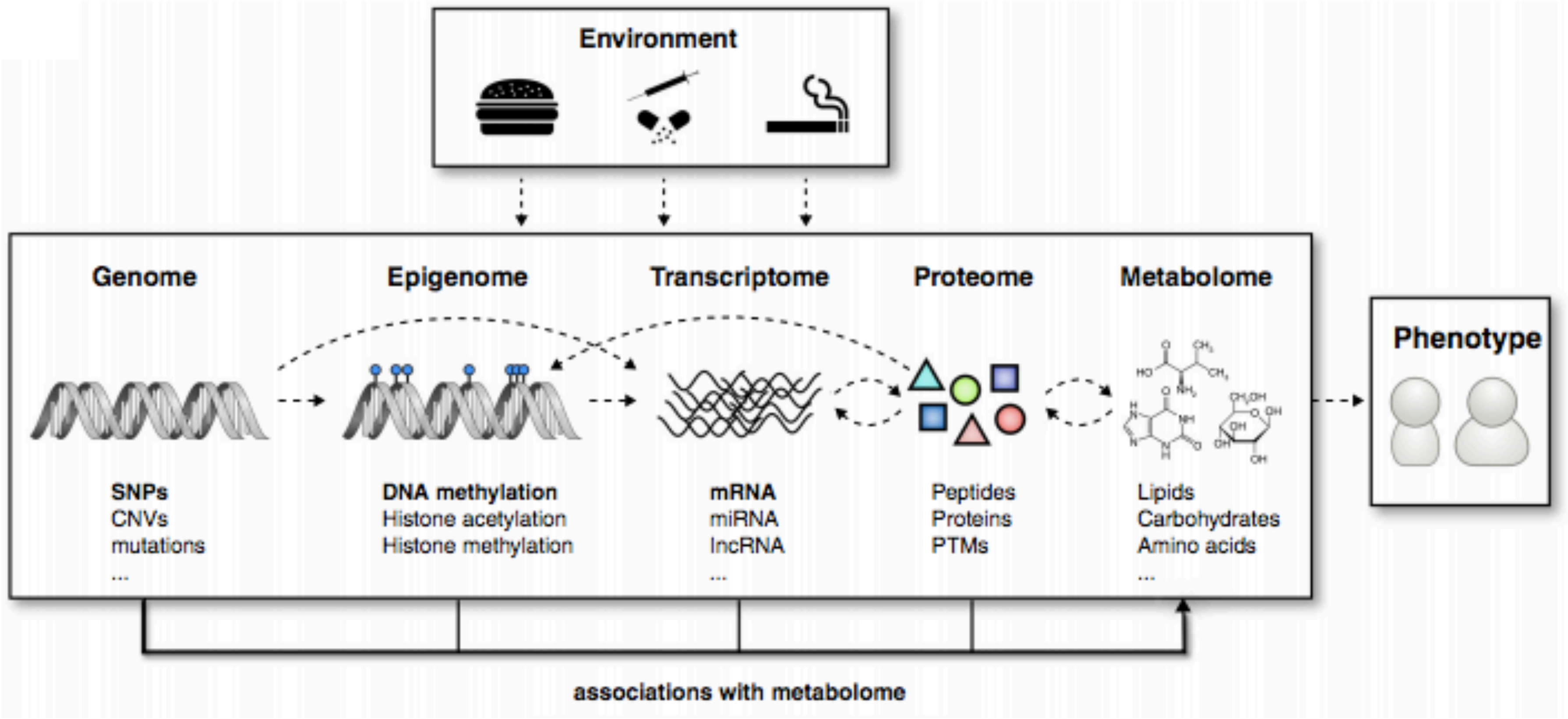
Intro to MS data processing for metabolomics and lipidomics

Ilya Kurochkin

May 6, 2017

Systems Biology

-omics, environment and phenotype



DNA tells what is possibly…

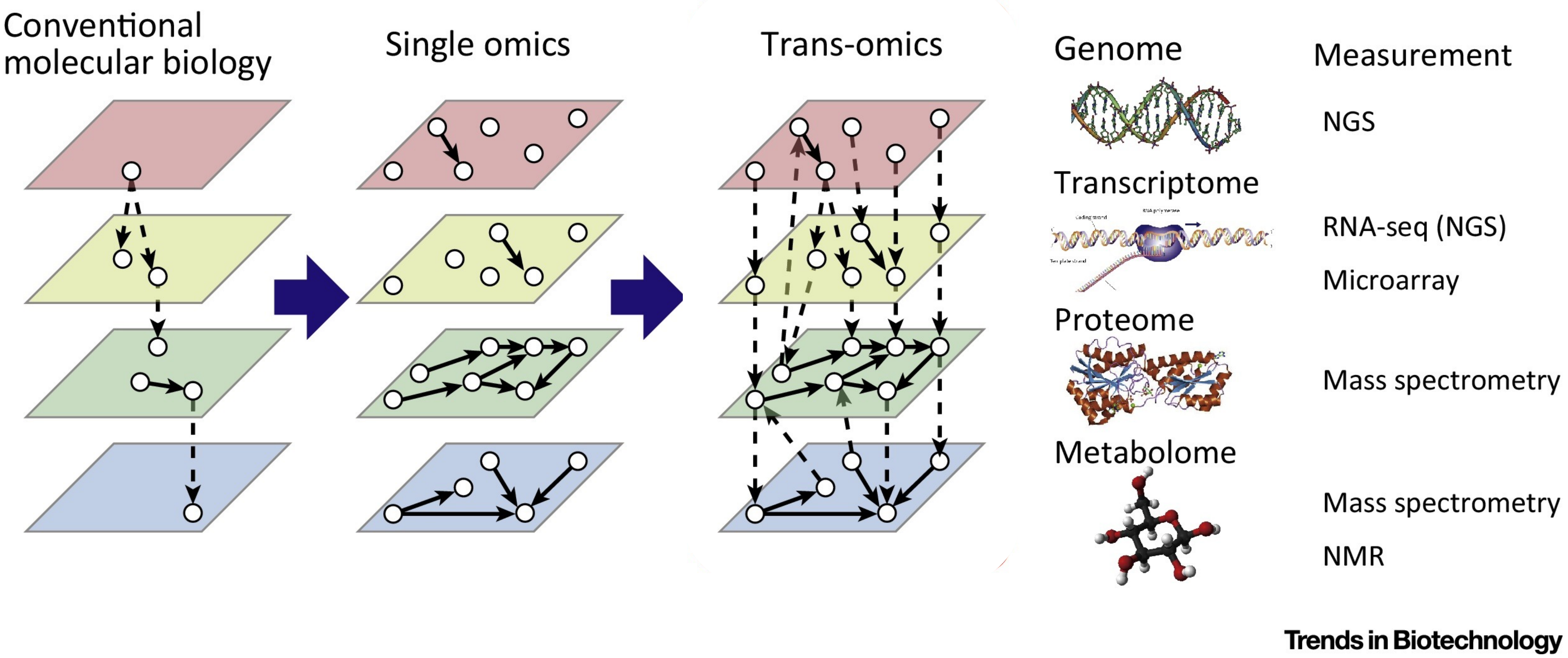
…RNA what is probably….

…proteins and metabolites what actually happens

Krumsiek 2016

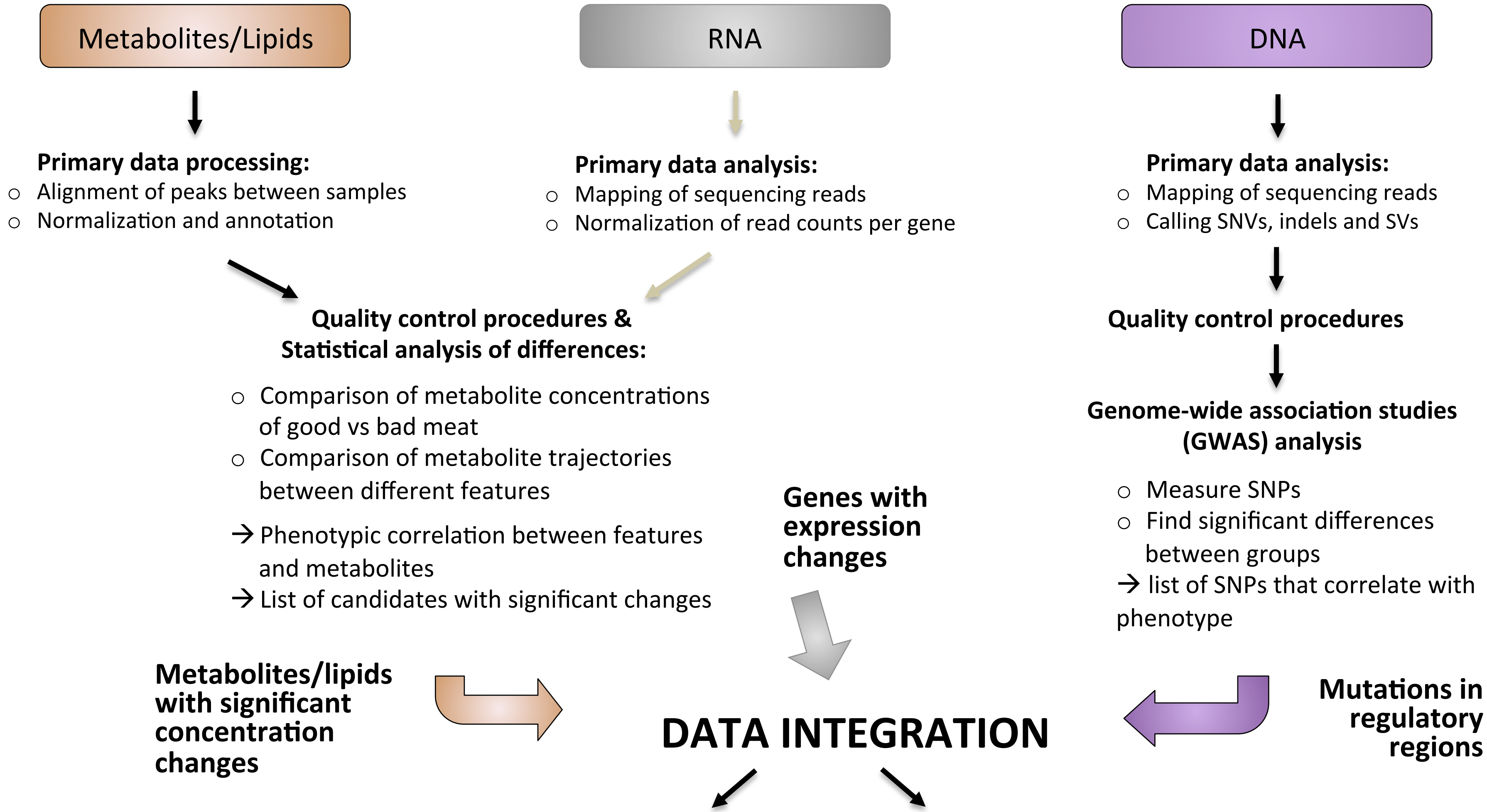
Systems Biology

Technologies for different -omics layers



	Measurement
Genome	NGS
Transcriptome	RNA-seq (NGS) Microarray
Proteome	Mass spectrometry
Metabolome	Mass spectrometry NMR

Yugi, Katsuyuki et al.
Trends in Biotechnology , Volume 34 ,
Issue 4 , 276 - 290



Metabolites/Lipids

RNA

DNA

Primary data processing:

- Alignment of peaks between samples
- Normalization and annotation

Primary data analysis:

- Mapping of sequencing reads
- Normalization of read counts per gene

Primary data analysis:

- Mapping of sequencing reads
- Calling SNVs, indels and SVs

Quality control procedures & Statistical analysis of differences:

- Comparison of metabolite concentrations of good vs bad meat
- Comparison of metabolite trajectories between different features

→ Phenotypic correlation between features and metabolites

→ List of candidates with significant changes

Quality control procedures

Genome-wide association studies (GWAS) analysis

- Measure SNPs
- Find significant differences between groups

→ list of SNPs that correlate with phenotype

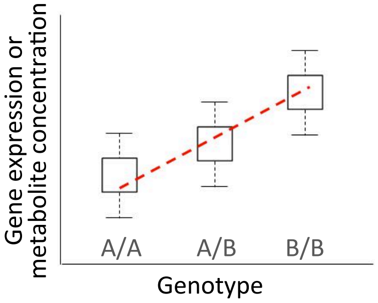
Genes with expression changes

Metabolites/lipids with significant concentration changes

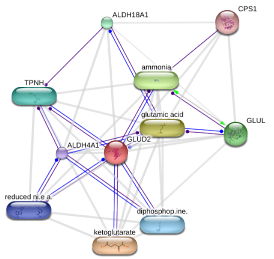
DATA INTEGRATION

Mutations in regulatory regions

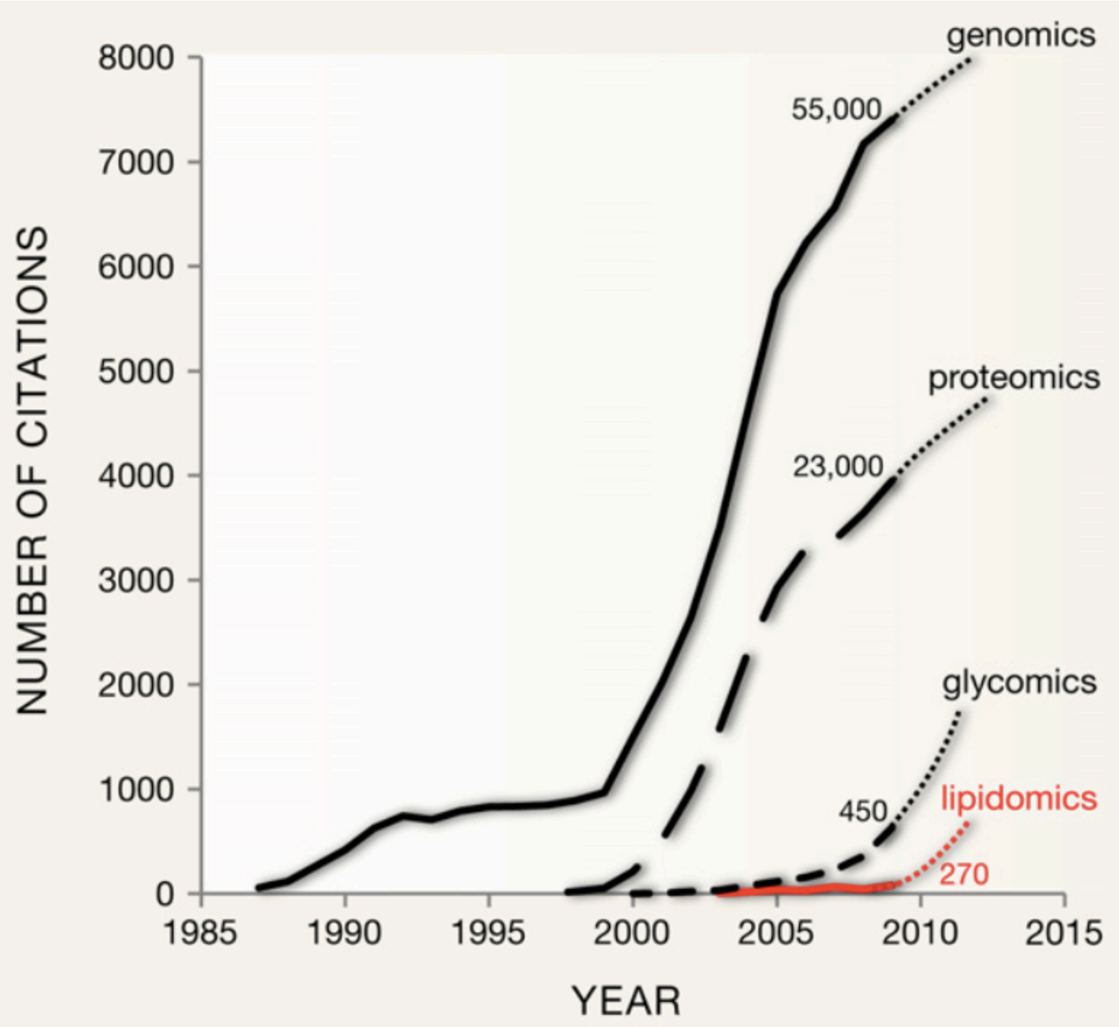
Metabolic QTL



Pathway Enrichment Analysis



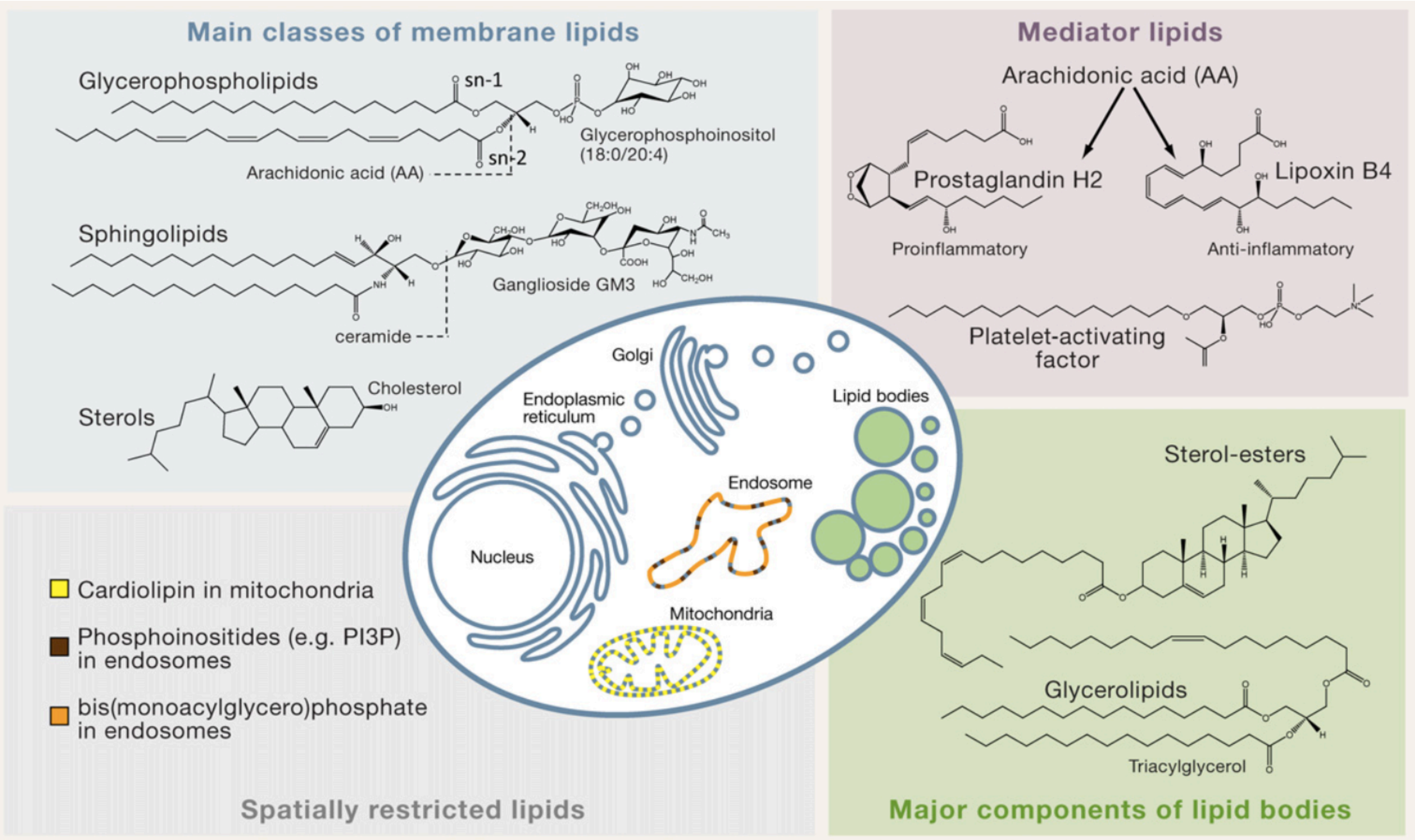
Lipidomics is a young, emerging field



Wenk 2010 – Cell 143(6):888-95
Lipidomics: New Tools and Applications

Lipids

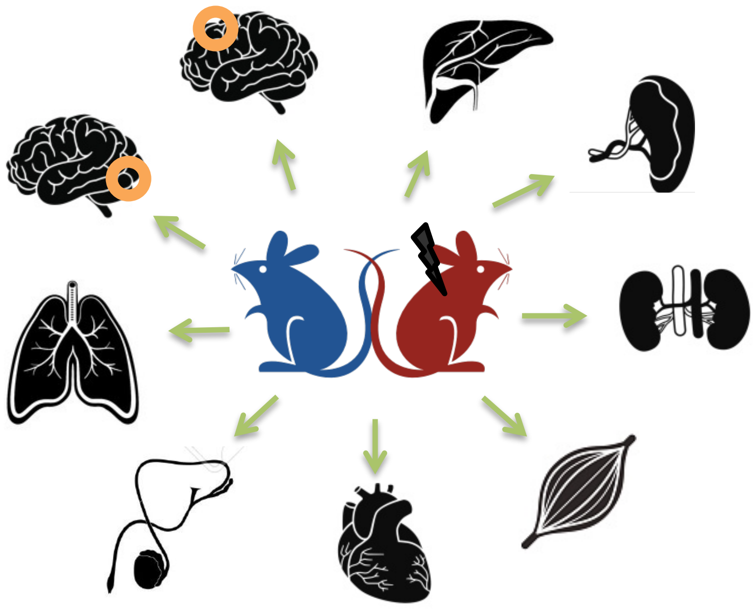
Cellular Compartments of Common Biological Lipids



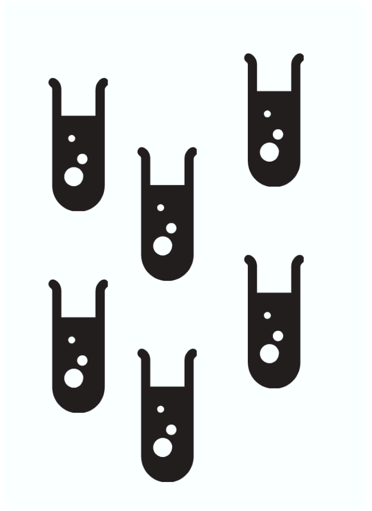
Wenk 2010 – Cell 143(6):888-95
Lipidomics: New Tools and Applications

Study designs

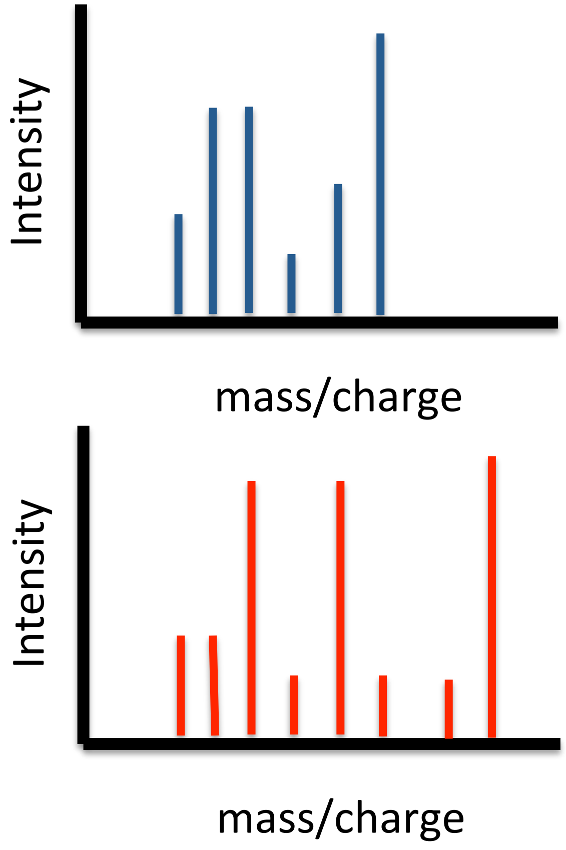
Tissue collection



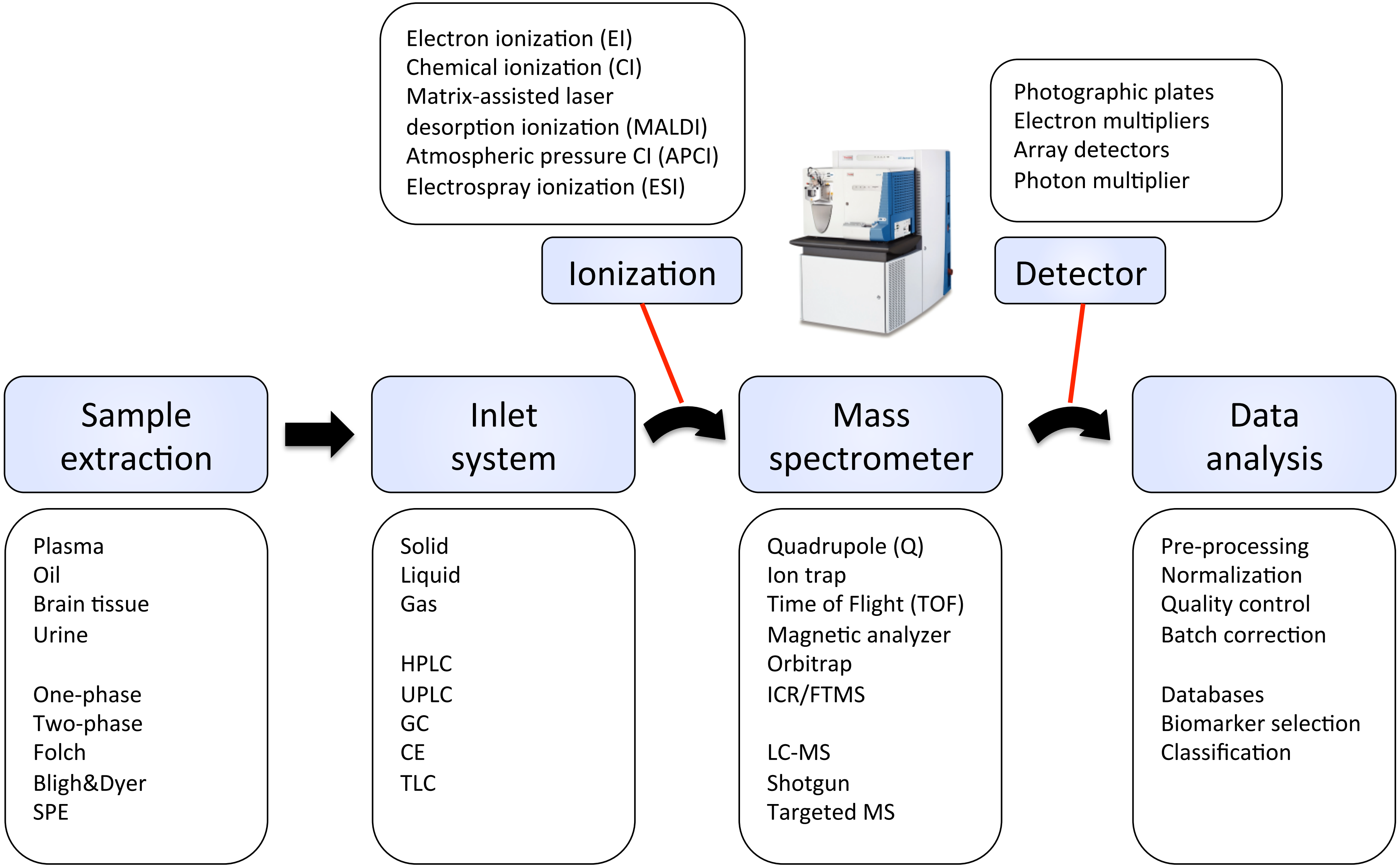
Extraction of metabolites



Mass spectrometry analysis



Mass spectrometry
Workflow and variety



Electron ionization (EI)
Chemical ionization (CI)
Matrix-assisted laser desorption ionization (MALDI)
Atmospheric pressure CI (APCI)
Electrospray ionization (ESI)

Ionization

Photographic plates
Electron multipliers
Array detectors
Photon multiplier

Detector

Sample extraction

Plasma
Oil
Brain tissue
Urine

One-phase
Two-phase
Folch
Bligh&Dyer
SPE

Inlet system

Solid
Liquid
Gas

HPLC
UPLC
GC
CE
TLC

Mass spectrometer

Quadrupole (Q)
Ion trap
Time of Flight (TOF)
Magnetic analyzer
Orbitrap
ICR/FTMS

LC-MS
Shotgun
Targeted MS

Data analysis

Pre-processing
Normalization
Quality control
Batch correction

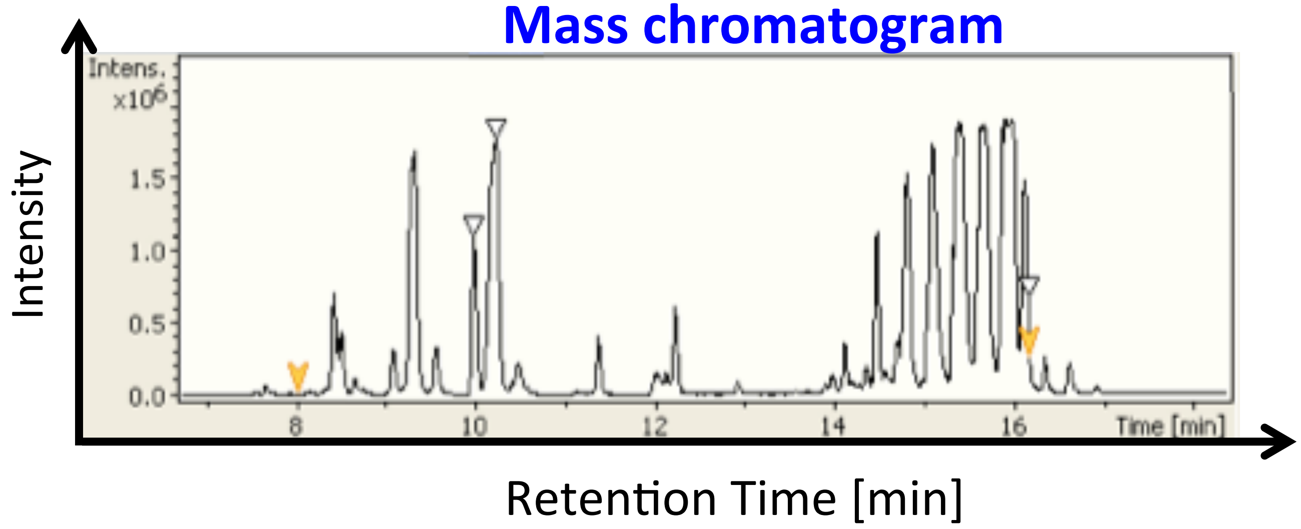
Databases
Biomarker selection
Classification

The data

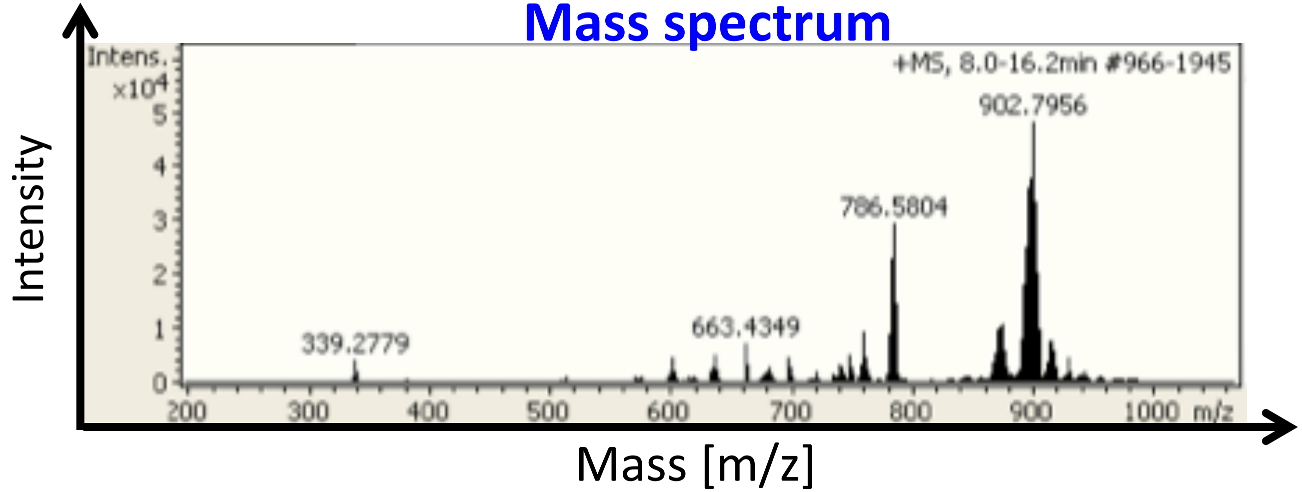
How does LC-MS data look like?

Chromatogram vs Spectrum

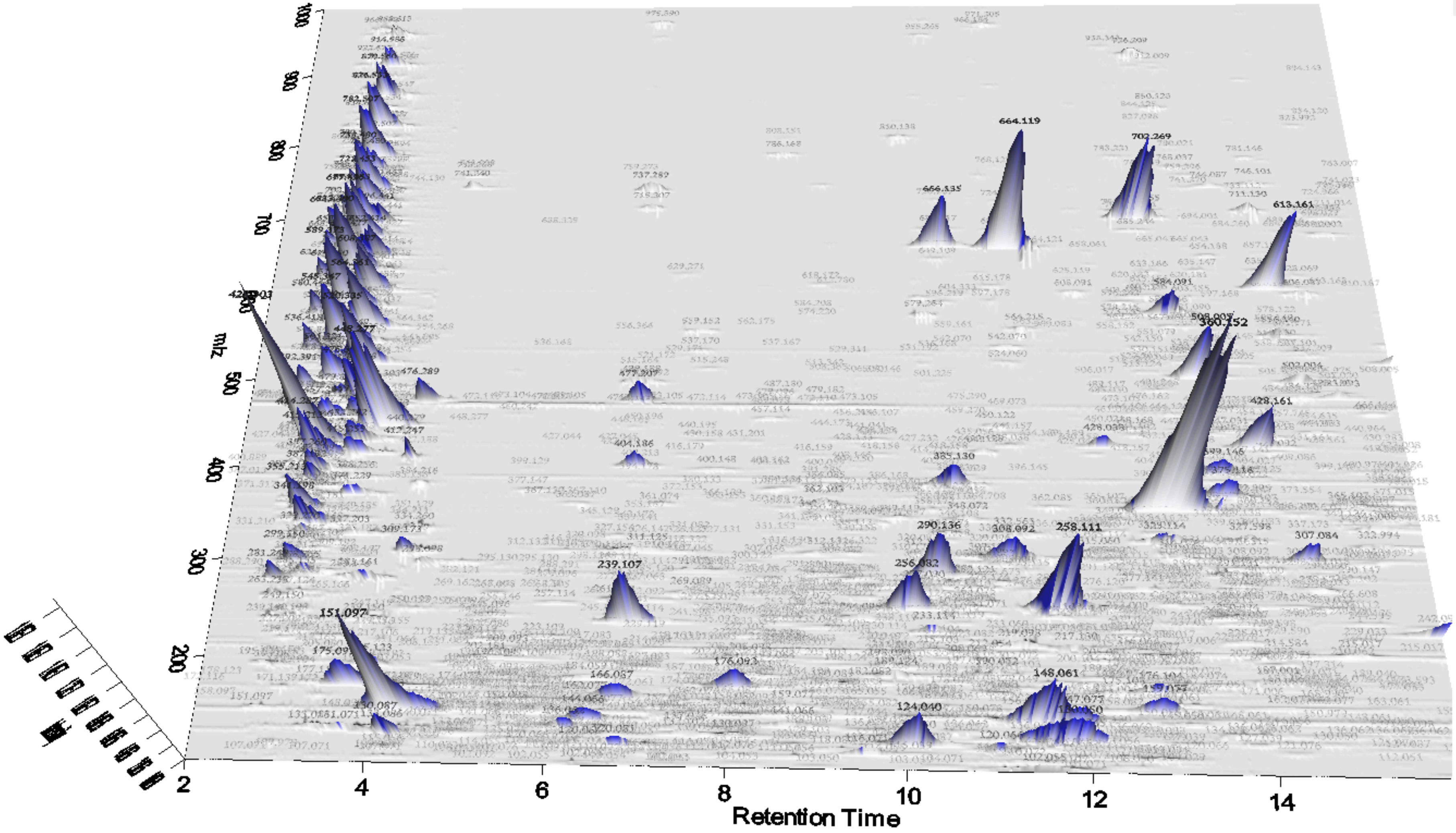
Mass chromatogram



Mass spectrum



Data representation



http://mzmine.github.io/

How does LC-MS data look like?

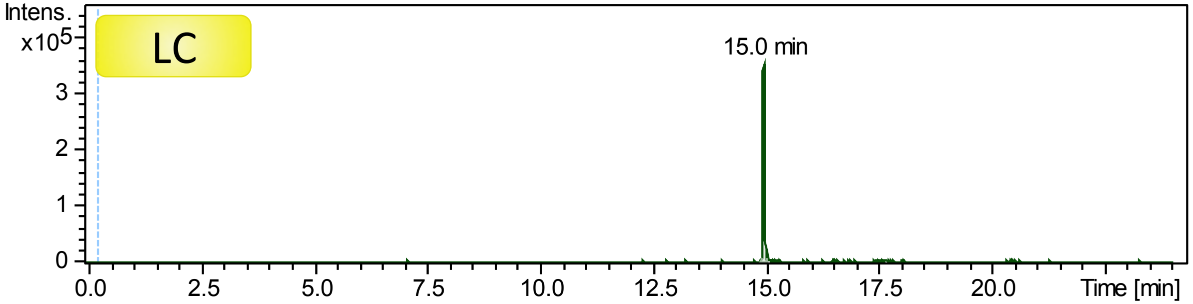
Zooming in.....

TAG(15:0/18:1–d7/15:0)

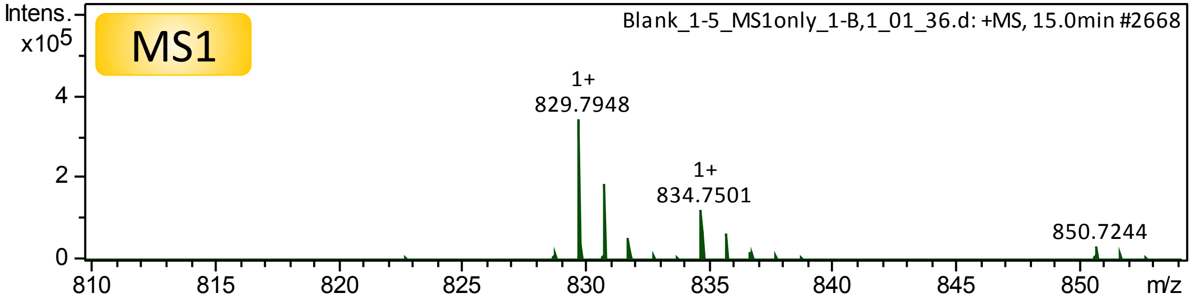
$C_{51}H_{89}D_7O_6$

Neutral Mass: 811.76465

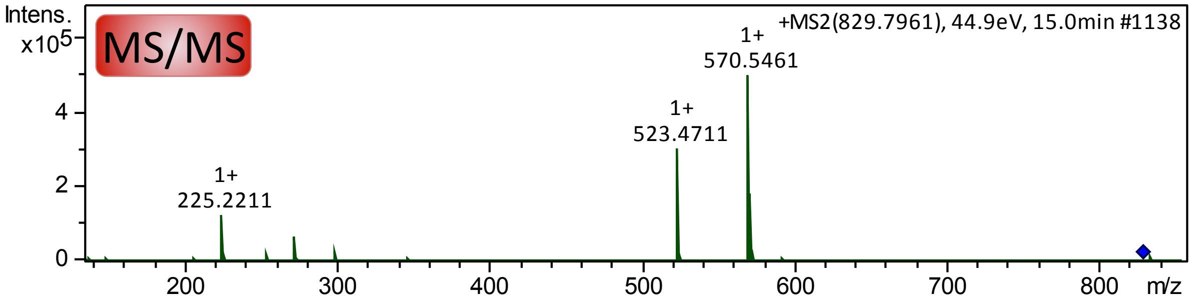
Chromatogram



Mass spectrum

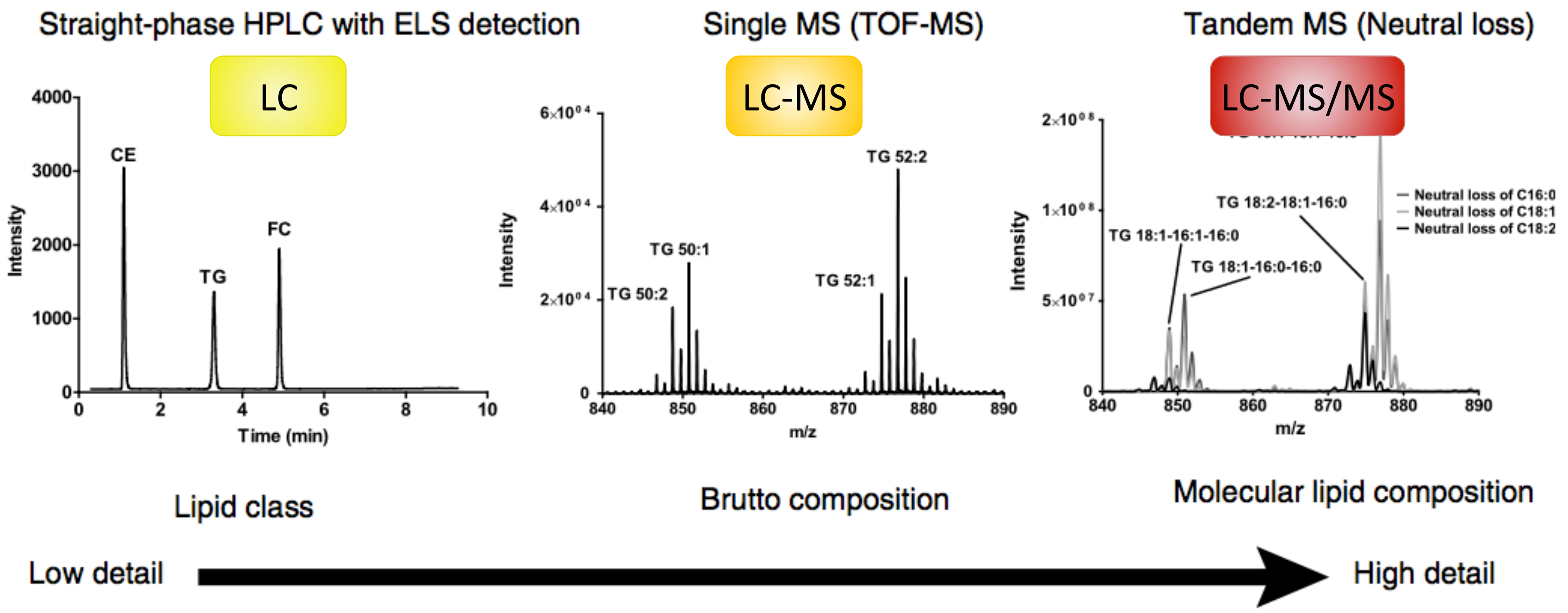


Fragment mass spectrum



How do you get from data to compound?

Why tandem MS (MS/MS)?



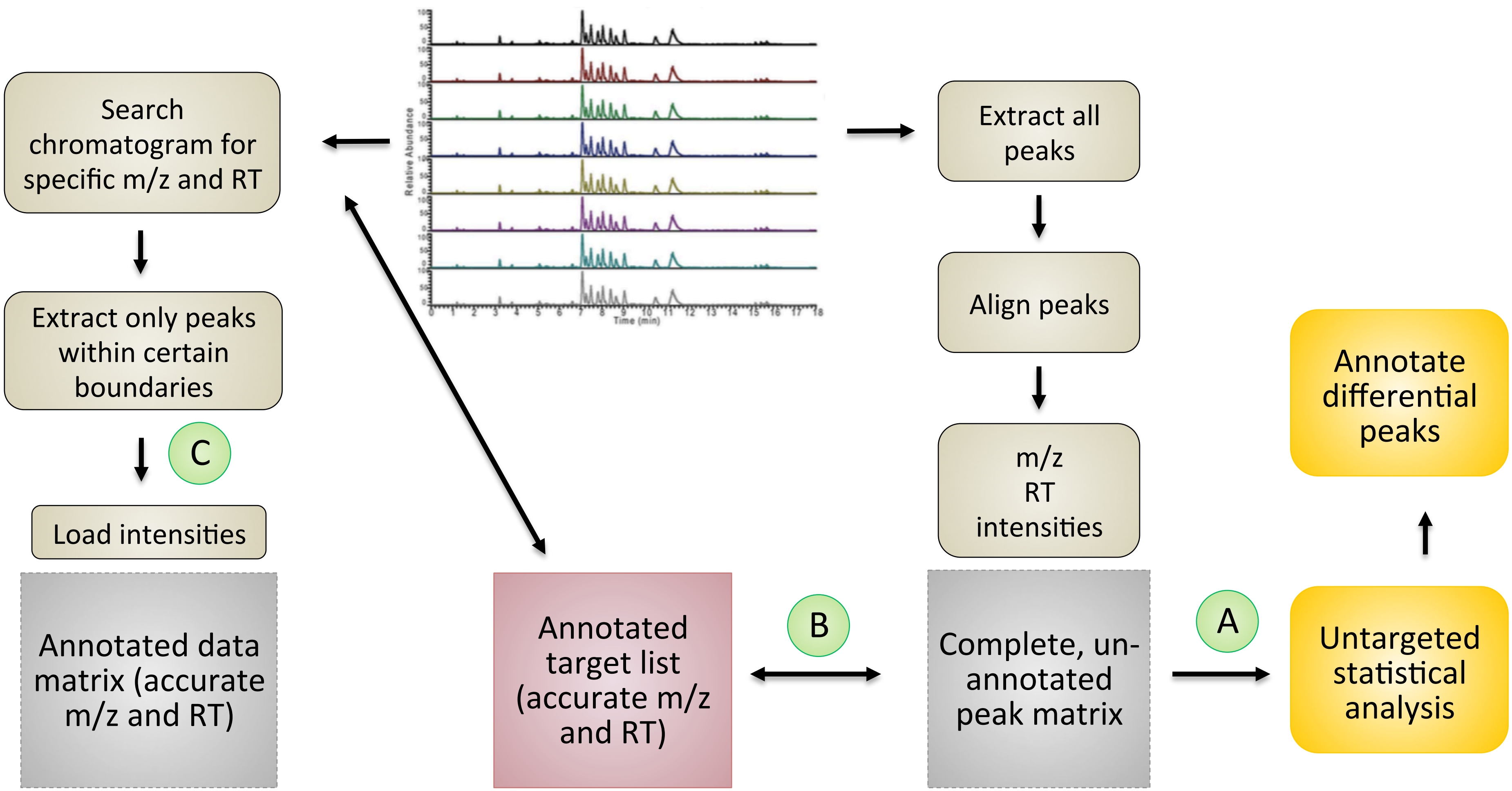
- LC-MS allows for elucidation of molecular mass and most of the times brutto composition, but little structural information

- LC-MS/MS allows for the 1) detection of structurally informative fragment ions, and 2) the confirmation of ambiguous annotation of lipid species

Ekroos, K. (2013). Lipidomics - Technologies and Applications. (K. Ekroos, Ed.) (pp. 1–345). Wiley-VCH Verlag GmbH & Co. KGaA.

How do you get from data to compound?

LC-MS strategies (MS1)



Chromatograms of lipids samples

Search chromatogram for specific m/z and RT

Extract only peaks within certain boundaries

C

Load intensities

Annotated data matrix (accurate m/z and RT)

Annotated target list (accurate m/z and RT)

B

Extract all peaks

Align peaks

m/z
RT
intensities

Complete, un-annotated peak matrix

A

Untargeted statistical analysis

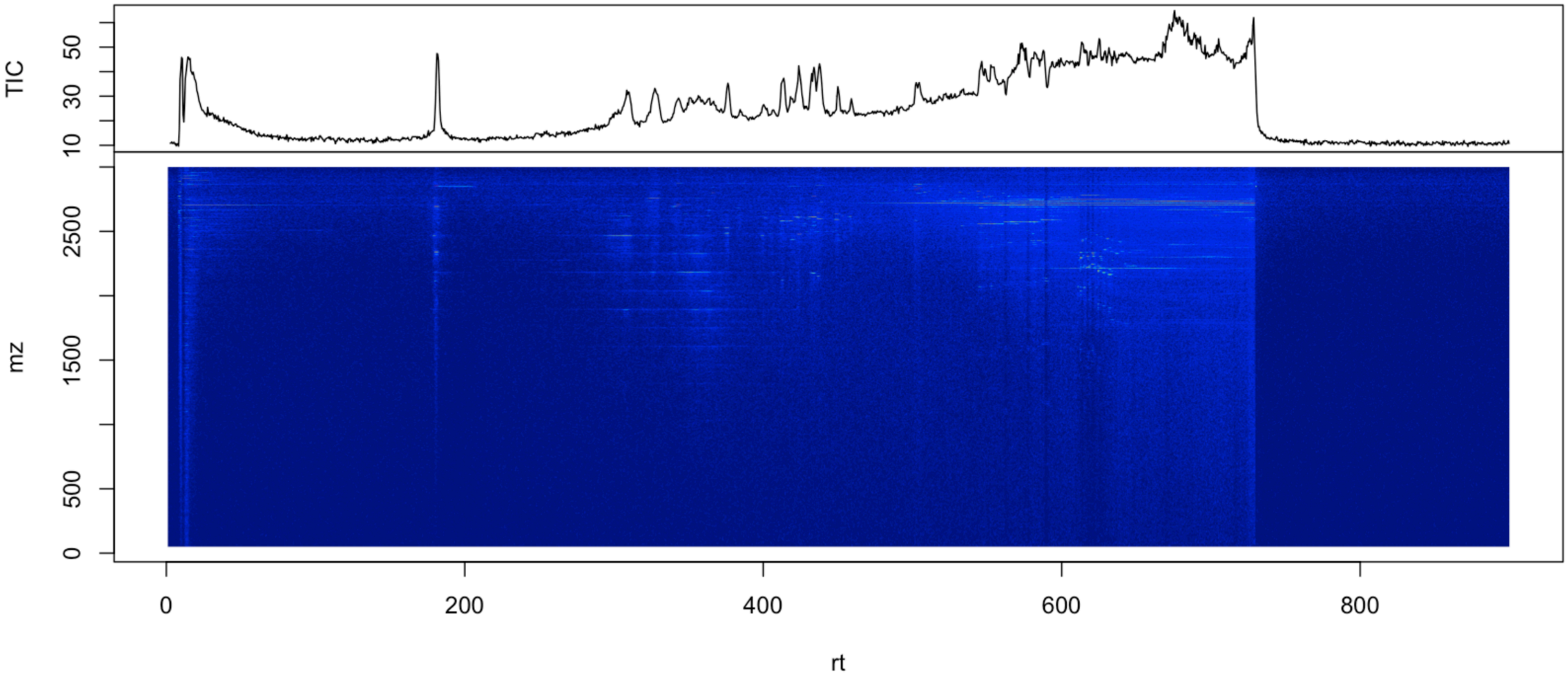
Annotate differential peaks

“Targeted” strategy

Untargeted strategy

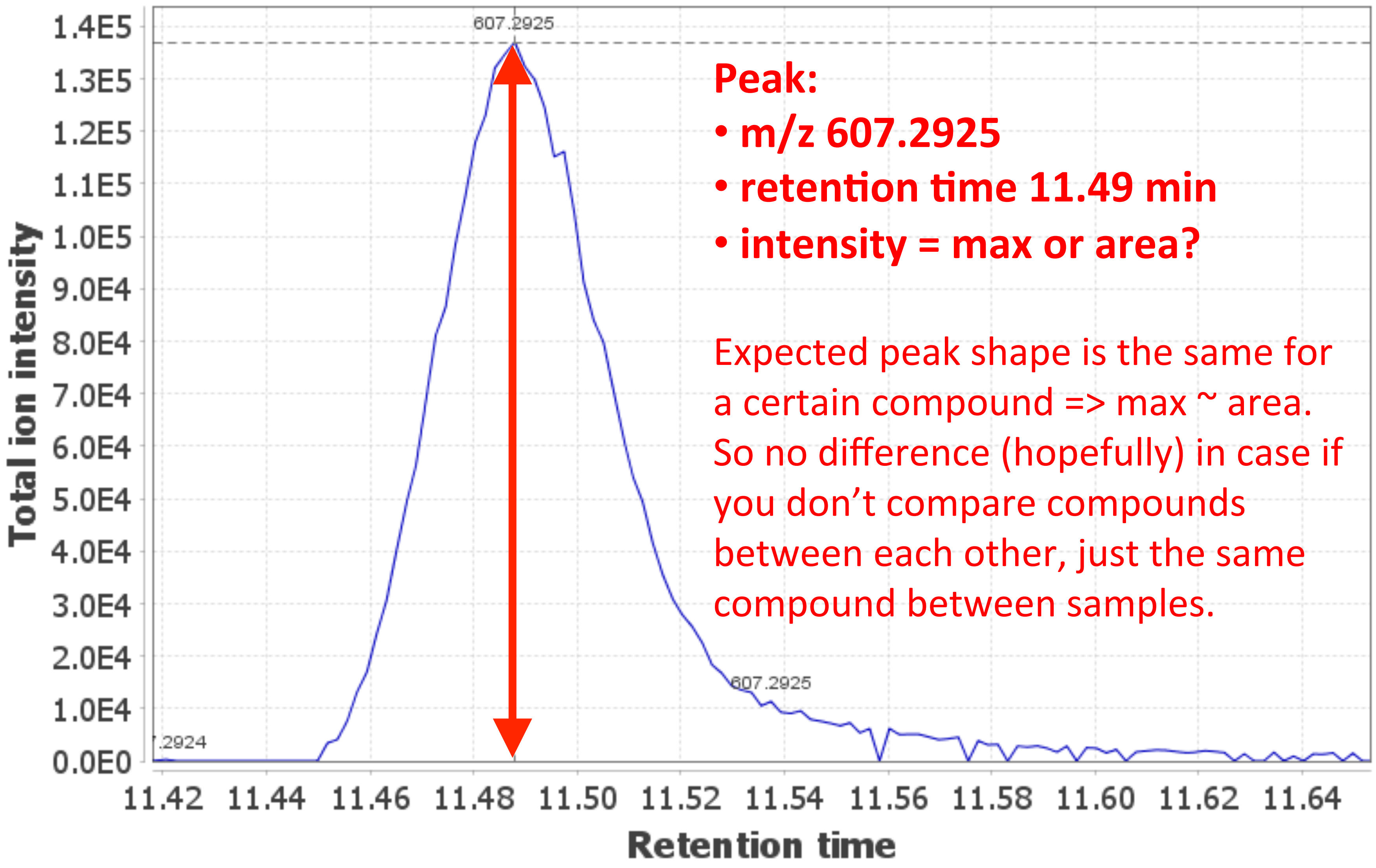
Data Processing

Data representation



https://www.ebi.ac.uk/training

LC/MS: Extracted Ion Chromatogram



Peak:

- **m/z 607.2925**
- **retention time 11.49 min**
- **intensity = max or area?**

Expected peak shape is the same for a certain compound => max ~ area. So no difference (hopefully) in case if you don't compare compounds between each other, just the same compound between samples.

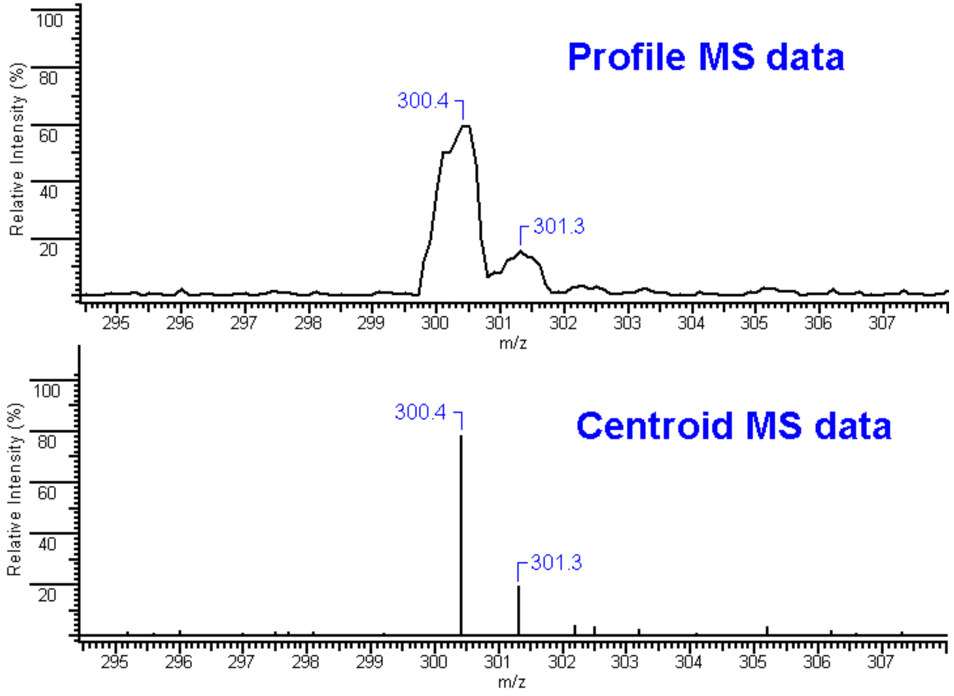
Data types

Profile vs Centroid

Profile data (aka continuous) – intensity records for all the range of mz and retention time (RT).

Centroided data – only local maximums are detected and saved.

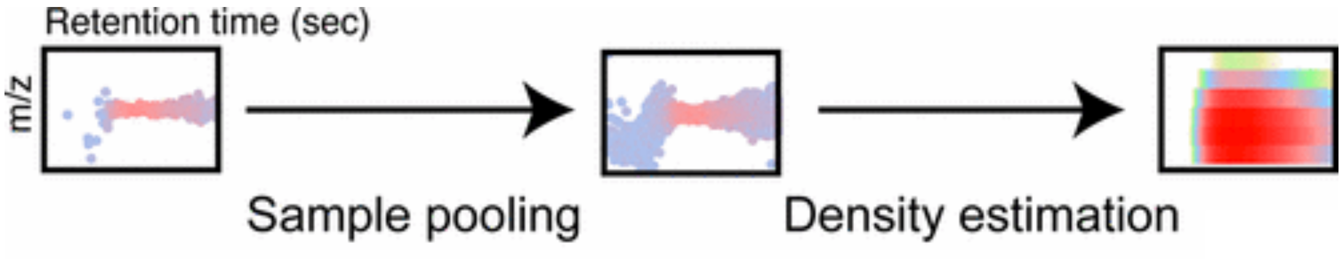
- Pros profile:
  - More options for peak detection, better detection
  - Less ambiguous => less false positive values
- Cons profile:
  - Big data volume
  - Slow conversion and analysis



Two paradigms

1. Peak picking then alignment (Do peak picking for each sample separately)

2. Alignment and peak picking (Do peak picking on each sample simultaneously)



Peak Picking / Peak Detection

Methods

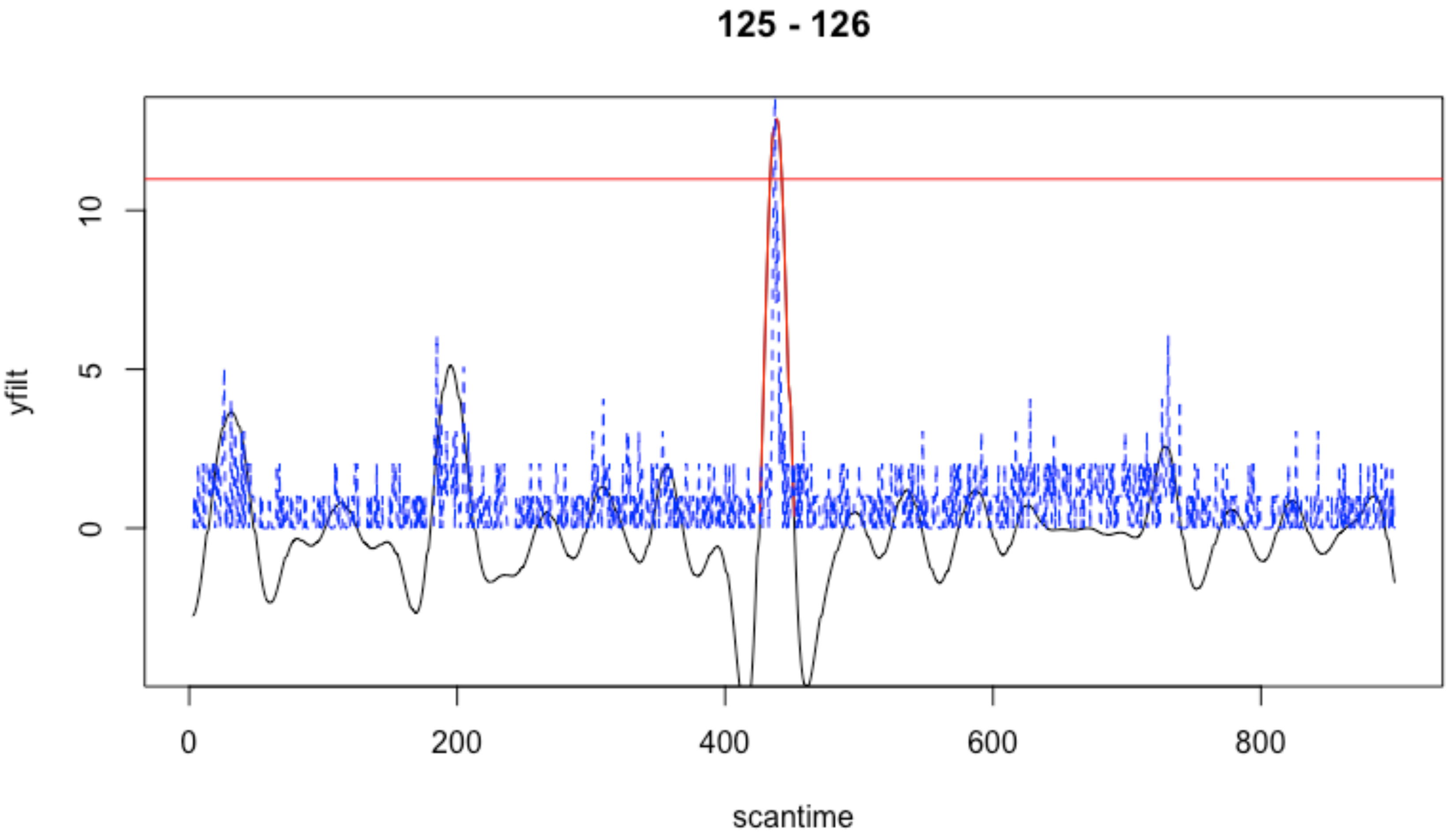
- In literature there are a lot of different peak picking algorithm. But no best solution, only better solutions.
- *Know your data!*
- Gaussian model peak width – standard

Smoothing, baseline correction may be applied, not for all methods.

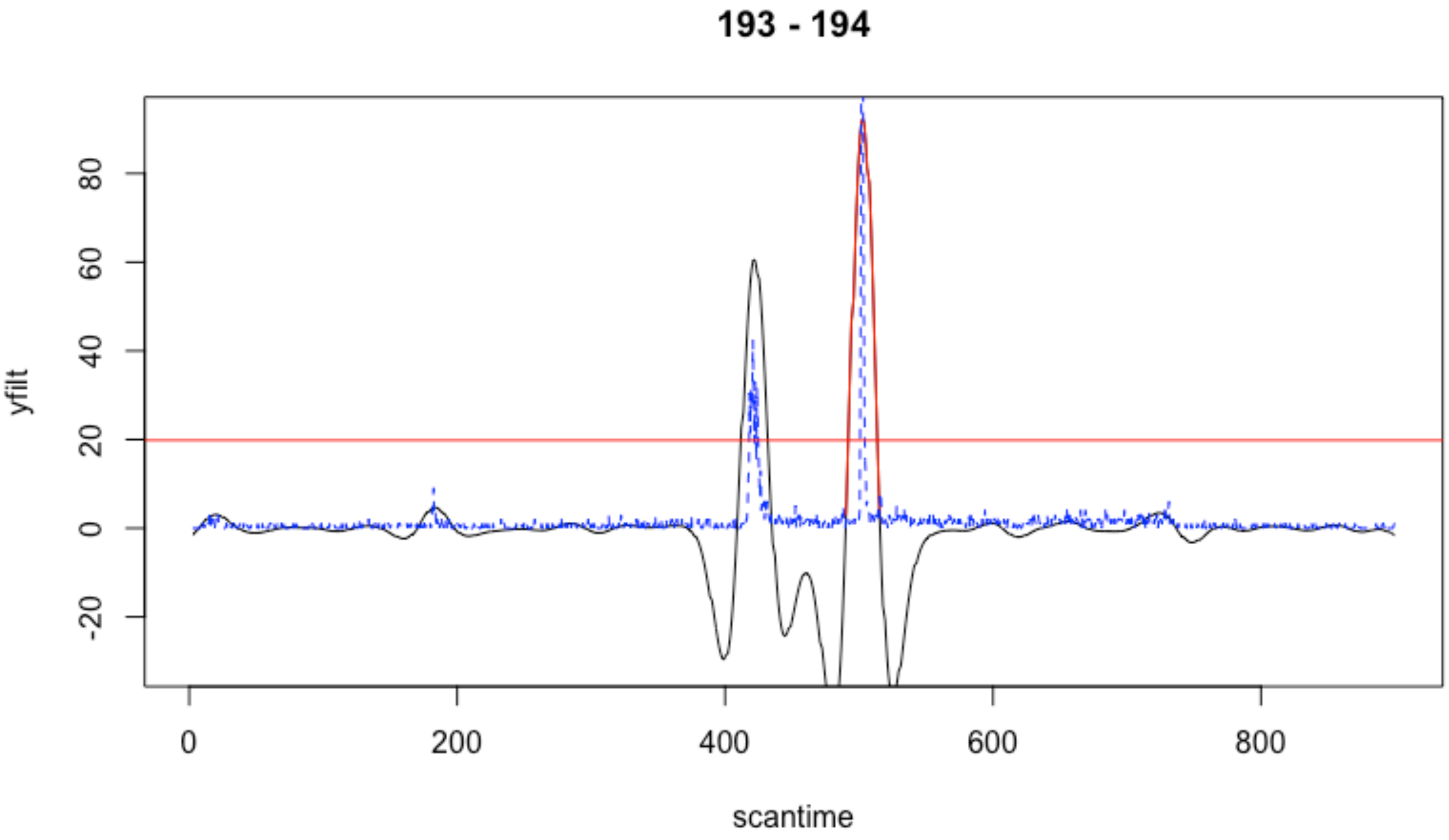
Peak picking is a crucial step of analysis. The main question: how to choose method and parameters?

- Tryout => tradition
- Repeating for others
- Attempt to define objective metrics of peak-picking quality and build a parameter selection based on their maximization: Brodsky L. et al. (2010) Evaluation of Peak Picking Quality in LC−MS Metabolomics Data. *Anal. Chem.*

Matched filter (gaussian model) – noisy example

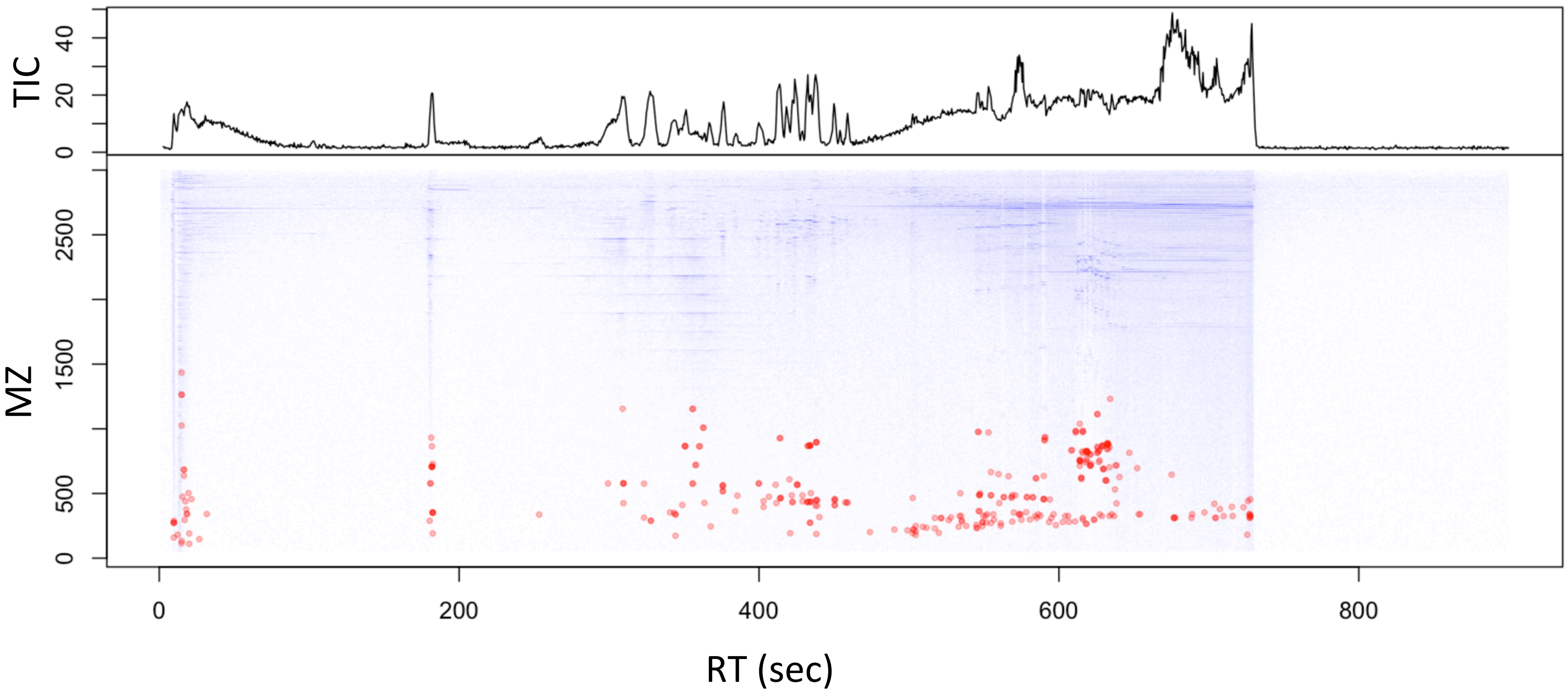


Matched filter (gaussian model) – good example



https://www.ebi.ac.uk/training

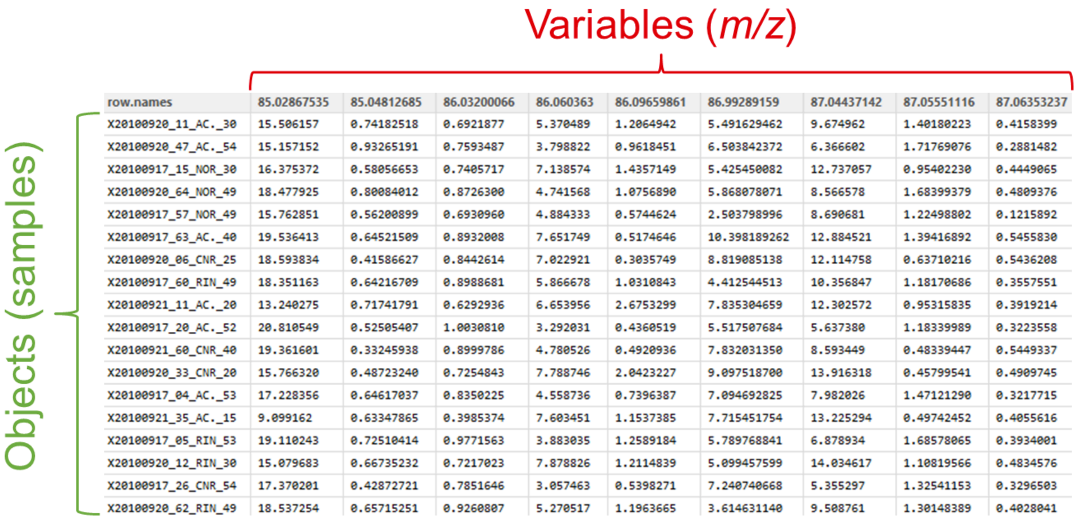
Peak Picking / Peak Detection



https://www.ebi.ac.uk/training

Align different samples

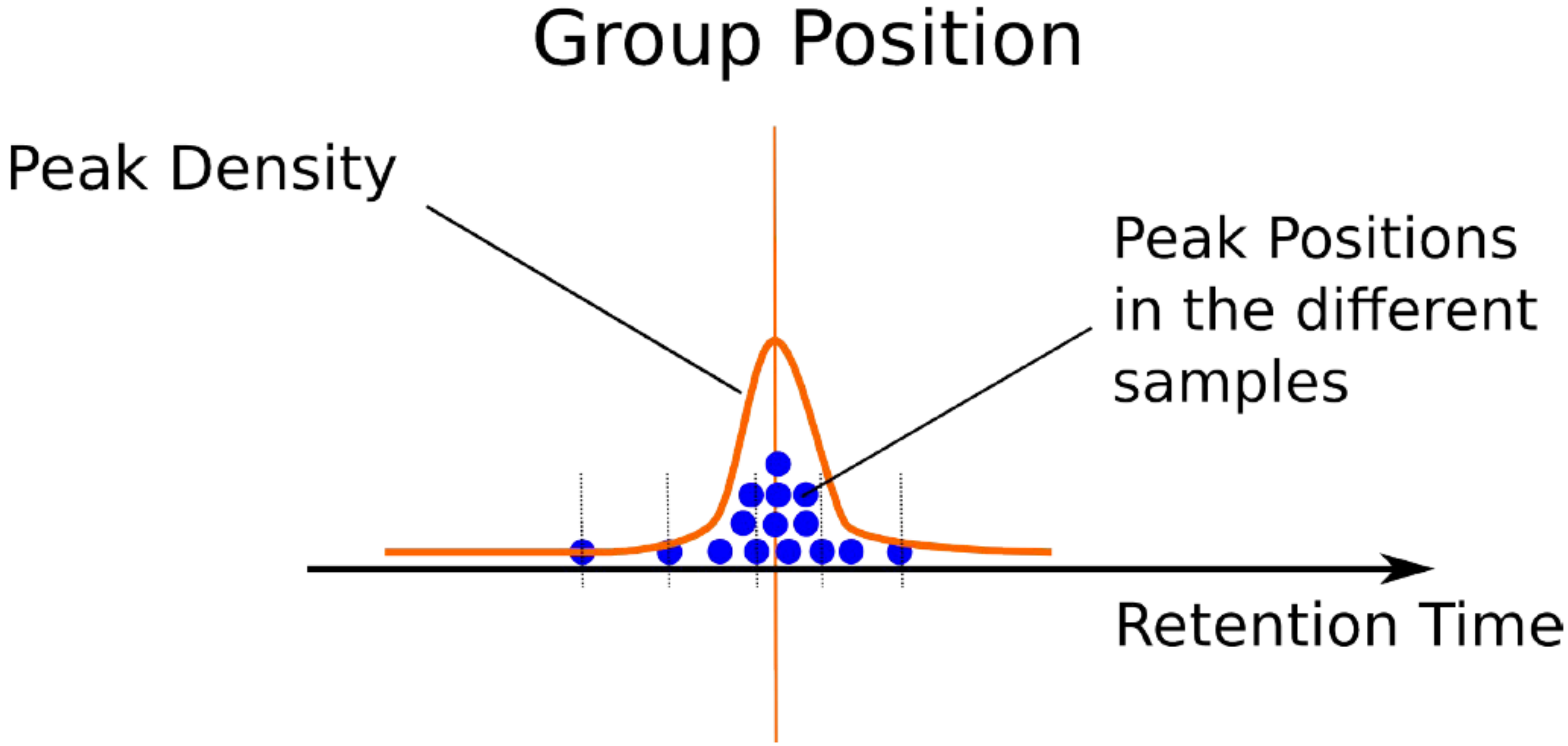
- Construct the data matrix
- Combine the single samples
- With the ultimate goal of correcting for retention time shifts


Variables (*m/z*)
Objects (samples)

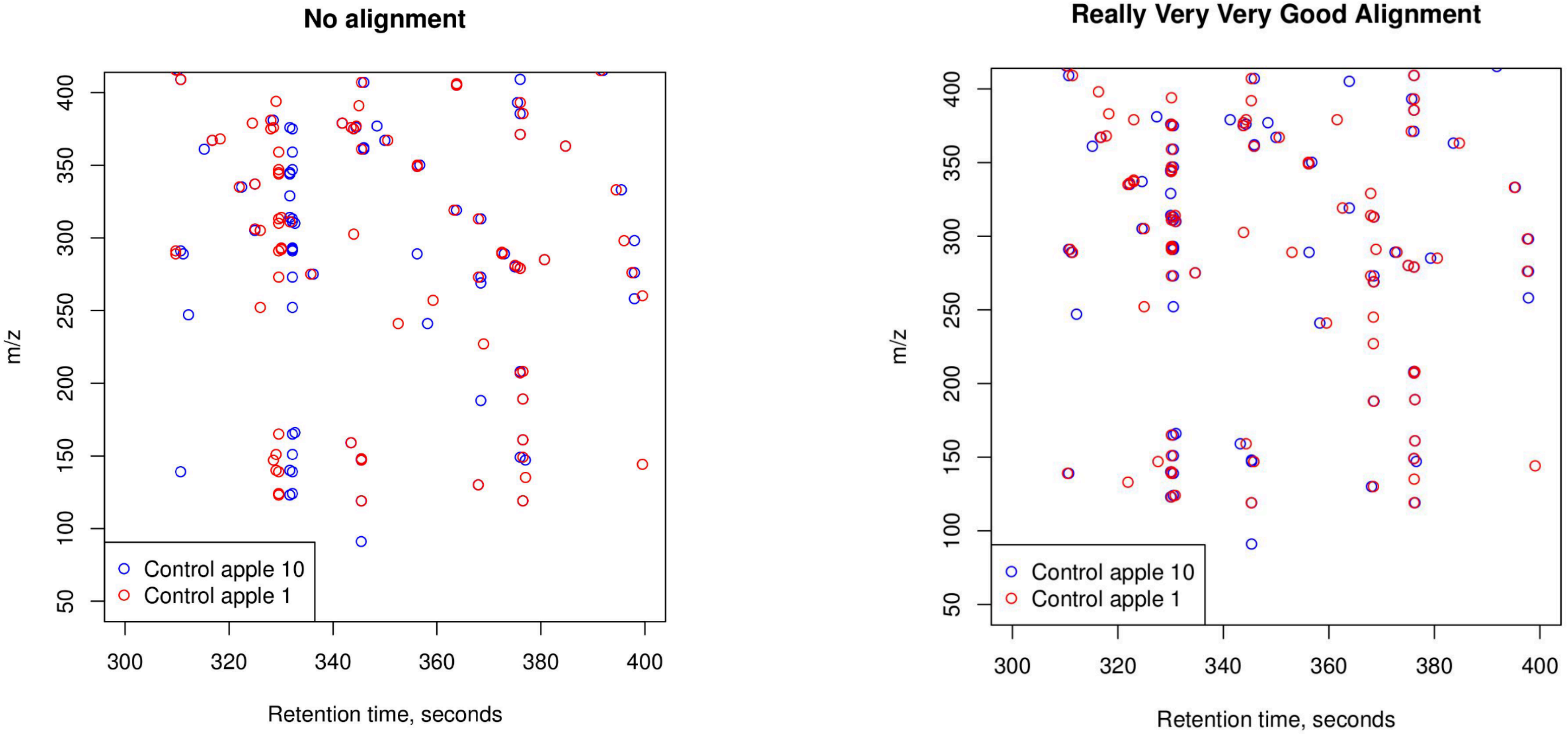
row.names	85.02867535	85.04812685	86.03200066	86.060363	86.09659861	86.99289159	87.04437142	87.05551116	87.06353237
X20100920_11_AC._30	15.506157	0.74182518	0.6921877	5.370489	1.2064942	5.491629462	9.674962	1.40180223	0.4158399
X20100920_47_AC._54	15.157152	0.93265191	0.7593487	3.798822	0.9618451	6.503842372	6.366602	1.71769076	0.2881482
X20100917_15_NOR_30	16.375372	0.58056653	0.7405717	7.138574	1.4357149	5.425450082	12.737057	0.95402230	0.4449065
X20100920_64_NOR_49	18.477925	0.80084012	0.8726300	4.741568	1.0756890	5.868078071	8.566578	1.68399379	0.4809376
X20100917_57_NOR_49	15.762851	0.56200899	0.6930960	4.884333	0.5744624	2.503798996	8.690681	1.22498802	0.1215892
X20100917_63_AC._40	19.536413	0.64521509	0.8932008	7.651749	0.5174646	10.398189262	12.884521	1.39416892	0.5455830
X20100920_06_CNR_25	18.593834	0.41586627	0.8442614	7.022921	0.3035749	8.819085138	12.114758	0.63710216	0.5436208
X20100917_60_RIN_49	18.351163	0.64216709	0.8988681	5.866678	1.0310843	4.412544513	10.356847	1.18170686	0.3557551
X20100921_11_AC._20	13.240275	0.71741791	0.6292936	6.653956	2.6753299	7.835304659	12.302572	0.95315835	0.3919214
X20100917_20_AC._52	20.810549	0.52505407	1.0030810	3.292031	0.4360519	5.517507684	5.637380	1.18339989	0.3223558
X20100921_60_CNR_40	19.361601	0.33245938	0.8999786	4.780526	0.4920936	7.832031350	8.593449	0.48339447	0.5449337
X20100920_33_CNR_20	15.766320	0.48723240	0.7254843	7.788746	2.0423227	9.097518700	13.916318	0.45799541	0.4909745
X20100917_04_AC._53	17.228356	0.64617037	0.8350225	4.558736	0.7396387	7.094692825	7.982026	1.47121290	0.3217715
X20100921_35_AC._15	9.099162	0.63347865	0.3985374	7.603451	1.1537385	7.715451754	13.225294	0.49742452	0.4055616
X20100917_05_RIN_53	19.110243	0.72510414	0.9771563	3.883035	1.2589184	5.789768841	6.878934	1.68578065	0.3934001
X20100920_12_RIN_30	15.079683	0.66735232	0.7217023	7.878826	1.2114839	5.099457599	14.034617	1.10819566	0.4834576
X20100917_26_CNR_54	17.370201	0.42872721	0.7851646	3.057463	0.5398271	7.240740668	5.355297	1.32541153	0.3296503
X20100920_62_RIN_49	18.537254	0.65715251	0.9260807	5.270517	1.1963665	3.614631140	9.508761	1.30148389	0.4028041

We HAVE to compare the right variable across the samples

Density Based Grouping



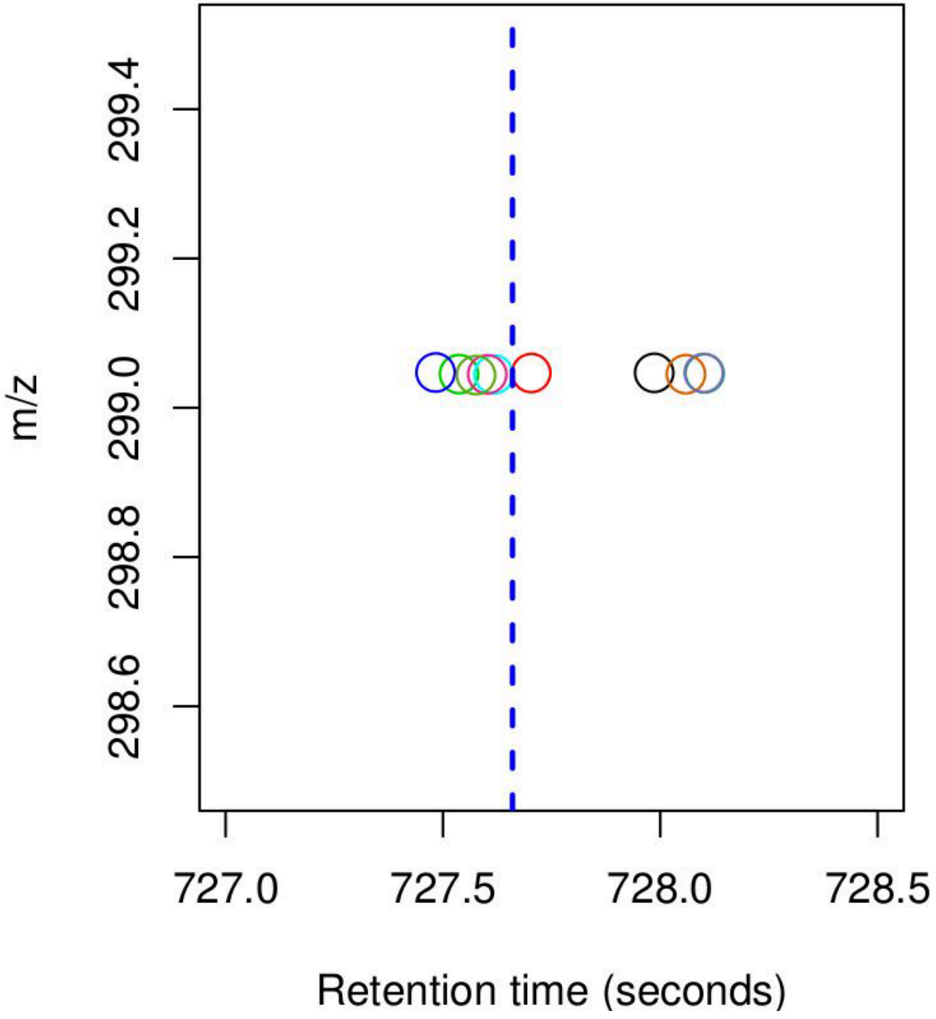
Peak alignment



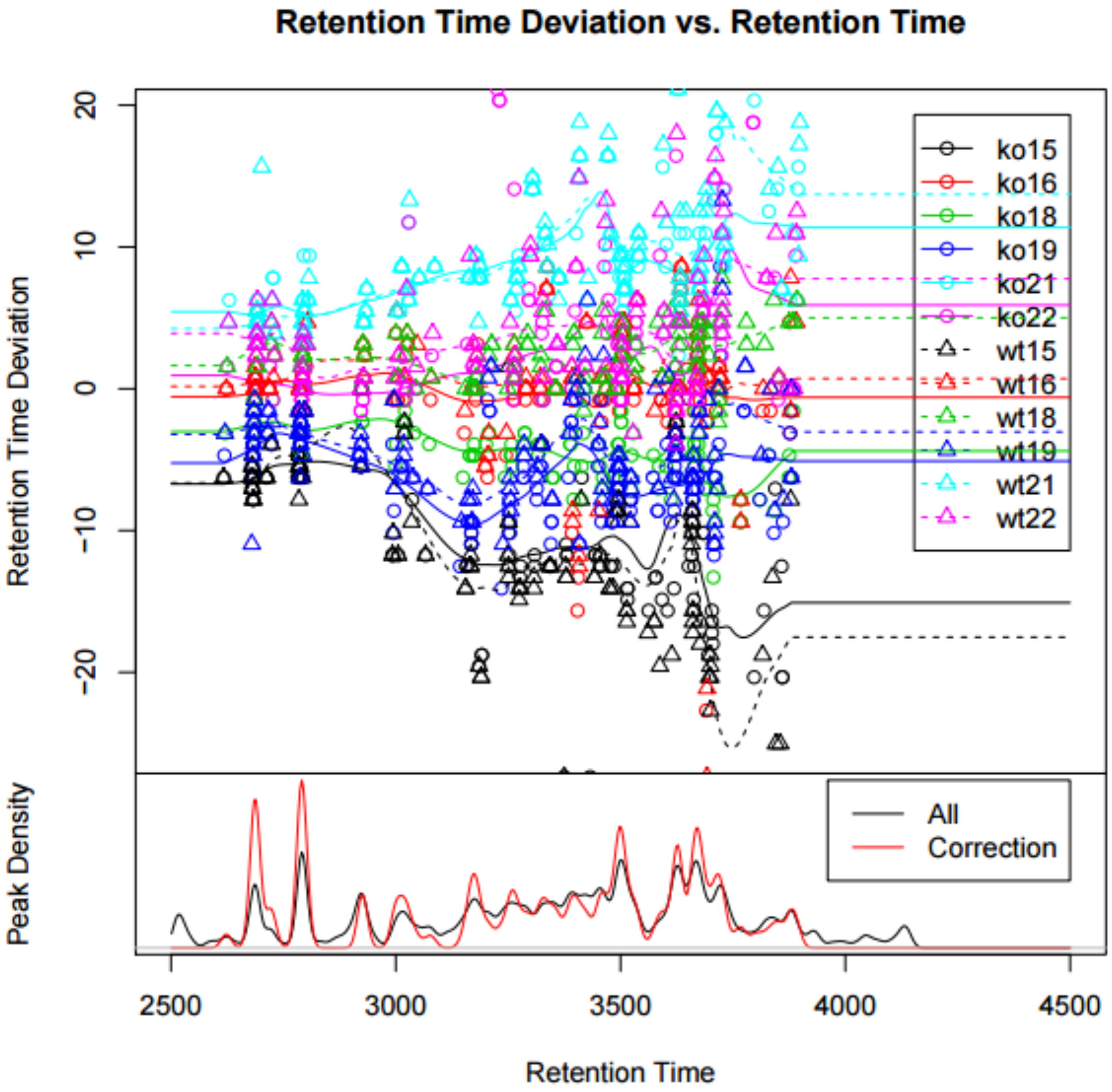
Franceschi et al., J. Chemom. (2012)

Rt correction

- Need “hook” groups.
- Ideally each sample is represented by one feature in a “hook” group.
- Correct the hooks and interpolate elsewhere
- Unfortunately You can have more or fewer features per group.



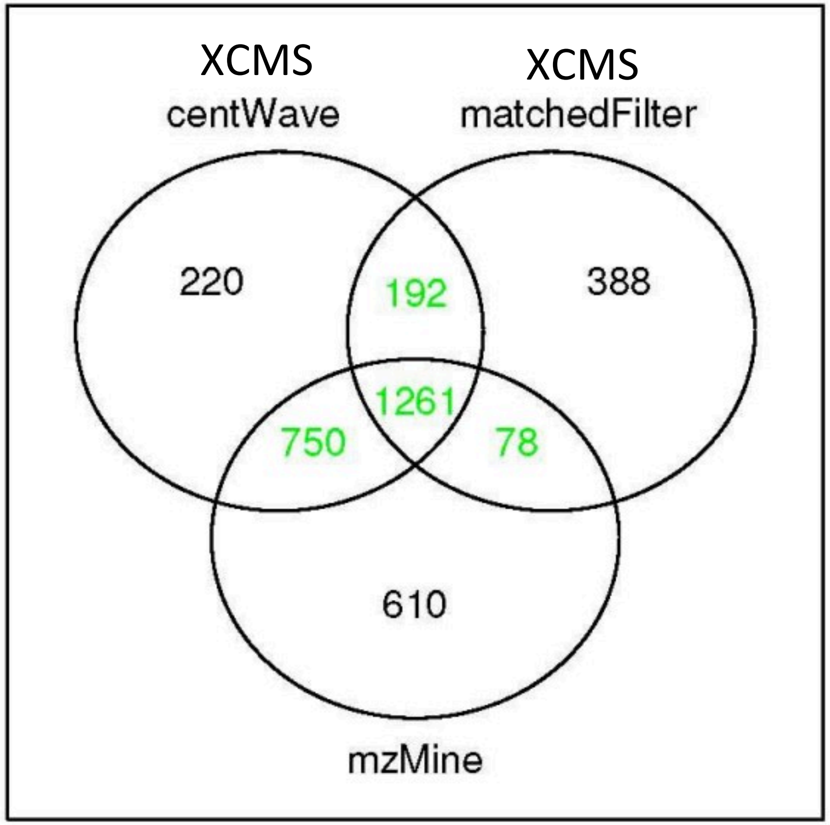
Rt correction



- Can be performed before peak picking (chromatogram alignment)
- Linear or polynomial or whatever correction
- May afford to exclude any ambiguous peaks
- You could run it iteratively till RT deviation is less than your window for peak grouping

Feature Detection Evaluation

- Compare mzMine, XCMS
- Gold standard via
  - Technical replicates
  - Democracy
- Evaluation via
  - Dilution series
  - Mix of complex samples
- F-Measure: sum of
  - Precision (TP/(TP+FP))
  - Recall or sensitivity (TP/P)



Tautenhahn, Böttcher, Neumann. High Sensitive Feature Detection For High Resolution LC-MS. BMC Bioinformatics (2008)

Peak Filtering

Remove peaks from data table based on:

- Number of missing values for a peak
- Max/mean/median intensity (total or within groups of replicates)
- Variability in intensity – coefficient of variance, standard deviation, interquartile range, etc. (total or within groups of replicates)
- …

Missing values
Discrimination

NAs could be:

- real zero/low concentration
- mispicked/misaligned peaks (in general feature is detected correctly)
- incorrectly detected feature

Considerations:

- Total % of NAs for a feature
- presence in replication groups
- amplitude, variability
- …

🟥	🟥	🟥	🟩	🟩	🟩	🟪	🟪	🟪
🟥	🟥	🟥	🟩	🟩	🟩	🟪	F	🟪
T	T	T	T	T	T	🟪	🟪	🟪
T	T	T	T	T	T	🟪	F	🟪
🟥	F?	🟥	F?	🟩	🟩	T?	🟪	T?

Missing values

Treatment

- **Unreliable features:**
  - Remove

- **True zeros:**
  - Look at raw specters data
  - Generate random baseline-level noise

- **False zeros:**
  - Replace by mean/median/etc. for this feature
  - Replace by mean/median/etc. for this feature & replication group
  - PCA-based (BPCA, PPCA, …), KNN-based imputation methods

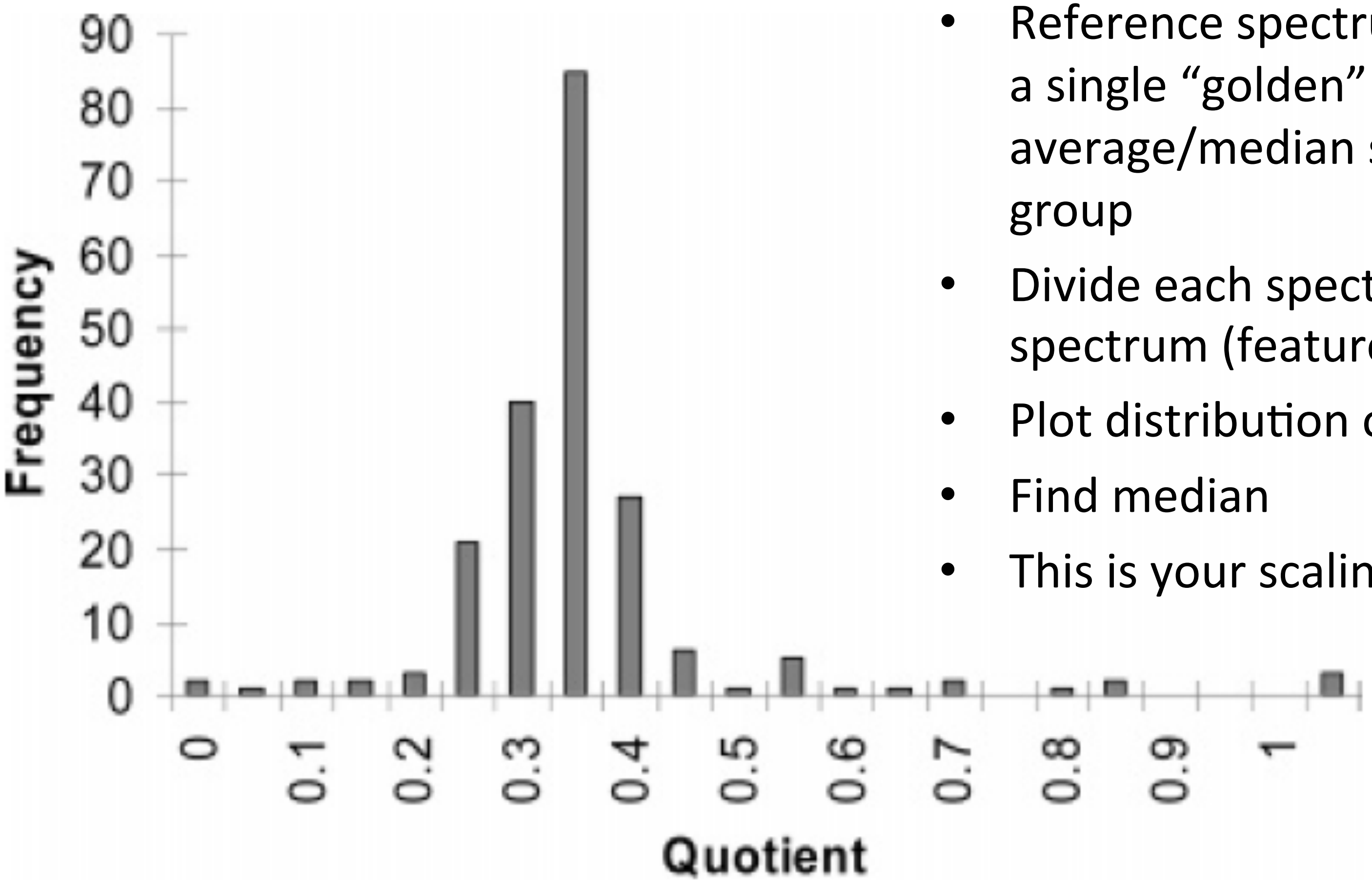
See:
Stacklies, W. et al. (2007). pcaMethods — a bioconductor package providing PCA methods for incomplete data. *Bioinformatics.*

Normalization
Methods

- Not changing intensity distribution – all intensities in one sample have the same normalization factor:
  - by biomass
  - by a single internal standard
  - by mean/median/sum intensity of features in this sample
  - probabilistic quotient normalization (PQN)
  - …
- Changing intensity distribution – each feature in each sample has it's own norm factor, i.e.:
  - by multiple internal standards (i.e. NOMIS)
  - quantile normalization – "stretching" distributions of all samples to make them similar
  - …
- General assumption for normalization is that most of the compounds are not affected. Is that true? For different treatment? For different species? For different tissues? Does it matter if we have no choice? :/

Normalization

Probabilistic Quotient Normalization (PQN)



- Reference spectrum could be a single "golden" spectra or an average/median spectrum of control group
- Divide each spectrum by a reference spectrum (feature by feature)
- Plot distribution of ratios (quotients)
- Find median
- This is your scaling factor

Dieterle, F. et al. (2006). **Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1H NMR Metabonomics**. *Anal Chem.*

Centering and Scaling

Applied to features across the samples

Nature of MS data:

- Features are extremely different in amplitude
- Heteroscedasticity – biological (induced and uninduced) and technical variance are higher for features with high intensity

Scaling:

- Equalizes contributions of features to separation in multivariate space
- Makes features comparable (i.e. for looking at time profile)

Types of scaling:

- Range scaling – by [max – min] – sensitive to outliers; undesirable
- Auto-scaling – by standard deviation (SD) – data loose dimensionality
- Pareto-scaling – by root of SD – features with higher intensity decrease more

Centering is subtracting mean/median from all the values:

- Necessary for some methods like PCA and makes no sense for others like fold change

Transformation

Certain function applies to all the values in a data table.

- Log-transformation
- General logarithmic transformation (glog) – approximately log for high values and linear close to zero
- Cube-root transformation

Why?

- Transformation has a scaling-like effect making features more comparable.
- Log/glog-transformation helps to reveal multiplicative relations between features.

Annotation

Retention indexing / Retention projection

RT is extremely variable. Idea of ***retention indexing:***

save an exemplary LC as a “scale” for the future and then align all the times by this database.

Limitations:

- limited number of tested compounds – extrapolate several compounds to a class?
- interactions between compounds => RT could depend on a sample composition – databases of complex mixtures?
- only certain LC system/conditions – retention projection? (see the next slide)

All additional experiments => time, money

Annotation

Databases

Annotation could be manual or with more or less automatic tools coupled with databases:

- Commercial – really?
- Open source
- In-house:
  - works for you, specified for your needs, possible to include retention indexing
  - but costs additional work, money, time

Fragment MS/MS (or GC/MS) databases:

- Experimental
  - specific: instrument, ionization parameters, etc.
- In-silico (e.g. LipidBlast)
  - theoretical, but wide coverage

Problems

Experimental

1. Batch effect (48 per run)
2. Platform-based effect
3. Poor correspondence between experiments
4. Concentration estimation

Data Analysis

1. Annotation (low percent of annotated compounds ~20-40%)
2. No golden software standard
3. Technical effects
4. Poor alignment of samples

Acknowledgments

Skoltech

Skolkovo Institute of Science and Technology

- Dr. Philipp Khaitovich
- Dr. Waltraud Mair
- Dr. Ekaterina Khrameeva
- Anna Tkachev
- Dmitry Zubkov
- Vita Stepanova

Questions?