Acta Scientiarum

http://periodicos.uem.br/ojs/acta
ISSN on-line: 1807-863X
Doi: 10.4025/actascibiolsci.v42i1.49877

Genetic diversity of *16S rRNA* and *mcrA* genes from methanogenic archaeas

Keyla VItória Marques Xavier[1*], Eden Silva e Souza[2], Erick de Aquino Santos[3] and Michely Correia Diniz[4*]

[1]Universidade Federal de Pernambuco, Recife, Pernambuco, Brasil. [2]Universidade Federal do Rio Grande do Norte, Natal, Rio Grande do Norte, Brasil. [3]Universidade Federal da Bahia, Salvador, Bahia, Brasil. [4]Universidade Federal do Vale do São Francisco, Campus Ciências Agrárias, Rod. BR-407, 12, Lote 543, Nilo Coelho C1, 56300-990, Petrolina, Pernambuco, Brasil. *Authors for correspondence. E-mail: keylaxavier01@gmail.com; michely.diniz@univasf.edu.br;

ABSTRACT. Methanogenic archaeas are found in aquatic and terrestrial environments and are fundamental in the conversion of organic matter into methane, a gas that has a potential use as renewable source of energy, which is also considered as one of the main agents of the greenhouse effect. The vast majority of microbial genomes can be identified by a conservative molecular marker, the *16S ribosomal* gene. However, the *mcrA* gene have been using in studies of methanogenic archaea diversity as an alternative marker, highly conserved and present only in methanogens. This gene allows the expression of the enzyme Methyl-coenzyme M reductase, the main agent in converting by-products of anaerobic digestion into methane. In this context, we aimed to study the genetic diversity of *mcrA* and *16S rRNA* genes sequences available in databases. The nucleotide sequences were selected from the NCBI. The heterozygosity and molecular diversity indexes were calculated using the Arlequin 3.5 software, with plots generated by package R v3.0. The diversity and heterozygosity indices for both genes may have been influenced by the number and size of the sequences. Descriptive analysis of genetic diversity generated by sequences deposited in databases allowed a detailed study of these molecules. It is known that the organisms in a population are genetically distinct, and that, despite having similarities in their gene composition, the differences are essential for their adaptation to different environments.

Keywords: heterozygosity; microbiomes; methane; anaerobic; arlerquin; databases.

Received on July 19, 2019.
Accepted on January 27, 2020.

Introduction

Archaea is a domain present in the tree of life. This group, considered polymorphic, is directly involved in the anaerobic digestion process and is responsible for performing methanogenesis steps. Hydrogenotrophic methanogenic archaea are capable of converting hydrogen and carbon dioxide into methane, while acetoclastic methanogenic archaea converts acetate to methane (Amaral, Steinmetz, & Kunz, 2019).

Studies of these organisms are carried out mainly through metagenomics, because cultivation in laboratory is difficult. However, recent studies accomplished by Japanese scientists (Imachi et al., 2020) isolated and cultivated, after almost fifteen years, in the laboratory an Asgard archaeon strain extracted from a deep marine sediment.

The current knowledge about archaea is due to the development of a way to study environmental samples without the need to cultivate microorganisms. The first application of metagenomics with sequencing was performed in the 70's decade, from which several techniques were developed and refined, in order to obtain increasingly reliable analyzes. There is currently a wide range of metagenomic data that can be accessed at several databases, such as NCBI – (National Center for Biotechnology Information), and which allows various microbiome studies to be performed.

Within the Archaea group, methanogenic archaeas are found in aquatic and terrestrial environments, which are fundamental in the conversion of organic matter into methane, a gas that has a potential use as renewable source of energy, which is also considered as one of the main agents of the greenhouse effect. Environmental sequence of data suggests that they are common constituents of anaerobic environments across the globe, ranging from hypersaline to freshwater sediments, acidic to alkaline soils, invertebrate to vertebrate gastrointestinal tracts, industrial bioreactors, hydrocarbon-rich deposits, and rice paddies at temperatures ranging from extreme thermal to near freezing conditions (Figure 1) (Evans et al., 2019).

In this context, genetic and evolutionary studies of various types of environmental microorganisms might be performed through a series of molecular genetics techniques. Metagenomics allowed the acquisition of microbial DNA without the need for culture isolation and became the main technique for the study of microbial community genomes of samples obtained directly from the environment (Handelsman, 2004).

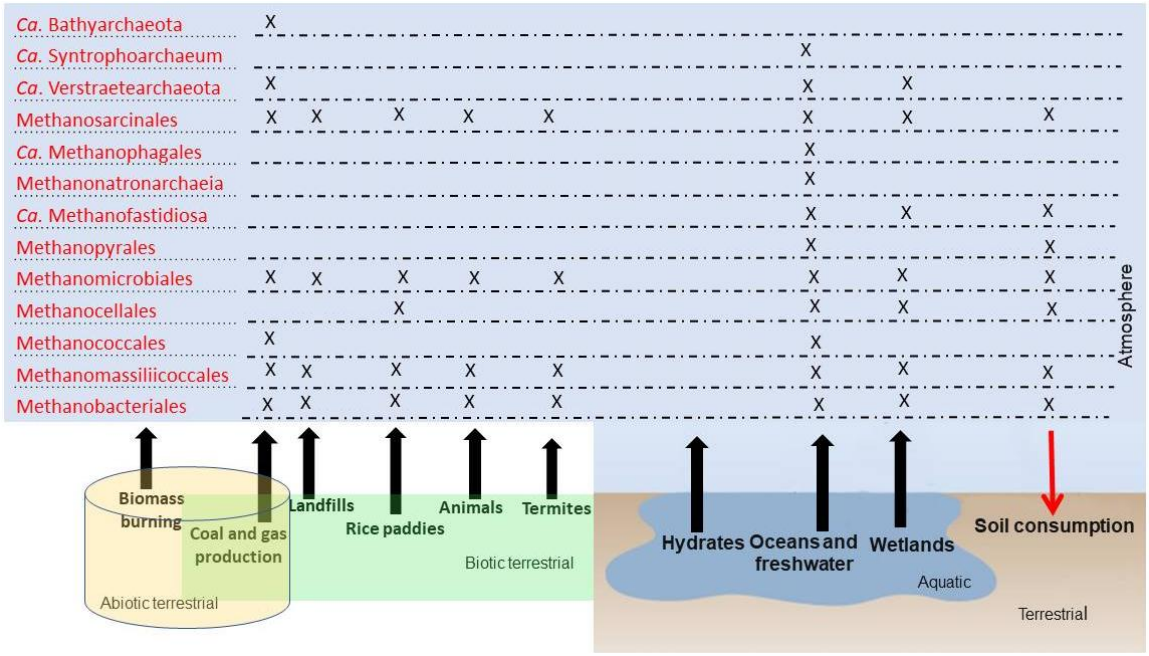

Figure 1. Methane Flux. Archaeal phyla and orders contribute to the global methane flux (red arrow is net consumption, and black arrows are net production) from anaerobic environments. X indicate the presence of methanogenic and methanotrophic archaeal strains in different environments based on *16S* ribosomal RNA (rRNA) gene sequences from publicly available methanogen isolates. Source: Based in Evans et al., 2019.

Archaeal genome is a circular double-stranded DNA molecule - l. 9 Mbp in length, or - 45% the size of the *E. coli* genome. Methanogen genomic DNA range overall from 26 to 68 mol% G+C, although intergenic regions are frequently more A+T rich than the average value for the other archaeal genomes (Reeve, 1992).

The vast majority of microbial genomes can be identified by a conservative molecular marker, the *16S* ribosomal gene. Through cloning and sequencing techniques, the *16S rRNA* gene becomes a great ally in studies on microbial diversity (Handelsman, 2004).

However, the *mcrA* gene have been using in studies of methanogenic archaea diversity as an alternative marker, highly conserved and present only in methanogens. This gene allows the expression of the enzyme Methyl-coenzyme M reductase, the main agent in converting by-products of anaerobic digestion into methane. One of the advantages of *mcr*A gene is that only one or two copies of *mcr*A have been found in sequenced methanogens genomes, making it a more precise tool for estimating the number of these archaeas in the environment than the *16S rRNA* gene, which can have up to four copies per genome (Lee, Kim, Hwang, O'Flaherty, & Hwang, 2009). Thus, *mcrA* has been replacing the *16S rRNA* gene in the study of methanogenic diversity and phylogenetics (Hallam, Girguis, Preston, Richardson, & DeLong, 2003; Evans et al., 2019).

Molecular studies of these environmental samples allow the understanding microbiomes and their interactions. Bioinformatics analysis will allow a greater functional and structural understanding of the microbiome. In this context, we aimed to study the genetic diversity of *mcrA* and *16S rRNA* genes sequences available in databases.

Material and methods

Sequences retrieval

mcrA and *16S rRNA* sequences genes were obtained from NCBI – (National Center for Biotechnology Information) https://www.ncbi.nlm.nih.gov/taxonomy, with the keyword "Methanobacteria" selecting the *mcrA* and *16S* sequences. The resulting sequences were classified according to information available from NCBI and were used in FASTA format.

Sequence alignment was carried out using BioEdit v.7.2.6 (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). Then, the sequences of each gene were grouped relating to the country of collection and its environment (Supplementary data). The nucleotide compositions of the groups were verified on MEGA-X (https://www.megasoftware.net/).

Descriptive statistics of genetic diversity

The analysis of genetic diversity was performed with the software ARLEQUIN v3.5X (Excoffier & Lischer, 2010). The sequences were organized into groups and converted into .arp file using the software DNA sequence polymorphism (DnaSP v6) (Rozas et al., 2017). The .arp format is used as the input format to ARLEQUIN. The data generated were analyzed and shown as plots of Heterozygosity and Molecular diversity index using the statistical ARLEQUIN for R package v3.0.

Results and discussion

We found 59 sequences for the *mcrA* gene, but after an analysis of them, 56 were selected for presenting complete data of their origin, totalizing 14 different origins. Regarding the *16S* gene, of the 27 found, 25 were selected, showing six types of environments.

These sequences were organized into groups, in which from the 56 sequences, 11 groups of *mcrA* gene were formed (Table 1); whereas from 25 sequences for *16S*, five groups of *16S rRNA* gene (Table 2) were constituted. After aligning the gene sequences into their groups, their nucleotide composition was conducted using MEGA-X.

In the *mcrA* groups (Table 1), the largest sequence number (14) belonged to the mesophilic sludge, with an average sequence size of 454.7 base pairs (bp). However, its GC content (49.9%) was not the largest among the groups. Thermal Waters group, with only two sequences in its composition, showed larger CG content (51.6%) in *mcrA* groups.

The Pool group with the largest average size (760.7 bp showed 46.3% of CG. The groups belonging to Marine Sediment, Terrestrial Sediment and Peninsula (two sequences belonging to the group) showed the lowest percentages of CG, being 38.9%, 39.2% and 42.6%, respectively.

Table 1. Nucleotide composition of groups for the *mcrA* gene estimated by MEGA-X.

mcrA Groups	T %	C%	A%	G%	GC%	Sequences Number	Average Sequence Size (bp)
Thermal Waters	20.2	26.1	28. 2	25.5	51.6	2	596.5
Rice	24.0	22.2	31.9	21.9	44.1	6	499
Biogas Reactor	19.9	25.2	30.5	24.5	49.7	3	477
Beer	22.6	24.8	30.1	22.6	47.4	3	371
Thermophilic Consortium	18.6	25.0	32.0	24.4	49.4	6	717.5
Mesophilic Sludge	21.3	25.6	28.8	24.3	49.9	14	454.7
Peninsula	23.4	21.1	34.0	21.5	42.6	2	428
Oil	19.1	25.0	32.0	23.9	48.9	3	397.3
Pool	23.3	23.4	30.4	22.9	46.3	3	760.7
Marine Sediment	31.2	17.6	29.9	21.3	38.9	8	718.9
Terrestrial Sediment	28.7	19.4	32.1	19.8	39.2	6	465.5

Table 2. Nucleotide composition of groups for *16S* rRNA *gene* estimated by MEGA-X.

16S Groups	T%	C%	A%	G%	GC%	Sequences Number	Average Sequence Size (bp)
Maize Reactor	20.3	23.8	23.9	32.0	55.8	13	1293.9
Seafloor Mud	20.3	22.4	23.0	34.2	56.6	7	380.9
BR Oil	18.0	24.8	22.2	35.0	59.8	3	258
JP Oil	17.9	25.7	22.9	33.5	59.2	1	1452
Sludge Water	20.2	22.4	23.2	34.2	56.6	1	406

In the *16S* groups (Table 2), the JP Oil group (only one sequence) had an average of 1452 bp, with 59.2% GC, representing the second largest percentage. The BR Oil group (three sequences) had an average of 258 bp and the largest amount of CG content (59.8%). The Maize Reactor group, despite having the largest number of sequences composing it (13 sequences, with an average of 1293.9 bp), showed the lowest CG content, 55.8%.

According to Yakovchuk, Protozanova, and Frank-Kamenetskii (2006), DNA sequences with low GC content are less stable than with high GC content. This stability is not dependent of the three hydrogen bonds, but is largely due to the base stacking.

It is also known that in PCR experiments, the GC content of the primers is counted by the software to calculate the annealing temperature to the DNA template. An elevated GC content indicates a relatively higher annealing temperature (Dieffenbach, Lowe, & Dveksler, 1993).

The expected heterozygosity (He) is defined as an estimated fraction of all individuals who may be heterozygous at a randomly chosen locus. Groups with a He close to one, indicate that the alleles present are very different within their group. However, when the sequences are more similar to each other in a group, the closest to zero will be the value, indicating a low heterozygosity (Nei, 1978).

The Figure 2 shows the expected heterozygosity (He) of the *mcrA* gene sequence groups from the haplotypes (each base position in the sequences) present in each cluster.

The Beer group (a) had a He average of 0.669 with 142 polymorphic loci. Thermal Waters (b) showed 284 nucleotides with polymorphism, varying in several positions within the sequence, reaching a high He, with average 1; the same value found in Peninsula (d), which may have been caused by introducing gaps (spaces) during alignment.

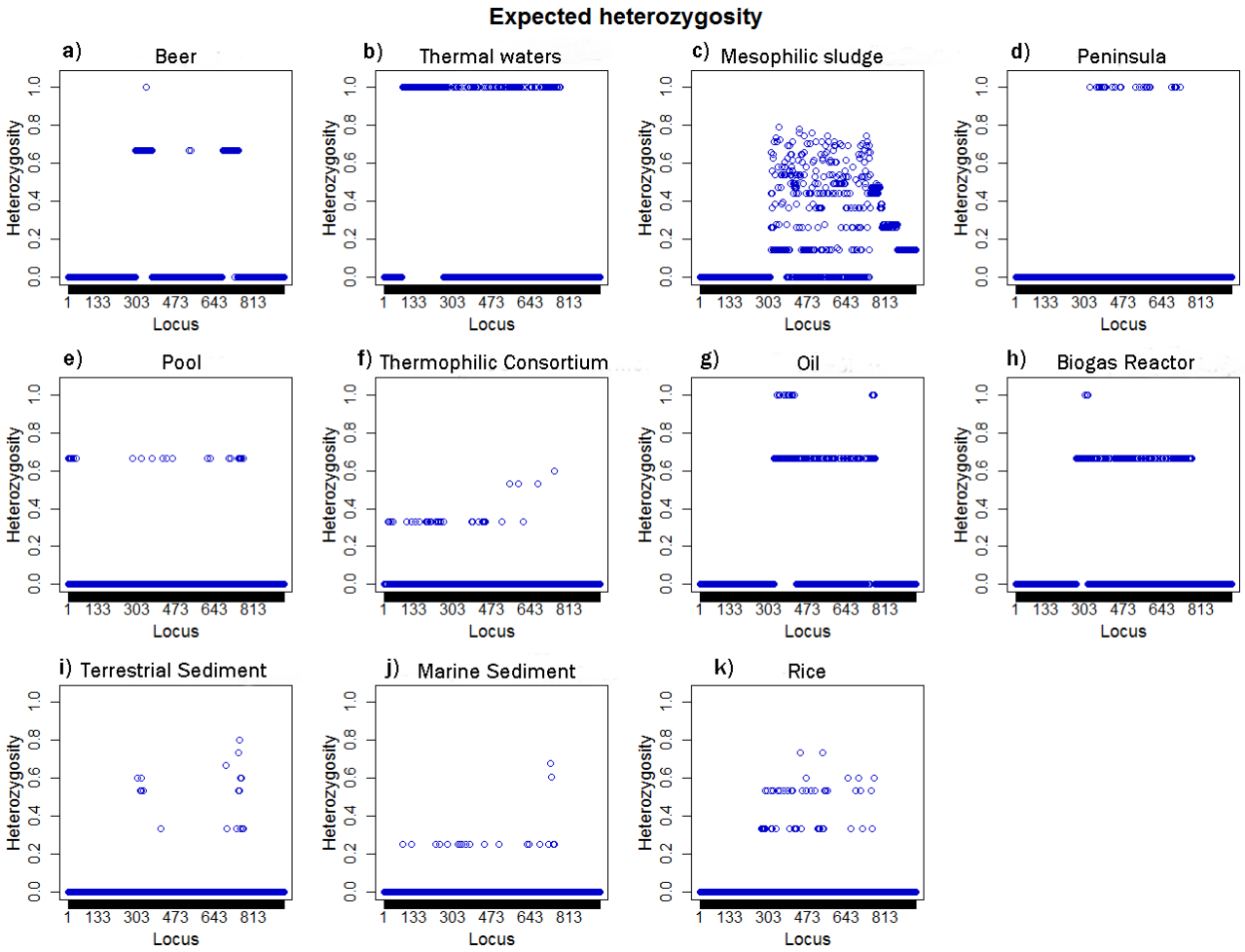

Figure 2. Expected heterozygosity of the *mcrA* gene groups.

Gaps, although are not nucleotides, interfere substantially at the calculation of heterozygosity, as it is a random choice method. The groups Pool (e) (0.666), Oil (g) (0.697), and Biogas Reactor (h) (0.693) also indicate high-expected heterozygosity (Figure 2).

The Mesophilic sludge group (c) showed a higher number of divergent loci (positions). However, the expected average heterozygosity was low, with a value of 0.361. The values found in Thermophilic Consortium (f) (0.366), Terrestrial Sediment (i) (0.498), Marine Sediment (j) (0.289) and Rice (k) (0.461) showed low Heterozygosity averages.

Heterozygosity averages for groups c, f, i, j and k was probably due to the presence of more similar sequences in alignment. Despite higher nucleotide content, the bases were repeated in the analyzed positions, showing a low divergence (Figure 2).

The molecular diversity index among the *mcrA* groups is shown in Figure 3. This index corresponds to the expected heterozygosity when having diploid information, that is, the probability that two randomly chosen haplotypes are different in the sample, which are calculations from Theta (θ), a population parameter of genetic differentiation (Zolet, Segatto, Turchetto, Silva, & Freitas, 2013).

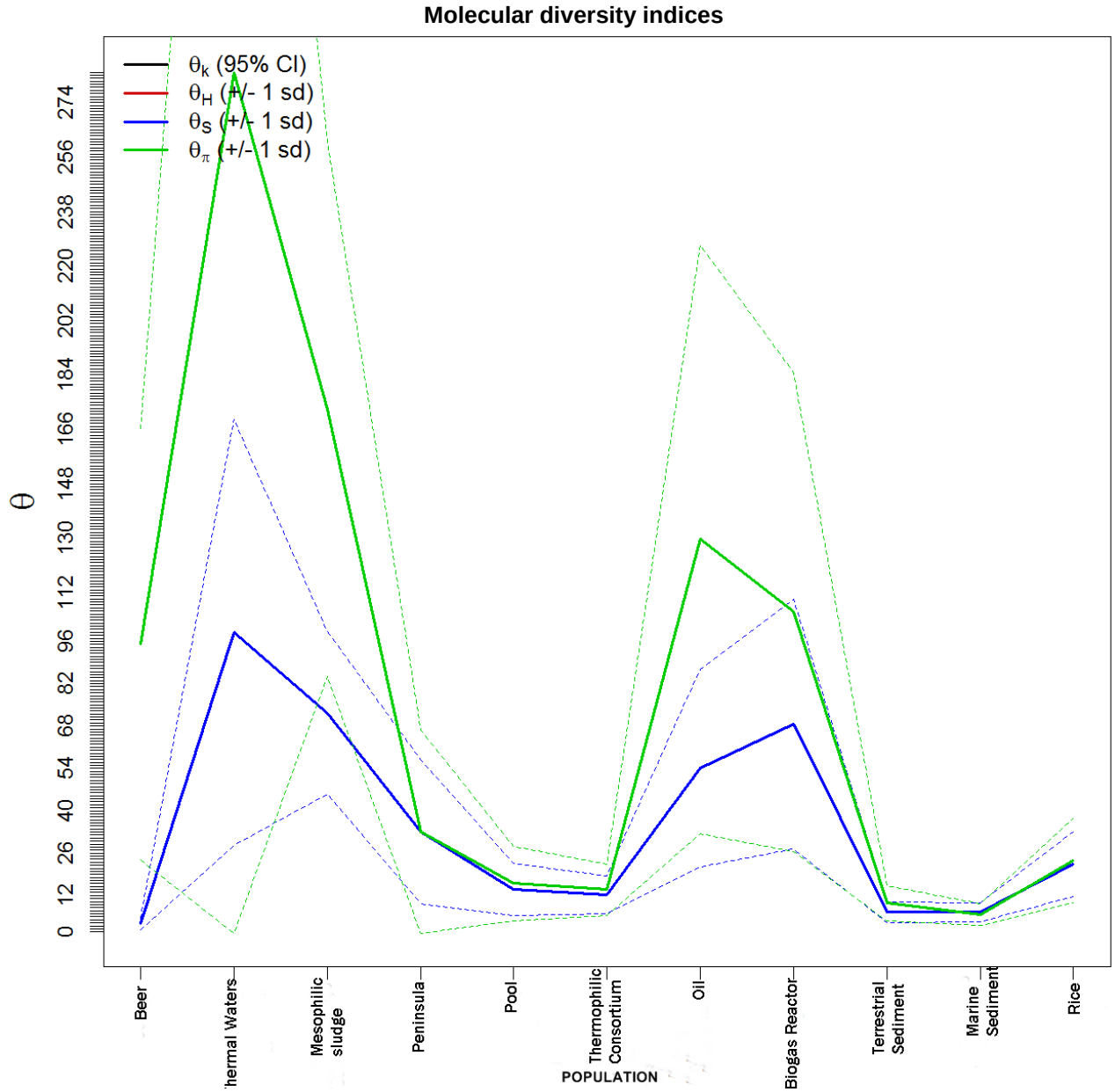

Figure 3. Molecular diversity index of groups belonging to the *mcrA* gene. Dotted lines show plus or minus standard deviation.

The blue line (θ_S) in the Figures 3 and Figure 5 is an estimate among the equilibrium relationship at an infinite locus (Watterson, 1975) and between the number of loci that suffered segregation (S), sample size (n) and theta (θ) for a non-recombinant DNA sample.

The green line (θ_π) is an estimated value from the equilibrium relationship between an infinite locus, the average number of differences between pairs (π) and theta (θ) (Tajima, 1983). The black and red lines do not apply, and they were not estimated for this type of DNA sequence study. The black line (θ_k) is an estimate from the equilibrium ratio of an infinite allele (Ewens, 1972) among the expected numbers of alleles, while in red (θ_H) calculates the relationship from homozygosity expected in an equilibrium population and between drift and mutation (Zolet, Segatto, Turchetto, Silva, & Freitas, 2013).

Based on the molecular diversity (Figure 3), the same pattern found in the results of (He) was observed, with Thermal waters and Mesophilic sludge showing higher diversity indexes among their sequences (θ_S), besides a higher average in differentiation among pairs (θ_π).

It is worth noting that in the diversity index, the gaps end up being analyzed, because the concept of molecular diversity is to estimate the possibility of two alleles being different from each other, so a nucleotide will always be different from a gap (Excoffier & Lischer, 2010). The lowest values of θ S and θ π were found in the terrestrial and marine sediment groups, corroborating the low heterozygosity found in Figure 2.

For the data generated from the *16S rRNA* gene sequences (Figure 4), only the Seafloor Mud (a), Maize Reactor (b) and BR Oil (c) groups were submitted to the diversity test, as they showed two or more sequences in the group, allowing an analysis of expected heterozygosity intragroup.

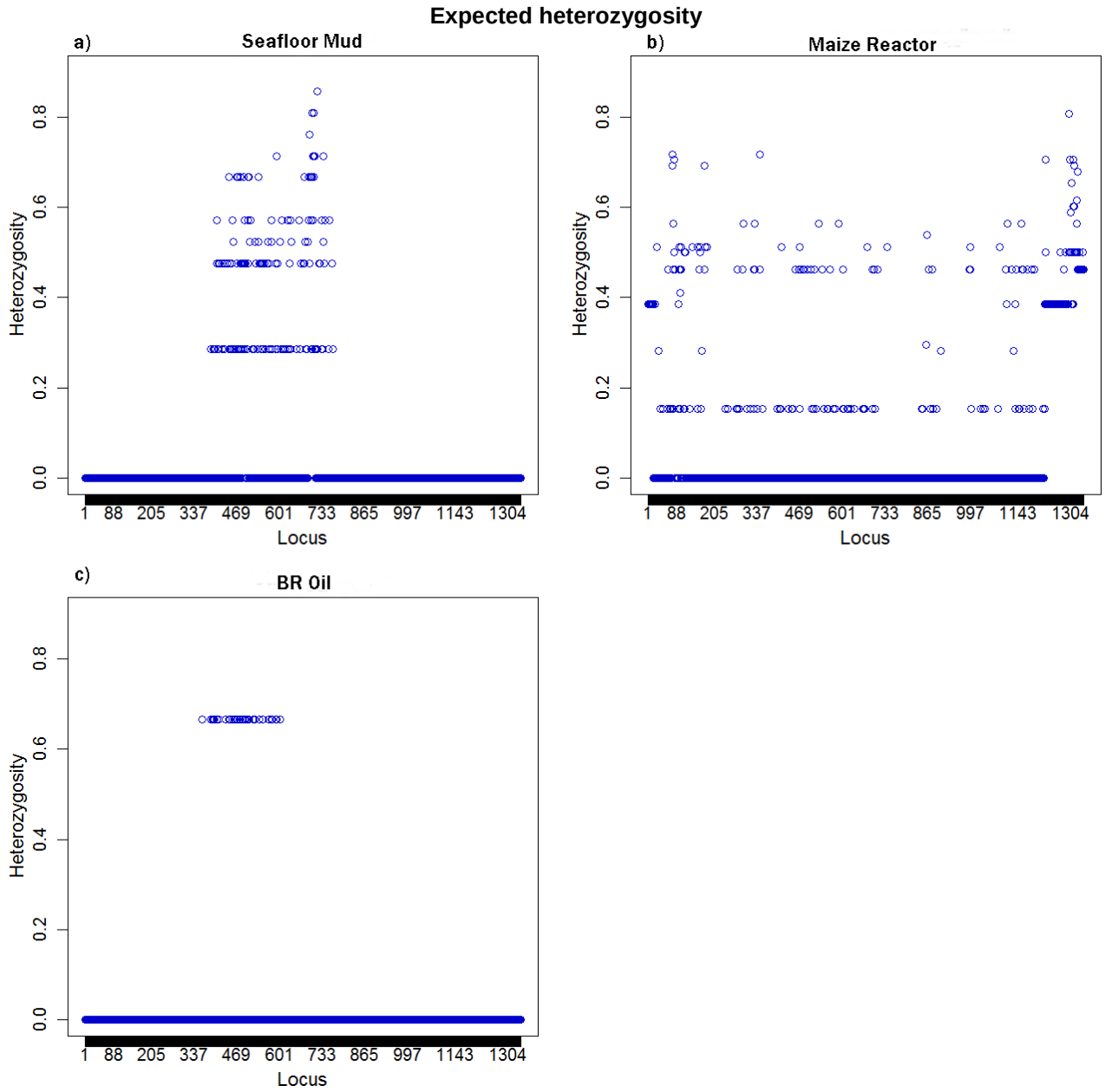

Figure 4. Expected heterozygosity of the *16S* rRNA gene groups.

The groups of Seafloor Mud (a) (composed of seven sequences) and Maize Reactor (b) (13 sequences) showed more divergent loci; however, the expected average heterozygosity was low, with a value of 0.460 and 0.376, respectively.

The value found in BR Oil (c) (three sequences) was 0.666 representing a higher average heterozygosity compared to the other groups. The observed pattern in (He) of the *mcrA* gene was also found for the *16S*

rRNA gene groups, i.e., groups with a higher number of sequences showed more similar positions, with lower heterozygosity.

Yet, the molecular diversity index for the *16S rRNA* gene did not show the same pattern obtained in Figure 5.

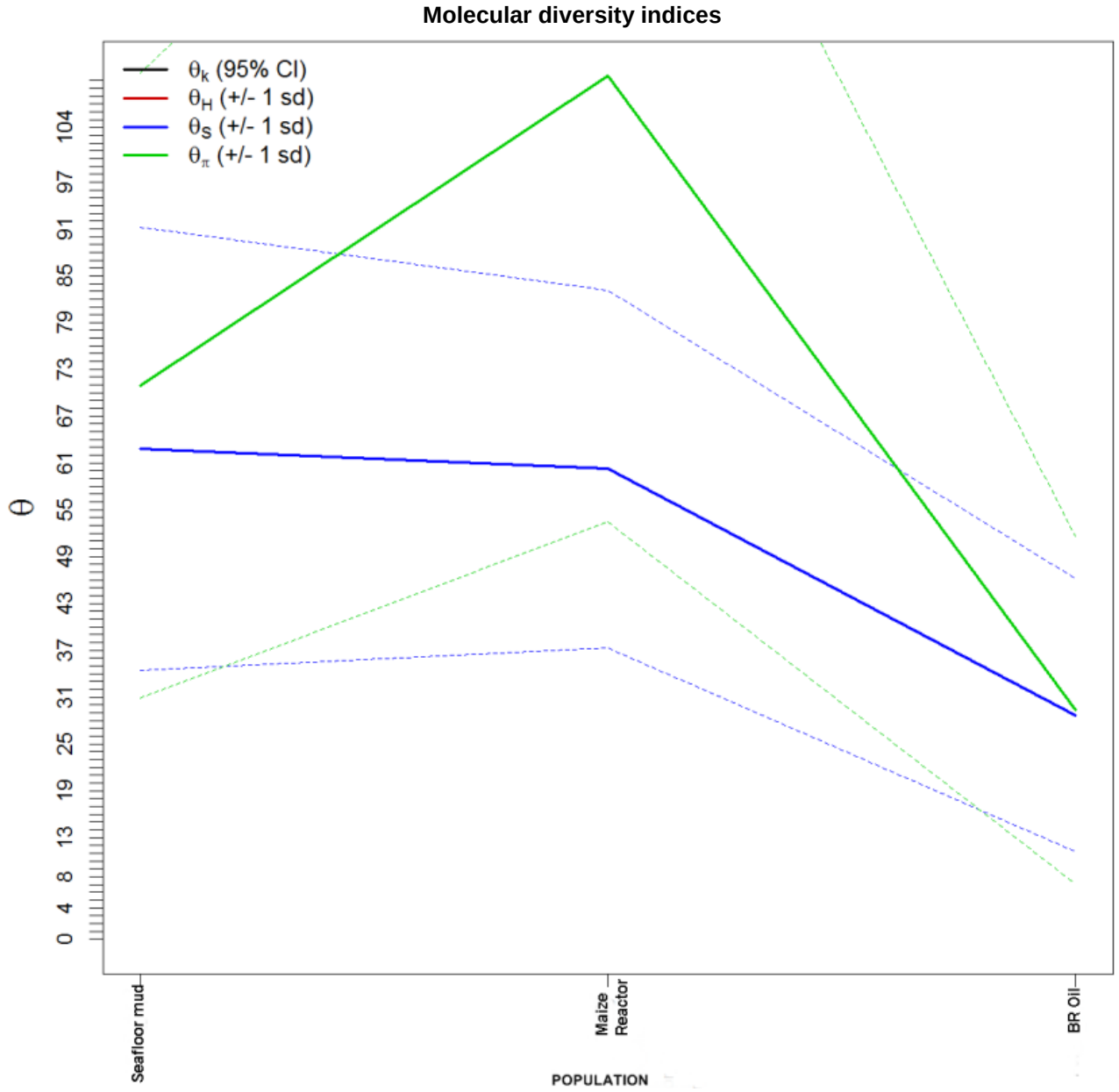

Figure 5. Molecular diversity index of groups belonging to the *16S* gene. Dotted lines show plus or minus standard deviation.

The Maize Reactor group had a high number of sequences in its group and with the largest size, with an average of 1293.9 bp, which significantly interfered with θ_{π}, since it calculates the largest average number of differences among pairs. In the group BR Oil, despite its high He (0.666), the average size of its sequences was 258 bp, with gaps present in the alignment that repeated in several positions, not indicating different haplotypes in the sample (Figure 5).

Conclusion

Descriptive analysis of genetic diversity generated by sequences deposited in databases allowed a detailed study of these molecules. It is known that the organisms in population are genetically distinct, and that, despite having similarities in their gene composition, the differences are essential for their adaptation to different environments.

Our findings for *16S* and *mcrA* sequences analyzed in this study suggest that groups with greater number of sequences, and longer base pair length were more similar; in the group with smaller base-pair sequences, automatic insertion of gaps by the tool used may have underestimated the showed differences.

Genetic diversity aims to understand the emergence and disappearance of alleles, due to spontaneous mutations or not. From population genomic studies, as in the present work, it is possible to establish a range of information about evolutionary processes, interactions between organisms and the environment, thus strengthening the context of population genetics.

Because archaeas are present in many environments and most often live in extremophilic conditions, understanding and contextualizing the vastness of molecular data that is added to databases (Big Data Era) becomes critical so that this information can be transformed into applicable knowledge.

Acknowledgements

The authors thank Dr. Eduardo Miranda Dantas, from UNIVASF, for the contributions with orthographic and grammatical revision.

References

Amaral, A. C., Steinmetz, R. L. R., & Kunz, A. (2019). O processo de biodigestão. In A. Kunz, R. L. R. Steinmetz & A. C. Amaral (Eds.), *Fundamentos da digestão anaeróbia, purificação do biogás, uso e tratamento do digestato (p. 13-26)*. Brasília, DF: Embrapa. doi: 10.21452/978-85-93823-01-5.2019.01_1

Dieffenbach, C. W., Lowe, T. M. J, & Dveksler, G. S. (1993). General concepts for PCR primer design. *PCR Methods and Applications, 3*, S30-S37. doi: 10.1101/gr.3.3.s30

Evans, P. N., Boyd, J. A., Leu, A. O., Woodcroft, B. J., Parks, D. H., Hugenholtz, P., & Tyson, G. W. (2019). An evolving view of methane metabolism in the Archaea. *Nature Reviews Microbiolgy, 17*(4), 219-232. doi: 10.1038/s41579-018-0136-7

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology, 3*(1), 87-112. doi: 10.1016/0040-5809(72)90035-4

Excoffier, L., & Lischer, H.E. L. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources, 10*(3), 564-567. doi: 10.1111/j.1755-0998.2010.02847.x

Hallam, S. J., Girguis, P. R., Preston, C.M., Richardson, P. M., & DeLong, E. F. (2003). Identification of methyl coenzyme M reductase A (*mcrA*) genes associated with methane-oxidizing Archaea. *Applied and Environmental Microbiology, 69*(9), 5483-5491. doi: 10.1128/AEM.69.9.5483-5491.2003

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews, 68*(4), 669-685. doi: 10.1128/MMBR.69.1.195.2005

Imachi, H., Nobu, M. K., Nakahara, N., Morono, Y., Ogawara, M., Takaki, Y., ... & Takai, K. (2020). Isolation of an archaeon at the prokaryote–eukaryote interface. *Nature, 577*(7791), 1-7. doi:10.1038/s41586-019-1916-6

Lee, C., Kim, J., Hwang, K., O'Flaherty, V., & Hwang, S. (2009). Quantitative analysis of methanogenic community dynamics in three anaerobic batch digesters treating different wastewaters. *Water Research, 43*(1), 157–165. doi: 10.1016/j.watres.2008.09.032

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics, 89*(3), 583-590.

Reeve, J. N. (1992). Molecular biology of methanogens. *Annual Review of Microbiology, 46*, 165–191. doi:10.1146/annurev.mi.46.100192.001121

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., & Sánchez-Gracia, A. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Molecular Biology and Evolution, 34*(12), 3299-3302. doi: 10.1093/molbev/msx248

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics, 105*(2), 437-460.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology, 7*(2), 256-276. doi: 10.1016/0040-5809(75)90020-9

Yakovchuk, P., Protozanova, E., & Frank-Kamenetskii, M. D. (2006). Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Research, 34*(2), 564-574. doi: 10.1093/nar/gkj454

Zolet, A. C. T., Segatto, A. L. A., Turchetto, C., Silva, C. P., & Freitas, L. B. (2013). *Guia prático para estudos filogeográficos*. Ribeirão Preto, SP: SBG.