Applying genome-resolved metagenomics to de-convolute the halophilic microbiome

Gherman Uritskiy[1] and Jocelyne DiRuggiero[1,†]

[1]Department of Biology, Johns Hopkins University, Baltimore MD 21218.

[†]Correspondence to JDR (jdiruggiero@jhu.edu)

Abstract

In the past decade, the study of microbial life through shotgun metagenomic sequencing has rapidly expanded our understanding of environmental, synthetic, and clinical microbial communities. Here, we review how shotgun metagenomics affected the field of halophilic microbial ecology, from functional potential reconstruction, virus-host interactions, pathway selection, strain dispersal, and novel genome discoveries. However, there still remain pitfalls and limitations from conventional metagenomic analysis being applied to halophilic microbial communities. Deconvolution of halophilic metagenomes has been difficult due to the high G+C content of these microbiomes and their high intraspecific diversity, which made both metagenomic assembly and binning a challenge. Halophiles are also underrepresented in public genome databases, which in turn slows progress. With this in mind, this review proposes experimental and analytical strategies to overcome the challenges specific to halophilic microbiome from experimental design, to data acquisition, to computational analysis of metagenomics sequences. Finally, we speculate on the potential applications of other next-generation sequencing technologies to halophilic communities. RNA sequencing, long read technologies, and chromosome conformation assays, no initially intended for microbiomes, are becoming available to study microbial communities. Together with recent analytical advancements, these new methods and technologies have the potential to rapidly advance the field of halophiles research.

Keywords

extremophiles, halophilic microorganisms, hypersaline habitats, metagenomics, microbiome, shotgun sequencing, genome assembly and binning, functional annotation

Introduction

Microbial life is one of the most diverse and energetically dominant forces in Earth's ecosphere (1), making microbiome research a critical component of modern ecology. The unparalleled taxonomic and functional diversity of microbiomes allowed them to populate all locations on the planet (2, 3), including environments unfit for habitation by other life forms. In hyper-saline environments, unique environmental pressures forced microbiota to evolve specific survival adaptations, resulting in highly resilient communities that push the boundaries of life's limit (Figure 1). Halophiles have been found to play important roles in soil bioenergetics processes (4), food storage and preservation (5, 6), and have also been detected in the human gut microbiota (7). Additionally, studying halophilic life-forms revealed many fundamental aspects of life's survival limits and strategies, including its potential to endure the harsh environments we are most likely to find on other planets (8, 9). Prior to the introduction of high-throughput sequencing our understanding of halophile genomics was limited to studying cultured organisms (10, 11). While next-generation sequencing technologies have become commonplace in microbiology, the halophile field lacks a critical analysis of prospects and potential applications of these technologies to halophilic microbiomes. In this review, we discuss key aspects of halophile community composition and function that metagenomics has revealed and provide examples of studies in various hyper-saline environments for perspective on analytical progress. We then examine the advantages and limitations of applying shotgun metagenomic sequencing to uncover the structure and functioning of halophilic microbiomes. We outline the factors and characteristics that make the de-convolution of halophilic metagenomes a major challenge, and propose analytical adjustments to be made when investigating these complex communities. Both experimental design and computation analysis approaches appropriate in halophilic metagenomics are summarized. Finally, we discuss novel sequencing technologies that show promise to further propel the halophile metagenomic field.

Shotgun sequencing in metagenomics

Rapid developments in high-throughput DNA sequencing technologies since 2008 have propelled our understanding of not only single-organism genetics, but also microbiome community structure and function. Marker gene (particularly 16S rRNA gene) amplicon sequencing revealed the taxonomic composition of a given community through sequencing a small target of the community's DNA. In contrast, whole-metagenomic sequencing (WMGS) theoretically allows for reconstruction of the entire microbial community DNA content. This has led to a number of important findings in microbiome research (12, 13), as biologists were able to thoroughly investigate microbial communities at the genetic level without the need for culturing (14). However, while sequencing technologies are rapidly developing, producing complete genomes of all the microorganisms found in a community is currently unattainable due to low sequencing coverage of the less abundant organisms. Additionally, sequence repeats and regions of homology between organisms limits genome recovery from short-read data, resulting in incomplete assemblies. Instead, long contiguous pieces (contigs) of genomes are produced, ranging in length from 1Kbp to 1Mbp (15, 16). These contigs then need to be grouped based on the genome they belong to, a process known as binning. It is only recently that binning has become reliable enough to produce reasonably high-quality metagenome-assembled genomes (MAGs). The ability to produce high quality MAGs has in turn led to the discovery of thousands of novel organisms and thus enabled many breakthroughs in characterizing the taxonomic and functional components of microbiomes (17-19). Shotgun metagenomics offers tremendous advantages in recovering taxonomic and functional potential components of microbial communities, however sequencing costs deter some researchers from deploying this approach in their studies. The high average read coverage required for the assembly of a genome from shotgun reads (20) presents a major challenge for the assembly of lowly-abundant organisms in a metagenomic context. These highly diverse but under-represented taxa often constitute significant proportions of microbial communities and play important roles in biome functioning (21). Despite these challenges, whole metagenomic sequencing (WMGS) carries tremendous benefits, empowering researchers to study previously unknown aspects of microbiomes. In particular, WMGS allowed for the reconstruction of a given community's gene content, which enabled ecologists to predict the functional potential of entire communities. This new angle of microbiome analysis enabled predic-

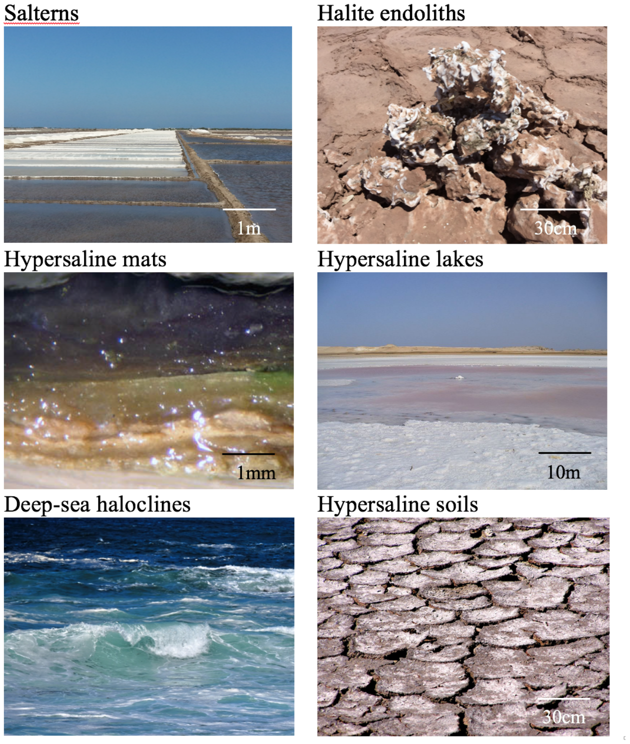


Figure 1: Photographs of commonly studied hyper-saline environments.

tion of metabolic processes potentially present in communities, and the study of community natural selection at the functional level (22, 23). The possibility of studying the functional potential of any organism in a community means that our understanding of microbial genetics, dynamics, evolution and function was no longer limited to cultured organisms. In many fields such as the human microbiome research, this has hailed a new era for research (24, 25).

Halophilic microbiome research powered by shotgun metagenomics

Numerous breakthroughs in halophilic microbiome research have been enabled by WMGS (11) (Table 1). This sequencing approach interrogates the taxonomic structure of microbiomes in high-salt environments with significantly less taxonomy-based biases than conventional ribosomal amplicon sequencing. Indeed, in conventional 16S rDNA amplicon sequencing, primer choices can have a massive impact on the taxonomic distribution of the data (26). While WMGS still has biases associated with G+C content, taxonomic annotation of shotgun reads usually results in more accurate and robust taxonomic profiles than amplicon sequencing (27). This is particularly important in high-salt environments where both Archaea and Bacteria are found in high abundance, because it is difficult to reliably amplify both domains with the same set of primers. WMGS also provides DNA sequences that are not targeted by 16S rDNA amplification, including Eukaryotic genomes, DNA viruses, and extra-chromosomal DNA such as plasmids. The reconstruction of viral genomes from hyper-saline environments (28) and halite endolithic communities (29) using WMGS has resulted in the characterization of a major aspect of halophilic microbiomes that was previously unexplored. Viruses infect and kill microorganisms, effectively playing the role of predators in many microbiomes and contributing to nutrient turnover (30, 31). Their lytic activity releases the contents of cells into the environment, making their nutrients available to other members of the community. Perfect alignments between CRISPR spacers of microorganisms and virus genomes have been used in solar salterns to infer previous infections, and thus establishing putative virus-host interactions (32). In endolithic halophiles, virus sequences encoded in CRISPR arrays have been used as a high-sensitivity strain signature, and allowed the tracking of strain dispersal (33). As previously mentioned, one of the biggest strengths of WMGS is the ability to reconstruct the functional potential of a microbial community. With WMGS, hypersaline water (8, 34), soil (4), and endolithic (35) microbiomes have been characterized in terms of their functioning, particularly their ability to use a range of energy sources. Building on previous culture-dependent methods, systematic functional analysis of halophilic metagenomes led to major improvements in our understanding of halophile osmotic adaptation and evolution (36). Functional annotation of longitudinal studies of halophiles from saltern, hypersaline lake, and salt mineral environments have also led to the characterization of horizontal gene transfers, evolutionary dynamics, and functional adaptations across time and space (34, 35, 37, 38). Functional potential profiling also uncovered the selective pressures and community functional dynamics, which are not possible to investigate through taxonomy alone due to high functional redundancy. With WMGS analysis rapidly improving and halophile databases rapidly growing (39), more breakthroughs will follow. Another major aspect of metagenomics facilitated by WMGS is the reconstruction of novel individual genomes of halophiles. This is particularly important because extreme halophiles, and extremophiles in general, have been difficult to isolate due to specific growth conditions requirements, symbiotic relationships, and cross-species functional pathways (40). The binning of metagenomics assemblies has enabled researchers to recover hundreds of halophilic MAGs in the past decade (39) with many belonging to previously unknown orders or even phyla. The recovery of near-complete genomes of Nanohaloarchaea and Halobacteria from metagenomics samples has improved our overall understanding of halophilic microbiomes, while empowering future research by expanding existing taxonomic and functional annotation databases (41, 42). In a positive-feedback loop, the rapidly increasing number of annotated reference halophile genomes is allowing for more accurate taxonomic and functional annotation in halophilic microbiomes (39).

Limitations of shotgun metagenomics in halophile research

In contrast to human and synthetic microbiomes, the reconstruction of environmental metagenomes has been complicated by their sheer diversity and microdiversity. This is especially true in high-salt environments, which often host microbial communities with low taxonomic diversity but very high intraspecific diversity and characteristically high G+C content (64, 65). The presence of a large number of highly similar strains presents major challenges for de-convoluting their DNA content through metagenomic assembly and binning, and the high G+C content reduces the fraction of unique sequences in the samples (48, 66). For example, halophilic endolith communities are typically dominated by Halobacteria and Salinibater, however their high strain diversity, and G+C content over 60%, leads to relatively poor assembly and MAG quality (33). In contrast, other community members that are less abundant and have low G+C content, such as Cyanobacteria, Actinobacteria, and Gammaproteobacteria, have yielded high quality MAGs (35). Due to the previously mentioned difficulties in culturing a diversity of halophiles, there is a relatively small number of genomes available. In 2018, there were just 1088 complete halophile genomes available in all databases – a tiny number in the era of high throughput sequencing, which thus far yielded over 200,000 Prokaryotic complete genomes (67). This leaves MAG extraction from environmental sequencing data the primary method for obtaining the genomes of halophilic organisms, which has also been difficult due to their metagenomic properties. In a negative feedback loop, this in turn further stalled progress of halophilic microbiome research, as the lack of available reference genomes made taxonomic and functional annotation difficult. As WMGS becomes commonplace in microbiome research, it is crucial that the halophile field takes full advantage of the new technology and the use of newly available bioinformatic tools to further its understanding of microbial community assembly and function. Since 2014-2015, improvements in analytical methods and assembly software such as metaSPAdes (68), binning software such as metaBAT (69), and processing pipelines such as metaWRAP (17) allowed for effective deconvolution of WMGS data from even the most complex microbiomes. These new progress will greatly benefit the halophile research field if applied effectively.

Experimental design considerations for sequencing halophilic metagenomes

There are two general approaches to metagenomic sequencing and analysis – (1) co-assembly of multiple shallowly sampled samples or (2) individual processing of a few deeply sequenced samples. Both approaches have their benefits and limitations, depending on the microbiome that is sequenced and the biological question to answer. In the first approach, samples are sequenced with relatively low read coverage and reads from all samples are combined during metagenomic assembly (Figure 2A). In research projects that demand a large number of samples, such as longitudinal studies, this results in low sequencing costs per sample, while also producing high quality MAGs from the co-assembly by leveraging differential abundances of the contigs across samples (17, 69). The taxonomic and functional composition of individual samples can be interrogated by linking the taxonomic and functional annotations of each contig with its abundance in each sample, allowing for easy comparison between large numbers of samples (35, 37). Finally, co-assembling data from multiple samples enhance the recovery of genomes from low-abundance organisms, which is not possible from individual samples due to low coverage (42). However, the use of co-assembly in metagenomics comes with

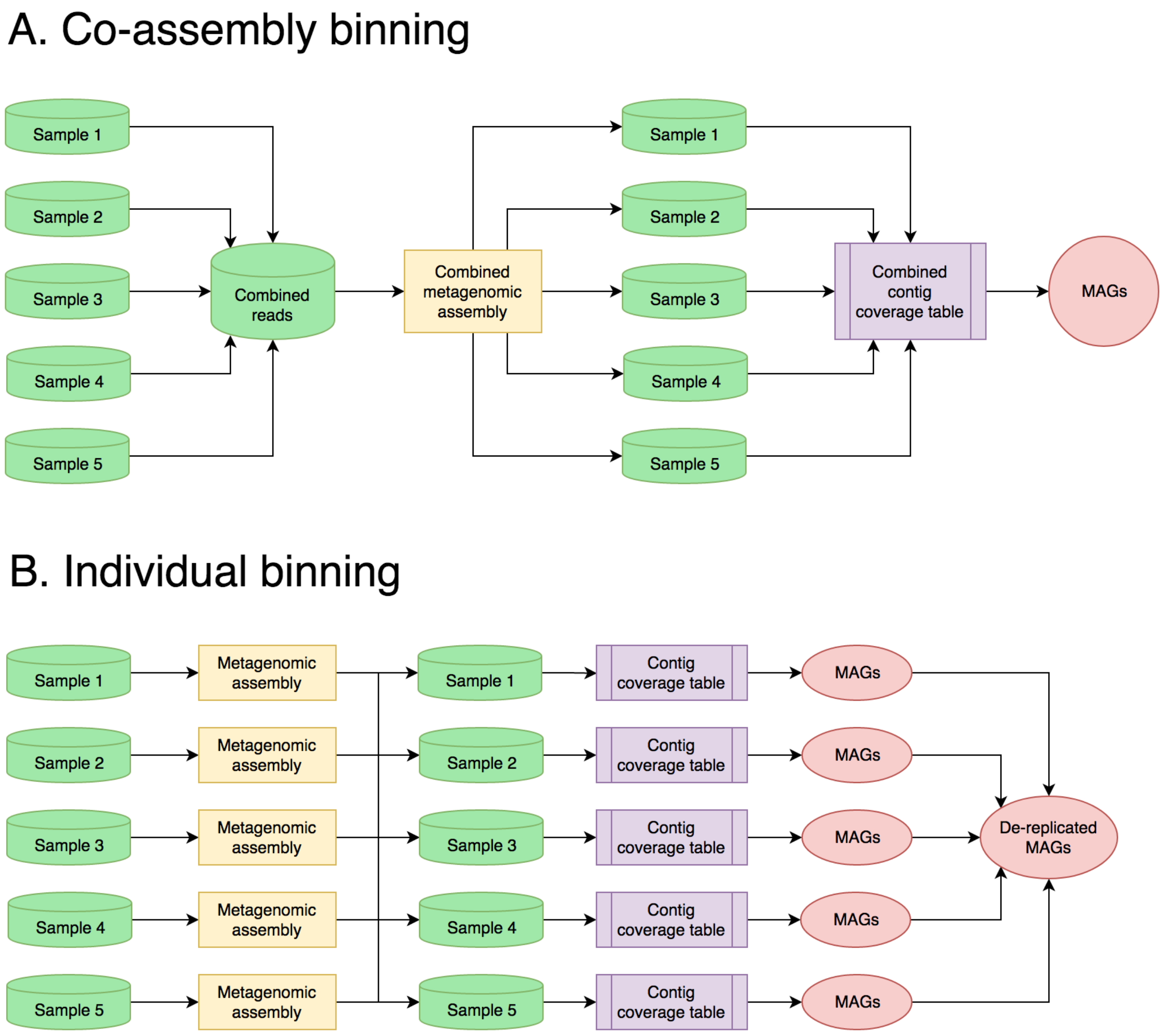
Environment	Longitudinal dynamics	MAG discovery	Functional potential	Virus analysis
Hypersaline lakes	Andrade (43), Tschitschko (38)	Narasingarao (42)	Vavourakis (44)	Emerson (45), Tschitschko (38), Ramos-Barbero (46)
Salterns	Di Meglio (47)	Ramos-Barbero (48)	Jayanath (49), Plominsky (50)	Moller (32), Di Meglio (47)
Hypersaline microbial mats	Berlanga (51)	Mobberley (52)	Mobberley (52), Ruvindy (53), Wong (54)	White (55)
Deep-sea haloclines	N/A	Speth (56)	Guan (57), Pachiadaki (58)	Antunes (59)
Halite endoliths	Uritskiy (35), Finstad (33)	Finstad (33), Uritskiy (35),	Crits-Christoph (60), Uritskiy (35)	Crits-Christoph (60)
Hypersaline soils	Narayan (61)	Vera-Gargallo (4)	Vera-Gargallo (4), Pandit (62)	Emerson (63)

Table 1: Studies that uncovered novel aspects of halophilic microbial communities through WMGS in hypersaline environments (list is not exhaustive).

significant drawbacks (48) including the high computational costs of co-assembling large data and the high level of microdiversity introduced by each new biological replicate. This later point might be counter-intuitive but it leads to poor assemblies of very abundant taxa because accumulated mismatches from strain heterogeneity complicate the De Bruijn graph during assembly. This is particularly problematic with halophilic microbiomes that are often dominated by highly diverse groups of Euryarchaeota and Bacteroidetes (41). The high population microdiversity of these taxa is exacerbated when using multiple biological replicates and results in poor, fragmented or chimeric assemblies (48). This in turn translates in poor-quality MAGs. However, when a broad capture of community diversity across many samples is the intent of the study, these limitations should then be considered in data interpretation. An alternative approach to co-assembly is to sequence a small number of samples with deep coverage, and process them individually (Figure 2B). Because of the reduced microdiversity, individual assemblies produce larger contigs given comparable sequencing depth (70). After binning each sample separately, MAGs can be combined into a single set through dereplication, removing duplicate MAGs that share high nucleotide identity (71). As with the co-assembly approach, differential contig coverage across samples may be used to improve the binning results (33). While this method is superior in highly heterogeneous communities such as halophilic microbiomes, it comes at a major increase in sequencing cost per sample. For most metagenomes, a meaningful assembly (N50>5Kbp) requires 25-50Gbp of sequencing data per sample, which limits the number of samples that can be multiplexed on a sequencing run. In turn, the limited replication reduces the effectiveness of binning, which leverages differential coverage of contigs across many samples to increase binning accuracy (72). For many studies that require a large number of replicates, such as longitudinal studies, the cost of this approach may become prohibitively expensive. An additional consideration in choosing a strategy for metagenomic sequencing and analysis is that of inter-sample community diversity. Communities in aquatic biomes, such as hyper-saline lakes or brine ponds, are often more homogenous, harboring the same microorganisms with different relative abundances at different sampling locations. Under those conditions, a co-assembly strategy for metagenomics, as discussed above, is often preferred (37, 42, 73). In contrast, in terrestrial microbiomes with limited dispersal, such as halite nodules in Salars of the Atacama Desert that contain unique taxonomic compositions, an individual assembly approach is more advantageous (29, 33). Hybrid approaches are also possible in many cases, as binning of the individual and grouped assemblies may be combined and de-replicated to obtain the most robust MAGs of both rare and abundant species (74). Regardless of the experimental design, it is critical to process samples, generate libraries, and sequence samples together to avoid batch effects (75). If more than one flowcell is required to achieve the desired read depth, it is usually better to sequence the pooled libraries on several flowcells than to sequence each sample on its own flowcell (75). For library preparation, it is recommended to use protocols that produce minimal G+C biases in coverage, particularly in halophilic communities that have high G+C-content variation in their metagenomes (76, 77). The take home message is that when conducting a halophile metagenomic study it is especially important to design the sampling and sequencing scheme with the statistical questions in mind. Because of the high strain-level diversity typically found in halophilic microbiome, the experimental design should avoid adding unnecessary replicates into the study, as each added biological replicate will introduce more microdiversity into the data, further complicating the assembly and binning stages of the analysis (48). In practical terms, unless the intent of the study is to capture maximum diversity, the experimental design should include the minimum number of biological replicates that will allow the intended statistical analysis downstream.

Best practices for halophilic metagenome analysis

When processing halophilic metagenome sequencing data, it is important to adjust existing pipelines to accommodate for high intraspecific diversity, G+C-content diversity, and underrepresentation in most sequence databases. While this section does not provide a step-by-step instruction of bioinformatics analysis, it outlines core considerations and adjustments scientists should be making while interrogating halophilic metagenomes. While automated metagenomic analysis pipelines such as metaWRAP (17) or SqueezeM (78) may be used to streamline and simplify analysis, pipelines that are specifically trained on/or intended for animal microbiomes such as the gut microbiota should be avoided. Indeed, these latter pipelines rely strongly on pre-existing taxonomic and functional databases of closely related organisms, as the majority of organisms found in host-associated microbiomes have been sequenced and characterized. The pre-processing of WMGS data, which typically includes read trimming, duplicate read removal, and metagenomic assembly, is standard for most types of metagenomes. We encourage testing a variety of software and comparing the results with evaluation programs such as FastQC (79) (for read quality) and MetaQUAST (80) (for assembly quality), as some methods may be more suited for specific types of microbial community types (81). For metagenomic assembly, metaSPAdes (68) is currently considered to be the best overall, while MegaHIT (82) is a better solution when resources are a limiting factor as it is significantly faster and requires less memory (83). In contrast to assembly, the annotation of halophilic metagenomes for taxonomies and functions can be somewhat compromised because halophiles have extremely limited representation in standard-distribution taxonomic databases (84, 85), which introduces significant biases in sequence annotation. As of 2018, there were only 942 published complete halophilic genomes available in NCBI (39) – the main database used as a reference in most taxonomic and functional annotation software. Regarding methods for taxonomic profiling, general alignment-based methods such as MegaBLAST(86) are usually too specific for annotating halophilic DNA sequences, especially non-assembled reads. To produce more balanced taxonomic annotation given the limited databases, it is recommended to assign taxonomy to assembled contigs based on genes that they carry, and then infer taxonomy of reads based on their alignment to the contigs. If the intent is to obtain the most accurate taxonomic distribution profile of the community, extracting and annotating marker genes such as 16S rRNA genes with EMIRGE is usually the best alternative (87), as rRNA gene databases are more established and encompass greater taxonomic diversity (88). Functional annotation – the functional categorization of genes – in halophile metagenomes is also severely limited by existing databases, especially compared to human microbiomes. Because many halophilic genes are not annotated in NCBI databases, metagenome-inclusive custom or specific databases are preferred, as they contain a greater variety of non-cultured organisms. In particular, services such as the Integrated Microbial Genomes systems (89) include taxonomic and functional annotation models that are trained on user-submitted metagenomic data, including high-quality MAGs. The annotation sensitivity resulting from using the newest metagenomic data is extremely



A. Co-assembly binning

B. Individual binning

Figure 2: Flowcharts showing two common experimental designs and analysis workflows: (A) co-assembly and (B) individual sample processing and binning.

valuable for both functional and taxonomic annotation in relatively understudied systems such as halophilic microbiomes. Annotation pipelines geared towards human microbiomes such as HUMANN2 (90) should be avoided, as they rely on the presence of closely-related organisms in databases. Finally, the success of metagenomic binning of assemblies will depend greatly on the software choice, as binning programs perform differently on various data types (17). Additionally, many popular binning software such as metaBAT1 are trained on gut microbiome data (69), potentially limiting their efficacy in complex halophilic communities. Furthermore, benchmarking of such algorithms is often done on real or synthetic gut microbial communities (81). Because of this, it is recommended to bin the metagenomic assembly with a variety of the most recent binning software such as metaBAT2 (69) and CONCOCT (91) and to use a binning consolidation tool such as metaWRAP or DAS_Tool to produce the best final bin set (17, 92). When estimating the read coverage of the contigs in a given sample to feed into the binning algorithms, it is important to remember that they represent collapsed averages of a number of strains. Given the high intraspecific diversity of halophilic microbiomes (48), more accurate abundance estimation could potentially be obtained with slightly relaxed read alignment parameters to allow more approximate matches. Considering the overwhelming number of metagenomic bioinformatics tools coming out each year, it is difficult to keep up to date with the best analytical methods. In general, we advise testing and benchmarking multiple software for each analytical step to determine the best option, as many conventionally used software behave unpredictably with halophilic sequence data. For annotation, emphasis should be placed on high sensitivity rather than high precision, given the database limitations.

The future of halophilic metagenomics

Beyond shotgun sequencing of a microbiome DNA content, there exist a number of other sequencing technologies that have become available and may further our understanding of halophilic ecosystems. Studies applying these technologies to more developed microbial fields such as human gut microbiomes show their great promise and their potential applications to halophilic microbial communities in the near future. Conventional Illumina sequencing is limited to short DNA fragments (50bp - 250bp), as errors accumulate rapidly at higher read lengths. However, read length, together with sequencing coverage, is undoubtedly a major limiting factor for metagenomics sequence assembly. Longer reads result in more accurate assembly and reduced chimeras, while improving the contiguity of the assembly by allowing assembly of repetitive DNA elements (93). Recent sequencing technologies – minION from Oxford Nanopore and SMRT from PacBio sequencing – produce longer DNA fragments compared to Illumina. PacBio is able to consistently produce long reads (N50 up to 10Kbp) with a relatively high degree of accuracy (94, 95), while Nanopore sequencing produces even longer reads (N50 up to 100Kbp) but with some sacrifices to accuracy (96, 97). Read lengths from these technologies enable for not only sequencing complete ribosomal genes for improved taxonomic annotation, but also for significantly improving the accuracy of metagenomics assembly and binning (95, 98). In highly diverse halophilic communities, long reads can help assemble ambiguous regions resulting from taxonomic heterogeneity, drastically improving the quality of the metagenome assembly (98). Pseudo-single cell technology from 10X Genomics, which tags each read with a barcode unique to the cell it came from, also show great promise in halophilic microbiome de-convolution, as they are able to produce strain-specific synthetic long reads originating from single cells (99). With reported maximum read lengths of over 1Mbp from Nanopore, long read technology is rapidly approaching the point where sequencing complete genomes in a single read will be theoretically possible (100). When this becomes reality, it will propel the field of metagenomics into a new post-assembly era. However, the recovery of lowly abundant taxa will remain a concern given the relatively low throughput of these methods. Chromosome conformation capture assays (Hi-C) is another sequencing technology that shows great promise for the field of halophilic metagenomics. Hi-C assays crosslink DNA based on spatial proximity; the chimeric segments resulting from the crosslink events are then sequenced, revealing sections of DNA that were proximal to each other. Conventionally used to indirectly measure the proximity between sections of a genome, HiC was successfully applied to microbiomes to improve binning predictions in 2017 (101). Considering the difficulty of binning halophilic metagenomes due to their heterogeneity, HiC could significantly improve halophile MAG extraction. HiC-based binning also enables recovery of extrachromosomic elements such as viral and plasmid DNA, which so far has been difficult to accomplish (102). HiC can also be used to produce DNA proximity maps in individual MAGs for the study of chromatin conformation in prokaryotes at the metagenomic and single-cell scale (102). Finally, genome-resolved metatranscriptomics – the analysis of a microbial community's RNA content – has been widely used in a variety of microbiomes to interrogate microbial transcriptional activities (24, 103). Metatranscriptomics have been used in halophile research to characterize carbon cycling in saline soils (104) and extensively to characterize activity in soil microbiomes (105, 106). However, it remains a largely under-deployed tool, partly due to difficulty in depleting ribosomal sequences in archaeal RNA. Another major deterrent has been the difficulty in standardizing transcript expression to the abundance of each individual organism in a sample. In other words, if a transcript is more abundant in a given sample, it can be difficult to determine if the organism carrying it is more abundant in that sample, or if it is truly highly expressed. However, with rapid improvements in genome-resolved metagenomic analysis of halophile communities, it is possible that the metatranscriptomic problem can be simplified down to more conventional transcriptome analysis by investigating the transcriptomes of individual MAGs.

Conclusions

Successful application of whole metagenomics to halophilic communities has already led to numerous breakthroughs in our understanding of their functional composition, virus-host interactions, strain diversity and dispersal, and uncovered thousands of novel halophile genomes. However, the genomic qualities and composition characteristics of halophilic communities made them difficult to de-convolute in a metagenomic context, limiting the information that can be extracted from halophilic shotgun metagenomes. Combined with relative low number of cultures of halophiles, this led to their underrepresentation in existing taxonomical and functional databases, which further complicated analysis. While in-silico de-convolution of halophilic metagenomes is a challenge, it can be accomplished with analysis workflows that account for specific characteristics of halophile communities. With proper utilization, the rapidly advancing sequencing technology has the potential to reconstruct the complete nucleic acid content of halophilic communities, allowing the halophile field to focus on microbial functional activity and interactions.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by grants NNX15AP18G from NASA and DEB1556574 from the NSF.

Author's contributions

GU and JD conceptualized the review; GU wrote the manuscript; JD edited the manuscript.

Acknowledgements

We thank Sarah Preheim for useful suggestions in the writing of this manuscript.

Bibliography

1. E. B. Graham et al., Microbes as Engines of Ecosystem Function: When Does Community Structure Enhance Predictions of Ecosystem Processes? Front Microbiol 7, 214 (2016).
2. J. Kallmeyer, R. Pockalny, R. R. Adhikari, D. C. Smith, S. D'Hondt, Global distribution of microbial abundance and biomass in subseafloor sediment. Proc Natl Acad Sci U S A 109, 16213-16216 (2012).
3. W. B. Whitman, D. C. Coleman, W. J. Wiebe, Prokaryotes: the unseen majority. Proc Natl Acad Sci U S A 95, 6578-6583 (1998).
4. B. Vera-Gargallo, A. Ventosa, Metagenomic Insights into the Phylogenetic and Metabolic Diversity of the Prokaryotic Community Dwelling in Hypersaline Soils from the Odiel Saltmarshes (SW Spain). Genes (Basel) 9, (2018).
5. A. Gibtan et al., Diversity of Extremely Halophilic Archaeal and Bacterial Communities from Commercial Salts. Front Microbiol 8, 799 (2017).
6. O. Henriet, J. Fourmentin, B. Delince, J. Mahillon, Exploring the diversity of extremely halophilic archaea in food-grade salts. Int J Food Microbiol 191, 36-44 (2014).
7. E. H. Seck, J. C. Dufour, D. Raoult, J. C. Lagier, Halophilic & halotolerant prokaryotes in humans. Future Microbiol 13, 799-812 (2018).
8. A. Oren, Halophilic archaea on Earth and in space: growth and survival under extreme conditions. Philos Trans A Math Phys Eng Sci 372, (2014).
9. Y. Ma, E. A. Galinski, W. D. Grant, A. Oren, A. Ventosa, Halophiles 2010: life in saline environments. Appl Environ Microbiol 76, 6971-6981 (2010).
10. C. Rinke et al., Insights into the phylogeny and coding potential of microbial dark matter. Nature 499, 431-437 (2013).
11. B. P. Hedlund, J. A. Dodsworth, S. K. Murugapiran, C. Rinke, T. Woyke, Impact of single-cell genomics and metagenomics on the emerging view of extremophile "microbial dark matter". Extremophiles 18, 865-875 (2014).
12. R. Ranjan, A. Rani, A. Metwally, H. S. McGee, D. L. Perkins, Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem Biophys Res Commun 469, 967-977 (2016).
13. M. Tessler et al., Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. Sci Rep 7, 6589 (2017).
14. C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, N. Segata, Corrigendum: Shotgun metagenomics, from sampling to analysis. Nat Biotechnol 35, 1211 (2017).
15. J. S. Ghurye, V. Cepeda-Espinoza, M. Pop, Metagenomic Assembly: Overview, Challenges and Applications. Yale J Biol Med 89, 353-362 (2016).
16. N. D. Olson et al., Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. Brief Bioinform, (2017).
17. G. V. Uritskiy, J. DiRuggiero, J. Taylor, MetaWRAP - a flexible pipeline for genome-resolved metagenomic data analysis. bioRxiv, (2018).
18. B. J. Tully, E. D. Graham, J. F. Heidelberg, The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci Data 5, 170203 (2018).
19. N. Sangwan, F. Xia, J. A. Gilbert, Recovering complete and draft population genomes from metagenome datasets. Microbiome 4, 8 (2016).
20. D. Sims, I. Sudbery, N. E. Ilott, A. Heger, C. P. Ponting, Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet 15, 121-132 (2014).
21. R. Zaheer et al., Impact of sequencing depth on the characterization of the microbiome and resistome. Sci Rep 8, 5890 (2018).
22. F. Sharifi, Y. Ye, From Gene Annotation to Function Prediction for Metagenomics. Methods Mol Biol 1611, 27-34 (2017).
23. J. Wang et al., Metagenomic sequencing reveals microbiota and its functional potential associated with periodontal disease. Sci Rep 3, 1843 (2013).

24. W. L. Wang et al., Application of metagenomics in the human gut microbiome. World J Gastroenterol 21, 803-814 (2015).
25. C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, N. Segata, Shotgun metagenomics, from sampling to analysis. Nat Biotechnol 35, 833-844 (2017).
26. R. Poretsky, R. L. Rodriguez, C. Luo, D. Tsementzi, K. T. Konstantinidis, Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. PLoS One 9, e93827 (2014).
27. J. R. White, N. Nagarajan, M. Pop, Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput Biol 5, e1000352 (2009).
28. S. Roux et al., Analysis of metagenomic data reveals common features of halophilic viral communities across continents. Environ Microbiol 18, 889-903 (2016).
29. A. Crits-Christoph et al., Functional interactions of archaea, bacteria and viruses in a hypersaline endolithic community. Environ Microbiol 18, 2064-2077 (2016).
30. C. Pedros-Alio et al., The microbial food web along salinity gradients. FEMS Microbiol Ecol 32, 143-155 (2000).
31. N. GuixaBoixareu, J. I. CalderonPaz, M. Heldal, G. Bratbak, C. PedrosAlio, Viral lysis and bacterivory as prokaryotic loss factors along a salinity gradient. Aquat Microb Ecol 11, 215-227 (1996).
32. A. G. Moller, C. Liang, Determining virus-host interactions and glycerol metabolism profiles in geographically diverse solar salterns with metagenomics. PeerJ 5, e2844 (2017).
33. K. M. Finstad et al., Microbial Community Structure and the Persistence of Cyanobacterial Populations in Salt Crusts of the Hyperarid Atacama Desert from Genome-Resolved Metagenomics. Front Microbiol 8, 1435 (2017).
34. J. A. Kimbrel et al., Microbial Community Structure and Functional Potential Along a Hypersaline Gradient. Front Microbiol 9, 1492 (2018).
35. G. Uritskiy et al., Response of extremophile microbiome to a rare rainfall reveals a two-step adaptation mechanism. bioRxiv, (2018).
36. E. A. Becker et al., Phylogenetically driven sequencing of extremely halophilic archaea reveals strategies for static and dynamic osmo-response. PLoS Genet 10, e1004784 (2014).
37. M. Z. DeMaere et al., High level of intergenera gene exchange shapes the evolution of haloarchaea in an isolated Antarctic lake. Proc Natl Acad Sci U S A 110, 16939-16944 (2013).
38. B. Tschitschko et al., Genomic variation and biogeography of Antarctic haloarchaea. Microbiome 6, 113 (2018).
39. A. Loukas, I. Kappas, T. J. Abatzopoulos, HaloDom: a new database of halophiles across all life domains. J Biol Res (Thessalon) 25, 2 (2018).
40. L. Solden, K. Lloyd, K. Wrighton, The bright side of microbial dark matter: lessons learned from the uncultivated majority. Curr Opin Microbiol 31, 217-226 (2016).
41. A. Ventosa, R. R. de la Haba, C. Sanchez-Porro, R. T. Papke, Microbial diversity of hypersaline environments: a metagenomic approach. Curr Opin Microbiol 25, 80-87 (2015).
42. P. Narasingarao et al., De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. ISME J 6, 81-93 (2012).
43. K. Andrade et al., Metagenomic and lipid analyses reveal a diel cycle in a hypersaline microbial ecosystem. ISME J 9, 2697-2711 (2015).
44. C. D. Vavourakis et al., A metagenomics roadmap to the uncultured genome diversity in hypersaline soda lake sediments. Microbiome 6, 168 (2018).
45. J. B. Emerson et al., Virus-host and CRISPR dynamics in Archaea-dominated hypersaline Lake Tyrrell, Victoria, Australia. Archaea 2013, 370871 (2013).
46. M. D. Ramos-Barbero et al., Prokaryotic and viral community structure in the singular chaotropic salt lake salar de uyuni. Environ Microbiol, (2019).
47. L. Di Meglio et al., Seasonal dynamics of extremely halophilic microbial communities in three Argentinian salterns. FEMS Microbiol Ecol 92, (2016).

48. M. D. Ramos-Barbero et al., Recovering microbial genomes from metagenomes in hypersaline environments: the Good, the Bad and the Ugly. Systematic and Applied Microbiology, (2018).
49. G. Jayanath et al., A novel solvent tolerant esterase of GDSGG motif subfamily from solar saltern through metagenomic approach: Recombinant expression and characterization. Int J Biol Macromol 119, 393-401 (2018).
50. A. M. Plominsky et al., Distinctive Archaeal Composition of an Artisanal Crystallizer Pond and Functional Insights Into Salt-Saturated Hypersaline Environment Adaptation. Front Microbiol 9, 1800 (2018).
51. M. Berlanga, M. Palau, R. Guerrero, Functional Stability and Community Dynamics during Spring and Autumn Seasons Over 3 Years in Camargue Microbial Mats. Front Microbiol 8, 2619 (2017).
52. J. M. Mobberley et al., Organismal and spatial partitioning of energy and macronutrient transformations within a hypersaline mat. FEMS Microbiol Ecol 93, (2017).
53. R. Ruvindy, R. A. White, 3rd, B. A. Neilan, B. P. Burns, Unravelling core microbial metabolisms in the hypersaline microbial mats of Shark Bay using high-throughput metagenomics. ISME J 10, 183-196 (2016).
54. H. L. Wong et al., Disentangling the drivers of functional complexity at the metagenomic level in Shark Bay microbial mat microbiomes. ISME J 12, 2619-2639 (2018).
55. R. A. White Iii, H. L. Wong, R. Ruvindy, B. A. Neilan, B. P. Burns, Viral Communities of Shark Bay Modern Stromatolites. Front Microbiol 9, 1223 (2018).
56. D. R. Speth et al., Draft Genome of Scalindua rubra, Obtained from the Interface Above the Discovery Deep Brine in the Red Sea, Sheds Light on Potential Salt Adaptation Strategies in Anammox Bacteria. Microb Ecol 74, 1-5 (2017).
57. Y. Guan, T. Hikmawan, A. Antunes, D. Ngugi, U. Stingl, Diversity of methanogens and sulfate-reducing bacteria in the interfaces of five deep-sea anoxic brines of the Red Sea. Res Microbiol 166, 688-699 (2015).
58. M. G. Pachiadaki, M. M. Yakimov, V. LaCono, E. Leadbetter, V. Edgcomb, Unveiling microbial activities along the halocline of Thetis, a deep-sea hypersaline anoxic basin. ISME J 8, 2478-2489 (2014).
59. A. Antunes et al., First Insights into the Viral Communities of the Deep-sea Anoxic Brines of the Red Sea. Genomics Proteomics Bioinformatics 13, 304-309 (2015).
60. A. Crits-Christoph et al., Phylogenetic and Functional Substrate Specificity for Endolithic Microbial Communities in Hyper-Arid Environments. Front Microbiol 7, 301 (2016).
61. A. Narayan et al., Response of microbial community structure to seasonal fluctuation on soils of Rann of Kachchh, Gujarat, India: Representing microbial dynamics and functional potential. Ecological Genetics and Genomics 6, 22-32 (2018).
62. A. S. Pandit et al., A snapshot of microbial communities from the Kutch: one of the largest salt deserts in the World. Extremophiles 19, 973-987 (2015).
63. J. B. Emerson et al., Dynamic viral populations in hypersaline systems as revealed by metagenomic assembly. Appl Environ Microbiol 78, 6309-6320 (2012).
64. S. Cuadros-Orellana et al., Genomic plasticity in prokaryotes: the case of the square haloarchaeon. ISME J. 1, 235-245 (2007).
65. R. T. Papke, J. E. Koenig, F. Rodriguez-Valera, W. F. Doolittle, Frequent Recombination in a Saltern Population of Halorubrum. Science 306, 1928-1929 (2004).
66. Y. C. Chen, T. Liu, C. H. Yu, T. Y. Chiang, C. C. Hwang, Effects of GC bias in next-generation-sequencing data on de novo genome assembly. PLoS One 8, e62856 (2013).
67. D. H. Haft et al., RefSeq: an update on prokaryotic genome annotation and curation. Nucleic Acids Res 46, D851-D860 (2018).
68. S. Nurk, D. Meleshko, A. Korobeynikov, P. A. Pevzner, metaSPAdes: a new versatile metagenomic assembler. Genome Res 27, 824-834 (2017).
69. D. D. Kang, J. Froula, R. Egan, Z. Wang, MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3, e1165 (2015).
70. J. M. Haro-Moreno et al., Fine metagenomic profile of the Mediterranean stratified and mixed water columns revealed by assembly and recruitment. Microbiome 6, 128 (2018).

71. M. R. Olm, C. T. Brown, B. Brooks, J. F. Banfield, dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J 11, 2864-2868 (2017).
72. J. K. Goodrich et al., Conducting a microbiome study. Cell 158, 250-262 (2014).
73. C. D. Vavourakis et al., Metagenomic Insights into the Uncultured Diversity and Physiology of Microbes in Four Hypersaline Soda Lake Brines. Front Microbiol 7, 211 (2016).
74. R. D. Stewart et al., Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. Nat Commun 9, 870 (2018).
75. S. M. Gibbons, C. Duvallet, E. J. Alm, Correcting for batch effects in case-control microbiome studies. PLoS Comput Biol 14, e1006102 (2018).
76. S. Paul, S. K. Bag, S. Das, E. T. Harvill, C. Dutta, Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. Genome Biol 9, R70 (2008).
77. M. B. Jones et al., Library preparation methodology can influence genomic and functional predictions in human microbiome research. Proc Natl Acad Sci U S A 112, 14024-14029 (2015).
78. J. Tamames, F. Puente-Sanchez, SqueezeM, a highly portable, fully automatic metagenomic analysis pipeline. bioRxiv, (2018).
79. J. Brown, M. Pirrung, L. A. McCue, FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. Bioinformatics, (2017).
80. A. Mikheenko, V. Saveliev, A. Gurevich, MetaQUAST: evaluation of metagenome assemblies. Bioinformatics 32, 1088-1090 (2016).
81. A. Sczyrba et al., Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. Nat Methods 14, 1063-1071 (2017).
82. D. Li et al., MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods 102, 3-11 (2016).
83. J. Vollmers, S. Wiegand, A. K. Kaster, Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters! PLoS One 12, e0169662 (2017).
84. D. L. Wheeler et al., Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 29, 11-16. (2001).
85. N. A. O'Leary et al., Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44, D733-745 (2016).
86. Y. Chen, W. Ye, Y. Zhang, Y. Xu, High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic Acids Res 43, 7762-7768 (2015).
87. C. S. Miller, B. J. Baker, B. C. Thomas, S. W. Singer, J. F. Banfield, EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. Genome Biol 12, R44 (2011).
88. C. Quast et al., The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41, D590-596 (2013).
89. I. A. Chen et al., IMG/M: integrated genome and metagenome comparative data analysis system. Nucleic Acids Res 45, D507-D516 (2017).
90. S. Abubucker et al., Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol 8, e1002358 (2012).
91. J. Alneberg et al., Binning metagenomic contigs by coverage and composition. Nat Methods 11, 1144-1146 (2014).
92. C. M. K. Sieber et al., Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat Microbiol, (2018).
93. K. Wommack, E., J. Bhavsar, J. Ravel, Metagenomics: read length matters. Appl Environ Microbiol. 74, 1453-1463 (2008).
94. A. Rhoads, K. F. Au, PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics 13, 278-289 (2015).

95. J. A. Frank et al., Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. Sci Rep 6, 25373 (2016).
96. B. L. Brown, M. Watson, S. S. Minot, M. C. Rivera, R. B. Franklin, MinION nanopore sequencing of environmental metagenomes: a synthetic approach. Gigascience 6, 1-10 (2017).
97. F. J. Rang, W. P. Kloosterman, J. de Ridder, From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol 19, 90 (2018).
98. C. B. Driscoll, T. G. Otten, N. M. Brown, T. W. Dreher, Towards long-read metagenomics: complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture. Stand Genomic Sci 12, 9 (2017).
99. E. Moss et al., De novo assembly of microbial genomes from human gut metagenomes using barcoded short read sequences. bioRxiv, (2017).
100. M. Jain et al., Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol 36, 338-345 (2018).
101. M. O. Press et al., Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions. bioRxiv, (2017).
102. J. N. Burton, I. Liachko, M. J. Dunham, J. Shendure, Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. G3 (Bethesda) 4, 1339-1346 (2014).
103. A. Lavelle, H. Sokol, Gut microbiota: Beyond metagenomics, metatranscriptomics illuminates microbiome functionality in IBD. Nat Rev Gastroenterol Hepatol 15, 193-194 (2018).
104. M. Ren et al., Diversity and Contributions to Nitrogen Cycling and Carbon Fixation of Soil Salinity Shaped Microbial Communities in Tarim Basin. Front Microbiol 9, 431 (2018).
105. A. Garoutte, E. Cardenas, J. Tiedje, A. Howe, Methodologies for probing the metatranscriptome of grassland soil. J Microbiol Methods 131, 122-129 (2016).
106. Y. Jiang, X. Xiong, J. Danska, J. Parkinson, Metatranscriptomic analysis of diverse microbial communities reveals core metabolic pathways and microbiome-specific functionality. Microbiome 4, 2 (2016).