MarMic Laboratory Rotation 3 Biweekly Report

M. Sc. student: Matthew Schechter
Supervisor: Manuel Liebeke

June 19 - July 3, 2017

Todo

Tasks

- upload data to compound discoverer
- Metabolite extraction with Dolma and learn how to use Xcalibur software

Tasks in progress

- successfully upload and annotate data to XCMS
- continue reading relevant literature
- write script to query potential metabolites to annotated metabolites
- read how to use SPLASH (chemical unique identifier generator)

Tasks Completed

- Composed and share instructions for msConvert and GNPS database
- Install Pathway tools
- upload data to GNPS
- compose lab rotation objectives and timeline
- initialize git repo (https://github.com/mschecht/MassSpec_labrotation) for lab rotation
- import sox and mox genomes to pathway tools
- output list of potential lipid metabolites from relevant genomes

Lab rotation objectives

- Read relevant literature on multi-omic data processing
- Explore alternative databases for annotating unknown spectra (GNPS)

- Learn metabolite extraction methods, how to run the LCMS, and how to work with the Chromealan/Xcalibur software
- Learn Pathway Tools software and output a potenital list of lipome metabolites from relevant genomes of Bathy and its symbionts
- Develop pipeline to query LCMS annotated spectra to potential metabolite list from Pathway Tools

Results and progress

MS databases

These past two weeks I explored the database GNPS and how to upload raw mass spec data to it. After I successfully uploaded and annotated LCMS data, I composed uploading instructions for the mass spec group. Later, I learned how to upload data to the ubiquitous XCMS database. Unfortunately, the data did not successfully annotated by the server due to this error:

```
2017-06-28 06:43:10 : CAMERA peak annotation failed (findIsotopes).
Processing data without annotation
```

I searched online to see if anyone has had this issue in the past but did not find a similar error. The XCMS customer service answered promptly saying that the data I uploaded may of been unresolvable. This is inconsistent because Maggie uploaded a similar Bathy LCMS data set successfully. I believe my issue may be due to a file conversion error and I will look into solving this is week.

The GNPS mass spec annotation server is unique because it uses molecular networking to annotate unknown mass spectra. The website encourages publishing of uploaded data to its "social network" to expand its molecular network. This aids the underlying database because as the network size increases, the more unknown spectra can be annotated due to the increasing ability of the database to find similarity of unknown spectra to known spectra . GNPS claims data on the server is "alive" because as the network continually annotates unknown spectra even after the original analyses due to the ever expanding network. This aspect is very relevant to the mass spec group because most of the compounds that differentiate host/symbiont metabolome are rare and if not now may be annotated in the future via GNPS molecular networking as its database expands.

Pathway Tools

Adrienn provided me with a detailed tutorial on utilizing the Pathway Tools software. This included installation, genome uploading, genome navigation, and querying the genome of pathways, enzymes, and metabolites. From this I was able to upload the SOX and MOX symbiont genomes from Bathy Azoricus and

output a list of all potential list of lipome metabolites. To do this I read up on membrane lipids and how they are characterized to generate a list of key lipid related words to query against the genomes ("Cell Membranes" Lukas Beuhler).

lipid, fatty, prenol, sterols, ketide, steroids, Terpene

MS Pipeline

The main goal of this lab rotation is to generate a robust pipeline that can resolve species specific lipome metabolites between Bathymodilus and its symbionts. To accomplish this, a list of species lipome metabolites will be generated via Pathway Tools. Next a *species specific* list will be generated by removing common metabolites between species. From here, the *species specific* lists of metabolites can be queried against the annotated spectra from the LCMS. This will potentially lead to annotated MALDI data of various Bathy symbiont species.

An image of the proposed pipeline can be viewed on the lab rotation git hub README.md

So far in the pipeline I generated a potential list of lipome metabolites for Bathy and its symbionts and outputed LCMS generated lipome metabolite annotations from GNPS and XCMS databases. The next step will be to bind the database outputs and query the two lists together. Binding the annotations lists has been challenging because each database chooses a different chemical notation style i.e. colloquial, IUPAC, SMILE, etc. The chemical name would ideally be the best binding key between the different annotations databases because other chemical properties such as molecular weight and mass/charge are may be the same in more than one compound. This issue has also delayed a side project to create a Venn diagram comparing the annotation output of various annotation databases.

One idea I came across is a database independent unique identifier generator called SPLASH. This script uses a combination of molecular mass and SMILE to generate a unique identifier for any compound. A large amount of annotation databases use this key (except XCMS) which will allow me to bind annotation output to make comparisons and query against the species specific lipome metabolite list.

Outlook

This week I want to complete the pipeline by using SPLASH in order to bind annotation outputs. I believe this chemical unique identifier generator has a lot of potential for connecting multi-omic data in the future. Additionally, I will be generating my own Bathy data set with the help of Dolma. By the end of the week I should be able to run my own data set through the pipeline to yield species unique lipids.