Insights into the phylogeny and coding potential of microbial dark matter

Christian Rinke[1], Patrick Schwientek[1], Alexander Sczyrba[1,2,] Natalia N. Ivanova[1], Iain J. Anderson[1Δ], Jan-Fang Cheng[1], Aaron Darling[3,4], Stephanie Malfatti[1], Brandon K. Swan[5], Esther A. Gies[6], Jeremy A. Dodsworth[7], Brian P. Hedlund[7],George Tsiamis[8], Stefan M. Sievert[9], Wen-Tso Liu[10], Jonathan A. Eisen[3], Steven J. Hallam[6], Nikos C. Kyrpides[1], Ramunas Stepanauskas[5], Edward M. Rubin[1], Philip Hugenholtz[11] & Tanja Woyke[1]

[1]DOEJointGenomeInstitute, Walnut Creek, California 94598, USA.

[2]Center forBiotechnology,Bielefeld University,33602 Bielefeld, Germany.

[3]Department of Evolution and Ecology, University of California

Davis, Davis, California 95616, USA.

[4]ithree institute, University of Technology Sydney, Ultimo NSW 2007, Australia.

[5]Bigelow Laboratory for Ocean Sciences, East Boothbay, Maine 04544-0380, USA.

[6]Department ofMicrobiology and Immunology and Graduate Program in Bioinformatics, University of British Columbia, Vancouver, British Columbia, V6T 1Z3 Canada.

[7]School of Life Sciences, University

of Nevada, Las Vegas, Nevada 89154-4004, USA.

[8]Department of Environmental and Natural Resources Management, University of Patras, Agrinio, T.K. 30100, Greece.

[9]Biology Department, Woods Hole

Oceanographic Institution, Woods Hole, Massachusetts 02543, USA.

[10]Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61802, USA.

[11]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences and Institute for Molecular Bioscience, The University of Queensland, St. Lucia QLD 4072, Australia.

[Δ] Deceased

**To whom correspondence should be addressed*: T.W. (twoyke@lbl.gov) and P.H. (p.hugenholtz@uq.edu.au).

Insights into the phylogeny and coding potential of microbial dark matter

October 16, 2013

ACKNOWLEDGMENTS:

We thank the DOE JGI production sequencing, IMG and GOLD teams for their support; J. Lee and E. Ng for experimental assistance; H.-P. Klenk and D. Gleim for providing a DSMZ inventory database dump and I. Letunic´ for his knowledge and support to make iTOL work for this project. We are very grateful to B. Schink for invaluable etymological advice. The work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231. We also thank the CeBiTec Bioinformatics Resource Facility, which is supported by BMBF grant 031A190. B.P.H. and J.A.D. were supported by the NASA Exobiology grant EXO-NNX11AR78G and NSFOISE 096842 and B.P.H. by a generous contribution from G. Fullmer through the UNLV Foundation. S.M.S was supported by NSF grants OCE-0452333 and OCE-1136727, and the WHOI's Andrew W. Mellon Fund for Innovative Research; and S.J.H. by the Canadian Foundation for Innovation, the British Columbia Knowledge Development Fund, the National Sciences and Engineering Research Council (NSERC) of Canada and the TULA foundation funded Centre for Microbial Diversity and Evolution (CMDE), and the Canadian Institute for Advanced Research (CIFAR). R.S. was supported by NSF grants DEB-841933, EF-826924, OCE-1232982, OCE-821374 and OCE-1136488, and the Deep Life I grant by the Alfred P. Sloan Foundation. P.H. was supported by a Discovery Outstanding Researcher Award (DORA) from the Australian Research Council, grant DP120103498.

DISCLAIMER:

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein

Insights into the phylogeny and coding potential of microbial dark matter

to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

ARTICLE

OPEN
doi:10.1038/nature12352

Insights into the phylogeny and coding potential of microbial dark matter

Christian Rinke[1], Patrick Schwientek[1], Alexander Sczyrba[1,2], Natalia N. Ivanova[1], Iain J. Anderson[1]‡, Jan-Fang Cheng[1], Aaron Darling[3,4], Stephanie Malfatti[1], Brandon K. Swan[5], Esther A. Gies[6], Jeremy A. Dodsworth[7], Brian P. Hedlund[7], George Tsiamis[8], Stefan M. Sievert[9], Wen-Tso Liu[10], Jonathan A. Eisen[3], Steven J. Hallam[6], Nikos C. Kyrpides[1], Ramunas Stepanauskas[5], Edward M. Rubin[1], Philip Hugenholtz[11] & Tanja Woyke[1]

Genome sequencing enhances our understanding of the biological world by providing blueprints for the evolutionary and functional diversity that shapes the biosphere. However, microbial genomes that are currently available are of limited phylogenetic breadth, owing to our historical inability to cultivate most microorganisms in the laboratory. We apply single-cell genomics to target and sequence 201 uncultivated archaeal and bacterial cells from nine diverse habitats belonging to 29 major mostly uncharted branches of the tree of life, so-called 'microbial dark matter'. With this additional genomic information, we are able to resolve many intra- and inter-phylum-level relationships and to propose two new superphyla. We uncover unexpected metabolic features that extend our understanding of biology and challenge established boundaries between the three domains of life. These include a novel amino acid use for the opal stop codon, an archaeal-type purine synthesis in Bacteria and complete sigma factors in Archaea similar to those in Bacteria. The single-cell genomes also served to phylogenetically anchor up to 20% of metagenomic reads in some habitats, facilitating organism-level interpretation of ecosystem function. This study greatly expands the genomic representation of the tree of life and provides a systematic step towards a better understanding of biological evolution on our planet.

Microorganisms are the most diverse and abundant cellular life forms on Earth, occupying every possible metabolic niche. The large majority of these organisms have not been obtained in pure culture and we have only recently become aware of their presence mainly through cultivation-independent molecular surveys based on conserved marker genes (chiefly small subunit ribosomal RNA; SSU rRNA) or through shotgun sequencing (metagenomics)[1,2]. As an increasing number of environments are deeply sequenced using next-generation technologies, diversity estimates for Bacteria and Archaea continue to rise, with the number of microbial 'species' predicted to reach well into the millions[3]. According to SSU rRNA-based phylogeny, these fall into at least 60 major lines of descent (phyla or divisions) within the bacterial and archaeal domains[4], of which half have no cultivated representatives (so-called 'candidate' phyla). This biased representation is even more fundamentally skewed when considering that more than 88% of all microbial isolates belong to only four bacterial phyla, the Proteobacteria, Firmicutes, Actinobacteria and Bacteroidetes (Supplementary Fig. 1a). Genome sequencing of microbial isolates naturally reflects this cultivation bias (Supplementary Fig. 1b). Recently, a systematic effort, the Genomic Encyclopaedia of Bacteria and Archaea (GEBA) Project[5], has been initiated to maximize coverage of the diversity captured in microbial isolates by phylogenetically targeted genome sequencing. However, GEBA does not address candidate phyla that represent a major unexplored portion of microbial diversity, and have been referred to as microbial dark matter (MDM)[6].

Metagenomics can obtain genome sequences from uncultivated microorganisms through direct sequencing of DNA from the environment[7]. In some instances, draft or even complete genomes of candidate phyla have been recovered solely from metagenomic data (Supplementary Table 1). A complementary cultivation-independent approach for obtaining genomes from candidate phyla is single-cell genomics; the amplification and sequencing of DNA from single cells obtained directly from environmental samples[8]. This approach can be used for targeted recovery of genomes and has been applied to members of several candidate phyla (Supplementary Table 1). In particular, natural populations that have a high degree of genomic heterogeneity will be more accessible through single-cell genomics than through metagenomics as co-assembly of multiple strains is avoided. Despite these advances in obtaining genomic representation of MDM, no systematic effort has been made to obtain genomes from uncultivated candidate phyla using single-cell whole genome amplification approaches.

Here, we present GEBA-MDM, the natural extension of the Genomic Encyclopaedia into uncultivated diversity by applying single-cell genomics to recover draft genomes from over 200 cells representing more than 20 major uncultivated archaeal and bacterial lineages. Genome-based phylogenetic analysis confirms the validity of rRNA-defined candidate phyla as monophyletic groups and resolves a number of associations among phyla not apparent by single gene analysis. We discovered several unexpected features, including archaeal sigma factors and stop codon reassignments that challenge established views of the microbial world. Furthermore, we show that single-cell genome references substantially improve the phylogenetic anchoring of about 340 million previously incorrectly or under-classified metagenomic reads.

[1]DOE Joint Genome Institute, Walnut Creek, California 94598, USA. [2]Center for Biotechnology, Bielefeld University, 33602 Bielefeld, Germany. [3]Department of Evolution and Ecology, University of California Davis, Davis, California 95616, USA. [4]ithree institute, University of Technology Sydney, Ultimo NSW 2007, Australia. [5]Bigelow Laboratory for Ocean Sciences, East Boothbay, Maine 04544-0380, USA. [6]Department of Microbiology and Immunology and Graduate Program in Bioinformatics, University of British Columbia, Vancouver, British Columbia, V6T 1Z3 Canada. [7]School of Life Sciences, University of Nevada, Las Vegas, Nevada 89154-4004, USA. [8]Department of Environmental and Natural Resources Management, University of Patras, Agrinio, T.K. 30100, Greece. [9]Biology Department, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA. [10]Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61802, USA. [11]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences and Institute for Molecular Bioscience, The University of Queensland, St. Lucia QLD 4072, Australia.
‡Deceased.

©2013 Macmillan Publishers Limited. All rights reserved

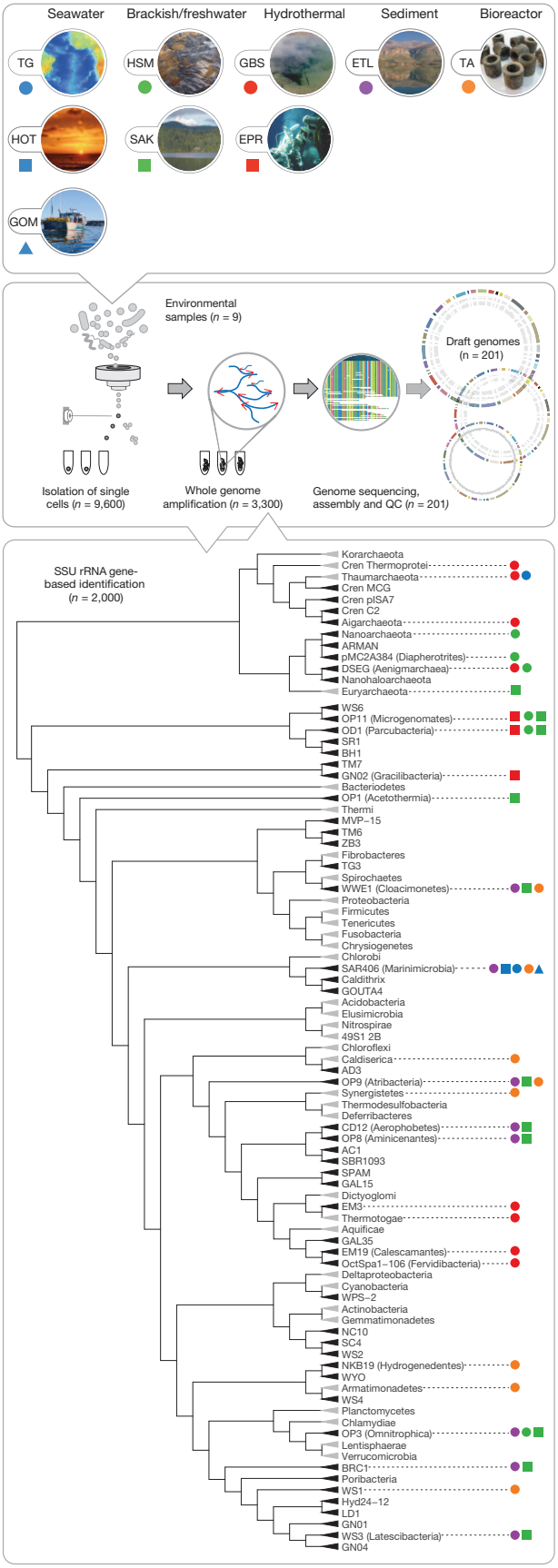

Figure 1 | Sampling sites and single-cell sequencing workflow. Upper panel, nine global sampling sites grouped into ocean samples (blue), fresh and brackish water samples (green), hydrothermal sites (red), sediment samples (magenta), and bioreactor samples (orange symbol). EPR, East Pacific Rise; ETL, Etoliko Lagoon; GBS, Great Boiling Spring; GOM, Gulf of Maine; HOT, Hawaii Ocean Time-series Project; HSM, Homestake Mine; SAK, Sakinaw Lake; TA, terephthalate degrading reactor; TG, tropical gyre in the South Atlantic. Middle panel, environmental samples were processed using a fluorescence-activated cell sorter allowing the isolation of 9,600 single cells. Each cell was lysed and the genome amplified yielding 3,300 successful amplifications. Resulting SAGs were screened by SSU rRNA gene PCR and sequencing to resolve taxonomic identities. SAGs belonging to major novel lineages were selected for genome sequencing and assembly resulting in 201 draft genomes. QC, quality control. Lower panel, cladogram showing the taxonomy of the SSU rRNA gene sequences, grouped into phyla. Candidate phyla are highlighted in black, and known phyla (according to the list of 'Prokaryote Names with Standing in Nomenclature' at http://www.bacterio.net/) are shown in light grey. For each phylum for which we retrieved one or more single-cell genomes the sampling sites are indicated according to the symbols in the upper panel. Note that marker gene phylogeny suggests that SAG JGI0000068-E11clusters within the PER group, a sister lineage to Gracilibacteria (Supplementary Fig. 3). This finding is not supported by the SSU rRNA gene phylogeny and will need further evaluation as more genome and SSU rRNA gene sequences become available.

Single-cell genomics at scale

We began by screening numerous physicochemically and geographically diverse environmental samples using SSU rRNA community profiling to identify habitats enriched in candidate phyla, and we targeted nine for in-depth single-cell analysis (Fig. 1, top panel, and Supplementary Fig. 2). Cells representing novel lineages were identified using high-throughput single-cell flow sorting, whole-genome amplification and SSU rRNA screening of single amplified genomes (SAGs; Fig. 1, middle panel; see Methods). A total of 201 SAGs representing 21 and 8 highly under-represented major bacterial and archaeal lineages were selected for whole genome sequencing (Fig. 1, bottom panel).

To improve assemblies, SAG sequence data was digitally normalized to reduce over-represented regions caused by amplification bias[9]. The fidelity of the resulting assemblies was validated using tetra-nucleotide frequency, BLAST (Basic Local Alignment Search Tool) and single copy marker gene analyses (Supplementary Methods and Supplementary Fig. 4). Draft SAGs ranged in size from 148 kilobase pairs (kb) to 2.4 Mb comprising an average of 59 major contigs per assembly (Supplementary Fig. 5a and Supplementary Table 2). Genome completeness was estimated to range from less than 10% to more than 90% (mean 40%) based on the presence or absence of 139 bacterial and 162 archaeal conserved marker genes (Supplementary Fig. 5a). Combining reads of single cells belonging to the same population, that is, with an average nucleotide identity of ≥97% (ref. 10) (see Methods), improved assemblies and produced seven population genomes with an estimated completeness of over 90% (Supplementary Fig. 6 and Supplementary Fig. 5a, b).

Genome-based phylogenetic inference

SSU rRNA trees are known to be sound predictors of phylogenetic novelty[5,11] despite the blurring of vertical descent by lateral gene transfer[12]. However, concatenated alignments of multiple universally distributed single copy marker genes are generally considered to provide greater phylogenetic resolution than any individual gene for estimating a species tree[13]. We constructed bootstrapped maximum likelihood trees based on a concatenation of up to 38 commonly used conserved marker genes[5,14] (Supplementary Methods and Supplementary Table 3) with 15 taxa configurations[15] (Supplementary Table 4). Substitution models were selected to address known issues, including long branch attraction[16] (discussed further in Supplementary Information). Congruency of the individual marker gene topologies to each other was independently assessed confirming the selection of these gene

©2013 Macmillan Publishers Limited. All rights reserved

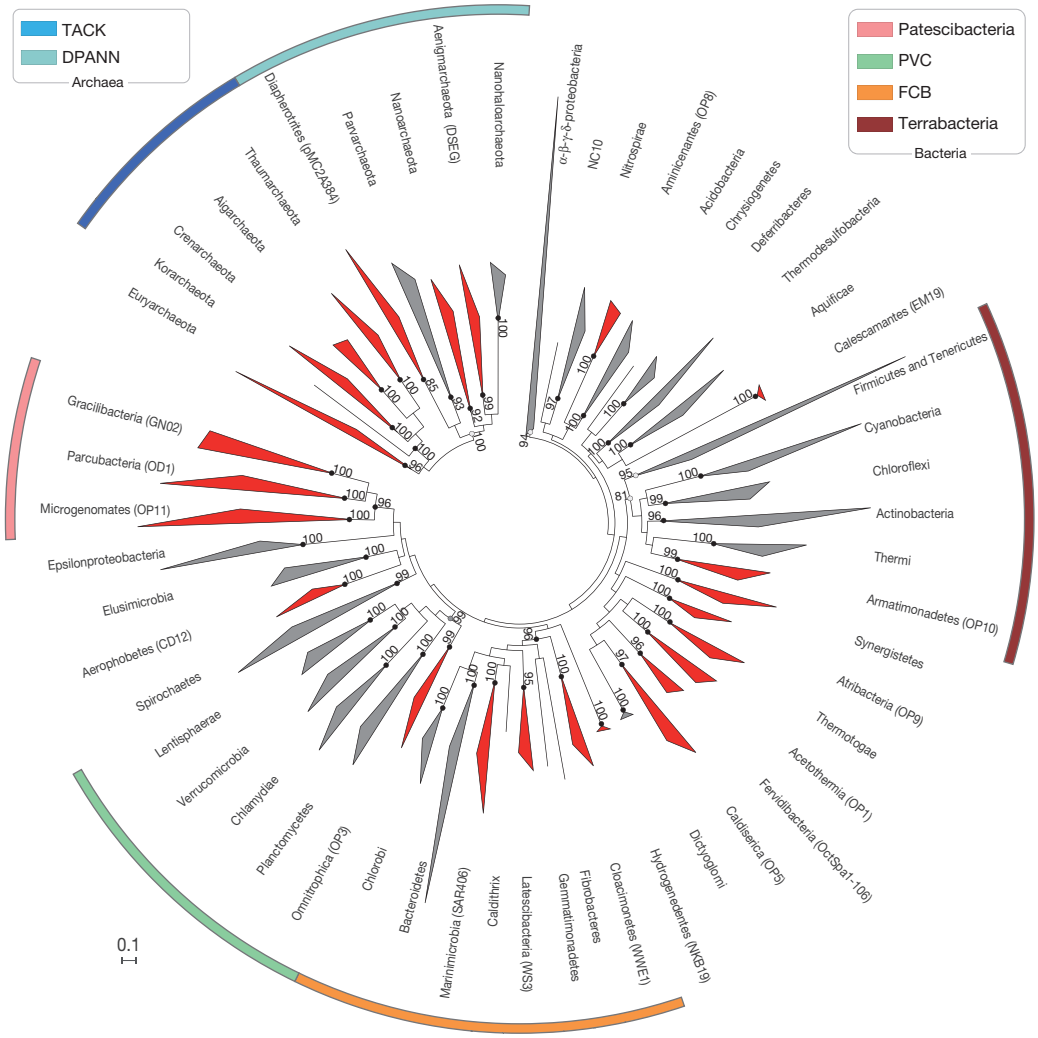

Figure 2 | Maximum-likelihood phylogenetic inference of Archaea and Bacteria. The phylogenetic trees are based on up to 38 marker genes and sequences are collapsed at the phylum level occluding subgroups such as the Geoarchaeota which clusters within the Crenarchaeota. Phyla containing SAGs from this study are highlighted in red. Superphyla (TACK, DPANN, Terrabacteria, FCB, PVC and Patescibacteria) are highlighted with colour ranges. The phylogenetic robustness (monophyly score) of phyla and superphyla is indicated by a small circle on the node: black circle (node was resolved in 100% of all tree calculations); grey circle (resolved in ≥90% of all calculations); light-grey circle (resolved in ≥50% of all calculations). Average bootstrap support values are provided for each phylum and superphylum when resolved. The underlying phylogenetic inference configurations as well as detailed branch support values and monophyly scores are provided in Supplementary Table 3. The two domain trees were independently calculated and are unrooted and the scale bar represents 10% estimated sequence divergence for both trees.

families for genome tree reconstruction (Supplementary Fig. 7). All candidate phyla with three or more SAG representatives were resolved as monophyletic groups consistent with their rRNA delineations (Fig. 2 and Supplementary Fig. 8). These are the first substantive genomic data for candidate bacterial phyla SAR406 (Marine Group A)[17], OP3, OP8 (ref. 18), WS1, WS3 (ref. 19), BRC1, CD12, EM19, EM3, NKB19, and Oct-Spa1-106 (ref. 20), as well as for several highly divergent archaeal groups related to the Nanoarchaeota (Fig. 2). We propose names for candidate phyla with two or more representatives based on their inferred physiology and distinguishing properties (Supplementary Table 5, see below).

Owing to the greater phylogenetic resolution afforded by the concatenated gene data sets, compared to rRNA phylogeny, we were able to identify a number of robust associations among phyla. These include the well-recognized Planctomycetes–Verrucomicrobia–Chlamydiae (PVC) superphylum that, based on rRNA analysis, was proposed to also include candidate phylum Omnitrophica (OP3) and the phylum Lentisphaerae[21]. Genome-based analysis confirms this grouping (Fig. 2) and we found a suggested PVC signature gene[22] in an Omnitrophica genome (Supplementary Information). The Fibrobacteres–Chlorobi–Bacteroidetes (FCB) superphylum[23] was robustly resolved together with Marinimicrobia (SAR406), Latescibacteria (WS3), Cloacimonetes (WWE1), Gemmatimonadetes[24] and Caldithrix[25]. Comparative genomics revealed that a conserved carboxy-terminal domain of extracellular proteinases (TIGR04183) is found exclusively (but not comprehensively) in members of the FCB superphylum. This includes the original phyla Fibrobacteres, Chlorobi, Bacteroidetes, as well as the candidate phyla Cloacimonetes, Marinimicrobia, Latescibacteria and the *Caldithrix* genome (Supplementary Information).

The Terrabacteria, proposed to comprise the 'terrestrial' bacterial phyla Actinobacteria, Cyanobacteria, Thermi (Deinococcus-Thermus), Chloroflexi and Firmicutes[26], was resolved in our analysis with the additional membership of Armatimonadetes (former candidate phylum OP10)[27] (Fig. 2). Perhaps more compelling than the assertion of ancient adaptations to life on land unifying the Terrabacteria[26] are commonalities in cell envelope architecture. This superphylum comprises monoderm (single membrane) and atypical monoderm lineages[28]. We assessed the additional proposed Terrabacteria phyla for genes most characteristic of monoderms and diderms[29] and confirmed that all had monoderm-like or atypical gene complements (Supplementary Fig. 9).

The phylogenetic placement of the Cloacimonetes (WWE1 clade) has been inconclusive based on rRNA comparative analysis. It was

©2013 Macmillan Publishers Limited. All rights reserved

originally proposed as a candidate phylum[30] and more recently as a class within the Spirochaetes phylum[28]. Our analysis, which substantially expands the genomic representation of this group, finds no support for a specific affiliation with the Spirochaetes (Fig. 2). It was suggested, based on a smaller data set, that the Acidobacteria reproducibly cluster with the Deltaproteobacteria[14] but this is not supported by our analyses. Instead, Acidobacteria reproducibly affiliate with the Aminacenantes (OP8) (Fig. 2). Candidate phylum OP11, as originally proposed[26], has not been resolved consistently as a monophyletic group leading to the proposal for subdivision into multiple phyla, including OP11 (former subdivisions 1 to 3 only), OD1 (former OP11 subdivision 5) and SR1 (ref. 31). Here we found that Microgenomates (OP11) and Parcubacteria (OD1) genomes were resolved reproducibly as a monophyletic group based on concatenated marker gene analysis together with Gracilibacteria (GN02)[32]. To recognize this affiliation, we propose the superphylum name 'Patescibacteria' (patesco (Latin), meaning bare) (Fig. 2), reflecting the reduced metabolic capacities of these lineages[33]. We found support for a specific association between the Patescibacteria and Terrabacteria using a larger bacteria-specific marker gene set (Supplementary Fig. 10). This association is consistent with a monoderm-like gene complement in the Patescibacteria (Supplementary Fig. 9) but will need to be verified when additional genomes belonging to these lineages are available.

Based on phylogenetic analysis of our archaeal single-cell genomes and several recently described genome-sequenced lineages of very small cells, such as *Candidatus* Parvarchaeum, *Candidatus* Micrarchaeum[34], *Candidatus* Nanosalina, *Candidatus* Nanosalinarum[35], we propose the following phyla; Diapherotrites (pMC2A384)[36], Parvarchaeota, Aenigmarchaeota (DSEG)[37] and Nanohaloarchaeota (Fig. 2 and Supplementary Table 5). The Nanohaloarchaeota include the recently proposed class Nanohaloarchaea that was incorrectly placed within the Euryarchaeota owing to inadequate outgroup representation[35]. We predict that small cell and genome size are unifying features of these phyla and in Archaea-only trees these lineages, together with the Nanoarchaeota, form a monophyletic superphylum for which we propose the identifier, DPANN (Fig. 2 and Supplementary Text). Our expanded genomic representation and analysis of the archaeal domain also supports the proposal for the TACK superphylum[38], but is not consistent with the eocyte hypothesis[39], which places the Eukaryota within the archaeal domain, recently reinvestigated using a 36-genome data set[40] (Supplementary Fig. 11). As more genomes and improved phylogenetic inference methods come to hand, our proposed lineage delineations can be further evaluated.

Functional diversity and novel findings

The numerous strategies that cultivated microorganisms use to obtain energy and nutrients suggest that many metabolic surprises remain to be discovered in the uncultivated microbial majority. Here we provide a first glimpse into the potential functional diversity of many of the investigated candidate phyla and novel lineages. The majority of bacterial and several archaeal single-cell genomes in our study possess a large array of genes for the degradation of amino acids and sugars (providing the basis for some candidate names for phyla; Supplementary Table 5), pointing to a heterotrophic lifestyle (Supplementary Fig. 12). We found evidence for an electron transport chain, and thus the ability to perform a more complete set of cellular respiration processes, in most bacterial SAGs with the notable exception of members of the Parcubacteria (OD1), Microgenomates (OP11), Gracilibacteria (GN02) and Latescibacteria (WS3). Genes necessary for carbon fixation were found in a wide range of archaeal SAGs (Wood–Ljungdahl pathway, adenosine nucleotide degradation pathway) with a more limited distribution in the bacterial SAGs (Supplementary Fig. 12). Hydrogen metabolism is widespread amongst the novel lineages, and two SAGs (belonging to Caldiserica and Aigarchaeota) have genes for sulphur utilization (Supplementary Fig. 12 and Supplementary Table 6).

A novel recoding of the opal stop codon UGA for glycine was identified in members of the Gracilibacteria (Fig. 3 and Supplementary Fig. 13a). The same recoding was found and biochemically validated in candidate phylum SR1 very recently[41], suggesting that this codon reassignment may be phylogenetically widespread in uncharacterized lineages. This expands the known alternative coding for UGA, which has previously been reported for selenocysteine[42] and tryptophan[43,44]. The very low guanine–cytosine content of the Gracilibacteria single-cell genomes (<24%) may have driven the recoding of UGA to a lower guanine–cytosine glycine codon alternative (UGA versus GGN) particularly as glycine is the third most commonly used amino acid (>7% average abundance per genome; Supplementary Fig. 13b).

Purine biosynthesis is highly conserved in the Bacteria and Archaea in terms of the penultimate step in the pathway responsible for ribonucleate formylation[45]. All bacteria sequenced so far use the PurH1 enzyme for this step, whereas the majority of Archaea use the PurP enzyme. However, members of the bacterial superphylum Patescibacteria lack the *purH1* gene and instead have an euryarchaeal *purP*-like gene (Fig. 3

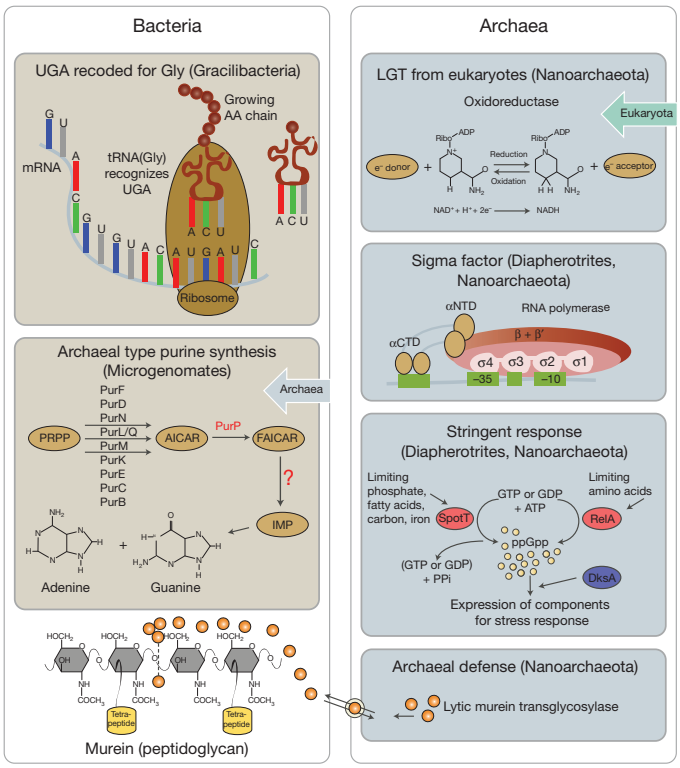

Figure 3 | Novel metabolic features found in the SAG data set. Left, features found in Bacteria: in a subgroup of the Gracilibacteria (GN02), the opal stop codon UGA codes for glycine and these genomes encode a transfer RNA (tRNA) for UGA. Two lineages of Microgenomates (OP11) bacteria use the archaeal pathway (PurH1 enzyme) for purine (adenine, guanine) biosynthesis, inferred to have been acquired by lateral gene transfer (LGT) from Euryarchaeota. AICAR, aminoimidazole carboxamide ribonucleotide; ATP, adenosine tri-phosphate; FAICAR, formyl aminoimidazole carboxamide ribonucleotide; IMP, inosine monophosphate; mRNA, messenger RNA; PRPP, phosphoribosyl pyrophosphate; PurH, bifunctional purine biosynthesis protein PurH. Right, features found in Archaea. A Nanoarchaeota genome encodes an oxidoreductase most closely related to the soil-living amoebae (slime mould) representing a lateral gene transfer from a eukaryote to an archaeon. Two members of the Diapherotrites (pMC2A384) and one representative of the Nanoarchaeota encode complete bacteria-like sigma factors (∂70). The bacterial stringent response based on deployment of signalling molecules (ppGpp) was identified in a member of the Diapherotrites and the Nanoarchaeota. A bacterial-like lytic murein transglycosylase was found in two members of the Nanoarchaeota. αCTD, α-subunit C-terminal domain; αNTD, α-subunit N-terminal domain; ADP, adenosine di-phosphate; GDP, guanosine di-phosphate; GTP, guanosine tri-phosphate.

©2013 Macmillan Publishers Limited. All rights reserved

and Supplementary Table 7) as a result of an ancient lateral transfer of most of the purine biosynthesis operon from a Thermococci-like donor to the ancestor of the Patescibacteria (Supplementary Fig. 14).

The DPANN superphylum contains a number of metabolic novelties pointing to a capacity for co-opting foreign genetic elements. A Nanoarchaeota genome encodes an oxidoreductase most closely related to the slime mould *Dictyostelium discoideum* and sits within the eukaryal evolutionary radiation for this gene (Supplementary Fig. 15). To our knowledge, this is the first instance of a lateral gene transfer from a eukaryote to an archaeon. Sigma factors are RNA transcription initiation factors found exclusively in Bacteria, although one conserved sigma factor domain (region four) has been reported in Archaea[46]. Here we report the first instance of complete bacteria-like sigma factors (∂70) in Archaea, specifically in two members of the Diapherotrites and one representative of the Nanoarchaeota (Fig. 3 and Supplementary Table 8). These appear to be the result of multiple lateral transfers from bacterial donors (Supplementary Fig. 16). All three sigma factors belong to the non-essential ∂70 groups (3 and 4)[47] and their hosts retain the standard archaeal TATA-binding protein gene regulation apparatus, suggesting that the co-opted full-length bacterial sigma factors are used for specialized instances of gene regulation or serve some other function (Supplementary Information).

The well-described bacterial stringent response based on deployment of multi-domain signalling molecules (guanosine tetraphosphate; ppGpp) called alarmones were identified in one member each of the Diapherotrites and Nanoarchaeota (Fig. 3). These seem to be the result of ancient transfers from bacterial donors of key ppGpp synthetic genes belonging to the RelA/SpoT homologue (RSH) superfamily[48] (Supplementary Fig. 17 and Supplementary Table 9). Although putative single domain alarmones (synthases and hydrolases) have been found in a number of Euryarchaeota[48], this is the first report of complete multi-domain archaeal alarmones comprising synthetase, hydrolase and regulatory domains, suggesting that some DPANN Archaea can produce ppGpp in response to the sensation of an intracellular signal. Finally, a bacterial-like lytic murein transglycosylase was found in two members of the Nanoarchaeota (Fig. 3 and Supplementary Fig. 18). This enzyme is ubiquitous in Bacteria and responsible for creating space within the peptidoglycan sacculus for its biosynthesis, recycling and cell division and is tightly regulated because of its potent activity[49]. As Archaea lack peptidoglycan and there is no evidence for peptidoglycan synthesis in the Nanoarchaeota, we speculate that the murein transglycosylase is secreted from the cell and used as a defensive mechanism against bacteria or possibly as a mechanism for facilitating cell-to-cell interaction with bacteria.

Phylogenetic anchoring of metagenomes

A major challenge in metagenomics is determining the phylogenetic origin of anonymous genome fragments, a process called binning or classification[50]. Our ability to classify metagenomic fragments is hampered by the enormous under-sampling of MDM reflected in a highly biased reference genome data set (Supplementary Fig. 1b). To determine whether our set of phylogenetically novel single-cell genomes improves metagenomic binning, we classified 893 publicly available metagenomes against the non-redundant database with and without the 201 SAGs (the single-cell genomes constitute a minimal increase in total database size of 0.7%). Over half (475) of these metagenomes showed new or improved read anchoring (Supplementary Table 10), which accounted for a total of 340 million reads (0.7%). Although this average percentage may seem small, up to 20% anchoring was achieved for some metagenomes, reinforcing the need for phylogenetically directed genomic characterization of microbial diversity. Metagenomes with MDM-SAG-enabled read anchoring of >2% are shown in Fig. 4, and all other metagenomes are shown in Supplementary Table 11. On average, BLASTX matches of the 340 million reclassified reads increased by approximately 27% amino acid identity, resulting in higher resolution assignments for two-thirds of these reads. Of these, 78% and 22% were newly assigned or re-assigned at the phylum level, respectively (Supplementary Fig. 19 and Supplementary Table 10). The most pronounced improvements were seen in habitats comprising dominant populations belonging to phyla that are well represented in the SAG data set including the Marinimicrobia (SAR406), Aminacenantes (OP8), Cloacimonetes (WWE-1), Parcubacteria (OD1), Atribacteria (OP9) and Microgenomates (OP11) (Fig. 4). Despite these improvements, the majority of reads in the 475 metagenomes could not be classified beyond domain level (up to 80% in some metagenomes) attesting to the continuing need for MDM exploration.

Outlook

Increasing genomic coverage of the microbial world has emerged as a major goal over the past decade and notable international efforts are underway; for example, the Microbial Earth Project, which aims to generate a comprehensive genome catalogue of all archaeal and bacterial type strains (http://www.microbial-earth.org), and the Earth Microbiome Project, which uses metagenomics, metatranscriptomics and amplicon sequencing to analyse microbial communities across the globe (http://www.earthmicrobiome.org). Although these projects will undoubtedly increase our understanding and appreciation of the microbial world, the phylogenetically targeted approach applied in the GEBA project[5] and in the present study complements these efforts and facilitates novel discovery. For example, our single-cell genome data set provides an 11% greater coverage of known phylogenetic diversity than currently available genomes according to SSU

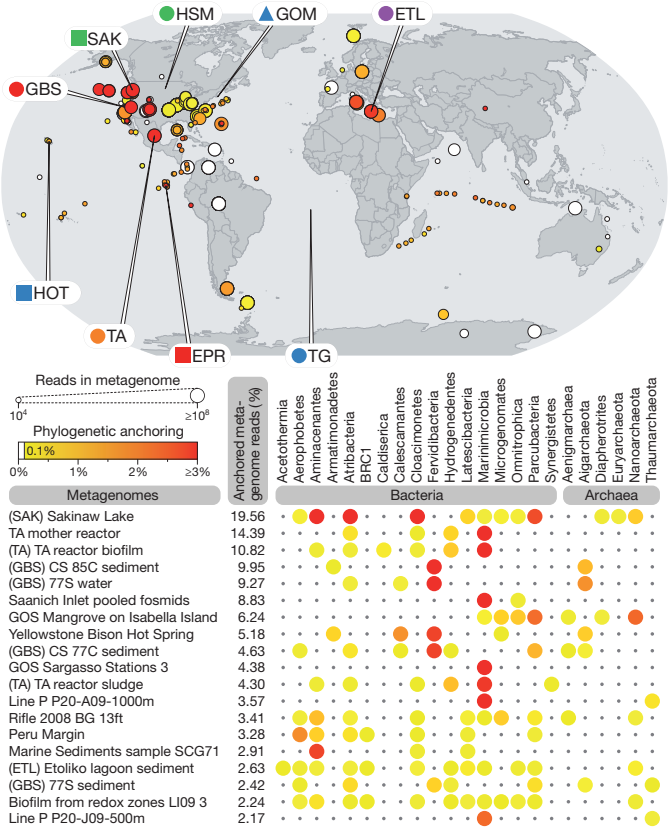

Figure 4 | Phylogenetic anchoring. The geographic location of all 475 metagenome sample sites (circles) and the origin of the MDM samples from which the SAGs were derived. The heatmap below the world map shows the details of 19 metagenomes whose phylogenetic anchoring could be improved for at least 2% of all reads. Phylogenetic anchoring was calculated as the percentage of reads that could be assigned to novel phyla using MEGAN4 results based on BLASTX analysis of all metagenomes against the NCBI non-redundant database before and after addition of MDM data. Statistical testing revealed a significant ($P = 0.00024$) increase in reads that were anchored beyond domain level after the addition of MDM data.

©2013 Macmillan Publishers Limited. All rights reserved

rRNA comparisons (Supplementary Fig. 20a). This represents a 4.5-fold increase in phylogenetic diversity per genome relative to the average phylogenetic diversity of genomes in the public database and a twofold phylogenetic diversity increase per genome afforded by GEBA[5] (Supplementary Fig. 20a). This increase is also reflected in overall protein novelty with nearly 20,000 new hypothetical protein families in the GEBA-MDM data set, representing an increase of 8.5% compared to the number of genomes sequenced to date (Supplementary Fig. 21). Although the phylogenetic diversity of microbial isolates has increased gradually over time as pure cultures accrue, the phylogenetic diversity of uncultivated microorganisms identified in SSU rRNA surveys has quadrupled since 2007 and currently represents >85% of known microbial diversity (Supplementary Fig. 20b). We estimate that a sequencing effort of at least 16,000 additional genomes from diverse environments is needed to cover 50% of the known phylogenetic diversity based on SSU rRNA profiling (Supplementary Fig. 20a). Single-cell genomics offers a means to inventory this genomic diversity at the organism level directly, bypassing the assembly and binning problems associated with plurality sequencing approaches. Further development of single-cell technologies should overcome known challenges such as fragmented genome recoveries[8] and will make this technique a more robust tool. As single-cell and other cultivation-independent genomic approaches are used, we anticipate robust improvements to the genomic tree of life that will supercede the single-locus resolution of the SSU rRNA tree. As the genomic tree is filled in, we will witness for the first time a global view of the evolutionary forces that have shaped life on Earth.

METHODS SUMMARY

Nine sites were sampled for single-cell sorting, whole-genome amplification and SSU rRNA screening. A total of 201 phylogenetically targeted SAGs were shotgun sequenced and assembled. Genome completeness was estimated based on universal, single-copy genes. Genome trees were calculated from concatenated alignments of up to 38 universally conserved protein-coding genes in Bacteria and Archaea, and phylogenetic inference was carried out via RAxML, RAxML-Light, and fasttree using 15 taxon configurations. Gene predictions, functional annotation, manual curation and pathway reconstruction were carried out within the Integrated Microbial Genomes (IMG) system (http://img.jgi.doe.gov). Phylogenetic anchoring of metagenomic reads was computed using protein blast and the lowest common ancestor approach. Phylogenetic diversity values were calculated from a SSU rRNA maximum likelihood tree. All steps are detailed in the Supplementary Information.

Received 14 January; accepted 4 June 2013.
Published online 14 July 2013.

1. Rajendhran, J. & Gunasekaran, P. Microbial phylogeny and diversity: small subunit ribosomal RNA sequence analysis and beyond. *Microbiol. Res.* **166,** 99–110 (2011).
2. Gilbert, J. A. & Dupont, C. L. Microbial metagenomics: beyond the genome. *Ann. Rev. Mar. Sci.* **3,** 347–371 (2011).
3. Pedrós-Alió, C. Marine microbial diversity: can it be determined? *Trends Microbiol.* **14,** 257–263 (2006).
4. Hugenholtz, P. & Kyrpides, N. C. A changing of the guard. *Environ. Microbiol.* **11,** 551–553 (2009).
5. Wu, D. *et al.* A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462,** 1056–1060 (2009).
6. Marcy, Y. *et al.* Dissecting biological 'dark matter' with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc. Natl Acad. Sci. USA* **104,** 11889–11894 (2007).
7. Handelsman, J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* **68,** 669–685 (2004).
8. Stepanauskas, R. Single cell genomics: an individual look at microbes. *Curr. Opin. Microbiol.* **15,** 613–620 (2012).
9. Swan, B. K. *et al.* Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science* **333,** 1296–1300 (2011).
10. Konstantinidis, K. T., Ramette, A. & Tiedje, J. M. The bacterial species definition in the genomic era. *Phil. Trans. R. Soc. Lond. B* **361,** 1929–1940 (2006).
11. Zaneveld, J. R., Lozupone, C., Gordon, J. I. & Knight, R. Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. *Nucleic Acids Res.* **38,** 3869–3879 (2010).
12. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405,** 299–304 (2000).
13. Szöllosi, G. J., Boussau, B., Abby, S. S., Tannier, E. & Daubin, V. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc. Natl Acad. Sci. USA* **109,** 17513–17518 (2012).
14. Ciccarelli, F. D. *et al.* Toward automatic reconstruction of a highly resolved tree of life. *Science* **311,** 1283–1287 (2006).
15. Dalevi, D., Hugenholtz, P. & Blackall, L. A multiple-outgroup approach to resolving division-level phylogenetic relationships using 16S rDNA data. *Int. J. Syst. Evol. Microbiol.* **51,** 385–391 (2001).
16. Bergsten, J. A review of long-branch attraction. *Cladistics* **21,** 163–193 (2005) CrossRef.
17. Gordon, D. A. & Giovannoni, S. J. Detection of stratified microbial populations related to *Chlorobium* and *Fibrobacter* species in the Atlantic and Pacific oceans. *Appl. Environ. Microbiol.* **62,** 1171–1177 (1996).
18. Hugenholtz, P., Pitulle, C., Hershberger, K. L. & Pace, N. R. Novel division level bacterial diversity in a Yellowstone hot spring. *J. Bacteriol.* **180,** 366–376 (1998).
19. Dojka, M. A., Hugenholtz, P., Haack, S. K. & Pace, N. R. Microbial diversity in a hydrocarbon- and chlorinated-solvent-contaminated aquifer undergoing intrinsic bioremediation. *Appl. Environ. Microbiol.* **64,** 3869–3877 (1998).
20. McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* **6,** 610–618 (2012).
21. Wagner, M. & Horn, M. The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. *Curr. Opin. Biotechnol.* **17,** 241–249 (2006).
22. Gupta, R. S., Bhandari, V. & Naushad, H. S. Molecular signatures for the PVC clade (Planctomycetes, Verrucomicrobia, Chlamydiae, and Lentisphaerae) of bacteria provide insights into their evolutionary relationships. *Front. Microbiol.* **3,** 327 (2012).
23. Gupta, R. S. The phylogeny and signature sequences characteristics of Fibrobacteres, Chlorobi, and Bacteroidetes. *Crit. Rev. Microbiol.* **30,** 123–143 (2004).
24. Zhang, H. *et al.* Gemmatimonas aurantiaca gen. nov., sp. nov., a gram-negative, aerobic, polyphosphate-accumulating micro-organism, the first cultured representative of the new bacterial phylum Gemmatimonadetes phyl. nov. *Int. J. Syst. Evol. Microbiol.* **53,** 1155–1163 (2003).
25. Miroshnichenko, M. L. *et al. Caldithrix abyssi* gen. nov., sp. nov., a nitrate-reducing, thermophilic, anaerobic bacterium isolated from a Mid-Atlantic Ridge hydrothermal vent, represents a novel bacterial lineage. *Int. J. Syst. Evol. Microbiol.* **53,** 323–329 (2003).
26. Battistuzzi, F. U. & Hedges, S. B. A major clade of prokaryotes with ancient adaptations to life on land. *Mol. Biol. Evol.* **26,** 335–343 (2009).
27. Tamaki, H. *et al. Armatimonas rosea* gen. nov., sp. nov., a Gram-negative, aerobic, chemoheterotrophic bacterium of a novel bacterial phylum, *Armatimonadetes* phyl. nov., formally called the candidate phylum OP10. *Int. J. Syst. Evol. Microbiol.* **61,** 1442–1447 (2011).
28. Sutcliffe, I. C. Cell envelope architecture in the Chloroflexi: a shifting frontline in a phylogenetic turf war. *Environ. Microbiol.* **13,** 279–282 (2011).
29. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnol.* **6,** 533–538 (2013).
30. Chouari, R. *et al.* Novel major bacterial candidate division within a municipal anaerobic sludge digester. *Appl. Environ. Microbiol.* **71,** 2145–2153 (2005).
31. Harris, J. K., Kelley, S. T. & Pace, N. R. New perspective on uncultured bacterial phylogenetic division OP11. *Appl. Environ. Microbiol.* **70,** 845–849 (2004).
32. Ley, R. E. *et al.* Unexpected diversity and complexity of the guerrero negro hypersaline microbial mat. *Appl. Environ. Microbiol.* **72,** 3685–3695 (2006).
33. Wrighton, K. C. *et al.* Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* **337,** 1661–1665 (2012).
34. Baker, B. J. *et al.* Enigmatic, ultrasmall, uncultivated Archaea. *Proc. Natl Acad. Sci. USA* **107,** 8806–8811 (2010).
35. Narasingarao, P. *et al.* De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. *ISME J.* **6,** 81–93 (2012).
36. Takai, K. & Horikoshi, K. Genetic diversity of Archaea in deep-sea hydrothermal vent environments. *Genetics* **152,** 1285–1297 (1999).
37. Takai, K., Moser, D. P., DeFlaun, M., Onstott, T. C. & Fredrickson, J. K. Archaeal diversity in waters from deep south african gold mines. *Appl. Environ. Microbiol.* **67,** 5750–5760 (2001).
38. Guy, L. & Ettema, T. J. G. The archaeal 'TACK' superphylum and the origin of eukaryotes. *Trends Microbiol.* **19,** 580–587 (2011).
39. Lake, J. A., Henderson, E., Oakes, M. & Clark, M. W. Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proc. Natl Acad. Sci. USA* **81,** 3786–3790 (1984).
40. Williams, T. A., Foster, P. G., Nye, T. M. W., Cox, C. J. & Embley, T. M. A congruent phylogenomic signal places eukaryotes within the Archaea. *Proc. R. Soc. B* **279,** 4870–4879 (2012).
41. Campbell, J. H. *et al.* UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proc. Natl Acad. Sci. USA* **110,** 5540–5545 (2013).
42. Johansson, L., Gafvelin, G. & Arnér, E. S. J. Selenocysteine in proteins—properties and biotechnological use. *Biochimica et Biophysica Acta* **1726,** 1–13 (2005).
43. Yamao, F. *et al.* UGA is read as tryptophan in *Mycoplasma capricolum*. *Proc. Natl Acad. Sci. USA* **82,** 2306–2309 (1985).
44. McCutcheon, J. P., McDonald, B. R. & Moran, N. A. Origin of an alternative genetic code in the extremely small and GC–rich genome of a bacterial symbiont. *PLoS Genet.* **5,** e1000565 (2009).
45. Zhang, Y., Morar, M. & Ealick, S. E. Structural biology of the purine biosynthetic pathway. *Cell. Mol. Life Sci.* **65,** 3699–3724 (2008).

©2013 Macmillan Publishers Limited. All rights reserved

46. Kyrpides, N. C. & Ouzounis, C. A. Bacterial sigma 70 transcription factor DNA-binding domains in the archaeon *Methanococcus jannaschii*. *J. Mol. Evol.* **45,** 706–707 (1997).
47. Paget, M. S. & Helmann, J. D. The σ70 family of sigma factors. *Genome Biol.* **4,** 203 (2003).
48. Atkinson, G. C., Tenson, T. & Hauryliuk, V. The RelA/SpoT homolog (RSH) superfamily: distribution and functional evolution of ppGpp synthetases and hydrolases across the tree of life. *PLoS ONE* **6,** e23479 (2011).
49. Scheurwater, E., Reid, C. W. & Clarke, A. J. Lytic transglycosylases: bacterial space-making autolysins. *Int. J. Biochem. Cell Biol.* **40,** 586–591 (2008).
50. Dröge, J. & McHardy, A. C. Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. *Brief. Bioinform.* **13,** 646–655 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the DOE JGI production sequencing, IMG and GOLD teams for their support; J. Lee and E. Ng for experimental assistance; H.-P. Klenk and D. Gleim for providing a DSMZ inventory database dump and I. Letunić for his knowledge and support to make iTOL work for this project. We are very grateful to B. Schink for invaluable etymological advice. The work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231. We also thank the CeBiTec Bioinformatics Resource Facility, which is supported by BMBF grant 031A190. B.P.H. and J.A.D. were supported by the NASA Exobiology grant EXO-NNX11AR78G and NSF OISE 096842 and B.P.H. by a generous contribution from G. Fullmer through the UNLV Foundation. S.M.S was supported by NSF grants OCE-0452333 and OCE-1136727, and the WHOI's Andrew W. Mellon Fund for Innovative Research; and S.J.H. by the Canadian Foundation for Innovation, the British Columbia Knowledge Development Fund, the National Sciences and Engineering Research Council (NSERC) of Canada and the TULA foundation funded Centre for Microbial Diversity and Evolution (CMDE), and the Canadian Institute for Advanced Research (CIFAR). R.S. was supported by NSF grants DEB-841933, EF-826924, OCE-1232982, OCE-821374 and OCE-1136488, and the Deep Life I grant by the Alfred P. Sloan Foundation. P.H. was supported by a Discovery Outstanding Researcher Award (DORA) from the Australian Research Council, grant DP120103498.

Author Contributions T.W., C.R. and E.M.R. designed the project, B.K.S., E.A.G., J.A.D., B.P.H., G.T., S.M.S., W.-T.L., S.J.H. and R.S. provided the samples, C.R., T.W. and J.-F.C. performed the experiments, C.R., P.S., A.S., N.N.I., I.J.A., A.D. and S.M. analysed the data, C.R., P.S. and I.J.A. created the figures and tables, and C.R., P.H. and T.W. wrote the manuscript with significant input from A.S., A.D., S.J.H., B.P.H., N.C.K., J.A.E., R.S. and E.M.R.

Author Information Whole-Genome Shotgun projects have been deposited at GenBank under the accession numbers AQPI00000000, AQRL00000000–AQRZ00000000, AQSA00000000–AQSZ00000000, AQTA00000000–AQTF00000000, AQYL00000000–AQYX00000000, ARTZ00000000, ARWO00000000, ASKJ00000000–ASKZ00000000, ASLA00000000–ASLZ00000000, ASMA00000000–ASMZ00000000, ASNA00000000–ASNZ00000000, ASOA00000000–ASOZ00000000, ASPA00000000–ASPH00000000, ASPJ00000000–ASPO00000000, ASWY00000000, ASZK00000000 and ASZL00000000. The annotated single-cell assemblies can be accessed via IMG (http://img.jgi.doe.gov). Single-cell genome assemblies are also available at the Microbial Dark Matter project webpage (http://genome.jgi.doe.gov/MDM). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.W. (twoyke@lbl.gov) and P.H. (p.hugenholtz@uq.edu.au).

This work is licensed under a Creative Commons Attribution-NonCommercial-Share Alike 3.0 Unported licence. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-sa/3.0

©2013 Macmillan Publishers Limited. All rights reserved