Establishing a computational screening framework to identify environmental exposures using untargeted GC-HRMS

Juni Kim[1], Anna R Robuck[2], Douglas I Walker[2]

1. Stanford Online High School, 415 Broadway, Redwood City, CA; 2. Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York NY

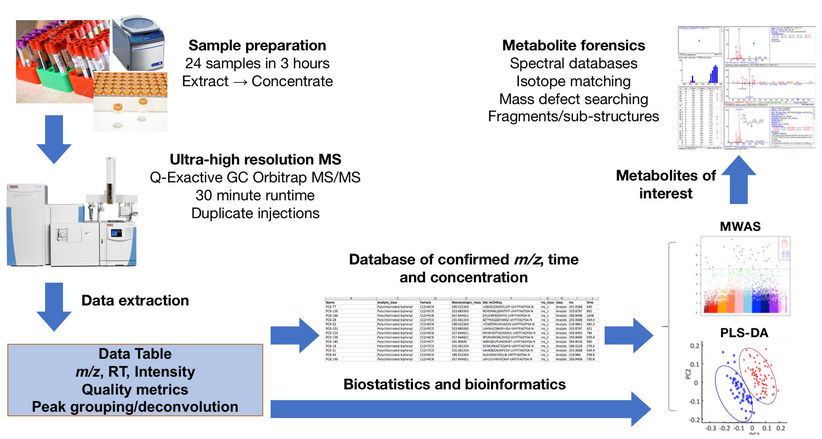
Introduction

- The human exposome, which includes chemical exposures over the lifetime and their effects, is now recognized as an important measure for understanding human health.
- Currently, there are no methods available that allow you to measure the full exposome.
- The High Resolution Exposomics Laboratory at Mount Sinai develops methods based upon high-resolution mass spectrometry to detect and measure the exposome.
- Gas-chromatography with high resolution mass spectrometry (GC-HRMS) can detect around 75,000 chemical signals, which include over 550 confirmed environmental pollutants.
- These analyses are completed using **untargeted methods**, meaning the levels of many chemicals can be detected, but the identity may not be known.
- While these measures are important for exposome research, the complexity of the data makes identification of environmental chemicals and key effects challenging.
- Many research efforts now focus on software tools that can quickly and accurately detect environmental chemicals in untargeted data.
- **The goal of this research project was to establish new tools for identifying halogenated chemicals, which most include chemicals arising from environmental exposures, in blood samples. These measures are important for determining which exposures are linked to health.**

How do we measure the exposome?

- Your blood includes a very large number of different chemicals, ranging from metabolites that ensure our body functions correctly, to environmental chemicals present from a very wide range of exposures.
- We know many different chemicals are in blood, but it is challenging to predict which ones exactly will be present.
- We now use untargeted methods to measure the exposome in blood samples because this gives us the greatest chance of measuring the most exposures.
- How this is done is shown in **Figure 1**. Blood samples are first measured using GC-HRMS. Detected chemicals are then extracted into a table that includes information on the mass of the compound and how much is present in the samples. These data are then analyzed to identify environmental chemicals and see how those are associated with a disease.
- **However, many naturally occurring metabolites are present. The question arises of how we determine which chemicals are from exposures.**

Figure 1. A workflow to measure the exposome



The research problem

- Untargeted GC-HRMS detects 1000's of signals, but only a few may be important to the disease we are studying.
- We need ways to remove chemicals that occur naturally, and only focus on environmental chemicals.
- Many pesticides and environmental pollutants contain chlorine (Cl) or bromine (Br) (these are called halogenated compounds).
- Cl and Br have special properties that allow them to be detected as two different isotopes, and these are present in HRMS data as the M and M+2 peaks. By searching for these, we can predict which signals have Cl or Br, and are likely environmental chemicals.

Goals

1. Develop an algorithm that can rapidly detect Cl and Br chemicals in HRMS data.
2. Make the script easily usable by others for analysis.
3. Apply this tool to samples from healthy participants and patients with liver disease to see if there were different exposures for sick individuals for selected signals.

Goal 1: Algorithms for Cl and Br detection

- We first searched for published R packages that allow searching for different isotopes, including Cl and Br
- The package "NonTarget" provided the base functionalities needed to identify Cl and Br chemicals.The function "pattern.search" filters out halogenated organic compounds by measuring the M and M + 2 mass/charge differences in peaks.
- The original algorithm is for all peaks detected in a sample and runs in quadratic time. We needed to modify the algorithm to work with the tens of 1000's of spectra detected using GC-HRMS.
- We created a new function based upon the pattern.search function that allows evaluation of multiple peak clusters, has multi-core support, and provides a simplified output identifying the Cl and Br compounds. The flow diagram for the function is provided in **Figure 2.**
- The algorithm was tested using a dataset containing 46,000 spectra from pesticide plant workers exposed to dioxins, which are very toxic chlorine containing chemicals. We hypothesized these workers would contain a large number of chlorine containing chemicals. The function performance and results are provided in **Table 1** and **Figure 3.**

Figure 2. Algorithm flow chart for detection of Cl and Br compounds

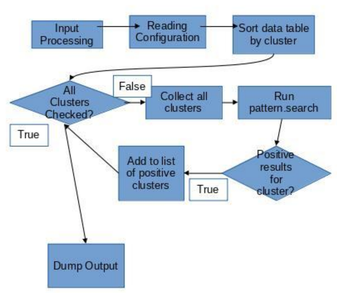
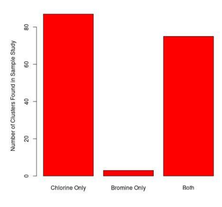


Table 1. Function runtime as a function of cores for 46,000 peaks

Cores Used	Execution Time
1	27s
2	16s
3	10s
4	8.9s
6	8.7s

Figure 3. Number of Cl/Br containing spectra from the pesticide plant workers



Goal 2: Publishing our function

- First, we created an easily accessible GitHub page that can be updated and includes example datasets: https://git.junickim.me/junikimm717/Patternmatch
- Next, we created to test data set with known compounds containing Cl and Br. This allows users to test their accuracy and false positive rate.
- All datasets are available on the releases page of Juni's Gitea Instance.
- Finally, we created a website that can be searched for, and contains information about the function, example datasets, and scripts (**Figure 4**). Users are recommended to check the website to learn about the script. The website can be accessed from the QR code below.

Figure 4. Website for demonstrating our function

Goal 3: Application to liver disease.

- Our final goal was to apply this algorithm to real sample data measured in individuals with liver disease, including primary schlerosing cholangitis (PSC) and primary biliary cholangitis.
- The cause if these diseases is unknown and may be caused by exposures.
- We first identified all Cl and Br containing chemicals in these samples (**Figure 5**) and tested to see if they were different between people with PSC or PBC, and controls.
- We identified 22 Cl or Br compounds. 6 of these were different for PBC patients at $p<0.05$. 4 were different for PSC patients at $p<0.05$. Plots for two of the compounds in cases and controls are in **Figure 6**.
- These chemicals may be important exposures contributing to the diseases.

Figure 5. Number of Cl/Br containing spectra from PSC/PBC samples

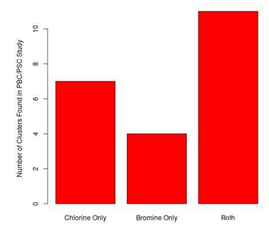
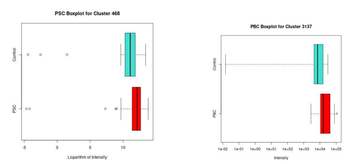


Figure 6. Chemicals different in PSC and PBC



Conclusions

- We have developed an improved algorithm for measuring Cl and Br compounds in GC-HRMS data
- Future studies will use these results to study exposures and disease.