Documentation for the CCMRI Metagenomic Study Corpus Curation

Corpus Creation and Content

The CCMRI Metagenomic Study Corpus (CCMRI corpus) comprises key textual information of 2.346 EBI MGnify metagenomics studies (and sample description thereof).

In particular, key pieces of information on 2.346 EBI MGnify metagenomics studies were collected on the 8th of August 2023. To capture studies from life-supporting ecosystems these studies belong to the *aquatic* (root:Environmental:Aquatic) and *terrestrial* (root:Environmental:Terrestrial) biomes (1761 and 585 studies respectively)[1]. These studies 2.346 were randomly split in ~50% evaluation and ~50% training per biome.

Upon the aforementioned corpus split the MGnify "Related study" has been taken into consideration. Once a study is binned into the "evaluation" or the "training" corpus-part, then all of its related studies (and the related studies thereof) would end up in the same part as well.

There above resulted in the following metagenomics study corpus:

- aquatic
    - aquatic_evaluation 713 studies
    - aquatic_training 1048 studies
- terrestrial
    - terrestrial_evaluation 234 studies
    - terrestrial_training 351 studies

The corpus text files organized as above are available at: https://github.com/lab42open-team/ccmri_corpus/tree/main/data

[1] According to the MGnify API three metagenomic studies were characterised as belonging to both the *aquatic* and the *terrestrial* biomes. In their web page, however, these three studies were all classified as *aquatic*. Thus, they were classified as *aquatic* in the CCMRI Corpus too.

Main Guidelines

Aim of this curation is to annotate if Climate Change (CC) influences a microbiome (MB) community (CC -> MB) or if microbial communities contribute to CC (MB -> CC). Both are considered CC-related in the CCMRI context, meaning that an annotation should be made if either or both is stated.

In particular, CC-related studies are further annotated with labels and tags as follows:

- **Label 1** indicates studies that actively study a **CC-caused phenomenon affecting** the microbiome (MB) (**1. CC-caused**).
- **Label 2** concerns studies in which a microbiome-mediated process **contributes** to CC (**2. CC-causing**).
- **Label 3** describes studies that **investigate mitigation strategies** to combat CC (**3. CC-mitigating**).

A study, according to what it explores, can be annotated with more than one label if needed (e.g. both **1. CC-caused** and **2. CC-causing**)

Beyond the labels, **tags** are assigned to each study describing the phenomena/processes being explored.

Tags for **Label 1** annotated studies typically include *desertification*, *soil/ocean acidification*, *permafrost thawing*, *extreme weather events*, *increased incidence of wildfires*, *sea level rise*, *temperature rise, elevated CO2 levels, coral bleaching*, *shifts in ecosystems and species distributions*.

Tags for **Label 2** annotated include *methane production from wetlands*, *soil carbon decomposition*, *methane production in ruminants*, *nitrous oxide emissions from agriculture*, *peatland degradation*, *biochar degradation*.

An example tag of **Label 3** studies is *methane-feeding microbiome*. Others are *sulfate-reduction* and CC-mitigating practices.

If there is enough evidence, studies can be annotated further with the following two sub-labels:

> **Sub-label a**. for a time-series study (given the temporal aspect of the planet warming) covering a period of more than a year.
>
> **Sub-label b**. For a study that involves a comparison of samples from transitional environmental states where the transition is linked to CC, like permafrost that changes into periodically freezing soil.

Note: Each study is to be annotated independently from the rest of the studies; this applies to “Related studies” too (thus upon annotation the “Related Study” link is not to be followed).

Curation Examples

Note: the following metagenomics studies are analysed in depth for curator training purposes. Shown in green are text segments that can drive a curator's decision on whether a study is CC-related or not. Shown in light blue are the fields that the curator has filled-in. Such highlighting is done only in this example and for illustration purposes only, no segment highlighting/annotation is included in the CCMRI corpus.

1. A CC-related corpus entry (CC-caused label)

Study

MGYS00000693

Title

Soil microbial diversity of 106 samples Metagenome

Description

These data contain the information about the response of soil bacterial diversity and composition to multi-factorial environmental changes, including increased precipitation, rising temperature, adding nitrogen, adding phosphorus, removing plant functional groups, grazing, and some of their combinations. Specifically, [samples] represent the treatment of removing no plant functional group; [samples] represent the treatment of removing one plant functional group; [samples] represent the treatment of removing two plant functional groups; and [samples] represent the treatment of removing three plant functional groups. Meanwhile, [samples] represent the control; [samples] represent the treatment of warming; [samples] represent the treatment of watering; [samples] represent the treatment of simultaneous warming and watering; [samples] represent the treatment of adding phosphorus; [samples] represent the treatment of adding nitrogen; [samples] represent the treatment of simultaneous adding N and watering; [samples] represent the treatment of simultaneous adding N and P; [samples] represent the treatment of grazing; [samples] represent the treatment of simultaneous grazing and watering; [samples] represent the treatment of simultaneous grazing and adding P; [samples] represent the treatment of simultaneous grazing and adding N; [samples] represent the treatment of simultaneous grazing and adding N and watering; [samples] represent the treatment of simultaneous grazing and adding N and P.

CC-relatedness

CC-related

Labels

CC-caused

Tags

Temperature rise

Explanation

Investigates the response of soil bacterial diversity and composition to multi-factorial environmental changes, including rising temperature among other parameters. CC-related since it mimics the temperature rise of Climate Change.

2. A CC-related corpus entry (CC-causing label)

Study

MGYS00000745

Title

Rice paddy soil Targeted Locus (Loci)

Description

Investigation of Methane Emission and Bacterial and Archaeal Communities in Rice Paddy Soil during Rice Cultivation.

CC-relatedness

CC-related

Labels

CC-causing

Tags

Greenhouse gas emission

Explanation

Investigates methane emission of bacterial and archaeal communities in rice paddy soil. Even though there is no direct mention of CC, it studies the emission of a greenhouse gas, which affects CC.

3a. A CC-related corpus entry (CC-caused & CC-causing label)

Study

MGYS00000529

Title

Permafrost 454 amplicons

Description

Study of microbial communities in permafrost, active layer and thermokarst bog in Alaska

Related publication

Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes.

Related publication abstract

Over 20% of Earth's terrestrial surface is underlain by permafrost with vast stores of carbon that, once thawed, may represent the largest future transfer of carbon from the biosphere to the atmosphere. This process is largely dependent on microbial responses, but we know little about microbial activity in intact, let alone in thawing, permafrost. Molecular approaches have recently revealed the identities and functional gene composition of microorganisms in some permafrost soils and a rapid shift in functional gene composition during short-term thaw experiments. However, the fate of permafrost carbon depends on climatic, hydrological and microbial responses to thaw at decadal scales. Here we use the combination of several molecular 'omics' approaches to determine the phylogenetic composition of the microbial communities, including several draft genomes of novel species, their functional potential and activity in soils representing different states of thaw: intact permafrost, seasonally thawed active layer and thermokarst bog. The multi-omics strategy reveals a good correlation of process rates to omics data for dominant processes, such as methanogenesis in the bog, as well as novel survival strategies for potentially active microbes in permafrost.

CC-relatedness

CC-related

Labels

CC-caused & **CC-causing**

Tags

permafrost thawing and **greenhouse gas emission**

Explanation

Determines the phylogenetic composition of the microbial communities, including several draft genomes of novel species, their functional potential and activity in soils representing different states of thaw: intact permafrost, seasonally thawed active layer and thermokarst bog, using multi -omics. Additionally, methanogenesis' process rates are found in good correlation with the -omics data, which justifies the "CC-causing" label.

3b. A CC-related corpus entry (CC-causing + CC-caused labels)
Study

Study

MGYS00000365

Title

Functional metagenomic profiling of Tibetan Plateau soils affected by permafrost or seasonal freezing

Description

Approximately two thirds of the Tibetan Plateau is affected by permafrost and this area reacts particularly sensitively to possible effects of climate change. However, little is known about the functional potential of the microbial communities inhabiting this environment, which is of key importance to predict potential feedback effects, such as increased emissions of greenhouse gas. A metagenomic analysis was performed on soil profiles from two meadow sites on the Tibetan Plateau, either affected by permafrost (site Huashixia [HUA]) or seasonal freezing (Site Haibei Station [HAI]). The goal was to determine how respiratory and fermentative pathways varied with soil depth and across sites.

CC-relatedness

CC-related

Labels

CC-caused & **CC-causing**

Tags

permafrost thawing and **greenhouse gas emission**

Explanation

Studies the functional metagenomic profiles of Tibetan Plateau soils affected by permafrost or seasonal freezing. It is CC-related as it studies an increase in gas emissions (CC-causing label) due to permafrost thawing (CC-caused label) by performing metagenomic analysis on two soil profiles one affected by permafrost and the other by seasonal freezing.

4. A CC-related corpus entry (CC-mitigating labels)

Study

MGYS00000368

Title

Metagenomic study targeting N cycling processes

Description

Sugarcane is a crop for bioenergy in Brazil and one of the main concerns in the production of this crop is the impact on the environment, in particular in greenhouse gases (GHG) emissions. Recent studies have shown that the N2O emissions related to sugarcane production are dependent on soil management practices (Carmo et al. 2013). The way to mitigate N2O emissions would be to understanding the conditions and the processes involved in the N2O production and consumption. However, no studies have linked the emissions of GHGs with soil-borne microbial communities, which are the main players in nutrient cycling.

CC-relatedness

CC-related

Labels

CC-mitigating

Tags

CC-mitigating and **greenhouse gas emission**

Explanation

Studies soil microbial communities related to the emission of Greenhouse Gases (GHG) in order to mitigate them by changing the soil management practices, which in turn will make sugarcane production more sustainable.

5. Non CC-related study examples

Study

MGYS00001007

Title

Microorganisms responsible for carbon uptake from ethylbenzene and its degradation products

Description

This study aimed at identifying microorganisms that are involved in the carbon uptake from ethylbenzene and its degradation products in soil microcosms. These microcosms were determined by combing stable isotope probing and high throughput sequencing techniques. Briefly total genomic DNA was extracted from microcosms that degraded approximately 80% of the added labeled or unlabeled ethylbenzene and was subject to gradient ultracentrifugation and were divided into fractions for further analysis. Heavy fractions and total genomic DNA samples from all microcosms were sampled. This BioProject contains the illumina sequences generated from the total genomic DNA samples from all microcosms.

Explanation

This study explores microorganisms that are involved in the carbon uptake from ethylbenzene and its degradation products in soil microcosms. No link to CC is mentioned

Study

MGYS00000678

Title

Metagenomes isolated from NE Brazil

Description

Mangroves are important and productive ecosystems found in tropical and subtropical environments, providing habitats for a variety of species. These ecosystems have been suffering from impacts through the years and consequently the activity of soil microorganisms, which are directly related to valuable processes in these environments, is affected. Understanding the diversity and function of microbial communities and their response to environmental changes is essential for the maintenance of significant functions in these ecosystems. Currently few studies in Brazil are devoted to the understanding of microbial diversity in mangrove soils. Thus it becomes essential to study the microbial diversity in these environments and to search for genes encoding for enzymes of biotechnological interest. For this reason, the aim of this study was to assess the soil microbial diversity from four mangroves in Ceara state, northeast Brazil, and its response to possible impacts, and search genes of biotechnological interest using metagenomics strategies. The four studied mangroves are located in the state of Ceara, northeastern Brazil. Two of them are located in the

extreme east (Jaguaribe Mangrove) and west (the Mangrove Timonha) coast of the state separated by 530 km and the other two are located in a central region, near the city of Fortaleza, state capital (Coco and Pacoti mangroves). Jaguaribe (JAG) mangrove is located at the largest river of the state, in a sub-urbanized region, and impacted by agriculture run-off and extensive shrimp farming. In contrast, Timonha (TIM) is an undisturbed mangrove, located in an island inside the estuary, with limited access to humans. The Coc? (COC) and Pacoti (PAC) mangroves, located in the metropolitan region of Fortaleza, suffer impacts due to water pollution, deforestation of native vegetation, especially vegetation of dunes and mangroves, extraction of sand, clay, stone and release of industrial effluents. The sampling locations TIM - S 02?56.587? W 041?19.064?; JAG - S 04?26.749? W 37?46.989?; PAC - S 03?49.226? W 038?24.286?; COC - S 03?46.482? W 38?26.552?. Sediment samples were collected in three sites inside the mangroves aiming to cover typical habitats in this ecosystems: near the river, Rhizophora mangle forest and the last one in an area covered by Avicennia schaueriana. In each habitat five sediment cores from 0-10 cm layer in an area of 10 m2, in the low tide of 0.0, were collected. The fifteen samples from each mangrove were pooled to form a single composite sample and DNA metagenomic extraction was carried out using the PowerMaxSoil DNA Extraction kit (MoBio Laboratories, Carlsbad, CA, USA) following the manufacturer's protocol. The extracted DNA was then subjected to 454 pyrosequencing.

Explanation

This mangrove/deforestation microbiome study explores a number of anthropogenic types of impact but there is no link to CC

6. Borderline case

Study

MGYS00001007

Title

Tundra soil Metagenome

Description

Investigation distribution of soil microbial community

Abstract

Vegetation-associated impacts on arctic tundra bacterial and microeukaryotic communities.

The Arctic is experiencing rapid vegetation changes, such as shrub and tree line expansion, due to climate warming, as well as increased wetland variability due to hydrological changes associated with permafrost thawing. These changes are of global concern because changes in vegetation may increase tundra soil biogeochemical processes that would significantly enhance atmospheric CO2 concentrations. Predicting the latter will at least partly depend on knowing the structure, functional activities, and distributions of soil microbes among the vegetation types across Arctic landscapes. Here we investigated the bacterial and microeukaryotic community structures in soils from the four principal low Arctic tundra vegetation types: wet sedge, birch hummock, tall birch, and dry heath.

Sequencing of rRNA gene fragments indicated that the wet sedge and tall birch communities differed significantly from each other and from those associated with the other two dominant vegetation types. Distinct microbial communities were associated with soil pH, ammonium concentration, carbon/nitrogen (C/N) ratio, and moisture content. In soils with similar moisture contents and pHs (excluding wet sedge), bacterial, fungal, and total eukaryotic communities were correlated with the ammonium concentration, dissolved organic nitrogen (DON) content, and C/N ratio. Operational taxonomic unit (OTU) richness, Faith's phylogenetic diversity, and the Shannon species-level index (H') were generally lower in the tall birch soil than in soil from the other vegetation types, with pH being strongly correlated with bacterial richness and Faith's phylogenetic diversity. Together, these results suggest that Arctic soil feedback responses to climate change will be vegetation specific not just because of distinctive substrates and environmental characteristics but also, potentially, because of inherent differences in microbial community structure.

Explanation

This study speaks about tundra vegetation types. If there were a statement reg. CC-related changes in the microbiome the study should have been characterised as CC-related (or if there were an explanation of a transition from one vegetation type to the other due to CC. No such statement is available in the corpus text (ie. metagenomics record text fields and related abstract text)).

To be more thorough, such a statement is made in the study full text manuscript. However the manuscript full text is not available nor to the curator nor to the computer system. It should be clarified that a curator should curate based on the corpus text, not based on what the user may infer based on background biology knowledge; a computer system has no chance to capture such a case.

Note

The following publications have been found to be associated with more than 15 studies each in the MGnify study dataset. Their abstracts have been found non-informative, in terms of the CC-relatedness (or not), and have been excluded from the corpus:

1. Mitchell AL, Scheremetjew M, Denise H, Potter S, Tarkowska A, Qureshi M, Salazar GA, Pesseat S, Boland MA, Hunter FMI, Ten Hoopen P, Alako B, Amid C, Wilkinson DJ, Curtis TP, Cochrane G, Finn RD. **EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies**. Nucleic Acids Res. 2018 Jan 4;46(D1):D726-D735. doi: 10.1093/nar/gkx967. PMID: **29069476**; PMCID: PMC5753268.
2. Weißbecker C, Schnabel B, Heintz-Buschart A. **Dadasnake, a Snakemake implementation of DADA2 to process amplicon sequencing data for microbial ecology**. Gigascience. 2020 Nov 30;9(12):giaa135. doi: 10.1093/gigascience/giaa135. PMID: **33252655**; PMCID: PMC7702218.
3. Mitchell AL, Almeida A, Beracochea M, Boland M, Burgin J, Cochrane G, Crusoe MR, Kale V, Potter SC, Richardson LJ, Sakharova E, Scheremetjew M, Korobeynikov A, Shlemov A, Kunyavskaya O, Lapidus A, Finn RD. **MGnify: the microbiome analysis resource in 2020**. Nucleic Acids Res. 2020 Jan 8;48(D1):D570-D578. doi: 10.1093/nar/gkz1035. PMID: **31696235**; PMCID: PMC7145632.
4. Větrovský T, Morais D, Kohout P, Lepinay C, Algora C, Awokunle Hollá S, Bahnmann BD, Bílohnědá K, Brabcová V, D'Alò F, Human ZR, Jomura M, Kolařík M, Kvasničková J, Lladó S,

López-Mondéjar R, Martinović T, Mašínová T, Meszárošová L, Michalčíková L, Michalová T, Mundra S, Navrátilová D, Odriozola I, Piché-Choquette S, Šťursová M, Švec K, Tláskal V, Urbanová M, Vlk L, Voříšková J, Žifčáková L, Baldrian P. **GlobalFungi, a global database of fungal occurrences from high-throughput-sequencing metabarcoding studies**. Sci Data. 2020 Jul 13;7(1):228. doi: 10.1038/s41597-020-0567-7. Erratum in: Sci Data. 2020 Sep 15;7(1):308. PMID: **32661237**; PMCID: PMC7359306.

Curation Procedure and Inter Annotator Agreement

The corpus has been annotated by two curators. Following the above curation guideline finalization, 95 studies of the *terrestrial training* corpus part (out of the 2,346 CCMRI corpus studies) have been annotated by both curators.

The curators annotated the same way 93 out of the 95 studies achieving a Cohen's Kappa score of 0.82; a score considered as almost perfect.