Extended Methodology
8th ICMBB 2025
Poster Presentation: PE007

Methanogenic Archaeal Class Bog-38 In the North Selangor Peat Swamp Forest: A Tropical Outlier In a Predominantly Arctic Lineage

0.1 Study Area

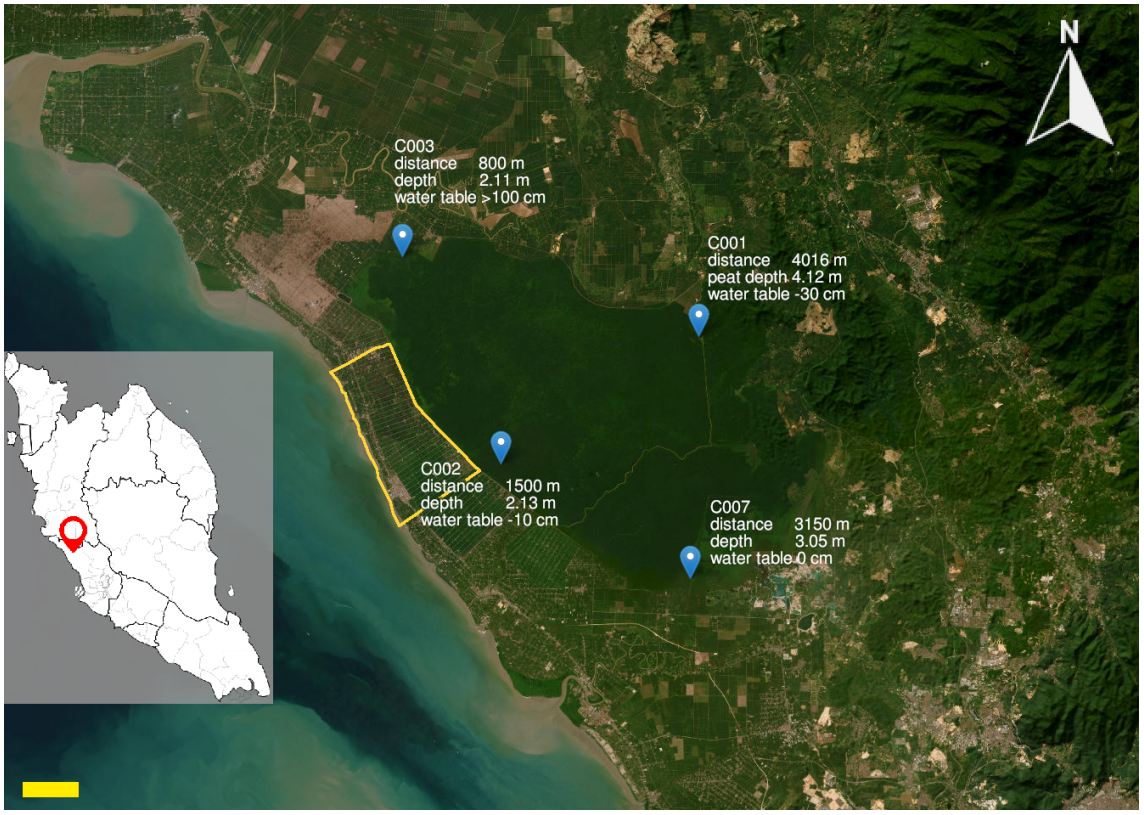

FIGURE 1: Map of the NSPSF, Selangor, Peninsular Malaysia. The four sampling sites are marked with blue indicators. The scale bar in the bottom left represents a distance of 5 kilometers. The yellow boundary outlines the nearby town of Sekinchan and surrounding paddy fields. The inset map shows the location of the NSPSF (red marker) within Peninsular Malaysia.

This study proposes to analyze the soil microbiome in the NSPSF, which is the largest remaining peatland of Peninsular Malaysia [1]. The NSPSF is located in northwestern part of the state of Selangor Darul Ehsan in Peninsular Malaysia, near the coastal town of Sekinchan. It spans approximately 81,304 hectares (Figure 1). As mentioned earlier, it was originally subjected to logging for timber products in the 1930s, and canals were excavated for drainage. However, it was officially gazetted as a protected area in the 1990s [1]. A few logging tracks and access roads cut through the reserve, many of which were originally built on mineral soils and bordered by drainage canals. To the west of the NSPSF lies a paddy field that receives its water from drainage originating in the surrounding peat swamp forest, and is part of the Tanjong Karang Irrigation Scheme, which spans the southwest and west of the reserve. Other surrounding land uses includes small-scale oil palm plantations, rehabilitated logging areas and protected forest areas under management by the Selangor Forestry Department.

The region experiences a typical humid equatorial climate, with a mean annual temperature of 27 °C, mean relative humidity of 79.3% and mean annual rainfall ranging from

1359 mm to 2480 mm [1]. The dominant vegetation across much of the NSPSF includes *Macaranga pruinosa* , *Campnosperma auriculatum*, *Stenochlaena palustris*, and *Pandanus* spp. [2], with the latter particularly prominent in wetter areas and at rehabilitated sites such as C007 1. The underlying soils are deep, ombrotrophic peat soils, formed from the accumulation of partially decomposed woody plant material. The maximum peat depth of 10.15 meters has been recorded at the site [1]. Peat swamps in Peninsular Malaysia, including the NSPSF, typically formed in former coastal lagoons and deltas created by rising sea levels after the last ice age around 10,000 years ago, with actual peat formation beginning approximately 5,000–7,000 years ago [3]. Along the sheltered west coast where the NSPSF is located, peat has developed primarily over fine clay deposits formed through sedimentation [1].

The NSPSF has been a site of research interest due to its critical role in understanding the impacts of land use change on GHG emissions [4]. Additionally, the NSPSF plays an important role in regulating water flow to nearby paddy fields, which affects agricultural productivity. Several areas experience occasional flooding, highlighting its significance in regional hydrological dynamics. However, other parts of peat soil of NSPSF are drier due to excessive drainage for agricultural purposes (oil palm, paddy). Furthermore, the NSPSF is recognized for its rich biodiversity, making it a valuable ecosystem for conservation efforts [5].

TABLE 1: Summary of site characteristics for peat samples collected in this study. Negative water table depth value indicates that the water level is above the soil surface.

Sample	**Coordinates**	**Sample collection date**	**Peat depth (cm)**	**Water peat depth (cm)**	**Distance from forest margin (km)**
C001	3.63 N, 101.34 E	2nd April 2023	412	-30	4.02
C002	3.53 N, 101.18 E	21st March 2023	213	-10	1.50
C003	3.69 N, 101.10 E	15th April 2023	211	>100	0.80
C007	3.43 N, 101.34 E	29th April 2023	305	0	3.15

0.2 Permits

The NSPSF is under the jurisdiction of the Selangor Forestry Department and research within the area requires permits. We obtained a permit from the Selangor State Forestry Department (Permit Number JH/100 Jld. 31 (59)) to conduct this research. Entry into the forest also required permits from the relevant districts and were obtained from the Klang District Forestry Office and the Rawang District Forestry Office. Additionally,

in accordance to the Nagoya Protocol, biological resources require the application for a permit under the Access to Biological and Benefit Sharing (ABS) Act 2017, Government of Malaysia. An application was submitted under the myABS portal on November 2022 and is still under review (Reference number: 246499).

0.3 Soil Sample Collection

This study focuses on four secondary peat swamp forest sites (C001, C002, C003, and C007) within the North Selangor Peat Swamp Forest (NSPSF), a lowland tropical peatland ecosystem on the west coast of Peninsular Malaysia (Figure 1). Each site was located at least 100 meters from the forest edge to minimize edge effects. Sites C001 and C007 are situated further inland, whereas C002 and C003 are located closer to the coastline. Site-specific descriptions are provided below to contextualize their environmental settings and potential influences on biogeochemical processes and microbial communities.

Site C001 is located inland near Sungai Tengi. Access was possible via a cleared path used by 4WD vehicles, which ascends a small hill and runs parallel to the river. However, to reach the actual sampling location, we had to exit the vehicle and proceed on foot through the forest. The soil surface at the site was consistently inundated, indicating a high water table and saturated conditions that support anaerobic processes.

Site C002, located nearer to the coast, required a 1.5-kilometer walk along a wide (5–10 m) cleared path built on mineral soil, historically used for logging operations. Small canals run along both sides of the path and were used for timber transport. For sampling, we left the path, crossed a canal, and walked approximately 5–10 meters into the forest to collect soil samples. The water table at this site was relatively high at the time of sampling.

Site C003, also coastal, lies adjacent to a small oil palm plantation. Reaching the sampling location involved walking through the plantation and into the forest. In contrast to other sites, the peat soil here was notably dry. Although we had previously installed perforated PVC piezometers, on the day of sampling the water level had dropped below the bottom of the pipe, indicating significantly reduced water retention and a lowered water table.

Site C007 represents a rehabilitated area that previously underwent logging and experienced forest fires. Restoration measures have included blocking the drainage canals, resulting in persistently high water levels and a water table at or above the soil surface. The area is now dominated by dense Pandanus vegetation, and the site remains saturated year-round.

Peat soil samples were collected using an Eijkelkamp peat auger and each sample was sectioned to 10 cm depths. Samples were collected until the mineral clay substrate was

reached. Perforated polyvinyl chloride pipes were installed in auger holes down at least 1 meter depth for groundwater level measurements using a measuring tape.

Aliquots of samples for genomic research were transported back to the Monash University Malaysia laboratory in a cold chain within 24 hours and subsequently stored at -20 ℃ until use. The remaining samples were transported under surrounding ambient conditions for bulk density and other measurements [6].

Peat soils at the center of the peat forest is supposedly deeper [1], but these sites were inaccessible due to practical and logistical challenges and due to the health and safety concerns.

0.4 DNA Extraction

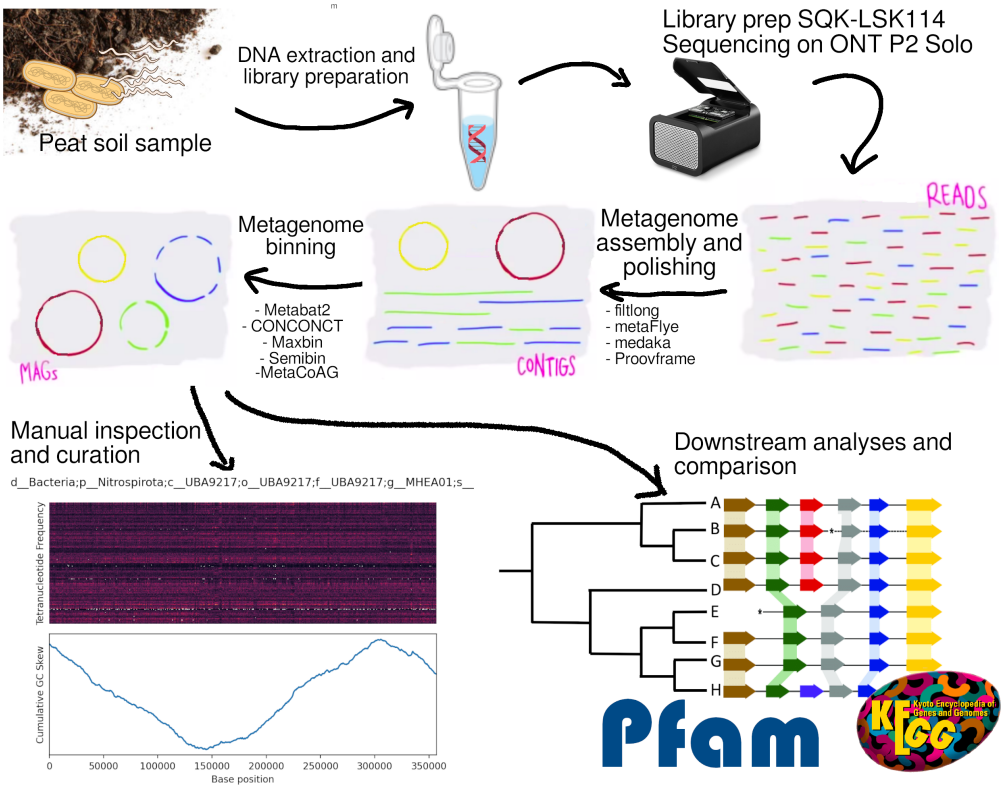

FIGURE 2: Flowchart summarizing the workflow DNA extraction, sequencing on Oxford Nanopore P2 Solo platform and bioinformatics analyses. Manual curation and inspection included genome circulation and reorientation, and display of tetranucleotide frequency and GC skew.

Total genomic DNA was extracted from peat soil samples using a modified traditional method [7]. Initially, the soil sample was flash frozen with liquid nitrogen and crushed using a mortar and pestle. The ground soil was then combined with a lysis buffer containing 1.5 M NaCl, 100 mM Tris-HCl at pH 8, 25 mM EDTA, and 3% (v/v) β-mercaptoethanol. The resulting extract was then incubated at 60°C for 3 hours. After

centrifugation at maximum speed for 10 minutes, the supernatant was collected and subjected to two rounds of purification using an equal volume of chloroform:isoamyl alcohol. The DNA was then precipitated by adding three volumes of absolute ethanol and subsequently incubated overnight at -20°C. The DNA precipitate was recovered through multiple rounds of centrifugation (20,000 g for 20 minutes) to precipitate the DNA through the whole volume. The DNA was then subsequently eluted in EB buffer. To further purify the crude DNA extract, multiple cleanup steps were performed using 0.5X AMPure XP beads. The concentration and quality of DNA was evaluated using the NanoDrop 2000 Spectrophotometer (Thermo Fisher, USA) and the Qubit dsDNA HS kit (Thermo Fisher Scientific, USA) with a Qubit 2.0 fluorometer (Thermo Fisher Scientific, USA).

0.5 Library Preparation and Oxford Nanopore Sequencing

Sequencing library preparation was conducted for 20 soil samples DNA samples that passed quality control. DNA quality was assessed based on the following criteria: A260/A280 ratio between 1.8 and 2.0, A260/A230 ratio above 2.0, total DNA yield exceeding 1 µg, and a Nanodrop to Qubit concentration ratio between 1.0 and 1.5, in accordance to Oxford Nanopore sample requirement. For library preparation, the SQK-LSK114 Ligation Sequencing kit (Oxford Nanopore Technologies, UK) following the manufacturer's protocols. Briefly, high-quality genomic DNA was first subjected to end-repair using the NEB End Repair Module (New England Biolabs, USA) to generate blunt-ended fragments suitable for adapter ligation. Then, it was ligated to ONT sequencing adapters. After ligation, a clean-up step using AMPure XP beads was performed to remove unligated adapters and short fragments. The resulting library was quantified to assess both the yield and quality prior to sequencing. The DNA library was then loaded into FLO-PRO114M Nanopore flow cells (Oxford Nanopore Technologies, UK) and sequenced in 400 bps sequencing speed mode using the Oxford Nanopore P2 Solo sequencers (Oxford Nanopore Technologies, UK). Real-time data quality monitoring was performed using the Oxford Nanopore MinKNOW software.

0.6 Sequencing Data Analysis

Most parts of bioinformatics analyses require heavy computational power and were performed using the M3 MASSIVE high-performance computer cluster in Victoria, Australia, as well as the Advanced Computing Platform in Monash University Malaysia.

The raw Nanopore pod5 sequencing data was processed in a series of processing steps to generate a high-quality assembly. Initially, the data was basecalled using Dorado 0.7.0 in super-accurate mode v5.0.0 (https://community.nanoporetech.com/downloads).

Sequencing reads with a Qscore below 8 were then filtered out using seqkit v2.6.1 [8]. The filtered sequences were assembled using metaFlye version 2.9.3-b1797 [9] with "–nano-hq" options to produce a flye assembly. Subsequently, the quality-filtered reads were aligned to the flye assembly using minimap2 version 2.28-r1209 [10]. The flye assembly and the resulting read alignments were then utilized to generate a polished medaka assembly, employing the medaka model r1041_e82_400bps_sup_v5.0.0 (https://community.nanoporetech.com/downloads). Finally, frame-shift correction was performed using proovframe v0.9.7 [11], resulting in the final proovframe assembly.

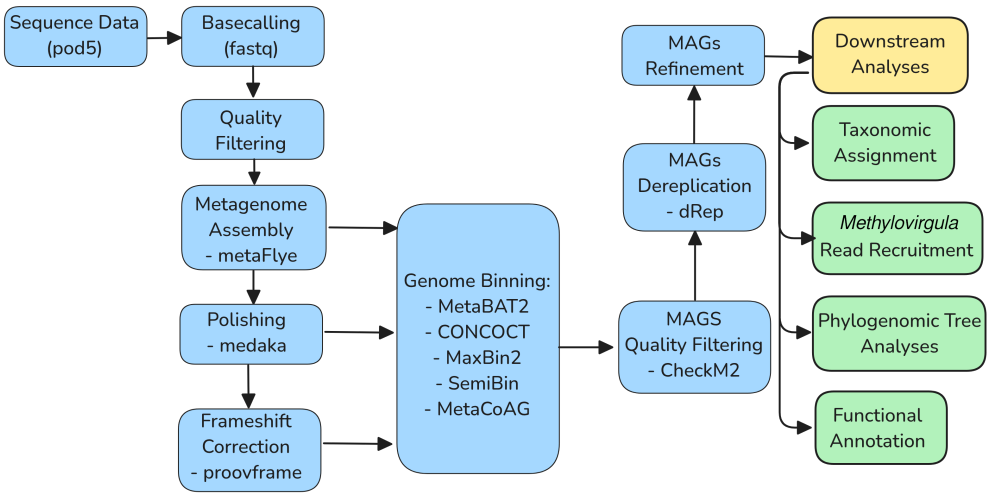

FIGURE 3: Flowchart summarizing the bioinformatics workflow, from sequence basecalling, metagenome assembly, genome binning and downstream analyses.

Using the three different assemblies (flye, medaka and proovframe), the read depth values per contig were calculated using minimap2 version 2.28-r1209 [10] and samtools version 1.19.1 [12]. To take advantage of the differential abundance information, we leverage the fact that for each site, there are five depths for five samples. While considering the genome assembly of a sample, read depths were calculated from all five samples of the corresponding site. Contig depth values was calculated using jgi_summarize_bam_contig_depths function in MetaBAT2 version 2.17 [13].

Co-assembly of shotgun metagenomes involves combining sequencing data from multiple related samples to produce a single assembly. This approach improves genome recovery, especially for low-abundance microbes and for highly diverse microbiome, like in soil. This is also practical if a microbe is present across different depths of a site; the genomic detection of this microbe can be enhanced using this strategy. However, the coassembly from multiple samples was not possible due to computational limitations.

0.7 Generating Metagenome-Assembled Genomes

Metagenome-assembled genomes (MAGs) are predicted genomes reconstructed from environmental samples by binning contigs based on genomic features like GC skew, tetranucleotide frequency, and differential abundance. For each of flye, medaka and proovframe assemblies and intermediate files, preliminary MAGs were generated using multiple binning softwares namely MetaBAT2 version 2.17 [13], CONCOCT version 1.1.0 [14], MaxBin2 version 2.2.7 [15], SemiBin version 2.0.2 [16] and MetaCoAG version 1.1.4[17]. For every MAGs generated, genome completeness and contamination was estimated using CheckM2 [18], which is a software based on pre-trained machine learning. Completeness and contamination of a MAG is estimated based on the presence, absence, and redundancy of lineage-specific SCGs that are expected to occur once per genome in closely related taxa.

According to Bowers et al. [19], a medium-quality MAG possesses a completeness of at least 50% and contamination of less than 10%. Meanwhile, a high-quality MAG possesses a completeness of at least 90%, contamination of less than 5% as well as the complete rRNA operon containing the 25S, 16S and 5S rRNA genes. Only MAGs of at least medium quality (at least 50% completeness and less than 10% contamination) were considered for downstream analyses. Afterwards, MAGs were further clustered using dRep version 3.4.5 [20] at (99% gANI). Scoring criteria was applied on each cluster using the following formula:

$$score = completion - 5(contamination) + log(contigN50) \tag{1}$$

From each cluster, the MAG of the highest score was selected as the final MAG from that cluster to produce the a set of non-redundant MAGs. Taxonomic classification of each MAG was done using GTDBtk version 2.4.0 [21] on GTDB release 220 [22]. GTDB Taxonomic classification was based on the placement of genome into a backbone bacterial or archaeal tree using pplacer version 1.1 [23] as well as average nucleotide identity (ANI) criteria using skani version 0.2.2 [24].

0.8 Refinement of Metagenome-Assembled Genomes

To further refine MAG assemblies, each preliminary MAG was subjected to reassembly, refinement, and curation. Sequencing reads from all twenty samples were first mapped to each MAG using Minimap2 v2.28-r1209 [10], applying a minimum identity threshold of 95% and a minimum alignment coverage of 80%. Mapped reads were then reassembled using two targeted assembly tools: metaFlye v2.9.3-b1797 [9] with the `--nano-hq` flag, and hifiasm version 0.24.0 [25] with parameters `--ont --primary --n-hap 1`.

Subsequently, read coverage across contigs was calculated using CoverM version 0.7.0 [26]. Complete MAGs were selected based on criteria such as full circularization or contigs exceeding 100 kb in length containing single-copy genes (SCGs). Contigs exhibiting anomalous read coverage were excluded from the final set of MAGs.

Contig selection from each assembly method were selected based on assembly graph visualization using Bandage [27]. However, visualization of individual graphs and capability of integrating of external data to screen through numerous assemblies to compare different assemblies is laborious. We designed a software enabling seamless direct comparison of assembly graphs to screen through MAGs, which will enable MAG inspection and curation and facilitate further decisions whether the MAG should be reassemble or fit for downstream analyses. The preliminary software is available through the GitHub link (https://github.com/ZarulHanifah/MAGGFA).

0.9 Phylogenomic Analysis

Phylogenomic analysis was conducted using GToTree version 1.8.8 [28]. Briefly, single copy genes (SCGs) of target taxa were extracted from each MAG using prodigal version 2.6.3 [29] and HMMER version 3.4 [30]. Gene sequences were aligned using Muscle version 5.1 [31], and spurious sites were removed using TrimAL version 1.5 [32]. Phylogenetic tree were constructed using FastTree 2 version 2.1.11 and validate using bootstrapping [33]. The phylogenomic trees was displayed using iTOL version 7.0 [34]. Bootstrap values of above 70%, indicative of strong support values, were displayed on the phylogenomic tree internal nodes.

0.10 Functional Annotation

Genome annotation was performed using Bakta version 1.10.3 [35]. Gene calling was done on pyrodigal version 3.5.0 and pyHMMER version 0.10.15 [36]. tRNAs were predicted using tRNAscan-SE version 2.0.11 [37]. Functional annotations were done against multiple databases namely UniProt [38], RefSeq [39], COG [40], KEGG KOFAM [41] and pFAM [42]. This was further supplemented with functional predictions from DRAM version 1.5.0 [43], which is built on databases UniRef90 [38], pFAM [42], dbCAN [44], RefSeq [39] and MEROPS peptidase [45] databases. 16S sequences were extracted from MAGs using barrnap (https://github.com/tseemann/barrnap).

0.11 *Methylovirgula* in Global Peat Metagenome Datasets

Due to the prevalence and abundance of *Methylovirgula* in the NSPSF dataset, we proceeded to look into its global distribution in publicly available global peat metagenome sequences (Table 2). Read recruitment was conducted using bowtie2 version 2.5.4 and samtools version 1.19.1 [12] against *Methylovirgula* genomes from GTDB as well as those newly generated from this study.

TABLE 2: Description of extended global peat metagenome dataset.

Research paper	Study site
Bahram et al. [46]	Peat soil sample collection from many different countries, including Maludam National Park in Sarawak.
Bandla et al. [47]	Oil palm plantation, Jambi, Indonesia.
Pavia et al. [48]	Pastaza Maranon Foreland Basin, Peruvian Amazon.
Piatkowski et al. [49]	Peat moss of Iceland, Sweden and France.
Romanowicz et al. [50]	Tussock and wet sedge tundra, Toolik Lake, Alaska, USA.
ter Horst et al. [51]	Boreal peatlands, northern Minnesota, USA. Part of SPRUCE project.
Trubl et al. [52]	Arctic peat soils from Bonanza Creek Long-Term ecological research site, Alaska, USA.
Tveit et al. [53]	Temperate gradient Arctic peat.
Woodcroft et al. [54]	Permafrost thaw gradient, Stordalen Mire, Sweden.
Midot et al. [55]	Bornean tropical peatlands, Sarawak, Malaysia. Peat Transition from secondary forest to oil palm plantation.

The insertion sequences in *Methylovirgula* sp. HY1 was predicted using ISEScan verson 1.7.3 [56].

0.12 Data Visualization

Python version 3.12.8 was the main language for most of the analyses performed in Jupyter Lab (`https://github.com/jupyterlab/jupyterlab`). Tabular manipulation was performed using pandas version 2.2.3 [57]. Data visualization was conducted using matplotlib version 3.9.3 [58], seaborn version 0.13.2 [59] and plotly version 5.24.1 [60].

Bibliography

[1] Centre G. E . *Integrated Management Plan for North Selangor Peat Swamp Forest 2014-2023*. Selangor State Forestry Department, 2014. ISBN 9789671026854.

[2] Yule C. M , Lim Y. Y , and Lim T. Y . Degradation of tropical malaysian peatlands decreases levels of phenolics in soil and in leaves of macaranga pruinosa. *Frontiers in Earth Science*, 4, April 2016. ISSN 2296-6463. doi: 10.3389/feart.2016.00045. URL http://dx.doi.org/10.3389/feart.2016.00045.

[3] Weiss D , Shotyk W , Rieley J , Page S , Gloor M , Reese S , and Martinez-Cortizas A . The geochemistry of major and selected trace elements in a forested peat bog, kalimantan, se asia, and its implications for past atmospheric dust deposition. *Geochimica et Cosmochimica Acta*, 66(13):2307–2323, July 2002. ISSN 0016-7037. doi: 10.1016/s0016-7037(02)00834-7. URL http://dx.doi.org/10.1016/s0016-7037(02)00834-7.

[4] Cooper H. V , Vane C. H , Evers S , Aplin P , Girkin N. T , and Sjögersten S . From peat swamp forest to oil palm plantations: The stability of tropical peatland carbon. *Geoderma*, 342:109–117, May 2019. ISSN 0016-7061. doi: 10.1016/j.geoderma.2019.02.021. URL http://dx.doi.org/10.1016/j.geoderma.2019.02.021.

[5] Beamish F. W. H , Beamish R. B , and Lim S. L.-H . Fish assemblages and habitat in a malaysian blackwater peat swamp. *Environmental Biology of Fishes*, 68(1):1–13, September 2003. ISSN 1573-5133. doi: 10.1023/a:1026004315978. URL http://dx.doi.org/10.1023/A:1026004315978.

[6] Melling L . *Peatland in Malaysia*, page 59–73. Springer Japan, 2016. ISBN 9784431556817. doi: 10.1007/978-4-431-55681-7_4. URL http://dx.doi.org/10.1007/978-4-431-55681-7_4.

[7] Sambrook J and Russell D. W . *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 3rd edition, 2001.

[8] Shen W , Le S , Li Y , and Hu F . Seqkit: A cross-platform and ultrafast toolkit for fasta/q file manipulation. *PLOS ONE*, 11(10):e0163962, October 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0163962. URL http://dx.doi.org/10.1371/journal.pone.0163962.

[9] Kolmogorov M , Yuan J , Lin Y , and Pevzner P. A . Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, 37(5):540–546, April 2019. ISSN 1546-1696. doi: 10.1038/s41587-019-0072-8. URL http://dx.doi.org/10.1038/s41587-019-0072-8.

[10] Li H . Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34 (18):3094–3100, May 2018. ISSN 1367-4811. doi: 10.1093/bioinformatics/bty191. URL http://dx.doi.org/10.1093/bioinformatics/bty191.

[11] Hackl T , Trigodet F , Eren A. M , Biller S. J , Eppley J. M , Luo E , Burger A , DeLong E. F , and Fischer M. G . proovframe: frameshift-correction for long-read (meta)genomics. August 2021. doi: 10.1101/2021.08.23.457338. URL http://dx.doi.org/10.1101/2021.08.23.457338.

[12] Danecek P , Bonfield J. K , Liddle J , Marshall J , Ohan V , Pollard M. O , Whitwham A , Keane T , McCarthy S. A , Davies R. M , and Li H . Twelve years of samtools and bcftools. *GigaScience*, 10(2), January 2021. ISSN 2047-217X. doi: 10.1093/gigascience/giab008. URL http://dx.doi.org/10.1093/gigascience/giab008.

[13] Kang D. D , Li F , Kirton E , Thomas A , Egan R , An H , and Wang Z . Metabat 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7:e7359, July 2019. ISSN 2167-8359. doi: 10.7717/peerj.7359. URL http://dx.doi.org/10.7717/peerj.7359.

[14] Alneberg J , Bjarnason B. S , Bruijn I de, Schirmer M , Quick J , Ijaz U. Z , Lahti L , Loman N. J , Andersson A. F , and Quince C . Binning metagenomic contigs by coverage and composition. *Nature Methods*, 11(11):1144–1146, September 2014. ISSN 1548-7105. doi: 10.1038/nmeth.3103. URL http://dx.doi.org/10.1038/nmeth.3103.

[15] Wu Y.-W , Simmons B. A , and Singer S. W . Maxbin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, 32(4):605–607, October 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv638. URL http://dx.doi.org/10.1093/bioinformatics/btv638.

[16] Pan S , Zhu C , Zhao X.-M , and Coelho L. P . A deep siamese neural network improves metagenome-assembled genomes in microbiome datasets across different environments. *Nature Communications*, 13(1), April 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-29843-y. URL http://dx.doi.org/10.1038/s41467-022-29843-y.

[17] Mallawaarachchi V and Lin Y . Accurate binning of metagenomic contigs using composition, coverage, and assembly graphs. *Journal of Computational Biology*, 29 (12):1357–1376, December 2022. ISSN 1557-8666. doi: 10.1089/cmb.2022.0262. URL http://dx.doi.org/10.1089/cmb.2022.0262.

[18] Chklovski A , Parks D. H , Woodcroft B. J , and Tyson G. W . Checkm2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nature Methods*, 20(8):1203–1212, July 2023. ISSN 1548-7105. doi: 10.1038/s41592-023-01940-w. URL http://dx.doi.org/10.1038/s41592-023-01940-w.

[19] Bowers R. M , Kyrpides N. C , Stepanauskas R , Harmon-Smith M , Doud D , Reddy T. B. K , Schulz F , Jarett J , Rivers A. R , Eloe-Fadrosh E. A , Tringe S. G , Ivanova N. N , Copeland A , Clum A , Becraft E. D , Malmstrom R. R , Birren B , Podar M , Bork P , Weinstock G. M , Garrity G. M , Dodsworth J. A , Yooseph S , Sutton G , Glöckner F. O , Gilbert J. A , Nelson W. C , Hallam S. J , Jungbluth S. P , Ettema T. J. G , Tighe S , Konstantinidis K. T , Liu W.-T , Baker B. J , Rattei T , Eisen J. A , Hedlund B , McMahon K. D , Fierer N , Knight R , Finn R , Cochrane G , Karsch-Mizrachi I , Tyson G. W , Rinke C , Lapidus A , Meyer F , Yilmaz P , Parks D. H , Murat Eren A , Schriml L , Banfield J. F , Hugenholtz P , and Woyke T . Minimum information about a single amplified genome (misag) and a metagenome-assembled genome (mimag) of bacteria and archaea. *Nature Biotechnology*, 35(8):725–731, August 2017. ISSN 1546-1696. doi: 10.1038/nbt.3893. URL http://dx.doi.org/10.1038/nbt.3893.

[20] Olm M. R , Brown C. T , Brooks B , and Banfield J. F . drep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *The ISME Journal*, 11(12):2864–2868, July 2017. ISSN 1751-7370. doi: 10.1038/ismej.2017.126. URL http://dx.doi.org/10.1038/ismej.2017.126.

[21] Chaumeil P.-A , Mussig A. J , Hugenholtz P , and Parks D. H . Gtdb-tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics*, 38(23):5315–5316, October 2022. ISSN 1367-4811. doi: 10.1093/bioinformatics/btac672. URL http://dx.doi.org/10.1093/bioinformatics/btac672.

[22] Parks D. H , Chuvochina M , Chaumeil P.-A , Rinke C , Mussig A. J , and Hugenholtz P . A complete domain-to-species taxonomy for bacteria and archaea. *Nature Biotechnology*, 38(9):1079–1086, April 2020. ISSN 1546-1696. doi: 10.1038/s41587-020-0501-8. URL http://dx.doi.org/10.1038/s41587-020-0501-8.

[23] Matsen F. A , Kodner R. B , and Armbrust E. V . pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, 11(1), October 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-538. URL http://dx.doi.org/10.1186/1471-2105-11-538.

[24] Shaw J and Yu Y. W . Fast and robust metagenomic sequence comparison through sparse chaining with skani. *Nature Methods*, 20(11):1661–1665, September 2023. ISSN 1548-7105. doi: 10.1038/s41592-023-02018-3. URL http://dx.doi.org/10.1038/s41592-023-02018-3.

[25] Cheng H , Concepcion G. T , Feng X , Zhang H , and Li H . Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, 18(2): 170–175, February 2021. ISSN 1548-7105. doi: 10.1038/s41592-020-01056-5. URL http://dx.doi.org/10.1038/s41592-020-01056-5.

[26] Aroney S. T. N , Newell R. J. P , Nissen J. N , Camargo A. P , Tyson G. W , and Woodcroft B. J . Coverm: read alignment statistics for metagenomics. *Bioinformatics*, 41(4), March 2025. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaf147. URL http://dx.doi.org/10.1093/bioinformatics/btaf147.

[27] Wick R. R , Schultz M. B , Zobel J , and Holt K. E . Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*, 31(20):3350–3352, June 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv383. URL http://dx.doi.org/10.1093/bioinformatics/btv383.

[28] Lee M. D . Gtotree: a user-friendly workflow for phylogenomics. *Bioinformatics*, 35 (20):4162–4164, March 2019. ISSN 1367-4811. doi: 10.1093/bioinformatics/btz188. URL http://dx.doi.org/10.1093/bioinformatics/btz188.

[29] Hyatt D , Chen G.-L , LoCascio P. F , Land M. L , Larimer F. W , and Hauser L. J . Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1), March 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-119. URL http://dx.doi.org/10.1186/1471-2105-11-119.

[30] Eddy S. R . Accelerated profile hmm searches. *PLoS Computational Biology*, 7(10): e1002195, October 2011. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002195. URL http://dx.doi.org/10.1371/journal.pcbi.1002195.

[31] Edgar R. C . Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nature Communications*, 13 (1), November 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-34630-w. URL http://dx.doi.org/10.1038/s41467-022-34630-w.

[32] Capella-Gutiérrez S , Silla-Martínez J. M , and Gabaldón T . trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–1973, June 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp348. URL http://dx.doi.org/10.1093/bioinformatics/btp348.

[33] Price M. N , Dehal P. S , and Arkin A. P . Fasttree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE*, 5(3):e9490, March 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0009490. URL http://dx.doi.org/10.1371/journal.pone.0009490.

[34] Letunic I and Bork P . Interactive tree of life (itol) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Research*, 52(W1):

W78–W82, April 2024. ISSN 1362-4962. doi: 10.1093/nar/gkae268. URL http://dx.doi.org/10.1093/nar/gkae268.

[35] Schwengers O , Jelonek L , Dieckmann M. A , Beyvers S , Blom J , and Goesmann A . Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification: Find out more about bakta, the motivation, challenges and applications, here. *Microbial Genomics*, 7(11), November 2021. ISSN 2057-5858. doi: 10.1099/mgen.0.000685. URL http://dx.doi.org/10.1099/mgen.0.000685.

[36] Larralde M . Pyrodigal: Python bindings and interface to prodigal, an efficient method for gene prediction in prokaryotes. *Journal of Open Source Software*, 7 (72):4296, April 2022. ISSN 2475-9066. doi: 10.21105/joss.04296. URL http://dx.doi.org/10.21105/joss.04296.

[37] Chan P , Lin B , Mak A , and Lowe T . trnascan-se 2.0: improved detection and functional classification of transfer rna genes. *Nucleic Acids Research*, 49 (16):9077–9096, August 2021. ISSN 1362-4962. doi: 10.1093/nar/gkab688. URL http://dx.doi.org/10.1093/nar/gkab688.

[38] Consortium U . Uniprot: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, November 2018. ISSN 1362-4962. doi: 10.1093/nar/gky1049. URL http://dx.doi.org/10.1093/nar/gky1049.

[39] Haft D. H , DiCuccio M , Badretdin A , Brover V , Chetvernin V , O'Neill K , Li W , Chitsaz F , Derbyshire M. K , Gonzales N. R , Gwadz M , Lu F , Marchler G. H , Song J. S , Thanki N , Yamashita R. A , Zheng C , Thibaud-Nissen F , Geer L. Y , Marchler-Bauer A , and Pruitt K. D . Refseq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Research*, 46(D1): D851–D860, November 2017. ISSN 1362-4962. doi: 10.1093/nar/gkx1068. URL http://dx.doi.org/10.1093/nar/gkx1068.

[40] Galperin M. Y , Kristensen D. M , Makarova K. S , Wolf Y. I , and Koonin E. V . Microbial genome analysis: the cog approach. *Briefings in Bioinformatics*, 20 (4):1063–1070, September 2017. ISSN 1477-4054. doi: 10.1093/bib/bbx117. URL http://dx.doi.org/10.1093/bib/bbx117.

[41] Aramaki T , Blanc-Mathieu R , Endo H , Ohkubo K , Kanehisa M , Goto S , and Ogata H . Kofamkoala: Kegg ortholog assignment based on profile hmm and adaptive score threshold. *Bioinformatics*, 36(7):2251–2252, November 2019. ISSN 1367-4811. doi: 10.1093/bioinformatics/btz859. URL http://dx.doi.org/10.1093/bioinformatics/btz859.

[42] El-Gebali S , Mistry J , Bateman A , Eddy S. R , Luciani A , Potter S. C , Qureshi M , Richardson L. J , Salazar G. A , Smart A , Sonnhammer E. L , Hirsh L , Paladin L , Piovesan D , Tosatto S. C , and Finn R. D . The pfam protein families database

in 2019. *Nucleic Acids Research*, 47(D1):D427–D432, October 2018. ISSN 1362-4962. doi: 10.1093/nar/gky995. URL http://dx.doi.org/10.1093/nar/gky995.

[43] Shaffer M , Borton M. A , McGivern B. B , Zayed A. A , La Rosa S , Solden L. M , Liu P , Narrowe A. B , Rodríguez-Ramos J , Bolduc B , Gazitúa M. C , Daly R. A , Smith G. J , Vik D. R , Pope P. B , Sullivan M. B , Roux S , and Wrighton K . Dram for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Research*, 48(16):8883–8900, August 2020. ISSN 1362-4962. doi: 10.1093/nar/gkaa621. URL http://dx.doi.org/10.1093/nar/gkaa621.

[44] Zheng J , Ge Q , Yan Y , Zhang X , Huang L , and Yin Y . dbcan3: automated carbohydrate-active enzyme and substrate annotation. *Nucleic Acids Research*, 51 (W1):W115–W121, May 2023. ISSN 1362-4962. doi: 10.1093/nar/gkad328. URL http://dx.doi.org/10.1093/nar/gkad328.

[45] Rawlings N. D , Barrett A. J , Thomas P. D , Huang X , Bateman A , and Finn R. D . The merops database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the panther database. *Nucleic Acids Research*, 46(D1):D624–D632, November 2017. ISSN 1362-4962. doi: 10.1093/nar/gkx1134. URL http://dx.doi.org/10.1093/nar/gkx1134.

[46] Bahram M , Espenberg M , Pärn J , Lehtovirta-Morley L , Anslan S , Kasak K , Kõljalg U , Liira J , Maddison M , Moora M , Niinemets , Öpik M , Pärtel M , Soosaar K , Zobel M , Hildebrand F , Tedersoo L , and Mander . Structure and function of the soil microbiome underlying n2o emissions from global wetlands. *Nature Communications*, 13(1), March 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-29161-3. URL http://dx.doi.org/10.1038/s41467-022-29161-3.

[47] Bandla A , Mukhopadhyay S , Mishra S , Sudarshan A. S , and Swarup S . Genome-resolved carbon processing potential of tropical peat microbiomes from an oil palm plantation. *Scientific Data*, 10(1), June 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02267-z. URL http://dx.doi.org/10.1038/s41597-023-02267-z.

[48] Pavia M. J , Finn D , Macedo-Tafur F , Tello-Espinoza R , Penaccio C , Bouskill N , and Cadillo-Quiroz H . Genes and genome-resolved metagenomics reveal the microbial functional make up of amazon peatlands under geochemical gradients. *Environmental Microbiology*, 25(11):2388–2403, July 2023. ISSN 1462-2920. doi: 10.1111/1462-2920.16469. URL http://dx.doi.org/10.1111/1462-2920.16469.

[49] Piatkowski B. T , Carper D. L , Carrell A. A , Chen I.-M. A , Clum A , Daum C , Eloe-Fadrosh E. A , Gilbert D , Granath G , Huntemann M , Jawdy S. S , Klarenberg I. J , Kostka J. E , Kyrpides N. C , Lawrence T. J , Mukherjee S , Nilsson M. B , Palaniappan K , Pelletier D. A , Pennacchio C , Reddy T. B. K , Roux S , Shaw A. J , Warshan D , Živković T , and Weston D. J . Draft metagenome sequences of the sphagnum (peat moss) microbiome from ambient

and warmed environments across europe. *Microbiology Resource Announcements*, 11(10), October 2022. ISSN 2576-098X. doi: 10.1128/mra.00400-22. URL http://dx.doi.org/10.1128/mra.00400-22.

[50] Romanowicz K. J , Crump B. C , and Kling G. W . Rainfall alters permafrost soil redox conditions, but meta-omics show divergent microbial community responses by tundra type in the arctic. *Soil Systems*, 5(1):17, March 2021. ISSN 2571-8789. doi: 10.3390/soilsystems5010017. URL http://dx.doi.org/10.3390/soilsystems5010017.

[51] Horst A. M ter, Santos-Medellín C , Sorensen J. W , Zinke L. A , Wilson R. M , Johnston E. R , Trubl G , Pett-Ridge J , Blazewicz S. J , Hanson P. J , Chanton J. P , Schadt C. W , Kostka J. E , and Emerson J. B . Minnesota peat viromes reveal terrestrial and aquatic niche partitioning for local and global viral populations. *Microbiome*, 9(1), November 2021. ISSN 2049-2618. doi: 10.1186/s40168-021-01156-0. URL http://dx.doi.org/10.1186/s40168-021-01156-0.

[52] Trubl G , Kimbrel J. A , Liquet-Gonzalez J , Nuccio E. E , Weber P. K , Pett-Ridge J , Jansson J. K , Waldrop M. P , and Blazewicz S. J . Active virus-host interactions at sub-freezing temperatures in arctic peat soil. *Microbiome*, 9(1), October 2021. ISSN 2049-2618. doi: 10.1186/s40168-021-01154-2. URL http://dx.doi.org/10.1186/s40168-021-01154-2.

[53] Tveit A. T , Urich T , Frenzel P , and Svenning M. M . Metabolic and trophic interactions modulate methane production by arctic peat microbiota in response to warming. *Proceedings of the National Academy of Sciences*, 112(19), April 2015. ISSN 1091-6490. doi: 10.1073/pnas.1420797112. URL http://dx.doi.org/10.1073/pnas.1420797112.

[54] Woodcroft B. J , Singleton C. M , Boyd J. A , Evans P. N , Emerson J. B , Zayed A. A. F , Hoelzle R. D , Lamberton T. O , McCalley C. K , Hodgkins S. B , Wilson R. M , Purvine S. O , Nicora C. D , Li C , Frolking S , Chanton J. P , Crill P. M , Saleska S. R , Rich V. I , and Tyson G. W . Genome-centric view of carbon processing in thawing permafrost. *Nature*, 560(7716):49–54, July 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0338-1. URL http://dx.doi.org/10.1038/s41586-018-0338-1.

[55] Midot F , Goh K. M , Liew K. J , Lau S. Y. L , Espenberg M , Mander , and Melling L . Temporal dynamics of soil microbial c and n cycles with ghg fluxes in the transition from tropical peatland forest to oil palm plantation. *Applied and Environmental Microbiology*, 91(1), January 2025. ISSN 1098-5336. doi: 10.1128/aem.01986-24. URL http://dx.doi.org/10.1128/aem.01986-24.

[56] Xie Z and Tang H . Isescan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics*, 33(21):3340–3347, July 2017. ISSN 1367-4811. doi: 10.1093/bioinformatics/btx433. URL http://dx.doi.org/10.1093/bioinformatics/btx433.

[57] team T pandas development. pandas-dev/pandas: Pandas, February 2020. URL https://doi.org/10.5281/zenodo.3509134.

[58] Hunter J. D . Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

[59] Waskom M . seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, April 2021. ISSN 2475-9066. doi: 10.21105/joss.03021. URL http://dx.doi.org/10.21105/joss.03021.

[60] Inc. P. T . Collaborative data science, 2015. URL https://plot.ly.