LEVERAGING PUBLIC UNTARGETED METABOLOMICS DATA TO PROPAGATE ANNOTATIONS TO MILLIONS OF MS/MS SPECTRA

Wout Bittremieux, Nicole Avalon, Sydney P. Thomas, Mingxun Wang, Pieter C. Dorrestein

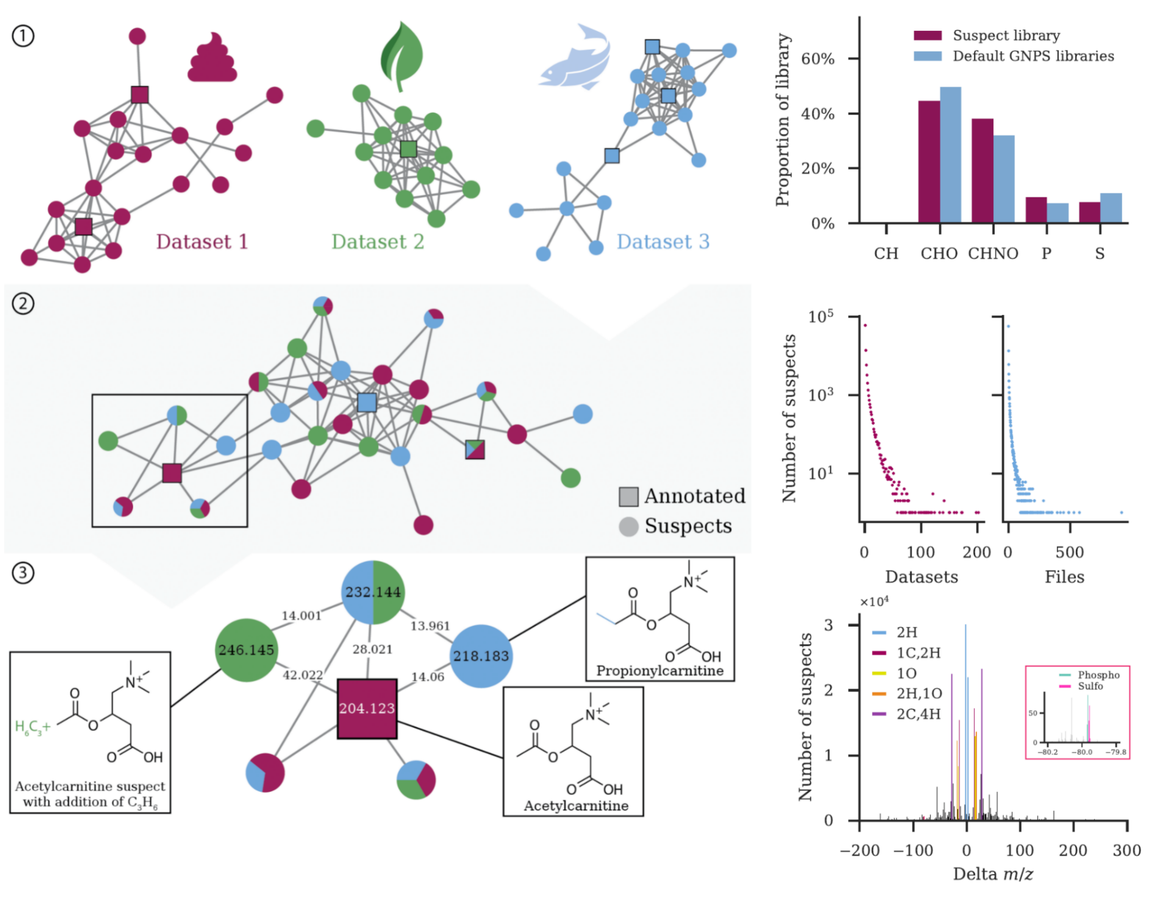
University of California San Diego, La Jolla, CA, USA
wbittremieux@health.ucsd.edu

Introduction

Despite important advances in the elucidation of MS/MS spectra over the past decade, the majority of information collected in mass spectrometry experiments still lacks a biological interpretation. For example, state-of-the-art spectral library searching of 1599 public **untargeted metabolomics** datasets on GNPS/MassIVE, corresponding to 485 million MS/MS spectra, resulted in 5.7% annotation rate. **Molecular networking** is an increasingly popular technique to propagate information to similar MS/MS spectra to increase the spectrum annotation rate.

Methods

Taking both direct peak matches and neutral loss peak matches into account, we have propagated annotations from the molecular networks associated with **1.2 billion MS/MS spectra from GNPS/MassIVE, Metabolights, and Metabolomics Workbench**, and created the open source GNPS nearest neighbor suspect spectral library.

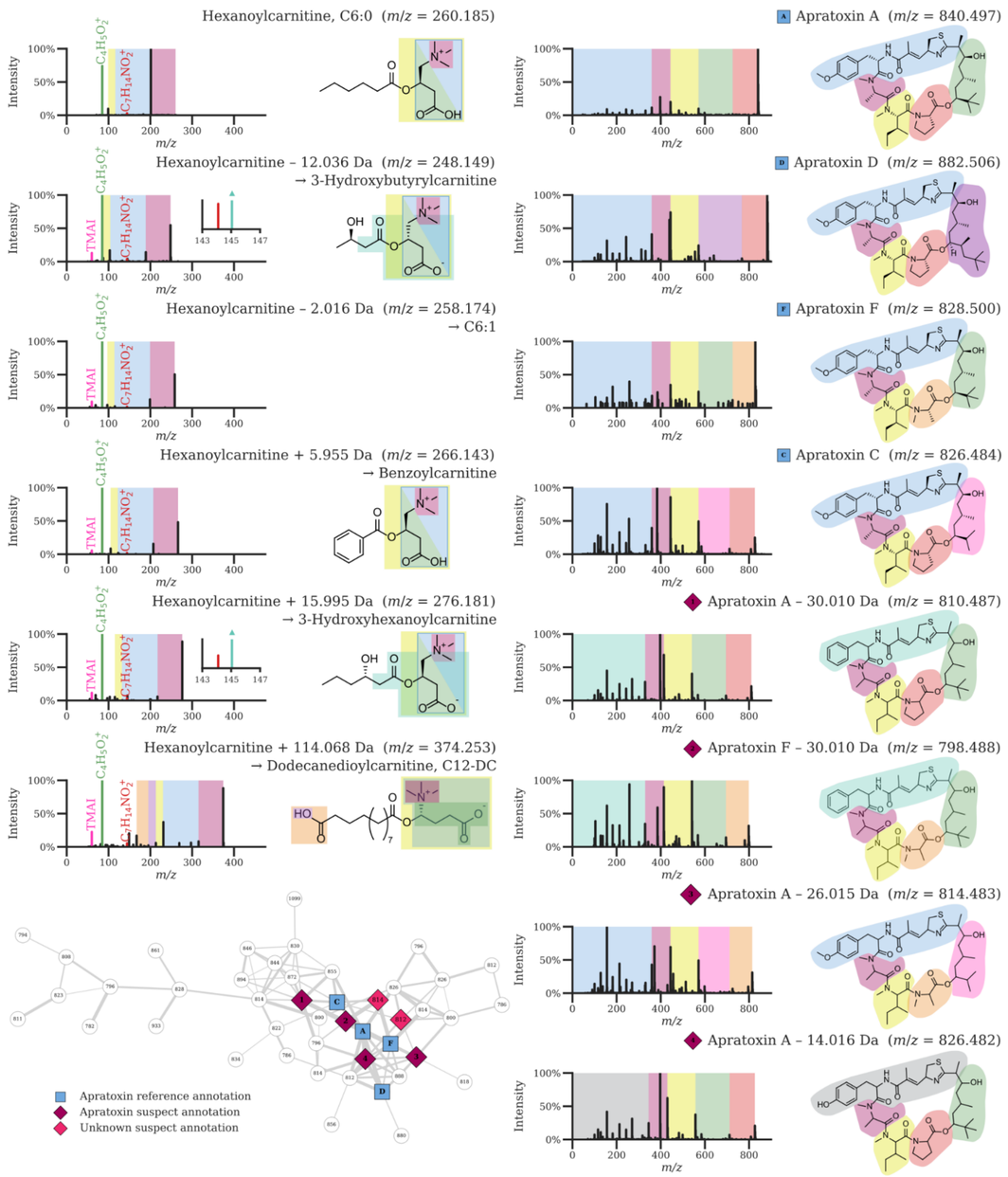


Molecular modifications: Repository-scale molecular networking to create the nearest neighbor suspect spectral library revealed 1350 **common modification mass differences**. These data give chemical insights into the processes that molecules undergo *in vivo* and during mass spectrometry analysis.

Results

Nearest neighbor suspect spectral library: The suspect spectral library consists of **87,916 novel reference spectra**, derived from matches to experimentally observed spectra. Suspects are molecules that are **structurally related to known reference molecules**, with the location of the structural modification unspecified.

Annotation performance: Library searching of heterogeneous sample types in publicly available datasets on GNPS/MassIVE using the suspect spectral library resulted in a **166% spectrum match rate** on average, compared to analysis using the default GNPS community spectral libraries only.



Conclusion

The nearest neighbor suspect spectral library is **freely available with an open source license** through GNPS, where it can be used for spectral library searching and molecular networking. Additionally, as part of the GNPS living data infrastructure, it is used to automatically annotate all newly deposited data to GNPS/MassIVE.

@bittremieux

bittremieux.be