1 **Mining mass spectrometry data: Using new computational tools to**
2 **find novel organic compounds in complex environmental mixtures**

3

4 Krista Longnecker* and Elizabeth B. Kujawinski

5 Woods Hole Oceanographic Institution, Department of Marine Chemistry and Geochemistry,

6 Woods Hole, MA, 02543 USA

7 *Corresponding author: klongnecker@whoi.edu

8 Short title: Lyso-sulfolipids in marine samples

12

13

Abstract

14

15   Untargeted metabolomics datasets provide ample opportunity for discovery of novel

16   metabolites. The major challenge is focusing data analysis on a short list of metabolites. Here,

17   we apply a combination of computational tools that serve to reduce complex mass spectrometry

18   data in order allow us to focus on new environmentally-relevant metabolites. In the first portion

19   of the project, we explored mass spectrometry data from intracellular metabolites extracted

20   from a model marine diatom, *Thalassiosira pseudonana*. The fragmentation data from these

21   samples were analyzed using molecular networking, an on-line tool that clusters metabolites

22   based on shared structural similarities. The features within each metabolite cluster were then

23   putatively annotated using MetFrag, an *in silico* fragmentation tool. Using this combination of

24   computational tools, we observed multiple lyso-sulfolipids, organic compounds not previously

25   known to exist within cultured marine diatoms. In the second stage of the project, we searched

26   our environmental data for these lyso-sulfolipids. The lyso-sulfolipid with a C14:0 fatty acid

27   was found in dissolved and particulate samples from the western Atlantic Ocean, and a culture

28   of cyanobacteria grown in our laboratory. Thus, the putative lyso-sulfolipids are present in both

29   laboratory experiments and environmental samples. This project highlights the value of

30   combining computational tools to detect and putatively identify organic compounds not

31   previously recognized as important within *T. pseudonana* or the marine environment. Future

32   applications of these tools to emerging metabolomics data will further open the black box of

33   natural organic matter, identifying molecules that can be used to understand and monitor the

34   global carbon cycle.

Introduction

Organic matter is a complex and heterogeneous mixture of compounds that challenges scientists investigating its role in global biogeochemical processes. Organic compounds are formed from inorganic carbon through the actions of primary producers. These compounds can then be transformed into new organic compounds through biological activity, and may ultimately be converted back to inorganic carbon. Through these actions, the biological processes and the composition of organic matter are tightly coupled (Azam et al., 1993). Each organic compound also has its own source and sink dynamic, which potentially varies with biotic and abiotic parameters in an ecosystem. Yet, we have only identified a small fraction of organic compounds that exist in the environment and have limited understanding of the roles of these compounds in the carbon cycle. Organic matter likely contains thousands of individual molecules, making comprehensive identification an elusive goal. However, biologically-derived molecules such as metabolites are likely to play an important role in the carbon cycle, either as growth substrates or growth factors for microbes. Thus identification of these molecules could provide insights into the function and metabolism of microbes that govern the ocean carbon cycle (Moran et al., 2016).

We present a novel combination of computational tools with the goal of more efficiently identifying individual compounds within a complex mixture of organic matter. This project expands our ability to analyze untargeted metabolomics data and is one of several methods that can be used to characterize organic matter in aquatic environments (e.g., Longnecker et al.,

55   2015a; Longnecker and Kujawinski, 2016; Treutler et al., 2016; van der Hooft et al., 2016). Here,

56   our analysis is based on two modes of analyzing fragmentation spectra from organic molecules.

57   These fragmentation spectra can be grouped based on the similarity of fragment m/z values

58   measured within a set of samples (Frank et al., 2007; Nguyen et al., 2013). This clustering of

59   fragmentation spectra, also called molecular networking (Yang et al., 2013), has proven useful in

60   finding known compounds within microbial colonies growing in the laboratory (Watrous et al.,

61   2012) and in environmental samples (Kharbush et al., 2016; Teta et al., 2015). Here, we combine

62   molecular networking with MetFrag, an *in silico* fragmentation tool (Wolf et al., 2010) that

63   presents potential compound identifications given a measured fragmentation spectra.

64   We introduce our approach through a comparison of laboratory data and environmental

65   data. The laboratory data were intracellular metabolites extracted from the centric diatom

66   *Thalassiosira pseudonana* which was grown under phosphate-limited and phosphate-replete

67   conditions. Thousands of intracellular metabolites are produced by *T. pseudonana*, yet our

68   previous research revealed that most of these metabolites cannot be identified (Longnecker et

69   al., 2015b). Here, we were able to identify a set of metabolites not previously known to be

70   important within diatom physiology and then expanded our analysis to investigate the extent to

71   which these compounds were found within marine ecosystems.

Materials and Methods

Untargeted metabolomics experiments with a cultured marine diatom

Thalassiosira pseudonana (CCMP #1335) was cultured axenically in modified L1 media. There were two treatments: phosphate-replete (36 μM PO_4^{-3}) and phosphate-limited (0.4 μM PO_4^{-3}). The experiment began with the addition of 30 ml of *T. pseudonana* in exponential phase to two-thirds of the flasks which contained 300 ml of media. The remaining one-third of the flasks were designated as cell-free controls. The cultures were maintained under a 12:12 light:dark regime (84 μmol m^{-2} s^{-1}). Two flasks with cells and one cell-free control for each treatment were sampled at four time points: 0, 2, 8, and 10 days. At each time point, cells were captured by gentle vacuum filtration on 0.2 μm Omnipore (Millipore) filters. The intracellular metabolites were extracted using a method modified from Rabinowitz and Kimball (2007), as described previously (Kido Soule et al., 2015). The extracts were re-dissolved in 95:5 water:acetonitrile and deuterated biotin (final concentration 0.05 μg ml^{-1}) and analyzed in negative ion mode with liquid chromatography (LC) coupled by electrospray ionization to a 7-Tesla Fourier-transform ion cyclotron resonance mass spectrometer (Thermo Scientific, FT-ICR MS). LC separation was performed using a Synergi Fusion reversed phase column (Phenomenex, Torrance, CA). The chromatography gradient was: an initial hold of 95% A (0.1% formic acid in water) : 5% B (0.1% formic acid in acetonitrile) for 2 minutes, ramp to 65% B from 2 to 20 minutes, ramp to 100% B from 20 to 25 min, and hold until 32.5 minutes. The column was re-equilibrated for 7 min between samples with solvent A. In parallel to the FT acquisition, four data-dependent fragmentation (MS/MS) scans were collected at nominal mass resolution in the ion trap (LTQ).

93   Samples were analyzed in random order with a pooled sampled run every six samples in order

94   to assess instrument variability.

95   The data files from the mass spectrometer were converted to the open-source mzXML

96   format using the MSConverter tool (Kessner et al., 2008). After this step, the data files were

97   processed in parallel using two different data analysis pipelines (Figure 1). The first pipeline

98   involves the use of XCMS (Smith et al., 2006) to conduct the peak picking, alignment, and

99   retention time correction that is standard for untargeted metabolomics data analysis (Johnson et

100   al., 2014). The output from this analysis is a list of 'mzRT features' which are defined as unique

101   combinations of *m/z* values and retention times that have passed our quality control checks. The

102   peak area for each mzRT feature provides an estimate of the relative levels of the feature during

103   the experiment. The freely-available XCMS2 code (Benton et al., 2008) was used to generate the

104   list of MS2 spectra obtained for the mzRT features and CAMERA provided details about

105   possible isotopologues within the dataset (Kuhl et al., 2012). The untargeted metabolomics data

106   from the *T. pseudonana* cultures are available with accession code MTBLS154 at MetaboLights

107   (Haug et al., 2013).

108   **Molecular networking**

109   Molecular networking is one of the functions available at the Global Natural Products

110   Social Molecular Networking site (GNPS, http://gnps.ucsd.edu) that has recently been described

111   by Wang et al. (2016). We used the molecular networking tool (Yang et al., 2013) to group our

112   mzRT features based on similarities in the MS2 fragmentation spectra. Molecular networking

113   takes advantage of the MS2 information within unprocessed mzXML files. The *T. pseudonana*

114   data used here are available at GNPS with MassIVE ID MSV000080990. The molecular

115   networking analysis was run with the following parameters: parent mass tolerance = 1 Da, ion

116   tolerance = 0.5 Da, minimum cluster size = 3, minimum pairs cosine = 0.7, score threshold = 0.5,

117   network topK = 10, run MSCluster = TRUE. The output is a network calculated based on the

118   overlap in peaks within the MS2 spectra. Within the network, each node is an mzRT feature

119   with MS2 data. The nodes are connected by edges, the width of each edge is a measure of the

120   similarity in the MS2 spectra between two nodes. We used Cytoscape (Smoot et al., 2011) as a

121   visualization tool to annotate the nodes to include information about each mzRT feature

122   (experimental treatment, *m/z* value, and strength of connection with other mzRT features within

123   the cluster).

124   *In silico* **calculation of fragmentation spectra**

125   We used MetFrag (Wolf et al., 2010) to putatively annotate mzRT features based on their

126   MS2 spectra. The analysis starts with a search for exact mass of the parent ion against a

127   database; the available database options are currently KEGG, PubChem, and ChemSpider. Once

128   potential matches are located, MetFrag generates *in silico* fragments from the parent compound

129   and compares the *in silico* fragments to the measured MS2 fragments uploaded by the user. We

130   used the data output from the XCMS processing (left side of Figure 1) to provide the *m/z* values

131   and peak intensities that were used in the MetFrag search. This required a computational step

132   that matched the *m/z* values and retention times from the XCMS output with the results file

133   produced by the molecular networking tool. While this required extra steps, the protocol

134   allowed us to use the data that passed our quality control checks within XCMS in lieu of relying

135   on the unprocessed mzXML files used by the molecular networking tool. The fragmentation

136   spectra from the mzRT features of interest were searched in MetFrag using a 1 ppm window for

137   the parent ion search in the ChemSpider database. For the fragments we set MZabs = 1, and

138   MZppm = 30. The output was manually inspected to assess possible annotations.

139   **Comparisons to existing untargeted metabolomics data**

140   We used the domdb database (Longnecker et al., 2015a) to compare select metabolites

141   from the *T. pseudonana* cultures with our existing untargeted metabolomics data. The database

142   allows searching by *m/z* and retention time. The search thus requires samples to be analyzed

143   using the same analytical methods to minimize variability in retention times that occur with

144   different liquid chromatography conditions. One of the metabolites discussed below was found

145   in four sets of samples processed within our laboratory: a culture experiment with *Synechococcus*

146   *elongatus*, incubation experiments with Atlantic Ocean seawater collected from 70 m and 700 m,

147   and sinking particles collected from net traps deployed for 24 hours at 150 m at select stations in

148   the Atlantic Ocean (Table 2).

149   **Results and Discussion**

150   In the following sections, we start with untargeted metabolomics data from a laboratory

151   experiment to reveal the value of integrated molecular networking and *in silico* analysis of

152   fragmentation spectra. Individually, each of these tools provides valuable insights into mass

153   spectrometry data. Collectively, the combination allows for efficient data mining into the

154   putative annotations of select metabolites. The time needed to putatively identify metabolites is

155   the most time consuming aspect of metabolomics experiments. Thus technical advances, such

156   our novel combination of computational tools, provide a new means for researchers to focus on

157   ecologically interesting results.

158   **Molecular networking: clustering of mzRT features by MS2 spectra**

159   Molecular networking formed clusters within the *T. pseudonana* data based on the

160   similarity between the MS2 spectra (Figure 2). With the parameters we selected at GNPS, we

161   obtained a set of thirteen clusters representing 90 mzRT features within our dataset. In

162   comparison, 2825 mzRT features were in the final, aligned dataset from the XCMS analysis and

163   1303 of these mzRT features had associated MS2 spectra. Thus, molecular networking helped to

164   constrain the dataset to a tractable number of features for additional analysis. Alternatively, a

165   user could set the input parameters at GNPS to allow a more conservative analysis with higher

166   numbers of linked mzRT features. In our analysis, the clusters linked a minimum of two mzRT

167   features (e.g, clusters K, L, or M) and maximum of 19 mzRT features (cluster A). The different

168   colors in the network diagram indicate whether an mzRT feature is present only in the

169   phosphate-replete or phosphate-limited conditions, or is present under both growth conditions.

170   For example, the cluster marked A contains mzRT features present under both growth

171   conditions individually as well as together, while cluster I only contains mzRT features present

172   during phosphate-limited growth. These mzRT features represent promising targets for

173   markers that are unique to specific ecological conditions. Yet, because molecular networking

174   considers unprocessed mzXML files, there are poor-quality mzRT features present as nodes in

175   the network (i.e., the mzRT features that do not pass the quality control checks in the XCMS

176   processing, see Methods). For example, all the mzRT features in cluster F were removed by

177   XCMS processing.

178   One benefit of the molecular networking tool is its ability to group a compound with its

179   isotopologues. The MS2 fragments from a compound and its isotopologues will have similar

180   fragmentation spectra, even though the m/z values for the fragments may differ by the mass

181   difference between ^{12}C and ^{13}C. In cluster A, six of the mzRT features are paired sets of

182   metabolites with the charged compound as one node and the charged compound with a single

183   ^{13}C atom as a second node within the cluster. The molecular networking tool does not annotate

184   these MS2 fragmentation spectra as originating from a compound and its isotopologues; rather

185   this distinction was identified by the CAMERA algorithm. In order to identify isotopologues

186   within the clusters, we combined the output from the molecular networking results with the

187   processed XCMS/CAMERA results. The integration of these two outputs streamlined our

188   identification efforts by removing ^{13}C compounds from further analysis with ^{12}C-based

189   computational tools such as MetFrag.

190   The GNPS website also provides users with the opportunity to compare measured

191   fragmentation spectra with fragmentation spectra stored at GNPS (Wang et al., 2016). The

192   fragmentation spectra stored at GNPS originate from any user consenting to the public use of

193   their data. However, in the case of the *T. pseudonana* dataset, none of the mzRT features had a

194   corresponding match to a metabolite in the GNPS database. Thus, while GNPS contains

195   increasing numbers of fragmentation spectra, the database is not yet a comprehensive survey of

196   organic compounds from environmental mixtures. Yet, even without the database match,

197   inspection of the nodes within cluster A revealed a set of mzRT features with direct relevance to

198   diatom physiology in marine environments, which could be putatively identified with MetFrag.

199   **Putatively annotating mzRT features within a cluster based on fragmentation spectra**

200   The identification of unknown metabolites is often a primary goal of untargeted

201   metabolomics projects as researchers seek to quantify the biogeochemical cycling of known

202   organic compounds. We used the classification scheme defined by Sumner et al. (2007) to guide

203   our descriptions of the putative metabolite annotations. Within this scheme, the metabolites we

204   discuss below are Level 2 identifications which are putatively annotated without chemical

205   reference standards, but are based on spectral similarities with data from public or commercial

206   libraries.

207   The similarities in the MS2 spectra in the mzRT features grouped by molecular

208   networking into cluster A is evident when the MS2 fragments are plotted together (Figure 3).

209   Note that all four of the mzRT features plotted in Figure 3 were observed as the charged ion and

210   the isotopologue with a single ^{13}C atom; Figure S1 shows the corresponding plots of the MS2

211   fragments from ^{13}C compounds. All four of these mzRT features have a sulfoquinovosyl head

212   group (sulfoquinovose, Figure 3A), which is a derivative of glucose with the 6-hydroxyl

213   replaced by a sulfonate group (Benning, 1998). Three of the mzRT features differ by the fatty

214   acid chain, with 14:0, 16:0, and 16:1 as potential options. The 14 and 16 refer to the number of

215   carbon atoms in the fatty acid while the 0 or 1 refers to the number of double bonds within the

216   fatty acid. Figure S2 shows the structure of each of these metabolites and Table S1 includes

217   images of the MS2 fragments and the distribution of the fragments across the mzRT features

218   from *T. pseudonana*. As further support of our putative annotation of these mzRT features, three

219   of the fragments we measured were noted as characteristic fragments of sulfoquinovosyl

220   monoacylglycerols (SQMG) by De Souza et al. (2006). More generally, these SQMG compounds

221   are known as lyso-sulfolipids. While analysis of these mzRT features with authentic standards

222   would be ideal, these compounds are not commercially available. Thus, in the absence of such

223   standards, the putative annotation of these compounds is state-of-the art.

224   **Lyso-sulfolipids in diatoms**

225   Lipids are the structural underpinning of the bilayer membrane surrounding a cell.

226   However, the lipids that comprise cell membranes have a polar head group and two non-polar

227   fatty acid tails, and this combination causes the lipids to self-assemble into a bilayer membrane.

228   In contrast, the lyso-sulfolipids observed in the present project have only a single fatty acid. The

229   biochemical origin of these lyso-sulfolipids is unknown, but here we consider several

230   possibilities. The lyso-sulfolipids could have been derived from sulfoquinovosyl diacylglycerols

231   (SQDG), the corresponding sulfolipid with two fatty acids which is an essential component of

232   photosynthetic membranes. This process has been observed to be enzymatically possible in

233   some (Gupta and Sastry, 1987; Wolfersberger and Pieringer, 1974; Yagi and Benson, 1962), but

234   not all organisms (Burns et al., 1977). Alternatively, only a single fatty acid could be combined

235   with the sulfoquinovosyl head group to form the lyso-sulfolipid. Finally, the lipids could have

236   degraded during sample processing (Allen et al., 1970), although we consider this option less

237   likely because the filters were stored frozen at -80° C and analyzed using mass spectrometry

238   within 10 days after extraction. Additional research will be needed to determine which process

239   is occurring within our samples.

240   Lyso-sulfolipids have been observed in cultures of marine algae (El Baz et al., 2013), but

241   not, to our knowledge, within cultures of *T. pseudonana*. Yet, SQDG lipids play a prominent role

242   in *T. pseudonana*'s physiological response to phosphorus limitation as the diatom switches from

243   phosphorus based lipids to sulfolipids in order to spare phosphorus for other cellular functions

244   (Martin et al., 2011; Van Mooy et al., 2009). SQDG has also been observed in single cell

245   measurements of *Chlamydomonas* grown under nitrogen limited conditions (Cahill et al., 2015),

246   which may indicate a broad physiological need for SQDG under nutrient limited growth. The

247   ecological role of lyso-sulfolipids within the metabolism of *T. pseudonana* is not known. Yet, as

248   with SQDG, these lipids are more prevalent under conditions of phosphate-limited growth

249   (Figure 4), which suggests the lyso-sulfolipids are also playing a role in phosphorus scavenging

250   within the cells.

251   In addition to the lyso-sulfolipids, we also putatively annotated the sulfoquinovosyl head

252   group attached to a 21-carbon sterol (Figure 5). On a per-cell basis, this metabolite was more 1.7

253   times more prevalent under phosphate-limited growth conditions, although a significant

254   amount of the compound was also found in the phosphate-replete cultures. *T. pseudonana*, like

255   all eukaryotes, makes sterol compounds and uses them to maintain the structural integrity of its

256   cell membrane. Sterols in *T. pseudonana* are primarily 27- or 28-carbon sterols (Rampen et al.,

257   2010; Véron et al., 1998), larger than the 21-carbon sterol we observed. In the marine

258   environment, 21-carbon sterols are not common, although they are present in marine sponges

259   (Ballantine et al., 1977). Given the novelty of a sulfoquinovosyl head group attached to a sterol,

260   we cannot speculate as to the role of this compound within the metabolism of *T. pseudonana*.

261   **Additional observations of lyso-sulfolipids in marine samples**

262   The lyso-sulfolipids are not unique to laboratory cultures with *T. pseudonana*. Using our

263   domdb database, we found the C14:0 lyso-sulfolipid in four sets of samples processed by our

264   laboratory using the same methods described for *T. pseudonana* (Table 2). In three cases, we have

265   extracts from paired filters and filtrate samples (Figure 6A, B, and C); the C14:0 lyso-sulfolipid

266   was always found at elevated levels in the filtrate compared to the filters. The C14:0 lyso-

267   sulfolipid was also found in particulate material captured by net traps deployed for 24 hours at

268   150 m (Figure 6D); no filtrate was processed for the net trap samples. None of the other

269   compounds listed in Table 1 were found in any of our samples, nor was the C14:0 lyso-

270   sulfolipid found in the filtrate from the experiment with *T. pseudonana* or in any of our sample

271   processing or instrumentation blanks (data not shown). Yet, the presence of the C14:0 lyso-

272   sulfolipid in samples spanning from the surface ocean, to deep seawater, and to laboratory

273   cultures hints at the prospect of a set of organic compounds that may provide information on

274   the physiological state of organisms in marine environments.

275   **Conclusions and ecological significance**

276   We used a combination of molecular networking and *in silico* fragmentation

277   computational tools to find a novel class of lipids within a set of ultrahigh resolution mass

278   spectrometry data. Without this combination of computational tools, we would not have

279   focused on putatively identifying these compounds, nor would we have known to look within

280   our existing data to find other sources for lyso-sulfolipids. The lyso-sulfolipids were the only

281   compounds to result from the computational tools described here. Beyond their classification as

282   lyso-sulfolipids, the compounds described here are sulfur-containing organic molecules which

283   are increasingly recognized as important within marine ecosystems (Ksionzek et al., 2016).

284   Organic sulfur compounds are transferred from autotrophic to heterotrophic microorganisms

285   (Durham et al., 2015; Malmstrom et al., 2004). In the process, select organic sulfur compounds

286   serve as signaling molecules to which heterotrophic bacteria respond (Johnson et al., 2016).

287   Lyso-sulfolipids have been shown to serve as signaling molecules and induce larval settlement

288   and metamorphosis in sea urchins (Takahashi et al., 2002) and corals (Tebben et al., 2015).

289   Finally, while the sulfoquinovosyl head group can be degraded to other organic sulfur

290   compounds (Denger et al., 2012; Felux et al., 2015), we did not observe any of the degradation

291   products within the particulate material sampled during these experiments (data not shown).

292   Given the presence of the lyso-sulfolipids in multiple experiments and field samples, we posit

293   that these compounds are serving an active role within the physiology of microbial cells.

294   Furthermore, the presence of the C14:0 lyso-sulfolipid in water samples from the Atlantic Ocean

295   provides a direct link between compounds found in laboratory cultures and compounds

296   observed in the marine environment.

References

Allen, C.F., Good, P., Holton, R.W., 1970. Lipid composition of *Cyanidium*. Plant Physiology 46, 748-751.

Azam, F., Smith, D.C., Steward, G.F., Hagström, Å., 1993. Bacteria-organic matter coupling and its significance for oceanic carbon cycling. Microbial Ecology 28, 167-179.

Ballantine, J.A., Williams, K., Burke, B.A., 1977. Marine sterols IV. C21 sterols from marine sources. Identification of pregnane derivatives in extracts of the sponge *Haliclona rubens*. Tetrahedron Letters 18, 1547-1550.

Benning, C., 1998. Biosynthesis and function of the sulfolipid sulfoquinovosyl diacylglycerol. Annual Review of Plant Physiology and Plant Molecular Biology 49, 53-75.

Benton, H.P., Wong, D.M., Trauger, S.A., Siuzdak, G., 2008. XCMS^2: Processing tandem mass spectrometry data for metabolite identification and structural characterization. Analytical Chemistry 80, 6382-6389.

Burns, D.D., Galliard, T., Harwood, J.L., 1977. Catabolism of sulfoquinovosyl diacylglycerol by an enzyme preparation from *Phaseolus multiflorus*. Phytochemistry 16, 651-654.

Cahill, J.F., Darlington, T.K., Fitzgerald, C., Schoepp, N.G., Beld, J., Burkart, M.D., Prather, K.A., 2015. Online analysis of single cyanobacteria and algae cells under nitrogen-limited conditions using aerosol time-of-flight mass spectrometry. Analytical Chemistry 87, 8039-8046.

De Souza, D.P., Saunders, E.C., McConville, M.J., Likic, V.A., 2006. Progressive peak clustering in GC-MS metabolomic experiments applied to *Leishmania* parasites. Bioinformatics 22, 1391 - 1396.

Denger, K., Huhn, T., Hollemeyer, K., Schleheck, D., Cook, A.M., 2012. Sulfoquinovose degraded by pure cultures of bacteria with release of C_3-organosulfonates: complete degradation in two-member communities. FEMS Microbiology Letters 328, 39-45.

Durham, B.P., Sharma, S., Luo, H., Smith, C.B., Amin, S.A., Bender, S.J., Dearth, S.P., Van Mooy, B.A.S., Campagna, S.R., Kujawinski, E.B., Armbrust, E.V., Moran, M.A., 2015. Cryptic carbon and sulfur cycling between surface ocean plankton. Proceedings of the National Academy of Sciences 112, 453-457.

El Baz, F.K., Baroty, G.S.E., Baky, H.H.A.E., El-Salam, O.I.A., Ibrahim, E.A., 2013. Structural characterization and biological activity of sulfolipids from selected marine algae. Grasas y Aceites 64, 561-571.

338 Felux, A.-K., Spiteller, D., Klebensberger, J., Schleheck, D., 2015. Entner–Doudoroff pathway for
339 sulfoquinovose degradation in *Pseudomonas putida* SQ1. Proceedings of the National Academy
340 of Sciences 112, E4298-E4305.

341 Fiore, C.L., Longnecker, K., Kido Soule, M.C., Kujawinski, E.B., 2015. Release of ecologically
342 relevant metabolites by the cyanobacterium, *Synechococcus elongatus* CCMP 1631. Environmental
343 Microbiology 17, 3949-3963.

344 Frank, A.M., Bandeira, N., Shen, Z., Tanner, S., Briggs, S.P., Smith, R.D., Pevzner, P.A., 2007.
345 Clustering millions of tandem mass spectra. Journal of Proteome Research 7, 113-122.

346 Gupta, S.D., Sastry, P.S., 1987. Metabolism of the plant sulfolipid-
347 Sulfoquinovosyldiacylglycerol: Degradation in animal tissues. Archives of Biochemistry and
348 Biophysics 259, 510-519.

349 Haug, K., Salek, R.M., Conesa, P., Hastings, J., de Matos, P., Rijnbeek, M., Mahendraker, T.,
350 Williams, M., Neumann, S., Rocca-Serra, P., Maguire, E., González-Beltrán, A., Sansone, S.-A.,
351 Griffin, J.L., Steinbeck, C., 2013. MetaboLights—an open-access general-purpose repository for
352 metabolomics studies and associated meta-data. Nucleic Acids Research 41, D781-D786.

353 Johnson, C.H., Ivanisevic, J., Benton, H.P., Siuzdak, G., 2014. Bioinformatics: the next frontier of
354 metabolomics. Analytical Chemistry 87, 147-156.

355 Johnson, W.M., Kido Soule, M.C., Kujawinski, E.B., 2016. Evidence for quorum sensing and
356 differential metabolite production by a marine bacterium in response to DMSP. ISME Journal.

357 Kessner, D., Chambers, M., Burke, R., Agus, D., Mallick, P., 2008. ProteoWizard: open source
358 software for rapid proteomics tools development. Bioinformatics 24, 2534-2536.

359 Kharbush, J.J., Allen, A.E., Moustafa, A., Dorrestein, P.C., Aluwihare, L.I., 2016. Intact polar
360 diacylglycerol biomarker lipids isolated from suspended particulate organic matter
361 accumulating in an ultraoligotrophic water column. Organic Geochemistry 100, 29-41.

362 Kido Soule, M.C., Longnecker, K., Johnson, W.M., Kujawinski, E.B., 2015. Environmental
363 metabolomics: analytical strategies. Marine Chemistry 177, Part 2, 374-387.

364 Ksionzek, K.B., Lechtenfeld, O.J., McCallister, S.L., Schmitt-Kopplin, P., Geuer, J.K., Geibert, W.,
365 Koch, B.P., 2016. Dissolved organic sulfur in the ocean: Biogeochemistry of a petagram
366 inventory. Science 354, 456-459.

367 Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T.R., Neumann, S., 2012. CAMERA: an integrated
368 strategy for compound spectra extraction and annotation of liquid chromatography/mass
369 spectrometry data sets. Analytical Chemistry 84, 283-289.

370 Longnecker, K., Futrelle, J., Coburn, E., Kido Soule, M.C., Kujawinski, E.B., 2015a.
371 Environmental metabolomics: databases and tools for data analysis. Marine Chemistry 177, Part
372 2, 366-373.

373 Longnecker, K., Kido Soule, M.C., Kujawinski, E.B., 2015b. Dissolved organic matter produced
374 by *Thalassiosira pseudonana*. Marine Chemistry 168, 114-123.

375 Longnecker, K., Kujawinski, E.B., 2016. Using network analysis to discern compositional
376 patterns in ultrahigh resolution mass spectrometry data of dissolved organic matter. Rapid
377 Communications in Mass Spectrometry 30, 2388-2394.

378 Malmstrom, R.R., Kiene, R.P., Kirchman, D.L., 2004. Identification and enumeration of bacteria
379 assimilating dimethylsulfoniopropionate (DMSP) in the North Atlantic and Gulf of Mexico.
380 Limnology and Oceanography 49, 597-606.

381 Martin, P., Van Mooy, B.A.S., Heithoff, A., Dyhrman, S.T., 2011. Phosphorus supply drives
382 rapid turnover of membrane phospholipids in the diatom *Thalassiosira pseudonana*. ISME Journal
383 5, 1057-1060.

384 Moran, M.A., Kujawinski, E.B., Stubbins, A., Fatland, R., Aluwihare, L.I., Buchan, A., Crump,
385 B.C., Dorrestein, P.C., Dyhrman, S.T., Hess, N.J., Howe, B., Longnecker, K., Medeiros, P.M.,
386 Niggemann, J., Obernosterer, I., Repeta, D.J., Waldbauer, J.R., 2016. Deciphering ocean carbon in
387 a changing world. Proceedings of the National Academy of Sciences 113, 3143-3151.

388 Nguyen, D.D., Wu, C.-H., Moree, W.J., Lamsa, A., Medema, M.H., Zhao, X., Gavilan, R.G.,
389 Aparicio, M., Atencio, L., Jackson, C., Ballesteros, J., Sanchez, J., Watrous, J.D., Phelan, V.V., van
390 de Wiel, C., Kersten, R.D., Mehnaz, S., De Mot, R., Shank, E.A., Charusanti, P., Nagarajan, H.,
391 Duggan, B.M., Moore, B.S., Bandeira, N., Palsson, B.Ø., Pogliano, K., Gutiérrez, M., Dorrestein,
392 P.C., 2013. MS/MS networking guided analysis of molecule and gene cluster families.
393 Proceedings of the National Academy of Sciences 110, E2611-E2620.

394 Rabinowitz, J.D., Kimball, E., 2007. Acidic acetonitrile for cellular metabolome extraction from
395 *Escherichia coli*. Analytical Chemistry 79, 6167-6173.

396 Rampen, S.W., Abbas, B.A., Schouten, S., Damste, J.S.S., 2010. A comprehensive study of sterols
397 in marine diatoms (Bacillariophyta): Implications for their use as tracers for diatom
398 productivity. Limnology and Oceanography 55, 91-105.

399 Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R., Siuzdak, G., 2006. XCMS: processing mass
400 spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and
401 identification. Analytical Chemistry 78, 779 - 787.

402   Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.-L., Ideker, T., 2011. Cytoscape 2.8: new features
403   for data integration and network visualization. Bioinformatics 27, 431-432.

404   Sumner, L., Amberg, A., Barrett, D., Beale, M., Beger, R., Daykin, C., Fan, T., Fiehn, O.,
405   Goodacre, R., Griffin, J., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A.,
406   Lindon, J., Marriott, P., Nicholls, A., Reily, M., Thaden, J., Viant, M., 2007. Proposed minimum
407   reporting standards for chemical analysis. Metabolomics 3, 211-221.

408   Takahashi, Y., Itoh, K., Ishii, M., Suzuki, M., Itabashi, Y., 2002. Induction of larval settlement
409   and metamorphosis of the sea urchin Strongylocentrotus intermedius by glycoglycerolipids
410   from the green alga Ulvella lens. Marine Biology 140, 763-771.

411   Tebben, J., Motti, C.A., Siboni, N., Tapiolas, D.M., Negri, A.P., Schupp, P.J., Kitamura, M., Hatta,
412   M., Steinberg, P.D., Harder, T., 2015. Chemical mediation of coral larval settlement by crustose
413   coralline algae. Scientific Reports 5.

414   Teta, R., Della Sala, G., Glukhov, E., Gerwick, L., Gerwick, W.H., Mangoni, A., Costantino, V.,
415   2015. Combined LC–MS/MS and molecular networking approach reveals new cyanotoxins from
416   the 2014 cyanobacterial bloom in Green Lake, Seattle. Environmental Science & Technology 49,
417   14301-14310.

418   Treutler, H., Tsugawa, H., Porzel, A., Gorzolka, K., Tissier, A., Neumann, S., Balcke, G.U., 2016.
419   Discovering Regulated Metabolite Families in Untargeted Metabolomics Studies. Analytical
420   Chemistry 88, 8082-8090.

421   van der Hooft, J.J.J., Wandy, J., Barrett, M.P., Burgess, K.E.V., Rogers, S., 2016. Topic modeling
422   for untargeted substructure exploration in metabolomics. Proceedings of the National Academy
423   of Sciences 113, 13738-13743.

424   Van Mooy, B.A.S., Fredricks, H.F., Pedler, B.E., Dyhrman, S.T., Karl, D.M., Koblížek, M., Lomas,
425   M.W., Mincer, T.J., Moore, L.R., Moutin, T., Rappé, M.S., Webb, E.A., 2009. Phytoplankton in the
426   ocean use non-phosphorus lipids in response to phosphorus scarcity. Nature 458, 69-72.

427   Véron, B., Dauguet, J.-C., Billard, C., 1998. Sterolic biomarkers in marine phytoplankton. II. Free
428   and conjugated sterols of seven species used in mariculture. Journal of Phycology 34, 273-279.

429   Wang, M., Carver, J.J., Phelan, V.V., Sanchez, L.M., Garg, N., Peng, Y., Nguyen, D.D., Watrous,
430   J., Kapono, C.A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A.V., Meehan, M.J., Liu,
431   W.-T., Crusemann, M., Boudreau, P.D., Esquenazi, E., Sandoval-Calderon, M., Kersten, R.D.,
432   Pace, L.A., Quinn, R.A., Duncan, K.R., Hsu, C.-C., Floros, D.J., Gavilan, R.G., Kleigrewe, K.,
433   Northen, T., Dutton, R.J., Parrot, D., Carlson, E.E., Aigle, B., Michelsen, C.F., Jelsbak, L.,
434   Sohlenkamp, C., Pevzner, P., Edlund, A., McLean, J., Piel, J., Murphy, B.T., Gerwick, L., Liaw,
435   C.-C., Yang, Y.-L., Humpf, H.-U., Maansson, M., Keyzers, R.A., Sims, A.C., Johnson, A.R.,

436 Sidebottom, A.M., Sedio, B.E., Klitgaard, A., Larson, C.B., Boya P, C.A., Torres-Mendoza, D.,
437 Gonzalez, D.J., Silva, D.B., Marques, L.M., Demarque, D.P., Pociute, E., O'Neill, E.C., Briand, E.,
438 Helfrich, E.J.N., Granatosky, E.A., Glukhov, E., Ryffel, F., Houson, H., Mohimani, H., Kharbush,
439 J.J., Zeng, Y., Vorholt, J.A., Kurita, K.L., Charusanti, P., McPhail, K.L., Nielsen, K.F., Vuong, L.,
440 Elfeki, M., Traxler, M.F., Engene, N., Koyama, N., Vining, O.B., Baric, R., Silva, R.R., Mascuch,
441 S.J., Tomasi, S., Jenkins, S., Macherla, V., Hoffman, T., Agarwal, V., Williams, P.G., Dai, J.,
442 Neupane, R., Gurr, J., Rodriguez, A.M.C., Lamsa, A., Zhang, C., Dorrestein, K., Duggan, B.M.,
443 Almaliti, J., Allard, P.-M., Phapale, P., Nothias, L.-F., Alexandrov, T., Litaudon, M., Wolfender,
444 J.-L., Kyle, J.E., Metz, T.O., Peryea, T., Nguyen, D.-T., VanLeer, D., Shinn, P., Jadhav, A., Muller,
445 R., Waters, K.M., Shi, W., Liu, X., Zhang, L., Knight, R., Jensen, P.R., Palsson, B.O., Pogliano, K.,
446 Linington, R.G., Gutierrez, M., Lopes, N.P., Gerwick, W.H., Moore, B.S., Dorrestein, P.C.,
447 Bandeira, N., 2016. Sharing and community curation of mass spectrometry data with Global
448 Natural Products Social Molecular Networking. Nature Biotechnology 34, 828-837.

449 Watrous, J., Roach, P., Alexandrov, T., Heath, B.S., Yang, J.Y., Kersten, R.D., van der Voort, M.,
450 Pogliano, K., Gross, H., Raaijmakers, J.M., Moore, B.S., Laskin, J., Bandeira, N., Dorrestein, P.C.,
451 2012. Mass spectral molecular networking of living microbial colonies. Proceedings of the
452 National Academy of Sciences 109, E1743-E1752.

453 Wolf, S., Schmidt, S., Muller-Hannemann, M., Neumann, S., 2010. In silico fragmentation for
454 computer assisted identification of metabolite mass spectra. BMC Bioinformatics 11, 148.

455 Wolfersberger, M.G., Pieringer, R.A., 1974. Metabolism of sulfoquinovosyl diglyceride in
456 *Chlorella pyrenoidosa* by sulfoquinovosyl monoglyceride:fatty acyl CoA acyltransferase and
457 sulfoquinovosyl glyceride:fatty acyl ester hydrolase pathways. Journal of Lipid Research 15, 1-
458 10.

459 Yagi, T., Benson, A.A., 1962. Plant sulfolipid. 5. Lysosulfolipid formation Biochimica et
460 Biophysica Acta 57, 601-&.

461 Yang, J.Y., Sanchez, L.M., Rath, C.M., Liu, X., Boudreau, P.D., Bruns, N., Glukhov, E., Wodtke,
462 A., de Felicio, R., Fenner, A., Wong, W.R., Linington, R.G., Zhang, L., Debonsi, H.M., Gerwick,
463 W.H., Dorrestein, P.C., 2013. Molecular networking as a dereplication strategy. Journal of
464 Natural Products 76, 1686-1699.

465

466

467 Table 1. Details on the lyso-sulfolipids putatively annotated in intracellular metabolites from *T.*

468 *pseudonana*. All of the metabolites have the sulfoquinovosyl head group, and the table provides

469 the details on the non-polar tail (fatty acid or sterol), elemental formula, exact mass, measured

470 *m*/z, retention time (RT), and ChemSpider identification number for each metabolite. The set of

471 fragments used to putatively annotate these metabolites are given in Table S1.

Non-polar tail	Elemental formula	Expected charged mass ([M-H]$^-$)	Measured *m*/z	RT (min)	ChemSpider #
C14:0	$C_{23}H_{44}O_{11}S$	527.253161	527.253276	28.5	8134199
C16:0	$C_{25}H_{48}O_{11}S$	555.284472	555.284665	31.6	10481089
C16:1	$C_{25}H_{46}O_{11}S$	553.268786	553.268738	29.4	8113163
Sterol	$C_{27}H_{44}O_{11}S$	575.253161	575.253434	27.2	9672866

472

473

474

475   Table 2. Metadata associated with the additional samples containing the C14:0 lyso-sulfolipid. The table includes a brief description

476   of each set of samples, the number of samples with the C14:0 lyso-sulfolipid, and details on how the peak areas were normalized for

477   each set of samples. All of the filters and filtrates from these studies were processed using the methods described in Kido Soule et al.

478   (2015).

Label	Description	Geographic region	# of samples	Peak area normalized by:	Citation
Synechococcus elongatus	laboratory experiment	(not applicable)	n = 12 (filters) n = 10 (filtrate)	abundance (cells ml^{-1})	(Fiore et al., 2015)
Incubation: 70 m	experiment with seawater collected from 70 m	10º North, 55º W	n = 15 (filters) n = 15 (filtrate)	concentration of total organic carbon (μM)	(unpublished)
Incubation: 700 m	experiment with seawater collected from 700 m	0º North, 34º W	n = 6 (filters) n = 6 (filtrate)	concentration of total organic carbon (μM)	(unpublished)
Net traps	net traps deployed for 24 hours at 150 m	0º N, 34º W 6ºN, 41ºW 6ºN, 45ºW 7ºN, 48ºW 8ºN, 50ºW 10ºN, 55ºW	n = 6 (filters)	wet weight of filter (g)	(unpublished)

479

Figure legends

Figure 1. Schematic summarizing the analysis of the untargeted mass spectrometry data using a combination of molecular networking and MetFrag.

Figure 2. The output from molecular networking as visualized using the Cytoscape network visualization tool. Each node in the figure is an mzRT feature with MS2 fragmentation spectra, the color of the nodes indicates the experimental conditions under which the mzRT feature was found. The nodes are connected by edges and the thickness of the line is a measure of the similarity between each pair of nodes. Letters (A–M) are used to label each cluster.

Figure 3. MS2 fragmentation spectra from the four metabolites with the sulfoquinovosyl head group. The compounds differ in the non-polar tail with three of the compounds having a single fatty acid (A) C14:0, (B) C16:0, (C) C16:1, and (D) one compound with a 21-carbon sterol. The inset in (A) is sulfoquinovose, the head group for each lipid. The structures corresponding to each metabolite are given in Figure S2. The numbers in each subplot are the nominal masses for the top six MS2 fragments.

Figure 4. Three lyso-sulfo lipids with a single fatty acid were putatively annotated in the experiment with *T. pseudonana*. The lipids contained (A) a 14:0 fatty acid, (B) a 16:0 fatty acid, or (C) a 16:1 fatty acid and all of them showed higher cell-specific levels at the conclusion of the experiment when *T. pseudonana* was grown under phosphate-limited conditions.
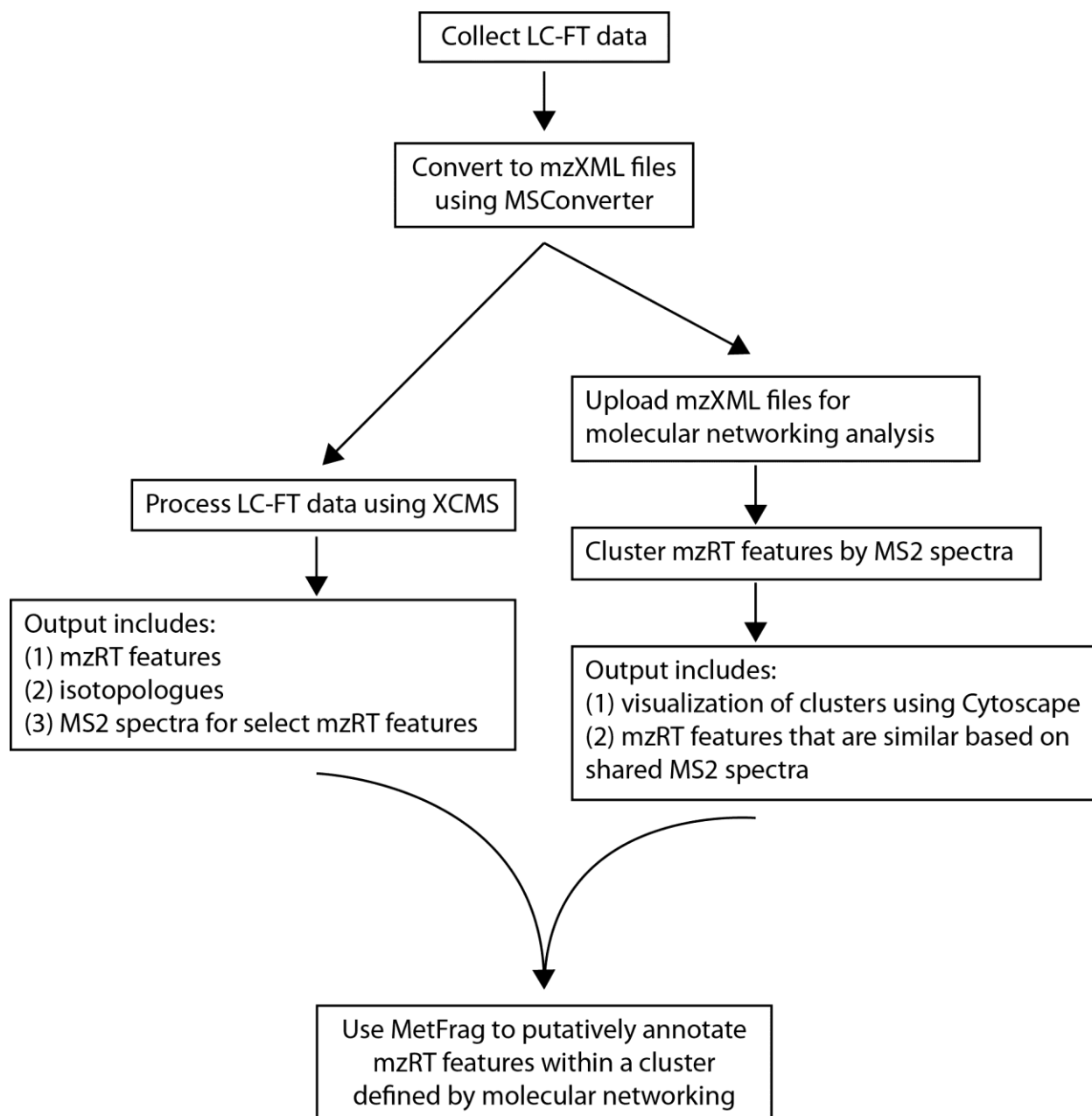
Figure 5. A sterol with the sulfoquinovosyl head group showed generally higher cell-specific levels under the phosphate-limited growth conditions. The box represents the middle 50% of

500   the data, or the inter-quartile range (IQR); whiskers extending above and below the box include

501   data within 1.5 IQRs of the box. The lines in boxes are median values.

502   Figure 6. The lyso-sulfo lipid with the C14:0 fatty acid was found in extracts from filters and

503   filtrates from four additional sets of samples processed within our laboratory: (A) a cultured

504   autotrophic microorganism, *S. elongatus* (Fiore et al., 2015), incubation experiments conducted

505   with seawater from (B) 70 m, and (C) 700 m, and (D) in particulate material captured by net

506   traps deployed at 150 m in the western equatorial Atlantic Ocean. The box represents the

507   middle 50% of the data, or the inter-quartile range (IQR); whiskers extending above and below

508   the box include data within 1.5 IQRs of the box; +: outliers, defined as normalized peak areas

509   between 1.5 and 3 IQRs distant from the box. The lines in boxes are median values.

510

511   Longnecker and Kujawinski
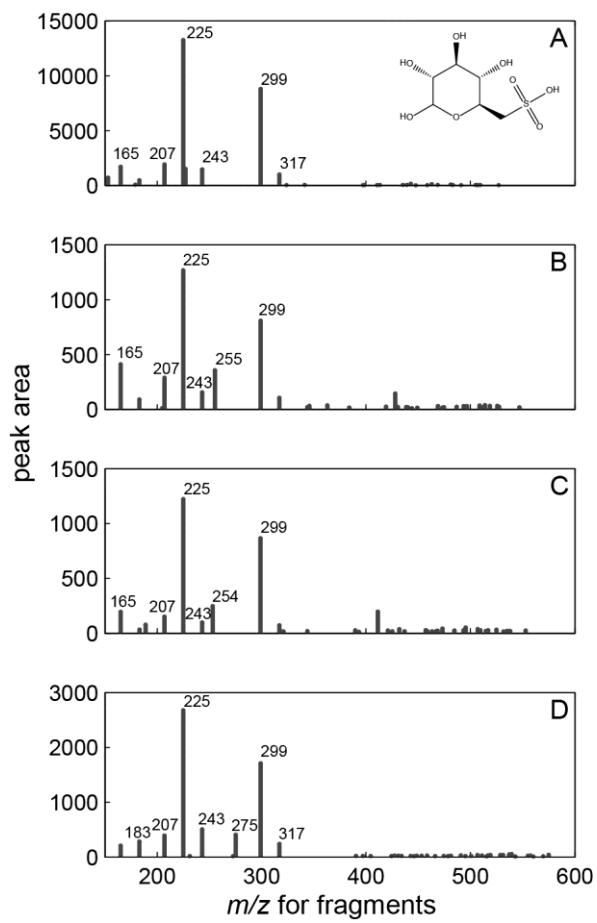
512   Figure 1

```
                    ┌─────────────────────┐
                    │  Collect LC-FT data │
                    └─────────────────────┘
                               │
                               ▼
                    ┌─────────────────────┐
                    │ Convert to mzXML files │
                    │  using MSConverter  │
                    └─────────────────────┘
                       ╱             ╲
                      ╱               ╲
                     ▼                 ▼
                          ┌──────────────────────────┐
                          │ Upload mzXML files for    │
                          │ molecular networking analysis │
                          └──────────────────────────┘
                                        │
                                        ▼
┌─────────────────────────────┐  ┌──────────────────────────┐
│ Process LC-FT data using XCMS │  │ Cluster mzRT features by MS2 spectra │
└─────────────────────────────┘  └──────────────────────────┘
              │                                 │
              ▼                                 ▼
┌─────────────────────────────┐  ┌──────────────────────────────────┐
│ Output includes:            │  │ Output includes:                  │
│ (1) mzRT features           │  │ (1) visualization of clusters using Cytoscape │
│ (2) isotopologues           │  │ (2) mzRT features that are similar based on │
│ (3) MS2 spectra for select mzRT features │  │ shared MS2 spectra        │
└─────────────────────────────┘  └──────────────────────────────────┘
                    ╲                      ╱
                     ╲                    ╱
                      ▼                  ▼
                    ┌──────────────────────────────┐
                    │ Use MetFrag to putatively annotate │
                    │ mzRT features within a cluster │
                    │ defined by molecular networking │
                    └──────────────────────────────┘
```

513

515   Figure 2
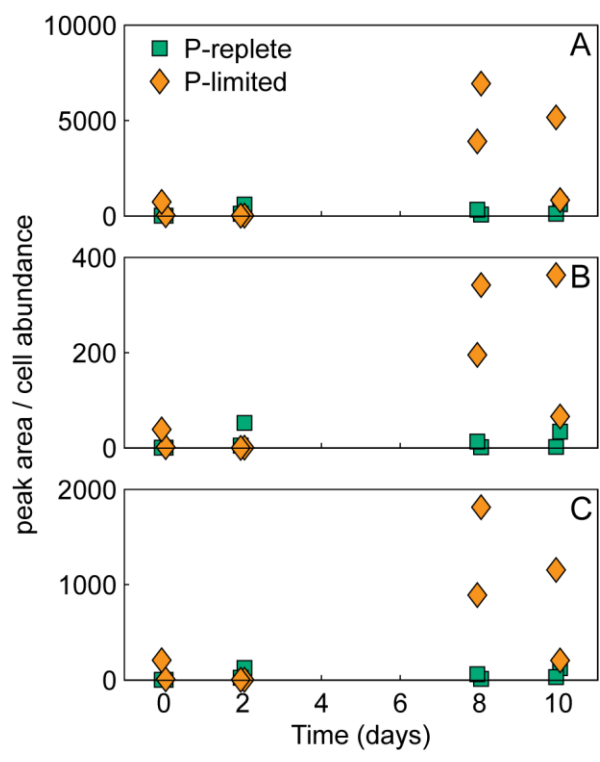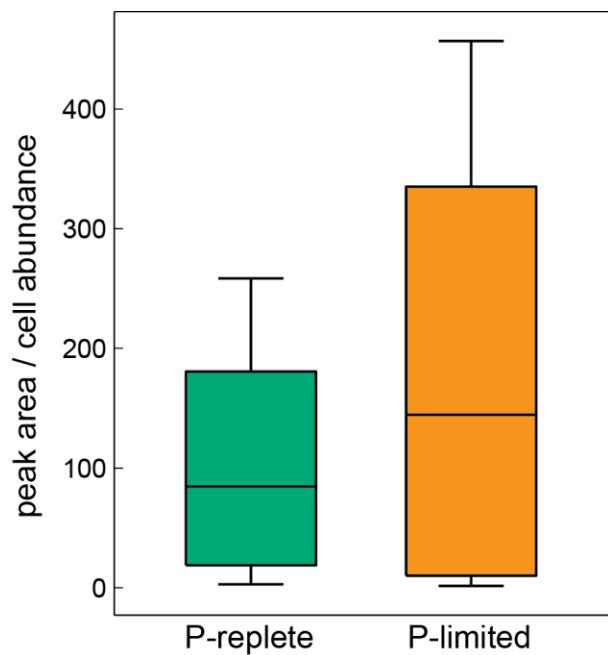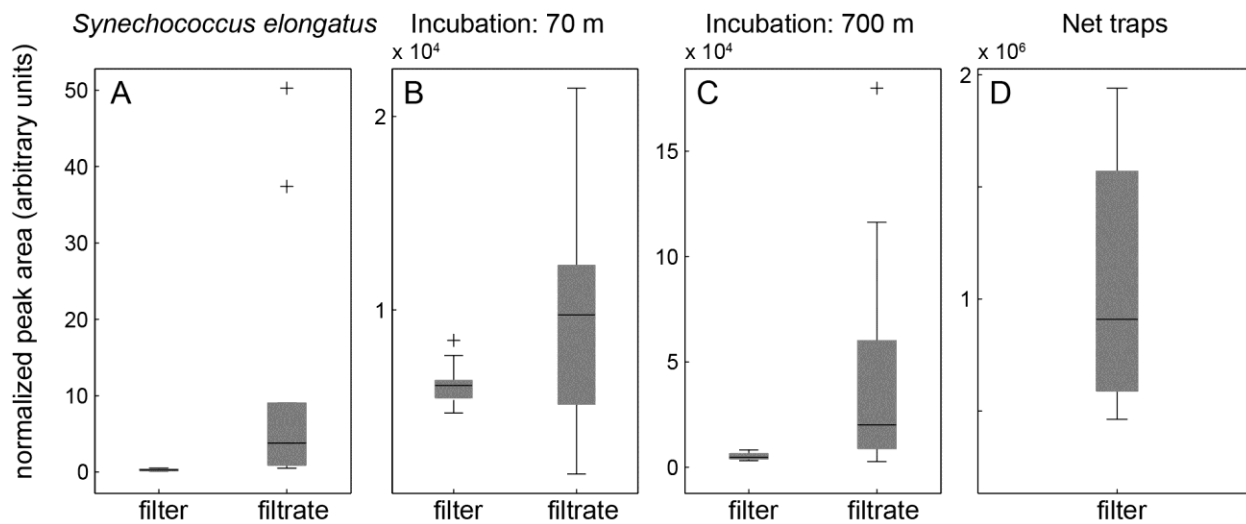
519   Figure 3

520

522   Figure 4

523

525   Figure 5

526

528   Figure 6

529