Title

Data and scripts associated with “Riverine dissolved organic matter transformations increase with watershed area, water residence time, and Damköhler numbers in nested watersheds” (v2)

Summary

This data package is associated with the publication “Riverine dissolved organic matter transformations increase with watershed area, water residence time, and Damköhler numbers in nested watersheds” submitted to Biogeochemistry by Ryan et al., 2024 (DOI: https://doi.org/10.1007/s10533-024-01169-5).

This study aims to investigate fundamental and transferable drivers of dissolved organic matter (DOM) diversity across five nested watersheds within the contiguous United States. DOM diversity was explored using ultrahigh-resolution Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS). The samples and the unprocessed FTICR-MS data used in this study are publicly available on the Environmental System Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) data repository (see DOIs below). The data for the Willamette, Gunnison, Connecticut, and Deschutes basins were collected as part of a collaboration between the Watershed Rules of Life (WROL) project and Worldwide Hydrobiogeochemistry Observation Network for Dynamic River Systems (WHONDRS). The data for the Yakima River basin (YRB) was collected by the PNNL River Corridor SFA. The raw, unprocessed FTICR-MS data with additional (meta)data can be found at https://data.ess-dive.lbl.gov/datasets/doi:10.15485/1895159 for WROL samples and https://data.ess-dive.lbl.gov/datasets/doi:10.15485/1898912 for YRB samples. This data package contains the processed data used in the associated manuscript. This package also contains ancillary geospatial, hydrological, and geochemical information that supports the interpretation of the FTICR-MS data within Ryan et al., 2024.

This data package is associated with the GitHub repository found at https://github.com/WHONDRS-Hub/rcsfa-RC4-WROL-YRB_DOM_Diversity.

This data package was originally published August 2024. It was updated January 2025 (modified files). See the change history section below for more details.

Brief Overview of Methods

For a full description of the methods, see the methods section in Ryan et al., 2024. Briefly, water samples were collected at 52 sites within five nested watersheds located in the contiguous United States. Samples were analyzed for dissolved organic carbon (DOC) and dissolved organic matter (DOM) chemistry via ultrahigh-resolution Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS). DOM data were used to infer putative biochemical transformations following methods previously published in Garayburu-Caruso et al., 2020. Patterns in DOM molecular diversity and putative biochemical transformations were assessed across gradients of explanatory variables associated with watershed characteristics (e.g., watershed area, water residence time, land cover) to investigate fundamental and transferable drivers of DOM diversity across watersheds.

Critical Details

The following steps were followed to generate the processed FTICR-MS data:

1 – The raw, unprocessed FTICR-MS data (XML files) were downloaded for WROL (https://data.ess-dive.lbl.gov/datasets/doi:10.15485/1895159) and YRB (https://data.ess-dive.lbl.gov/datasets/doi:10.15485/1898912) samples from ESS-DIVE.

2 – The instructions in the “FTICR_Instructions-Report_Generation_SOP_v3.pdf” document found in the original data packages were followed. Molecular formulae were assigned using Formultitude software. The data were further processed, and sample molecular properties were assigned using the R package “ftmsRanalysis”.

- The outputs from these steps were put into a folder called " Formultitude_Output_Folder". This folder is not found within this data package but is referenced in “Removing_poorly_calibrated_and_merge_reps.R" (see next step). If the user wants to create the processed data files with all replicates and before poorly calibrated samples were removed, they will need to follow these steps to create the processed data within a folder called " Formultitude_Output_Folder".

3 – Poorly calibrated samples were removed from the dataset and sample replicates from each site were merged such that a peak was kept in the merged sample if it was present in at least one of the reps. This step is computed within the “Removing_poorly_calibrated_and_merge_reps.R” script.

4 – Putative biochemical transformations were inferred following protocols previously described in Garayburu-Caruso et al. (2020). Briefly, pairwise mass differences were calculated between every peak (with and without molecular formula assigned) in a merged sample and compared to a reference list of reference transformations (“Transformation_Database_07-2020.csv”). Mass differences were matched to the compounds in the reference list (within 1 ppm) to infer the gain or loss of that compound via a biochemical transformation.

5 – Total number of transformations per sample and total number of transformations normalized per number of peaks in a sample were calculated.

Data Package Structure

At the directory level, the data package is comprised of three folders: (1) data, (2) output, and (3) src; and five additional files including the data dictionary (file ending in "_dd.csv”) and file-level metadata (file ending in “_flmd.csv”). The “src” folder contains the scripts used to process the FTICR data, conduct the analyses, and produce the manuscript figures. The inputs for these scripts are in the “data” folder and the returned outputs in the “output” folder. Inputs include temporal and spatial metadata associated with the sampling efforts, processed FTICR data, and total and normalized putative biochemical transformations per sample. Outputs include cleaned and combined data presented as tables, descriptive statistics, and plots. The file-level metadata file lists all files contained in this data package and descriptions for each. The data dictionary describes the units and definitions for each tabular data column or row header.

Citations and Acknowledgements

A portion of this research was supported in part by the U.S. Department of Energy (DOE) Biological and Environmental Research (BER) Environmental System Science (ESS) program (https://ess.science.energy.gov/) through the Pacific Northwest National Laboratory River Corridor Science Focus Area (SFA). PNNL is operated by Battelle Memorial Institute for the U.S. Department of Energy under Contract No. DE-AC05-76RL01830. Support was provided to P.A. Raymond, T. Bambakidis and B.C. Crump by National Science Foundation award DEB-1840243. Support was provided to S. Liu by National Natural Science Foundation of China (52379057). We thank the United States Forest Service, Washington Department of Natural Resources, and Washington Department of Fish and Wildlife for access to field locations where these samples were collected.

Cite this data package with the appropriate DOI. Cite the associated manuscript in any work that that uses analyses or conclusions presented in the manuscript:

Ryan, K.A., Garayburu-Caruso, V.A., Crump, B.C. et al. Riverine dissolved organic matter transformations increase with watershed area, water residence time, and Damköhler numbers in nested watersheds. Biogeochemistry (2024). https://doi.org/10.1007/s10533-024-01169-5

Please acknowledge the Worldwide Hydrobiogeochemistry Observation Network for Dynamic River Systems (WHONDRS, https://whondrs.pnnl.gov) and the U.S. Department of Energy (DOE) Biological and Environmental Research (BER) Environmental System Science (ESS) program (https://ess.science.energy.gov/) — which generously provides funding to WHONDRS — in your papers, presentations, proposals, etc. If using FTICR-MS data, please also acknowledge the Environmental Molecular Sciences Laboratory (EMSL; grid.436923.9). There is no obligation to include WHONDRS members as co-authors.

Citations:

- Garayburu-Caruso, V.A.; Danczak, R.E.; Stegen, J.C.; Renteria, L.; Mccall, M.; Goldman, A.E.; Chu, R.K.; Toyoda, J.; Resch, C.T.; Torgeson, J.M.; et al. Using Community Science to Reveal the Global Chemogeography of River Metabolomes. *Metabolites* 2020, *10*, 518. https://doi.org/10.3390/metabo10120518
- Kathryn Willi, & Matthew R. V. Ross. (2023). Geospatial Data Puller for Waters in the Contiguous United States (Version v1). Zenodo. https://doi.org/10.5281/zenodo.8140272
- Otenburg O ; Barnes M ; Borton M A ; Chen X ; Chu R ; Farris Y ; Forbes B ; Fulton S G ; Garayburu-Caruso V A ; Goldman A E ; Gonzalez B I ; Grieger S ; Kaufman M H ; McKever S A ; Myers-Pigg A ; Pelly A ; Renteria L ; Scheibe T D ; Son K ; Torgeson J M ; Toyoda J G ; Stegen J C (2022): Temporal Study 2021-2022: Sample-Based Surface Water Chemistry and Organic Matter Characterization across Watersheds in the Yakima River Basin, Washington, USA (v2). River Corridor and Watershed Biogeochemistry SFA, ESS-DIVE repository. Dataset. doi:10.15485/1898912 accessed via https://data.ess-dive.lbl.gov/datasets/doi:10.15485/1898912
- Pennino, M. J., Leibowitz, S. G., Compton, J. E., Hill, R. A., & Sabo, R. D. (2020). Patterns and predictions of drinking water nitrate violations across the conterminous United States. Science of the Total Environment, 722, 137661.
- Torgeson J M ; Bambakidis T ; Bates T L ; Chu R ; Crump B C ; Danczak R E ; Forbes B ; Garayburu-Caruso V A ; Goldman A E ; Logozzo L ; Maavara T ; Martin E W ; McKever S A ; Powers-McCormack B ; Raymond P A ; Renteria L ; Toyoda J G ; Stegen J C (2022): WHONDRS Surface Water Dissolved Organic Carbon and FTICR-MS across Stream Orders in Four United States Watersheds in 2019 and 2020 (v3). River Corridor and Watershed Biogeochemistry SFA, ESS-DIVE repository. Dataset. doi:10.15485/1895159 accessed via https://data.ess-dive.lbl.gov/datasets/doi:10.15485/1895159

Contact

Kevin A. Ryan, karyan@usgs.gov
Vanessa Garayburu-Caruso, vanessa.garayburu-caruso@pnnl.gov

Change History

Approach to change history and versioning:

Updates to **data package** version: When any file within a data package is updated, the data package version number is updated. The data package version number is indicated in the title of the data package, the data package folder name, and in the change history table below. You can access previous versions of the data package by sending a request to ESS-DIVE.

The change history below describes each file revised during versioning. If you are interested in seeing the exact cells within a file that have changed, you can utilize the daff package in R (https://github.com/edwindj/daff) to compare a previously downloaded file to a newly downloaded file.

In the change history table below, the sub-headers and bullets indicate the type of change in each file:

- New files: Describes new files added that were not present in previous data package versions
- Bulk changes to files: Describes a change to many files within the data package. The indicated superscript with be added to each file name that the change applies to.
- Modified files:
    - Corrected: Describes existing information modified or removed to prevent sharing of incorrect information
    - Added: Describes new information inserted into an existing file (e.g., appending new columns/rows)
    - Updated: Describes modifying existing information to maintain accuracy though version changes. (e.g., changing version number to new version number)

Data Package Version	Changes
Version 1 *August 2024*	Original data package publication
Version 2 *January 2025*	**MODIFIED FILES** **RC2_NPOC_TN_DIC_TSS_Ions_Summary_2021-2022.csv** (v2) • Replaced negative values in the "Mean_00530_TSS_mg_per_L" column with "-9999". **readme_Ryan_2024_WROL_YRB_DOM_Diversity.pdf** (v2) • Added new version number in data package title. • Added new information to data package summary to describe the new changes that were made. • Added new information to change history table. **Ryan_2024_WROL_YRB_DOM_Diversity_flmd.csv** (v2) • Updated rows to reflect changes of "modified" files and folders. • Updated standard column to reflect changes to ESS-DIVE's preferred controlled vocabulary.