Rapid MS1 Formulae - Isotope Pattern Matching Using a Novel MS1 Search Engine for Metabolomics

Scott Walmsley[1,2], Sam Bokatzian[1], Hyungwon Choi[3], Richard Reisdorph[1], Nichole Reisdorph[1]
[1]Department of Pharmaceutical Sciences, University of Colorado Denver Anschutz Medical Campus; [2]Computational Biosciences Program, University of Colorado Denver Anschutz Medical Campus; [3]Saw See Hawk School of Public Health, National University of Singapore

Skaggs School of Pharmacy and Pharmaceutical Sciences
UNIVERSITY OF COLORADO
ANSCHUTZ MEDICAL CAMPUS

Overview

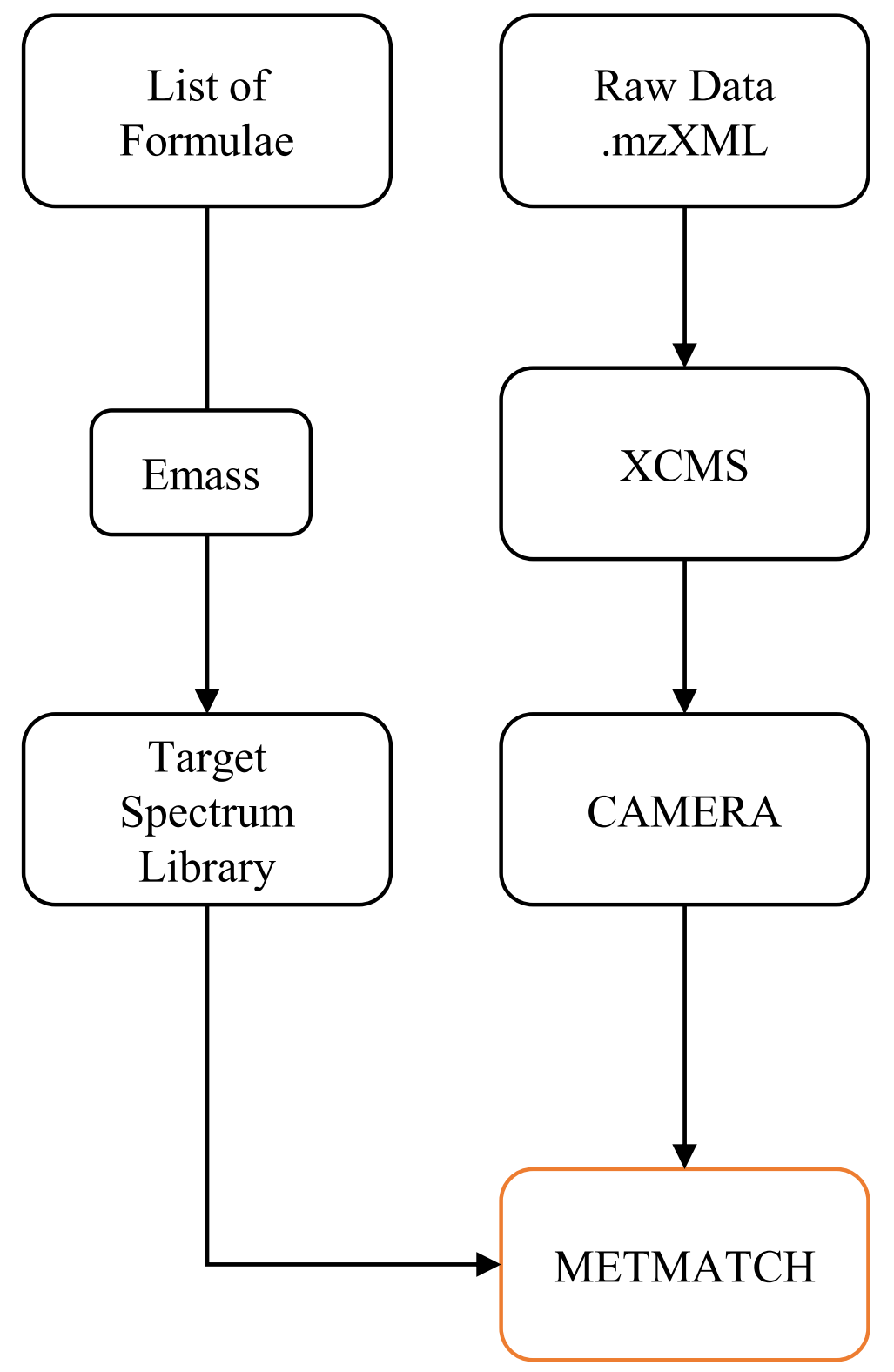
- **We present the R package, *MetMatch*, a novel global MS1 metabolomics search engine.**
- **The search engine *MetMatch* enables global modelling and detection of good isotope clusters likely to produce a correct formulae assignment.**
- **MS1—formula to spectral matching are completed using a target spectral library (TSL).**
- **The library is built using Emass for prediction of expected molecular formulae in a sample.**
- **Experimental isotope masses and relative abundances are measured against the TSL.**
- **A pseudo-spectrum library (PSL, eg. decoys) is used to determine the threshold score.**
- **The algorithm was evaluated using Aqueous and Lipid extracts of HEK293t cells (ESI-QTOF MS).**
- **Additional evaluation was performed against a published MS1-MS2 dataset (Orbitrap-MS/MSMS).**
- **Results indicated the algorithm can reliably and reproducibly assign MS1 isotope clusters a formula**
- **Quality of MS1-formula matches are dependent on factors including ion abundance.**

Introduction

The goal of this work was to develop a MS1 metabolite to formula matching algorithm that simultaneously matches all compounds with isotopic peaks in a sample to a potential formula. Current methods attempt to match these isotopes to all possible formulae for a given mass, and can be computationally slow. The advantage is that this MetMatch allows global modelling of the matches which produces a score that is directly correlated to the quality of the ion and it's formulae assignment. Our method departs from the traditional metabolomics workflow (differential analysis, feature grouping between samples, and selecting small groups of molecules for annotation) and adopts a more proteomics style interrogation of the data (attempting to annotate all molecules in each sample, followed by direct comparison of like annotated molecules across samples).

- MetMatch is written in R, and integrates with the XCMS / CAMERA feature finding and isotope annotating software.
- We adopt the concept of an metabolome formulae database containing isotope masses and abundances, here called a target spectrum library (TSL).
- The library is built using a list of expected unique molecular formulae in a given sample.
- MetMatch takes as input the TSL and the annotated peaklists generated from CAMERA.

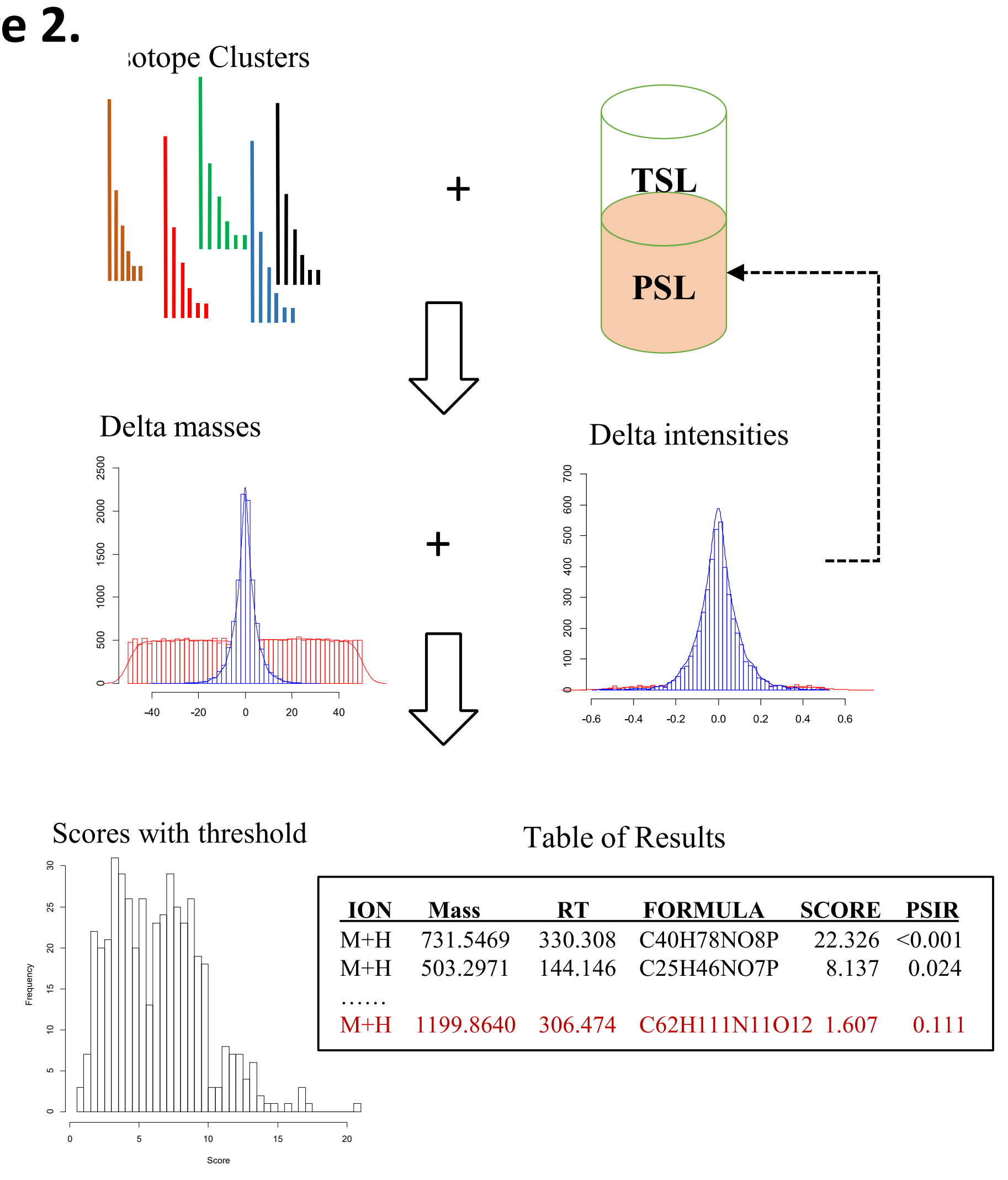
Figure 1.



Algorithm

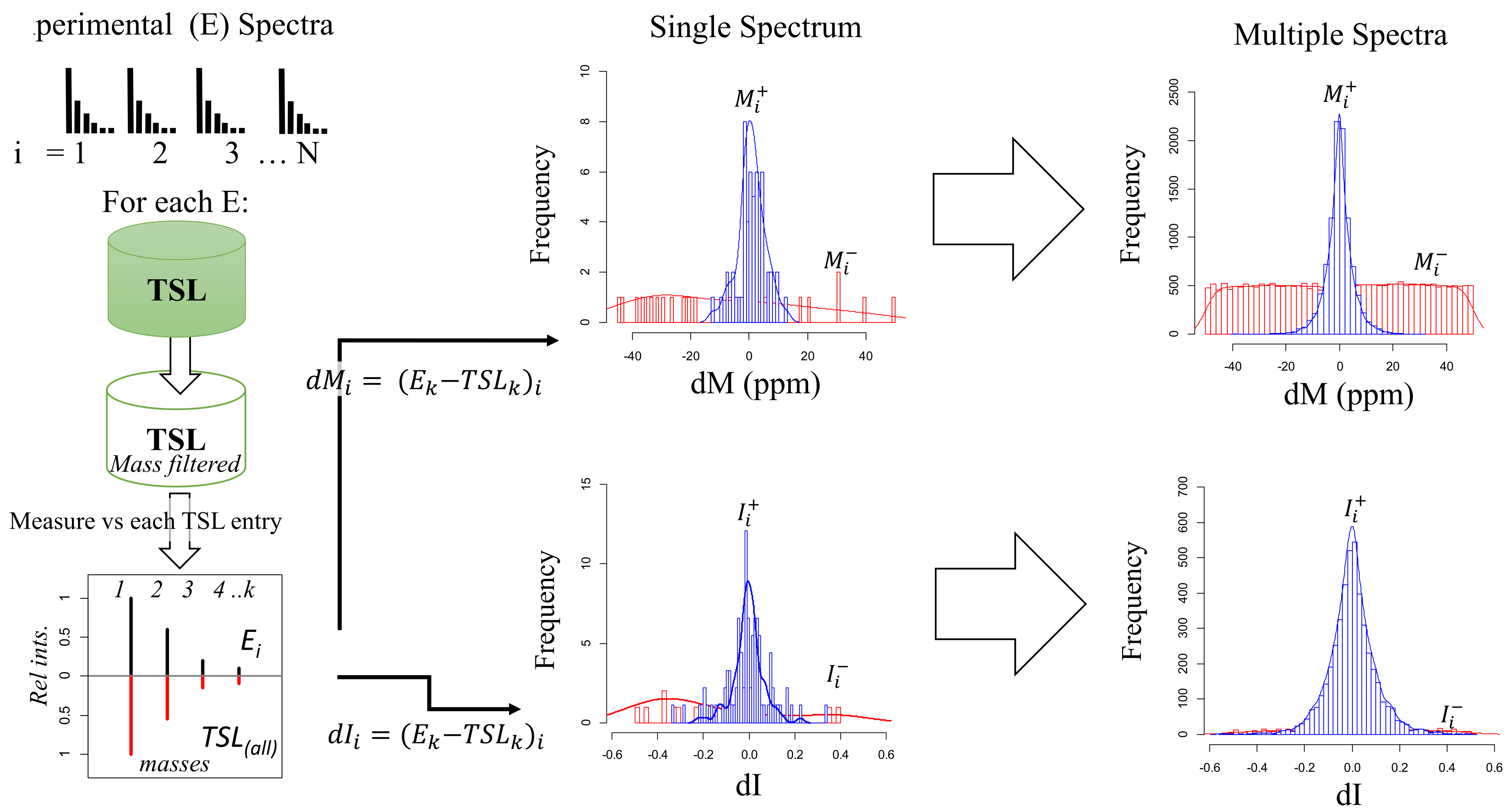
- MetMatch measures the error between experimental isotope clusters and potential candidates contained within the TSL (Figure 2). The TSL is constructed from the Human Metabolome and LipidMaps databases using the ~14000 unique formulae from these two databases.
- Two scoring modules, one for isotope masses and one for isotope intensities, are used to compute and store all measured isotopes matches to the TSL within a mass (50ppm) and intensity (±0.5 rel abundance) tolerance window.
- Distributions for all matches in the sample are computed using expectation maximization from these delta masses and delta intensites, and a score computed for masses and intensities.
- A final score is the sum of the mass and intensity score. Each molecule has a score linking it to the mostly likely correct formula match from the TSL. The higher the score, the better the match.
- A PseudoSpectrum Library (similar in concept to decoy databases in proteomics) is built using information learned from the delta mass and intensity distributions and aids selecting a threshold score.
- Figure 3 further visualizes how the models are produced. Isotope delta masses (dM) and intensities (dI) built between each experimental MS1 isotope cluster and closest TSL entries are binned. The Good matches have dM and dI values approaching 0 for each isotope. The models are fit to these distributions using EM . Scores are the sum of the log(M^+/M^-) and log(I^+/I^-) values produced for each experimental spectra's best formula match.

Figure 2.



ION	Mass	RT	FORMULA	SCORE	PSIR
M+H	731.5469	330.308	C40H78NO8P	22.326	<0.001
M+H	503.2971	144.146	C25H46NO7P	8.137	0.024
……					
M+H	1199.8640	306.474	C62H111N11O12	1.607	0.111

Figure 3.



Testing and Results

- Data from aqueous and lipid extracts of HEK293t cells grown in culture were analyzed using MetMatch. N=10 (aqueous MeOH, MtBE) or N=14 (Lipid) technical replicates were injected.
- Data were acquired using online C18-HPLC Agilent QTOF 6520 mass spectrometer with a 15minute gradient.
- After analysis with XCMS/CAMERA/MetMatch, formula assigned to isotope clusters were aligned between samples.

- Figure 4 shows results produced for one individual replicate. Included are the models built using the delta mass and delta intensity distributions.
- Included are a heat map of those two scores (lower left).
- The distribution of Pseudo Spectra (eg. the decoys) are shown as black dots in the map which aid determination of the thresholds.
- Figure 5 shows the aligned compounds and their intensities for the Lipid fraction.
- Figure 6A shows that the distribution of scores amongst all replicates are different by extraction type, but highly similar within fractions.
- Logistic regression model built using the data has indicated that first isotope peak abundance is important for a higher score.
- Figure 6B exemplifies this point for EIC traces of the detected 3 isotopes for a high scoring, score near the threshold, and poor scoring isotope clusters.

Figure 4.

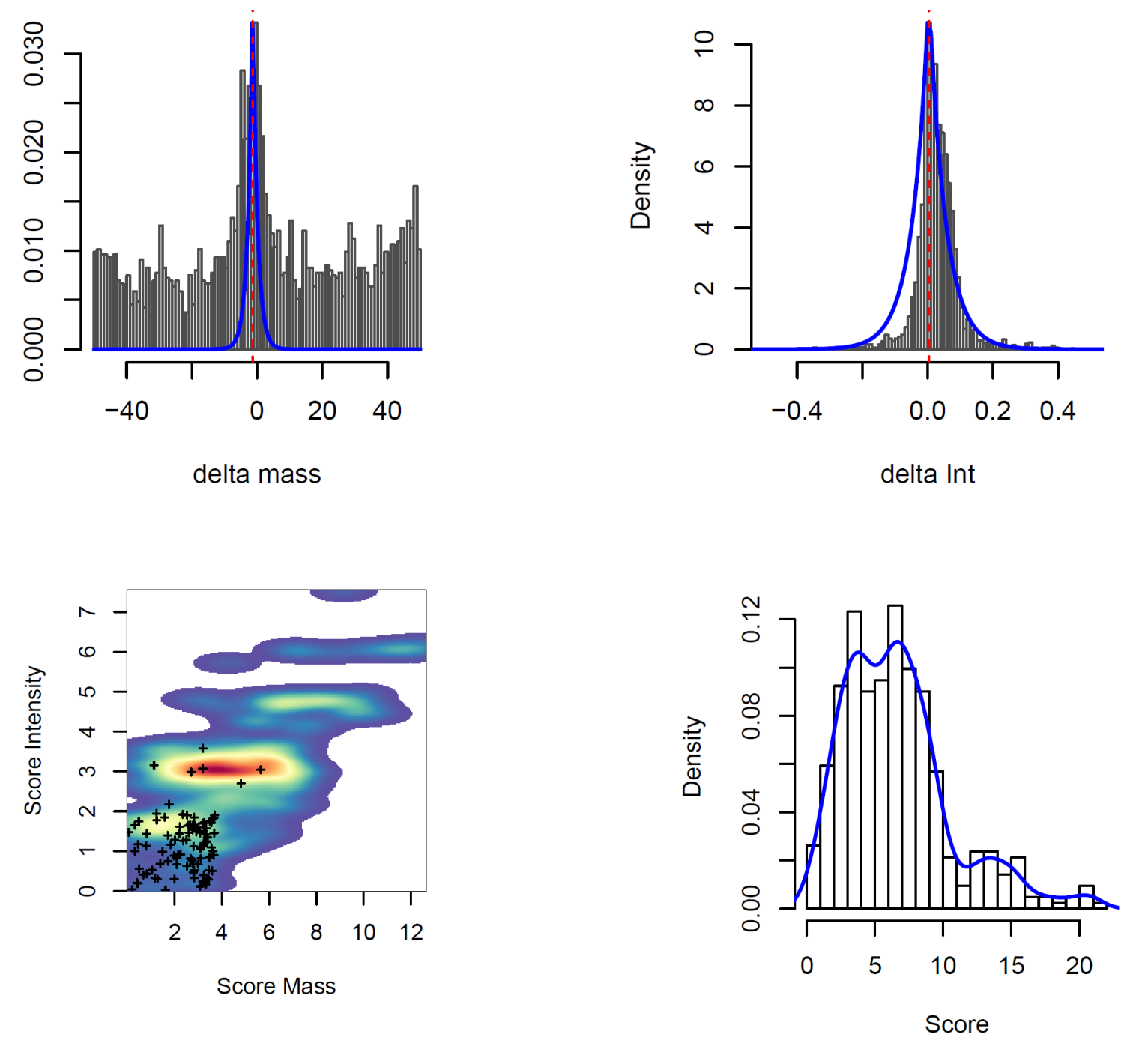


Figure 5.

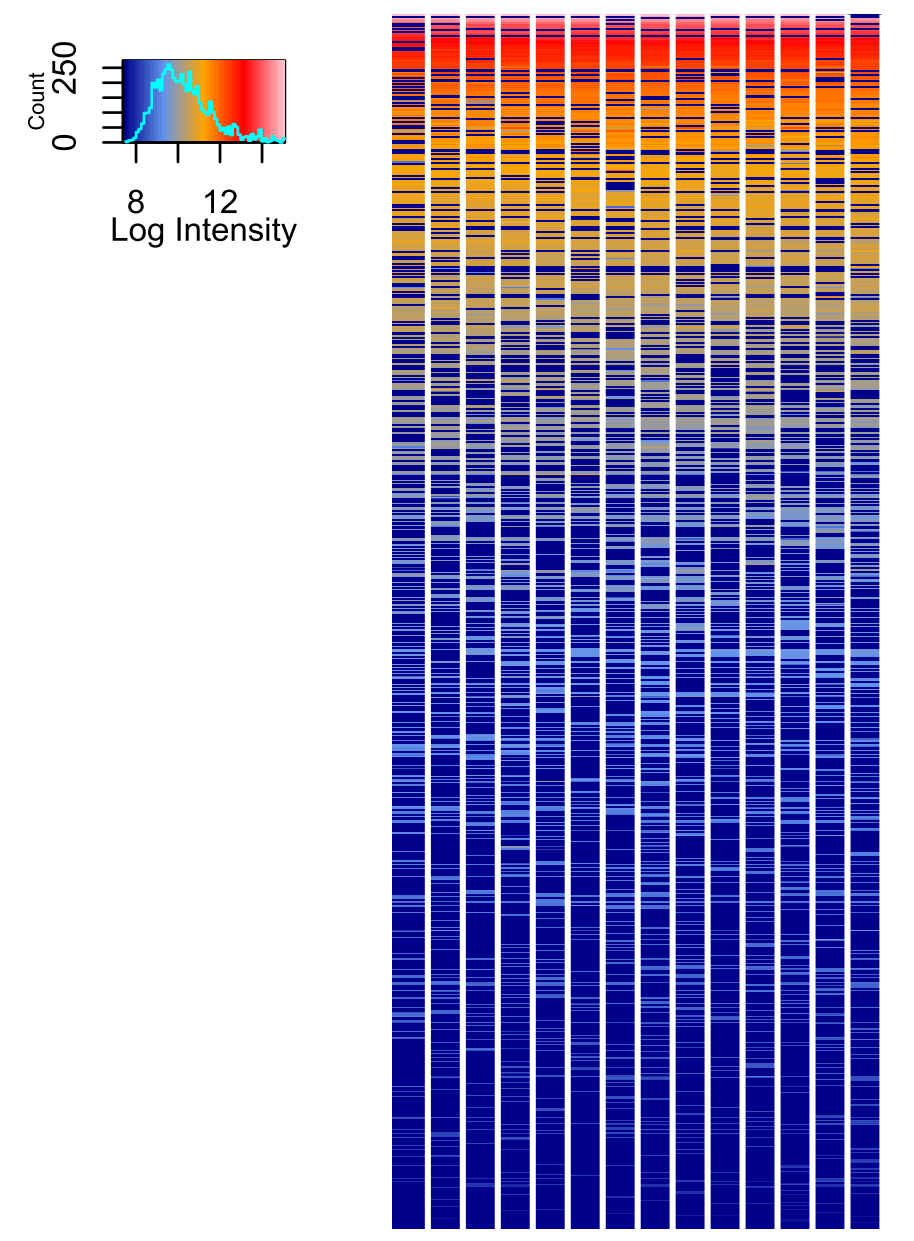
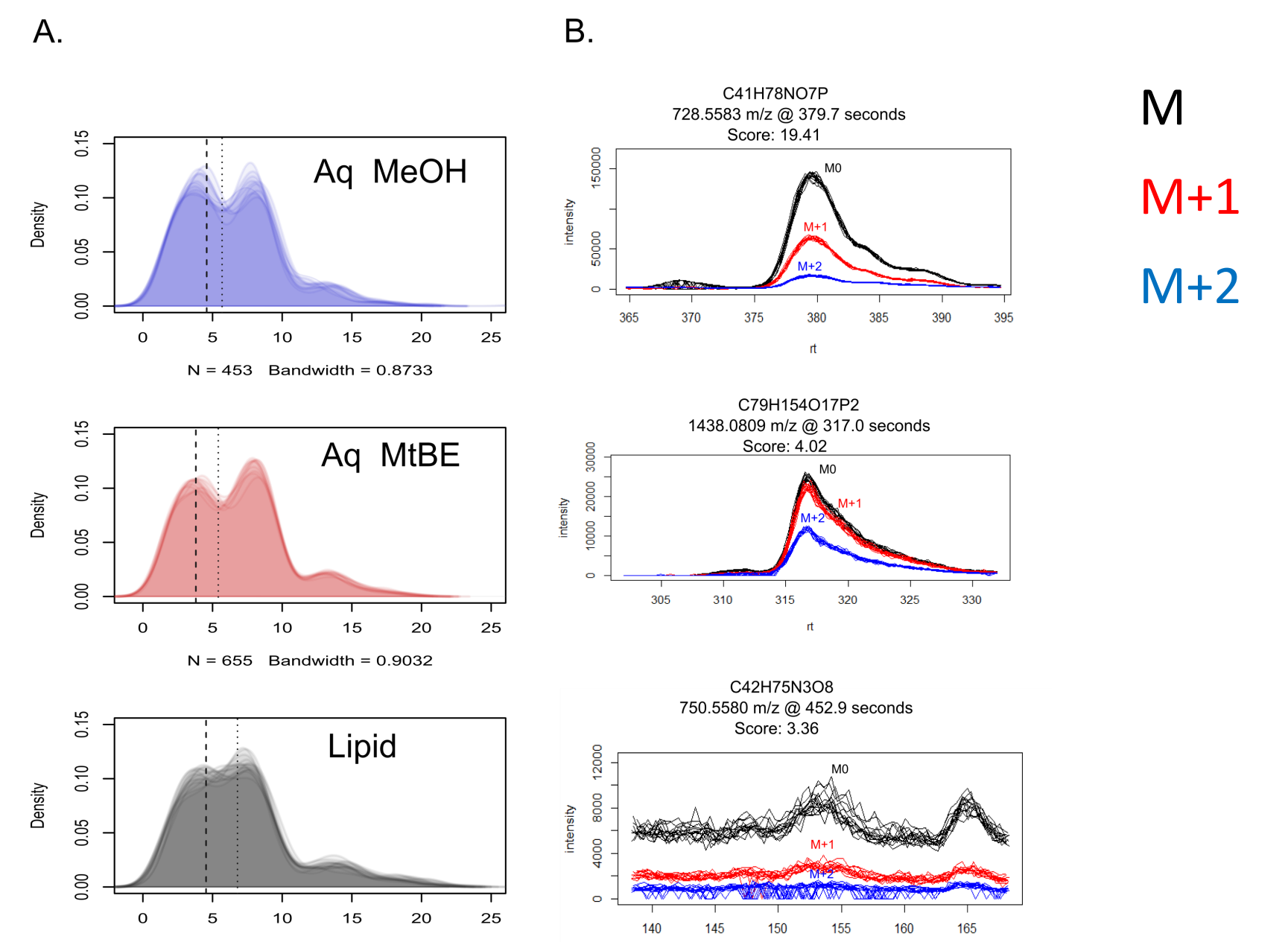


Figure 6.



Conclusions: Here we present MetMatch, a metabolomics MS1 search engine. Results indicate reproducible scoring assigning formulae to MS1 acquired isotope clusters. Searches are rapid (1-2 min), global (all spectra modelled in a sample)and are sensitive to the quality of the detected isotope clusters. Future work will include further validation with MSMS and standards.

This work was funded in part by grants NIH/NHLBI P20HL113445 and DARPA-BAA-13-14