Supporting Information:

Molecular Composition and Biodegradability of Soil Organic Matter: A Case Study Comparing Two New England Forest Types

Tsutomu Ohno[1]*, Thomas B. Parr[2], Marie–Cécile I. Gruselle[3], Ivan J. Fernandez[3], Rachel L. Sleighter[4] and Patrick G. Hatcher[4]

[1]School of Food and Agriculture, University of Maine, Orono, Maine 04469-5722, USA

[2]School of Biology and Ecology, University of Maine, Orono, Maine 04469-5751, USA

[3]School of Forest Resources, University of Maine, Orono, Maine 04469-5722, USA

[4]Department of Chemistry and Biochemistry, Old Dominion University, Norfolk, Virginia 23529, USA

*Corresponding author: e-mail: ohno@maine.edu; phone 207-581-2975; fax 207-581-2999

Experimental Section, 1 Figure, Reference Page, 7 Pages total.

Experimental Section

Field Site and Sample Analysis. The overarching experimental design is a four-compartment model resulting from each watershed being composed of ~95 year-old softwood stands dominated by red spruce (*Picea rubens* Sarg.) in the higher elevation range and of 65-70 year-old mixed hardwood stands with American beech *(Fagus grandifolia* Ehrh.), yellow birch (*Betula alleghaniensis* Britton), red maple (*Acer rubrum* L.), and sugar maple (*Acer saccharum* Marshall) in the lower elevations. Elevation ranges between 210 and 475 m. Sparse red spruce and balsam fir (*Abies balsamea* (L.) Mill.) seedlings make up the softwoods understory vegetation. Beneath the hardwood, the understory consists of seedlings of the named tree species and striped maple (*Acer pensylvanicum* L.), hay-scented fern (*Dennstaedtia punctilobula* (Michx.) T. Moore), Solomon's seal (*Polygonatum pubescens* (Willd.) Pursh), hobble bush (*Viburnum alnifolium* Michx.), and wild sarsaparilla (*Aralia nudicaulis* L.). The forest floor is three times as thick in conifer stands (mor-like type) than in hardwood stands (moder-like type). Soils are coarse-loamy, isotic, frigid, Typic Haplorthods formed from till (~1 m thick) primarily over a quartzite and gneiss bedrock. Soils were sampled in the reference watershed (East Bear, 11.0 ha) during August 2012 from a single soil pedon (~70 x 70 cm wide) in each of the deciduous and coniferous forest types. One sample (~500 g B horizon, <500 g E horizon) from the O_a, E, and B horizons per pedon were included in this study. They were manually sampled with a soil knife from one side of the pedon to obtain a grab sample with proportional representation throughout the horizon. B horizons were sampled in 5 cm increments from the top of the B horizon, where there is an abrupt boundary with the E horizon, to 25 cm depth followed by a 25-50 cm depth increment sample. The samples were placed in plastic bags and put in coolers during the sampling. In the laboratory, they were refrigerated until further processing. All samples were sieved field moist through a 2-mm sieve for mineral horizons and 6-mm sieve for the organic horizons, within 48 hours of collection in the field. Sieved soils were stored in paper bags and air-dried at room temperature.

FT-ICR-MS. The molecular formula calculator developed at the National High Magnetic Field Laboratory in Tallahassee, FL, (Molecular Formula Calc v.1.0 ©NHMFL, 1998) was used to generate empirical formula matches for the resolved peaks using combinations of C (8-50 atoms), H (8-100 atoms), O (1-30 atoms), N (0-5 atoms), S (0-3 atoms), and P (0-2 atoms) as the limiting atomic values. Only m/z values with a signal to noise above 5 were imported into the molecular formula calculator. The formula calculations were done sequentially by using the peak list obtained from the instrument and initially assigning formulas for CHO containing components. The peaks that were assigned as CHO components were then removed from the initial peak list, and the resulting reduced peak list was used for calculating formulas containing CHON. The removal of assigned peaks and next sequential assignment calculations were conducted for CHONS (which includes CHOS and CHONS formulas) and then CHONSP (which includes CHOP, CHONP, CHOSP, and CHONSP formulas). The removal of peaks that have been assigned ensures robust formula assignment, since the calculations done were constrained to only the selected elements present in the formula.

The output from the molecular formula calculator includes the peak abundance, instrument determined m/z, the number of pre-selected elements, and the ppm deviation of the proposed formula from the measured m/z. The resulting formula list was passed through a MATLAB script to constrain the formulas to chemically feasible organic matter molecules using the following criteria: $O/C < 1.2$, $H/C < 2.25$, $H/C > 0.3$, $N/C < 0.5$, $S/C < 0.2$, $P/C < 0.1$, (S+P)/C < 0.2, and double bond equivalents (DBE) ≥ 0 and must be a whole number.[1] The ^{13}C containing isotopic peaks that would appear exactly 1.0034 m/z units higher and <50% relative peak abundance than their ^{12}C containing peak were removed from the peak list, since they give redundant molecular information. Molecular formula calculator programs are typically set to ± 1 ppm mass error windows in the assignment of molecular formulas, which generally leads to unique formula assignments to peaks with m/z < 500, but peaks above this m/z value may have multiple potential formula assignments.

A MATLAB script was used to select the most appropriate molecular formula by parsing the proposed formulas into two arrays, those with unique formula assignments and those with two or more potential formula assignments.[2] The hierarchy for determining the correct assignment was selected based on: 1) KMD analysis, 2) least number of non-oxygen heteroatoms, and 3) lowest ppm m/z deviation. The script calculated additional descriptors for each formula: O/C ratio, H/C ratio, DBE, and DBE/C. The script also parsed the assigned peaks into the appropriate van Krevelen space which consisted of 7 discrete regions[3]: lipids (H:C, 1.5-2.0; O:C, 0-0.3); proteins (H:C, 1.5-2.2, O:C, 0.3-0.67); lignins (H:C, 0.7-1.5, O:C, 0.1-0.67); carbohydrates (H:C, 1.5-2.4, O:C, 0.67-1.2); unsaturated hydrocarbons (H:C, 0.7-1.5, O:C, 0-0.1); condensed aromatics (H:C, 0.2-0.7, O:C, 0-0.67); and tannins (H:C, 0.7-1.5, O:C, 0.67-1.2). Formulas that did not align in any of these pre-defined regions were removed from further analysis.

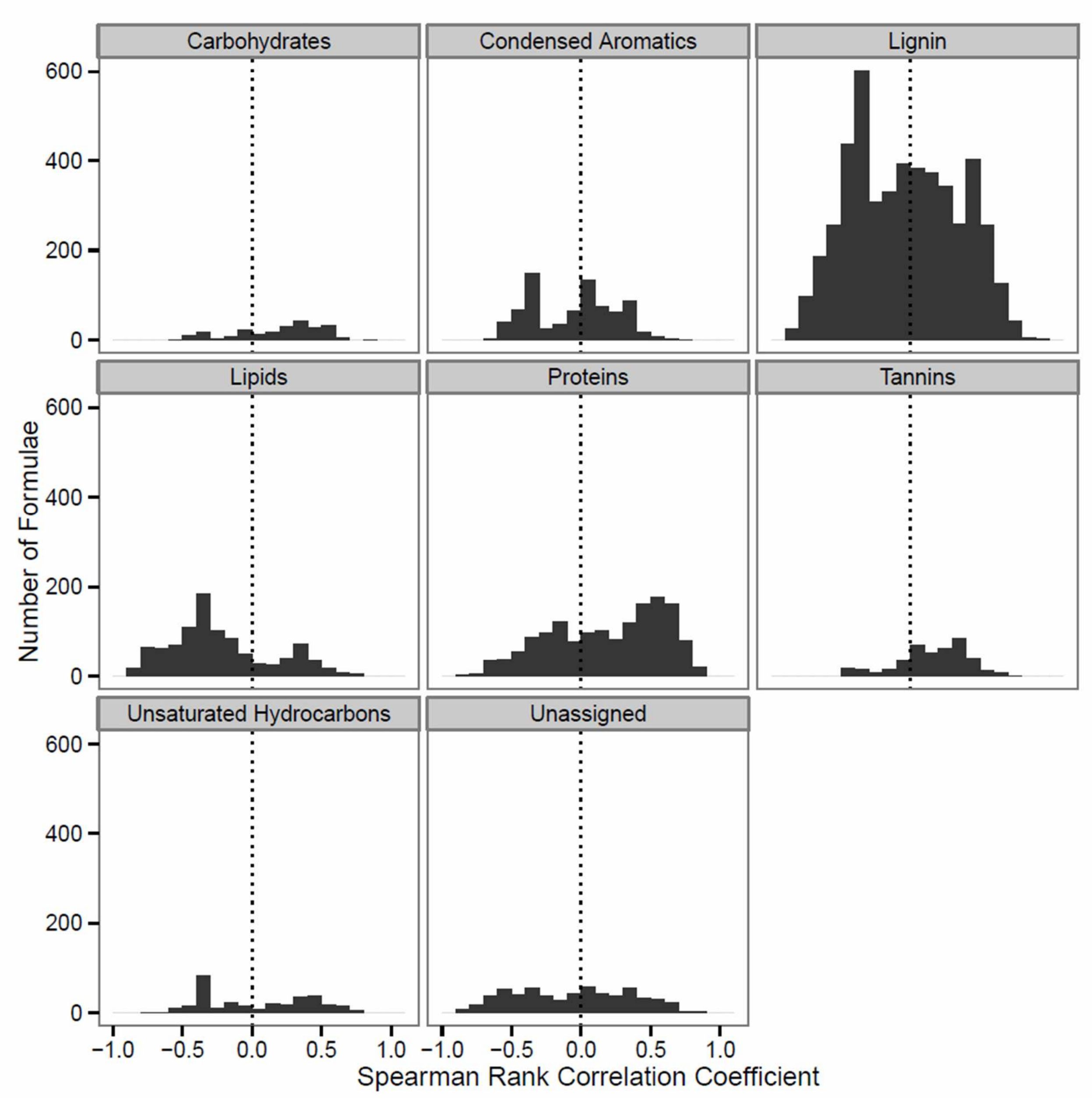
Principal Coordinate Analysis and Redundancy Analysis. First, all data were merged into a matrix by their formula assignment. Absence of a molecular formula in a sample was recorded as a 0. Thus, the SOM composition was described by a mass spectrum of ~7,500 formulas, ~4,000 of which are > 0. Peak abundances were normalized to the summed total abundance for each sample. Next, capscale() was used to perform principal coordinate analysis

(PCoA) with the relative peak abundance as the dependent variable or "species" and biodegradable dissolved organic carbon (BDOC) as the independent environmental variable in redundancy analysis (RDA). A Bray-Curtis dissimilarity index was used to describe the dissimilarity in composition among the samples. A Bray-Curtis distance measure was used, because adding zero-values for unobserved peaks introduces the "double-zero problem".[42] This is a frequently observed problem in ecology where, depending on the index used, rare species observed at a few sites increase the number of zero values observed at other sites, resulting in an overestimation of similarity. Euclidean distance is one index that will overestimate similarity in the presence of high zero-abundance observation. The Bray-Curtis index, conversely, does not use "double-zero" abundance observations. Thus, similarity is only calculated based on observations where at least one abundance is > 0 and this is defined as:

$$b_{ii'} = \sum_{j=1}^{J} \frac{\left|n_{ij} - n_{i'j}\right|}{n_i + n_{i'}}$$

where $b_{ii'}$ is the calculated Bray-Curtis dissimilarity, j is the molecule of interest, n_{ij} is the abundance of the jth molecule at the first site, and $n_{i'j}$ is the abundance of the jth molecule at the second site. Similarly, n_i and $n_{i'}$ are the summed total peak abundance at each site, and because we first normalized each site to the sum of peaks for each site, the sum of $n_i + n_{i'}$ is always 2. Finally, Spearman rank correlations were used to relate the relative abundances of individual formulas to BDOC.

Figure S1: Histogram of the Spearman rank correlation coefficients binned in 0.1 units for each of the van Krevelen diagram classification groups.



References

1. Stubbins, A.; Spencer, R. G. M.; Chen, H.; Hatcher, P. G.; Mopper, K.; Hernes, P. J.; Mwamba, V. L.; Mangangu, A. M.; Wabakanghanzi, J. N.; Six, J. Illuminated darkness: molecular signatures of Congo River dissolved organic matter and its photochemical alteration as revealed by ultrahigh precision mass spectrometry. *Limnol. Oceanogr.* **2010**, *55*, 1467-1477.
2. Ohno, T.; Ohno, P. E. Influence of heteroatom pre-selection on the molecular formula assignment of soil organic matter components determined by ultrahigh resolution mass spectrometry. *Anal. Bioanal. Chem.* **2013***, 405*, 3299-3306.
3. Hockaday, W. C.; Purcell, J. M.; Marshall, A. G.; Baldock, J. A.; Hatcher, P. G. Electrospray and photoionization mass spectrometry for the characterization of organic matter in natural waters: a qualitative assessment. *Limnol. Oceanogr.: Methods* **2009**, *7*, 81-95.
4. Legrendre, P.; Legrendre, L. *Numerical ecology, 2nd ed.*; Elsevier Science: New York, 1998.