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

Gherman Uritskiy, Jocelyne DiRuggiero

**Abstract**

The study of microbial life in their native communities through shotgun metagenomic sequencing offers tremendous advantages over culture-dependent techniques. In the past decade, this technique has rapidly propelled our understanding of environmental, artificial, and clinical microbial communities. In halophilic microbiome research however, this progress has been slowed by genomic characteristics that make halophiles difficult to distinguish in a metagenomic setting. Many key steps in shotgun metagenomic data analysis, including assembly and putative genomes recovery, are much more difficult in communities with a high GC content and complex strain heterogeneity, as seen in high-salt environments. This caused halophiles to be underrepresented in public sequence databases, which in turn further stalled the field. However, as metagenomic bioinformatics algorithms and pipelines rapidly improved, reliable de-convolution of even the most complex communities has become possible. Functional potential reconstruction, virus-host interactions, pathway selection, strain dispersal, and novel genome discovery are only some of the possible analytical angles enabled by shotgun metagenomics enables for halophilic research. However, there still remain potential pitfalls and limitations of conventional metagenomic analysis being applied to these complex communities, which require more specialized experimental and analytical strategies to overcome. Several technologies such as RNA sequencing, long read technologies, and chromosome conformation assays, which were initially not 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 halophilic research.

**Introduction**

Microbial life is one of the most diverse and bio-energetically dominant forces in Earth’s ecosphere (*1*), making microbiome research a critical aspect of modern ecology. The unparalleled taxonomic and functional diversity of microbiomes allowed them to populate the vast majority of potentially habitable locations on the planet (*2, 3*), including environments unfit for habitation by other life-forms. In hyper-saline environments, the unique environmental pressures forced microbiota to evolve unique survival adaptations, resulting in streamlined and highly resilient organisms that push the boundaries of life’s limit in the most extreme environments (*4*). Studying halophilic life-forms revealed many fundamental aspects of life’s survival limits and strategies, including its potential endurance in harsh environments such as other planets (*5, 6*). Additionally, halophiles have been found to play important roles in soil bioenergetics processes (*7*), food storage and preservation (*4, 8*), and human gut microbiota (*9*). Before application of high-throughput sequencing technology to microbial communities, our knowledge of halophiles remained limited to studying individual artificially cultured organisms, and most aspects of their activities in natural environments remained a mystery (*10, 11*). While next-generation sequencing technologies have become commonplace in the microbiology field, there has been lack of critical analysis of the prospects of the application of these technologies in halophilic microbiomes. In this review, we discuss the advantages and limitations of applying shotgun metagenomic sequencing to uncover the structure and functioning of halophilic microbiomes, while also discussing unique analytical adjustments that are required to de-convolute these complex communities.

**Shotgun sequencing in metagenomics**

The rapid improvements in high-throughput DNA sequencing technologies since 2008 has propelled our understanding of not only single-organism genetics, but also microbiome community structure and function. While marker gene (particularly 16S rRNA) amplicon sequencing revealed the taxonomic composition of a given community through sequencing a small targeted fraction of the community’s DNA, the remaining microbial genomic content was left unexplored. Since 2012, however, the rapidly growing field of shotgun sequencing has been applied to microbiomes, theoretically allowing for reconstruction of the entire microbial community metagenome by randomly sampling and sequencing all of its DNA content. Since then, whole metagenomic sequencing (WMGS) let 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 being confined to studying culturable species (*14*).

In WMGS, the extracted DNA content of a microbial community is randomly broken into short (100-1000bp) fragments, and sequenced. Most sequencing platforms produce short reads (50-200 base pairs), which come from random locations of various organisms in the sequenced community (*15, 16*). While sequencing technologies are rapidly developing, they are still far from being able to produce the complete genomes of all the organism found in a microbial community. Even in an isolation culture of a single strain, assembling its genome from short-read shotgun sequencing data is a major computational challenge, and is often compared to completing a jigsaw puzzle with a million pieces (*17*). On the other hand, producing complete separate genomes from WMGS of a complex microbial community, where each read belongs in a random position on a random genome can be equated to trying to assemble hundreds of such jigsaw puzzles from a mixed pile of billions of pieces.

Unfortunately, even the best metagenomic assembly process does not produce complete, circular genomes of all the community members from short-read data. Instead, long contiguous pieces (contigs) of the genomes are produced, typically ranging in length from 1Kbp to 1Mbp (*16, 18*). These contigs then need to be grouped based on the genome they belong to. This process, known as binning, takes advantage of subtle characteristic similarities between contigs belonging to the same genome to group them together (*19*). Binning prediction has only become reliable enough to produce reasonably high-quality metagenome-assembled genomes (MAGs) since 2014. Since then, genome-resolved metagenomic analysis has already led to the discovery of thousands of novel organisms and thus enabled many breakthroughs in characterizing the taxonomic and functional components of microbiomes (*20-22*).

While shotgun metagenomics offers tremendous advantages in being able to recover the more complete taxonomic and functional potential components of microbial communities, sequencing costs deter some researchers from deploying this method in their studies. Genomic assembly relies on overlaps between read sequences to attempt to reconstruct the complete genome sequence. Because reads are randomly dispersed along the genome length, a high read coverage is required for reliably reconstruction (*23*). When sequencing the DNA of a microbial community, this presents a major challenge for the assembly of lowly-abundant majority of the community, which often plays an important role in the biome functioning (*24*). Effective reconstruction of a meaningful fraction of a microbiome requires very deep sequencing of the sample, which comes at a much greater cost compared to amplicon sequencing, in which only a single gene is sequenced from each organism (*25, 26*).

Despite these challenges WMGS carries tremendous benefits, as it allows 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 microbiomes. This new angle of microbiome analysis allowed for prediction of metabolic processes present in communities, the study of community natural selection at the functional level (*27, 28*). The possibility of studying the functional potential of any organism in a microbiome meant that our understanding of microbial evolution and function was no longer limited to cultured organisms. In many fields such as human microbiome research, this hailed a new era for microbial research (*29, 30*).

**Limitations of shotgun metagenomics in halophile research**

Compared to clinical and synthetic microbiomes, the reconstruction of environmental metagenomes was stalled by their sheer complexity. This is especially true in high-salt environments, which often host microbial communities with low taxonomic diversity but very high strain heterogeneity and characteristically high GC-content (*31, 32*). 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 GC content reduces the kmer diversity of the samples (*33, 34*). For example, halophilic endolith community composition is typically dominated by *Halobacteria* and *Salinibater*, however their high strain diversity and high GC-content leads to relatively poor assembly and MAG quality compared to that of the low GC and less abundant *Cyanobacteria,* *Actinobacteria*, and *Gammaproteobacteria* (*35*).

These challenges led to the field halophile WMGS analysis falling behind more simple microbiomes, leading to relatively under-developed taxonomy databases and functional annotation models (*36*). In negative feedback loop, this in turn further stalled progress of halophilic microbiome research. As WMGS becomes commonplace microbiome research, it is crucial that the halophile field takes advantage of the new technology to catch up in its understanding of microbial community assembly and function. Since 2015, analytical method improvements allowed for effective de-convolution of WMGS data from even the most complex microbiomes, which will greatly benefit the halophile research field if applied effectively.

**Experimental design considerations for sequencing a halophilic metagenome**

While important in any microbiome study, it is especially important to keep then end goal in mind when sequencing halophilic microbiomes. Because their high strain-level diversity, the experimental design should avoid adding unnecessary replicates into the study, as each added biological replicate will further complicate the assembly and binning stages of analysis (*34*). 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 intended statistical analysis downstream. There are two general approaches to metagenomic sequencing and analysis – co-assembly of manly shallow samples or individual processing deeply sequenced samples. Both approaches have their benefits and limitations in halophilic metagenomics depending on the microbiome in question, which we discuss below.

The first approach is sequencing many samples with relatively low read coverage and combining the reads during metagenomic assembly. This allows for low sequencing costs per sample, while also producing many MAGs from the co-assembly by leveraging differential abundances of the contigs across samples (*20, 37*). 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, 38*). Finally, co-assembling many samples together grants the ability to recover genomes of low-abundance organisms, which are not possibly to assemble from individual samples due to low coverage (*39*). While attractive, the use of co-assembly in halophilic metagenomes comes with major drawbacks (*34*). In addition to high computational costs of co-assembling large data, the high strain heterogeneity introduced from each new biological replicate can non-intuitively lead to poor assemblies of very abundant taxa in the samples. Halophile microbiomes are often dominated by highly diverse groups of *Euryarchaeota* and *Bacteroidetes* (*40*). Introducing their strain heterogeneity from across multiple biological replicates confuses metagenomic assemblers and leads to fragmented or chimeric assemblies (*34*), which translates into poor-quality MAGs. If broadly capturing the community diversity across many samples is the intent of the study, then these limitations should be considered in data interpretation.

The alternative approach to co-assembly is to sequence a few samples with deep coverage, and process them individually. Because of the reduced strain heterogeneity, individual assemblies produce a more contiguous assemblies given comparable sequencing depth (*41*). After binning each sample separately, the MAGs can be combined into a single set through de-replication, removing duplicate MAGs that share a high nucleotide identity (*42*). As with the co-assembly approach, differential contig coverage across all the samples may be used to improve the binning results (*43*). While this method is superior in heterogeneous communities such as halophilic microbiomes, it comes at a major increase in sequencing cost per sample. For most metagenomes, meaningful assembly requires 25-50Gbp of sequencing data from each sample, which limits the number of samples that can be multiplexed on a sequencing run, which in turn reduces binning accuracy due to limited replication by differential coverage (*44*). For many studies requiring a large numbers of replicates, such as longitudinal studies, the cost of this approach may become prohibitively expensive.

In addition to cost considerations, the inter-sample diversity of the community of interest needs to be taken into account. Communities in homogeneous biomes such as water often contain very similar organisms present at different abundances, which reduces the drawbacks of co-assembly (*38, 39, 45*). On the other hand, attached microbiomes with limited dispersal, such as the highly segregated salt halite nodules, often contain unique taxonomic compositions, making individual assembly advantageous (*43, 46*). 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 (*47*). Regardless of the experimental design, it is critical to process samples, generate libraries, and sequence the samples together to avoid batch effects (*48*). 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 (*48*). For library preparation, it is recommended to use protocols that produce minimal GC biases in coverage, particularly in halophilic communities that have high GC-content variation in their metagenomes (*49, 50*).

**Best bioinformatics practices for halophilic metagenome analysis**

When processing halophilic metagenome sequencing data, it is important to adjust existing pipelines to accommodate for their high strain heterogeneity, high GC-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 bioinformaticians should be making while interrogating halophilic metagenomes. Using automated metagenomic analysis pipelines such as metaWRAP (*20*) or SqueezeM (*51*) may also be used to streamline and simplify analysis, although pipelines that are intended for clinical microbiomes such as gut microbiota should be avoided, as they rely strongly on pre-existing taxonomic and functional databases of closely related organisms.

The pre-processing of WMGS data, which typically includes read trimming, duplicate read removal, and metagenomic assembly, is not notably different from that of other microbiomes. We encourage testing a variety of software and comparing the results with evaluation programs such as FastQC (*52*) (for reads) and MetaQUAST (*53*) (for assembly), as some methods may be more suited for specific types of microbial community types (*54*). For metagenomic assembly, metaSPAdes (*55*) is currently considered to the the best overall, while MegaHIT (*56*) is the best solution when resources are a limiting factor (*57*).

During taxonomic and functional annotation of halophilic metagenomes, it is important to keep in mind that halophiles have extremely limited representation in standard-distribution taxonomic databases (*58, 59*), which will produce significant biases in the sequence annotation, i.e. determining the taxonomy of the sequences. During taxonomy annotation, alignment-based methods such as MegaBLAST(*60*) are usually too specific for halophile microbiomes, while approximate kmer matching search tools such as Kraken (*61*) can produce more balanced taxonomic profiles. If then intention is to obtain the most accurate taxonomic distribution profile of the community, extracting and annotating marker genes such as 16S rRNA genes is usually the best alternative (*62*), as rRNA gene databases are more established and constitute greater taxonomic variety (*63*).

Functional annotation – functional categorization of genes – in halophiles is also severely limited by existing databases. Because many halophilic genes are not annotated in NCBI databases, metagenome-inclusive custom or private databases are preferred, as they contain a greater variety of non-culturable organisms. In particular, services such as the Integrated Microbial Genomes systems (*64*) include taxonomic and functional annotation models that are trained on user-submitted metagenomic data, including from high-quality MAGs. The annotation sensitivity resulting from using this newest metagenomic data is extremely valuable for both functional and taxonomic annotation in relatively understudied systems such as halophilic microbiomes. Annotation pipelines geared towards clinical microbiomes such as HUMANN2 (*65*) should be avoided, as they rely on the presence of closely-related organisms in databases.

Finally, the success of metagenomic binning of the assemblies will depend greatly on the software choice, as binning programs perform differently on various data types (*20*). Additionally, many popular binning software such as metaBAT1 are trained on gut microbiome data (*37*), potentially limiting their efficacy in complex halophilic communities. Furthermore, benchmarking of such algorithms is often also done on real or synthetic gut microbial communities (*54*). Because of this, it is recommended to bin the metagenomic assembly with a variety of modern binning software such as metaBAT2 (*37*) and CONCOCT (*66*) and use a binning consolidation tool such as metaWRAP or DAS\_Tool to produce the best final bin set (*20, 67*). 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 strain heterogeneity of halophilic microbiomes (*34*), more accurate abundance estimation could potentially be obtained with slightly relaxed read alignment parameters to allow more approximate matches.

Despite setbacks and difficulties of implementing WMGS in halophilic communities(*68*). While WMGS still has biases associated with GC content, taxonomic annotation of shotgun reads usually results in more accurate and robust taxonomic profiles compared to amplicon sequencing (*69*).

In hyper-saline aquatic environments environments (*70*) and halite endolithic communities (*46*), recovery of viral sequences allowed for the characterization and study of a major previously unexplored aspect of halophilic microbiomes. In addition to viral sequence extraction, CRISPR arrays encoded in the halophilic genomes have been analyzed to investigate strain dispersal (*43*). Additionally, perfect alignments between CRISPR spacers and virus sequences, which are indicative of previous infections, have been used in solar salterns to predict virus-host interaction in microbial communities (*71*).

With WMGS, high-salt water (*5, 72*), soil (*7*), and endolithic (*35*) microbiomes have been characterized in terms of their functioning, particularly their ability to metabolize substrates of interest. 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 (*73*). Functional annotation of longitudinal studies of halophilic water microbiomes have also led to characterization of horizontal gene transfer, evolutionary dynamics, and functional adaptations across time and space (*35, 38, 72*). With WMGS analysis rapidly improving and halophile databases rapidly growing (*36*), more breakthroughs will follow.

(*74*). Binning of metagenomics assemblies has allowed researchers to recover thousands of halophilic MAGs in the past decade (*36*), many of which 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 (*39, 40*). In a positive-feedback loop, the rapidly increasing number or annotate reference halophile genomes is allowing for more accurate taxonomic and functional annotation in halophilic microbiomes (*36*).

**The future of sequencing in halophilic metagenomics**

Beyond shotgun sequencing of a microbiome DNA content, there exist a number of other technologies that may further our ability to understand halophilic microbiome function in the near future. These technologies have already been applied in more developed microbial fields such as clinical gut microbiomes, and show great promise to be applied to halophilic microbial communities.

Together with sequencing coverage, read length is undoubtedly a major limiting factor for metagenomics sequence assembly. Longer reads allow for more accurate assembly and reduced chimeras, while also improving the contiguity of the assembly by allowing assembly of repetitive DNA elements (*75*). Conventional Illumina sequencing is limited to only short DNA fragments (50bp-250bp), as errors accumulate rapidly at higher read lengths. However, two other sequencing technologies – Nanopore and PacBio sequencing – allow for sequencing of much longer DNA fragments. PacBio, is able to consistently produce long reads (N50 up to 10Kbp) with a relatively high degree of accuracy (*76, 77*), while Nanopore sequencing produces even longer reads (N50 up to 100Kbp), with some sacrifices made to accuracy (*78, 79*). These read lengths allows for not only sequencing of complete ribosomal genes for improved taxonomic annotation, but also for significantly improving the accuracy of metagenomics assembly and binning (*77, 80*). In highly diverse halophilic communities, long reads can help assemble ambiguous regions resulting from high taxonomic heterogeneity, drastically improving metagenomic assembly (*80*). Pseudo-single cell technology such as 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 (*81*). 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 (*82*). When this becomes possible, it would propel the field of metagenomics into a new post-assembly era.

Chromosome conformation capture assays (Hi-C) is another sequencing technology that shows great promise in halophilic metagenomics. Conventionally used to indirectly measure the proximity between sections of a genome, HiC was successfully applied to microbiomes to improve binning predictions in 2017 (*83*). Considering the difficulty of binning halophilic metagenomes due to their heterogeneity, HiC could significantly improve halophile MAG extraction. HiC-based binning also allows for recovery of extra-genomic elements such as viral and plasmid DNA, which so far has been impossible to accomplish (*84*). Finally, HiC can be used to produce DNA proximity maps in individual MAGs, allowing for the study of chromatin conformation in prokaryotes at the metagenomic scale (*84*).

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 (*29, 85*). While metatranscriptomics have been used in halophile research to characterize carbon cycling in saline soils (*86*), and used extensively to characterize activity in soil microbiomes (*87, 88*), it remains a largely under-deployed tool. One major deterrent has been difficulty in standardizing transcript expression to the abundance of each individual organism in that sample. 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 filtering RNA reads to their respective MAG.

**Conclusion:**

Shotgun sequencing resulted in rapid advancement in metagenomic research in a variety of fields, as it allowed for genomic and functional interrogation of novel unculturable organisms. However, the genomic qualities and composition characteristics of halophilic communities made them difficult to de-convolute in a metagenomic context, limiting the application of shotgun metagenomics in halophilic microbiome research. This led to underrepresentation of halophiles in existing taxonomical and functional databases, which further complicated analysis. In this review we covered the benefits and limitations of next-generation sequencing in microbial research, and presented the core experimental and analytical considerations for applying it to halophilic communities. 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. Successfully application of whole metagenomics to these complex communities was 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. With the proper utilization, rapidly advancing sequencing technology has the potential to rapidly advance the field of halophilic research.

**Possible figures for the paper (just ideas at this point):**

*Figure 2: Two flowcharts showing the two common experimental designs and analysis workflows.*



Table 1: Outline of benefits and drawbacks of individual and merged metagenomic assembly

|  |  |  |
| --- | --- | --- |
|  | Individual assemblies | Co-assembly |
| Chimera potential | Low | Medium-High (depending on strain diversity) |
| Assembly accuracy | Good (except highly heterogeneous taxa) | Medium-Poor (depending on strain diversity) |
| Abundant taxa recovery | Very good | Poor (for highly heterogeneous taxa) |
| Analysis costs (time and memory) | Cheap | Expensive (for assembly) |
| Binning accuracy | Poor/Good (depends on number of samples) | Good |
| Rare taxa recovery | Poor (limited by coverage) | Good |
| Sequencing cost per sample | Expensive | Very cheap |

**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. A. Gibtan *et al.*, Diversity of Extremely Halophilic Archaeal and Bacterial Communities from Commercial Salts. *Front Microbiol* **8**, 799 (2017).

5. A. Oren, Halophilic archaea on Earth and in space: growth and survival under extreme conditions. *Philos Trans A Math Phys Eng Sci* **372**, (2014).

6. 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).

7. 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).

8. 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).

9. E. H. Seck, J. C. Dufour, D. Raoult, J. C. Lagier, Halophilic & halotolerant prokaryotes in humans. *Future Microbiol* **13**, 799-812 (2018).

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. A. Howe, P. S. Chain, Challenges and opportunities in understanding microbial communities with metagenome assembly (accompanied by IPython Notebook tutorial). *Front Microbiol* **6**, 678 (2015).

16. J. S. Ghurye, V. Cepeda-Espinoza, M. Pop, Metagenomic Assembly: Overview, Challenges and Applications. *Yale J Biol Med* **89**, 353-362 (2016).

17. M. Pop, Genome assembly reborn: recent computational challenges. *Brief Bioinform* **10**, 354-366 (2009).

18. 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).

19. K. Sedlar, K. Kupkova, I. Provaznik, Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Comput Struct Biotechnol J* **15**, 48-55 (2017).

20. G. V. Uritskiy, J. DiRuggiero, J. Taylor, MetaWRAP - a flexible pipeline for genome-resolved metagenomic data analysis. *bioRxiv*, (2018).

21. 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).

22. N. Sangwan, F. Xia, J. A. Gilbert, Recovering complete and draft population genomes from metagenome datasets. *Microbiome* **4**, 8 (2016).

23. 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).

24. R. Zaheer *et al.*, Impact of sequencing depth on the characterization of the microbiome and resistome. *Sci Rep* **8**, 5890 (2018).

25. M. Hamady, R. Knight, Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res* **19**, 1141-1152 (2009).

26. T. Thomas, J. Gilbert, F. Meyer, Metagenomics - a guide from sampling to data analysis. *Microb Inform Exp* **2**, 3 (2012).

27. F. Sharifi, Y. Ye, From Gene Annotation to Function Prediction for Metagenomics. *Methods Mol Biol* **1611**, 27-34 (2017).

28. J. Wang *et al.*, Metagenomic sequencing reveals microbiota and its functional potential associated with periodontal disease. *Sci Rep* **3**, 1843 (2013).

29. W. L. Wang *et al.*, Application of metagenomics in the human gut microbiome. *World J Gastroenterol* **21**, 803-814 (2015).

30. 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).

31. S. Cuadros-Orellana *et al.*, Genomic plasticity in prokaryotes: the case of the square haloarchaeon. *ISME J.* **1**, 235-245 (2007).

32. 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).

33. 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).

34. 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).

35.

36. A. Loukas, I. Kappas, T. J. Abatzopoulos, HaloDom: a new database of halophiles across all life domains. *J Biol Res (Thessalon)* **25**, 2 (2018).

37

38.

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).

39.

40. 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).

41. 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).

42. 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).

43.

44. J. K. Goodrich *et al.*, Conducting a microbiome study. *Cell* **158**, 250-262 (2014).

45. 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).

46.

47. R. D. Stewart *et al.*, Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat Commun* **9**, 870 (2018).

48. S. M. Gibbons, C. Duvallet, E. J. Alm, Correcting for batch effects in case-control microbiome studies. *PLoS Comput Biol* **14**, e1006102 (2018).

49. 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).

50. 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).

51. J. Tamames, F. Puente-Sanchez, SqueezeM, a highly portable, fully automatic metagenomic analysis pipeline. *bioRxiv*, (2018).

52. 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).

53. A. Mikheenko, V. Saveliev, A. Gurevich, MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* **32**, 1088-1090 (2016).

54. A. Sczyrba *et al.*, Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. *Nat Methods* **14**, 1063-1071 (2017).

55. S. Nurk, D. Meleshko, A. Korobeynikov, P. A. Pevzner, metaSPAdes: a new versatile metagenomic assembler. *Genome Res* **27**, 824-834 (2017).

56. 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).

57. 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).

58. D. L. Wheeler *et al.*, Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **29**, 11-16. (2001).

59. 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).

60. Y. Chen, W. Ye, Y. Zhang, Y. Xu, High speed BLASTN: an accelerated MegaBLAST search tool. *Nucleic Acids Res* **43**, 7762-7768 (2015).

61. D. E. Wood, S. L. Salzberg, Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* **15**, R46 (2014).

62. 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).

63. 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).

64.

65. S. Abubucker *et al.*, Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol* **8**, e1002358 (2012).

66. J. Alneberg *et al.*, Binning metagenomic contigs by coverage and composition. *Nat Methods* **11**, 1144-1146 (2014).

67. C. M. K. Sieber *et al.*, Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol*, (2018).

68.

69.

70.

71.

72.

73.

74

75. K. Wommack, E., J. Bhavsar, J. Ravel, Metagenomics: read length matters. *Appl Environ Microbiol.* **74**, 1453-1463 (2008).

76. A. Rhoads, K. F. Au, PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* **13**, 278-289 (2015).

77. J. A. Frank *et al.*, Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. *Sci Rep* **6**, 25373 (2016).

78. 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).

79. 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).

80. 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).

81. E. Moss *et al.*, De novo assembly of microbial genomes from human gut metagenomes using barcoded short read sequences. *bioRxiv*, (2017).

82. M. Jain *et al.*, Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **36**, 338-345 (2018).

83. 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).

84. 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).

85. A. Lavelle, H. Sokol, Gut microbiota: Beyond metagenomics, metatranscriptomics illuminates microbiome functionality in IBD. *Nat Rev Gastroenterol Hepatol* **15**, 193-194 (2018).

86. 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).

87. A. Garoutte, E. Cardenas, J. Tiedje, A. Howe, Methodologies for probing the metatranscriptome of grassland soil. *J Microbiol Methods* **131**, 122-129 (2016).

88. 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).