Themis: Pre-processing of complex mixture data

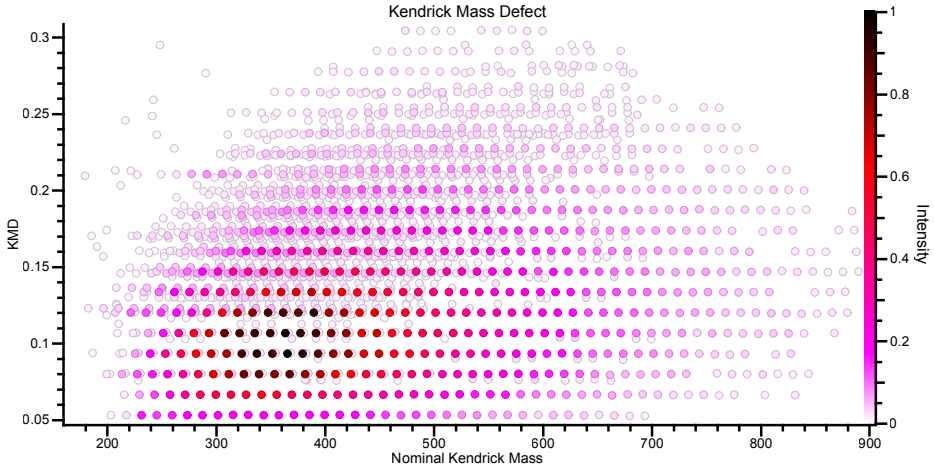
Remy Gavard,[1] David Rossell,[2] Simon E. F. Spencer,[1] and Mark P. Barrow[1]

[1] University of Warwick, Coventry, United Kingdom

[2] Universitat Pompeu Fabra, Barcelona, Spain

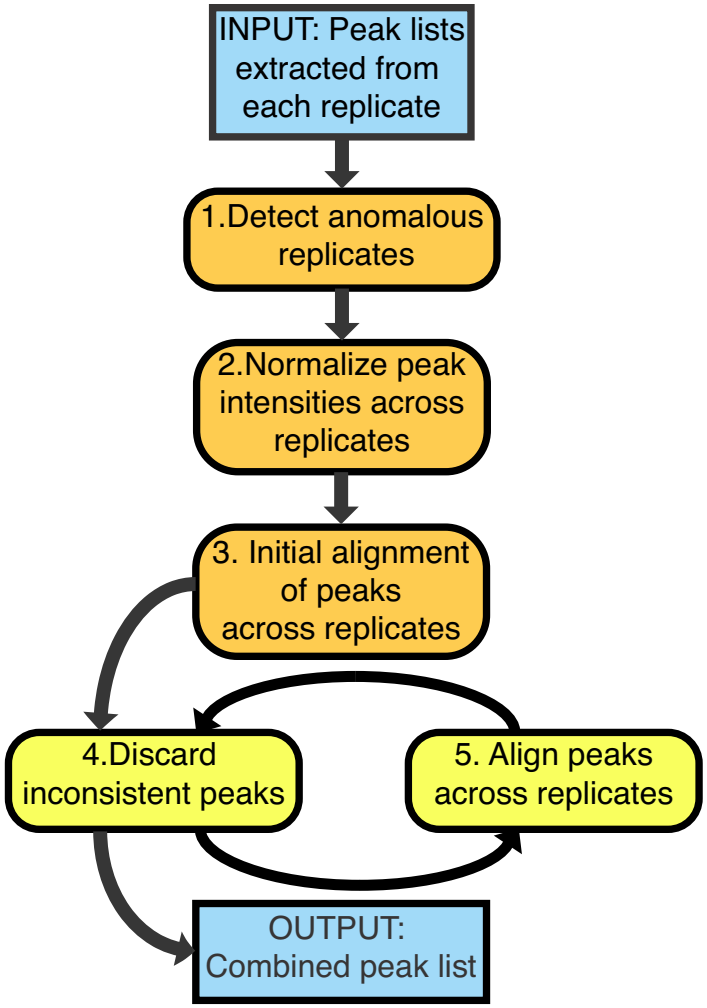
Data analysis challenges

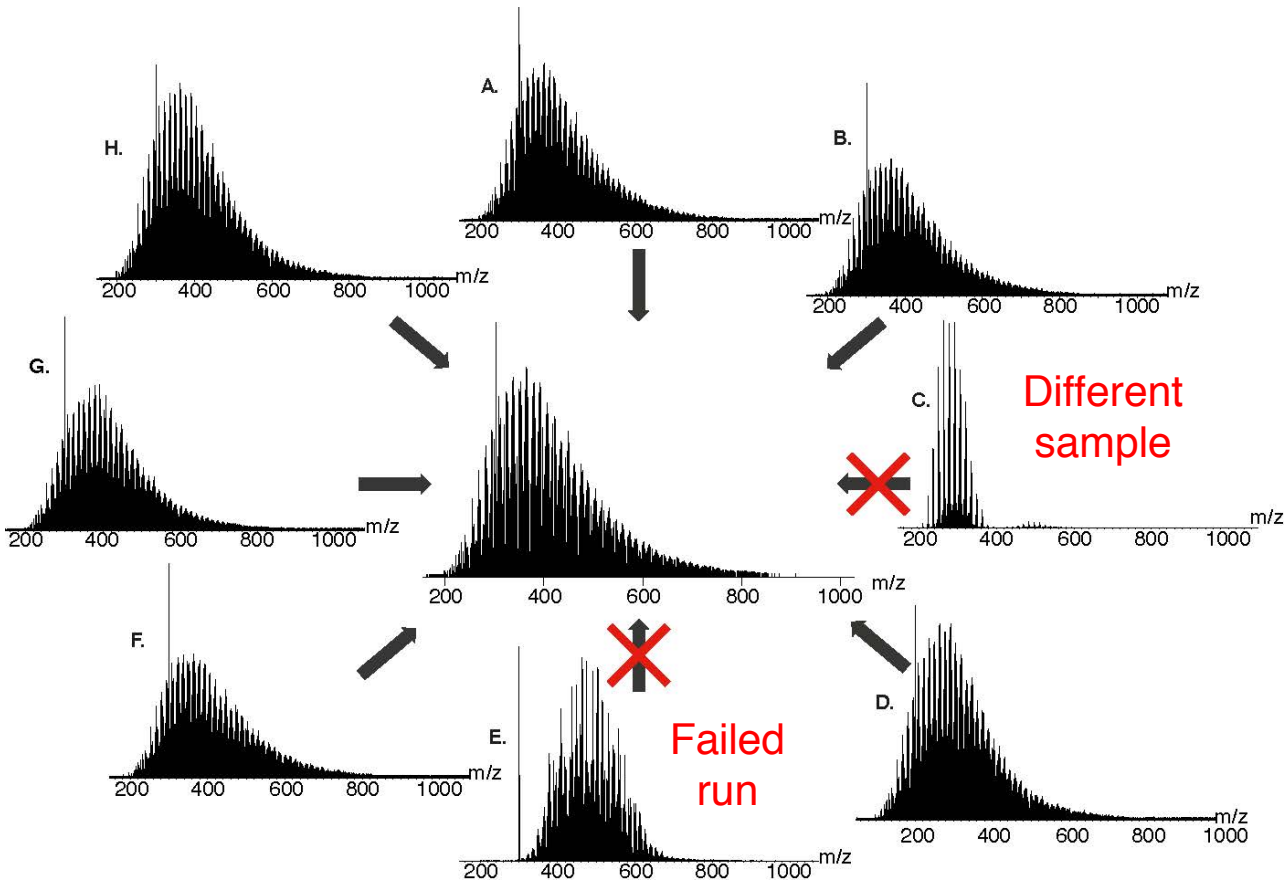
- **Environmental** and **petroleum** samples are highly **complex mixtures**
- **Instrument advances** enable acquisition of more complex data
- **Data analysis** then becomes the **bottleneck**
- **Peak picking risks**:
  - high threshold → miss out genuine peaks
  - low threshold → include noise → **misassignments**
- Need for **data reliability**: decision making
- Need for improved **workflows** using **complex data sets**

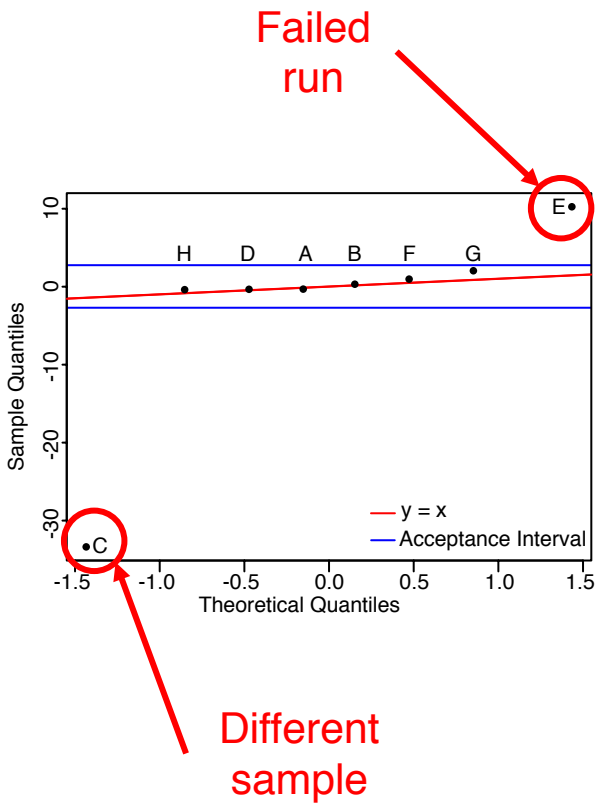


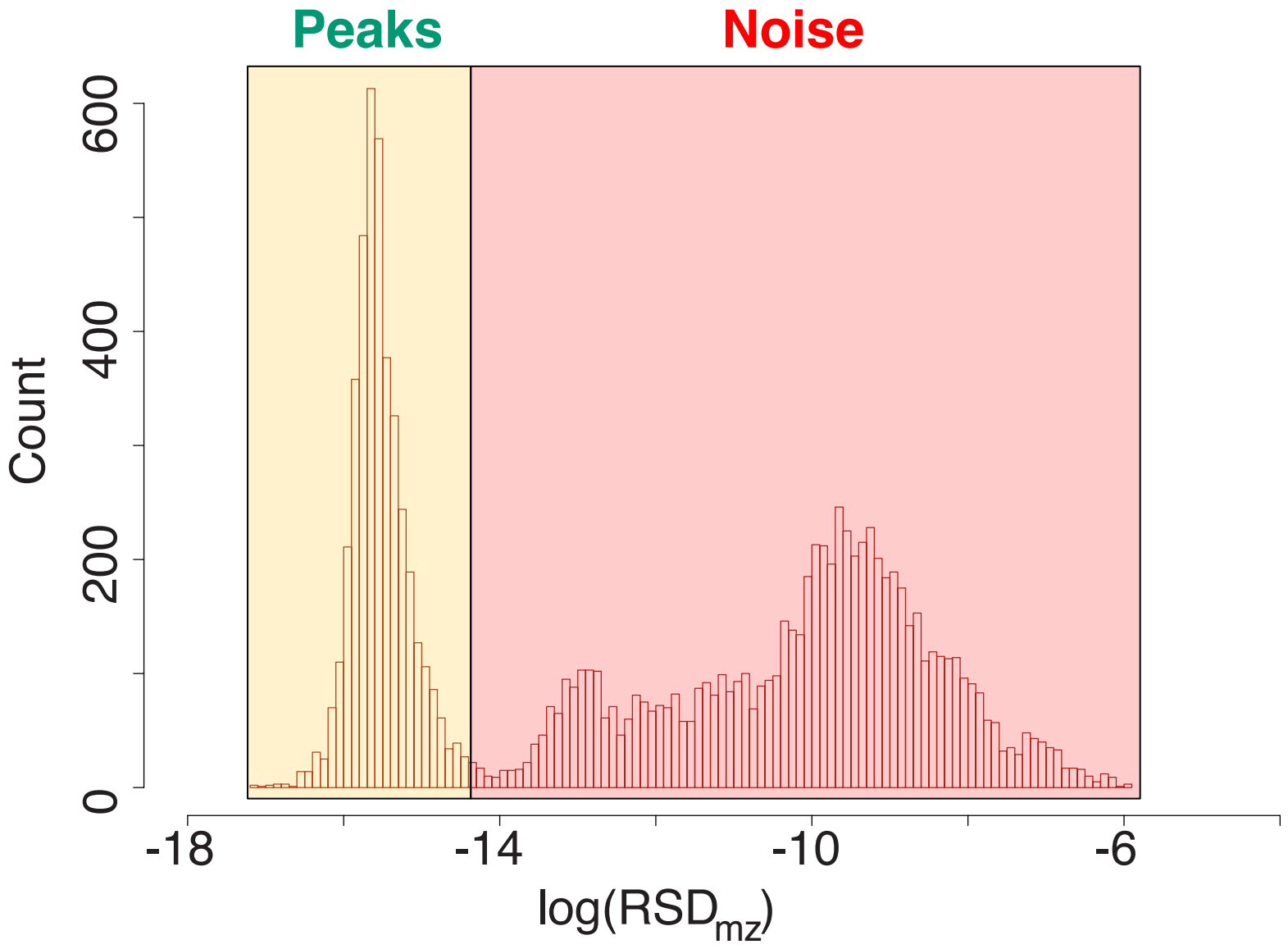
New algorithm: Themis

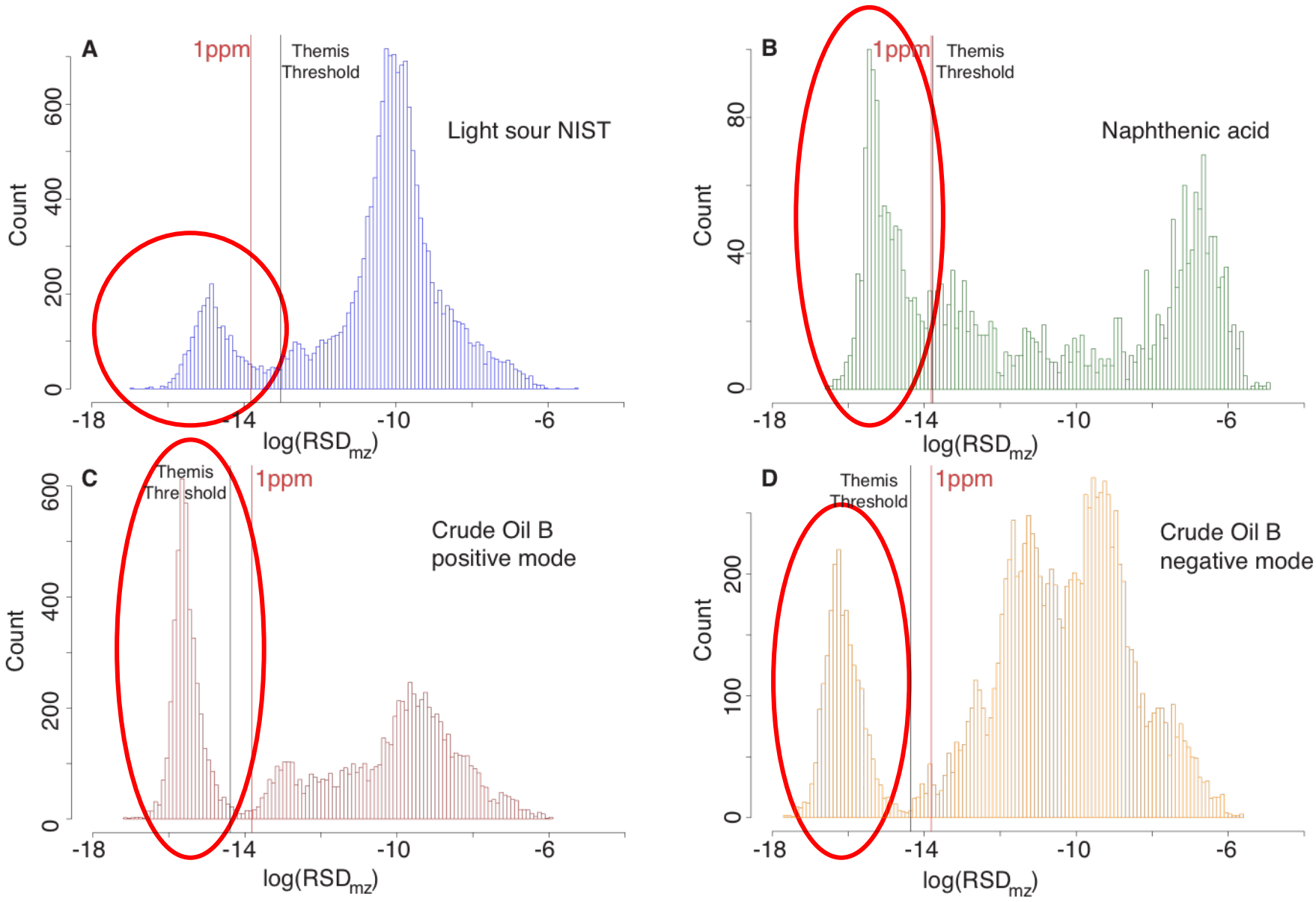
- Users can **generate peak lists** in preferred software (with low S/N threshold) and **import**
- **Themis** uses **replicate data sets** without prior data analysis or knowledge of composition
- Batch **pre-processes** replicate spectra (repeat measurements)
- **Detects** and can remove **outlier datasets**
- **Improved** differentiation between **noise** and **genuine peaks**
- Produces **averaged peak list** as output
- Soon available to academic community online

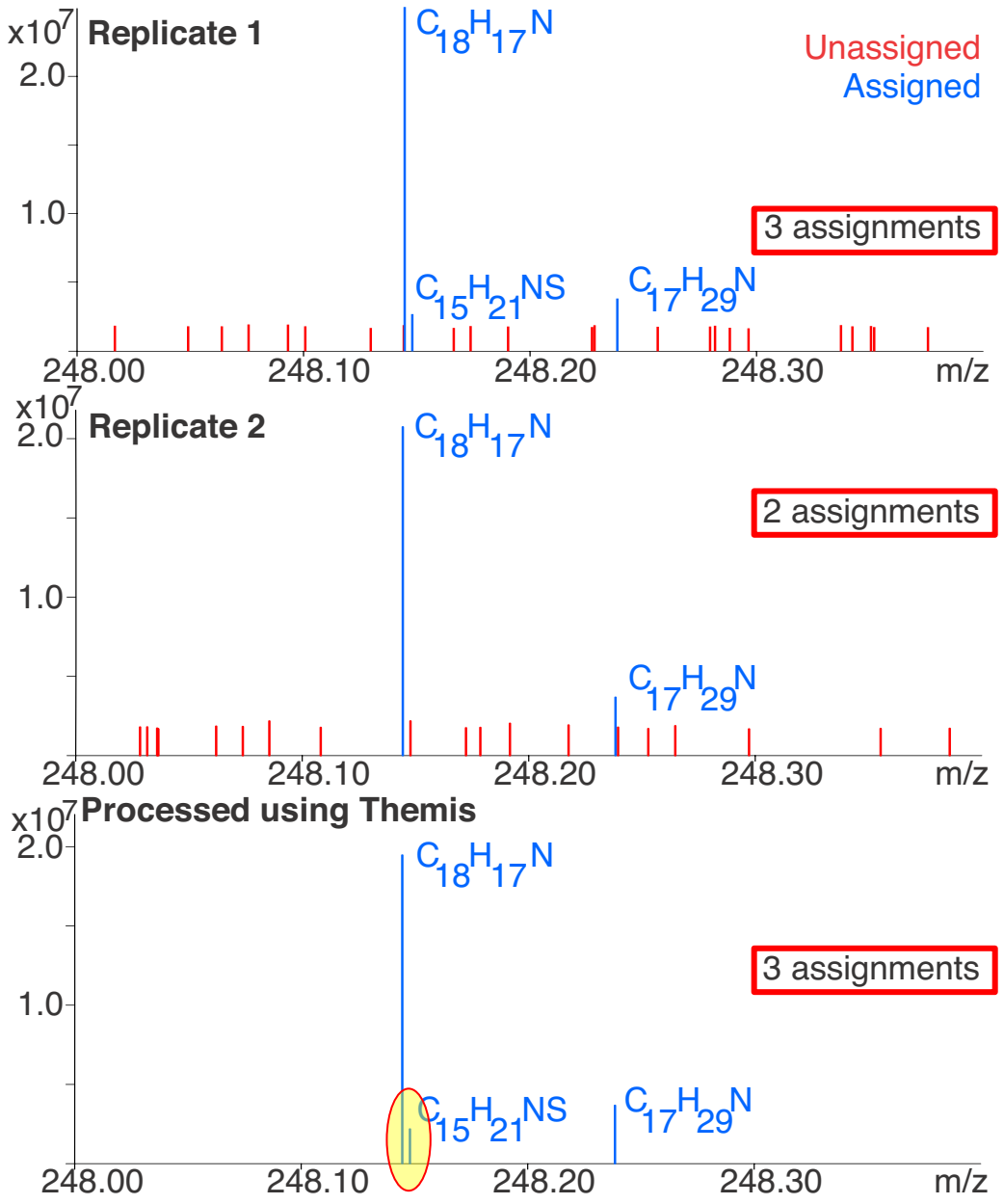



A.
B.
C.
D.
E.
F.
G.
H.
m/z
200
400
600
800
1000
Different sample
Failed run

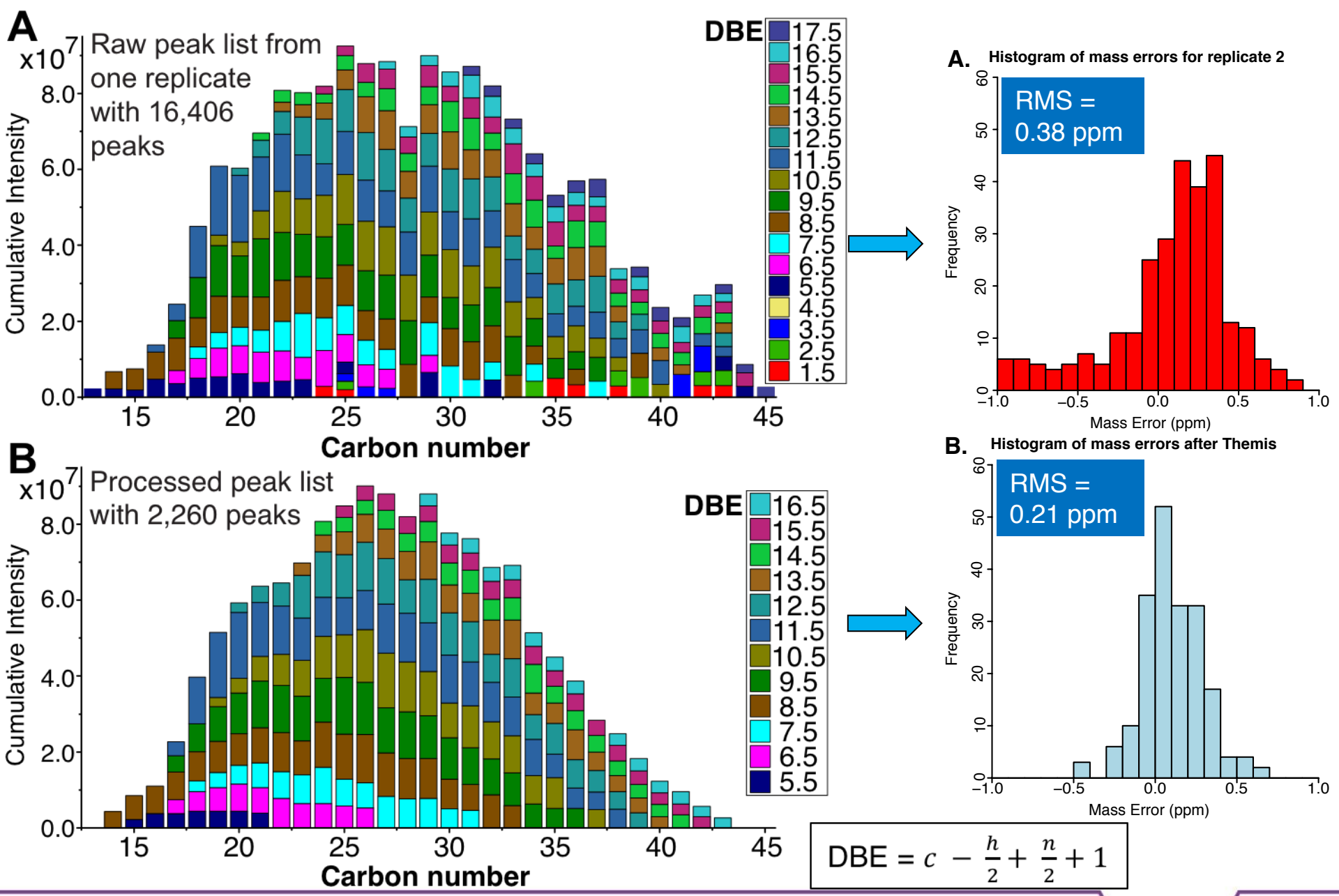

Failed run
Different sample
Sample Quantiles
Theoretical Quantiles
H
D
A
B
F
G
E
C
y = x
Acceptance Interval
10
0
-10
-20
-30
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5


Peaks
Noise
Count
0
200
400
600
-18
-14
-10
-6
log(RSD_mz)


A
1ppm
Themis Threshold
Light sour NIST
Count
0
200
400
600
-18
-14
-10
-6
log(RSD$_{mz}$)
B
1ppm
Themis Threshold
Naphthenic acid
Count
0
40
80
-18
-14
-10
-6
log(RSD$_{mz}$)
C
Themis Thre shold
1ppm
Crude Oil B positive mode
Count
0
200
400
600
-18
-14
-10
-6
log(RSD$_{mz}$)
D
Themis Threshold
1ppm
Crude Oil B negative mode
Count
0
100
200
-18
-14
-10
-6
log(RSD$_{mz}$)



Thresholding: example of low S/N peaks



$$\mathrm{DBE} = c - \frac{h}{2} + \frac{n}{2} + 1$$

NS class: example of low S/N peaks

Themis summary

- **Themis**: produces **averaged single peak list** from replicates, **differentiates** peaks from noise, improves **workflow** and **data reliability**
- **Web interface** created to run algorithm easily
- **Access** to the academic community will be provided shortly
- **Contact us** for more information or to provide feedback

Poster on Thursday: ThP327

Questions

- How are replicates currently handled?
- What are the remaining bottlenecks for analysis complex mixtures?
- Thoughts on use of Themis for Environmental community?
- What features would be useful to add?

Poster on Thursday: ThP327

Acknowledgments

Dr. David Stranz

ROYAL SOCIETY OF CHEMISTRY | ANALYTICAL DIVISION

BMSS BRITISH MASS SPECTROMETRY SOCIETY

MAS CDT
Molecular Analytical Science Centre for Doctoral Training

EPSRC
Pioneering research and skills

Contact: R.Gavard@warwick.ac.uk ; M.P.Barrow@warwick.ac.uk