Data in Brief 36 (2021) 107124

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data article

# Metagenomic datasets of air samples collected during episodes of severe smoke-haze in Malaysia

Grace Lymie James[a], Mohd Talib Latif[c], Mohd Noor Mat Isa[b], Mohd Faizal Abu Bakar[b], Nurul Yuziana Mohd Yusuf[c], William Broughton[d], Abdul Munir Murad[a], Farah Diba Abu Bakar[a,*]

[a] *Department of Biological Sciences and Biotechnology, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, UKM, Bangi 43600, Selangor, Malaysia*
[b] *Malaysia Genome Institute, Ministry of Science, Technology and Innovation, Jalan Bangi, 43000 Kajang, Selangor, Malaysia*
[c] *Department of Earth Sciences and Environment, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, UKM, Bangi 43600, Selangor, Malaysia*
[d] *Department 4 (Materials & Environment), Federal Institute of Materials Research and Testing, Berlin, Germany*

ARTICLE INFO

*Article history:*
Received 10 March 2021
Revised 27 April 2021
Accepted 29 April 2021
Available online 9 May 2021

*Keywords:*
Forest fires
Haze samples
NGS
Multiple Displacement Amplification

ABSTRACT

Transboundary emissions of smoke-haze from land and forest fires have recurred annually during the dry period (June to October, over the past few decades) in South East Asia. Hazardous air quality has been recorded in Malaysia during these episodes. Agricultural practices such as slash-and-burn of biomass and peat fires particularly in Sumatera and Kalimantan, Indonesia, have been implicated as the major causes of the haze. Past findings have shown that a diversity of microbes can thrive in air including in smoke-haze polluted air. In this study, metagenomic data were generated to reveal the diversity of microorganisms in air during days with and without haze. Air samples were collected during non-haze (2013A01) and two haze (2013A04 and 2013A05) periods in the month of June 2013. DNA was extracted from the samples, subjected to Multiple Displacement Amplification and whole genome sequencing (Next Generation Sequencing) using the HiSeq 2000 Platform. Extensive

* Corresponding author.
*E-mail address:* fabyff@ukm.edu.my (F.D. Abu Bakar).
*Social media:* (F.D. Abu Bakar)

https://doi.org/10.1016/j.dib.2021.107124

bio-informatic analyses of the raw sequence data then followed. Raw reads from these six air samples were deposited in the NCBI SRA databases under Bioproject PRJNA662021 with accession numbers SRX9087478, SRX9087479 and SRX9087480.



## Specifications Table

| | |
|---|---|
| Subject | Biology |
| Specific subject area | Environmental Biology |
| Type of data | Metagenomic Data, Tables, Figures |
| How data were acquired | Paired-end metagenomes of air samples were sequenced using Illumina HiSeq 2000 at the Malaysia Genome Institute (MGI). De novo metagenome assembly was performed using the Velvet 1.2.09 and MetaVelvet V1.2.01 assemblers. Taxonomic analyses were performed using MEGAN 6.19.9 and visualised using Krona (https://github.com/marbl/Krona). |
| Data format | Raw (FASTQ) and analysed sequences |
| Parameters for data collection | Samples were collected from the roof top of a meteorological station with different air qualities (polluted air quality during haze days and good air quality during non-haze days) based on the $PM_{10}$ pollution concentration, in June 2013. |
| Data source location | City/Town/Region: Petaling Jaya<br>Country: Malaysia |
| Data accessibility | Raw FASTQ files were deposited in the NCBI SRA database under BioProject PRJNA662021. The accession number for the samples are SRX9087478, SRX9087479 and SRX9087480.<br>https://www.ncbi.nlm.nih.gov/sra/PRJNA662021 |

## Value of the Data

- Metagenomic sequences collected from Petaling Jaya, Malaysia, in conjunction with the recurring forest fire induced transboundary smoke-emission episodes in South East Asia, provide evidence of the microbial load, types and diversity transported by this smoke and haze. These data are the first metagenomic haze samples in South East Asia obtained during haze periods caused by the burning of biomass. Of particular concern are the potential pathogens that may spread by convection due to the burning of diseased flora (and fauna). In this respect, the agriculture and health sectors could be under threat if circumventing measures are not undertaken.
- The data can be used as a reference by environmental biologists/microbiologists to investigate microbial community structures in haze tainted air brought about by the burning of biomass.
- The data contain DNA sequences from various types of organisms (prokaryotes and eukaryotes; some of which are extremophiles and potential pathogens). They may also be used to discover novel genes/sequences that may be beneficial in other applications including industry.

## 1. Data Description

The data describe the microbial diversity in the air samples collected during days with and without haze. Metagenome data of sample 2013A01 ($PM_{10}$: 58.50 μg/$m^3$) describe the

**Table 1**
Meteorological conditions during sampling and average concentrations of suspended particulate, $PM_{10}$ pollutants of air samples collected in 2013.

| Sample | 2013A01 | 2013A04 | 2013A05 |
|---|---|---|---|
| Weather conditions | No haze | Hazy | Hazy |
| Suspended particulate, $PM_{10}$ concentration (μg/m$^3$) | 58.5 | 287.9 | 244.2 |
| Date sample collected | 17.06.13 | 23.06.13 | 24.06.13 |
| Average daily wind speed (km/h) | 4.5 | 4.1 | 4.4 |
| Average daily humidity (%) | 57 | 54 | 53 |
| Average daily ambient temperature (°C) | 30.0 | 30.6 | 30.9 |

**Table 2**
Statistical analyses of DNA sequences generated from three libraries.

| Samples | Total reads | Total clean reads | Total Paired- reads | Total Singletons | Average read length (bp) | Total # base pair (bp) |
|---|---|---|---|---|---|---|
| **2013A01** | 33,499,214 | 17,466,247 | 13,128,848 | 4337,399 | 89 | 1555,667,282 |
| **2013A04** | 28,870,012 | 11,122,757 | 7477,502 | 3645,255 | 89 | 994,285,720 |
| **2013A05** | 31,255,894 | 20,814,137 | 16,155,962 | 4658,175 | 89 | 1859,371,625 |

**Table 3**
Statistical analyses of DNA sequences assembled using Velvet and MetaVelvet.

| Sample | 2013A01 | 2013A04 | 2013A05 |
|---|---|---|---|
| **Number of scaffolds** | 1045 | 864 | 3994 |
| **Total Scaffold Length (bp)** | 701,711 | 540,296 | 3279,821 |
| **Longest Length (bp)** | 9483 | 10,572 | 20,391 |
| **Average scaffold length (bp)** | 671 | 625 | 821 |
| **N50 size** | 953 | 924 | 1574 |

diversity of microbes in the sample without haze, whilst metagenome data of sample 2013A04 and 2013A05 show the microbial diversity during days with haze where each sample contained $PM_{10}$ pollutant concentrations of 287.90 μg/m$^3$ and 244.20 μg/m$^3$, respectively (Table 1). The wind speeds during sampling were 4.5 km/h, 4.1 km/h and 4.4 km/h, for samples 2013A01, 2013A04 and 2013A05, respectively, whilst average daily temperatures (and humidity) were 30 °C (57%), 30.6 °C (54%) and 30.9 °C (53%), respectively (Table 1).

Table 2 describes the statistical analyses of DNA sequences generated from three libraries, 2013A01, 2013A04 and 2013A05. A total of 33,499,214 reads were generated from the sample without haze (2013A01), whilst 28,870,012 and 31,255,894 reads were generated from haze samples 2013A04 and 2013A05, respectively (Table 2).

The number of scaffolds generated after assembly of the reads for the three air samples were 1045 (2013A01), 864 (2013A04) and 3994 (2013A05) (Table 3). The assembled reads were then used in taxonomic analyses. The data presented here focus on the taxonomic profile of bacteria and fungi found in all three samples.

Taxonomic analysis using MEGAN software generated 1045, 864 and 3994 reads from samples 2013A01, 2013A04 and 2013A05, respectively. More than 70% of the reads from each sample did not produce hits against the NCBI taxonomic tree in MEGAN 6.19.9 software (Fig. 1). A total of 16.9%, 10.2% and 22.5% of the reads of sample 2013A01, 2013A04 and 2013A05, respectively, were assigned to a bacterial domain (Fig. 1). Three reads were unassigned to the taxonomic tree for sample 2013A01 while no reads were unassigned for samples 2013A04 and 2013A05.

Figs. 2–4 show the taxonomic distribution of microorganism for samples 2013A01, 2013A04 and 2013A05, respectively. Most of the bacteria found in the sample without haze (2013A01) were Firmicutes (52% of total assigned reads, 114 reads), Proteobacteria (19% of total assigned reads, 42 reads) and Actinobacteria (5% of total assigned reads, 11 reads) (Fig. 2). Whilst 4%


**Fig. 1.** Percent (numbers within bars) distribution of total assigned reads after taxonomic analyses of each air sample using MEGAN version 6.19.9.

of the total assigned reads (8 reads) belonged to the fungal phyla Ascomycota and 1% of total assigned reads (3 reads) belonged to Basidiomycota (Fig. 2).

Totals of 20% (31 reads), 17% (26 reads) and 16% (25 reads) of the assigned reads in the air sample with haze, 2013A04, assigned to Proteobacteria, Firmicutes and Actinobacteria, respectively (Fig. 3). Most of the fungi found in the sample belonged to the phylum Ascomycota (8% of total assigned reads, 12 reads), followed by Basidiomycota (0.6% of total assigned reads, 1 read) (Fig. 3). A total of 448 (39% of total assigned reads), 215 (19% of total assigned reads) and 167 reads (14% of total assigned reads) in the air sample with haze, 2013A05, belonged to the phyla Firmicutes, Actinobacteria and Proteobacteria, respectively (Fig. 4). In addition, 5% (56 reads) and 4% (46 reads) of total reads in the sample were assigned to the fungal phyla Ascomycota and Basidiomycota, respectively (Fig. 4). These data are the first to showcase the microbial profile in air contaminated with the haze brought about by burnt biomass in South East Asia.

## 2. Experimental Design, Materials and Method

### *2.1. Sampling of haze samples*

Three air samples were collected by the Malaysian Meteorology Department at the following coordinates - latitude 3°6′, 0″ U, longitude 101°39′, 0″ T. Meteorological conditions were also recorded by the Malaysian Meteorological Department whereby wind speeds were determined using Met-One 010C sensor (Met-One Instrument, Inc., USA), and ambient temperatures as well as relative humidity were measured using Met-One 062 and Met-One 083D sensor (Met-One Instrument, Inc., USA), respectively. The samples were collected on 8 × 10 inches Whatman® glass microfibre filters with a pore size 10 µm. Before sampling, the filter papers were placed in a drying chamber at 25 °C to 30 °C with relative humidity of 40% for 24 h to remove moisture. Weight differences of filter papers before and after exposure were measured to determine the gravimetric mass concentrations in deriving the $PM_{10}$ values. A high-volume sampler

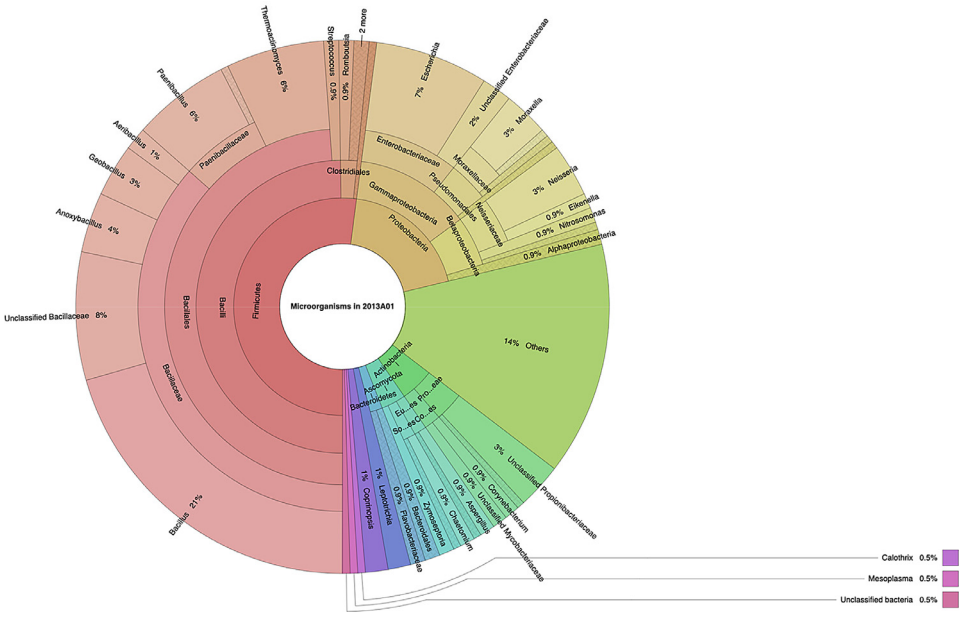


**Fig. 2.** Taxonomic classification of bacteria and fungi in the non-haze sample, 2013A01 visualised using a Krona chart.

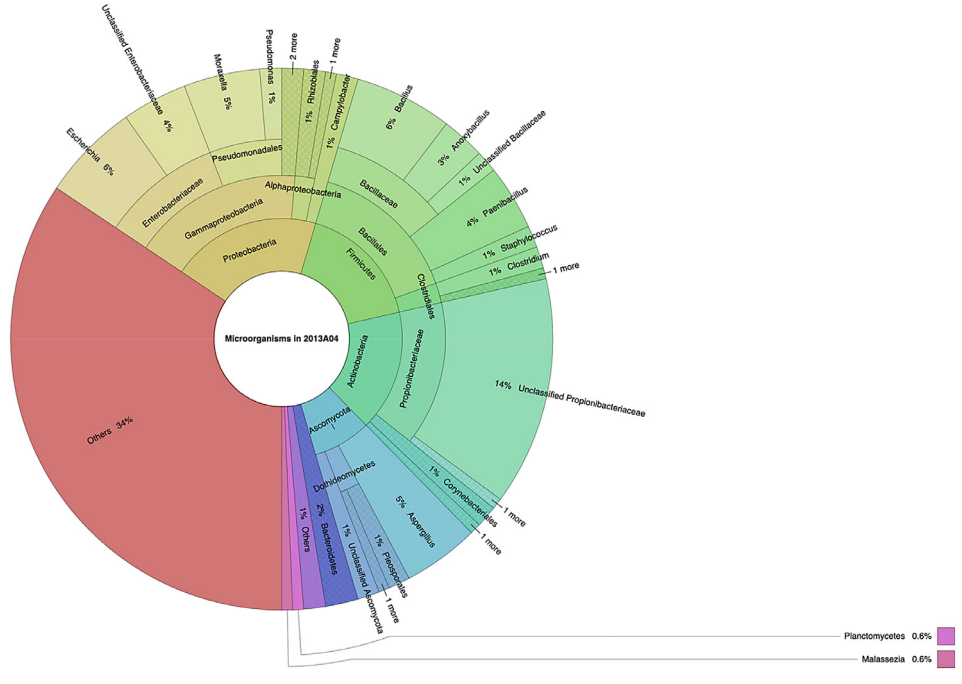


**Fig. 3.** Taxonomic classification of bacteria and fungi in the haze sample, 2013A04, visualised using a Krona chart.

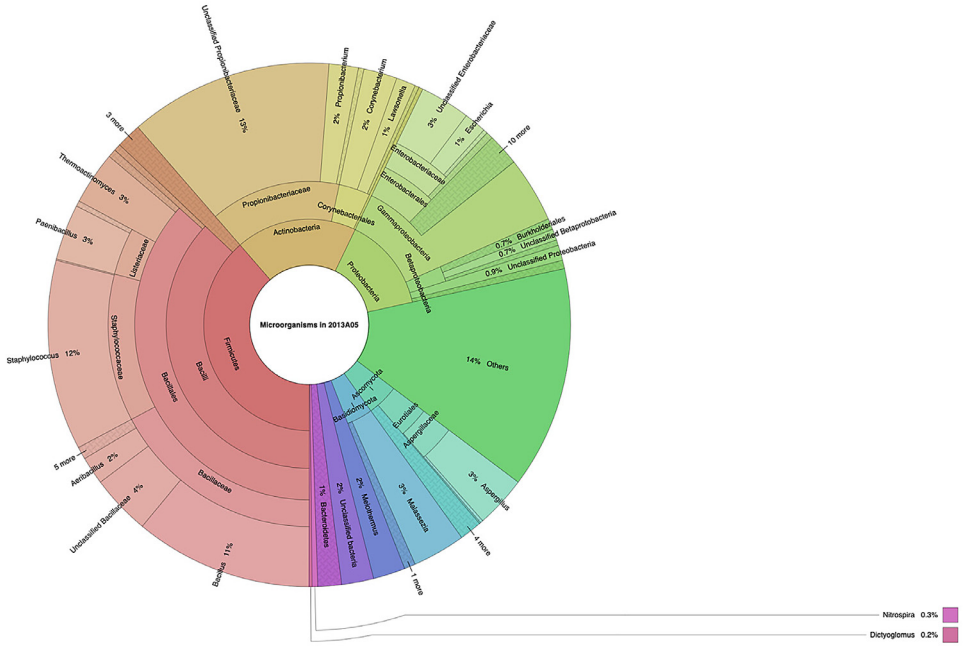


**Fig. 4.** Taxonomic classification of bacteria and fungi found in the haze sample, 2013A05, visualised using a Krona chart.

(Sierra-Andersen/GMW Model 1200) installed 58.6 M MSL with an air flow of 1.13 $m^3$ $min^{-1}$ for 24 h was used to collect samples. Samples were stored at −80 °C until use.

### *2.2. DNA extraction, Multiple Displacement Amplification (MDA) and sequencing*

Metagenomic DNA was extracted using the PowerWater® DNA Isolation Kit (MO BIO Laboratories Inc., Carlsbad, CA, USA). The filter papers used to collect the samples and were cut into smaller pieces (~ 0.3 $cm^2$) and placed in the beaded tubes provided in the kit. Extraction was carried out following the manufacturer's protocols. The concentration of DNA extracted was quantified using a spectrophotometer. Isolated DNA was then subjected to Multiple Displacement Amplification [4] to increase the concentration of isolated DNA using REPLI-g Single Cell (Qiagen, Venlo, The Netherlands) kits. Metagenomic DNA library preparation was carried out using the TrueSeq DNA PCR Free sample preparation kit (Illumina, https://www.illumina.com) as per the manufacturer's instructions. Prepared library samples of 10 pM were used in paired end (2 × 100 bp) sequencing that was carried out on the HiSeq 2000 system.

### *2.3. De-novo assembly and DNA-seq analysis*

The raw metagenomic DNA sequences were trimmed, cleaned and filtered using SolexaQA [3]. The parameters for Phred quality were set to $Q_{20}$, while the read length was set to 50 bp. Paired-end reads were determined using Perl script select_paired.pl. Statistical analyses of the metagenomic DNA sequences are shown in Table 2. De-novo assembly of metagenomic DNA sequences were then performed using Velvet V1.2.09 and MetaVelvet V1.2.01 [2,5] with default parameters. The statistical analyses of DNA sequences are shown in Table 3.

After alignment, the assembled reads were matched with the sequences available in the NCBI (National Centre of Biotechnology Information) Nucleotide Collection (nr/nt) database using BLASTN software (https://www.ncbi.nlm.nih.gov/). The best matching reads with E-values less than 1.0 $E^{-5}$ were used for further analyses. Next, BLASTN.xml files generated after comparitive sequencing were used for taxonomic analyses using MEGAN version 6.19.9 [1] with the default parameter and the LCA algorithm set to naive mode (Fig. 1). These taxonomic analyses were then visualised using Krona (https://github.com/marbl/Krona) (Figs. 2–4).

## CRediT Author Statement

**Grace Lymie James:** Writing - Original Draft, Investigation, Visualization; **Mohd Talib Latif:** Supervision, Validation, Resources; **Mohd Noor Mat Isa:** Resources, Validation; **Mohd Faizal Abu Bakar:** Software, Resources, Validation, Visualization; **Nurul Yuziana Mohd Yusuf:** Resources; **William Broughton:** Conceptualization, Manuscript Review & Editing; **Munir Abdul Murad:** Supervision, Validation; **Farah Diba Abu Bakar:** Conceptualization, Supervision, Methodology, Validation, Manuscript Review & Editing.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relation-ships that could have appeared to influence the work reported in this paper.

## Acknowledgment

This research was funded by Ministry of Higher Education, Malaysia FRGS/1/2014/SG03/UKM/02/3.

## References

[1] D.H. Huson, A.F. Auch, J. Qi, S.C. Schuster, MEGAN analysis of metagenomic data, Genome Res. 17 (3) (2007) 377–386, doi:10.1101/gr/5969107.
[2] D.R. Zerbino, E. Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs, Genome Res. 18 (2008) 821–829, doi:10.1101/gr.074492.107.
[3] M.P. Cox, D.A. Peterson, P.J. Biggs, SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data, BMC Bioinform. 11 (2010) 485, doi:10.1186/1471-2105-11-485.
[4] S. Yilmaz, M. Allgaier, P. Hugenholtz, Multiple displacement amplification compromises quantitative analysis of metagenomes, Nat. Methods 7 (12) (2010) 943, doi:10.1038/nmeth1210-943.
[5] T. Namiki, T. Hachiya, H. Tanaka, Y. Sakakibara, MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads, Nucl. Acids Res 40 (20) (2012) e155, doi:10.1093/nar/gks678.