EMERGE Sample Metadata Sheet for Samples with Microbiomes: Variable definitions and other FAQs - Public version

This is the FAQ file for the Sample Metadata Sheet (filename beginning with "coring_geochem_sequenced_samples").

About the Sample Metadata Sheet

The EMERGE Sample Metadata Sheet is a data compilation published by the EMERGE Biology Integration Institute, providing basic (meta)data (including field sample info and biogeochemistry data) for samples with linked metagenomes (metaGs). It was generated by parsing and cleaning up the "Depth-Info" and "Biogeochemistry" query output from the EMERGE Database (EMERGE-DB; https://emerge-db.asc.ohio-state.edu/), using R scripts in the EMERGE-BII sample metadata_sheet repository on GitHub (available to EMERGE project members at https://github.com/emerge-bii/sample_metadata_sheet; this internally-shared repository also includes another version of this FAQ for project members, with additional information on data sharing guidelines and related resources).

This sheet is a compilation of the following data sources (listed with specific contacts) from the EMERGE-DB:

- Field Sampling/Coring Sheets (2010-2017): Virginia Rich (virginia.isabel.rich@gmail.com), Suzanne Hodgkins (suzanne.b.hodgkins@gmail.com)
- Lab-Measured Biogeochemistry:
 - o Geochemistry (2010-2012): Suzanne Hodgkins (suzanne.b.hodgkins@gmail.com), Jeffrey Chanton (jchanton@fsu.edu)
 - Geochemistry (2013-2017): Rachel Wilson (rachelmywilson@gmail.com), Jeffrey Chanton (jchanton@fsu.edu)

In the table below, the "Data Source" column indicates the category (above) from which each variable derives, with blank values indicating standardized metadata that are applied to all applicable nodes in the EMERGE-DB (most of these were originally generated via standardization of the field sampling sheets). Note that each of the above categories actually represents several individual datasets as demarcated in the EMERGE-DB; some of these are not yet fully public, but are having a portion of their data published as part of this combined sheet.

Variable Definitions and Availability Notes

Below is a table of definitions and notes on data availability for each of the column names in the Sample Metadata Sheet. This table is also available in CSV format (filename beginning with "variables coring geochem sequenced samples").

In the Sample Metadata Sheet, all values within the same column have been ensured to be in the same basic format and units. Additionally, variables ending in a double underscore (__) represent "official" variable names that have been fully standardized for name, format, and units across data sources (this is done either during initial import to the DB, or [for some geochemistry variables] in the R script that generates this Sample Metadata Sheet, with the latter soon to be migrated to the DB import scripts). If a variable name does *not* end in "__", then it can still be considered consistent enough for data analysis using this sheet; it just might have some slight formatting inconsistencies in the broader EMERGE-DB.

For each variable, the "Fraction of Samples" refers to the percentage of field samples from distinct depths (i.e., with solid peat and corresponding porewater counted as the same "depth") that have that measurement (defined as any value other than "NA" or "Not Measured", with other non-numeric values indicating that an observation was made but could not be interpreted as a number [e.g., a water table depth of "Not Detected" indicates that no water table was detected at that site]). To give additional context for these measurement coverages, the variables are also categorized as either "intended to be comprehensive" (i.e., the measurement should exist for 100% of depths, even if this wasn't the case in practice) or "measured episodically" (i.e., the measurement was not expected to exist for all samples, which could be due to several reasons: e.g., it was only measured for a specific sub-project, it's expensive and/or time-consuming, or it's not expected to vary much over seasonal and interannual timescales). Thus, the majority of non-data-containing cells are not actually "missing data," but are either not applicable or intended to be episodic.

Variable	Туре	Definition	Data Source (Depth-Info = field sheets; Biogeochemistry = lab-measured biogeochemistry)	Sample Type Measured (if applicable)	Availability Notes	Fraction of Samples (%)
DepthID	text (unique for each row)	Unique random IDs for each "Soil-Depth" node in the EMERGE-DB, where each "Soil-Depth" node corresponds to a unique core section along with its corresponding porewater/poregas sample. This column is included here, as some coders may find it useful; otherwise it can be ignored.			Intended to be comprehensive.	100%
SampleID	text (unique for each row)	Sample IDs designed to be consistent thoughout the DB. These use the following standardized format: [core group].[YYYYMM]_[site abbrev]_[core # or name]_[min depth]to[max depth]			Intended to be comprehensive.	100%
SampleID_old	text (unique for each row)	Old sample IDs used in previous versions (pre-1.0.0) of the Sample Metadata Sheet.				100%
CoreGroup	categorical	Broad core groupings, useful for subsetting samples by subproject (e.g. "MainAutochamber" for the main yearly autochamber site cores). These groupings match the first part of the SampleID			Intended to be comprehensive.	100%

Variable	Туре	Definition	Data Source (Depth-Info = field sheets; Biogeochemistry = lab-measured biogeochemistry)	Sample Type Measured (if applicable)	Availability Notes	Fraction of Samples (%)
Month	numeric	Month of field sample collection.			Intended to be comprehensive.	100%
Year	numeric	Year of field sample collection.			Intended to be comprehensive.	100%
Date	date (YYYY- MM-DD)	Date that the sample was collected in the field.			Intended to be comprehensive.	100%
FieldSampling	categorical, date (YYYY-MM)	Month and year (YYYY-MM) of the beginning of the field sampling campaign during which the sample was collected. This tag is added for quick distinction between field campaigns that may not be obvious from simply parsing their sampling dates (e.g., continuous field campaigns that encompassed >1 calendar month).			Intended to be comprehensive.	100%
Habitat	categorical	labitat type. This is distinct from the specific ite names (which are under the "Site_" plumn), and represents the site's categorization within the superset of possible habitat types		Intended to be comprehensive.	100%	
Site	categorical	Full name of the site from which a sample or neasurement was taken. A full list of site names and their synonyms (compiled from https://bitbucket.org/MAVERICLab/isogeniedbools/raw/HEAD/tools/data/site_syn.json) can be found in this FAQ document under "Site Synonyms."			Intended to be comprehensive.	100%
Core	text	Core number (i.e., replicate), or core name + number if it's a non-canonical core (e.g., "SHR1" for Sphagnum High Resolution Core #1).			Intended to be comprehensive.	100%
GPS_	text			Intended to be comprehensive.	97.4%	
image_url	text	URL(s) of core image(s).			Intended to be comprehensive.	65.5%
DepthCode	text	Standard IsoGenie depth letter codes (S, M, D, etc.), used for field samples collected between 2010-2013.			Intended to be comprehensive.	69.5%
DepthMin	numeric			Intended to be comprehensive.	100%	
DepthMax	numeric	Deepest depth in the sample's depth range (cm). For solid-phase core sections, the width of the depth intervals is typically 3cm (2010-2012) or 4cm (2013-present). For porewater, they are typically ~3cm (based on the range encompassed by the swampsucker holes). If DepthMin_ and DepthMax_ are missing, this is likely because the sample was porewater and the full depth range was not recorded, so the full depth range can be approximated as DepthMin_ =DepthAvg1.5 and DepthMax_ =DepthAvg_ +1.5. Average depth of the sample (cm), represented			Intended to be comprehensive.	100%
DepthAvg	numeric	as a single number for easy plotting.			comprehensive.	100/0

Variable	Туре	Definition	Data Source (Depth-Info = field sheets; Biogeochemistry = lab-measured biogeochemistry)	Sample Type Measured (if applicable)	Availability Notes	Fraction of Samples (%)
metaG_ACE	text	Whether this sample has a metagenome sequenced at ACE: yes = sample name used; no = NA.		solid phase		82.4%
metaG_JGI_HiSeq	text	Whether this sample has a metagenome sequenced at JGI using HiSeq: yes = sample name used; no = NA.		solid phase		6.6%
metaG_JGI_NovaSeq	text	Whether this sample has a metagenome sequenced at JGI using NovaSeq: yes = sample name used; no = NA.		solid phase		59.9%
FTIR	categorical	Whether there was FTIR (Fourier Transform Infrared Spectroscopy) measured on this sample: yes = X; no = NA.	Biogeochemistry	solid phase	Measured episodically.	28.6%
FTICRMS	categorical	Whether there was FT-ICR MS (Fourier Transform Ion Cyclotron Resonance Mass Spectrometry) measured on this sample: yes = X; no = NA.	Biogeochemistry	porewater	Measured episodically.	1.2%
EEMS	categorical	Whether there was EEMS (Excitation-Emission Matrix Spectroscopy) measured on this sample: $yes = X$; $no = NA$.	Biogeochemistry	porewater	Measured episodically.	23.9%
ALD.cm_	numeric (some text values)	Active layer depth at time of sampling (cm below peat surface).	Depth-Info		Intended to be comprehensive.	93%
WTD.cm_neg_is_below_sfc	numeric (some text values)	Water table depth at time of sampling (cm), with negative=below and positive=above peat surface.	Depth-Info		Intended to be comprehensive.	94.1%
T_air.deg_C	numeric	Air temperature (deg C) at time of sampling.	Depth-Info		Intended to be comprehensive.	73.5%
T_soil.deg_C	numeric (some text values)	Soil temperature (deg C). These were mostly measured at the same depth as the corresponding peat and porewater samples, but here are a few that are a few cm off; contact suzanne for the exact depths.			Intended to be comprehensive.	64.3%
Weather_Conditions_on_Day_of_Coring	text	Weather conditions on day of coring, as noted by the field team.	Depth-Info		Intended to be comprehensive.	56.1%
Conductivity.uS	numeric	Conductivity (microsiemens). Depth-Info		porewater	Measured episodically.	1.6%
Redox_Potential_ORP. uncalibrated_relative_values_only	numeric	Oxidation-reduction potential. This was measured on an uncalibrated instrument, and I'm not sure of the units (it was a last-minute addition), so I would interpret these values tentatively as relative potentials across samples. More negative indicates more reducing conditions.		porewater	Measured episodically.	1.6%
O2.percent	numeric	Percent oxygen, as measured with an oxygen probe in the field.	Depth-Info	porewater, poregas	Measured semi-episodically.	6.6%
pH_porewater	numeric	pH of porewater.	Depth-Info (mostly)	porewater	Intended to be comprehensive.	51.4%
pH_peat	numeric	oH of peat, measured by mixing a small amount of wet peat with DI water for a couple minutes, hen measuring the pH with the same pH probe		solid phase	Measured episodically.	10.1%
BulkDensity	numeric	Peat bulk density (g / cm^3).	Biogeochemistry	solid phase	Measured episodically.	7.5%
percentWater_wt	numeric	%water (by weight) of wet peat. Biogeochemistry solid phase			Measured episodically.	17.6%
CH4.mM	numeric	Dissolved CH4 concentration (mM). Calculated using Henry's Law constant of 0.0015 mol/(L*atm).	Biogeochemistry	porewater	Intended to be comprehensive for waterlogged samples.	45.3%
CO2.mM	numeric	Dissolved inorganic carbon (DIC, or "CO2") concentration (mM). Calculated using Henry's Law constant of 0.037 mol/(L*atm).	Biogeochemistry	porewater	Intended to be comprehensive for	47.4%

Variable	Туре	Definition	Data Source (Depth-Info = field sheets; Biogeochemistry = lab-measured biogeochemistry)	Sample Type Measured (if applicable)	Availability Notes	Fraction of Samples (%)
					waterlogged samples.	
CH4.percent	numeric	Percent CH4 by volume (mole fraction * 100).	ercent CH4 by volume (mole fraction * 100). Biogeochemistry poregas		Intended to be comprehensive for non- waterlogged samples.	11%
CO2.percent	numeric	Percent CO2 by volume (mole fraction * 100).	Percent CO2 by volume (mole fraction * 100). Biogeochemistry porega			11%
d13C_CH4	numeric	d13C value of CH4.	Biogeochemistry	porewater, poregas	Intended to be comprehensive.	42.7%
d13C_CO2	numeric	d13C value of DIC or CO2.	Biogeochemistry	porewater, poregas	Intended to be comprehensive.	59.4%
dD_CH4	numeric	dD value of dissolved CH4.	Biogeochemistry	porewater	Measured episodically.	13.8%
dD_H2O	numeric	dD value of water.	Biogeochemistry	porewater	Measured episodically.	16.4%
d18O_H2O	numeric	d18O value of water.	Biogeochemistry	porewater	Measured episodically.	16.4%
DOC.mM	numeric	DOC concentration (mM). For the samples up through August 2011, these values may be slightly overestimated relative to values measured in Oct. 2011 and later, due to differences in sample prep (the later dates were acidified first whereas the earlier dates were not). The acidified samples (Oct. 2011 and later) probably have more accurate values, but the non-acidified samples are still accurate within reason.	Biogeochemistry	porewater	Measured semi-episodically.	26.3%
TN.mM	numeric	Total nitrogen concentration (mM), including both organic and inorganic N. Since this was measured at the same time as the DOC and the DOM has very high C/N ratios, many of the TN values were very close to the detection limit, so would use them with caution. Biogeochemistry poreway would use them with caution.			Measured semi- episodically.	22.8%
Acetate.uM	numeric	Acetate concentration (µM).	Biogeochemistry	porewater	Measured episodically.	16.2%
Propionate.uM	numeric	Propionate concentration (μM).	Biogeochemistry	porewater	Measured episodically.	5.2%
Butyrate.uM	numeric	Butyrate concentration (µM).	Biogeochemistry	porewater	Measured episodically.	0.5%
Formate.uM	numeric	Formate concentration (µM).	Biogeochemistry	porewater	Measured episodically.	4.2%
Sulfate.uM	numeric	Sulfate concentration (μM).	Biogeochemistry	porewater	Measured episodically.	22.5%
Phosphate.uM	numeric	Phosphate concentration (μM).	Biogeochemistry	porewater	Measured episodically.	7%
Nitrate.uM	numeric	Nitrate concentration (µM). Biogeochemistry pore		porewater	Measured episodically.	18.3%
Ammonia.uM	numeric	Ammonia concentration (uM) Riogeochemistry, porawater		Measured episodically.	18.3%	
Ca.uM	numeric	Calcium concentration (µM).	Biogeochemistry	porewater	Measured episodically.	2.1%
Mg.uM	numeric	Magnesium concentration (uM) Riogeochemistry porawater Measu		Measured episodically.	2.1%	
C.percent	numeric	%C in peat (by weight).	Biogeochemistry	solid phase	Measured semi- episodically.	49.3%
N.percent	numeric	%N in peat (by weight).	Biogeochemistry	solid phase	Measured semi-episodically.	48.8%

Variable	Туре	Definition	Data Source (Depth-Info = field sheets; Biogeochemistry = lab-measured biogeochemistry)	Sample Type Measured (if applicable)	Availability Notes	Fraction of Samples (%)
d13C_peat	numeric	d13C value of peat C.	Biogeochemistry	solid phase	Measured semi- episodically.	49.3%
d15N_peat	numeric	d15N value of peat N.	Biogeochemistry	solid phase	Measured semi-episodically.	48.8%
D14C_peat	numeric	Radiocarbon of peat, expressed as D14C.	Biogeochemistry	solid phase	Measured episodically.	0.9%
pool.DOM	categorical	Whether or not the DOM sample was pooled across all replicate porewater profiles from the same Site_and FieldSampling The DOM sample includes the following variables: FTICRMS_, EEMS_, DOC.mM_, TN.mM, Sulfate.uM, Nitrate.uM, Ammonia.uM, Ca.uM, Mg.uM	Biogeochemistry	porewater	This was done episodically. If not explicitly noted, assume FALSE.	26.1%
pool.dD_CH4	categorical	Whether or not the dD_CH4 sample was pooled across all replicate porewater profiles from the same Site and FieldSampling	Biogeochemistry	porewater	This was done episodically. If not explicitly noted, assume FALSE.	26.1%
UpdateDateDepthInfo	date (YYYYMMDD)	Date when the graph DB node for the coring netadata was last updated. Depth-Info				100%
UpdateDateBiogeochemistry	date (YYYYMMDD)	Date when the graph DB node for the lab- measured biogeochemistry was last updated.	Biogeochemistry			90.1%
NotesDepthInfo	text	Notes on field sampling.	Depth-Info			73.9%
NotesBiogeochemistry	text	Notes on lab-measured biogeochemistry.	Biogeochemistry			25.6%

Site Synonyms

The table below was generated by parsing the same lists of site <u>synonyms</u> and <u>abbreviations</u> used directly by the DB import scripts.

Offical Site Name	Abbreviation Synonyms (note: do not infer ecology from species tags in these names)		
Avni Transect A (collar 1,2,3)	AvAPW1	 PW1 Avni Transect A (collar 1,2,3) Avni Transect A (collar 1,2,3) - fen Avni Transect A - PW1 (collar 1,2,3) Avni Transect A - PW1 (collar 1,2,3) - fen 	
Avni Transect A (collar 4,5,6)	AvAPW2	 PW2 Avni Transect A (collar 4,5,6) Avni Transect A - PW2 (collar 4,5,6) 	
Avni Transect B (collar 10,11,12)	AvBPW3	 PW3 Avni Transect B (collar 10,11,12) Avni Transect B (collar 10,11,12) - bog Avni Transect B - PW3 (collar 10,11,12) Avni Transect B - PW3 (collar 10,11,12) - bog 	
Avni Transect B (collar 13,14,15)	AvBPW4	 PW4 Avni Transect B (collar 13,14,15) Avni Transect B (collar 13,14,15) - bog Avni Transect B - PW4 (collar 13,14,15) Avni Transect B - PW4 (collar 13,14,15) - bog 	
Avni Transect B (collar 16,17)	AvBPW5	 PW5 Avni Transect B (collar 16,17) Avni Transect B - PW5 (collar 16,17) 	
Avni Transect B (collar 20,21,22)	AvBPW6	PW6 Avni Transect B (collar 20,21,22)	

Offical Site Name	Abbreviation	Synonyms (note: do not infer ecology from species tags in these names)
		Avni Transect B (collar 20,21,22) - bog Avni Transect B - PW6 (collar 20,21,22) Avni Transect B - PW6 (collar 20,21,22) Avni Transect B - PW6 (collar 20,21,22) - bog
Avni Transect C-extension (P, palsa-stayed-palsa; collar 1,2,3)	AvCeP	 Avni Transect C-extension (P, palsa-stayed-palsa; collar 1,2,3) AvC-P AvC-1 AvC-P1 AvnC-P
Avni Transect C-extension (C, palsa-now-fen; collar 4,5,6)	AvCeC	Avni Transect C-extension (C, palsa-now-fen; collar 4,5,6) AvC-C AvCe-C
Avni Transect C-extension (F, fen-stayed-fen; collar 7,8,9)	AvCeF	 Avni Transect C-extension (F, fen-stayed-fen; collar 7,8,9) AvC-F AvCe-F AvC-F1
Bog1	Bog1	Bog1 Bog 1 - SPHAGNUM
Collapsed Palsa, New Hole, First Site, Collapse Hole	CPN1H	 CPN1-H CPNH1-H CPN1 Collapsed Palsa New 1 - Hole Collapsed Palsa New - Hole Collapsed Palsa, New Hole, First Site, Collapse Hole COLLAPSED PALSA, NEW HOLE, FIRST SITE, COLLAPSE HOLE
Collapsed Palsa, New Hole, First Site, Palsa Ring	CPN1R	 CPNH1-R CPNO1 Collapsed Palsa New 1 - Rim Collapsed Palsa New - Rim Collapsed Palsa, New Hole, First Site, Palsa Ring COLLAPSED PALSA, NEW HOLE, FIRST SITE, PALSA RING
Collapsed Palsa, New Hole, Second Site, Collapse Hole	CPN2H	 CPNH2-H CPN2 Collapsed Palsa, New Hole, Second Site, Collapse Hole COLLAPSED PALSA, NEW HOLE, SECOND SITE, COLLAPSE HOLE
Collapsed Palsa, New Hole, Second Site, Palsa Ring	CPN2R	 CPNH2-R CPNO2 Collapsed Palsa, New Hole, Second Site, Palsa Ring COLLAPSED PALSA, NEW HOLE, SECOND SITE, PALSA RING
Eriophorum Autochamber Site	Е	E ERIOPHORUM ANGUSTIFOLIUM ERIOPHORUM Eriophorum Eriophorum Autochamber Eriophorum Autochamber Site AutochamberEriophorum E.time Stordalen fen Stordalen Eriophorum fen
Eriophorum Optimization Site	EOS	EOS Eriophorum Optimization Site (EOS) - ERIOPHORUM VAGINATUM + SPHAGNUM Eriophorum Optimization Site

Offical Site Name	Abbreviation	Synonyms (note: do not infer ecology from species tags in these names)
Fen1	Fen1	• Fen1 • Fen 1 - CAREX • FEN 1
Fen2	Fen2	• Fen2 • Fen 2 - CAREX • FEN 2 • Fen 2
Palsa Autochamber Site	P	 P PALSA Palsa Palsa Autochamber Palsa Autochamber Site AutochamberPalsa Stordalen palsa Stordalen Palsa
Palsa Hole Big - Hole	РНВН	 PHB-H PHBH Palsa Hole - Big (PHB) - COLLAPSED PALSA Palsa Hole Big (PHB) - COLLAPSED PALSA PHB Palsa Hole Big Palsa Hole Big - Hole
Palsa Hole Big - Rim	PHBR	 PHB-R PHBR PHB Rim Palsa Hole Big - Rim
Palsa Hole Small - Hole	PHSH	 PHS Palsa Hole - Small (PHS) - COLLAPSED PALSA Palsa Hole Small (PHS) - COLLAPSED PALSA CP CPH PHS-H Palsa Hole Small - Hole Palsa Hole Small
Palsa Hole Small - Rim	PHSR	 CPP Collapsed Palsa Rim CP Rim PHS Rim PHS-R Collapsed Palsa Permafrost
Palsa Optimization Site	POS	POS Palsa Optimization Site
Sphagnum Autochamber Site	s	S SPHAGNUM Sphagnum Sphagnum Autochamber Sphagnum Autochamber Site AutochamberSphagnum Stordalen bog Stordalen Sphagnum Bog
Sphagnum Optimization Site	sos	SOS Sphagnum Optimization Site (SOS) - SPHAGNUM Sphagnum Optimization Site
AJ-Palsa	AJP	AJ-Palsa AJ-P

Offical Site Name	Abbreviation	Synonyms (note: do not infer ecology from species tags in these names)
AJ-Sphagnum	AJS	AJ-SphagnumAJ-SphagAJ-S
AJ-Eriophorum	AJE	AJ-ErioAJ-EAJ-Eriophorum
Inc-Palsa	IncP	IncPInc-PInc-Palsa
Inc-Sphagnum	IncS	IncSInc-SInc-Sphagnum
Inc-Eriophorum	IncE	 IncE Inc-E Inc-E Auto Inc-Eriophorum
Bioreactor-Palsa	BioP	Bioreactor-Palsa B- Palsa
Bioreactor-Fen, Avni I	BioFAvI	Bioreactor-Fen, Avni I B- Fen Avni I
Bioreactor-Fen, Avni II	BioFAvII	Bioreactor-Fen, Avni II B- Fen Avni II