Systematic and Applied Microbiology 38 (2015) 223–230

Contents lists available at ScienceDirect

Systematic and Applied Microbiology

journal homepage: www.elsevier.de/syapm

Classifying the uncultivated microbial majority: A place for metagenomic data in the *Candidatus* proposal

Konstantinos T. Konstantinidis[a,b,*], Ramon Rosselló-Móra[c,**]

[a] School of Civil and Environmental Engineering, Georgia Institute of Technology, 311 Ferst Dr. NW, Atlanta, GA 30332, USA
[b] School of Biology, Georgia Institute of Technology, 311 Ferst Dr. NW, Atlanta, GA 30332, USA
[c] Marine Microbiology Group, Institut Mediterrani d'Estudis Avançats (IMEDEA; CSIC-UIB), E-07190 Esporles, Spain

ARTICLE INFO

Keywords:
Candidatus
Species concept
Taxonomy
Metagenomics
16 rRNA gene
Uncultivated diversity

ABSTRACT

Microbial taxonomists have generally been reluctant to accept the valid publication of names of uncultured taxa given that only pure cultures allow for a thorough description of the genealogy, genetics and phenotype of the putative taxa to be classified. The classification of conspicuous uncultured organisms has been considered into the *Candidatus* provisional status, but this is only possible with organisms for which it is possible to retrieve basic data on phylogeny, morphology, ecology and some metabolic traits that unequivocally identify them. The current developments on modern sequencing techniques, and especially metagenomics, allow the recognition of discrete populations of DNA sequences in environmental samples, which can be considered to belong to individual closely related populations that may be identified as members of yet-to-be described species. The recognition of such populations of (meta)genomes allow the retrieval of valuable taxonomic information, i.e. genealogy, genome, phenotypic coherence with other populations, and ecological relevant traits. Such traits may be included in the *Candidatus* proposals of environmentally occurring, yet uncultured species not exhibiting exceptional morphologies, phenotypes or ecological relevancies.

The uncultured diversity and the *Candidatus* status for putative taxa

It is nowadays a commonplace to say that the vast majority of the microbial diversity has never been brought to culture in the laboratory. This realization is primarily due to the continuous development of molecular techniques applied to, in the first place taxonomy, and in the second place microbial molecular diversity studies of environmental samples. The taxonomy of prokaryotes has empirically developed in parallel to the technological advances that have allowed the retrieval of valuable genetic and phenotypic information beyond simple morphological traits [29]. However, the real success in the application of taxonomic methods directed to the identification of microorganisms has occurred in environmental microbiology. There, culture-independent technologies have been extensively used to characterize naturally occurring communities. Microbial molecular ecology studies have overwhelmingly surpassed in efforts, scientific production and funding those on taxonomy. This is well reflected in the larger number of journals with higher citation indexes publishing microbial molecular ecology than taxonomic studies. The extensive interest in the discovery of environmental taxonomic novelty has resulted in the generation of millions of sequences that are deposited in public repositories, providing a better appreciation of the extent of genetic diversity on earth. In this regard, the number of deposited 16S rRNA gene sequences is currently close to 4 million [26]. On the other hand, the fraction of these sequences that corresponds to cultured organisms is far below 1% of the total entries, and that of the type strains of the hitherto classified species even smaller (n = 12,000) [7,50].

There are neither official rules for the classification of prokaryotic taxa, nor an official classification [36]. Only the nomenclature of taxa is regulated under the International Code of Nomenclature of *Bacteria* [18]. Actually, for most taxonomists, the official recognition that a species is accepted as a new taxon is the valid publication of its name in the official journal of the International Committee for Systematics of Prokaryotes (ICSP). The effective publication of a name (i.e. the description and classification of the new taxon) occurs by either its publication in the International

* Corresponding author at: School of Civil and Environmental Engineering, Georgia Institute of Technology, 311 Ferst Dr., Atlanta, GA 30332-0512, USA.
Tel.: +1 404 385 3628.
** Corresponding author. Tel.: +34 971 611 826.
E-mail addresses: kostas@ce.gatech.edu (K.T. Konstantinidis), rossello-mora@uib.es (R. Rosselló-Móra).

http://dx.doi.org/10.1016/j.syapm.2015.01.001
0723-2020/

Journal of Systematic and Evolutionary Microbiology (IJSEM; published and listed in the "Notification Lists"), or when published in other non-official journals, the name should appear (after request of the authors) in the IJSEM "Validation List" [43]. However, in order to validly publish a name, its effective publication should meet several requirements raised in the Bacteriological Code. Some of them are formalities related to the etymological correctness of the proposed name, and other related to the extent of phenotypic and genetic information provided in the protologue. However, the most important cornerstone for the valid publication of a specific name under the Bacteriological Code, which is covered by its rule 18a, is the requirement for a pure culture to be designated as the type strain [18]. This requirement is one of the most controversial issues among microbiologists as it seems to hamper the "classification" of uncultured organisms or microbial consortia (the latter covered by the rule 31b of the code). The benefits of having an organism in pure culture in the laboratory are obvious for microbiological studies. In addition, the special requirement covered by rule 30a of the code, which is the deposition of the type strain in at least two international collections [43] guarantees access and reproducibility of the results in taxonomic studies. On the other hand, the need to isolate pure cultures has several major limitations. At first, and given that most of the environmentally occurring microorganisms are resistant to cultivation [3], the extent of known classifiable taxa surpasses that currently classified by orders of magnitude. Second, the isolation and characterization of organisms needs skills, time and effort, which currently leads to a rate of classification of about 700 species yearly [29], a speed that given the vast diversity awaiting to be classified, is slow. Finally, culture techniques tend to recover a tiny part of the real diversity of a given sample, but also very rarely the most abundant taxa that may play the key roles [24].

Given the difficulties in culturing most of the environmentally occurring organisms, there is an urgent need to classify the extent of diversity discovered by molecular techniques. Actually, the development of methods that allow the identification, observation, quantification, and (in some extent) assessment of metabolic properties, have brought taxonomists to introduce the *Candidatus* status [21,22]. *Candidatus* is a category with no standing in the Bacteriological Code, thus cannot be considered a rank, but a status that is permitted to be listed in the "notification or validation" lists, and denotes that the putative taxon is awaiting for a formal validation once its representatives are brought to pure culture and extensively described. Given that a *Candidatus* name can be formally recorded as susceptible to be validly published, the condensed description should meet some minimal etymological and protologue formalities [22].

Candidatus is the closest status to species with a name that stands in nomenclature. It is generally (and probably wrongly) understood as an incomplete species classification. It is indeed incomplete for the requirements of the Bacteriological Code to recognize the name as validly publishable; yet, it is the closest recognition that a species is formally classified within a hitherto unofficial taxonomy. As long as the data provided in the description (i.e. molecular, morphological, metabolic and ecological traits), unequivocally identifies it, the classification of the organism should be understood as acceptable. It is a somewhat a separate issue whether or not the ICSP considers the name to be provisional.

The activities of environmental microbiologists using molecular methods to describe uncultured taxa has led to the description of almost 360 *Candidatus* taxa (http://www.bacterio.net/-candidatus.html), but in all cases the organisms exhibited very conspicuous traits (morphology, size, environmental uniqueness and isolation, extraordinary metabolic properties, etc.) that allowed their unequivocal recognition. Good examples are the giant size of the nitrate oxidizer "*Candidatus* Thiomargarita joergensenii" [33]; or the magnetic body inclusions of "*Candidatus* Magnetobacterium bavaricum" [37]. However, most of the organisms inhabiting environmental niches do not exhibit such conspicuous properties that simplify their recognition. Standard molecular data such as 16S rRNA gene sequence, rRNA targeted probes to identify and quantify intact cells, and even sometimes the assessment of metabolic properties, do not produce enough discriminative data for the unequivocal identification of the majority of uncultivated taxa. However, as exemplified below, the new –omics technologies can contribute substantially in this direction and facilitate the recognition of new taxa as an extension of the parameters to be provided in *Candidatus* proposals.

Table 1
Numbers of validly published names listed in the LPSN, putatively detected taxa in the public repositories (using the thresholds calculated for species (98.7%), genus (94.5%), family (86.5%), order (82.0%), class (78.5%) and phylum (75.0%)) and ambiguous sequences that according to SILVA 114 (December 2012) showed a pintail score below 75% [51]. The data has been calculated with the entries up to year 2012.

	Validly published	SILVA REF 114 (2012)	Ambiguous
Sequences	–	1,306,670	31,469
Species	10,015	241,254	28,983
Genera	2029	80,939	19,607
Families	317	14,369	4803
Orders	137	5366	1148
Classes	88	2573	408
Phyla	29	1356	84

The achievability of a global classification system

16S rRNA gene sequences may be one of the most, if not the most, extensively deposited information in public repositories, currently accounting for almost 4 millions entries [26]. A survey based on sequence identity thresholds has been performed to enumerate putative taxa discovered by environmental studies [51]. The thresholds were calculated based on the Living Tree Project (LTP) database that contains only high-quality, curated sequences of type strains with validly published names [49]. The numbers of detected taxa using the SILVA REF 114 database released in December 2012 [26], considering only good quality sequences longer than 900 nucleotides, were surprising, and revealed numbers close to 250,000 species (with the conservative threshold of 98.7%; [38]), or over 1300 phyla (Table 1). Based on the discovery rates and current sequence deposits, it was calculated that the total number of putative prokaryotic species inhabiting the biosphere would range between 5×10^5 and 2×10^6, but in no case would exceed 10^7 [51]. These numbers were quite encouraging, as the global classification of prokaryotic taxa seemed like an unachievable task previously.

The vast, but finite, extent of taxonomic diversity hitherto detected in the public repositories requires classification for obvious reasons. Among them, it is of paramount relevance to facilitate the understanding of the uncultured diversity using common criteria with cultured organisms. With this in mind, the "Candidatus Taxonomic Unit" (CTU) was proposed [51] as a way to categorize environmental sequences, which, by means of taxonomic thresholds and phylogenetic uniqueness (monophyly), would be reminiscent of the categories recognized in the bacteriological code. A CTU was considered as "*a biological entity delineated by a monophyletic set of sequences with a minimum identity that stays within, or very close to, the taxonomic threshold proposed for a given rank*". Moreover, the CTU was proposed to fit in a hierarchical system with the same categories as that of the cultured organisms, but with an alphanumerical nomenclature that would be computer readable. Applying these criteria, a hierarchical layout was proposed for most of the environmental clades, which could also facilitate the reclassifications of phyla such as the *Spirochaetes* [51].

If the CTU proposal is accepted among the scientific community, the complete dataset of 16S rRNA gene sequences can be structured in a hierarchical system that would reflect taxa that could mirror the cultured hierarchy. Given that the classification (or better the naming of taxa) of uncultured microorganisms is not covered by the Bacteriological Code yet, two parallel taxonomies would coexist, as a case of pluralism [8], but with a tendency to converge. In any case, the *Candidatus* status, goes far beyond the recognition of the uniqueness of a phylogenetic lineage, as it requires the acquisition of ecological, genetic or genomic, and phenotypic data that guarantees that a population represents a new taxon in the frame of the current taxa definition. In this regard, great expectations are based on the new –omics technologies.

The *Candidatus* based on metagenomic data

The last decade has brought the development of highly efficient sequencing approaches known as the Next Generation Sequencing (NGS). The application of NGS in microbiology has, for instance, allowed the determination of full (or nearly full) genome sequences of pure cultures with relative low cost. Further, taxonomy has enormously benefitted from such developments as for the first time in silico genome to genome reassociation studies were possible [16], and permitted the substitution of the wet-lab and error-prone DNA–DNA hybridization (DDH) experiments with more accurate and portable, database-based measures of their gene/genome relatedness [11,27]. In this regard, the Average Nucleotide Identity (ANI, [16]), which is a measure of the average sequence identity of all genes shared between two genomes, has been used to genomically circumscribe prokaryotic species. In general, two organisms sharing ANI values above 94–96% may be considered as members of the same genospecies [11,27]. There are discrepancies between what may be a species [30] from the taxonomic point of view (i.e., coherent groups of organisms identifiable as a unit by means of "polyphasic" studies), and from the ecological point of view (i.e. non interchangeable entities with different fitness colonizing different niches) [5], but the discussion of this topic is beyond the scope of the present manuscript. Instead, here we focused on how a *Candidatus* proposal, which has been created to accommodate uncultured taxa to the framework used in taxonomy, can be improved by adding metagenomic data to describe the majority of taxa that lack conspicuous morphological and/or phenotypic characteristics and is resistant to isolation in pure culture. The benefits of applying this molecular ecology's approach for the improvement of *Candidatus* proposals had been already predicted more than five years ago [34]. Now, given the important technological improvements in NGS and the theoretical advances in the species definition for cultured taxa, the implementation of (meta)genomics in the *Candidatus* proposal seems to be timely.

High throughput sequencing has been applied to study complex microbial assemblages of environmental samples by whole-genome shotgun metagenomics. Contrary to the 16S rRNA gene sequencing approach, which is just based on a single gene of the genome to catalog diversity, metagenomics offers the possibility to describe diversity at the genome level. Given that available bioinformatic tools provide now the possibility to discern gene populations belonging to distinct microbial populations (by means of e.g. gene recruiting [14], or tetranucleotide signature similarities [25,42]), metagenomics can reveal the gene content of individual populations. Despite the facts that it is yet challenging to recover the complete genome sequence of each single microbial population, and the assembled sequences typically represent the average ("mosaic") genome of the populations rather than any cell actually present in the sample [12], the amount of information retrieved is frequently good enough as that used for taxonomic descriptions. In this regard, the new metagenomic technologies can improve the *Candidatus* proposal as a provisional status of uncultivated taxa.

By synthesizing the findings from all large shotgun metagenomic surveys of ocean, freshwater, engineered bioreactors, human gut, and soil environments performed to date, it was concluded that natural prokaryotic communities are predominantly composed of sequence-discrete populations [5]. In other words, the individual cells that belong to a single population within an environment show typically high sequence similarities among themselves, e.g., >95% ANI, [11]. In contrast, ANI values between such cells and those of (different) co-occurring populations in the same sample or habitat are typically lower than 85–90% (Fig. 1). Conspicuously, genomes (cells) showing values between 85% and 95% ANI to members of another discrete population were rare, and typically showed different abundances compared to the members of the same discrete population, indicating that they were ecologically differentiated since under the same conditions (i.e., present in the same sample) they differed in abundance [14]. The range of intra-population sequence diversity varied between ~1% and a maximum of about 5% genome-aggregate average sequence divergence, depending on the age of the population, with younger populations showing lower sequence diversity (i.e. the number of mutations accumulated is proportional to the time the population has been evolving within the given habitat). In almost all cases examined, the members of a population showed similar in situ abundances and small gene content differences (typically, less than 5% of the total genes in the genome differed), which contrasts with examples such as *Escherichia coli*, where two genomes may differ in up to 30–35% of their genes [5,14,35]. Furthermore, the (same) sequence-discrete population is typically detectable by metagenomics over time and space as long as similar environmental conditions prevail, i.e., the same ecological niche is available [14,28]. In other words, when populations are being dispersed between similar habitats, these populations are likely to be maintained, and habitats that are characterized by different physicochemical properties (i.e. niches) typically contain distinct populations of the same (i.e. higher gene content differences, but with high ANI values) or different species. The former is the case in connected habitats such as similar depths in the oceans, whereas distinct depths of the water column typically harbor distinct populations of the same species (i.e., different ecotypes) that are adapted to the physicochemical properties of the depth (e.g., light intensity, hydrostatic pressure) [13]. These findings suggest that the sequence-discrete populations are not ephemeral, clonal amplifications of one or a few cells, but long-lived entities that may encompass substantial (intra-population) genetic diversity. Furthermore, non-discrete populations (i.e. unclear genetic discontinuities between populations) are rare, at least for the abundant members of natural communities that can be robustly evaluated by metagenomics. When present, these non-discrete populations are typically associated with predictable environmental perturbations such as the mixing of distinct populations that are adapted to living in different depths in the ocean caused by deep-water upwelling events [13,14]. In summary, the sequence-discrete populations represent an important unit of microbial communities, and are easily tractable and distinguishable from other co-occurring populations [5].

The analysis of discrete metagenomic sequence bins representing populations permits accurate genealogic affiliation, and also the prediction of metabolic traits. With respect to the gene content (core genome vs. accessory genome) high quality metagenomics seems to be more informative than the analysis of single (cultured) strains, as the latter just informs about single individual cells whereas recruited metagenomes represent a gene-mosaic that reflects the pan-genome of the total population. In addition, ecological properties can likely be better detected by meta-omics in

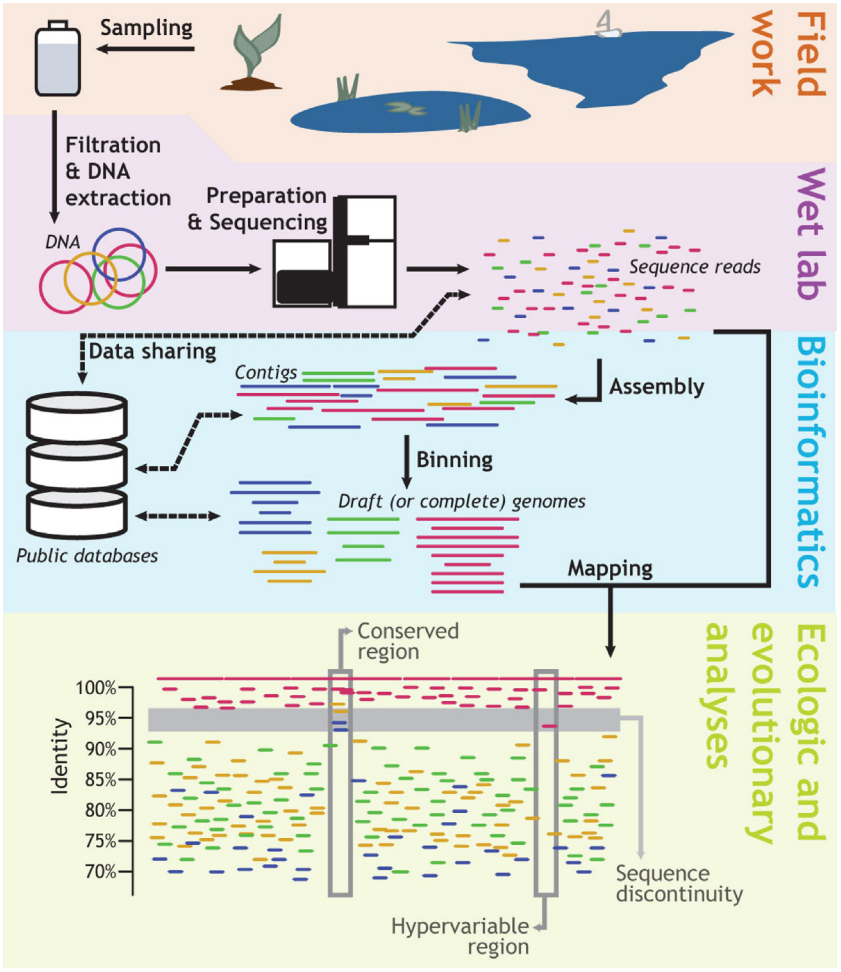

Fig. 1. Schematic of the metagenomic pipeline to identify sequence-discrete populations. Reads from metagenomic sequencing of microbial community DNA can be assembled into consensus contig sequences that represent sequence-discrete populations. Contigs originating from the same population can be identified based on their sequence characteristics and then grouped into nearly closed draft population genomes (binning). When the original reads of the metagenome are mapped against the contigs of a reference population (recruitment analysis; bottom), it becomes apparent that each population is sequence-discrete compared to its co-occurring populations. In this hypothetical example, reads originating from members of the reference population (red) evenly match the assembled contigs that represent the population with high nucleotide sequence identities (>97%). In contrast, reads from other populations (other colors) match the reference contigs at lower sequence identities, forming a sequence discontinuity ("gap") in the recruitment plot. Areas that deviate from this pattern are limited to highly conserved regions of the genome (e.g., rRNA operons), where reads from related but distinct populations are recruited due to their highly sequence identity to the reference sequences, or regions characterized by intrapopulation heterogeneity, which typically show lower coverage.

Source: Figure is modified with permission from [28].

well characterized environmental samples than on pure cultures in the laboratory. The availability of all this information is ideal for incorporation into the *Candidatus* proposal. To identify such populations and assess their genetic relatedness to other populations in the same or different habitat, it is recommended (Fig. 1) to perform a fragment recruitment plot of metagenomic reads against a reference genome sequence, complete or draft, of a representative of the population [13,14,32]. The reference sequence could be derived from metagenomic assemblies (e.g., long assembled contigs or recovered draft genomes) [19], single-cell techniques [40] or best from cultures (when available). A high-draft genome sequence is typically assemblable from a complex community based on short-read shotgun sequencing when the target population is covered at 20X or higher [20]. Further, several bioinformatics tools have recently become available that can bin together assembled contigs that belong to the same population to assist the recovery of the draft genome, e.g. [1,2]. Genealogy can be revealed by phylogenetic analysis of rRNA [6] or other universal housekeeping [15,41,48] genes encoded on the reference sequence used. With this respect, the single cell approaches, which can recover the partial genome of individual cells in a sample, may be advantageous because complete rRNA genes are usually recovered in such cases. In contrast, in metagenomic datasets the rRNA genes are infrequently assemblable due to their high sequence conservation, even between moderately related populations. However, caution needs to be exercised when applying single-cell techniques as they are still vulnerable to DNA contamination and chimera issues, (e.g. [47]). To facilitate such efforts, the Konstantinidis' research group at the Georgia Institute of Technology has also developed online implementations of the ANI and fragment recruitment tools (available through http://enve-omics.gatech.edu/).

By providing refined genealogic and gene content information, the (meta)genome-based approach offers, for the first time, an invaluable opportunity to genomically describe naturally occurring populations to such an extent which meets, and even surpasses, some of the requirements of a *Candidatus* proposal. Once a sequence-discrete microbial population is detected (Fig. 2), the recovered gene sequences can be used to reveal the phylogenetic affiliation, and in addition, the rRNA sequences (when available) can be used to design phylogenetic probes for fluorescence in situ

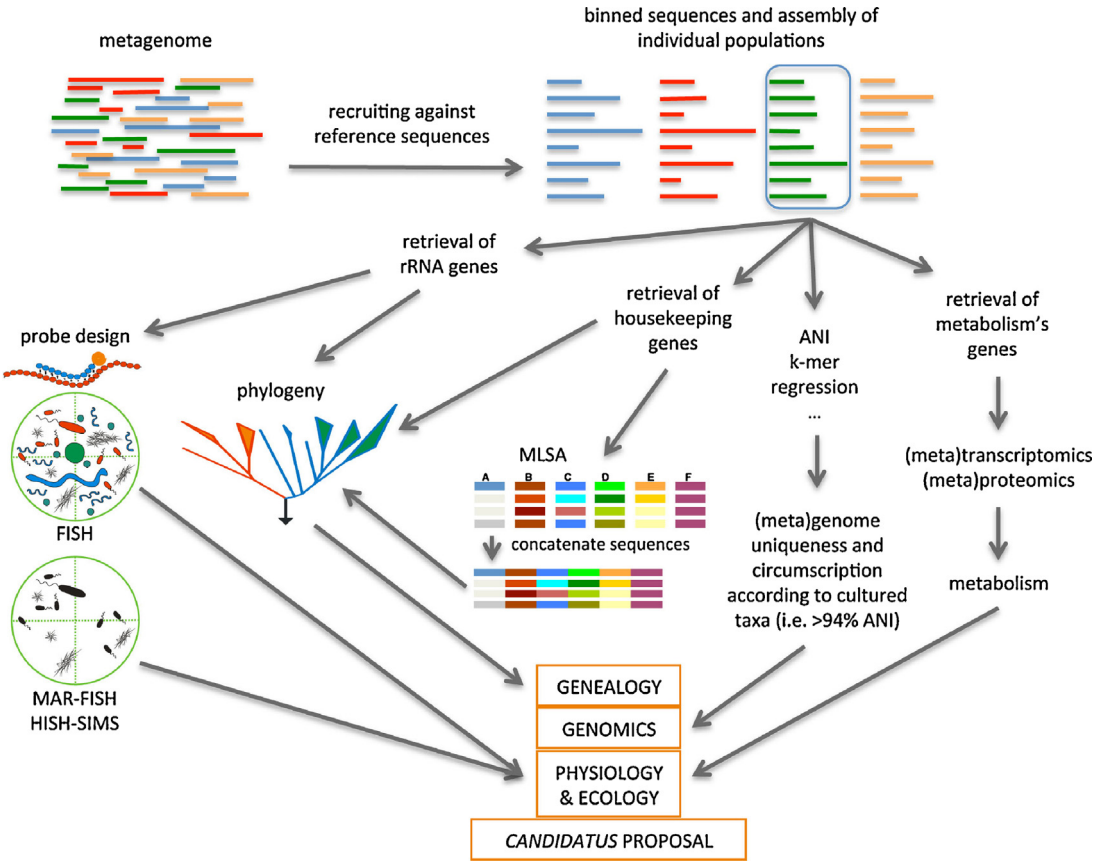

Fig. 2. Schematic of the procedure to achieve a *Candidatus* proposal based on environmentally identified individual (meta) populations. When an individual population is identified and a high quality genome assembly that represents it becomes available, genetic information can be retrieved for: (i) assessing morphology, abundances and other ecological properties, and designing probes for FISH based on (recovered) rRNA genes; (ii) performing phylogenetic reconstruction by means of either rRNA or other housekeeping genes (using the multilocus sequence analysis (MLSA) approach [39]); (iii) performing (meta)genome comparisons with analogous entries in databases and/or individual genomes to understand the uniqueness within the frame of proposed species or *Candidatus* species; and (iv) assessing physiology inferred from genetic (genomic and metatranscriptomic) data, as well as (when possible) additional information derived from approaches such as proteomics, MAR-FISH or HISH-SIMS (for metabolic properties). The final *Candidatus* proposal must show that the new population represents real individual cells in the environment and a monophyletic group of yet unclassified taxa, which genomic and metabolic properties guarantee their identification and placement within the classification schema.

hybridization application (FISH; [3]) in order to microscopically observe the morphology, abundances and distribution of the candidate taxa. The ANI, and the average amino acid identity (AAI) for more divergent populations [17], can be used to assess the genomic uniqueness of the newly identified populations within the existing classification schema.

However, the classification of a *Candidatus* taxon requires, in addition to genetic distinctiveness, also a description of phenotypic properties [21]. There are a series of molecular methods that can be used to reveal metabolic properties of uncultured organisms such as microautoradiography using radioactive isotopes combined with FISH (MAR-FISH; [31]), or stable isotopes combining secondary ion mass spectroscopy with FISH (HISH-SIMS; [23]). In addition, the (meta)genome approach can also provide valuable information about the potential metabolic and physiological properties of the population/taxon under study through the bioinformatics analysis of the recovered gene content, and hence aid cultivation efforts by providing hypotheses about nutrient requirements of the population, e.g. [44]. Therefore, it might be preferable to combine the metagenomic, and other omic-based approaches such as metatranscriptomics and proteomics, with single-cell and traditional 16S rRNA gene cloning approaches for a more complete description of the candidate taxon. As metatranscriptomics and metaproteomics techniques advance in the near future [45], it would be also possible to detect taxon-specific signatures at the gene expression or activity level (e.g., [46]), in addition to sequence-based distinctiveness (DNA metagenomics level), resulting in more accurate and comprehensive descriptions of putative taxa.

Authors' recommendations

To our opinion, the scientific community of taxonomists should be aware that allowing the classification of uncultured microorganisms by means of new molecular techniques represents probably the only realistic approach to describe the large, yet finite prokaryotic diversity, and would benefit communication between scientists from different disciplines (i.e. ecology, systematics and evolutionary microbiology) and the public. The finite extent of diversity makes possible to cover the taxonomic diversity based on the CTU approach, which will harmonize cultured and uncultured classification. For the time being, and given that only a *Candidatus* status is accepted by taxonomists, we suggest that modern proposals of candidate taxa should be preferably accompanied by a good description of their genome sequence (Fig. 2 and Table 2), at least at the high draft status or >95% coverage of the complete genome [4], retrieved either through metagenomic or single-cell approaches. The quality of such sequences should be verified following the recommendations of the Genome Standards Consortium [9,10], and by the (consistent) phylogenies of the universally conserved protein-coding genes encoded in the genome [1], and ribosomal RNA sequences when available. We also recommend

Table 2
Steps to follow when recognizing a discrete (meta)population to be classified as *Candidatus*:.

- **Assemble a high quality genome of the population** (i.e., >95% of the total genome recovered) based on either de-novo assembly, when sequencing effort is not limiting, or reference-guided assembly when a representative of the population is available, e.g., it has been retrieved by single-cell approaches.
  - Bin contigs to obtain a better representation of the population genome, using high values of regression of k-mers frequencies (e.g., tetra-, penta- or hexa-nucleotide) and/or contig coverage variation in multiple metagenomes (e.g., time series).
  - Recruit DNA contigs against reference genome(s) with nucleotide identity values >94% and re-assembly the retrieved contigs to obtain (whenever possible) longer contigs or scaffolds.
- **Assess intra-population genetic diversity** within the same metagenome/habitat and genetic relatedness to genomes of populations from other similar metagenomes/habitats.
  - Recruit DNA raw reads from the same metagenome (no cut-off on sequence identity employed) against the recovered (assembled) population genome sequence and check for the presence of areas of genetic discontinuities in the resulting recruitment plot, similar to that of Figure 1.
  - Do a recruitment plot with reads from other metagenome(s) of interest against the recovered population genome sequence or compare population genome sequences assembled from different metagenomes, similar to comparative isolate genome analysis (e.g. ANI calculation).
- **Identify the phylogenetic position and genomic uniqueness** of the putative recognized populations by:
  - Using (when available) complete or partial rRNA genes.
  - Using the available housekeeping sequences, reconstruct individual or concatenated (strongly recommended) phylogenetic trees.
  - Identify the closest relative taxa and evaluate the phylogenetic distance to recognize whether the (meta)population may correspond to an unclassified species.
  - Compare the ANI values with the closest available genomes of type strains.
- **Identify the shape, abundances and growth dynamics** of individual cells by:
  - Designing phylogenetic probes based on rRNA gene sequences and use them in in situ hybridizations such as FISH for morphology examination and quantification of cell abundance. Provide a picture showing the morphology of the identified cells belonging to the new putative *Candidatus* taxon. When possible, try to obtain several samples, e.g., time series or replicated experiments, to observe dynamics of the relative abundance of the population identified by the metagenomic approach.
  - Design probes or PCR primers against rRNA, housekeeping genes or unique ORFs to detect the presence and abundance of cells using semi-quantitative techniques such as qPCR, dot/slot-blots, and DGGE.
- **Identify putative physiologic data** by:
  - Bioinformatically inferring metabolic functions encoded in the recovered population genome, and, when possible, validating some of these inferences by experimental procedures such as those listed below:
  - Using metabolic tests combined with optical microscopy such as MAR-FISH, HISH-SIMS.
  - Using metatranscriptomics and metaproteomics combined with cell dynamics to infer putative metabolism.
- **Identify ecological traits** by:
  - Assessing population stability in the same environment through time series.
  - Assessing the occurrence in other environments.
- Once the identification of a (meta)population has been complemented with the understanding of (i) the genealogical position and uniqueness of their (meta)genomes, (ii) observation and quantification of individual cells in their original environment, (iii) inferred metabolism and validation of some of the bioinformatics predictions about metabolism with alternative molecular techniques, and (iv) assessment of the ecological traits of the uncultured population, you can **proceed with the *Candidatus* proposal** as recommended by Murray and Stackebrandt [22] and in the 2008 revision of the International Code of Nomenclature of Prokaryotes (to be published soon). **Take into account that**:
  - The new name of the *Candidatus* taxon should be written between quotes and preceded with the word *Candidatus* written in italics and the new genus (e.g. "*Candidatus* Salinibacter" when only describing a candidate genus) or the genus name with specific epithet (e.g. "*Candidatus* Magnetobacterium bavaricum" when describing a candidate species) should not be written in italics.
  - The most important properties (phylogenetic, genetic, physiologic and ecologic) should be recorded and any relevant probe or primer sequence should be provided.
  - The (meta)genome sequences should be deposited in the public repositories and the accession number should be provided.
  - The *Candidatus* name is provisional and has no standing in the nomenclature for prokaryotes.
  - Upon publication of the description, the new *Candidatus* names should be published in the validation or notification lists of the IJSEM in order to be part of the record list kept by the Judicial Commission of the ICSP.

generating a *Candidatus* (meta)genome database for ANI comparisons and record these as "type" material for further proposals. The recognition of putative *Candidate* taxa based on (meta)genome data must be accompanied by the description of their morphology, abundances and major metabolic and ecological properties to guarantee their identification (Fig. 2). As a minimum level of information related to metabolic properties, the bioinformatically-inferred metabolic functions encoded in the genome should be provided, following the standards in the genome announcement descriptions [9]. As the ANI may become the gold standard for cultured species circumscription, we suggest that the same thresholds (i.e. 94–96% ANI) are applied to initially circumscribe what an uncultured and environmentally occurring species may be.

Acknowledgements

The authors acknowledge Rudolf Amann for his comments for improvements of the manuscript. KTK's work related to the prokaryotic species issue has been supported by the US National Science Foundation (awards DEB 0516252 and DEB 1241046).

R.R.M. acknowledges the scientific support given by the Spanish Ministry of Economy with the project CGL2012-39627-C03-03, which is also supported with European Regional Development Fund (FEDER) funds, and the preparatory phase of Microbial Resource Research Infrastructure (MIRRI) funded by the EU (grant number 312251).

References

[1] Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K.L., Tyson, G.W., Nielsen, P.H. (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat. Biotechnol. 31, 533–538.

[2] Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Lahti, L., Loman, N.J., Andersson, A.F., Quince, C. (2014) Binning metagenomic contigs by coverage and composition. Nat. Methods 11, 1144–1146, http://dx.doi.org/10.1038/nmeth.3103 (Epub 2014 Sep 14).

[3] Amann, R.I., Ludwig, W., Schleifer, K.H. (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev. 59, 143–169.

[4] Branscomb, E., Predki, P. (2002) On the high value of low standards. J. Bacteriol. 184, 6406–6409 (discussion 6409).

[5] Caro-Quintero, A., Konstantinidis, K.T. (2012) Bacterial species may exist, metagenomics reveal. Environ. Microbiol. 14, 347–355.

[6] Cole, J., Konstantinidis, K.T., Farris, R.J., Tiedje, J.M. (2010) Microbial diversity and phylogeny: extending from rRNAs to genomes. In: Liu, W.-T., Jansson, J. (Eds.), Environmental Molecular Biology, Horizon Scientific Press, Norwich, UK, pp. 1–20.

[7] Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., Tiedje, J.M. (2014) Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucl. Acids Res. 42, D633–D642, http://dx.doi.org/10.1093/nar/gkt1244.

[8] Ereshefsky, M. (1998) Species pluralism and anti-realism. Phylos. Sci. 65, 103–120.

[9] Field, D., Amaral-Zettler, L., Cochrane, G., Cole, J.R., Dawyndt, P., Garrity, G.M., Gilbert, J., Glockner, F.O., Hirschman, L., Karsch-Mizrachi, I., Klenk, H.P., Knight, R., Kottmann, R., Kyrpides, N., Meyer, F., San Gil, I., Sansone, S.A., Schriml, L.M., Sterk, P., Tatusova, T., Ussery, D.W., White, O., Wooley, J. (2011) The genomic standards consortium. PLoS Biol. 9, e1001088.

[10] Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., Tatusova, T., Thomson, N., Allen, M.J., Angiuoli, S.V., Ashburner, M., Axelrod, N., Baldauf, S., Ballard, S., Boore, J., Cochrane, G., Cole, J., Dawyndt, P., De Vos, P., DePamphilis, C., Edwards, R., Faruque, N., Feldman, R., Gilbert, J., Gilna, P., Glockner, F.O., Goldstein, P., Guralnick, R., Haft, D., Hancock, D., Hermjakob, H., Hertz-Fowler, C., Hugenholtz, P., Joint, I., Kagan, L., Kane, M., Kennedy, J., Kowalchuk, G., Kottmann, R., Kolker, E., Kravitz, S., Kyrpides, N., Leebens-Mack, J., Lewis, S.E., Li, K., Lister, A.L., Lord, P., Maltsev, N., Markowitz, V., Martiny, J., Methe, B., Mizrachi, I., Moxon, R., Nelson, K., Parkhill, J., Proctor, L., White, O., Sansone, S.A., Spiers, A., Stevens, R., Swift, P., Taylor, C., Tateno, Y., Tett, A., Turner, S., Ussery, D., Vaughan, B., Ward, N., Whetzel, T., San Gil, I., Wilson, G., Wipat, A. (2008) The minimum information about a genome sequence (MIGS) specification. Nat. Biotechnol. 26, 541–547.

[11] Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., Tiedje, J.M. (2007) DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. Int. J. Syst. Evol. Microbiol. 57, 81–91.

[12] Hallam, S.J., Konstantinidis, K.T., Putnam, N., Schleper, C., Watanabe, Y., Sugahara, J., Preston, C., de la Torre, J., Richardson, P.M., DeLong, E.F. (2006) Genomic analysis of the uncultivated marine crenarchaeote Cenarchaeum symbiosum. Proc. Natl. Acad. Sci. U. S. A. 103, 18296–18301.

[13] Konstantinidis, K.T. (2011) Metagenomic insights into bacterial species. In: De Bruijn, (Ed.), Handbook of Molecular Microbial Ecology. I: Metagenomics and Complementary Approaches, John Wiley & Sons Inc., Hoboken, NJ, USA, pp. 89–98.

[14] Konstantinidis, K.T., DeLong, E.F. (2008) Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. ISME J. 2, 1052–1065.

[15] Konstantinidis, K.T., Ramette, A., Tiedje, J.M. (2006) Toward a more robust assessment of intraspecies diversity, using fewer genetic markers. Appl. Environ. Microbiol. 72, 7286–7293.

[16] Konstantinidis, K., Tiedje, J.M. (2005) Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. U. S. A. 102, 2567–2592.

[17] Konstantinidis, K.T., Tiedje, J.M. (2007) Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. Curr. Opin. Microbiol. 10, 504–509.

[18] Lapage, S.P., Sneath, P.H.A., Lessel, E.F., Skerman, V.B.D., Seeliger, H.P.R., Clark, W.A. 1991 International Code of Nomenclature of Bacteria (1990 Revision), American Society for Microbiology, Washington, DC.

[19] Luo, C., Rodriguez, R.L., Konstantinidis, K.T. (2013) A user's guide to quantitative and comparative analysis of metagenomic datasets. Methods Enzymol. 531, 525–547.

[20] Luo, C., Tsementzi, D., Kyrpides, N.C., Konstantinidis, K.T. (2012) Individual genome assembly from complex community short-read metagenomic datasets. ISME J. 6, 898–901.

[21] Murray, R.G.E., Schleifer, K.-H. (1994) Taxonomic notes: a proposal for recording the properties of putative procaryotes. Int. J. Syst. Bacteriol. 44, 174–176.

[22] Murray, R.G.E., Stackebrandt, E. (1995) Taxonomic note: implementation of the provisional status *Candidatus* for incompletely described procaryotes. Int. J. Syst. Bacteriol. 45, 186–187.

[23] Musat, N., Halm, H., Winterholler, B., Hoppe, P., Peduzzi, S., Hillion, F., Horreard, F., Amann, R., Jørgensen, B.B., Kuypers, M.M. (2008) A single-cell view on the ecophysiology of anaerobic phototrophic bacteria. Proc. Natl. Acad. Sci. U. S. A. 105, 17861–17866.

[24] Pedrós-Alió, C. (2006) Marine microbial diversity: can it be determined? Trends Microbiol. 14, 257–263.

[25] Quaiser, A., Zivanovic, Y., Moreira, D., López-García, P. (2011) Comparative metagenomics of bathypelagic plankton and bottom sediment from the Sea of Marmara. ISME J. 5, 285–304.

[26] Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucl. Acid Res. 41, D590–D596.

[27] Richter, M., Rosselló-Móra, R. (2009) Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl. Acad. Sci. U. S. A. 106, 19126–19131.

[28] Rodriguez, R.L.-M., Konstantinidis, K.T. (2014) Bypassing cultivation to identify bacterial species. Microbe 9, 111–118.

[29] Rosselló-Móra, R. (2012) Towards a taxonomy of *Bacteria* and *Archaea* based on interactive and cumulative data repositories. Environ. Microbiol. 14, 318–334.

[30] Rosselló-Móra, R., Amann, R. (2015) Past and future species definitions for *Bacteria* and *Archaea*. Syst. Appl. Microbiol (in press).

[31] Rosselló-Mora, R., Lee, N., Antón, J., Wagner, M. (2003) Substrate uptake in extremely halophilic microbial communities revealed by microautoradiography and fluorescence in situ hybridisation. Extremophiles 5, 409–413.

[32] Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., Wu, D., Eisen, J.A., Hoffman, J.M., Remington, K., Beeson, K., Tran, B., Smith, H., Baden-Tillson, H., Stewart, C., Thorpe, J., Freeman, J., Andrews-Pfannkoch, C., Venter, J.E., Li, K., Kravitz, S., Heidelberg, J.F., Utterback, T., Rogers, Y.H., Falcon, L.I., Souza, V., Bonilla-Rosso, G., Eguiarte, L.E., Karl, D.M., Sathyendranath, S., Platt, T., Bermingham, E., Gallardo, V., Tamayo-Castillo, G., Ferrari, M.R., Strausberg, R.L., Nealson, K., Friedman, R., Frazier, M., Venter, J.C. (2007) The Sorcerer II global ocean sampling expedition: Northwest Atlantic through eastern tropical Pacific. PLoS Biol. 5, e77.

[33] Salman, V., Amann, R., Girnth, A.-C., Polerecky, L., Bailey, J.V., Høgslund, S., Jessen, G., Pantoja, S., Schulz-Vogt, H.N. (2011) A single-cell sequencing approach to the classification of large, vacuolated sulfur bacteria. Syst. Appl. Microbiol. 34, 243–259.

[34] Schleifer, K.-H. (2009) Classification of *Bacteria* and *Archaea*: past, present and future. Syst. Appl. Microbiol. 32, 533–542.

[35] Shapiro, B.J., Friedman, J., Cordero, O.X., Preheim, S.P., Timberlake, S.C., Szabo, G., Polz, M.F., Alm, E.J. (2012) Population genomics of early events in the ecological differentiation of bacteria. Science 336, 48–51.

[36] Sneath, P.H.A., Brenner, D.J. (1992) Official nomenclature lists. ASM News 58, 175.

[37] Spring, S., Amann, R., Ludwig, W., Schleifer, K.-H., Van Gemerden, H., Petersen, N. (1993) Dominating role of an unusual magnetotactic bacterium in the microaerobic zone of a fresh water sediment. Appl. Environ. Microbiol. 59, 2397–2403.

[38] Stackebrandt, E., Ebers, J. (2006) Taxonomic parameters revisited: tarnished gold standards. Microbiol. Today 8, 6–9.

[39] Stackebrandt, E., Frederiksen, W., Garrity, G.M., Grimont, P.A.D., Kämpfer, P., Maiden, M.C.J., Nesme, X., Rosselló-Mora, R., Swings, J., Trüper, H.G., Vauterin, L., Ward, A., Whitman, W.B. (2002) Report of the Ad Hoc Committee for the re-evaluation of the species definition in bacteriology. Int. J. Syst. Evol. Microbiol. 52, 1043–1047.

[40] Stepanauskas, R. (2012) Single cell genomics: an individual look at microbes. Curr. Opin. Microbiol. 15, 613–620.

[41] Sunagawa, S., Mende, D.R., Zeller, G., Izquierdo-Carrasco, F., Berger, S.A., Kultima, J.R., Coelho, L.P., Arumugam, M., Tap, J., Nielsen, H.B., Rasmussen, S., Brunak, S., Pedersen, O., Guarner, F., de Vos, W.M., Wang, J., Li, J., Dore, J., Ehrlich, S.D., Stamatakis, A., Bork, P. (2013) Metagenomic species profiling using universal phylogenetic marker genes. Nat. Methods 10, 1196–1199.

[42] Teeling, H., Meyerdierks, A., Bauer, M., Amann, R., Glöckner, F.O. (2004) Application of tetranucleotide frequencies for the assignment of genomic fragments. Environ. Microbiol. 6, 938–947.

[43] Tindall, B.J., Kämpfer, P., Euzéby, J.P., Oren, A. (2006) Valid publication of names of prokaryotes according to the rules of nomenclature: past history and current practice. Int. J. Syst. Evol. Microbiol. 56, 2715–2720.

[44] Tyson, G.W., Lo, I., Baker, B.J., Allen, E.E., Hugenholtz, P., Banfield, J.F. (2005) Genome-directed isolation of the key nitrogen fixer *Leptospirillum ferrodiazotrophum* sp nov from an acidophilic microbial community. Appl. Environ. Microbiol. 71, 6319–6324.

[45] VerBerkmoes, N.C., Denef, V.J., Hettich, R.L., Banfield, J.F. (2009) Systems biology: functional analysis of natural microbial consortia using community proteomics. Nat. Rev. Microbiol. 7, 196–205.

[46] Vital, M., Chai, B., Ostman, B., Cole, J., Konstantinidis, K.T., Tiedje, J.M. (2014) Gene expression analysis of *E. coli* strains provides insights into the role of gene regulation in diversification. ISME J., http://dx.doi.org/10.1038/ismej.2014.204 (Epub ahead of print).

[47] Woyke, T., Sczyrba, A., Lee, J., Rinke, C., Tighe, D., Clingenpeel, S., Malmstrom, R., Stepanauskas, R., Cheng, J.F. (2011) Decontamination of MDA reagents for single cell whole genome amplification. PLoS ONE 6, e26161.

[48] Wu, M., Eisen, J.A. (2008) A simple, fast, and accurate method of phylogenomic inference. Genome Biol. 9, R151.
[49] Yarza, P., Ludwig, W., Euzéby, J., Amann, R., Schleifer, K.-H., Glöckner, F.O., Rosselló-Móra, R. (2010) Update of the all-species living-tree project based on 16S and 23S rRNA sequence analyses. Syst. Appl. Microbiol. 33, 291–299.
[50] Yarza, P., Spröer, C., Swiderski, J., Mrotzek, N., Spring, S., Tindall, B.J., Gronow, S., Pukall, R., Klenk, H.-P., Lang, E., Verbarg, S., Crouch, A., Lilburn, T., Beck, B., Unosson, C., Cardew, S., Moore, E.R.B., Gomila, M., Nakagawa, Y., Janssens, D., De Vos, P., Peiren, J., Suttels, T., Clermont, D., Bizet Ch Sakamoto, M., Iida, T., Kudo, T., Kosako, Y., Oshida, Y., Ohkuma, M., Arahal, D.R., Spieck, E., Pommerening Roeser, A., Figge, M., Park, D., Buchanan, P., Cifuentes, A., Munoz, R., Euzéby, J., Schleifer, K.-H., Ludwig, W., Amann, R., Glöckner, F.O., Rosselló-Móra, R. (2013) Sequencing orphan species initiative (SOS): filling the gaps in the 16S rRNA gene sequence database for all species with validly published names. Syst. Appl. Microbiol. 36, 69–73.
[51] Yarza, P., Yilmaz, P., Prüße, E., Glöckner, F.O., Ludwig, W., Schleifer, K-H., Whitman, W.B., Euzéby, J., Amann, R., Rosselló-Móra, R. (2014) Uniting the classification of cultured and uncultured Bacteria and Archaea by means of 16S rRNA gene sequences. Nat. Rev. Microbiol. 12, 635–645.