**RESEARCH PRODUCT SHARING PLAN – MSCOLLABORATORY.**

## 1 Data type

### 1.A Types and amount of scientific data expected to be generated in the project

This project will produce metabolomics data for the marine symbiosis community. The following data files will be used or produced in the course of the project: technical metadata for all experiment types (batch number, well location, date, etc.) in tabular format (.csv); mass spectrometry data and MS/MS data (.d, .raw, .mzml and .MGF format). We will further provide metabolite annotation data (that can be downloaded as.csv or other data table formats as needed - see below). Each data that is generated through the Moore Marine Metabolite Data Science Collaboratory set will be stored in GNPS/MassIVE and a DOI will be created that can be referenced together with the MassIVE accession number.

Metabolomics data is made available as .mzml or .MGF data formats. mzml and .MGF data are not proprietary, and are analyzable by multiple software. It may be that during the course of the grant improved open formats emerge (as the community is working on them to improve the ability to analyze larger and larger data sets) and in that case such open format will be adapted. Of note, the data could be analyzed within the GNPS/MassIVE infrastructure in which it is shared[1]. Common open-source tools for feature finding tools will be leveraged for feature based molecular networking within the GNPS ecosystem are OpenMS and MZmine 3.0 (refs. [2,3]. To ease access, we will share processed tabular data via hyperlinks, in GNPS/MassIVE, in addition to the raw metabolomics files. GNPS/MassIVE is designed to retain all analysis jobs, the analysis parameters, annotations and resulting tables with provenance of where this information comes from[1,4]. The data tables can be exported from GNPS into formats compatible for analysis in QIIME[5], MetaboAnalyst[6], Cytoscape[7], Jupyter notebooks, or R including in the Galaxy environment or similar third party tools. Should this project need new scripts or code, this will be documented in Github and archived in Zenodo or equivalent code repository.

### 1.B Metadata, other relevant data, and associated documentation

Sample covariates, data collection instruments, will be made accessible that is provided by the community that will use the collaboratory for their work. These can include organisms names, links to sequencing data, controls, environmental data such as oxygen measurements, salinity measurements, chlorophyll measurements, depth, longitude latitude coordinates, name of country, or location and other metadata the data generators have available. We will share full technical metadata (processing batches, well locations, negative and positive controls; reagent batches; date of processing; etc.) to enable full replication and accounting for technical artifacts (if necessary).

## 3 Standards

Metabolomics data will be shared in mzml and .MGF formats which are open, XML-based formats for mass spectrometer output, and are an accepted standard by both labs and vendors in the metabolomics field[8]. It may be that during the course of the grant improved open formats emerge (as the community is working on them to improve the ability to analyze larger and larger data sets) and in that case such open format will be adapted.

## 4 Data preservation, access, and associated timelines

### 4.A Repository where scientific data and metadata will be archived

All untargeted mass spectrometry data will be shared via GNPS/MassIVE - an NIH supported data repository for untargeted mass spectrometry data (proteomics and metabolomics) that is recommended by both the Nature Journal Publishing Family and the American Chemical Society. Should NMR data be created in this project it will be shared via the more generalist metabolomics repository Metabolomics workbench[9] or for pure molecules we will use NP-MRD[10], both are repositories supported by the NIH.

### 4.B How scientific data will be findable and identifiable

Metabolomics data will be finable through accession numbers and DOI's generated by GNPS/MassIVE, as well as discoverable with MASST[11], MassQL[12] and ReDU[13], search engines that are part of the GNPS ecosystem and are discoverable through the Omics Discovery Index[14]. Spectral annotations are findable in Massbank of North America (MONA) and are indexed in Pubchem.

**4.C When and for how long the scientific data will be made available**

Data will be preserved and made available through the public repositories (GNPS/MassIVE, Metabolomics Workbench and NP-MRD), for as long as the respective repositories persist after the end of the funding period.

**5 Access, distribution, or reuse considerations**

**5.A Factors affecting subsequent access, distribution, or reuse of scientific data**

Mass spectrometry data will be publicly available in GNPS/MassIVE without restrictions for academic use, the rest of the licensing will be defined by the sample creators of the Moore Marine Metabolite Data Science Collaboratory. Minimum license requirement for the use of the Moore Marine Metabolite Data Science Collaboratory will be for academic use but we encourage a fully open unrestricted license. The preferred licensing options will be dictated by the sample generators when initiating the project before the data is generated.

**5.B Whether access to scientific data will be controlled**

It is possible for the sample generators to request a delay in making the data public for publication purposes. Even in such circumstances, we will encourage the public deposition of the data but keep the metadata locked until publication.

**6 Oversight of data management and sharing**

Lead PI Dr. Pieter Dorrestein, ORCID: 0000-0002-3003-1030, will be responsible for management and overseeing the sharing of metabolomic data with the sample generators and the broader community. Broader issues of Data Management Plan compliance oversight and reporting will be handled by the PI and Co PI-team team as part of general stewardship, reporting, and compliance processes.

**7 Moore Foundation Specific Research Product Sharing Plan.**

**7.A Citing the grants DOI and other DOI's in research products.**

We will ensure that all research products, research products, such as publications, protocols, datasets, websites, and software, the DOI to the grant that the Moore Foundation will provide using the following language "This work was supported by the Gordon and Betty Moore Foundation, GBMF12120 and https://doi.org/10.37807/GBMF12120." In addition to your grant-specific DOI, each research product should have a separate DOI number generated by a relevant repository or platform (such as protocols.io or a peer-reviewed journal).

**7.B All researchers on this project will have to create and use ORCID.**

All researchers on your project should have a persistent digital identifier issued by ORCID.org. We aim for all grant relevant project research products to be included on team ORCID profiles. For new hires, we will ensure these individuals create ORCID accounts promptly after starting work on the project if they don't already have one.

**7.C. This product sharing plan will be shared with all scientists working on the project.**

The research product sharing plan should be reviewed with all scientists working on the grant, including the postdocs, graduate students, and technicians responsible for generating data and other products. We plan on also posting this document on the website for all collaboratory users to review as well.

1. Wang, M., Carver, J. J., Phelan, V. V., Sanchez, L. M., Garg, N., Peng, Y., Nguyen, D. D., Watrous, J., Kapono, C. A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A. V., Meehan, M. J., Liu, W.-T., Crüsemann, M., Boudreau, P. D., Esquenazi, E., Sandoval-Calderón, M., Kersten, R. D., Pace, L. A., Quinn, R. A., Duncan, K. R., Hsu, C.-C., Floros, D. J., Gavilan, R. G., Kleigrewe, K., Northen, T., Dutton,

R. J., Parrot, D., Carlson, E. E., Aigle, B., Michelsen, C. F., Jelsbak, L., Sohlenkamp, C., Pevzner, P., Edlund, A., McLean, J., Piel, J., Murphy, B. T., Gerwick, L., Liaw, C.-C., Yang, Y.-L., Humpf, H.-U., Maansson, M., Keyzers, R. A., Sims, A. C., Johnson, A. R., Sidebottom, A. M., Sedio, B. E., Klitgaard, A., Larson, C. B., P, C. A. B., Torres-Mendoza, D., Gonzalez, D. J., Silva, D. B., Marques, L. M., Demarque, D. P., Pociute, E., O'Neill, E. C., Briand, E., Helfrich, E. J. N., Granatosky, E. A., Glukhov, E., Ryffel, F., Houson, H., Mohimani, H., Kharbush, J. J., Zeng, Y., Vorholt, J. A., Kurita, K. L., Charusanti, P., McPhail, K. L., Nielsen, K. F., Vuong, L., Elfeki, M., Traxler, M. F., Engene, N., Koyama, N., Vining, O. B., Baric, R., Silva, R. R., Mascuch, S. J., Tomasi, S., Jenkins, S., Macherla, V., Hoffman, T., Agarwal, V., Williams, P. G., Dai, J., Neupane, R., Gurr, J., Rodríguez, A. M. C., Lamsa, A., Zhang, C., Dorrestein, K., Duggan, B. M., Almaliti, J., Allard, P.-M., Phapale, P., Nothias, L.-F., Alexandrov, T., Litaudon, M., Wolfender, J.-L., Kyle, J. E., Metz, T. O., Peryea, T., Nguyen, D.-T., VanLeer, D., Shinn, P., Jadhav, A., Müller, R., Waters, K. M., Shi, W., Liu, X., Zhang, L., Knight, R., Jensen, P. R., Palsson, B. O., Pogliano, K., Linington, R. G., Gutiérrez, M., Lopes, N. P., Gerwick, W. H., Moore, B. S., Dorrestein, P. C. & Bandeira, N. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34,** 828–837 (2016).

2. Nothias, L.-F., Petras, D., Schmid, R., Dührkop, K., Rainer, J., Sarvepalli, A., Protsyuk, I., Ernst, M., Tsugawa, H., Fleischauer, M., Aicheler, F., Aksenov, A. A., Alka, O., Allard, P.-M., Barsch, A., Cachet, X., Caraballo-Rodriguez, A. M., Da Silva, R. R., Dang, T., Garg, N., Gauglitz, J. M., Gurevich, A., Isaac, G., Jarmusch, A. K., Kameník, Z., Kang, K. B., Kessler, N., Koester, I., Korf, A., Le Gouellec, A., Ludwig, M., Martin H, C., McCall, L.-I., McSayles, J., Meyer, S. W., Mohimani, H., Morsy, M., Moyne, O., Neumann, S., Neuweger, H., Nguyen, N. H., Nothias-Esposito, M., Paolini, J., Phelan, V. V., Pluskal, T., Quinn, R. A., Rogers, S., Shrestha, B., Tripathi, A., van der Hooft, J. J. J., Vargas, F., Weldon, K. C., Witting, M., Yang, H., Zhang, Z., Zubeil, F., Kohlbacher, O., Böcker, S., Alexandrov, T., Bandeira, N., Wang, M. & Dorrestein, P. C. Feature-based molecular networking in the GNPS analysis environment. *Nat. Methods* **17,** 905–908 (2020).

3. Schmid, R., Heuckeroth, S., Korf, A., Smirnov, A., Myers, O., Dyrlund, T. S., Bushuiev, R., Murray, K. J., Hoffmann, N., Lu, M., Sarvepalli, A., Zhang, Z., Fleischauer, M., Dührkop, K., Wesner, M., Hoogstra, S. J., Rudt, E., Mokshyna, O., Brungs, C., Ponomarov, K., Mutabdžija, L., Damiani, T., Pudney, C. J., Earll, M.,

Helmer, P. O., Fallon, T. R., Schulze, T., Rivas-Ubach, A., Bilbao, A., Richter, H., Nothias, L.-F., Wang, M., Orešič, M., Weng, J.-K., Böcker, S., Jeibmann, A., Hayen, H., Karst, U., Dorrestein, P. C., Petras, D., Du, X. & Pluskal, T. Integrative analysis of multimodal mass spectrometry data in MZmine 3. *Nat. Biotechnol.* **41,** 447–449 (2023).

4. Aron, A. T., Gentry, E. C., McPhail, K. L., Nothias, L.-F., Nothias-Esposito, M., Bouslimani, A., Petras, D., Gauglitz, J. M., Sikora, N., Vargas, F., van der Hooft, J. J. J., Ernst, M., Kang, K. B., Aceves, C. M., Caraballo-Rodríguez, A. M., Koester, I., Weldon, K. C., Bertrand, S., Roullier, C., Sun, K., Tehan, R. M., Boya P, C. A., Christian, M. H., Gutiérrez, M., Ulloa, A. M., Tejeda Mora, J. A., Mojica-Flores, R., Lakey-Beitia, J., Vásquez-Chaves, V., Zhang, Y., Calderón, A. I., Tayler, N., Keyzers, R. A., Tugizimana, F., Ndlovu, N., Aksenov, A. A., Jarmusch, A. K., Schmid, R., Truman, A. W., Bandeira, N., Wang, M. & Dorrestein, P. C. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nat. Protoc.* **15,** 1954–1991 (2020).
5. Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., Cope, E. K., Da Silva, R., Diener, C., Dorrestein, P. C., Douglas, G. M., Durall, D. M., Duvallet, C., Edwardson, C. F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J. M., Gibbons, S. M., Gibson, D. L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G. A., Janssen, S., Jarmusch, A. K., Jiang, L., Kaehler, B. D., Kang, K. B., Keefe, C. R., Keim, P., Kelley, S. T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M. G. I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B. D., McDonald, D., McIver, L. J., Melnik, A. V., Metcalf, J. L., Morgan, S. C., Morton, J. T., Naimey, A. T., Navas-Molina, J. A., Nothias, L. F., Orchanian, S. B., Pearson, T., Peoples, S. L., Petras, D., Preuss, M. L., Pruesse, E., Rasmussen, L. B., Rivers, A., Robeson, M. S., 2nd, Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S. J., Spear, J. R., Swafford, A. D., Thompson, L. R., Torres, P. J., Trinh, P., Tripathi, A., Turnbaugh, P. J., Ul-Hasan, S., van der Hooft, J. J. J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K. C., Williamson, C. H. D., Willis, A. D., Xu, Z. Z., Zaneveld, J. R., Zhang, Y., Zhu, Q., Knight, R. & Caporaso, J. G. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat.*

*Biotechnol.* **37,** 852–857 (2019).

6. Pang, Z., Zhou, G., Ewald, J., Chang, L., Hacariz, O., Basu, N. & Xia, J. Using MetaboAnalyst 5.0 for LC-HRMS spectra processing, multi-omics integration and covariate adjustment of global metabolomics data. *Nat. Protoc.* **17,** 1735–1761 (2022).
7. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504 (2003).
8. Deutsch, E. W. File formats commonly used in mass spectrometry proteomics. *Mol. Cell. Proteomics* **11,** 1612–1621 (2012).
9. Sud, M., Fahy, E., Cotter, D., Azam, K., Vadivelu, I., Burant, C., Edison, A., Fiehn, O., Higashi, R., Nair, K. S., Sumner, S. & Subramaniam, S. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* **44,** D463–70 (2016).
10. Wishart, D. S., Sayeeda, Z., Budinski, Z., Guo, A., Lee, B. L., Berjanskii, M., Rout, M., Peters, H., Dizon, R., Mah, R., Torres-Calzada, C., Hiebert-Giesbrecht, M., Varshavi, D., Varshavi, D., Oler, E., Allen, D., Cao, X., Gautam, V., Maras, A., Poynton, E. F., Tavangar, P., Yang, V., van Santen, J. A., Ghosh, R., Sarma, S., Knutson, E., Sullivan, V., Jystad, A. M., Renslow, R., Sumner, L. W., Linington, R. G. & Cort, J. R. NP-MRD: the Natural Products Magnetic Resonance Database. *Nucleic Acids Res.* **50,** D665–D677 (2022).
11. Wang, M., Jarmusch, A. K., Vargas, F., Aksenov, A. A., Gauglitz, J. M., Weldon, K., Petras, D., da Silva, R., Quinn, R., Melnik, A. V., van der Hooft, J. J. J., Caraballo-Rodríguez, A. M., Nothias, L. F., Aceves, C. M., Panitchpakdi, M., Brown, E., Di Ottavio, F., Sikora, N., Elijah, E. O., Labarta-Bajo, L., Gentry, E. C., Shalapour, S., Kyle, K. E., Puckett, S. P., Watrous, J. D., Carpenter, C. S., Bouslimani, A., Ernst, M., Swafford, A. D., Zúñiga, E. I., Balunas, M. J., Klassen, J. L., Loomba, R., Knight, R., Bandeira, N. & Dorrestein, P. C. Mass spectrometry searches using MASST. *Nat. Biotechnol.* **38,** 23–26 (2020).
12. Jarmusch, A. K., Aron, A. T., Petras, D., Phelan, V. V., Bittremieux, W., Acharya, D. D., Ahmed, M. M. A., Bauermeister, A., Bertin, M. J., Boudreau, P. D., Borges, R. M., Bowen, B. P., Brown, C. J., Chagas, F. O., Clevenger, K. D., Correia, M. S. P., Crandall, W. J., Crüsemann, M., Damiani, T., Fiehn, O., Garg, N.,

Gerwick, W. H., Gilbert, J. R., Globisch, D., Gomes, P. W. P., Heuckeroth, S., Andrew James, C., Jarmusch, S. A., Kakhkhorov, S. A., Kang, K. B., Kersten, R. D., Kim, H., Kirk, R. D., Kohlbacher, O., Kontou, E. E., Liu, K., Lizama-Chamu, I., Luu, G. T., Knaan, T. L., Marty, M. T., McAvoy, A. C., McCall, L.-I., Mohamed, O. G., Nahor, O., Niedermeyer, T. H. J., Northen, T. R., Overdahl, K. E., Pluskal, T., Rainer, J., Reher, R., Rodriguez, E., Sachsenberg, T. T., Sanchez, L. M., Schmid, R., Stevens, C., Tian, Z., Tripathi, A., Tsugawa, H., Nishida, K., Matsuzawa, Y., van der Hooft, J. J. J., Vicini, A., Walter, A., Weber, T., Xiong, Q., Xu, T., Zhao, H. N., Dorrestein, P. C. & Wang, M. A Universal Language for Finding Mass Spectrometry Data Patterns. *bioRxiv* 2022.08.06.503000 (2022). doi:10.1101/2022.08.06.503000

13. Jarmusch, A. K., Wang, M., Aceves, C. M., Advani, R. S., Aguirre, S., Aksenov, A. A., Aleti, G., Aron, A. T., Bauermeister, A., Bolleddu, S., Bouslimani, A., Caraballo Rodriguez, A. M., Chaar, R., Coras, R., Elijah, E. O., Ernst, M., Gauglitz, J. M., Gentry, E. C., Husband, M., Jarmusch, S. A., Jones, K. L., 2nd, Kamenik, Z., Le Gouellec, A., Lu, A., McCall, L.-I., McPhail, K. L., Meehan, M. J., Melnik, A. V., Menezes, R. C., Montoya Giraldo, Y. A., Nguyen, N. H., Nothias, L. F., Nothias-Esposito, M., Panitchpakdi, M., Petras, D., Quinn, R. A., Sikora, N., van der Hooft, J. J. J., Vargas, F., Vrbanac, A., Weldon, K. C., Knight, R., Bandeira, N. & Dorrestein, P. C. ReDU: a framework to find and reanalyze public mass spectrometry data. *Nat. Methods* **17,** 901–904 (2020).
14. Perez-Riverol, Y., Bai, M., da Veiga Leprevost, F., Squizzato, S., Park, Y. M., Haug, K., Carroll, A. J., Spalding, D., Paschall, J., Wang, M., Del-Toro, N., Ternent, T., Zhang, P., Buso, N., Bandeira, N., Deutsch, E. W., Campbell, D. S., Beavis, R. C., Salek, R. M., Sarkans, U., Petryszak, R., Keays, M., Fahy, E., Sud, M., Subramaniam, S., Barbera, A., Jiménez, R. C., Nesvizhskii, A. I., Sansone, S.-A., Steinbeck, C., Lopez, R., Vizcaíno, J. A., Ping, P. & Hermjakob, H. Discovering and linking public omics data sets using the Omics Discovery Index. *Nat. Biotechnol.* **35,** 406–409 (2017).