over a million jobs were performed during the last 2 years.
6271 users have registered on the site as of Feb, 2015.
https://xcmsonline.scripps.edu



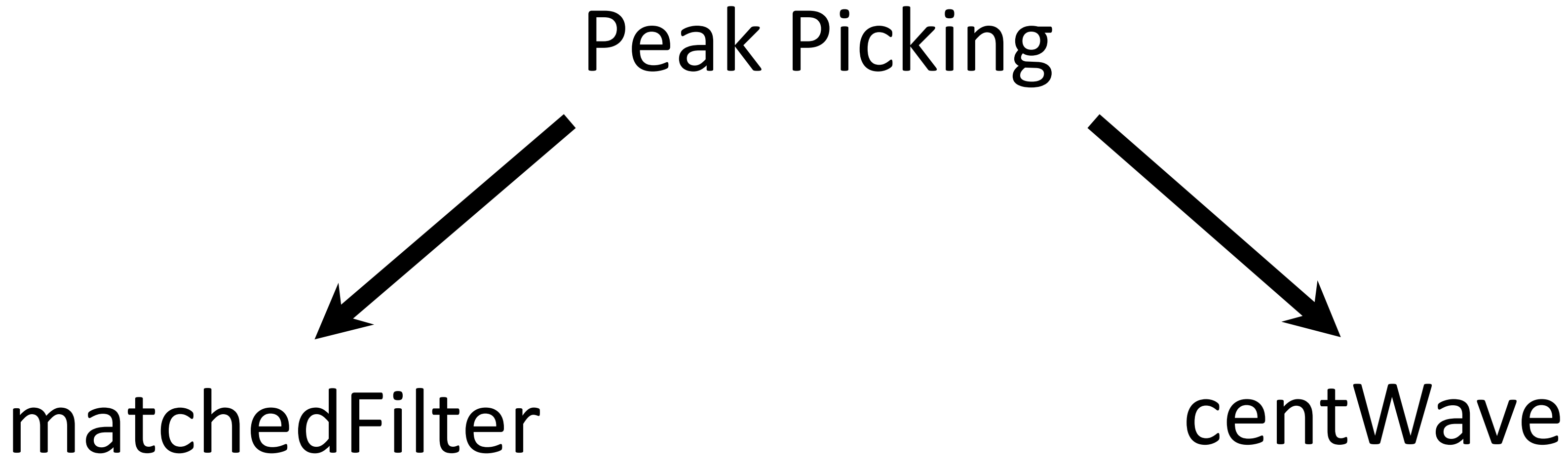
- Smith et al. *Anal. Chem*. (2006) 78:779
- Tautenhahn et al. *Anal. Chem*. (2012) 84: 5035
- Gowda et al. *Anal. Chem*. (2014) 86: 6931
- Patti et al. *Nature Protocols* (2012) 7: 508

Finding related features‡

- Peak listing translates each continuous mass spectrum to a list of m/z and intensity values.
- A particular feature appears in multiple successive MS scans, varying in m/z and intensity.
- Most features are present in multiple isotopes, may be found at multiple charges, or can be observed with a variety of adducts.
- Hopefully, different experiments will overlap in the features that they perceive.

‡ a “feature” is a bounded, two-dimensional (m/z and retention time) LC/MS signal
Tautenhahn et al. *BMC Bioinformatics* (2008) 9:504

Peak detection choice



Peak Picking

matchedFilter

- Profile Data
- Low resolution data
- Original algorithm

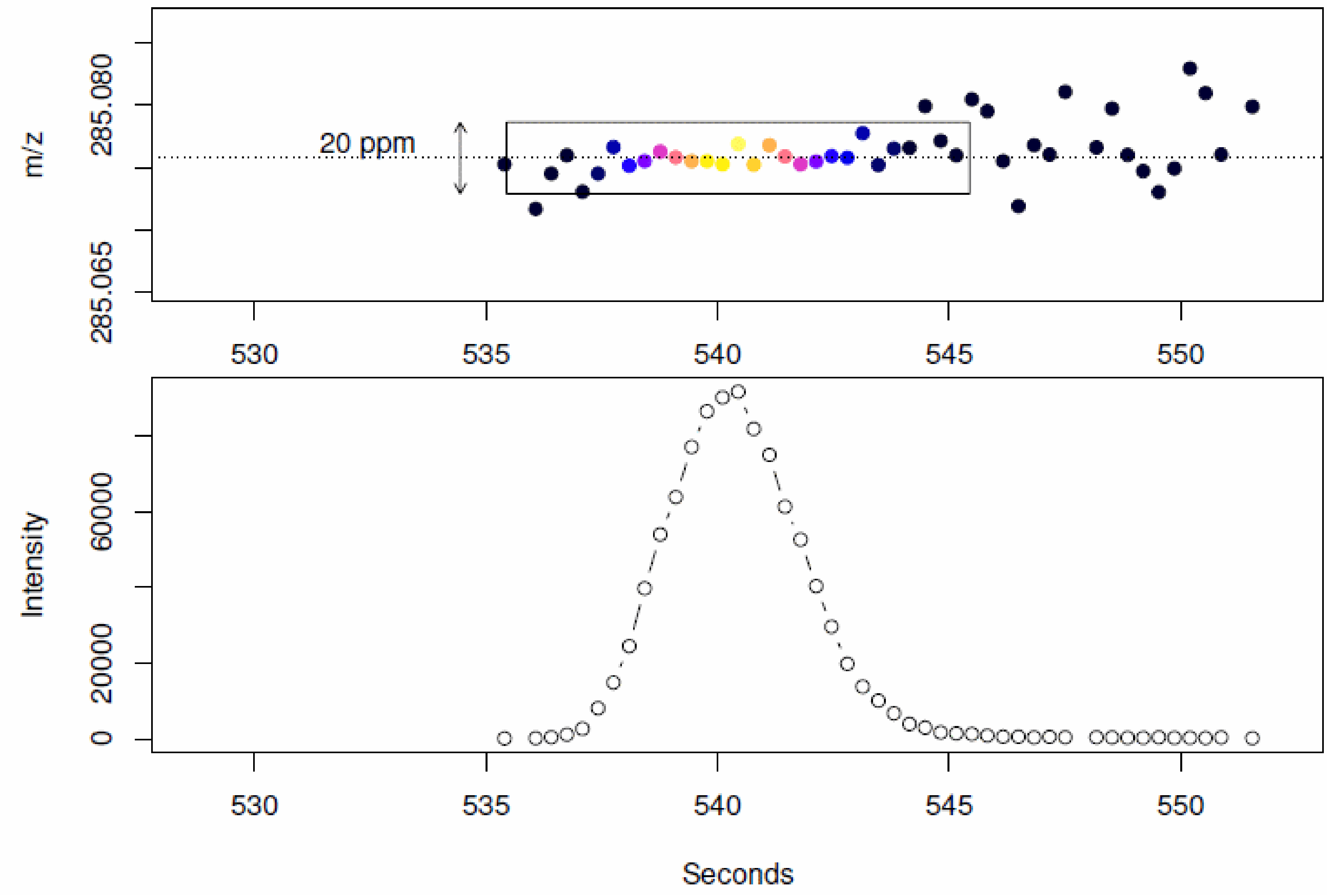
centWave

- Centroid data
- High resolution data
- New published algorithm

Smith et al. *Anal. Chem*. (2006) 78: 779

Tautenhahn et al. *BMC Bioinformatics* (2008) 9:504

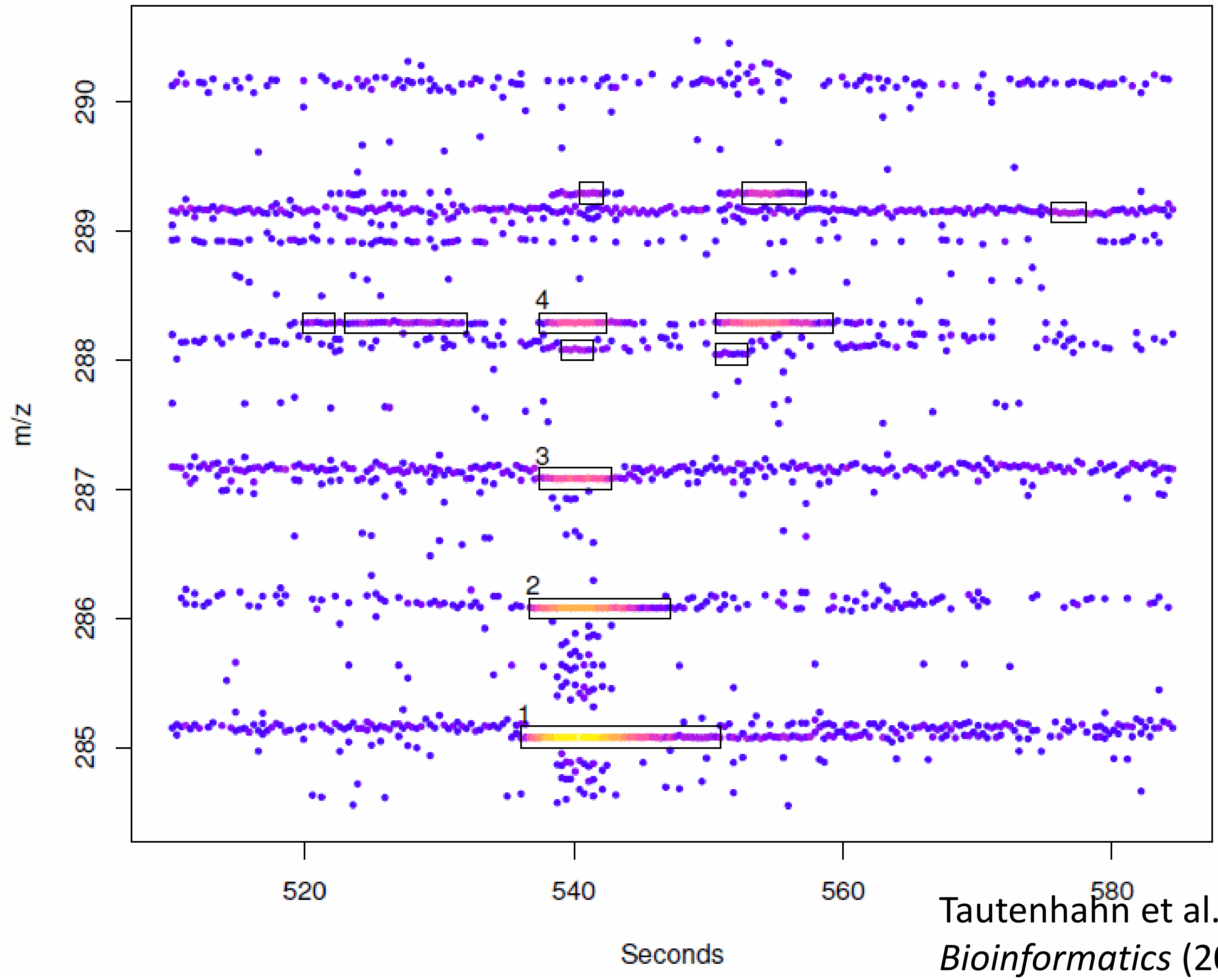
ROI‡ for biochanin A monoisotope



‡ ROI= “Region of Interest”

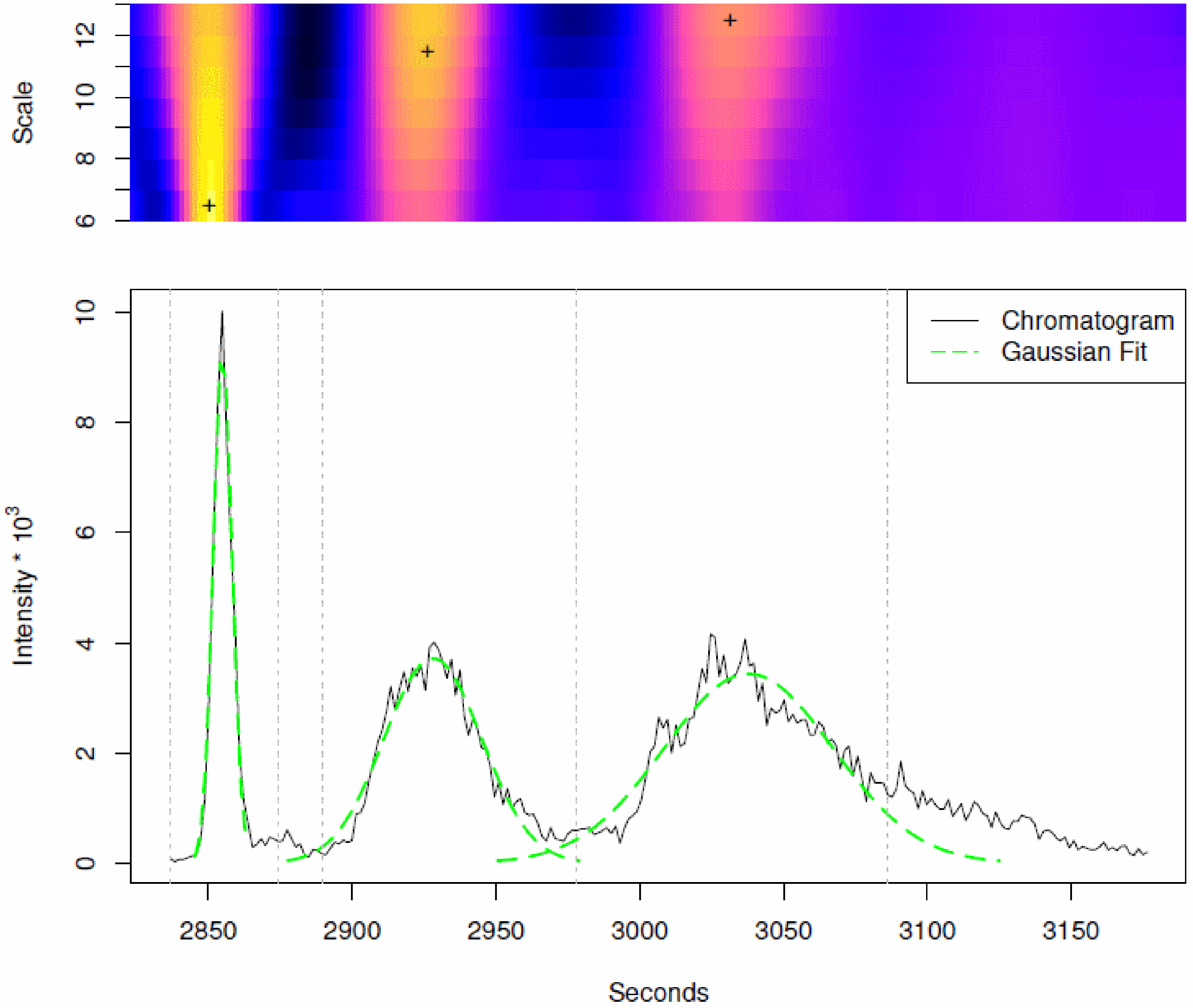
Tautenhahn et al. *BMC Bioinformatics* (2008) 9:504

Linking isotopes of biochanin A



Tautenhahn et al. *BMC Bioinformatics* (2008) 9:504

Matching wavelets to chromatograms



Tautenhahn et al. *BMC Bioinformatics* (2008) 9:504

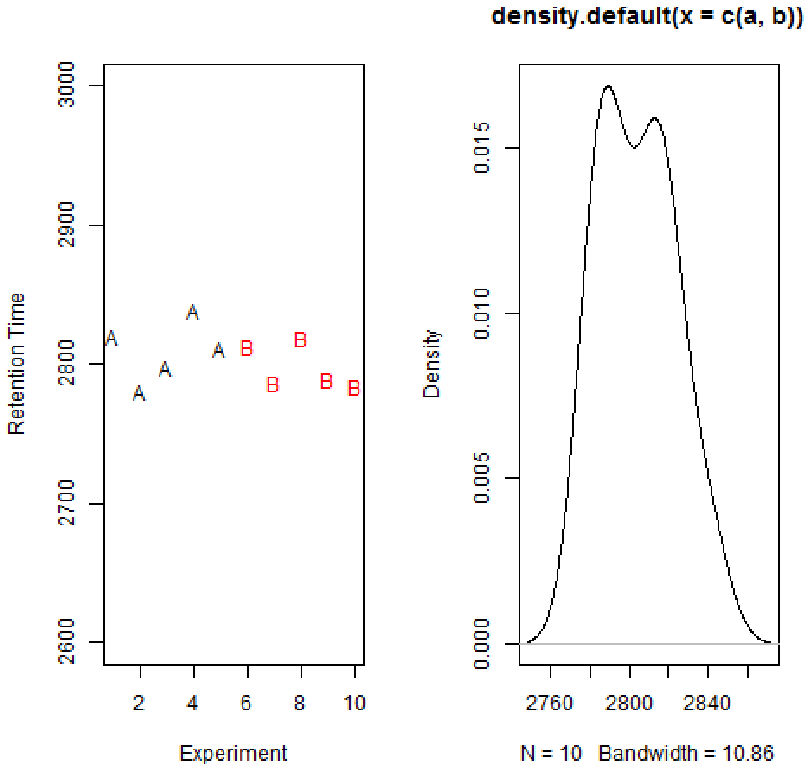
Why map features among expts?

- Ultimately, we want to compare intensity for a particular analyte among experiments.
- Intense features will act as landmarks to help manage LC variability.
- This initial “coarse matching” of “well-behaved” peak groups is the starting point for later retention time correlation.

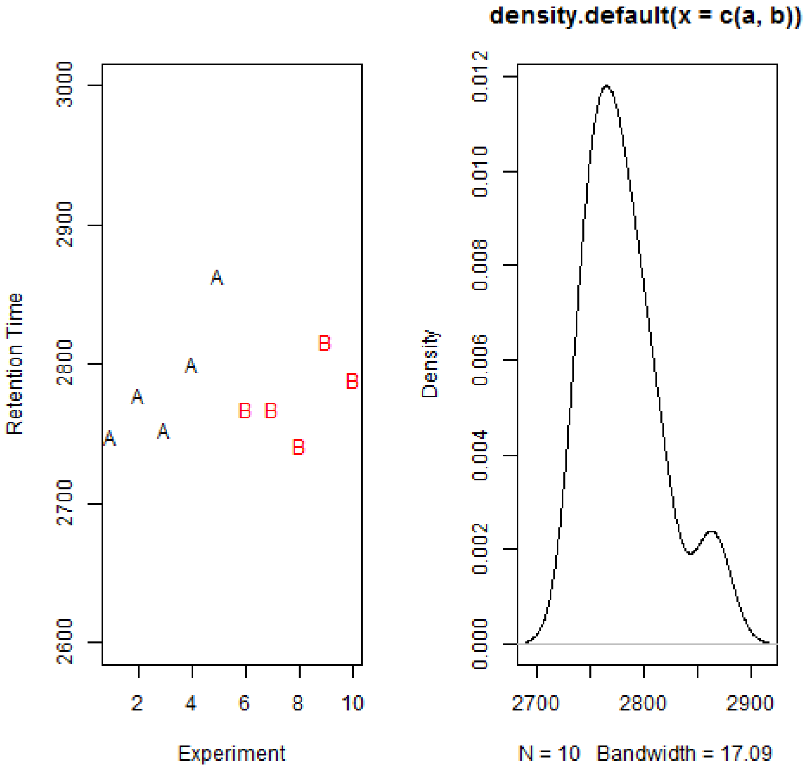
Smith et al. *Anal. Chem*. (2006) 78: 779

Cross-LC feature mapping

Low RT variability (sd=15 sec)



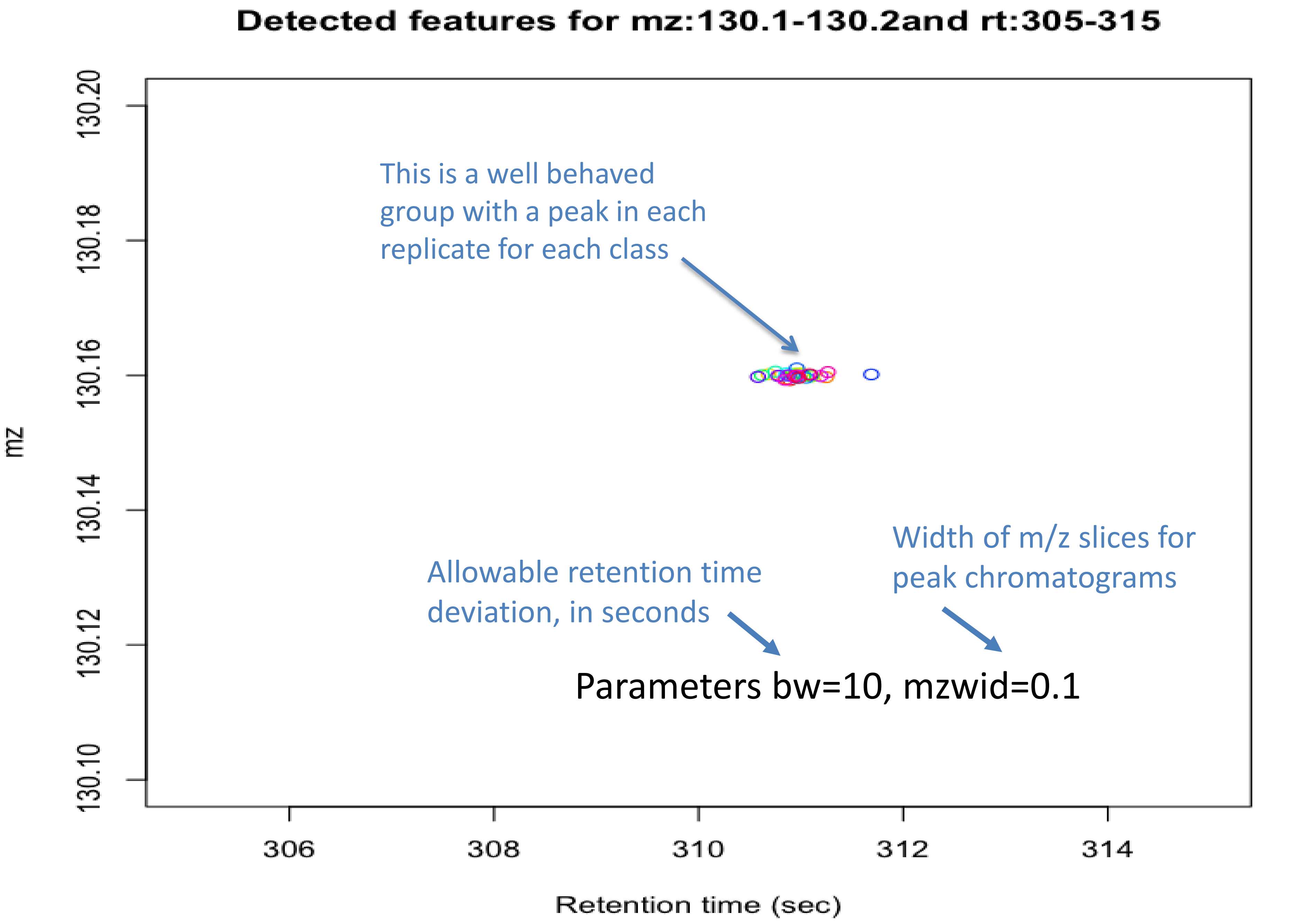
High RT variability (sd=45 sec)



Higher density implies greater commonality of retention times across files for this feature.
“MinFrac” defines the minimum fraction of samples in one cohort that must contain feature.

Issues with grouping features across experiments

- Feature grouping is highly dependent upon the parameters guiding the process.
  - Too tight a m/z tolerance could fail to find a feature even thought it was observed.
  - Too wide a RT tolerance could erroneously group an unrelated ion with others.
- Compounds of similar m/z and RT vex us
  - A major ion at the same RT may reduce other signals or be convoluted with our biomarker signal



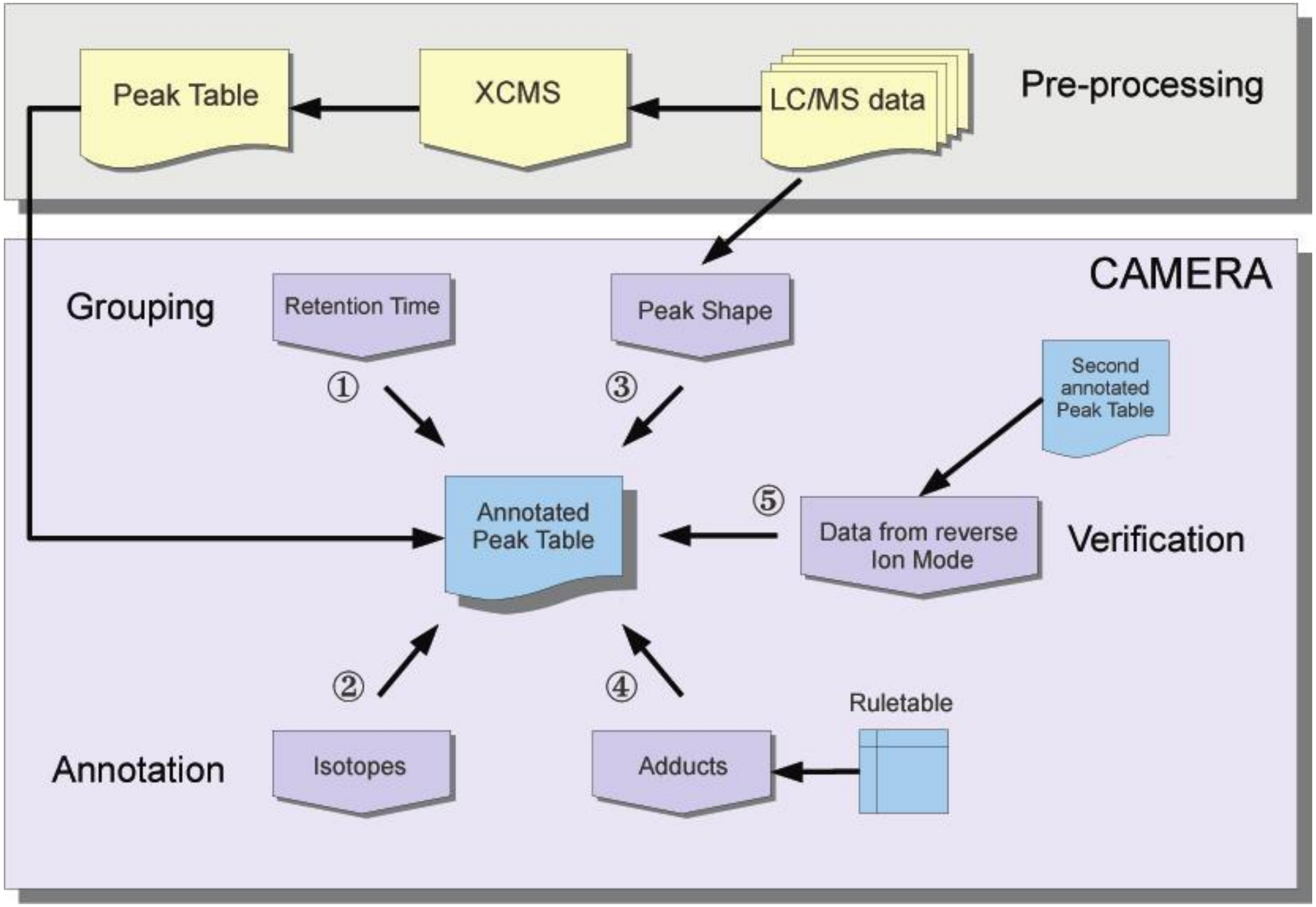
This is a well behaved group with a peak in each replicate for each class

Allowable retention time deviation, in seconds

Width of m/z slices for peak chromatograms

Parameters bw=10, mzwid=0.1

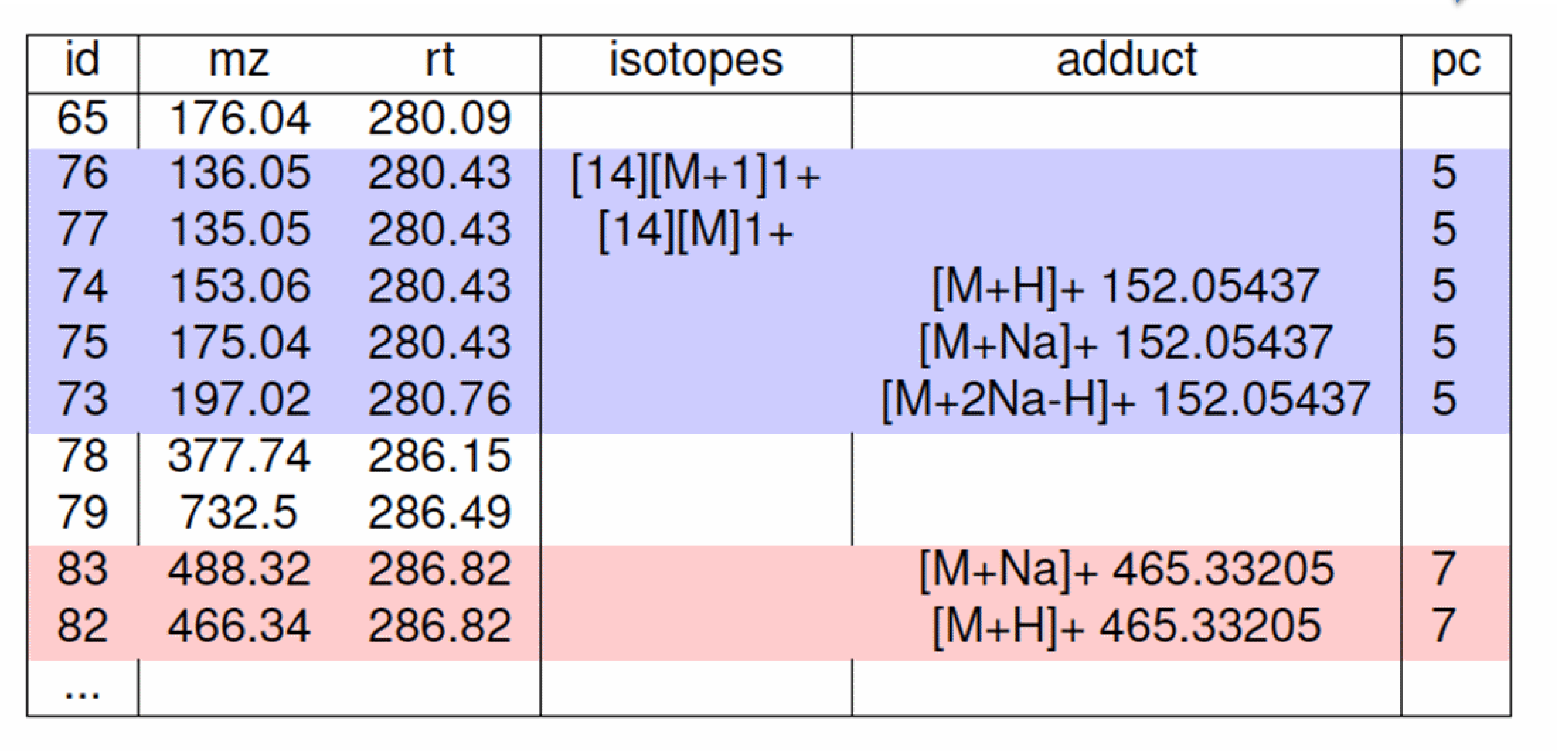
CAMERA associates sets of related ions



By recognizing multiple ions that correspond to a single chemical, CAMERA reduces the later burden of identifying differences.
Kuhl et al. *Anal. Chem*. (2012) 84: 283

One ion may produce multiple isotopes and adducts

PCA finds correlations

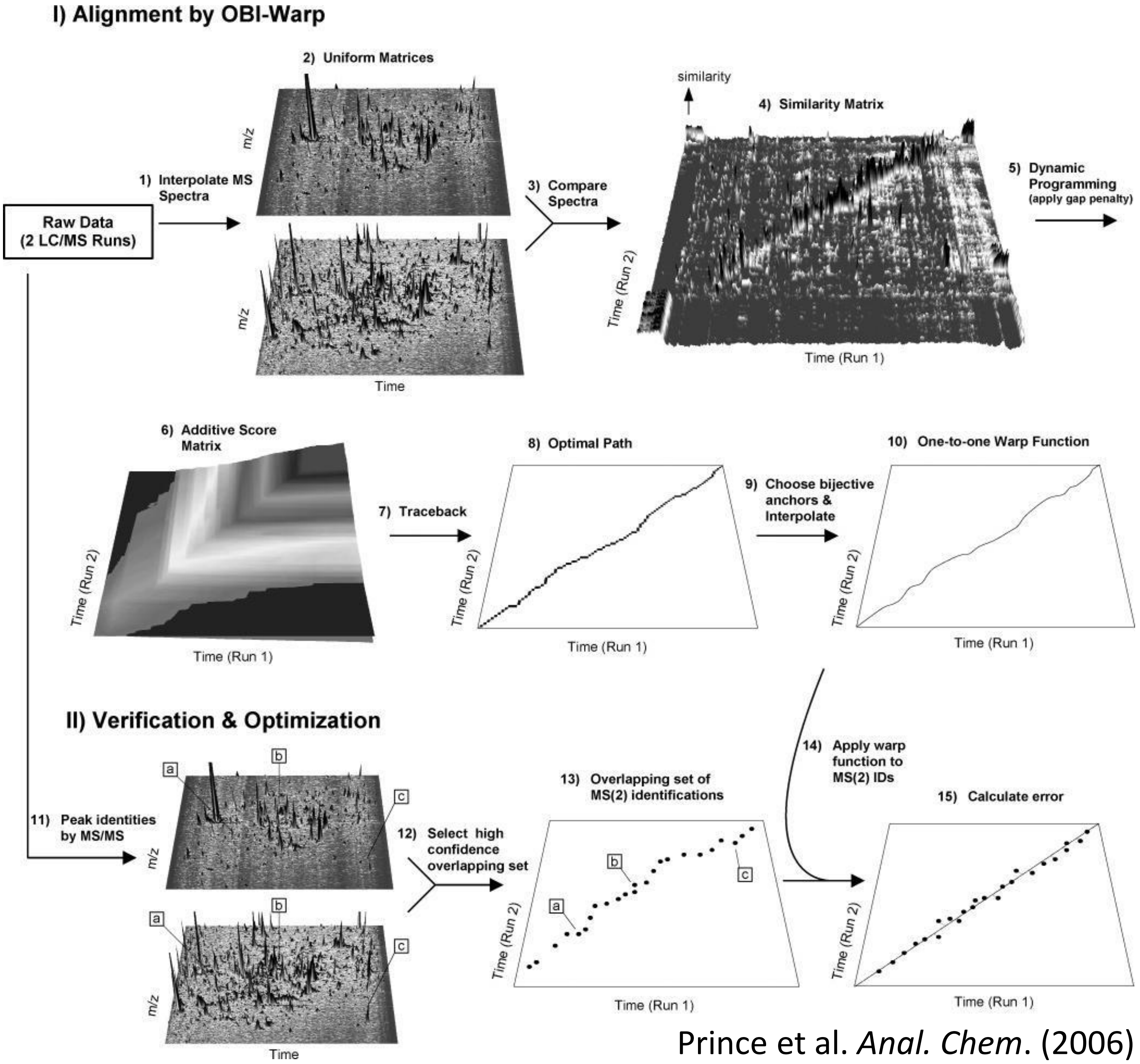


id	mz	rt	isotopes	adduct	pc
65	176.04	280.09			
76	136.05	280.43	[14][M+1]1+		5
77	135.05	280.43	[14][M]1+		5
74	153.06	280.43		[M+H]+ 152.05437	5
75	175.04	280.43		[M+Na]+ 152.05437	5
73	197.02	280.76		[M+2Na-H]+ 152.05437	5
78	377.74	286.15			
79	732.5	286.49			
83	488.32	286.82		[M+Na]+ 465.33205	7
82	466.34	286.82		[M+H]+ 465.33205	7
...					

What comes next?

From the initial grouping, the algorithm typically identifies hundreds of “well-behaved” peak groups in which very few samples have no peaks assigned and very few samples have more than one peak assigned. Such well-behaved groups have a high probability of being properly matched and can be used as temporary standards… Because the well-behaved peak groups are generally distributed evenly over the significant portions of the chromatographic profile, a detailed, nonlinear retention time deviation contour can be built for each sample.

Smith et al. *Anal. Chem*. (2006) 78: 779



Prince et al. *Anal. Chem*. (2006) 78: 6140

Why did we do all of that?

- Accounting for run-to-run LC variability lets us find differences much more sensitively.
- Dynamic programming adjusts the data in reasonable amounts of time.
- A simple linear fit (scaling retention time linearly and adjusting the intercept) would not account for small time-scale variances.

Good XCMS habits

- Visualize your RAW data before starting XCMS.
- If you use XCMS on your own computer, copy your R scripts to a file for reproducibility.
- Save your objects, not the whole workspace
  - `save(xset, file=“xset.RData”)`
- Optimize your parameters to get the best results
- Rubbish in, rubbish out

Takeaway messages

- To find biomarkers, XCMS must first reduce retention time variability, increasing comparability among experiments.
- Ions with high signal-to-noise act as landmarks to relate each LC-MS experiment to others.
- Compounds may be found in multiple isotopes, charge states, and adduct variants.