Journal of Biotechnology and Natural Science
Vol. 4, No. 1, June 2024, pp. 47-54
ISSN 2808-2508

Establishing Geographical Indicator of Fermented Cacao Beans Using Microbiome Fingerprinting

[1,2*]Imam Bagus Nugroho, [3]Abdul Rahman Siregar

[1]Bioindustry Lab., Department of Agro-Industrial Technology, Faculty of Agricultural Technology, Universitas Gadjah Mada
[2]Genetic Engineering Lab., Research Center for Biotechnology, Universitas Gadjah Mada
[3]Microbiology Lab., Department of Tropical Biology, Faculty of Biology, Universitas Gadjah Mada
[1]imam.bagus.n@ugm.ac.id*; [3] ar.siregar@ugm.ac.id.
*Corresponding author

ARTICLE INFO

Article history
Received: June 12nd 2024
Revised: June 30th 2024
Accepted: June 30th 2024

Keywords
Bioinformatics
Clustering
PCA
UMAP
KNN

ABSTRACT

Geographical indication is an essential label for industrial products. Herein, we aimed to explore a method for establishing geographical indications based on microbial diversity data. We collected two groups of datasets available on the public server of the European Nucleotide Archive. These datasets contain 12 (twelve) NGS-generated reads (amplicon sequencing metagenomes) of fermented cacao beans from Brazil and Mexico. We extracted the microbiome profile using bioinformatic tools in the SHAMAN server. We analyzed further using Principal Component Analysis, Clustering (Ward's Method of Hierarchical Agglomerative Clustering), and UMAP (Uniform Manifold Approximation and Projection) combined with KNN (K-Nearest Neighbor). We discovered differences in microbial diversity and unique taxa in the fermented cacao beans from Brazil and Mexico. Lactic acid bacteria (LAB), such as *Liquorilactobacillus*, *Tatumella*, *Leuconostoc*, *Companilactobacillus*, and *Limosilactobacillus*, are unique genera in samples from Mexico, while *Bacillus* is a unique genus found in samples from Brazil. We have demonstrated the separation of the microbiome profiles between fermented cacao beans from Brazil and Mexico using PCA, clustering analysis and UMAP-KNN. We have successfully developed the proof of concept in establishing geographical indicators based on microbial diversity data or microbiome profiles. In the future, we will extend this research to analyze samples from Indonesia and establish a microbial diversity database of Indonesian fermented cacao. This database is essential for the authentication assay of Indonesian fermented cacao and for developing fine cacao and specialty products.

How to Cite: Nugroho, I.B. & Siregar, A.R. (2024). Establishing Geographical Indicator of Fermented Cacao Beans Using Microbiome Fingerprinting. Journal of Biotechnology and Natural Science, 4(1):47-54.

1. Introduction

Microbial processing relies heavily on microbial activity to convert raw materials, such as agro-industrial biomass, resulting in a value-added product (Rabha et al., 2023). Fermentation relies on microbial activity (Da Silva et al., 2021) and is also used for post-harvest processing of cacao beans (Ferreira et al., 2022; Streule et al., 2022; Taylor et al., 2022). Cacao beans are fermented dominantly by three groups of microbes: yeast, lactic acid bacteria (LAB), and acetic acid bacteria (AAB) (De Vuyst and Leroy, 2020). Fermentation by yeasts usually occurs in the first phase, followed successively by the occurrence of lactic acid bacteria and acetic acid bacteria (De C. Lima et al.,

2021; Rahayu et al., 2021). These microbes produce organic acids and convert organic components in cacao beans, such as amino acids and sugars, resulting in the unique flavor of fermented cacao beans (Fang et al., 2020; Calvo et al., 2021).

Studies have shown cacao beans from different origins have a unique profile of fermenting microbes. A study has demonstrated that fermented cacao beans from the Amazonian region have a distinctive microbial profile (Serra et al., 2019). This study also reported specific bacterial species found only in sub-regions/states in the Amazon. Another research by Papalexandratou et al. (2019) showed distinct microbial signatures of the fruity or berry-like flavor. Floral notes were associated with *Pichia kudriavzevii*, a yeast species, whereas *Bacillus* spp. was responsible for the lower flavor quality flavor. These studies showed the connection between the microbial profiles to the origin and flavor of the fermented cacao bean.

The microbial profiles of fermented cacao beans can be examined using Next Generation Sequencing (NGS) analysis (Nema, 2019; Schmidt et al., 2022), followed by bioinformatic data processing (Viesser et al., 2021). The result of bioinformatic data processing is the microbial community structure consisting of the microbial taxa (e.g., genus or species) and the relative abundance of each identified taxa. The community structure and relative abundance value are unique to the samples' attributes, such as the origin of fermented cacao beans. Therefore, the microbiome profile is a fingerprint, differentiating various samples' origins.

We analyzed two groups of NGS-generated datasets of fermented cacao beans originating from Mexico and Brazil, available on the European Nucleotide Archive (ENA) public server. After analyses using pipelines on the SHAMAN (Volant et al., 2020) server, we retrieved the dominant taxa from each dataset. We employed Principal Component Analysis (PCA) (Maćkiewicz and Ratajczak, 1993), clustering analysis by Ward's Method (Murtagh and Legendre, 2014), and UMAP (Uniform Manifold Approximation and Projection) (McInnes et al., 2018) combined with K-Nearest Neighbor (KNN) analysis to separate each group. PCA resulted in the separation of the microbiome profiles of fermented cacao beans from Mexico and Brazil.

We aimed to explore an alternative to establishing geographical indicators by comparing different microbial profiles of fermented cacao beans from various origins. This study demonstrates the proof-of-concept of differentiating fermented cacao bean origins by using microbiome profiling methods and profile separation by PCA, clustering analysis, and UMAP-KNN. We also retrieved the microbial signature successfully differentiating the Brazil and Mexico fermented cacao beans. In the future, we will extend this method for analyzing samples from Indonesian regions which are known to produce fermented cacao beans, such as Sumatera, Java, and Sulawesi (Brugman et al., 2022).

2. Methods

2.1 Data collection

We accessed the European Nucleotide Archive (ENA) (https://www.ebi.ac.uk/ena/browser/) and searched for datasets using the keywords: "fermented cacao beans" and "cocoa fermentation". We downloaded two NGS datasets of fermented cacao beans from Brazil and Mexico. Each dataset contains six paired-ended sequencing reads or run datasets (see Table 1).

Table 1. Datasets used in this study

Sample	SRR Number	BioProject Accession Number	Region	Reference
M1	SRR23473202	PRJNA935329	Mexico	-
M2	SRR23473203			
M3	SRR23473204			
M4	SRR23473205			
M5	SRR23473206			
M6	SRR23473207			
B1	SRR6159231	PRJNA407677	Brazil	Serra et al., (2019)
B2	SRR6159232			
B3	SRR6162951			
B4	SRR6162952			
B5	SRR6162953			
B6	SRR6162954			

We selected the datasets based on several criteria, i.e., the Illumina platform for generating raw sequence reads and amplicon sequencing metagenomic datasets. For feasible data handling, we selected projects based on BioProject Accession Number that consists of at least two SRR Numbers and not more than ten SRR Numbers.

2.2 Microbiome data analysis

We conducted cloud computing in the SHAMAN (https://shaman.pasteur.fr/) server. SHAMAN's (Volant et al., 2020) analysis comprises contig assembly, adapters clipping, and chimeric sequence removal. We matched clean reads against the SILVA database for genus or species identification. Reads belonging to the same taxa (e.g. genus or species) were counted and merged. This analysis resulted in the generation of a table containing all identified bacterial Genus or Species within the sample and the corresponding count.

2.3 PCA, clustering, and UMAP analyses

We carried out Principal Component Analysis (Maćkiewicz and Ratajczak, 1993) and clustering analysis using R programming built-in within SHAMAN (Volant et al., 2020). We performed the clustering analysis using Ward's Method (Murtagh and Legendre, 2014) with the implementation of the Bray-Curtis (Bray and Curtis, 1957; Ricota and Podani, 2017) and Chao (Chao, 1984; Hughes et al., 2001) distance matrix. We conducted the UMAP (McInnes et al., 2018) analysis in combination with the K-Nearest Neighbor (KNN) algorithm (UMAP-KNN).

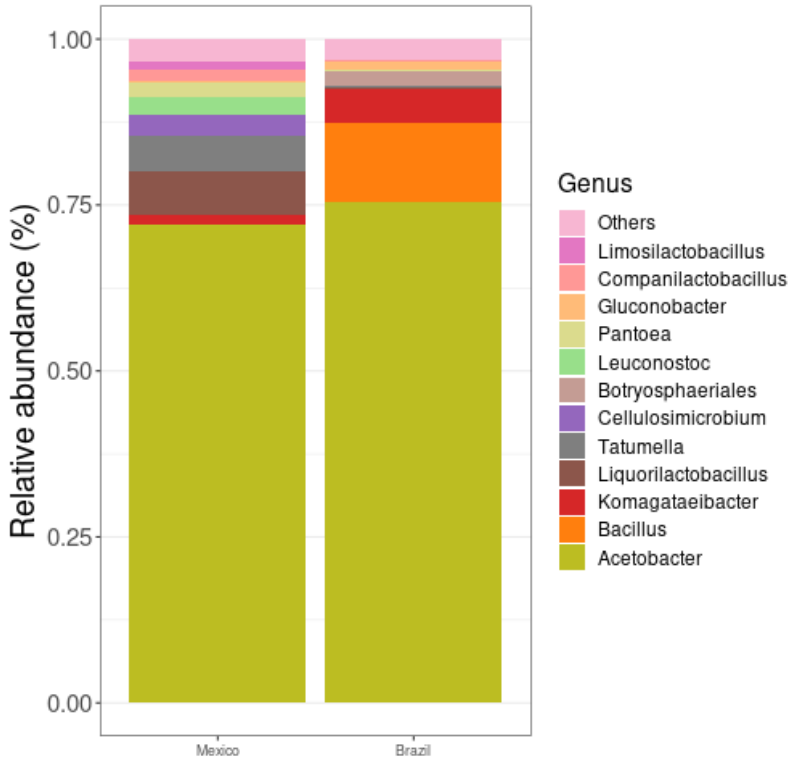
3. Results and Discussion

3.1. Microbiome Profiles and Signatures of Fermented Cacao Bean from Brazil and Mexico

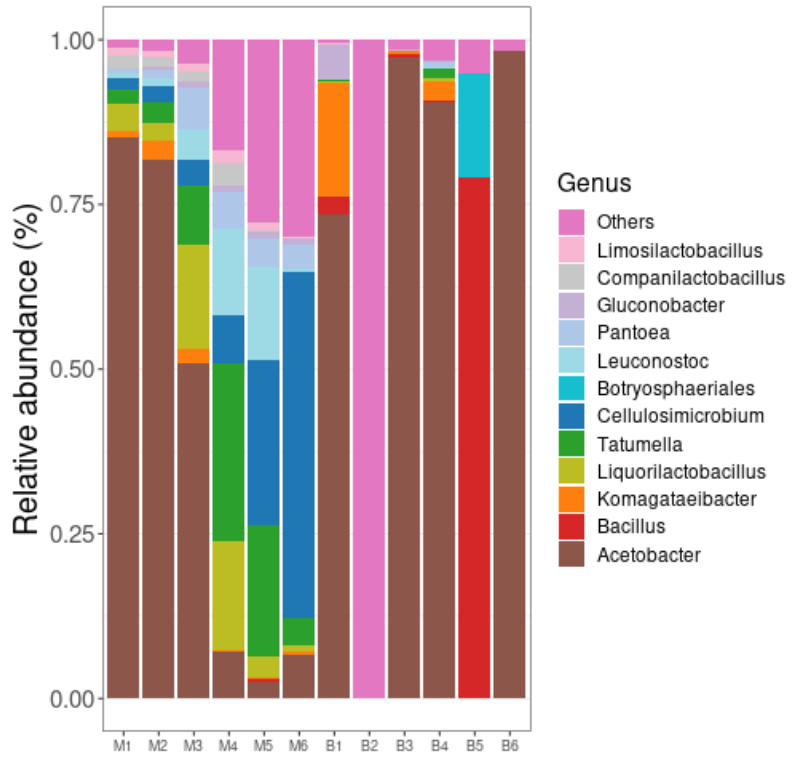
This study aims to explore an alternative method for establishing industrial products' origin or geographical indicator based on the analysis of microbial diversity data generated by the Next Generation Sequencing (NGS) machine. Raw reads were processed using bioinformatic tools in SHAMAN, followed by multivariate analysis and a machine learning approach. The employed analyses were to generate a microbiome profile, including microbial taxa unique from each origin. Furthermore, multivariate and machine learning analysis, namely, PCA, clustering, and UMAP-KNN were utilized to separate samples from different origins and cluster them into distinctive partitions.

The result in Figure 1 depicts the difference between the microbiome profile of fermented cacao beans from Brazil and Mexico. The average value of each contributing genus to the overall microbiome profile (Figure 1.a) showed the genus *Acetobacter* as the dominant taxa found in both samples from Brazil and Mexico. However, upon a detailed examination, *Acetobacter* is only present predominantly in four out of six samples from Brazil and only three from Mexico (Figure 1.b & 2.a). The involvement of *Acetobacter* in cacao bean fermentation is well-known. The genus *Acetobacter* belongs to the Acetic Acid Bacteria (AAB) and is a dominant genus in cacao fermentation (Soumahoro et al., 2020). AAB mostly consumes ethanol and lactic acid produced during the early phase of the fermentation by yeasts and lactic acid bacteria (LAB). One of the members of this genus, namely *Acetobacter pasteurianus* is a well-known bacterium, which dominates the last phase of the cacao bean fermentation (Tigrero-Vaca et al., 2022).

The analysis result also depicted the difference in microbial diversity of fermented cacao beans from Brazil and Mexico. Samples from Mexico showed higher diversity than those from Brazil. Ten dominant genera, i.e., *Acetobacter*, *Komagataeibacter*, *Liquorilactobacillus*, *Tatumella*, *Gluconobacter*, *Cellulosimicrobium*, *Pantoea*, *Leuconostoc*, *Companilactobacillus*, and *Limosilactobacillus* were in samples from Mexico. Meanwhile, six dominant genera, i.e., *Acetobacter*, *Bacillus*, *Komagataeibacter*, *Gluconobacter*, *Botryosphaeriales*, and *Pantoea* were in samples from Brazil. Other genera were at low abundance (< 0.01 %). *Liquorilactobacillus*, *Tatumella*, *Leuconostoc*, *Companilactobacillus* and *Limosilactobacillus* are unique in samples from Mexico, while *Bacillus* and *Botryosphaeriales* are unique signatures of samples from Brazil. All unique genera observed in samples from Mexico are lactic acid bacteria classified in the order of *Lactobacillales*. Moreover, *Bacillus* is a genus classified in the order *Bacillales*, while *Botryospaheriales* are an order of fungi.



a)



b)

Figure 1. The microbial diversity profile of fermented cacao beans from Mexico and Brazil. a. aggregate profile; b. profile breakdown per sample.

These results demonstrated that the fermentation conditions can influence the diversity and unique microbial genera discovered in samples from Brazil and Mexico. Lactic acid bacteria predominate the fermenting environment for up to 72 hours (3 days) from the starting point of cacao bean

fermentation. Meanwhile, *Bacillus* and *Botryospaheriales* predominate the fermenting environment during the later phase of fermentation (72-120 hours or 3-5 days). Studies have also demonstrated that *Bacillus* and other microbes, such as fungi, are known for spoilage-causing microbes that introduce unwanted flavour or off notes in fermented cacao beans.

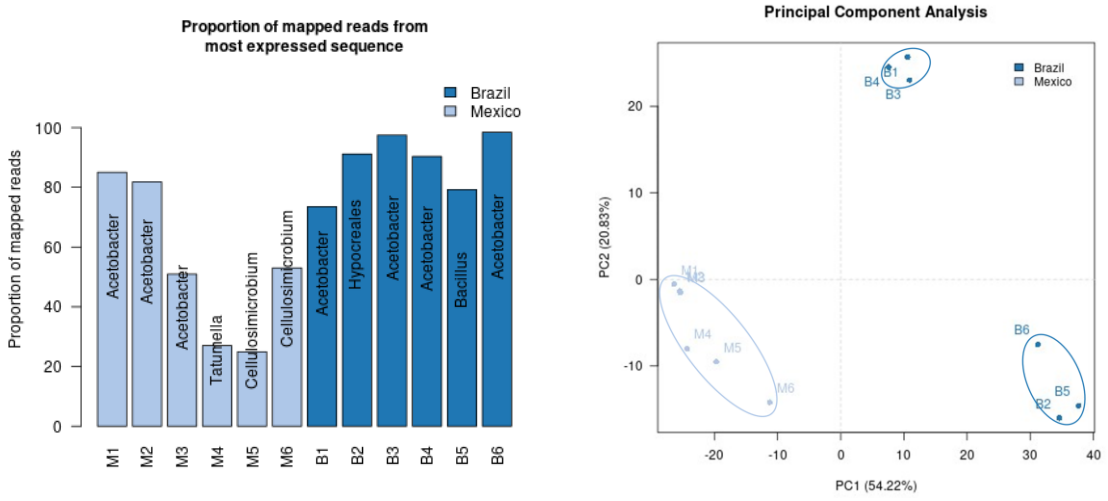
Based on the microbial diversity results, we can presume that fermented cacao beans from Mexico were finer and better in quality than other fermented cacao beans from Brazil. Samples from Mexico were homogenous and predominated by key microbes, such as LAB and AAB, while samples from Brazil were rather heterogeneous. The genus *Bacillus* was found in samples from Brazil, indicating over-fermentation and the possibility of spoilage leading to lesser fermented bean quality.

3.2 Microbiome Fingerprinting of Different Origins of Fermented Cacao Beans

The microbial profile provides a distinction between samples from Brazil and Mexico. In this study, we successfully made a separation based on the microbial diversity data using several multivariate analyses and machine learning algorithms, such as PCA, clustering, and UMAP-KNN analysis. Principal Component Analysis (PCA) separates samples from Brazil and Mexico distinctively (Figure 2. b).

The microbiome profile of samples from Mexico is more likely homogenous than samples from Brazil. The PCA plot also supported such findings and grouped samples from Mexico as one cluster. A comparison of individual samples from Mexico (Figure 1. b) revealed that every sample contains similar taxa even though each taxon has variable relative abundance.

Conversely, samples from Brazil consist of dissimilar taxa. Sample B1 is relatively invariable to sample B3 and B4, whereas sample B2 is close to sample B5 and B6; thus, it results in two clusters split (Figure 2. b).



a) **b)**

Figure 2. a. Major taxa of fermented cacao beans from Mexico and Brazil and b. Principal Component Analysis (PCA) result.

We conducted further examination of the microbial profile through clustering analysis. We implemented Ward's Method for clustering and then visualized using a cluster tree or dendrogram. (Figure 3. a and 3. b). Clustering by Bray-Curtis distance (Figure 3. a) resulted in a dendrogram comprising three clades (branches). The dendrogram visualized different clades of samples from Brazil (B1-B6) and Mexico (M1-M6). Two subclades within Clade 1 contain mixed individual samples from Brazil and Mexico. Clade 2 only consists of samples from Mexico, while Clade 3 consists of samples from Brazil.

Clustering by Chao distance resulted in more relatively efficient separation of samples from Brazil and Mexico. This dendrogram consists of three distinct clades. Clade 1 and Clade 3 contain only samples from Brazil, while Clade 2 contains only samples from Mexico.

Lastly, UMAP analysis using the KNN algorithm with k=4 also showed a distinct group split of samples from Brazil and Mexico. Both Euclidean and Manhattan distances implemented during

computation result in clear segregation of groups of samples from Brazil and Mexico (Figure 3. c and 3.d). Moreover, Euclidean distance is likely more optimal than Manhattan distance to handle highly variable data. Euclidean distance has successfully grouped all samples from Brazil in a quadrant while samples from Mexico in another different quadrant (Figure 3. c). However, Manhattan distance can only restrict samples from Mexico, which is less variable than samples from Brazil, in a single quadrant. Meanwhile, using the same distance, samples from Brazil are placed into two adjacent quadrants (Figure 3.d).

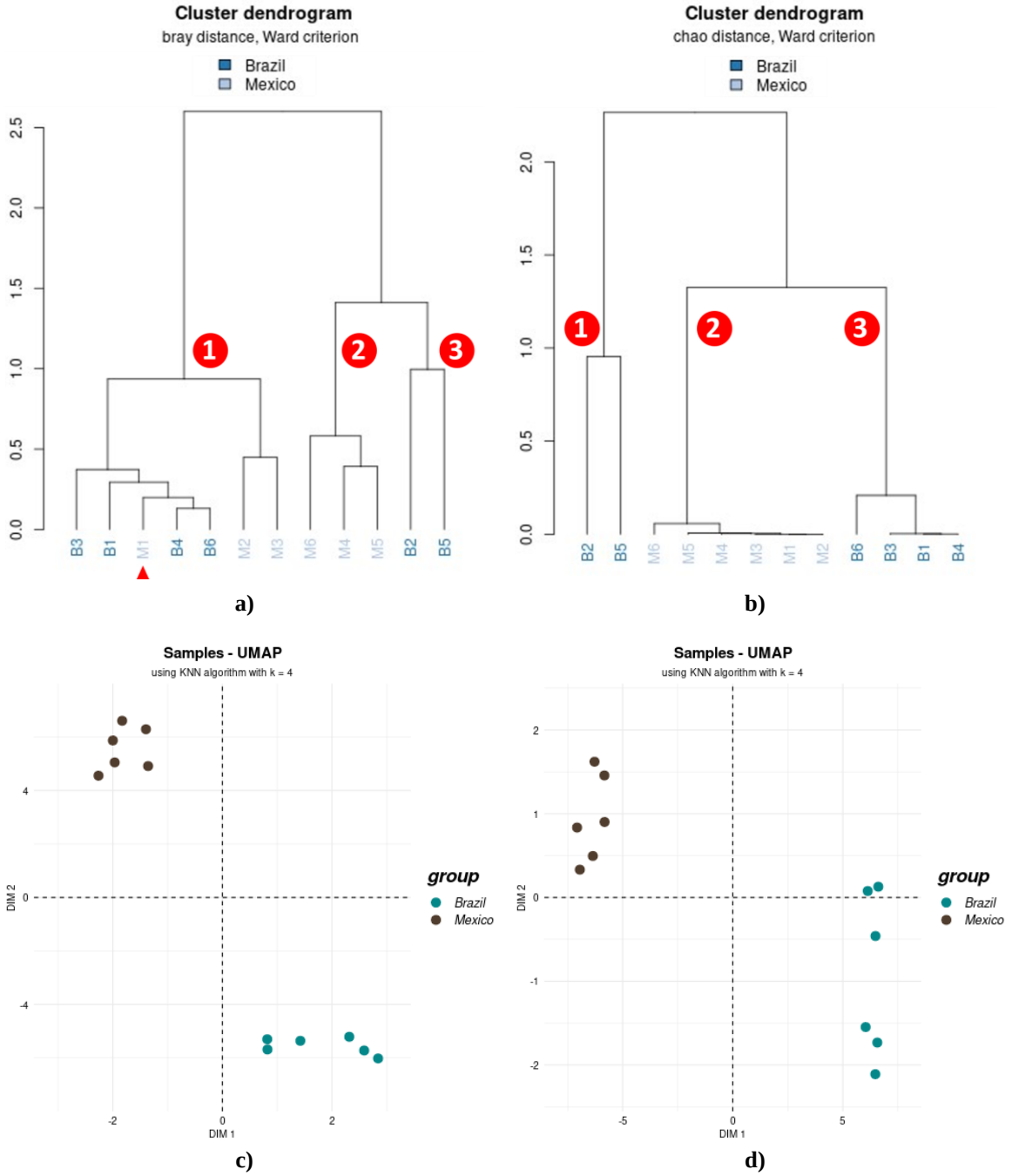


Figure 3. Clustering and UAMP analysis results. Ward's Method clustering using Bray-Curtis (a) and Chao (b) distance. UMAP-KNN analysis using Euclidean (c) and Manhattan (d) distance.

This study explores an alternative method for establishing geographical indications. Analysis of amplicon sequencing metagenomes produced microbial diversity data or microbiome profile. Further analyses, namely PCA, clustering analysis and UMAP-KNN, showed the utilization of microbiome

profiles to differentiate samples from different origins. Therefore, this study has successfully demonstrated the usage of microbiome profiles as geographical indicators. In terms of convenience, UMAP-KNN analysis is relatively robust in separating the microbiome profile of samples from different origins. Further analysis using broader datasets and applications, e.g., using datasets from Indonesian samples, is necessary to test the reliability of the methods implemented here.

4. Conclusion

This study has successfully served as a proof of concept showing the possibility of establishing geographical indicator or origin of fermented cacao beans using analysis based on microbial diversity data. Fermented cacao beans have microbiome profile corelated to the origin, such as genera of lactic acid bacteria (LAB) uniquely found in samples from Mexico. PCA, clustering, and UMAP-KNN analysis can be used to classify and separate samples from Brazil and Mexico distinctively. According to this study, UMAP-KNN is the most preferable choice to cluster samples from Brazil and Mexico as it can also handle relatively high variable data.

REFERENCES

Bray, J.R., & Curtis, J.T. (1957). An Ordination of The Upland Forest Communities of Southern Wisconsin. *Ecological Monographs*, 27(4):326–349. https://doi.org/10.2307/1942268.

Brugman, E., Wibowo, A., & Widiastuti, A. (2022). *Phytophthora palmivora* from Sulawesi and Java Islands, Indonesia, Reveals High Genotypic Diversity and Lack of Population Structure. *Fungal Biology*, 126(4):267–276. https://doi.org/10.1016/j.funbio.2022.02.004.

Calvo, A.M., Botina, B.L., García, M.C., Cardona, W.A., Montenegro, A.C., & Criollo, J. (2021). Dynamics of Cocoa Fermentation and Its Effect on Quality. *Scientific Reports*, 11(1):16746. https://doi.org/10.1038/s41598-021-95703-2.

Chao, A. (1984). Nonparametric Estimation of The Number of Classes in A Population. *Scandinavian Journal of Statistics*, 11(4):265–270. Retrieved from https://www.jstor.org/stable/4615964.

Da Silva, B.L., Pereira, P.V., Bertoli, L.D., Silveira, D.L., Batista, N.N., Pinheiro, P.F., Carneiro, J.S., Schwan, R.F., Silva, S.A., Coelho, J.M., & Bernardes, P.C. (2021). Fermentation Of *Coffea canephora* Inoculated with Yeasts: Microbiological, Chemical, And Sensory Characteristics. *Food Microbiology*, 98:103786. https://doi.org/10.1016/j.fm.2021.103786.

De C. Lima, C.O., Vaz, A.B.M., De Castro, G.M., Lobo, F., Solar, R., Rodrigues, C., Pinto L.R.M., Vandenberghe, L., Pereira, G., da Costa, A.M., Benevides R.G., Azevedo, V., Uetanabaro, A.P.T., Soccol, C.R., & Góes-Neto, A. (2021). Integrating Microbial Metagenomics and Physicochemical Parameters and A New Perspective on Starter Culture for Fine Cocoa Fermentation. *Food Microbiology*, 93:103608. https://doi.org/10.1016/j.fm.2020.103608.

De Vuyst, L., & Leroy, F. (2020). Functional Role of Yeasts, Lactic Acid Bacteria and Acetic Acid Bacteria In Cocoa Fermentation Processes. *FEMS Microbiology Reviews*, 44(4):432–453. https://doi.org/10.1093/femsre/fuaa014.

Fang, Y., Li, R., Chu, Z., Zhu, K., Gu, F., & Zhang, Y. (2020). Chemical And Flavor Profile Changes of Cocoa Beans (*Theobroma Cacao* L.) During Primary Fermentation. *Food Science & Nutrition*, 8(8):4121–4133. https://doi.org/10.1002/fsn3.1701.

Ferreira, O.D.S., Chagas-Junior, G.C.A., Chisté, R.C., Martins, L.H.D.S., Andrade, E.H.D. A., Nascimento, L.D.D., & Lopes, A.S. (2022). *Saccharomyces cerevisiae* and *Pichia manshurica* from Amazonian Biome Affect the Parameters of Quality and Aromatic Profile of Fermented and Dried Cocoa Beans. *Journal of Food Science*, 87(9):4148–4161. https://doi.org/10.1111/1750-3841.16282.

Hughes, J.B., Hellmann, J.J., Ricketts, T.H., & Bohannan, B.J.M. (2001). Counting The Uncountable: Statistical Approaches to Estimating Microbial Diversity. *Applied and Environmental Microbiology*, 67(10):4399–4406. https://doi.org/10.1128/AEM.67.10.4399-4406.2001.

Maćkiewicz, A., & Ratajczak, W. (1993). Principal Components Analysis (PCA). *Computers & Geosciences*, 19(3):303–342. https://doi.org/10.1016/0098-3004(93)90090-R.

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861. https://doi.org/10.21105/joss.00861.

Murtagh, F. & Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, 31(3):274–295. https://doi.org/10.1007/s00357-014-9161-z.

Nema, V. (2019). The Role and Future Possibilities of Next-Generation Sequencing in Studying Microbial Diversity. In: Das S & Dash HR (Eds.). *Microbial Diversity in the Genomic Era*, pp. 611–630. Academic Press Elsevier London UK. https://doi.org/10.1016/B978-0-12-814849-5.00034-4.

Papalexandratou, Z., Kaasik, K., Kauffmann, L.V., Skorstengaard, A., Bouillon, G., Espensen, J.L., Hansen, L.H., Jakobsen, R.R., Blennow, A., Krych, L., & Nielsen, D.S. (2019). Linking Cocoa Varietals and Microbial Diversity of Nicaraguan Fine Cocoa Bean Fermentations and Their Impact on Final Cocoa Quality Appreciation. *International Journal of Food Microbiology*, 304:106–118. https://doi.org/10.1016/j.ijfoodmicro.2019.05.012.

Rabha, J., Devi, S.P., Das, S., Roy, N., & Jha, D.K. (2023). Microbial conversion of biomass to value-added chemicals. In: Kuddus M & Ramteke P (Eds.). *Value-Addition in Agri-food Industry Waste Through Enzyme Technology*, pp. 37–64. Academic Press Elsevier London UK. https://doi.org/10.1016/B978-0-323-89928-4.00018-3.

Rahayu, E.S., Triyadi, R., Khusna, R.N.B., Djaafar, T.F., Utami, T., Marwati, T., & Hatmi, R.U. (2021). Indigenous Yeast, Lactic Acid Bacteria, And Acetic Acid Bacteria from Cocoa Bean Fermentation in Indonesia Can Inhibit Fungal-Growth-Producing Mycotoxins. *Fermentation*, 7(3):192. https://doi.org/10.3390/fermentation7030192.

Ricotta, C., & Podani, J. (2017). On Some Properties of The Bray-Curtis Dissimilarity and Their Ecological Meaning. *Ecological Complexity*, 31:201–205. https://doi.org/10.1016/j.ecocom.2017.07.003.

Schmidt, P.J., Cameron, E.S., Müller, K.M., & Emelko, M.B. (2022). Ensuring That Fundamentals of Quantitative Microbiology Are Reflected in Microbial Diversity Analyses Based on Next-Generation Sequencing. *Frontiers in Microbiology*, 13:728146. https://doi.org/10.3389/fmicb.2022.728146.

Serra, J.L., Moura, F.G., Pereira, G.V.D.M., Soccol, C.R., Rogez, H., & Darnet, S. (2019). Determination Of the Microbial Community in Amazonian Cocoa Bean Fermentation by Illumina-Based Metagenomic Sequencing. *LWT*, 106:229–239. https://doi.org/10.1016/j.lwt.2019.02.038.

Soumahoro, S., Ouattara, H.G., Droux, M., Nasser, W., Niamke, S.L., & Reverchon, S. (2020). Acetic Acid Bacteria (AAB) Involved in Cocoa Fermentation from Ivory Coast: Species Diversity and Performance In Acetic Acid Production. *Journal of Food Science and Technology*, 57(5):1904–1916. https://doi.org/10.1007/s13197-019-04226-2.

Streule, S., Freimüller Leischtfeld, S., Galler, M., & Miescher Schwenninger, S. (2022). Monitoring Of Cocoa Post-Harvest Process Practices on A Small-Farm Level at Five Locations in Ecuador. *Heliyon*, 8(6):e09628. https://doi.org/10.1016/j.heliyon.2022.e09628.

Tigrero-Vaca, J., Maridueña-Zavala, M.G., Liao, H.-L., Prado-Lince, M., Zambrano-Vera, C.S., Monserrate-Maggi, B., & Cevallos-Cevallos, J.M. (2022). Microbial Diversity and Contribution to The Formation of Volatile Compounds During Fine-Flavor Cacao Bean Fermentation. *Foods*, 11(7):915. https://doi.org/10.3390/foods11070915.

Viesser, J.A., De Melo Pereira, G.V., De Carvalho Neto, D.P., Favero, G.R., De Carvalho, J.C., Goés-Neto, A., Rogez, H., & Soccol, C.R. (2021). Global Cocoa Fermentation Microbiome: Revealing New Taxa and Microbial Functions by Next Generation Sequencing Technologies. *World Journal of Microbiology and Biotechnology*, 37(7):118. https://doi.org/10.1007/s11274-021-03079-2.

Volant, S., Lechat, P., Woringer, P., Motreff, L., Campagne, P., Malabat, C., Kennedy, S., Ghozlane, A. (2020). SHAMAN: A User-Friendly Website for Metataxonomic Analysis from Raw Reads to Statistical Analysis. *BMC Bioinformatics*, 21(1):345. https://doi.org/10.1186/s12859-020-03666-4.