ISSN 1062-3590, Biology Bulletin, 2021, Vol. 48, Suppl. 3, pp. S107–S117. © The Author(s), 2021. This article is an open access publication.
Russian Text © The Author(s), 2022, published in Izvestiya Akademii Nauk, Seriya Biologicheskaya, 2022, No. 1, pp. 70–81.

ECOLOGY

Metagenomics: A New Direction in Ecology

M. V. Vecherskii[a, *], M. V. Semenov[b], A. A. Lisenkova[c, d], and A. A. Stepankov[a]

[a] *Severtsov Institute of Ecology and Evolution, Russian Academy of Sciences, Moscow, 119071 Russia*
[b] *Dokuchaev Institute of Soil Science, Moscow, 119017 Russia*
[c] *Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, 119991 Russia*
[d] *Moscow State University, Moscow, 119234 Russia*
**e-mail: vecherskomy@mail.ru*

Received March 19, 2020; revised July 24, 2020; accepted July 24, 2020

Abstract—The prospects for application of metagenomic technologies in environmental studies are discussed. The advantages in investigating the taxonomic composition of aquatic and terrestrial ecosystems, as well as examples of trophic and phoric relationships found in ecosystems using the metagenomic approach, are described. The capabilities of metagenomics to study prokaryotic communities in complicated environments such as soils or animal intestines are shown. The role of relic DNA in the metagenome and the possibilities to study ancient organisms are highlighted. Particular attention is paid to the criticism of metagenomic technologies related to the low reproducibility of the sequencing data. Common methodological mistakes in bioinformatics processing of metagenomic data leading to misleading results are considered.

Keywords: metagenomics, NGS, bioinformatics, relic DNA
DOI: 10.1134/S1062359022010150

INTRODUCTION

The development of DNA sequencing technology provides the opportunity to study not only the genome of a particular organism, but also the metagenome, i.e., the entire set of genomes of a community. The metagenome is the set of genomes of all organisms found in an environmental sample, as well as extracellular DNA. Metagenomes of soil, water bodies, physiological material (intestinal content, pus, dental plaque, etc.), or industrial fermenters (Borbón-García et al., 2017; Chai et al., 2018; Wilson et al., 2019) are common objects of research. There are two approaches to defining metagenomics. In a strict sense, metagenomics means the analysis of the totality of genomes of a community of organisms (Riesenfeld et al., 2004). A broader interpretation of the term includes the study of not only genomes, but also individual genes within genomes as objects of metagenomics. When analyzing individual genes, the most crucial is phylogenetically significant sites used for taxonomic identification of community members (Ranjan et al., 2016). For example, a set of 16S rRNA genes is used for the determination of the taxonomic composition of prokaryotes; ITS sequences are used for fungi; 18S rRNA or a fragment of the mitochondrial *COI* gene, for animals. This method is called metabarcoding (from the analogy of species barcoding). Less commonly, a variety of functional genes are studied, for example, cellulases, nitrogenases, cytochromes, and antibiotic resistance factors (Ngara and Zhang, 2018). The data obtained by metagenomic analysis are predominantly qualitative; however, with a correctly designed experiment, it is possible to make quantitative assumptions. Metagenomic studies provide information on the taxonomic composition, trophic structure, and even phoric relationships in the community. In all cases, the researcher is not required to observe and identify individual members of the community. Moreover, the sample does not even have to include living organisms. The presence of their fragments or extracellular DNA is enough.

According to the method applied, metagenomic technologies are divided into two large groups: amplicon and shotgun techniques (Ma et al., 2014). For the first technique, the metagenome of a sample is only a template for amplification of one of the genes by the polymerase chain reaction (PCR). The sequencing itself is carried out on the obtained PCR products, and not on the original metagenomic DNA template. Amplification requires the use of primers flanking the target region of the nucleotide sequence. The imperfection of existing primers leads to an incomplete analysis of the community, and PCR produces a large number of artifacts. However, this method is relatively low-priced and relatively fast. In contrast, shotgun sequencing analyzes the entire metagenome. In this case, there are no mismatches associated with the choice of primers and imperfect PCR. However, this method is more laborious and expensive. In addition, if an object is characterized by a very high genetic

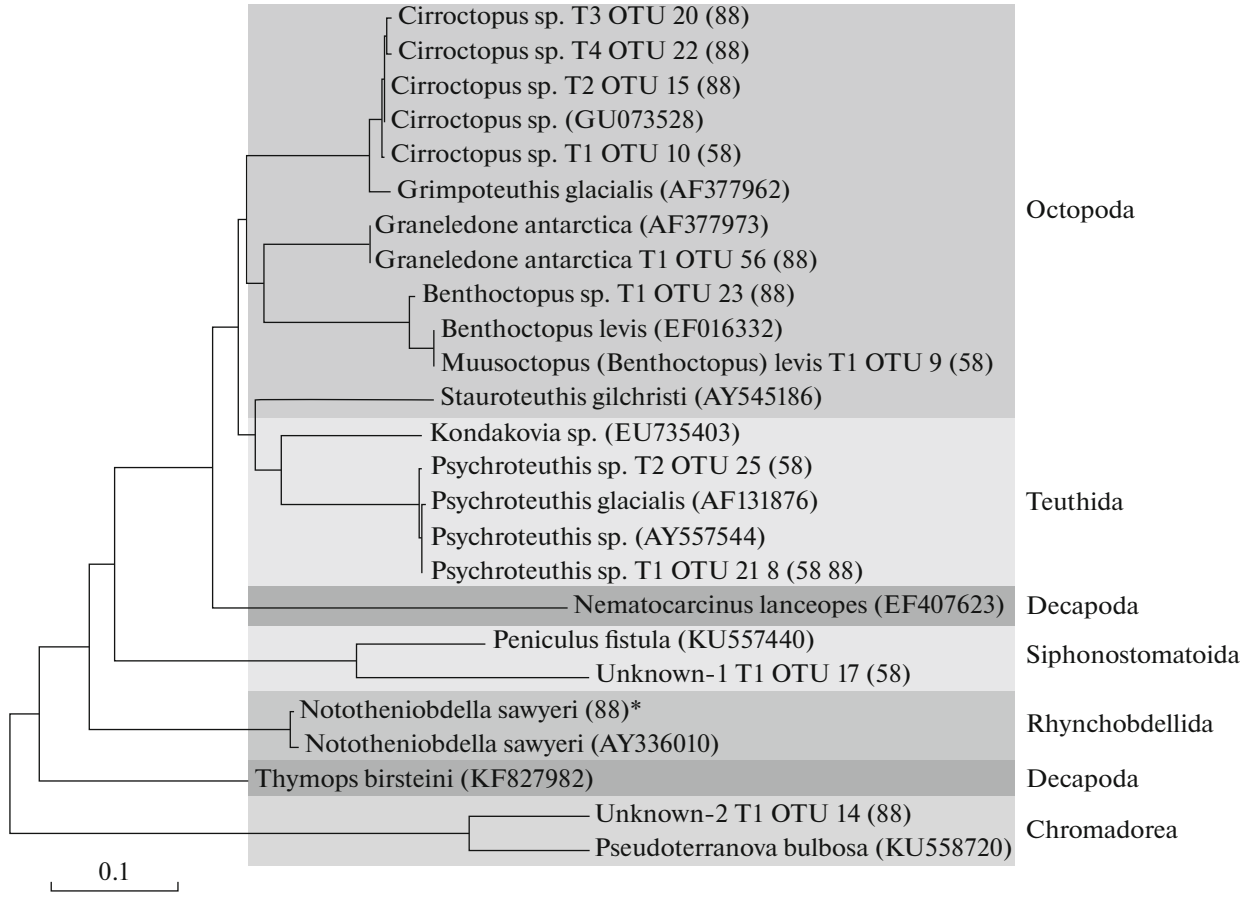

Fig. 1. Metagenomic analysis of the gut content of the Antarctic toothfish, *Dissostichus mawsoni* (according to Yoon et al., 2017).

diversity (for example, soil DNA), then shotgun sequencing faces the problem of completeness of reading the metagenome.

METAGENOMIC ANALYSIS FOR THE STUDY OF THE ECOLOGY OF COMMUNITIES

One of the important issues of the ecology of communities is the study of trophic interspecific interactions. While it is impossible to observe the nutritional process directly, such studies rely on the contents of the digestive system. However, the state of remains does not always allow one to determine the victim by morphological characteristics. At the same time, genetic markers in the gut content are preserved quite well (Pompanon et al., 2012). For example, it was shown (Yoon et al., 2017) that the diet of the largest south-polar fish, the Antarctic toothfish (*Dissostichus mawsoni*), includes at least 16 cephalopod prey species. In addition, four species of multicellular parasites of different types have been found in its stomach (Fig. 1).

Analysis of exogenous DNA provides a quantitative assessment of the composition of aquatic communities. Exogenous DNA can be released from the animal organism as part of exfoliating integuments, epithelial mucus, and feces. Exogenous DNA degrades in the external environment, but with sufficient numbers of the organism, its amount can be maintained at a technically determined level. In the metagenome of seawater samples, the number of copies of whale shark mitochondrial DNA was determined and then compared with the number of whale sharks themselves (Sigsgaard et al., 2016). The same samples were used to determine the number of mitochondrial DNA copies of the mackerel tuna, the eggs of which are consumed by the whale shark. It turned out that the copy number of mtDNA (the number of copies in the sample) reflects well the real ratio of the numbers of the corresponding fish (Fig. 2).

Metagenomic techniques are applied not only to animals, but also to plants. It became possible to study the composition of pollinated plants by the pollen metagenome on the integument of the pollinator insect. Thus, it was shown (Lucas et al., 2018) that the hoverfly pollinates flowers of 17 plant species (Fig. 3). Its food preferences were established for different biotopes in different seasons.

Metagenomic studies also provide a more accurate determination of the composition of symbiotic associations, including well-studied lichens. It turned out

that a significant part of the biomass of ascomycete lichens is represented not only by mycelial and bacterial components, but also by a yeast component (Spribille et al., 2016). The function of this basidiomycete yeast in the lichen is not yet clear, but it is obvious that its cells represent a significant proportion of cells of this organism.

Thus, analysis of ribosomal genes provides information on the taxonomic composition and diversity of the communities studied. However, these genes make up less than a tenth of a percent of the metagenome size. The other part of the metagenome can also be highly informative. The most interesting in this respect is the study of functional genes involved in various biochemical pathways. An example of such work is the study of biochemical pathways for butyrate synthesis by the intestinal microbiome. Butyrate is known to be a key nutrient for colon cells (Vital et al., 2017). A deficiency of butyrate leads to degradation of the large intestine, disbalance in the water and electrolyte balance, and an increased risk of cancer. Therefore, the changes in the microbiome that cause butyrate deficiency are actively studied. However, it is almost impossible to predict the actual yield of butyrate based on the taxonomic composition of the microbiome alone. More informative is the quantity of the functional genes involved in this biosynthesis. The full genome sequencing made it possible to establish the key metabolic pathways of butyrate in intestinal bacteria, as well as to link individual metabolic pathways with specific taxa of microorganisms (Vital et al., 2017).

One of the key general biological discoveries of recent years was made precisely by the full genome shotgun sequencing of the metagenome of prokaryotic communities of deep-sea hydrothermal vents (Zaremba-Niedzwiedzka et al., 2017). Metagenomic analysis revealed the presence of genes specific to eukaryotes and not found in prokaryotes. In particular, specific variants of cytoskeletal proteins have been found. Nevertheless, it turned out that these genes do not belong to eukaryotes, but to endemic archaea, which form a separate superphylum of Asgard or Asgardarchaeota. Subsequent bioinformatics analysis showed that Asgardarchaeota has traits specific for both prokaryotes and eukaryotes. It is very likely that precisely representatives of Asgardarchaeota made a successful endosymbiosis with Proteobacteria (future mitochondria) and formed the first eukaryotic organism about two billion years ago. Among the present-day Asgardarchaeota, the taxon Heimdallarchaeota is the closest to eukaryotes (Fig. 4). Such a late discovery of these unique organisms results from the impossibility of Asgardarchaeota cultivation and the inaccessibility of their habitats. This outstanding confirmation of the theory of symbiogenesis could be made only by metagenomics. However, it cannot be argued that now such works will be published regularly. The study of Asgardarchaeota cost an astronomical sum, significantly exceeding the annual budget of an ordinary

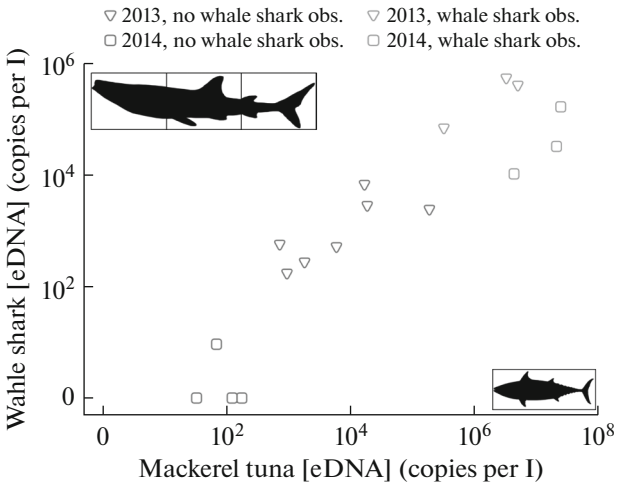

Fig. 2. Amount of exogenous DNA from the whale shark and mackerel tuna in seawater samples (according to Sigsgaard et al., 2016).

Russian scientific research institute. Furthermore, like any metagenomic study, this work does not involve the whole organisms, but only characterizes the set of genes in the environment. Therefore, the claim of whether or not the analyzed genes belong to a specific species is inevitably under criticism. There is the possibility of the existence of close associations of species that share some common set of genes. The contribution of extracellular DNA, including eukaryotic and degraded DNA, cannot be underestimated either.

METAGENOMIC ANALYSIS OF SOILS

Soil is one of the most difficult objects for ecological research (Lombard et al., 2011). There are three large groups of soil features that can be distinguished as an object of metagenomics study and the corresponding methodological problems: (1) heterogeneity of the soil, uneven distribution of microorganisms in soil aggregates, and the resulting problem of soil sampling; (2) adsorption of cells on soil particles, inhibition of amplification by humic substances of the soil, and the problem of DNA extraction and purification; (3) a very high diversity of communities, their different physiological status, and the presence of extracellular DNA.

Most of the soil metagenome is represented by prokaryotic DNA. Therefore, soil metagenomics is most often aimed at studying the taxonomic and functional-genetic structure of soil microbial communities. The high-throughput sequencing of soil DNA provides a list of taxa with their relative representation in the community, which can be used to obtain many indicators of the biological state of soils: (1) the taxonomic structure of the community; (2) the quantitative ratio of individual taxa and alpha diversity indices; (3) differences between communities of different soils (beta diversity); (4) the composition and proportion of indi-

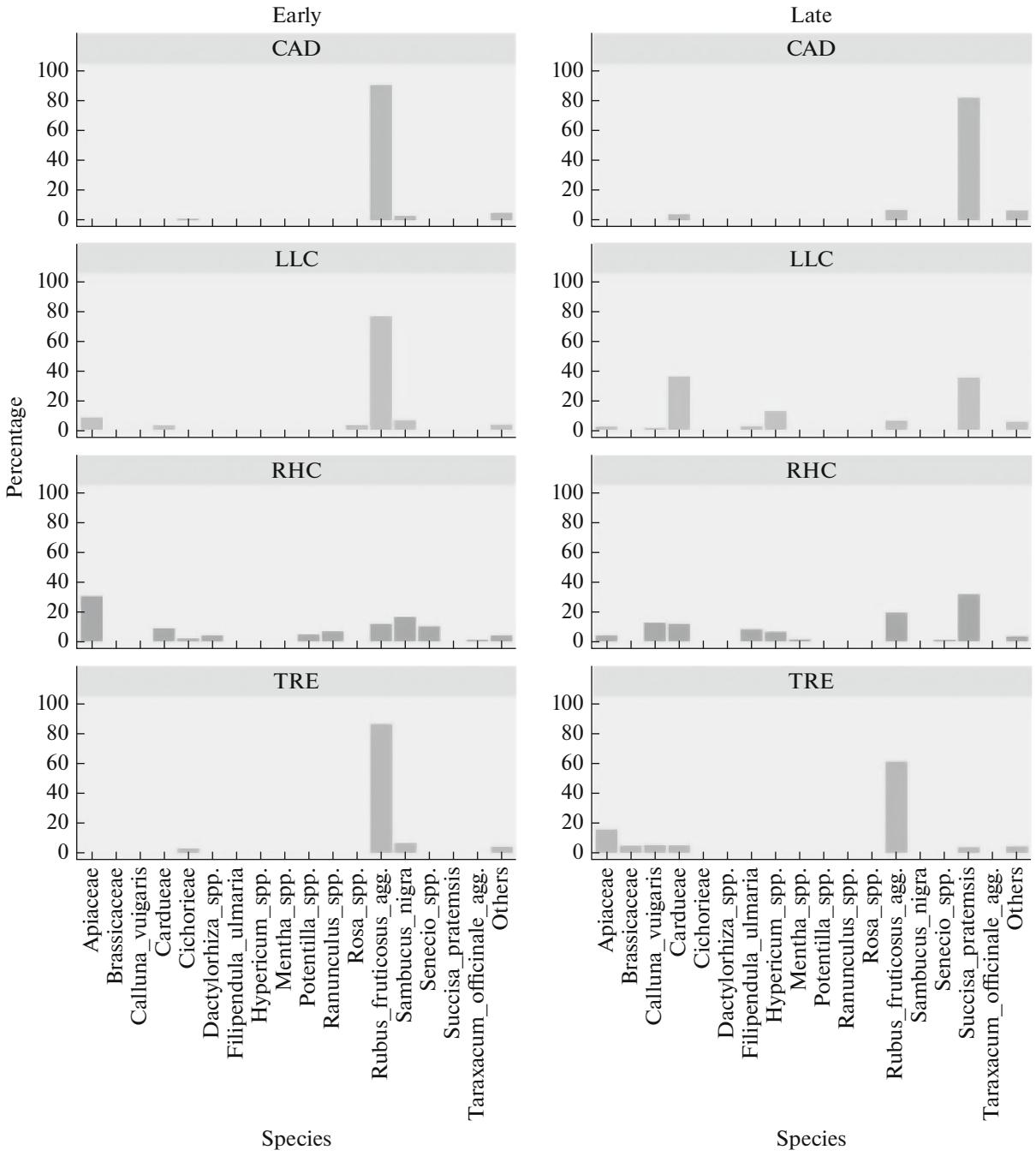

Fig. 3. Taxonomic composition of pollen on the covers of the hoverfly in early and late summer in different biotopes in western Greenland (according to Lucas et al., 2018).

cator taxa in the community; (5) the complexity and nature of interspecific networks; (6) the composition of communities associated with a certain biological process (analysis of functional genes). In addition, the analysis of functional genes reveals the correlations of individual taxa or communities as a whole with particular soil and ecological conditions.

Metagenomics plays a significant role in the study of ecology and the distribution of insufficiently studied taxa of soil microorganisms. The most striking examples of such studies are the phyla Thaumarchaeota and Verrucomicrobia, the representatives of which are not easily cultured in the lab. Until now, only three species of Thaumarchaeota have been isolated in a pure culture. Using RT-PCR and metabarcoding, it was shown that Thaumarchaeota is the most widespread archaea in the soil and on Earth in general, and, most likely, comprise the main oxidizers of soil

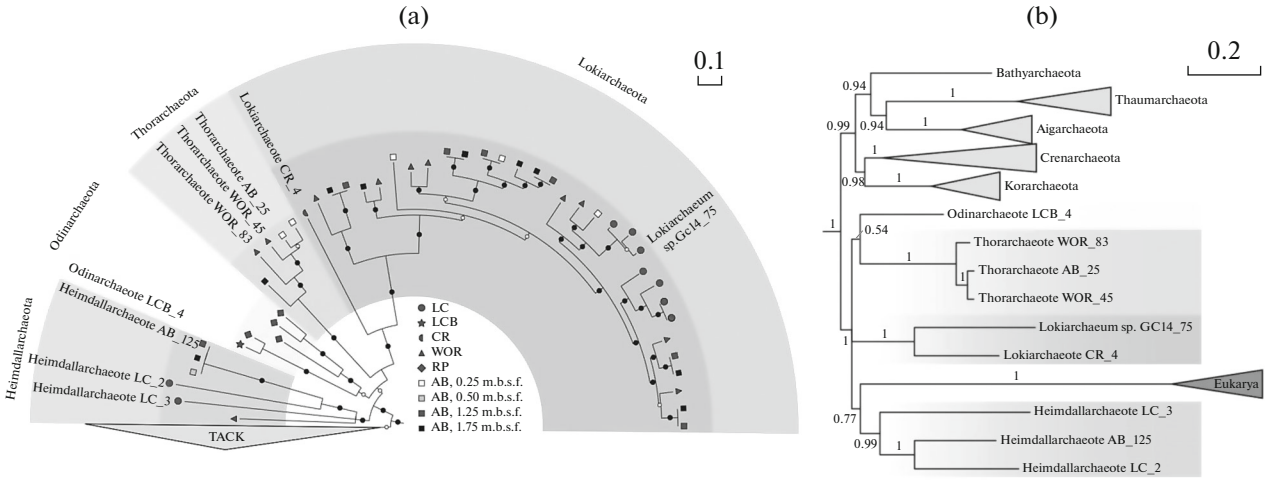

Fig. 4. Phylogenetic tree of modern Asgardarchaeota with an indication of the presumptive place of separation of the branch of eukaryotes (according to Zaremba-Niedzwiedzka et al., 2017).

ammonium (nitrifying agents) (Pester et al., 2011). Compared to nitrifying bacteria, Thaumarchaeota are well adapted to low ammonium concentrations, which gives them an advantage in oligotrophic environments typical for soils (Valentine, 2007). The proportion of Thaumarchaeota is, on average, 5–15% of the entire prokaryotic community.

Like Thaumarchaeota, the overwhelming majority of Verrucomicrobia species have not yet been isolated into pure cultures. Nevertheless, metabarcoding shows that, along with the Proteobacteria, Actinobacteria, and Acidobacteria, representatives of this phylum are the main dominants of prokaryotic communities, the proportion of which varies from 5 to 15%, and in chernozems, up to 25% and above (Semenov et al., 2018). The factors responsible for the distribution of Verrucomicrobia in soils are still unclear. For a long time, it was believed that Verrucomicrobia belongs to oligotrophic bacteria capable of growing under conditions of low carbon availability (Rocha et al., 2010; Senechkin et al., 2010; Eilers et al., 2012). Nevertheless, metabarcoding showed that Verrucomicrobia is characterized by high abundance in the upper soil horizons, with a higher availability of organic matter (Semenov et al., 2018). Metagenomic analysis revealed a relationship between the spatial distribution of Verrucomicrobia and carbon flux (Fierer et al., 2012), which casts doubt on the hypothesis of their oligotrophic nature. In addition, the number of Verrucomicrobia decreases during plowing and also reacts sharply to a decrease in the content of soil organic matter (Navarrete et al., 2015; Semenov et al., 2018).

In addition to taxonomic diversity, metagenomic analysis can provide information on the functionality of the soil microbiome. To archive this, a hierarchical structure is formed in which the identified genes of soil-inhabiting organisms are gathered into functional subsystems according to the principle of the unity of the function performed. Such a metagenome includes functional gene subsystems responsible for the metabolism of proteins, fats, and carbohydrates; virulence; respiration; the stress response, etc. At a lower hierarchical level, it is possible to analyze the proportion of genes responsible for the processes of the carbon and nitrogen cycle (for example, nitrification) and the synthesis or decomposition of certain compounds.

Comparison of the functional profiles of metagenomes of arctic and sand desert soils showed that, in terms of the ratio of gene subsystems, the difference between soils of such contrasting ecosystems is insignificant. Sand desert communities are characterized by a higher proportion of genes associated with osmoregulation and dormancy, as well as the metabolism of carbohydrates and aromatic substances. In the metagenome of arctic desert soils in comparison with sandy ones, a greater number of genes associated with the cycle of nutrients and catabolism of compounds associated with plants has been revealed (Fierer et al., 2012). At the same time, metagenomic analysis successfully has revealed the differences in soils of different types of land use. Compared to natural ecosystems, chernozem under arable land shows a lower diversity of archaea and fungi, as well as functional indicators (Gorbacheva et al., 2018). In the series tropical forest, ecosystem after deforestation, and arable land and pasture the lower abundance of microorganisms and the highest taxonomic and functional diversity were revealed in agricultural and pasture soils. These characteristics are important attributes for maintaining the functioning of the ecosystem after deforestation (Mendes et al., 2015). On the other hand, the ecosystem balance in the native forest ecosystem is maintained based on a lower diversity but a higher abundance of microorganisms. Another study showed that tillage and crop rotation significantly affect the subsystems of functional genes within the

full metagenome (Souza et al., 2015). In the soil with real tillage, there were more microorganisms associated with the decomposition of plant residues and the cycles of carbon and nitrogen, as well as eukaryotes. Soils with minimal tillage are characterized by a higher abundance of nitrogen-fixing rhizobia and archaea. It should be noted that in this work it was not possible to annotate almost half of the metagenomic sequences, and bacteria accounted for the bulk of all reads. The total contribution of archaea and eukaryotes to the annotated part of the studied metagenome was only 0.5% (Souza et al., 2015).

Additionally, soil microbial communities strongly respond to other agrogenic influences, for example, fertilization. The use of nitrogen fertilizers increases the number of copiotrophs, Proteobacteria and Firmicutes, and decreases the proportion of oligotrophic Acidobacteria, Nitrospirae, and Chloroflexi. The representatives of the order Rhizobiales, which include many associative nitrogen fixers, have been strongly negatively affected. Phosphorus fertilization increases the numbers of Armatimonadetes and Chlorobi (Ling et al., 2017).

METAGENOMIC ANALYSIS OF ANCIENT DNA

Metagenomics can study the genetic material not only of living organisms, but also of deceased creatures. Ancient DNA is extracted from samples dating back centuries and millennia. From a biochemical point of view, ancient DNA is characterized by extremely small fragment sizes, the length of which in most cases does not exceed 100–150 base pairs, although, of course, cases of better preservation of this type of DNA are known. In addition, ancient DNA has a large number of specific postmortem mutations. Ancient DNA is stored in many organs and tissues that can withstand time and environmental influences. The most popular research objects are bones, teeth, and hair, as they are best preserved under a wide range of conditions. The enamel and the jaw bone serve as protection for the teeth from rough mechanical influences. Different types of DNA (nuclear, mitochondrial, and in plants, DNA of chloroplasts) also have different properties and are preserved to varying degrees in certain parts of the organs and tissues. In the case of teeth, nuclear DNA is best preserved in the cells of the cementum, encapsulated in its mineral matrix, while in the pulp and dentin, the source of nuclear DNA is soft tissues, which are most susceptible to degradation at the beginning of the postmortem period. In contrast, multi-copy mtDNA is best preserved in dentin, especially in the area of the tooth roots (Higgins et al., 2015). Hair, due to the hydrophobic structure of keratin, is highly resistant to both exogenous DNA contamination and water (Gilbert et al., 2006). However, due to the peculiarities of hair development, the main remaining type of DNA in hair is mtDNA, which in large quantities (meaning copy number) can be found in the hair shaft along its entire length. Nuclear DNA is well preserved only in the root of living hair and in the nearest several centimeters of the shaft; although in hair that ahs fallen out, nuclear DNA is often not detected at all (Andréasson et al., 2006).

Soft tissues can also be preserved if there are favorable conditions for this, for example, permafrost. Due to low temperatures, leading to both better preservation of cellular components and DNA and the low activity of microorganisms and cellular enzymes, the biological material from permafrost is characterized by the highest resistance compared to other types of paleoDNA. In addition, permafrost preserves other study objects: fungi (Bellemain et al., 2013), bacteria (Willerslev et al., 2004), and ancient human populations (Green et al., 2006; Noonan et al., 2006; Fu et al., 2015). Ancient DNA is widely used in the fields of biology related to medicine. To trace the evolution and spread of the plague bacterium, the genomes of *Yersinia pestis* strains from 2800 to 4000 years old have been sequenced and phylogenetic trees for these strains have been built. The study revealed that the Yersinia murin toxin (*ymt*) gene appeared in the *Y. pestis* genome ~3000 years ago. The gene encodes phospholipase D, which protects the plague bacterium in the intestines of arthropods (in this case, fleas). Thus, the authors were able to predict fairly accurately the time when fleas became a vector for the spread of plague. Other authors have examined samples of Pleistocene and Holocene sediments from the Siberia permafrost, as well as samples of the cave and coastal sediments of New Zealand (Willerslev et al., 2003). DNA was extracted from samples weighing ~2 g, and amplified using primers for the chloroplast genes and mtDNA of animals, which resulted in DNA isolation from plants 300000–400000 years old and vertebrates of 20000–30000 years old. These made it possible to assess the species diversity in each locality, as well as to trace its dynamics for plants over long periods of time.

THE REPRODUCIBILITY PROBLEM AND INTERPRETATION OF METAGENOMIC DATA

The efficiency and high productivity of metagenomic analysis have led to a sharp increase in the number of works in this area. However, inaccuracy in the elaboration of metagenomic techniques leads to the irreproducibility of results, which can be observed even in highly ranked works. The greatest number of contradictions in the results obtained falls on the share of studies of the human gut microbiome (Poussin et al., 2018). Moreover, in the overwhelming majority of cases, the authors understand the human intestinal microbiome as the microbiome of human feces. Interest in this object is explained by the publication of a series of reports on the exceptional influence of the intestinal microbiome on the physiology of the organ-

ism as a whole. It was indicated that the human microbiota affects the development of diabetes mellitus, obesity, cancer, autoimmune diseases, human psychology through the synthesis of neurotransmitters, and, of course, immunity (Yan and Charles, 2017; Malan-Muller et al., 2018; Dicks et al., 2018). The logic of researchers was based on the correlation between the metagenome parameters and the frequency of some pathology. However, it is clear that correlation does not guarantee a direct relationship. For example, there is a good correlation between the age of the respondent and his/her microbiota. But so far, no article has appeared where the aging of a person is explained by the activity of gut bacteria. In addition, the observed differences may well be explained by the high microbiological heterogeneity of the stool samples. This heterogeneity, as well as numerous methodological assumptions, leads to the non-reproducibility of metagenomic results. For example, the journal *Cell* raised the question of the heritability of the human microbiome, as an organ that supposedly has the broadest functions (Goodrich et al., 2014). After conducting a twin experiment using metagenomic technologies, it turned out that the composition of the microbiome is genetically inherited. However, not long after, the opposite data was published in *Nature* (Rothschild et al., 2018). Such a discrepancy in the results published in journals with an impact factor of over 30 can be explained precisely by the impossibility of literal interpretation of metagenomic data. The reasons for the low reproducibility of metagenomic data are rooted in all stages of the experiment (Hoopen et al., 2017).

THE STAGE OF SELECTING A BIOLOGICAL SAMPLE

Samples for metagenomic studies are characterized by high heterogeneity. If we are talking about the microbiome, then its composition in the sample depends on the microstructure of the object, the presence of anaerobic zones, the presence of microscopic cavities, and the distance to active zones (for example root hairs, intestinal walls, earthworm coprolites). It is very difficult to perform randomization that takes into account all these factors. Sometimes randomization is simply impractical since the differences in the microbiome in the two nearest microloci (for example the surface of the intestinal wall and the intestinal lumen) are so great that they, in principle, should not be considered together. Furthermore, one should not forget about the speed of microbiome changes. For example, a one-hour difference in sampling can drastically change the nature of the microbiome. The idea of the microbiome as an unchanging and constant system is completely incorrect. The situation with the metagenome of animal communities is not much simpler. Different parts of an animal's body contain different amounts of DNA. In addition, DNA degrades at different rates in different parts of the animal. Its optimal preservation is achieved in the pulp of the tooth; the worst, in the desquamated epithelium. Therefore, differences in the representation of genes of two animals in the environment can be associated with both different numbers of animals and differences in the mechanisms of occurrence and deposition of their DNA in the environment. Unfortunately, many researchers, impressed by the technical novelty of the method, follow the methodological protocols too formally, which deceive with their simplicity. However, the protocols are developed by specialists unfamiliar with the specifics of each particular type of sample. An error in the choice of a biological sample for metagenomic analysis preliminarily causes biases in the results, which cannot be compensated for at later stages of the experiment.

THE STAGE OF TOTAL DNA ISOLATION

This stage is highly formalized. An error can occur only when comparing the results obtained using different methods of DNA extraction. Quite good reproducibility is achieved within one method.

AMPLIFICATION STAGE

In shotgun metagenomic studies, this stage and the corresponding errors are absent. Amplicon studies rely on multiple amplification of the target DNA fragment using PCR. At this stage, in addition to copying the target fragments, artifact fragments (for example, chimeras), which were initially absent in the sample, appear. In case of unsuccessful PCR, the proportion of artifacts may exceed 50%. Furthermore, it is necessary to select primers, special DNA molecules that mark the beginning and end of the target fragment. In different organisms, these areas are slightly different. Even the most versatile primers are well suited for amplifying the required DNA regions in some organisms and not suitable for others. Therefore, in practice, a "cocktail" of dozens of primers is used, which should potentially be suitable for all organisms. But the greater the number of primers used in parallel, the higher the number of PCR artifacts. It is impossible to distinguish an artifact from a target fragment. Artifact control is carried out at the stage of bioinformatics processing exclusively by indirect methods.

SEQUENCING STAGE

Sequencing can be performed on one of several hardware platforms. Most popular are Illumina, Pacific Bioscience, and IonTorrent. The sequencing process itself includes error occurrence. The IonTorrent platform has the highest level of hardware errors, but it is the most attractive in terms of cost. The results obtained on different platforms are characterized by low reproducibility with each other (Allali et al., 2017). The key way to increase reproducibility at the hardware level is to increase the sequencing depth, mea-

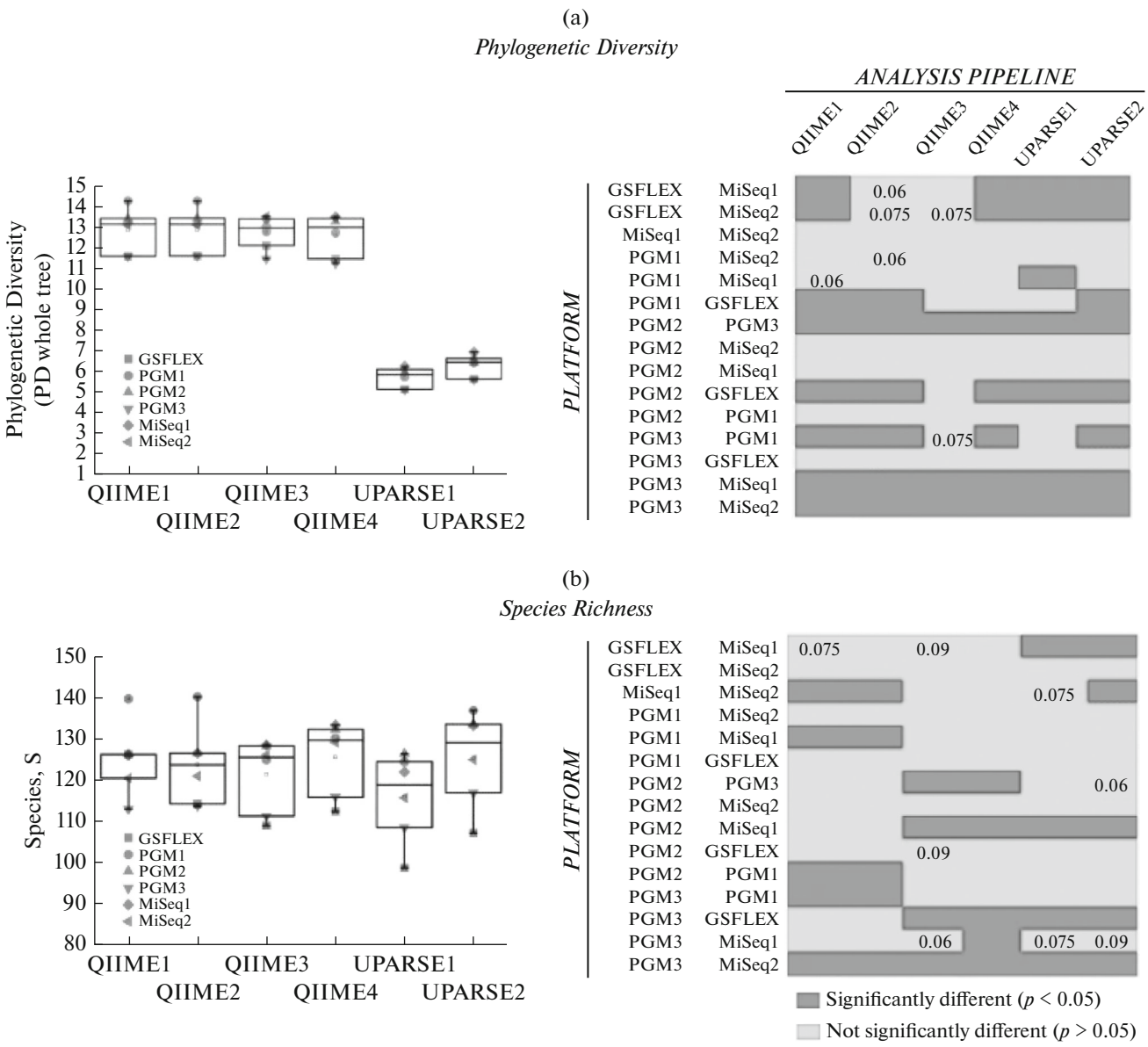

Fig. 5. Reproducibility of the results of metagenomic analysis of the sample using different hardware platforms for sequencing (vertical) and various mathematical data processing packages (horizontal) (according to Allali et al., 2017). (a) Phylogenetic diversity; (b) species richness.

sured in the number of reads per sample (Zaheer et al., 2018). If the works on the microbiome in the early 2010s were limited to 10000 reads, now the requirements have grown up to 50000 and even 120000 reads per sample. So it is preferable that all the sequences of the experiment are obtained on the same platform (at least in the same study). Moreover, this is also relevant when comparing our own results with the literature data.

THE BIOINFORMATIC DATA PROCESSING STAGE

The operation of the sequence device results in a list of nucleotide sequences that were in the sample and were flanked by primers, as well as a set of all sequencing artifacts. As a rule, each actually existing target fragment with a length of 200–400 bp due to errors is represented by a group of fragments, all members of which differ from each other by several nucleotides. Therefore, it is wrong to perceive each variant of a fragment as actually existing in the environment. This will falsely multiply the community biodiversity. To combat this (as well as to take into account intraspecific polymorphism), sequences are clustered into operational taxonomic units, OTUs. There are three main groups of clustering algorithms:

(1) Dynamic *de novo* algorithms. Such algorithms are suitable for clustering sequences when analyzing

new, previously unexplored communities that include rare taxa and even those absent in databases. They cannot be implemented on common computers and require high-performance processors.

(2) Greedy *de novo* algorithms. They are suitable for clustering sequences when analyzing new, previously unexplored communities, including rare and absent species in the databases. They can be implemented on any computer. However, they give a false overestimation of the diversity and distort the quantitative representation of clusters.

(3) Reference algorithms. They are suitable for clustering sequences when analyzing well-known communities, all members of which were previously analyzed separately, and the sequences were deposited into the database. Can be implemented on any computer. They give minimal distortion of results. They perform poorly with a low threshold of identity (see below).

(4) No clusterization. This is a method that is attracting increasingly greater attention due to the highest resolution. It requires a lot of personal involvement of the operator and therefore cannot be implemented on a sequenced-flow basis.

It is obvious that clustering seriously reduces the resolution of the method; therefore, the researcher must set the identity threshold (in % of differences in the nucleotide sequence) within which the OTU will be formed. A large number of artifacts requires an increased identity threshold. The standard threshold for barcoding for ribosomal genes is 97%, which corresponds to the genus/species level. It is very important that working with non-universal highly specialized primers requires an increased identity threshold; otherwise, it will devalue the high resolution of the primer. In rare cases, it can be raised to 100%, i.e., entirely without clustering. When working with very diverse or low-quality metagenomic data, using highly universal primers, the threshold is reduced by up to 93% (order–family level). When working with functional genes (cellulases, toxins, etc.), the threshold is usually also reduced to 93–95% (Ngara and Zhang, 2018). A poor choice of the clustering algorithm and identity threshold can be very detrimental to the performance. It is fundamentally important that all comparisons must be made only between results obtained using the same clustering method (Clooney et al., 2016). This is also true for comparisons with the literature data. In the work by Allali et al. (2017), it was clearly demonstrated that the analysis of the same sample using different mathematical packages and/or different sequencing platforms leads to data irreproducibility (Fig. 5). It should only be noted that the main tool of control is the exclusion of all nucleotide sequences occurring in a single copy (singletons). Another approach is to shift attention to higher taxa. Small errors in the nucleotide sequence of a fragment can transfer it to another genus of organisms, but not to another family or order.

Errors occurring at the listed stages of metagenomic research can unjustifiably increase the biodiversity of the analyzed communities and cause false correlations. The most effective way to deal with such errors is a control study of samples without the use of metagenomics technologies. Usually, high-ranked journals require the conclusion of any work to be confirmed by at least two experiments based on different methodological principles. For example, an increase in gene expression (shown by total RNA sequencing) should be confirmed by an increase in the amount of the target metabolite (shown by GC-MS). The same principle should be observed for metagenomic studies, but this does not happen in the overwhelming majority of them. Therefore, any reasoning about the composition and functions of the microbiome, arising only from metagenomic data, is largely speculative. Probably, the further development of metagenomic studies will consist in the control of phenotype parameters, since any genomic data are only indirectly related to the actual phenotype. More detailed information is provided by meta*transcriptomics*, the totality of all template RNA in the sample; by meta*proteomics,* the totality of all proteins; and, finally, by meta-*metabolomics*, which characterizes all organic substances, including small molecules. These approaches are sometimes called omics technologies, from the common English ending for all four terms: metagen*omics*, metatranscript*omics*, metaprote*omics*, and metametabol*omics*. The combination of omics technologies will serve as an internal control for any research. For example, the presence of characteristic protein isoforms will confirm the presence of a specific species in metabarcoding. The presence of a particular metabolite will confirm the detection of the corresponding functional gene. Without this, metagenomics will remain a promising but not yet reliable method.

ACKNOWLEDGMENTS

The authors are grateful to Academician of the RAS V.V. Rozhnov and O.L. Makarova for their support during the preparation of the manuscript.

FUNDING

The section of the article METAGENOMIC ANALYSIS OF SOILS was prepared with the financial support of the Russian Foundation for Basic Research (project no. 19-04-00315); the section of the article METAGENOMIC ANALYSIS OF ANCIENT DNA was prepared with the support of the Russian Foundation for Basic Research (project no. 18-34-00895).

COMPLIANCE WITH ETHICAL STANDARDS

The authors declare that they have no conflicts of interest. This article does not contain any studies involving animals or human participants performed by any of the authors.

OPEN ACCESS

REFERENCES

Allali, I., Arnold, J.W., Roach, J., Cadenas, M.B., Butz, N., Hassan, H.M., Koci, M., Ballou, A., Mendoza, M., Ali, R., and Azcarate-Peril, M.A., A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome, *BMC Microbiol.*, 2017, vol. 17, no. 1, p. 194.

Andréasson, H., Nilsson, M., Budowle, B., Lundberg, H., and Allen, M., Nuclear and mitochondrial DNA quantification of various forensic materials, *Forensic Sci. Int.*, 2006, vol. 164, no. 1, pp. 56–64.

Bellemain, E., Davey, M.L., Kauserud, H., Epp, L.S., and Boessenkool., S., Fungal palaeodiversity revealed using high-throughput metabarcoding of ancient DNA from arctic permafrost, *Environ. Microbiol.*, 2013, vol. 15, no. 4, pp. 1176–1189.

Borbón-García, A., Reyes, A., Vives-Flórez, M., and Caballero, S., Captivity shapes the gut microbiota of andean bears: insights into health surveillance, *Front. Microbiol.*, 2017, vol. 8, p. 1316.

Chai, Z.Y., He, Z.L., Deng, Y.Y., Yang, Y.F., and Tang, Y.Z., Cultivation of seaweed *Gracilaria lemaneiformis* enhanced biodiversity in a eukaryotic plankton community as revealed via metagenomic analyses, *Mol. Ecol.*, 2018, vol. 27, no. 4, pp. 1081–1093.

Clooney, A.G., Fouhy, F., Sleator, R.D., O'Driscoll, A., Stanton, C., Cotter, P.D., and Claesson, M.J., Comparing apples and oranges?: next generation sequencing and its impact on microbiome analysis, *PLoS One*, 2016, vol. 11, no. 2. e0148028.

Dicks, L.M.T., Geldenhuys, J., Mikkelsen, L.S., Brandsborg, E., and Marcotte, H., Our gut microbiota: a long walk to homeostasis, *Benef. Microbes*, 2018, vol. 9, no. 1, pp. 3–20.

Eilers, K.G., Debenport, S., Anderson, S., and Fierer, N., Digging deeper to find unique microbial communities: the strong effect of depth on the structure of bacterial and archaeal communities in soil, *Soil Biol. Biochem.*, 2012, vol. 50, pp. 58–65.

Fierer, N., Leff, J.W., Adams, B.J., Nielsen, U.N., Bates, S.T., Lauber, C.L., Owens, S., Gilbert, J., Wall, D., and Caporaso, J.G., Cross-biome metagenomic analyses of soil microbial communities and their functional attributes, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, vol. 109, no. 52, pp. 21390–21395.

Fu, Q., Hajdinjak, M., Moldovan, O.T., Constantin, S., Mallick, S., et al., An early modern human from Romania with a recent Neanderthal ancestor, *Nature*, 2015, vol. 524, no. 7564, pp. 216–219.

Gilbert, M.T.P., Menez, L., Janaway, R.C., Tobin, D.J., Cooper, A., and Wilson, A.S., Resistance of degraded hair shafts to contaminant DNA, *Forensic Sci. Int.*, 2006, vol. 156, nos. 2–3, pp. 208–212.

Goodrich, J.K., Waters, J.L., Poole, A.C., Sutter, J.L., Koren, O., Blekhman, R., Beaumont, M., Van Treuren, W., Knight, R., Bell, J.T., Spector, T.D., Clark, A.G., and Ley, R.E., Human genetics shape the gut microbiome, *Cell*, 2014, vol. 159, no. 4, pp. 789–799.

Gorbacheva, M.A., Melnikova, N.V., Chechetkin, V.R., Kravatsky, Y.V., and Tchurikov, N.A., DNA sequencing and metagenomics of cultivated and uncultivated chernozems in Russia, *Geoderma Regional*, 2018, vol. 14. e00180.

Green, R.E., Krause, J., Ptak, S.E., Briggs, A.W., Ronan, M.T., Simons, J.F., et al., Analysis of one million base pairs of Neanderthal DNA, *Nature*, 2006, vol. 444, p. 330.

Higgins, D., Rohrlach, A.B., Kaidonis, J., Townsend, G., and Austin, J.J., Differential nuclear and mitochondrial DNA preservation in post-mortem teeth with implications for forensic and ancient DNA studies, *PLoS One*, 2015, vol. 10, no. 5. e0126935.

Hoopen, P., Finn, R.D., Bongo, L.A., Corre, E., Fosso, F., Meyer, F., Mitchell, A., Pelletier, E., Pesole, G., Santamaria, M., and Willassen, N.P., The metagenomic data lifecycle: standards and best practices, *Gigascience*, 2017, vol. 6, no. 8, pp. 1–11.

Ling, N., Chen, D., Guo, H., Wei, J., Bai, Y., Shen, Q., and Hu, S., Differential responses of soil bacterial communities to long-term N and P inputs in a semi-arid steppe, *Geoderma*, 2017, vol. 292, pp. 25–33.

Lombard, N., Prestat, E., van Elsas, J.D., and Simonet, P., Soil-specific limitations for access and analysis of soil microbial communities by metagenomics, *FEMS Microbiol. Ecol.*, 2011, vol. 78, no. 1, pp. 31–49.

Lucas, A., Bodger, O., Brosi, B.J., Ford, C.R., Forman, D.W., Greig, C., Hegarty, M., Neyland, P.J., and de Vere, N., Generalisation and specialisation in hoverfly (Syrphidae) grassland pollen transport networks revealed by DNA metabarcoding, *J. Anim. Ecol.*, 2018, vol. 87, no. 4, pp. 1008–1021.

Ma, J., Prince, A., and Aagaard, K.M., Use of whole genome shotgun metagenomics: a practical guide for the microbiome-minded physician scientist, *Semin. Reprod. Med.*, 2014, vol. 32, pp. 5–13.

Malan-Muller, S., Valles-Colomer, M., Raes, J., Lowry, C.A., Seedat, S., and Hemmings, S.M.J., The gut microbiome and mental health: implications for anxiety- and trauma-related disorders, *OMICS*, 2018, vol. 22, no. 2, pp. 90–107.

Mendes, L.W., Tsai, S.M., Navarrete, A.A., De Hollander, M., van Veen, J.A., and Kuramae, E.E., Soil-borne microbi-

ome: linking diversity to function, *Microb. Ecol.*, 2015, vol. 70, pp. 255–265.

Navarrete, A.A., Soares, T., Rossetto, R., van Veen, J.A., Tsai, S.M., and Kuramae, E.E., Verrucomicrobial community structure and abundance as indicators for changes in chemical factors linked to soil fertility, *Antonie van Leeuwenhoek*, 2015, vol. 108, no. 3, pp. 741–752.

Ngara, T.R. and Zhang, H., Recent advances in function-based metagenomic screening, *Genom. Prot. Bioinform.*, 2018, vol. 16, no. 6, pp. 405–415.

Noonan, J.P., Coop, G., Kudaravalli, S., Smith, D., Krause, J., et al., Sequencing and analysis of Neanderthal genomic DNA, *Science*, 2006, vol. 314, no. 5802, pp. 1113–1118.

Pester, M., Schleper, C., and Wagner, M., The Thaumarchaeota: an emerging view of their phylogeny and ecophysiology, *Curr. Opin. Microbiol.*, 2011, vol. 14, no. 3, pp. 300–306.

Pompanon, F., Deagle, B.E., Symondson, W.O., Brown, D.S., Jarman, S.N., and Taberlet, P., Who is eating what: diet assessment using next generation sequencing, *Mol. Ecol.*, 2012, vol. 21, no. 8, pp. 1931–1950.

Poussin, C., Sierro, N., Boué, S., Battey, J., Scotti, E., Belcastro, V., Peitsch, M.C., Ivanov, N.V., and Hoeng, J., Interrogating the microbiome: experimental and computational considerations in support of study reproducibility, *Drug Discov. Today*, 2018, vol. 23, no. 9, pp. 1644–1657.

Ranjan, R., Rani, A., Metwally, A., McGee, H.S., and Perkins, D.L., Analysis of the microbiome: advantages of whole genome shotgun versus 16S amplicon sequencing, *Biochem. Biophys. Res. Commun.*, 2016, vol. 469, no. 4, pp. 967–977.

Riesenfeld, C.S., Schloss, P.D., and Handelsman, J., Metagenomics: genomic analysis of microbial communities, *Ann. Rev. Genet.*, 2004, vol. 38, pp. 525–552.

da Rocha, U.N., Andreote, F.D., de Azevedo, J.L., van Elsas, J.D., and van Overbeek, L.S., Cultivation of hitherto-uncultured bacteria belonging to the Verrucomicrobia subdivision 1 from the potato (*Solanum tuberosum* L.) rhizosphere, *J. Soils Sediments*, 2010, vol. 10, pp. 326–339.

Rothschild, D., Weissbrod, O., Barkan, E., Kurilshikov, A., Korem, T., Zeevi, D., Costea, P.I., Godneva, A., Kalka, I.N., Bar, N., Shilo, S., Lador, D., Vila, A.V., Zmora, N., Pevsner-Fischer, M., Israeli, D., Kosower, N., Malka, G., Wolf, B.C., Avnit-Sagi, T., Lotan-Pompan, M., Weinberger, A., Halpern, Z., Carmi, S., Fu, J., Wijmenga, C., Zhernakova, A., Elinav, E., and Segal, E., Environment dominates over host genetics in shaping human gut microbiota, *Nature*, 2018, vol. 555, no. 7695, pp. 210–215.

Semenov, M., Blagodatskaya, E., Stepanov, A., and Kuzyakov, Y., DNA-based determination of soil microbial biomass in alkaline and carbonaceous soils of semi-arid climate, *J. Arid Environ.*, 2018, vol. 150, pp. 54–61.

Senechkin, I.V., Speksnijder, A.G.C.L., Semenov, A.M., van Bruggen, A.H.C., and van Overbeek, L.S., Isolation and partial characterization of bacterial strains on low organic carbon medium from soils fertilized with different organic amendments, *Microb. Ecol.*, 2010, vol. 60, pp. 829–839.

Sigsgaard, E.E., Nielsen, I.B., Bach, S.S., Lorenzen, E.D., Robinson, D.P., Knudsen, S.W., Pedersen, M.W., Jaidah, M.A., Orlando, L., Willerslev, E., Møller, P.R., and Thomsen, P.F., Population characteristics of a large whale shark aggregation inferred from seawater environmental DNA, *Nat. Ecol. Evol.*, 2016, vol. 1, no. 1, p. 4.

Souza, R.C., Hungria, M., Cantao, M.E., Vasconcelos, A.T.R., Nogueira, M.A., and Vicente, V.A., Metagenomic analysis reveals microbial functional redundancies and specificities in a soil under different tillage and crop-management regimes, *Appl. Soil. Ecol.*, 2015, vol. 86, pp. 106–112.

Spribille, T., Tuovinen, V., Resl, P., Vanderpool., D., Wolinski, H., Aime, M.C., Schneider, K., Stabentheiner, E., Toome-Heller, M., Thor, G., Mayrhofer, H., Johannesson, H., and McCutcheon, J.P., Basidiomycete yeasts in the cortex of ascomycete macrolichens, *Science*, 2016, vol. 353, pp. 488–492.

Valentine, D.L., Adaptations to energy stress dictate the ecology and evolution of the archaea, *Nat. Rev. Microbiol.*, 2007, vol. 5, no. 4, pp. 316–323.

Vital, M., Karch, A., and Pieper, D.H., Colonic butyrate-producing communities in humans: an overview using omics data, *Systems*, 2017, vol. 2, no. 6.

Willerslev, E., Hansen, A.J., Binladen, J., Brand, T.B., Gilbert, M.T., et al., Diverse plant and animal genetic records from Holocene and Pleistocene sediments, *Science*, 2003, vol. 300, no. 5620, pp. 791–795.

Willerslev, E., Hansen, A.J., Rønn, R., Brand, T.B., Barnes, I., et al., Long-term persistence of bacterial DNA, *Curr. Biol.*, 2004, vol. 14, no. 1, pp. 9–10.

Wilson, J.J., Brandon-Mong, G.J., Gan, H.M., and Sing, K.W., High-throughput terrestrial biodiversity assessments: mitochondrial metabarcoding, metagenomics or metatranscriptomics?, *Mitochondrial DNA A DNA Mapp. Seq. Anal.*, 2019, vol. 30, no. 1, pp. 60–67.

Yan, J. and Charles, J.F., Gut microbiome and bone: to build, destroy, or both?, *Curr. Osteoporos. Rep.*, 2017, vol. 15, no. 4, pp. 376–384.

Yoon, T.H., Kang, H.E., Lee, S.R., Lee, J.B., Baeck, G.W., Park, H., and Kim, H.W., Metabarcoding analysis of the stomach contents of the Antarctic toothfish (*Dissostichus mawsoni*) collected in the Antarctic Ocean, *Peer J.*, 2017, vol. 5. e3977.

Zaheer, R., Noyes, N., Polo, R.O., Cook, S.R., Marinier, E., Van Domselaar, G., Belk, K.E., Morley, P.S., and McAllister, T.A., Impact of sequencing depth on the characterization of the microbiome and resistome, *Sci. Rep.*, 2018, vol. 8, no. 1, p. 5890.

Zaremba-Niedzwiedzka, K., Caceres, E.F., Saw, J.H., Bäckström, D., Juzokaite, L., Vancaester, E., Seitz, K.W., Anantharaman, K., Starnawski, P., Kjeldsen, K.U., Stott, M.B., Nunoura, T., Banfield, J.F., Schramm, A., Baker, B.J., Spang, A., and Ettema, T.J., Asgard archaea illuminate the origin of eukaryotic cellular complexity, *Nature*, 2017, vol. 541, no. 7637, pp. 353–358.

Translated by T. Kuznetsova