Sequence File Naming Convention

This is the proposed file naming convention to use for sequencing data generated in the CliMush project. The following naming method allows for mixing of raw sequencing files across sequencing runs, sequencing platforms, and collection dates without file name duplication.

Formatting of file names was guided by Jenny Bryan's (RStudio) content on best practices in file naming. Here, I refer to individual sections of the file name (e.g., treatment) as a *label*, and the entire file name as *file name*. Click on each section of the file name outline below for a detailed description, justification, and label. PacBio sporocarp sequence files have a slightly different format, so also consult sporocarp exceptions.

<seq-platform>_<compartment>_<coll-date>_<ecoregion>_<treatment>_<subplot>

File names are designed to be as concise as possible while maintaining clear distinctions between each file. The order of labels in the file name are arranged from the broadest label (sequencing method) to the most specific (subplot).

examples

filename	sequences are from…
illumina_soil_2023-05_D19_UC_01	a soil sample that was collected in May 2023 from Alaska's unburned conifer plot at subplot 01
sanger_leaf-sp01_2022-10_D14_BO_09	a leaf sample from plant species 01 that was collected in October 2022 from Arizona's burned oak plot at subplot 09
illumina_spore_2023-05_D03_UG_05	a spore sample that was collected beginning in May 2023 from Florida's unburned grassland site at subplot 05
pacbio_sporocarp-f_2023-04_pool01	fresh sporocarps that were sequenced during a PacBio run that took place in April 2023 in pool 01* that have not yet been demultiplexed
pacbio_sporocarp-f_2023-05_D16_603	fresh Oregon sporocarp 603 that was sequenced during a PacBio run that took place in May 2023 (after demultiplexing)

*PacBio runs are distributed across two pools (one per half of a SMRT cell) in order to maximize multiplexing per run (192 samples/cell, 384 samples/run)

sequencing platform

description

The sequencing method/technology that produced the sequencing data. Do not include the specific machine (e.g., MiSeq, SMRT); that can be included in metadata.

justification

This allows sequencing files to be sorted in the bioinformatics pipeline when certain processes or parameters are method-dependent. It will also help distinguish the Illumina and PacBio sequencing runs for soil and litter.

options

- sanger
- illumina
- pacbio

compartment

description

The material from which DNA samples were extracted. Also incorporates the species number from which leaf, seed, and root tissues originated, and whether a sporocarp is from fresh or museum archive material.

justification

It's best to be as explicit as possible by spelling the full compartment name, especially with difficulties in abbreviating spore and sporocarp in distinct ways. Demultiplexed Illumina reads for soil and litter are usually returned with L###_R# tags on the files, so fully spelling out the compartment avoids any further confusion.

Preferably, 'sp01, sp02, …' will be the extent to which the tissue source is described; the identity of the plant can remain in the metadata. This makes the file names much easier to sort if no additional data is included.

options

compartment	sub-compartment	label
litter	—	litter
soil	—	soil
spore	—	spore
sporocarp	fresh	sporocarp-f
	archive	sporocarp-a
leaf	plant species 01	leaf-sp01

compartment	sub-compartment	label
	plant species 02	leaf-sp02
seed	plant species 01	seed-sp01
	plant species 02	seed-sp02
root	plant species 01	root-sp01
	plant species 02	root-sp02

collection date

description

Year and month that the sample was collected following ISO 8601 formatting of YYYY-MM (no day). If sampling occurs over a transition between months, use the month that the sampling started. If samples were aggregated across several collection periods, like in the case of PacBio sporocarp sequencing, then use the date that the sequencing library was submitted for sequencing (see sporocarp exceptions).

justification

We need a way to distinguish between collections made at each subplot across sampling seasons. This date format is preferable over the use of spring/fall because it can be chronologically sorted (whereas ‘fall’ would be first alphabetically). We can also keep the date formats consistent across sequencing platforms, even if the meaning of the date included in the file name may differ (see sporocarp exceptions).

options

year: [YYYY]

- 2022, 2023, 2024

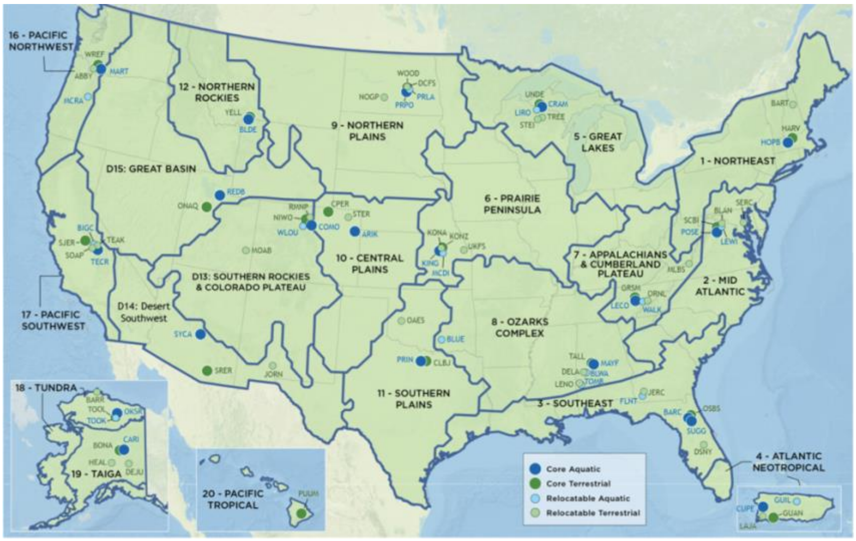
month: [MM]

- 01 = January
- 06 = June
- 12 = December

ecoregion

description

Geographic location from which the sample was collected using the NEON’s abbreviation for ecoclimatic domains (ecoregions). The ecoclimatic domain abbreviation is used across all NEON-generated data (see table below).



justification

A core component of the CliMush project is the ability to incorporate NEON data within our analysis. The only NEON-accepted abbreviation at the ecoregion level is the domain name (D##). If at any point we wanted to combine our data with NEON data, or contribute our data to NEON, a consistent method of naming the ecoregion would be beneficial.

If the mental gymnastics of converting the NEON domain label (D##) to site name makes the file name less human-readable, I would propose sticking with the site abbreviations over ecoregion abbreviations. It's far easier to read file names when the number of characters per label is consistent (all 3-letters long, except for Harvard Forest which can be shortened to HAF). It also makes the use of regular expressions much simpler.

options

state	site name	site abbr.	ecoregion name	ecoregion abbr.*	domain label
AK	Bonanza Creek	BNZ	Taiga	TAIG	D19
AZ	Santa Rita	SRE	Desert Southwest	DSW	D14
CO	Niwot Ridge	NWT	Southern Rockies & Colorado Plateau	SROC	D13
FL	Ordway-Swisher	ORD	Southeast	SE	D03
KS	Konza Prairie	KON	Prairie Peninsula	PRAI	D06
MA	Harvard Forest	HFMA	Northeast	NE	D01
MN	Cedar Creek	CDR	Great Lakes	GRTL	D05

state	site name	site abbr.	ecoregion name	ecoregion abbr.*	domain label
OR	HJ Andrews	HJA	Pacific Northwest	PNW	D16
OR	Mt. Pisgah	PIS	Pacific Northwest	PNW	D16

*I can't find the source of these ecoregion abbreviations, such as if they originated outside of this project; these could also be changed to be a consistent number of characters if they are preferable over both the domain label or the site abbreviation

treatment

description

Combination of the burn history and habitat type of the plot from which the sample was collected. Burn history abbreviation is first, then habitat, with no separation or punctuation between these abbreviations. PacBio sporocarp samples will not have a treatment, so this section of the file name should be omitted (see sporocarp exceptions).

justification

It is preferable, when possible, to keep the number of characters per label consistent. I therefore replaced UB (unburned) with simply U; burn and habitat abbreviations did not change from the previous sample name format. I did not include any punctuation between the burn history and habitat because treatment is a simple two-letter code. Punctuation doesn't add any clarity for human readability nor does it help with computer readability.

options

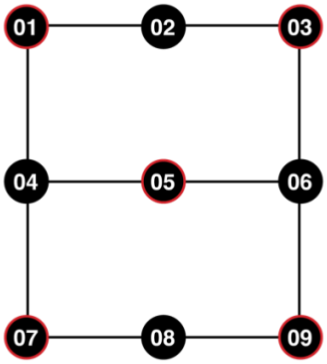
burn history	label
unburned	U
burned	B

habitat	label
conifer	C
oak	O
grassland	G

subplot

description

The number of the subplot from which a sample was collected. The figure below shows the general arrangement of subplots within a plot. Subplots circled in red are the remaining subplots sampled after sample reduction in 2022. PacBio sporocarp samples will not have a subplot, so this section of the file name should be omitted (see sporocarp exceptions).



justification

I omitted the ‘S’ because I don’t think it adds any additional information to the file name, but if others disagree I can include it. I believe it was included before to distinguish it from the transect number, which is now omitted (you can infer the transect number based on the subplot number).

options

- samples collected in spring of 2022 will contain subplot numbers 01–09
- samples collected after sample reduction (subplots circled in red above) will contain numbers 01, 03, 05, 07, 09

sporocarp exceptions

Sporocarp collections have different metadata than other samples. For example, foray collections don’t consistently include habitat or burn history information, nor is a sporocarp associated with a subplot. Additionally, PacBio sequencing of sporocarps is done in batches that are independent of sporocarp collection date, unlike Illumina sequencing of other compartments such as soil and litter, which are collectively run at the end of a sampling season. Therefore, sporocarp file names are slightly different from other compartments. A sporocarp file name does *not* have treatment and subplot values but instead has the collection number associated with the sporocarp (see below). The date in the file name does *not* correspond to the date the sporocarp was collected but instead has the date the DNA library was submitted to the sequencing core.

other compartments (as shown at start of document):

seq-platform_compartment_coll-date_ecoregion_treatment_subplot

sporocarps:

seq-platform_compartment_***seq-date***_ecoregion_***coll-number***

collection number

Ideally, collection number (coll-number) would be a four-digit numeric (0001—9999) that is sequentially assigned to a sporocarp collection on a per-ecoregion basis (e.g., first collection made as part of this project at an ecoregion is 0001, and so on). However, I think numbering has already occurred for most sporocarp collections, so there may be variation in what a collection number looks like across ecoregion sites.

Regardless of how a sporocarp collection is numbered, the collection number should match the plate number for that sporocarp on iNaturalist. For example, this collection of *Arrhenia* made in Alaska has a plate number (957) on its iNaturalist page under observation fields. Given this information, the sequence file from this collection (if sequenced in December 2023) would be:

pacbio_sporocarp_2023-12_D19_957

sequence submission date

The reason that the sequence submission date was used instead of collection date for PacBio sporocarp sequences is two-fold.

First, sporocarps are grouped for sequencing in the order they are received by the University of Oregon. This often does not correspond to the order in which they were collected. The date of collection for samples on a PacBio run is therefore highly variable. In contrast, compartments such as soils are grouped for sequencing by collection season, so that all samples collected in the spring, for example, are grouped together on the same Illumina run. It is much more tenable to include collection date in Illumina runs, because all samples from an ecoregion should have the same collection date, and often samples from different ecoregions will share the same collection date (since it's a monthly scale).

Second, the ultimate role of the PacBio sequences is different than the other compartments. Sample collection in other compartments are snap-shot representations of the fungal communities at a given place and time. While the iNaturalist observations of sporocarps may provide the same granularity, PacBio sporocarp sequencing is primarily meant to support taxonomic identification across all compartments. Given these differences, collection date is more contextually informative for other compartments than it is for sporocarps.