AJLS

Asean Journal of Life Sciences, Vol 2 (2), 2022, pp. 13-21

Accepted date: 30 October 2022
Publish date: 28 December 2022

Soil Metagenomics Analysis Reveals the Compositions of Bacteria Communities in The World's Oldest Tropical Rainforest

Mohd Fadzli Ahmad[a]*, Nor Suhaila Yaacob[b,c], Hasdianty Abdullah[a,b], Naim Hassan[a] , Imran Jamaludin[a], Irni Suhayu Sapian[d], Halimah Alias[d], Mohd Noor Mat Isa[d],.

[a] University Selangor, Faculty of Engineering & Life Sciences, Department of Science & Biotechnology, 45600 Bestari Jaya, Selangor, Malaysia.
[b] Institute of Bio-IT Selangor, Universiti Selangor, Jalan Zirkon A7/A, Seksyen 7, 40000, Shah Alam, Selangor, Malaysia
[c] Centre for Foundation and General Studies, Universiti Selangor, Jalan Zirkon A7/A, Seksyen 7, 40000, Shah Alam, Selangor, Malaysia
[d]Malaysia Genome Institute, National Institute of Biotechnology Malaysia, Jalan Bangi, 43000 Kajang Selangor, Malaysia

shuhaila@unisel.edu.my

Abstract

This study was initiated to investigate the complex of soil microbiomes that resides in three different pristine ecosystems in the world's oldest tropical rainforest, Royal Belum Reserved Forest (RB). 27 soil samples from 3 sites have been collected and classified as Lithosol formed on loess material. The composition of the bacterial communities was determined via Shotgun Whole Genome Sequencing (WGS) by Illumina technology and undergone data pre-processing that includes alignments and quality trimming with Solexa QA++, contigs assembly with Metaspades before functionally analyzed by MgRast server in order to extract the genetic repertoire of the microbiomes. Each gene sequence that consists of the whole genome data was assigned to its specific roles and functions and clustered into OTU based on a 99% similarity threshold. The RB soils tested were distinctly dominated by α-Proteobacteria made up of 30.57% abundance in Sungai Kooi (SK), 34.75% in Sungai Papan (SP), and 26.14% in Sungai Ruok (SR). Bacteria from Actinobacteria class and Solibacteres class were subdominants in every sample but slightly different in genus level where SK top 3 most abundant genus other than *Candidatus solibacter* and *Candidatus koribacter* were *Bradyrhizobium* (6.42%), *Streptomyces* (4.59%) and *Acidobacterium* (3.59%) while SP shows *Mycobacterium* appearance for 4.26% abundances. For SR samples, the *Variovorax, Chthoniobacter* and *Bradyrhizobium* were among of the 5 most abundant genus with 6.75%, 4.26% and 3.38%

respectively. A total of 620 different bacterial species from 269 genus were detected, of which 588 were present in all three sites. Among all the genus found relatively, SK showed higher statistical differences compared to other sites except for *Acidovorax, Delftia, Polaromonas, Stenotrophomonas, Variovorax, Xanthomonas, Xylella* and *Methylophilales* which were more abundance in SR sample.

Keyword: Next-Generation Sequencing; Pristine Rainforest; Soil metagenomes; Royal Belum

INTRODUCTION

The Royal Belum forest is the world's largest and most significant natural ecosystems and habitats to in situ conservation of biodiversity (Lazarus et al., 2019). The study of the content of bacterial species that have lived in symbiosis for millions of years and formed a balanced community in the soil of this conserved area is very interesting and significantly important because the ability of these microbes to transform the biosphere around them. Soils contain some of the most diverse microbiomes on Earth and are important for their working, it is important to model global patterns of distribution and functional gene repertoire of soil microorganisms, as well as environmental relations between the diversity and composition of soil populations (Bahram, et al., 2018). Most soil bacteria do not fit those found in pre-existing 16S ribosomal RNA (rRNA) gene databases because of very little genomic details and most soil bacteria have not been successfully grown in vitro (Schloss, et al., 2016; Lok, 2015). For these reasons, there is a lack of predictive understanding of the ecological properties of most soil bacterial taxa, with their environmental preferences, characteristics, and metabolic capabilities still largely unknown. Due to this matter, the study of environmental DNA has been revolutionized by adopting Shotgun Whole Genome Sequencing technology that be able to use direct environmental sample extraction technique and thorough genome characterization. Through this metagenomics approach DNA of microbial community can be access directly to the fully extend including the uncultivated majority. In Malaysia, there is very limited knowledge about the types of symbiotic and nonsymbiotic soil bacteria inhabiting forest soils. The research data available nowadays is mainly limited to non-tropical environments and constraints only to the culturable bacteria. Therefore, we have tried to provide new knowledge about the biodiversity of bacteria that resides in Malaysian oldest pristine rainforest soils. To answer this question, we used Shotgun Whole Genome Sequencing (WGS) technology whereby it became possible to omit the inefficient laboratory culture step and acquire knowledge about the enormous microbial groups termed as viable but not cultivable (VBNC).

METHODOLOGY

Experimental Sites

The study site was in the northern part of Malaysia in Perak State where soils were taken in May 2018. This forest is the largest continuous forest complex in Peninsular Malaysia covering an area of 117,500 hectares of thick forest stretching into the Thailand – Malaysia border as presented in Fig 1. The area has high rainfall of approximately 2,560 mm/year with the average temperature at the collection time was 28–30℃ and the pH value of the sample was between 4.42 to 6.12. The total soil samples collected were 27 from three distinct locations where nine replicates of soil

samples per site (1 kg each) were randomly collected from top surface sediments to a depth of 5 cm. The samples contained mainly soil and plant biomass residues were mixed to represent one site and used for all experiments in this study. Samples were stored at −80°C until DNA extraction was performed and preserved accordingly to maintain both quality and accuracy of metagenomics data. Site A soil sample has been collected at Sungai Kooi (SK) while site B is located at Sungai Papan (SP) and site C soil sample was located at Sungai Ruok (SR). The major coordinates and elevation for these three sites are as presented in Table. 1.

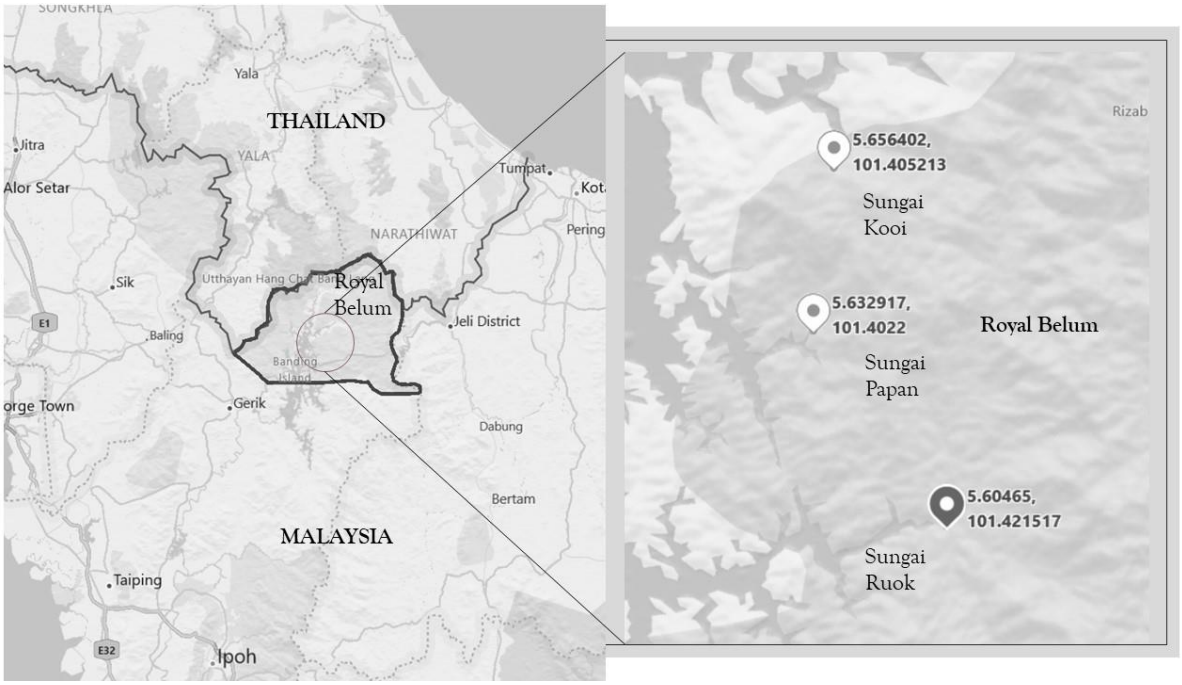

Figure 1 Location of the study site in Royal Belum Reserved Forest (RB)

Table 1 The coordinates and elevation for Royal Belum Reserved Forest Sampling Sites
(a) Sungai Kooi, (b) Sungai Papan and (c) Sungai Ruok

Sites	**Type of soil**	**GPS**	**Lat, Long**	**Elevation (m)**
Sungai Kooi	Lithosol and shallow red yellow	5°39'23.1"N 101°24'18.8"E	5.656402, 101.405213	345
Sungai Papan	Lithosol and shallow red yellow	5°37'58.5"N 101°24'07.9"E	5.6329167,101.4022	345
Sungai Ruok	Lithosol and shallow red yellow	5°36'16.7"N 101°25'17.5"E	5.604650, 101.421517	540

Soils Bacterial Metagenome Extraction

The DNA samples extracted within 24 hours after sample collection according to the modified procedure elaborated for soil material by using DNeasy PowerSoil Kit and undergone three quality controls (QC) procedure by Nanodrop spectrophotometer at UV length A260/280, gel electrophoresis profiling and Qubit fluorescent dye to ensure the intactness of the DNA samples (Desjardins & Conklin, 2010). DNA samples that passed the QC were purified by using magnetic beats and agent court to remove tannin and undergone mechanical shearing procedure by using Covaris instrument to make sure all the DNA samples length are between 300 bp to 350 bp measured by Bioanalyzer (Oberacker et al., 2019; Kong et al., 2014). Samples have been measured by Qubits to ensure the total recovery after purification is more than 70% and the total DNA per 50 µl volume is 1 µg before the adenylation of 3' end. Individual barcode of index adapter sequences was added to each DNA fragment during library preparation so that each read can be identified and sorted before the final data analysis in flow cells chip. Samples were sequenced by using Illumina NGS Paired-end technology (Sevim et al., 2019) and reads were align and gone through preprocessing phase in bioinformatics pipeline that includes alignments and quality trimming with Solexa QA++ (Cox et al., 2010) contigs assembly with Metaspades (Nurk et al., 2017) and finally analyzed by using Megan6 (Huson et al., 2016) and MgRast (Meyer et al., 2008) to determine genetics repertoire of the microbiome.

Gene Annotation and Sequence Analysis

The raw sequence quality was checked and the reads were trimmed accordingly by using FastQC (Version 0.11.5 released); it is a tool provided by Babraham Institute which makes the quality control of high-throughput sequencing pipelines an easy matter. In Command Line Interface (CLI), SolexaQA++ was used (Boetzer et al., 2011) in DynamicTrim application where sequences have been trimmed based on Qphred <20 and in LengthSort command sequences that shorter than 50 bp have been removed. The cleaned sequences were paired together and shuffled to produce high quality sequences before assembled using the metaSPAdes (version 3.13.0). The high-quality sequences were mapped to the final assembly and coverage information was generated using bbmap (Version 38.25) using default parameters, with the exception of ambiguous = random. Diamond software was used to blast predicted genes against the nonredundant protein sequences database of NCBI (https://www.ncbi.nlm.nih.gov/) with default settings (Buchfink et al., 2015), using BLASTP (best hit with $E < 0.001$). Functional annotation was conducted by aligning sequencing reads against KEGG database (Release 84.1) (Kanehisa & Goto, 2000) using MEGAN6 software (Version 6.11.7) (Huison et al., 2016) with the parameter setting of blastp (Buchfink et al., 2015) based on the LCA algorithm. Finally, the gene read numbers for each sample were normalized based on median read number. The relative abundances (percentage) of genes were calculated related to the annotated reads and used for subsequent analyses. The M5 non-redundant protein database (M5NR) was used for taxonomic annotation and the SEED database and Clusters of Orthologous Groups database for functional annotation. To identify the sequences, the best BLASTx hit was used with a minimum alignment length of 15 bp and an e-value cut-off of $e < 1 \times 10{-}5$ and 95% confidence interval. Functional annotation of the most abundant taxa was performed using the filter option. The same analysis was done for select group of genes to reveal the responsible taxa. The shotgun metagenomics sequence data used in this study are deposited in the MG-RAST server (Meyer et al., 2008) under project ID mgp94971 and mgp94737.

RESULTS AND DISCUSSION

The Shotgun WGS sequencing yields were 69,378,986 sequences from SK, 48,671,928 sequences from SP, 67,850,530 sequences from SR, 33,119,856 sequences for RM1, 48,593,274 sequences for RM2 and 31,164,018 sequences for RM3. The average lengths were 344 bp, 357 bp, 283 bp, 481 bp, 517 bp and 354 bp respectively, giving an overall average of 389 bp and a total of 298,778,592 bp. The metagenomes data represent thousands of species of bacteria, belonging to many phyla. The retained sequences were Blast against NCBI-nr database using Megan 6 and shows that around 80% of the total retained sequences from Royal Belum data sets had matches against the NCBI-nr database (614,523 matches for the SK sample, 516,589 for the SP sample and 252,255 matches for the SR sample), whereas about 80% to 86% of the total Raja Muda Musa sequences were significantly assigned, with 389,195 matches for RM1, 2,277,578 matches for RM2 and 321,571 matches for RM3. Megan conducted a process where reads were compared with database of known sequence (NCBI-nr). The program parses files generated by Blastx and saved the result as a series of read-taxon matches in a specific metafile. Based on the lowest common ancestor (LCA) algorithm, the metafiles that contained read-taxon matches have been plotted into taxonomy. This approach consists of several thresholds where the minimum alignment score was set at 50.0 according to the length of reads and the top percent filter was assigned to 10. The top percent filer was best set around 10 to 20 that make it more specific during taxa assignment. The result of Shotgun WGS raw sequence, recovery sequences after pre-processing, total sequences assigned, predicted protein features, total number of bases and α diversity of each sample are as presented in Table 2.

Table 2 Samples and number of raw and filtered reads obtained for each sample. For each sample, the number of raw reads and the numbers of reads surviving each processing step is indicated along with predicted protein features, total number of contigs, total bases of each sample and α diversity measurements. The percentages in bracket indicate the numbers of reads after each step relative to the number of raw reads.

Site	Raw Sequences	Pre-processing (Trimmed Sequences)	Predicted Protein Features	Total of Contigs (sequence)	Total No of Bases (bp) (Mg-RAST)	α diversity (Mg-RAST)
Sungai Kooi	69378986	51028698	765717	774188	226513266	352
Sungai Papan	48671928	44735602	631602	646194	230634531	395
Sungai Ruok	67850530	54502982	295742	303564	85911188	388

Taxonomic comparisons between Royal Belum Selected Sites

The analysis of the taxonomic community showed that Royal Belum soils were dominated by Bacteria (96% in Sungai Kooi, 98% in Sungai Papan, and 99 % in Sungai Ruok). The remaining sequences matched with the Archaea (0.47% in Sungai Kooi, 0.42% in Sungai Papan, and 0.47% in Sungai Ruok), and the remaining were classed under Eukaryota and Viruses. The bacterial composition of samples from SK, SP, and SR was further investigated using the MgRast server where a total of 27 phyla within the bacterial domain were detected among the three samples namely Acidobacteria, Actinobacteria, Aquificae, Bacteroidetes, Candidatus Poribacteria, Chlamydiae, Chlorobi, Chloroflexi, Chrysiogenetes, Cyanobacteria, Deferribacteres, Deinococcus-Thermus, Dictyoglomi, Elusimicrobia, Fibrobacteres, Firmicutes, Fusobacteria,

Gemmatimonadetes, Lentisphaerae, Nitrospirae, Planctomycetes, Proteobacteria, Spirochaetes, Synergistetes, Tenericutes, Thermotogae, and Verrucomicrobia.
The relative abundance of the phyla Proteobacteria, Actinobacteria, Cyanobacteria, Planctomycetes, and Bacteroidetes was significantly higher in SK compared to SP and SR. The phyla Chloroflexi was greater in relative abundance in SP compared to SK and SR. The other remaining phyla were not shown any significant difference in the gene abundances.

After identifying the dominant 27 phyla, we found that within the Proteobacteria, Alphaproteobacteria were relatively more abundant in the SP sample with 61.59%, while SR had more Betaproteobacteria with 28.35% and SK had more Deltaproteobacteria with 9.28% of abundances.

A total of 620 different bacterial species from 269 genus were detected, of which 588 were present in all three sites. Among all the genus found relatively SK showed higher statistical differences compared to other sites except for Acidovorax, Delftia, Polaromonas, Stenotrophomonas, Variovorax, Xanthomonas, Xylella and Methylophilales which were more abundant in SR sample.

From all 269 genus detected, Bradyrhizobium dominated the population with 13.54% in SP followed by SK 12.29% and SR 7.12%. The second largest genus found was Burkholderia and it consist of 6.56% in SP bacterial population followed by SK 6.04% and SR 5.46% (Fig2).

These findings are in line with previous research demonstrating that climatic factors and soil pH are often highly correlated with observed differences in overall soil bacterial community composition (Ramirez et al., 2014; Fierer et al., 2012; Zhou et al., 2016).

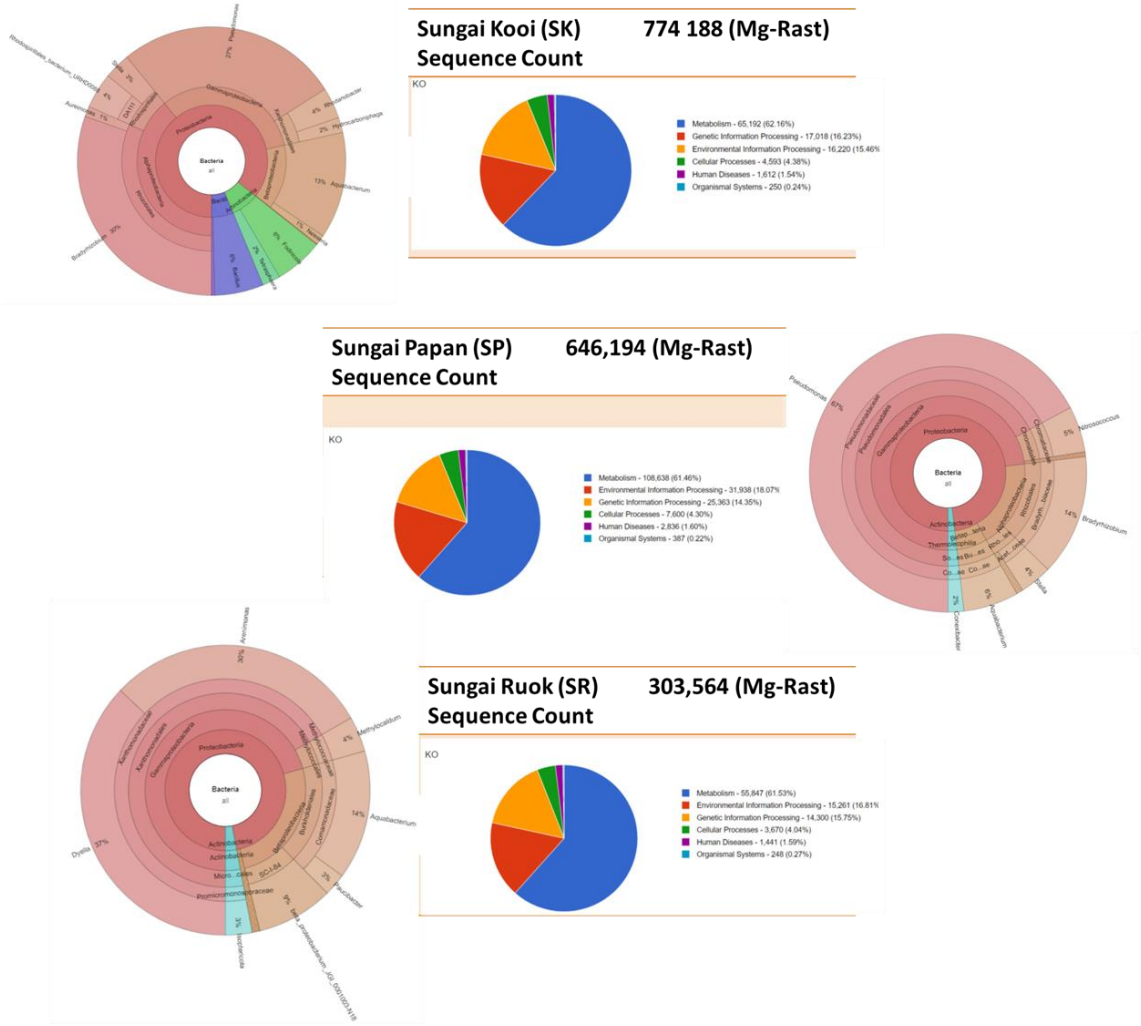

Figure 2 The Estimation of Bacteria Abundances and Their Genes Predicted Functions in Royal Belum Soils.

CONCLUSION

In this study, the bacterial composition of three different locations of Royal Belum Reserved Forest, Malaysia, was determined using Shotgun WGS metagenomics. We provided full details of bacteria structure based on metagenomic sequencing. The goal was to draw a correlation between the microbiota diversity within the pristine soil environment. The Royal Belum soils inhibited with a total of 620 different bacterial species from 269 genus and distinctly dominated by Proteobacteria at phylum level made up of 53%, followed by Actinobacteria 20% and Acidobacteria 12%. The most populous bacterial genus of this soil structure is *Bradyrhizobium* which dominates 6% of the entire community and is followed by *Candidatus Solibactor* by 5% as well as *Candidatus koribacter* by 4%.

ACKNOWLEDGEMENT

This research was supported by Japan Science and Technology Agency (JST)/Japan International Cooperation Agency (JICA), Science and Technology Research Partnership for Sustainable Development (SATREPS) through the project for Continuous Operation System for Microalgae Production Optimized for Sustainable Tropical Aquaculture (COSMOS), and the SATREPS-COSMOS Matching Fund from the Ministry of Higher Education Malaysia (MOHE). We also thank rangers at Forest Reserve to guide us to sampling points and lab members who support sampling as well sample analyses.

REFERENCES

Bahram, M., Hildebrand, F., Forslund, S.K. *et al.* (2018). Structure and function of the global topsoil microbiome. *Nature, 560,* 233–237. https://doi.org/10.1038/s41586-018-0386-6.

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. & Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics, 27*(4), 578-579. https://doi.org/10.1093/bioinformatics/btq683.

Buchfink, B., Chao, X. & Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat Methods, 12,* 59-60. https://doi.org/10.1038/nmeth.3176.

Cox, M.P., Peterson, D.A. & Biggs, P.J. (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics, 11*(485), https://doi.org/10.1186/1471-2105-11-485.

Desjardins, P., & Conklin, D. (2010). NanoDrop microvolume quantitation of nucleic acids. *Journal of visualized experiments, (*45), 2565. https://doi.org/10.3791/2565.

Fierer, N., Leff, J.W., Adams, B.J., Nielsen, U.N., Bates, S.T., Lauber, C.L., Owens, S., Gilbert, J.A., Wall, D.H., & Caporaso, J.G. (2012). Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *PNAS, 109*(52), 21390-21395. https://doi.org/10.1073/pnas.1215210110.

Huson, D., Beier, S., Flade, I., Gorska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H. & Rewati Tappu, D. (2016). MEGAN Community Edition - Interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Computational Biology, 12*(6), 1-12. https://doi.org/10.1371/journal.pcbi.1004957.

Jizhong, Z., Ye, D., Lina, S., Chongqing, W., Qingyun, Y., Daliang, N., Yujia, Q., Kai, X., Liyou, W., Zhili, H., Voordeckers, J. W., Nostrand, J. D. V., Buzzard, V., Michaletz, S.T., Enquist, B. J., Weiser, M. D., Kaspari, M., Waide, R., Yunfeng, Y., & Brown, J. H.Temperature mediates continental-scale diversity of microbes in forest soils. *Nature communications, 7*(12083). https:// doi.org/10.1038/ncomms12083.

Kanehisa, M. & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Ressearch, 28*(1), 27-30. https://doi.org/10.1093/nar/28.1.27.

Kong, Nguyet; Kao Thao; Carol Huang; Storey, Dylan; Weimer, Bart; Appel, Maryke; et al. (2015): Automated Library Construction Using KAPA Library Preparation Kits on the Agilent NGS Workstation Yields High-Quality Libraries for Whole-Genome Sequencing on the Illumina Platform. Journal contribution. 1-8. https://doi.org/10.6084/m9.figshare.1386854.v1.

Lazarus, B. A., M. Muzammil, Abdul Halim Shah, Azwan Hamdan, A. Najmi, Nik Hassan, M. S. Mohammad, H. Hassim, M. Noor, Tengku Rinalfi, Putra Tengku Azizan, & Hafandi Ahmad. (2020). Topographical differences impacting wildlife dynamics at natural saltlicks in the Royal Belum rainforest. *Asian Journal of Conservation Biology, 8*(2), 97-101. http://www.ajcb.in/journals/full_papers_dec_2019/AJCB-Vol8-No2-Lazarus%20et%20al.pdf.

Lok, C. (2015). Mining the microbial dark matter. *Nature, 522,* 270–273. https://doi.org/10.1038/522270a.

Meyer, F., Paarmann, D., D'Souza, M, Olson, R., Glass, E.M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J. & Edwards, R.A. (2008). The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics, 9*(386). https://doi.org/10.1186/1471-2105-9-386.

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome research*, *27*(5), 824–834. https://doi.org/10.1101/gr.213959.116.

Oberacker, P., Stepper, P., Bond, D. M., Höhn, S., Focken, J., Meyer, V., Schelle, L., Sugrue, V. J., Jeunen, G. J., Moser, T., Hore, S. R., von Meyenn, F., Hipp, K., Hore, T. A., & Jurkowski, T. P. (2019). Bio-On-Magnetic-Beads (BOMB): Open platform for high-throughput nucleic acid extraction and manipulation. *PLoS biology*, *17*(1), 1-16. https://doi.org/10.1371/journal.pbio.3000107.

Ramirez Kelly S., Leff Jonathan W., Barberán Albert, Bates Scott Thomas, Betley Jason, Crowther Thomas W., Kelly Eugene F., Oldfield Emily E., Shaw E. Ashley, Steenbock Christopher, Bradford Mark A., Wall Diana H. & Fierer Noah. (2014). Biogeographic patterns in below- ground diversity in New York City's Central Park are similar to those observed globally. Proceedings of the Royal Society B: Biological Sciences, 281(1795), http://doi.org/10.1098/rspb.2014.1988.

Schloss, P.D., Girard, R.A., Martin, T., Edwards, J., & Thrash, J.C. (2016). Status of the Archaeal and Bacterial Census: An Update. *mBio*, *7*(3), 1-10. https://doi.org/10.1128/mBio.00201-16.

Sevim, V., Lee, J., Egan, R. *et al.* (2019). Shotgun metagenome data of a defined mock community using Oxford Nanopore, PacBio and Illumina technologies. *Sci Data*, *6*(285), 1-9. https://doi.org/10.1038/s41597-019-0287.