Genomes From Uncultivated Microorganisms

Tanja Woyke, Devin FR Doud, and Emiley A Eloe-Fadrosh, DOE Joint Genome Institute, Walnut Creek, CA, United States

A Brief History of Microbial Genomics

In 1995, the first complete genome of a free-living microorganism, that of the bacterium *Haemophilus influenzae*, was sequenced by J. C. Venter and colleagues. This achievement proved the utility of shotgun genome sequencing and the discipline of microbial genomics was born. The years that followed were marked by sequencing genomes of bacterial and archaeal cultured isolates, and nearly 25 years later, well over 90,000 bacterial and nearly 900 archaeal isolate genome sequences are available in the public domain (**Fig. 1**). Due to our inability to cultivate the majority of microorganisms, cultivation-independent approaches to microbial genome discovery and identification, namely metagenomic sequencing, came to light in 2004 (Tyson *et al.*, 2004; Venter *et al.*, 2004) and have since been incredibly popular. Initially, cultivation-independent approaches were necessarily restricted to gene-centric analyses unless the microbial diversity of the sampled environment was very low; however, in recent years, genome-resolved metagenomics has become feasible through advances in sequencing technologies, metagenome assembly, and, importantly, computational binning algorithms (Wrighton *et al.*, 2012; Albertsen *et al.*, 2013). Genome-resolved metagenomics provides clear links between phylogeny and function, and offers population-level information on genome variability.

Complementary to genome-resolved metagenomics is microbial single-cell genomics, the sequencing of the genome from an individual cell directly isolated from the environment (for a review, see (Woyke *et al.*, 2017)). Single-cell genomics emerged in 2005 when A. Raghunathan and colleagues demonstrated that sequence data from a single *Escherichia coli* cell could be obtained. Two years later the first genomes were recovered from the candidate phylum Saccharibacteria (formerly TM7) using single-cell sequencing. Since then, single-cell genomics methods have been widely adopted to complement metagenomics. To date, more than 5,000 bacterial and archaeal single amplified genomes (SAGs) and nearly 13,000 metagenome-assembled genomes (MAGs) from bacteria and archaea are in the public domain (**Fig. 1**). These genomes provide a rich resource for the phylogenetic and functional interrogation of the uncultivated majority within the tree of life.

Technical Aspects for Accessing Genomes From Uncultivated Microorganisms

Genome-Resolved Metagenomics

The term 'metagenomics' was coined in 1998 by J. Handelsman and colleagues to describe an approach for exploring phylogenetic and functional diversity by directly extracting and cloning environmental DNA. The basic experimental methodologies have remained stable, while the field has seen major advances in high-throughput sequencing technologies and downstream computational analyses. Mechanical or chemical lysis approaches are used to extract total nucleic acids from cells from a given environmental sample (**Fig. 2**). Extensive literature exists detailing impacts of sample preservation protocols, microbial biomass considerations, DNA extraction methodologies, and contamination considerations (for a review, see Quince *et al.*, 2017). Based on sample quantification, subsequent library preparation approaches are predominantly geared towards sequencing on the Illumina platform. These library preparation methods include transposase-based 'tagmentation,' used in the Illumina Nextera and Nextera XT products, and PCR-free methods relying on physical fragmentation, used for Illumina TruSeq preparations. Known biases based on library preparation methods and the number of PCR cycles have been documented and are known to impact taxonomic abundance profile calculations. Increasing sequence output by leveraging the newly released Illumina NovaSeq platform promises upwards of 6 Tb in a dual flow cell run, enabling even deeper sequencing of environmental samples.

Single-molecule long read sequencing technologies, including Oxford Nanopore and Pacific Biosciences, hold potential for metagenomic applications, yet are not widely utilized because it is challenging to extract high-quality, high molecular weight DNA.

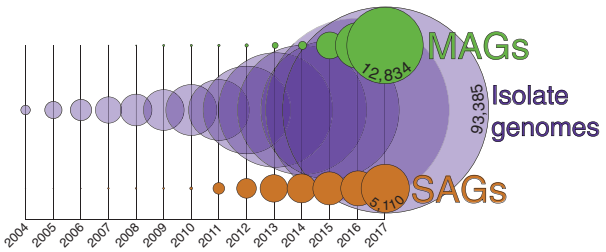

Fig. 1 The number of microbial isolate genomes, single amplified genomes (SAGs), and metagenome-assembled genomes (MAGs) sequenced over time. Data were extracted from the Genomes OnLine Database (GOLD; https://gold.jgi.doe.gov) on 11/21/2017.

 doi:10.1016/B978-0-12-809633-8.90682-4

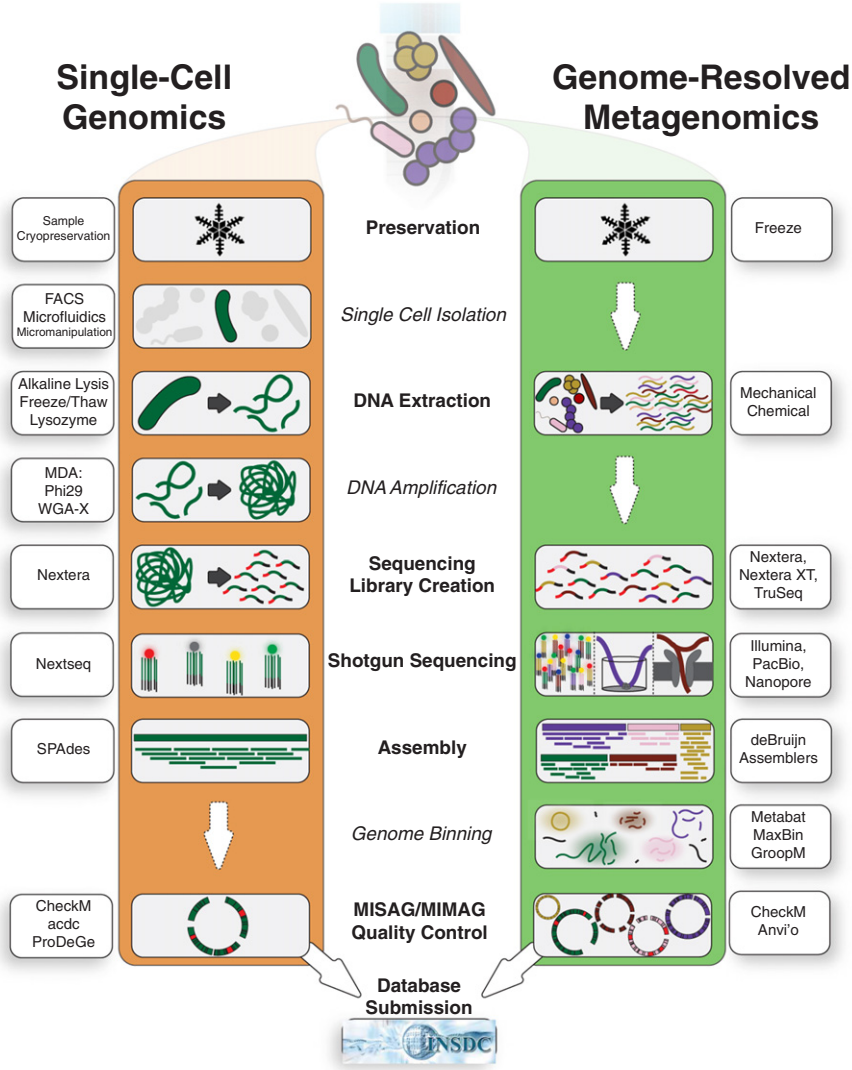

Fig. 2 Workflow for laboratory and analysis approaches of single-cell genomics and genome-resolved metagenomics.

Alternatives to true single molecule sequencing generate 'synthetic long reads' from barcoded short reads or apply high-throughput chromosome conformation capture (Hi-C) technology. These and other technological developments on the horizon will undoubtedly shift the current paradigm of genome-resolved metagenomics away from computationally intense short-read assembly dependencies. For instance, a recently described novel approach from J. Beaulaurier and colleagues utilizes DNA methylation signatures derived from single-molecule real-time sequencing to resolve species- and strain-level bins, as well as link mobile genetic elements such as plasmids to their host.

Two avenues for computational analyses can be pursued with short-read shotgun metagenomics, namely *de novo* metagenome assembly and assembly-free metagenomic profiling to estimate taxonomic abundances. Here, we focus on *de novo* metagenome assembly as this approach enables subsequent reconstruction of population genomes to enable genome-resolved metagenomics. The de Bruijn graph approach underpins the majority of current metagenome assembly algorithms, although at present, there is no community consensus as to which assembly method is superior (Quince *et al.*, 2017). While significant improvements in metagenome assembly workflows have been made, challenges remain for highly complex environments that consist of hundreds to thousands of strains.

To resolve metagenome-assembled genomes, termed MAGs, methods to link contigs to their respective genomes – termed binning – are required post-assembly. Binning methods can exploit sequence composition, species abundance, chromosome organization, or other inherent properties of the shotgun metagenomic data. A myriad of binning tools and approaches are available; however all tools are currently limited in their ability to distinguish closely related species and strains. Lastly, MAGs are evaluated for estimated genome completeness and contamination using universal single copy genes. As can be observed in **Fig. 1**, the deposition of MAGs within public databases has proliferated in the last three years and has led to major scientific insights.

Single-Cell Genomics

Although the overall approach to single-cell genomics is rather simple (**Fig. 2**), several technical aspects should be considered when generating and analyzing genomes from individual environmental cells. On the experimental side, sample preparation, cell isolation, cell lysis and whole genome amplification are critical steps in the procedure (Rinke *et al.*, 2014). Unless fresh samples are available for immediate processing, cryopreservation of the sample material using cryoprotectants such as glycerol, betaine, or DMSO is important to minimize cell damage and maximize maintenance of the integrity of the cellular structure and the genomic DNA. Further, samples should be prepared so that cells are well dispersed, facilitating efficient single-cell isolation. Although various methods are available for this next step (for a review see (Blainey, 2013)), including micromanipulation and microfluidics, fluorescence-activated cell sorting (FACS) has been used most prevalently by flow-sorting cells based on their size and level of fluorescence (stain-based or auto-fluorescence). A chief advantage of FACS is its accuracy and speed and thus high throughput. Following isolation, cells are lysed to make DNA accessible to whole genome amplification (WGA). To date, no universal lysis method exists and each sample may require custom adjustments based on the target taxa of interest. Most commonly used methods include alkaline lysis, though chemical lysis has more recently been combined with physical (i.e., freeze-thawing) and enzymatic (i.e., lysozyme treatment) lysis (Blainey, 2013; Rinke *et al.*, 2014). Despite the continued development of a broad range of WGA methods (for a review, see (Blainey, 2013)), Phi29-mediated multiple displacement amplification (MDA) has remained the key technique used for microbial single-cell sequencing. Recently, a thermostable Phi29 polymerase mutant was shown by Stepanauskas and colleagues to yield improved genome recovery for single cells with high GC genomes. Both enzymes are strand-displacement polymerases with high processivity that rely on random hexamers to prime the reaction before amplifying long (>10 kb) DNA products to generate the SAG. Taxonomic identification of SAGs can be performed via direct PCR amplification and Sanger-based sequencing of the small subunit (SSU) rRNA gene. Illumina's Nextera protocol is recommended for generating the shotgun sequencing library because it minimizes sample handling, and thus cross-contamination, and reduces cost. For SAG sequencing, the Illumina NextSeq platform has proven most useful, as bleed-over between poolmates is minimal and short reads/ short insert libraries minimize the occurrence of chimeric reads or read pairs, which hamper the sequence assembly process (see following section for more details).

Genome amplification causes single-cell sequence data to have coverage biases and chimeric junctions occurring approximately every 20 kb. SAG-specific *de novo* genome assemblers, such as SPAdes are therefore recommended (**Fig. 2**), as they are optimized to correct for such data artifacts. Further, due to the amplification of low femtogram-range DNA via random hexamer primers, the single-cell genomics process is prone to contamination, and thus, thorough quality assessment and assurance (QA) of the data is advised (Bowers *et al.*, 2017a). Several tools have been developed in recent years to facilitate single-cell data QA, such as ProDeGe (see "Relevant Websites section"), CheckM (see "Relevant Websites section"), acdc (see "Relevant Websites section"), and Anvio (see "Relevant Websites section").

Quality Considerations for Genomes From Uncultivated Microorganisms

MAGs and SAGs are often of draft quality, i.e., they are less complete and more fragmented than isolate genomes. The median completeness estimate of single-cell genomes from environmental cells is approximately 40%, though combined assembly of SAGs from the same species, as defined by average nucleotide identity (ANI), can yield more complete composite genomes. Only one complete, finished single-cell genome has been obtained to date. For MAGs, genomes are comparably incomplete, though several finished genomes have been reported (Albertsen *et al.*, 2013; Brown *et al.*, 2015). While most SAGs contain the SSU rRNA genes in the assembly, many MAGs do not due to the challenge of assembling and binning this highly conserved gene in a complex metagenome. When SSU rRNA genes are present, taxonomic inferences can be made based on the SSU rRNA gene based phylogeny, though phylogenetic placement of MAGs and SAGs is most often made through concatenated alignments of protein-coding single-copy marker genes.

Considering the variable quality of SAGs and MAGs, it is important to critically assess and report genomes quality. Recently, a set of community standards was put forward for reporting genome sequences of uncultivated microorganisms: the minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea (Bowers *et al.*, 2017b). MISAG and MIMAG encompass a minimal set of mandatory genome quality criteria, such as the reporting of genome completeness and contamination estimates. Considering the various different downstream analyses for SAGs and MAGs (e.g., basic metabolic reconstructions for comparative genomics, fragment recruitment for biogeographic analyses, and phylogenetic inference for evolutionary analysis and population genomics), reporting completeness and contamination estimates is imperative (**Fig. 2**), as some analyses require finished or high-quality genomes, while others may be feasible with medium- or low-quality genomes. Additionally, the consistent reporting of basic environmental metadata is suggested (Bowers *et al.*, 2017b), as downstream comparative genomics is only as good as the metadata of a genomic dataset. Adhering to these standards will facilitate more reproducible and robust data interpretation and comparative genomic analysis to make solid inferences about evolutionary relationships and ecosystem functions of uncultivated bacteria and archaea.

From Individual Cells to Ecosystems

The tremendous diversity of microbes that impact the environment, animal and plant health, and serve as major drivers of global biogeochemical cycles have remained largely uncatalogued and underexplored. Major scientific advances within the

past decade have been achieved through applying single-cell genomics and genome-resolved metagenomics. In particular, the microbial tree of life has greatly expanded with genomic information from newly identified candidate phyla (those phyla for which no cultivated isolate exists). The Patescibacteria were first described using single-cell genomics (Rinke *et al.*, 2013) and later expanded through genome-resolved metagenomics to the Candidate Phyla Radiation (CPR) (Brown *et al.*, 2015). This group of bacteria has small reduced genomes and lacks certain ribosomal proteins, implying divergent and unusual ribosome structures and biogenesis mechanisms (Brown *et al.*, 2015). Complementary to the Patescibacteria, C. J. Castelle and colleagues utilized genome-resolved metagenomics to characterize deep-branching lineages constituting a large radiation within the Archaea with similarly small genomes that encode metabolic capacity for carbon and hydrogen metabolism relevant to Earth's biogeochemical cycles. Further, metagenomic reconstructions of candidate phyla from the Archaea provided unique insight into the origin and evolution of the eukaryotes through the recovery of the Asgard superphylum (Zaremba-Niedzwiedzka *et al.*, 2017). While the exact phylogenomic placement of this clade is currently hotly debated, these genomes are enriched for proteins of eukaryotic origin, including homologues of eukaryotic membrane-trafficking machinery components and vesicle biogenesis.

Beyond cataloguing phylogenomic diversity, single-cell genomics and genome-resolved metagenomics have enabled important insights into how microbial metabolisms impact elemental fluxes and how those transformations influence ecosystem processes. For example, novel modes of metabolic coupling in the candidate phylum Marinimicrobia (formerly SAR406 and Marine Group A) were recently revealed by A. K. Hawley and colleagues through single-cell genomics with important implications for sulfur and nitrogen cycling. Similarly, single-cell genomic approaches utilized by D. Tsementzi and colleagues uncovered genes for respiratory nitrate reductases (Nar) encoded within certain clades of the highly abundant marine bacterium SAR11. Heterologous expression of the putative SAR11 *nar* operons verified functionality for the first step of denitrification, linking this important marine microbe to nitrogen loss pathways within oxygen minimum zones and expanding its known ecological niche. Furthermore, genome-resolved metagenomic approaches have led to major advances in understanding carbon fluxes mediated by uncultivated lineages, including methane metabolism in the archaeal phylum Bathyarchaeota and carbon fixation in the bacterial phyla *Candidatus* Eremiobacteraeota and *Candidatus* Dormibacteraeota (formerly WPS-2 and AD3, respectively). Importantly, functional capabilities can be tracked through better genomic representation of uncultivated lineages to shed light on key evolutionary origins. An elegant example by Soo *et al.* (2017) demonstrated independent acquisition of aerobic respiratory complexes within the three classes of Cyanobacteria (*Oxyphotobacteria*, *Melainabacteria*, and *Sericytochromatia*), supporting the hypothesis that aerobic respiration evolved after oxygenic photosynthesis approximately 2.3 billion years ago.

Owing to their synergy, single-cell genomics and genome-resolved metagenomic approaches are increasingly being used in combination (**Fig. 3**). Single-cell genomics offers genome-level resolution from an individual cell and can provide better spatial resolution. However, partial genome recovery and the need for specialized instrumentation and whole genome amplification results in low throughput and biases. On the other hand, genome-resolved metagenomics provides a "consensus" genome which is typically more complete compared to SAGs (provided sufficient sequencing and a high-quality

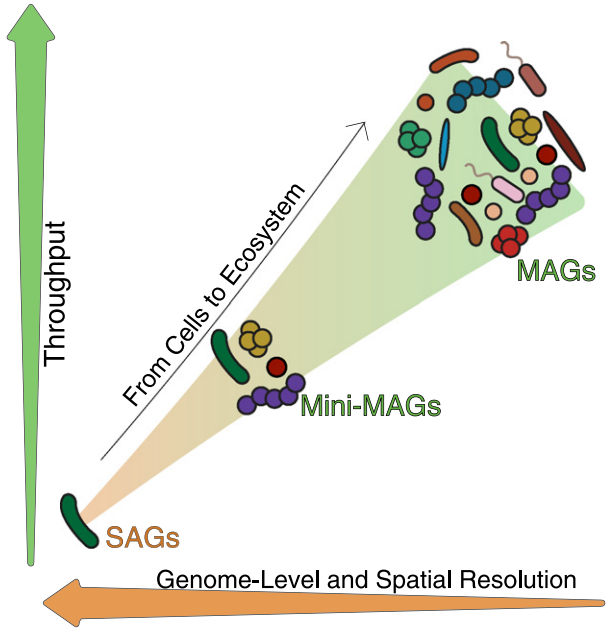

Fig. 3 From cells to ecosystems. Single cells provide genome-level and spatial resolution, which are critical to understand population structure in heterogeneous communities. MAGs from bulk metagenomes are produced at much higher throughput and without the bias of cell isolation and whole genome amplification. Mini-MAGs can be generated by subsampling an environmental sample.

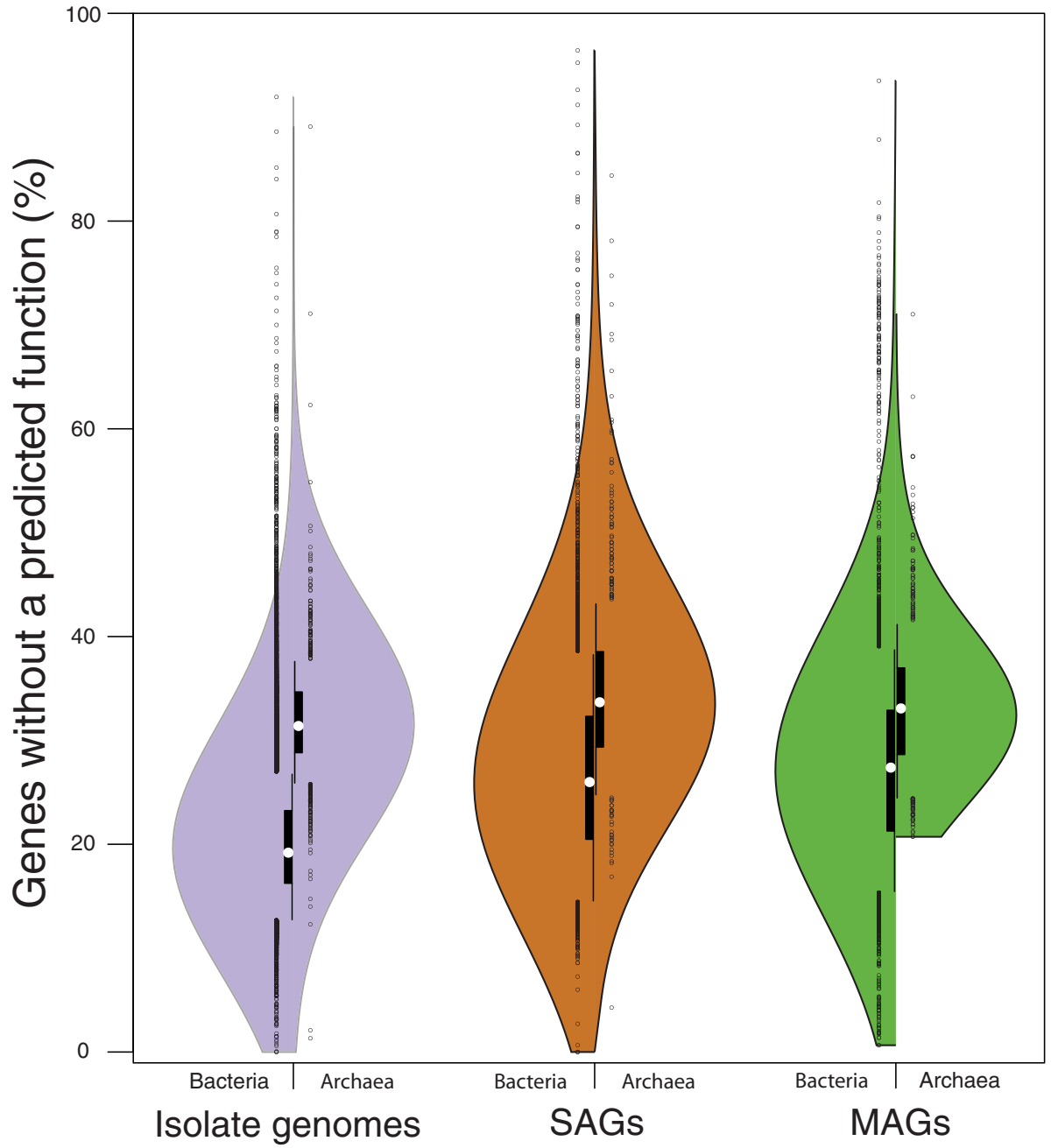

Fig. 4 Percentage of genes without a predicted function for bacterial and archaeal isolate genomes, SAGs and MAGs, illustrating a large knowledge gap in function assignment. Data were extracted from the Integrated Microbial Genomes and Microbiomes platform (https://img.jgi.doe.gov, IMG/M) on 1/16/2018.

metagenome assembly), incorporating genetic information from genotypically heterogeneous populations. It is currently challenging to resolve strain-level variation from MAGs, although new approaches have recently been developed to address these challenges. Early studies applying both technologies predominantly utilized the partial SAGs to recruit metagenomic data, but more recent studies leverage both approaches to examine species diversification, adaptive properties and ecological patterns within microbial communities. Further, aspects of both single-cell and shotgun metagenome strategies have been combined through a newly described microfluidics-based mini-metagenomic method from F. B. Yu and colleagues, which allows single-cell resolution and improved genome recovery. We anticipate future studies leveraging both single-cell genomics and metagenomics will make significant strides towards addressing fundamental questions in microbial ecology, niche partitioning and microbial evolution.

Future Outlook

The tremendous increase in genomes from uncultivated microorganisms (**Fig. 1**), including from candidate phyla with no cultivated representatives, provides an exciting foundation for the functional interrogation of this microbial dark matter. The down side of the data deluge, however, is the increasing gap in our ability to assign function to many of the newly discovered genes, proteins, and pathways (**Fig. 4**). Steps to fill the function gap are most commonly achieved by moving from sequence to function via designing experiments to validate sequence-based predictions of function; these experiments rely on technologies such as DNA synthesis and protein expression followed by a functional assay. An alternate approach is "function-driven" (meta)genomics. Here moving from function to sequence, prior to sequencing, cells or DNA from organisms are selected and enriched based on a particular phenotype. Function-driven single-cell genomics (Doud and Woyke, 2017) may target cells of general metabolic activity or highly specific activities, as demonstrated in a recent study by M. Martinez-Garcia and colleagues, where populations degrading the substrate laminarin were captured and genome sequenced following the addition of a fluorescently labeled form of laminarin. Analogous to these approaches is the application of stable isotope probing (SIP) to link microbial activity (function) to taxonomic identity within an environmental sample. This function-driven metagenomic approach relies on the incorporation of heavy isotopes (for example, ^{13}C) into microbial DNA during growth on labeled substrates. Early successes of the technology include the identification of dimethyl sulphide (DMS) functional capacity within previously unknown bacterial groups by O. Eyice and colleagues, and most recently, the application of DNA-SIP for genome-resolved metagenomics by R. M. Ziels and colleagues. While a very promising approach, DNA-SIP is currently not widely utilized because it requires specialized laboratory equipment and technical expertise. However, we anticipate with advanced high-throughput protocols relying on robotic automation, along with improvements in depth of coverage for low-biomass samples, this technology will become more broadly accessible. While there is no magic bullet for validating function, and the aforementioned approaches are often rather custom, tedious and low throughput, it is important to continue making progress on the functional verification of genome sequence space. The integration of functional and phenotypic data with genomics will ultimately move us towards a better systems biology and ecosystem understanding of uncultivated bacterial and archaeal taxa.

References

Albertsen, M., Hugenholtz, P., Skarshewski, A., *et al*., 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nature Biotechnology 31 (6), 533–538.
Blainey, P.C., 2013. The future is now: Single-cell genomics of bacteria and archaea. FEMS Microbiology Reviews 37 (3), 407–427.
Bowers, R.M., Doud, D.F.R., Woyke, T., 2017a. Analysis of single-cell genome sequences of bacteria and archaea. Emerging Topics in Life Sciences 1 (3), 249–255.
Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., *et al*., 2017b. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nature Biotechnology 35 (8), 725–731.
Brown, C.T., Hug, L.A., Thomas, B.C., *et al*., 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523 (7559), 208–211.
Doud, D.F.R., Woyke, T., 2017. Novel approaches in function-driven single-cell genomics. FEMS Microbiology Reviews 41 (4), 538–548.
Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J., Segata, N., 2017. Shotgun metagenomics, from sampling to analysis. Nature Biotechnology 35, 833.
Rinke, C., Lee, J., Nath, N., *et al*., 2014. Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. Nature Protocols 9 (5), 1038–1048.
Rinke, C., Schwientek, P., Sczyrba, A., *et al*., 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499 (7459), 431–437.
Soo, R.M., Hemp, J., Parks, D.H., Fischer, W.W., Hugenholtz, P., 2017. On the origins of oxygenic photosynthesis and aerobic respiration in Cyanobacteria. Science 355 (6332), 1436–1440.
Tyson, G.W., Chapman, J., Hugenholtz, P., *et al*., 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428 (6978), 37–43.
Venter, J.C., Remington, K., Heidelberg, J.F., *et al*., 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304 (5667), 66–74.
Woyke, T., Doud, D.F.R., Schulz, F., 2017. The trajectory of microbial single-cell sequencing. Nature Methods 14 (11), 1045–1054.
Wrighton, K.C., Thomas, B.C., Sharon, I., *et al*., 2012. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. Science 337 (6102), 1661–1665.
Zaremba-Niedzwiedzka, K., Caceres, E.F., Saw, J.H., *et al*., 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541, 353.

Relevant websites

https://github.com/mlux86/acdc
Acdc.
http://merenlab.org/software/anvio/
Anvio.
http://ecogenomics.github.io/CheckM/
CheckM.
https://gold.jgi.doe.gov
GOLD.
http://ecogenomics.github.io/GroopM/
GroopM.
https://prodege.jgi.doe.gov/
JGI.
https://www.illumina.com
Illumina.

https://img.jgi.doe.gov
IMG.
https://sourceforge.net/projects/maxbin/
MaxBin.
https://bitbucket.org/berkeleylab/metabat
MetaBAT.
http://bioinf.spbau.ru/spades
SPAdes.