ISSN 1064-2293, Eurasian Soil Science, 2015, Vol. 48, No. 4, pp. 410–415. © Pleiades Publishing, Ltd., 2015.
Original Russian Text © T.I. Chernov, A.K. Tkhakakhova, O.V. Kutovaya, 2015, published in Pochvovedenie, 2015, No. 4, pp. 462–468.

SOIL BIOLOGY

Assessment of Diversity Indices for the Characterization of the Soil Prokaryotic Community by Metagenomic Analysis

T. I. Chernov, A. K. Tkhakakhova, and O. V. Kutovaya

Dokuchaev Soil Science Institute, per. Pyzhevskii 7, Moscow, 119017 Russia
e-mail: chern-off@mail.ru

Received September 1, 2014

Abstract—The diversity indices used in ecology for assessing the metagenomes of soil prokaryotic communities at different phylogenetic levels were compared. The following indices were considered: the number of detected taxa and the Shannon, Menhinick, Margalef, Simpson, Chao1, and ACE indices. The diversity analysis of the prokaryotic communities in the upper horizons of a typical chernozem (Haplic Chernozem (Pachic)), a dark chestnut soil (Haplic Kastanozem (Chromic)), and an extremely arid desert soil (Endosalic Calcisol (Yermic)) was based on the analysis of 16S rRNA genes. The Menhinick, Margalef, Chao1, and ACE indices gave similar results for the classification of the communities according to their diversity levels; the Simpson index gave good results only for the high-level taxa (phyla); the best results were obtained with the Shannon index. In general, all the indices used showed a decrease in the diversity of the soil prokaryotes in the following sequence: chernozem > dark chestnut soil > extremely arid desert soil.

Keywords: biodiversity, soil microbiota, Shannon index, Menhinick index, Margalef index, Simpson index, Chao1, ACE, Haplic Chernozem (Pachic), Haplic Kastanozem (Chromic), Endosalic Calcisol (Yermic)

DOI: 10.1134/S1064229315040031

INTRODUCTION

The diversity of soil microbiota is an important factor of the biological stability of soils and the intensity and direction of many biochemical processes in the soil; this is a biodiagnostic tool with wide potential. Most data on the microbial diversity in different soils were obtained by classical methods [2, 6]; however, molecular biological methods, which account for the multitude of uncultivated and rare soil microorganisms, have become more common in recent years [9, 10]. Metagenomic methods occupy a particular place. The study of soil metagenomes is based on the isolation of the total DNA from the sample followed by the sequencing of nucleotide sequences for their identification at different taxonomic levels. The development of sequencing methods of a new generation resulted in the wide application of metagenomics for assessing the diversity of microbial communities in samples of different nature, including soils, which are unique sources of the genetic and phenotypic diversity of microorganisms [18, 20].

Most attention of researchers is focused on the 16S rRNA gene, which is used for the determination of the taxonomic (phylogenetic) position of prokaryotes [21]. The similar nucleotide sequences of the 16S rRNA gene obtained by sequencing are combined into operational taxonomic units, which can be then classified as phylogenetic taxa of different levels; this is the principle of biodiversity analysis in metagenomics.

The diversity of biological communities is assessed using different indices: numerical parameters calculated from the number of taxa in the community and the number of organisms (in megagenomics, the number of sequences) in different taxa. It should be noted that the diversity includes two components: the richness (number of taxa) and the evenness (relative abundance of taxa) [7]. Some diversity indices combine these two parameters into a common measure, which allows comparing different communities in terms of diversity.

In metagenomics, numerous diversity indices are used, including the classical Shannon and Simpson indices widely used in ecology for the analysis of communities of higher organisms and the recent specific Chao1 and ACE indices [8–10]. However, while the suitability of any diversity indices, their variation ranges, and other parameters are well known in classical ecology, their applicability for metagenomics data is still insufficiently clear. The soil metagenome has a specific structure; it is characterized by an extremely high taxonomic diversity of microorganisms [20] and usually the absence of distinct dominants at the low taxonomic level; therefore, some classical diversity indices are poorly applicable for its analysis.

The simplest diversity indices ignoring the relative abundance of taxa include the Margalef index $D_{\mathrm{Mg}} = \dfrac{S-1}{\ln N}$ and the Menhinick index $D_{\mathrm{Mn}} = \dfrac{S}{\sqrt{N}}$ [7];

only the number of detected taxa S and the total number of organisms N are used for their calculation.

The Shannon index is most frequently used to characterize the diversity of communities; it is sometimes referred to as the Shannon–Weaver or Shannon–Wiener index: $H' = \sum_{i=1}^{S} p_i \log p_i$, where p_i is the relative abundance of the ith taxon, and S is the number of detected taxa. The logarithm base is frequently 2 or e, although any other number is also acceptable [7]. For the Shannon index, the number of taxa is a more important factor at $S < 10$; the role of evenness increases with the number of taxa [11]. In the case of the soil prokaryotic metagenome, where even the number of high-level taxa is relatively large, the Shannon index is mainly determined by the evenness of the taxon abundances. The evenness measure from the Shannon index is sometimes calculated separately using the ratio of the observed evenness to its maximum value $E = \frac{H}{H_{\max}} = \frac{H}{\ln S}$ [7], which is sometimes referred to as the Pielou index.

Another diversity index frequently used in ecology is the Simpson index, which is frequently determined as the probability of belonging to different taxa for two organisms randomly selected from an indefinitely large community [11]. The Simpson index is calculated from the formula $D = \sum \left[\frac{n_i(n_i - 1)}{N(N - 1)}\right]$, where n_i is the number of organisms in the taxon, and N is the total number of organisms [7]. The value of D decreases when the evenness of the taxa increases; therefore, the Simpson index is frequently used in the form $1 - D$ (called the probability of interspecies encounter) or in the form $1/D$ (called the inverse Simpson index or the Williams polydominance index) [11]. Both indices increase with increasing evenness of the taxon abundances in the community (i.e., with increasing diversity). Some authors note that the Simpson index is almost completely determined by the proportion of the one to two most abundant species [11]; therefore, the Simpson index and its derivatives ($1 - D$, $1/D$) can be considered as dominance measures: indicators of dominance of one or several species.

The simplest method of characterizing the community richness is the use of the number of detected taxa (usually, species). However, the use of this parameter in soil metagenomics is difficult because of, first, the uncertainty in the selection of a specific taxonomic level and, second, the extremely large number of species and genera, the complete detection of which requires the analysis of 20000 sequences and more per sample [20]. Therefore, the analysis of the bacterial metagenome frequently includes different estimates of the actual number of taxa, e.g., the Chao1 index, which is calculated from the formula $\text{Chao1} = S_{obs} + \frac{a^2}{2b}$ [13, 15], where S_{obs} is the number of detected taxa, a is the number of taxa containing one sequence (singletons), and b is the number of taxa containing two sequences (doubletons). Another parameter estimating the actual number of taxa—the abundance coverage estimator (ACE) [14, 19]—is calculated by a significantly more complex method using the taxon abundance level selected by the researcher (usually it is equal to 10 species) to separate the rare and common taxa. Both parameters, Chao1 and ACE, are widely used and considered as giving good results in the analysis of diversity [16].

The suitability of the above-mentioned diversity indices for characterizing the soil prokaryotic community from the pyrosequencing data and the description of the metagenome of the studied soils was assessed. In our opinion, a suitable index should be characterized by a low spread of values for random sample sets, little depend on the sample set size, and be applicable for diversity analysis at different taxonomic levels.

OBJECTS AND METHODS

Three soil types of high biological activity that compose a zonal series from the steppe to the desert zone were selected for the analysis. Samples from the upper humus-accumulative horizons of a typical chernozem (Haplic Chernozem (Pachic) [5]) and a dark chestnut soil (Haplic Kastanozem (Chromic) [5]), as well as the surface K horizon (biological crust) of an extremely arid desert soil (Endosalic Calcisol (Yermic) [3]), were analyzed. The WRB soils names are given according to [4] with consideration for the third edition of the WRB [17].

The samples of a typical chernozem were taken in Kursk oblast (51°34′27.8″ N, 36°05′67.2″ E); the dark chestnut soil was sampled in Volgograd oblast (49°13′29″ N, 42°56′32″ E), and the extremely arid desert soil was sampled near the settlement of Taskarasu in the Uigurskii district of Almaty oblast in Kazakhstan (43°42′44.8″ N, 79°22′21.1″ E).

The samples were stored at –70°C. The DNA was isolated from 0.5 g of soil after mechanical breaking with glass beads in an extracting buffer containing 350 μL of solution A (200 mM sodium phosphate buffer and 240 mM guanidine isocyanate, pH = 7), 350 μL of solution B (500 mM Tris–HCl and 1% (w/v) SDS, pH = 7), and 400 μL of a 1 : 1 phenol–chloroform mixture. The sample breaking was performed using a Precellys 24 homogenizer (Bertin Technologies, France) with 3D motion at maximum power (speed of 6500 rpm) for 40 s. The obtained preparation was centrifuged at a maximum speed of 16000 rpm for 5 min. The water phase was separated and reextracted with chloroform. The DNA was precipitated by adding a similar volume of isopropanol.

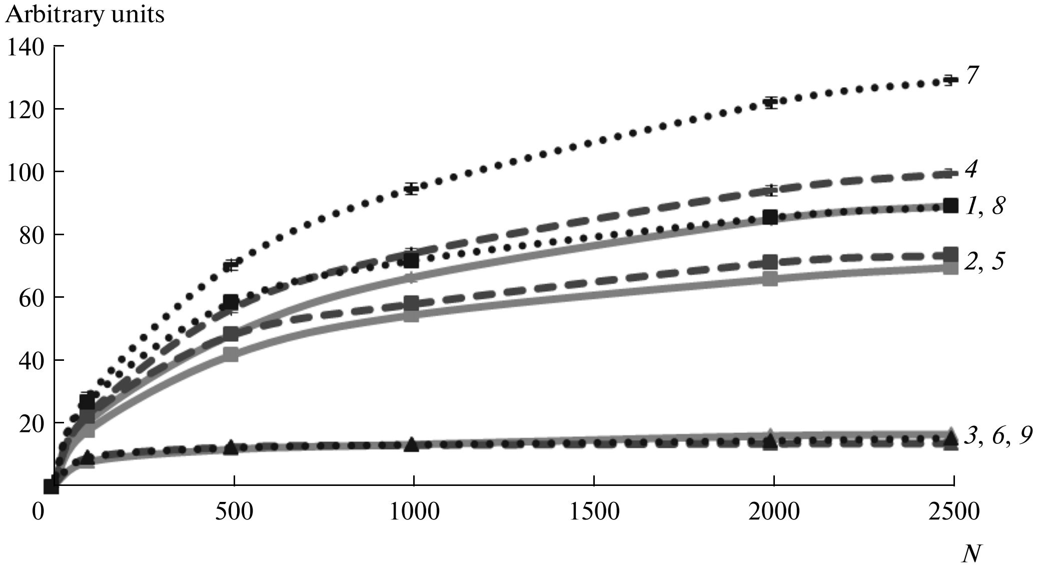

Fig. 1. Number of detected taxa (richness) as a function of the number of sequences (*N*). Here and below, extremely arid desert soil: (*1*) genera, (*2*) families, (*3*) phyla; dark chestnut soil: (*4*) genera, (*5*) families, (*6*) phyla; typical chernozem: (*7*) genera, (*8*) families, (*9*) phyla.

After centrifugation, the precipitate was washed with 70% ethanol and dissolved in water at 65°C for 5–10 min. The DNA was purified by electrophoresis in a 1% agarose gel followed by the separation of the DNA from the gel by sorption on silicon oxide [1].

Then, the purified DNA (10–15 ng) was entered into a polymerase chain reaction with the Encyclo polymerase (Evrogen, Russia) and universal primers to the V4 variable region of the 16S rRNA gene: F515 (GTGCCAGCMGCCGCGGTAA) and R806 (GGACTACVSGGGTATCTAAT) [12]. The reaction had the following temperature profile: 95°C, 30 s; 50°C, 30 s; 72 °C, 30 s; a total of 30 cycles were used. Oligonucleotide identifiers for the samples and service sequences necessary for the Roche pyrosequencing protocol were introduced into the primers. The sample preparation and sequencing were performed using a GS Junior instrument (Roche, Switzerland) according to the manufacturer's guidelines. At least 2500 sequences were obtained for each soil.

The alignment of the sequences and the determination of the taxonomic position from the RDP database were performed online using VAMPS (Visualization and Analysis of Microbial Population Structures, http://vamps.mbl.edu). From the obtained sequences, random sets of 100, 500, 1000, 2000, and 2500 sequences (30 sets for each size) were selected. For each set, the following diversity parameters were calculated: the number of taxa and the Chao1, ACE, Shannon, Margalef, Menhinick, $1/D$, and $1-D$ (where D is the Simpson index) indices. The diversity indices were considered for three taxonomic levels: genera, families, and phyla. The average value of each index was then determined, as well as the confidence interval ($p = 0.05$) for 30 sets of similar size.

The diversity indices as functions of the number of sequences in the set were plotted for the three analyzed taxonomic levels.

RESULTS AND DISCUSSION

The analysis of rarefaction curves describing the richness (the number of taxa) as a function of the number of sequences N showed the incomplete determination of the taxonomic composition for the sets of 2500 sequences at the level of genera and families (Fig. 1). For all three soils, the curves do not reach a plateau, which indicates the presence of a significant number of undetected taxa in the community. However, the gently sloping shapes of the rarefaction curves point to the evenness of the community and the absence of reliably dominant bacterial genera or families. A sufficiently complete determination of the phyla is observed already for the set of 500 sequences.

The Chao1 and ACE indices (Figs. 2a, 2b), which estimate the total number of taxa in the community, show about 10–15% of the families and genera unrevealed by the analysis. The Chao1 index (taking into account the number of taxa with one and two sequences) gives a higher estimate for the number of taxa in the community than the ACE index (which considers the taxa containing less than 10 sequences as rare). The distribution of soils by these indices is similar to that by the numbers of the detected taxa.

The analysis of the Shannon index (Fig. 2c) showed a low dependence of this parameter on the size of the

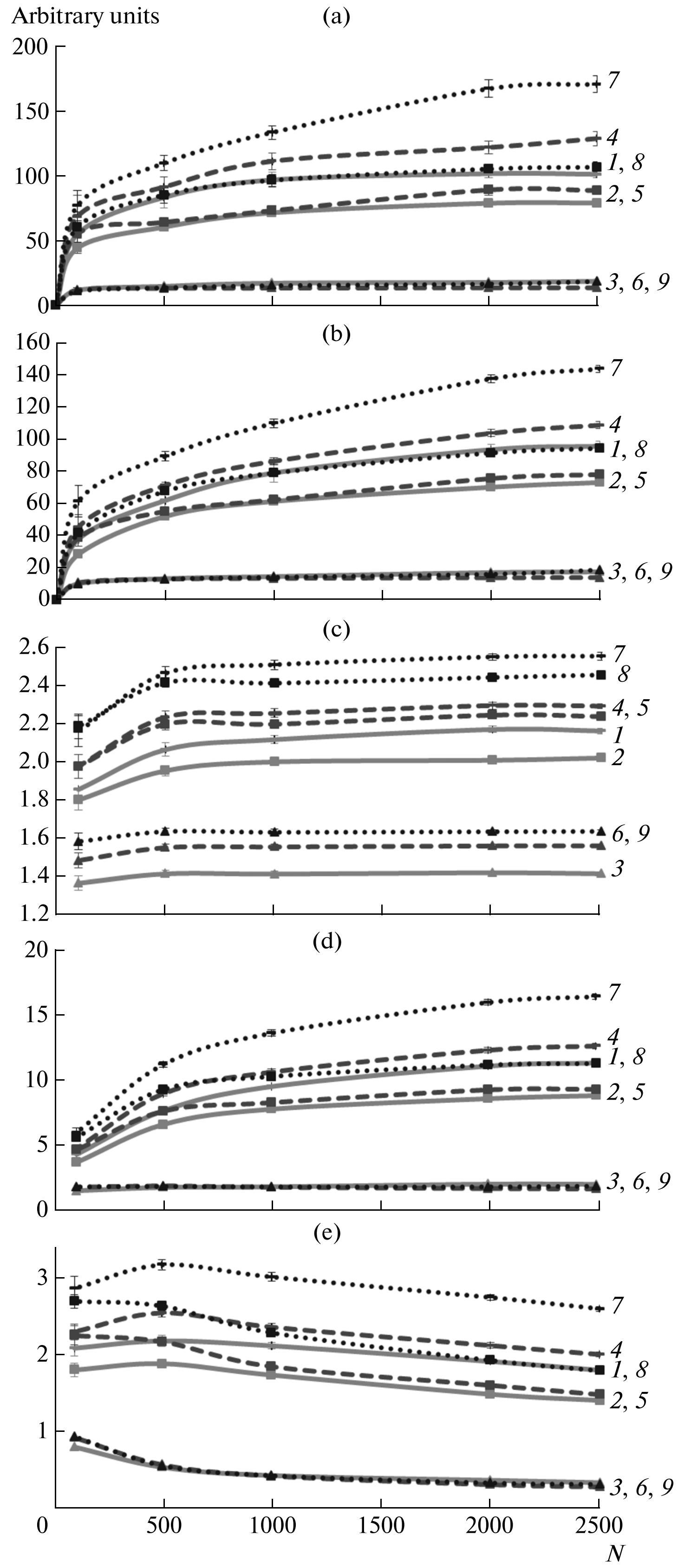

Fig. 2. Index values as functions of the number of sequences: (a) Chao1; (b) ACE; (c) Shannon index; (d) Margalef index; (e) Menhinick index.

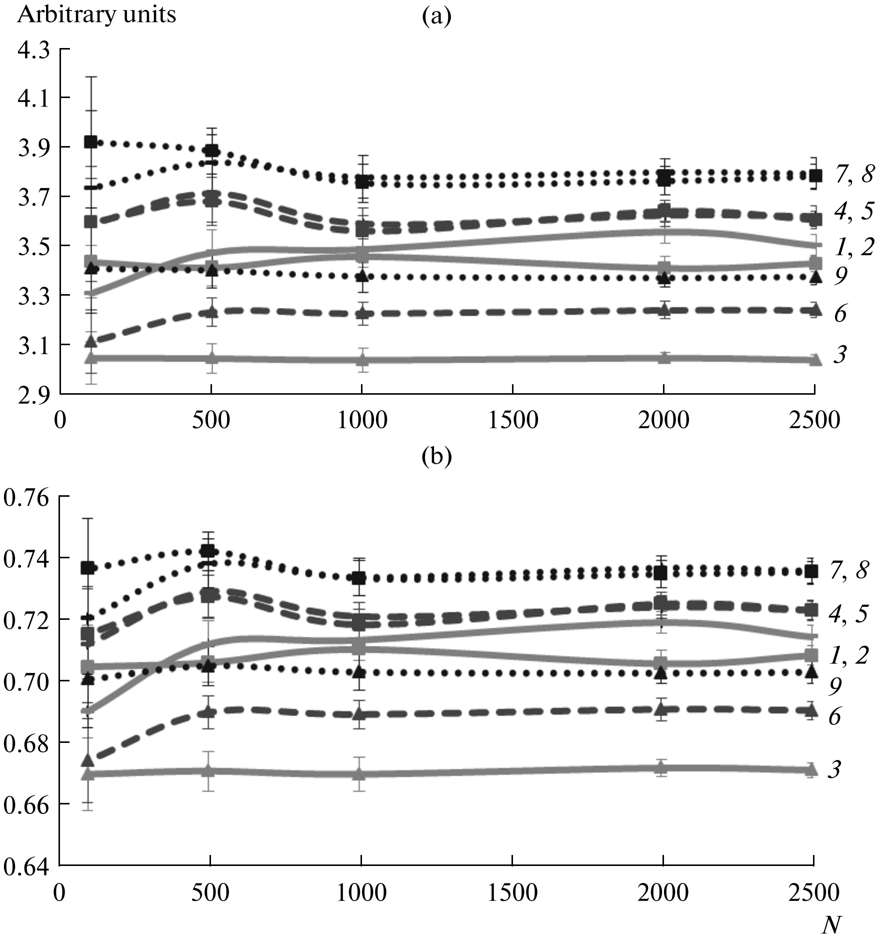

Fig. 3. Values of indices (a) $1/D$ and (b) $1 - D$ as functions of the number of sequences.

sample set and a narrow spread of values between the sets of the same size. The sets of 500–1000 sequences were sufficient for the determination of the Shannon index, the value of which remains almost unchanged when the set size increases to 2500 sequences. The communities of the three soils reliably differ in the Shannon index and compose the following series by decreasing diversity for all the taxonomic levels (genera, families, and phyla): typical chernozem > dark chestnut soil > extremely arid desert soil.

The Margalef and Menhinick indices (Figs. 2d, 2e) show that the distribution of the communities by diversity is almost similar to those for the species abundance parameters (the number of detected taxa and the Chao1 and ACE indices) for the set of 2500 sequences. For large sets (common in metagenomics), the number of sequences little affects the values of these indices, which are almost completely determined by the number of detected taxa. In distinction from the other studied indices, the decrease of the set size results in the overestimation of the diversity at the use of the Menhinick index, which can indicate the low suitability of this index for the analysis of large sets with high $\sqrt{N}$ values.

The inverse Simpson index and the value of $1 - D$ (Fig. 3) little vary among the metagenomes of the three soils at the levels of families and genera. This is related to the absence of distinct dominant taxa on these taxonomic levels, which is typical for the soil microbial community, although these are the few dominant taxa that determine the value of the Simpson index. The analysis at the phylum level showed significantly better separation by diversity on the basis of both Simpson index derivatives: the absolute dominants can be separated among the large taxa, in distinction from the levels of genera and families.

CONCLUSIONS

The Shannon index showed the least dependence on the size of the sample set (more than 500 sequences), a narrow spread of the values within the community, and good discrimination of all the soils at all the taxonomic levels. The number of taxa and the Chao1, ACE, Margalef, and Menhinick indices showed almost similar discriminations of the communities by diversity at the levels of genera or families; however, these indices are the least suitable for the discrimination of communities at the level of phyla. The Simpson index derivatives gave good results for the discrimination of the communities at the level of phyla, but they are unsuitable for the analysis at the low taxonomic levels.

In spite of the different methods of diversity assessment, all of the indices used showed a decrease of the diversity in the following soil series: typical chernozem > dark chestnut soil > extremely arid desert soil, which agrees with the general concepts of the microbial communities in these soils; however, we consider the Shannon index to be the most suitable for assessing the diversity of the soil bacterial metagenome.

ACKNOWLEDGMENTS

This work was supported in part by the Russian Foundation for Basic Research, project no. 14-26-00097.

REFERENCES

1. E. E. Andronov,, A. G. Pinaev, E. V. Pershina, and E. P. Chizhevskaya, *Isolation of DNA from Soil Samples: Methodological Guidelines* (All-Russia Research Institute of Agricultural Microbiology, St. Petersburg, 2011) [in Russian].
2. E. S. Vasilenko, O. V. Kutovaya, A. K. Tkhakakhova, and A. S. Martynov, "Changes in the intensity of soil biological processes as dependent on the size of aggregates in the migration-micellar chernozem," Byull. Pochv. Inst. im. V.V. Dokuchaeva, No. 73, 85–97 (2014).
3. Yu. G. Evstifev, "Soils of extremely arid regions in the People's Republic of Mongolia," in *The V Congress of Soviet Soil Science Society, Abstracts of Papers* (Minsk, 1977), No. 6, pp. 172–175.
4. *Unified State Register of Soil Resources of Russia, Ver. 1.0* (Dokuchaev Soil Science Institute, Moscow, 2014) [in Russian].
5. *Classification and Diagnostics of Soils in the Soviet Union* (Kolos, Moscow, 1977) [in Russian].
6. O. V. Kutovaya, E. S. Vasilenko, and M. P. Lebedeva, "Microbiological and micromorphological characteristics of extremely arid desert soils in the Ili Depression (Kazakhstan)," Eurasian Soil Sci. **45** (12), 1147–1158 (2012).
7. A. Magurran, *Ecological Diversity and Its Measurement* (Princeton University Press, New Jersey, 1988).
8. V. V. Parfenova, A. S. Gladkikh, and O. I. Belykh, "Comparative analysis of biodiversity in the planktonic and biofilm bacterial communities in Lake Baikal," Microbiology (Moscow) **82** (1), 91–101 (2013).
9. E. V. Pershina, G. S. Tamazyan, A. S. Dol'nik, A. G. Pinaev, N. Kh. Sergaliev, and E. E. Andronov, "Analysis of the structure of microbial community in saline soils using high-performance sequencing," Ekol. Genet. **10** (2), 31–38 (2012).
10. E. L. Chirak, E. V. Pershina, A. S. Dol'nik, O. V. Kutovaya, E. S. Vasilenko, B. M. Kogut, Ya. V. Merzlyakova, and E. E. Andronov, "Taxonomic structure of microbial communities in different types of soils according to the high-performance sequencing of the libraries of 16S rRNA gene," Sel'skokhoz. Biol., No. 3, 100–109 (2013).
11. V. K. Shitikov and G. S. Rozenberg, "Analysis of biological diversity: an attempt of formal generalization, in *Structural Analysis of Ecological Systems. Quantitative Methods Applied in Ecology and Hydrobiology* (Samara Scientific Center, Russian Academy of Sciences, Tolyatti, 2005), pp. 91–129.
12. S. Bates, D. Berg-Lyons, J. G. Caporaso, W. A. Walters, R. Knight, and N. Fierer, "Examining the global distribution of dominant archaeal populations in soil," ISME J., No. 5, 908–917 (2011).
13. A. Chao, "Nonparametric estimation of the number of classes in a population," Scand. J. Stat. **11**, 265–270 (1984).
14. A. Chao, M.-C. Ma, and M. C. K. Yang, "Stopping rules and estimation for recapture debugging with unequal failure rates," Biometrika **80**, 193–201 (1993).
15. R. K. Colwell and J. A. Coddington, "Estimating terrestrial biodiversity through extrapolation," Philos. Trans. R. Soc., B **345** (1311), 101–118 (1994).
16. J. Hortal, P. A. V. Borges, and C. Gaspar, "Evaluating the performance of species richness estimators: sensitivity to sample grain size," J. Anim. Ecol. **75**, 274–287 (2006). doi: 10.1111/j.1365-2656.2006.01048.x
17. IUSS Working Group WRB, *World Reference Base for Soil Resources 2014. International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*, in *World Soil Resources Reports No. 106* (FAO, Rome, 2014).
18. K. S. Kakirde, L. C. Parsley, and M. R. Liles, "Size does matter: application-driven approaches for soil metagenomics," Soil Biol. Biochem. **42**, 1911–1923 (2010). doi: 10.1016/j.soilbio.2010.07.021
19. S.-M. Lee and A. Chao, "Estimating population size via sample coverage for closed capture-recapture models," Biometrics **50**, 88–97 (1994).
20. L. F. W. Roesch, R. R. Fulthorpe, A. Riva, et al., "Pyrosequencing enumerates and contrasts soil microbial diversity," ISME J **1**, 283–290 (2007).
21. S. G. Tringe and P. Hugenholtz, "A renaissance for the pioneering 16S rRNA gene," Curr. Opin. Microbiol. **11**, 442–446 (2008). doi: 10.1016/j.mib.2008.09.011

Translated by K. Pankratova