Insights into the phylogeny and coding potential of microbial dark matter

Christian Rinke1, Patrick Schwientek1, Alexander Sczyrba1,2, Natalia N. Ivanova1, Iain J. Anderson1,4, Jan–Fang Cheng1, Aaron Darling1,4, Stephanie Malfatti1, Brandon K. Swan5, Esther A. Gies6, Jeremy A. Dodsworth7, Brian P. Hedlund7, George Tsiamis8, Stefan M. Sievert9, Wen–Tso Liu10, Jonathan A. Eisen3, Steven J. Hallam6, Nikos C. Kyprides1, Ramunas Stepanauskas9, Edward M. Rubin1, Philip Hugenholtz11 & Tanja Woyke3

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 more intra– and inter–phyllum–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 prokaryotes (Supplementary Fig. 1a). Genome sequencing of microbial isolates naturally reflects this culti

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 millions3. According to SSU rRNA–based phylogeny, these fall into at least 60 major lines of descent (phyla or divisions) within the bacterial and archaeal domains4, 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) Project5, 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)6,7. Metagenomics can obtain genome sequences from uncultivated microorganisms through direct sequencing of DNA from the environment7.

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 samples8. 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 GEBADMD, 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.
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. 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 49%) 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 despite the blurring of vertical descent by lateral gene transfer. 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. We constructed bootstrapped maximum likelihood trees based on a concatenation of up to 38 commonly used conserved marker genes with 15 taxa configurations (Supplementary Table 3) with 15 taxa configurations (Supplementary Table 4). Substitution models were selected to address known issues, including long branch attraction (discussed further in Supplementary Information).

Congruency of the individual marker gene topologies to each other was independently assessed confirming the selection of these gene

**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 Pacific; WP, Western Pacific; WC, Western Coastal; WO, West Oahu; OS, Oceanic Shelf; TA, terephthalate degrading reactor; TG, tropical gyre in the South Pacific. 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 profiles 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. 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 49%) 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 despite the blurring of vertical descent by lateral gene transfer. 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. We constructed bootstrapped maximum likelihood trees based on a concatenation of up to 38 commonly used conserved marker genes with 15 taxa configurations (Supplementary Table 3) with 15 taxa configurations (Supplementary Table 4). Substitution models were selected to address known issues, including long branch attraction (discussed further in Supplementary Information).

Congruency of the individual marker gene topologies to each other was independently assessed confirming the selection of these gene

**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 Pacific; WP, Western Pacific; WC, Western Coastal; WO, West Oahu; OS, Oceanic Shelf; TA, terephthalate degrading reactor; TG, tropical gyre in the South Pacific. 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 profiles 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. 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 49%) 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 despite the blurring of vertical descent by lateral gene transfer. 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. We constructed bootstrapped maximum likelihood trees based on a concatenation of up to 38 commonly used conserved marker genes with 15 taxa configurations (Supplementary Table 3) with 15 taxa configurations (Supplementary Table 4). Substitution models were selected to address known issues, including long branch attraction (discussed further in Supplementary Information).

Congruency of the individual marker gene topologies to each other was independently assessed confirming the selection of these gene
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 monopholy 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.

Together with Marinimicrobia (SAR406), Latescibacteria (WS3), Cloacimonetes (WWE1), Gemmatimonadetes24 and Caldithrix25. 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 Firmicutes26, was resolved in our analysis with the additional membership of Armamentimonadetes (former candidate phylum OP10)27 (Fig. 2). Perhaps more compelling than the assertion of ancient adaptations to life on land unifying the Terrabacteria28 are commonalities in cell envelope architecture. This superphylum comprises monoderm (single membrane) and atypical monoderm lineages29. We assessed the additional proposed Terrabacteria phyla for genes most characteristic of monoderms and dierms26 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
originally proposed as a candidate phylum\(^{36}\) and more recently as a class within the Spirochaetes phylum\(^{29}\). 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 Aminacatenantes (OP8) (Fig. 2). Candidate phylum OP11, as originally proposed\(^{36}\), 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)\(^{37}\). To recognize this affiliation, we propose the superphylum name ‘Patescibacteria’ (patesco (Latin), meaning bare) (Fig. 2), reflecting the reduced metabolic capacities of these lineages\(^{37}\). 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-sequence lineages of very small cells, such as Candidatus Parvarchaeum, Candidatus Micarchaeum\(^{38}\), Candidatus Nanosalin, Candidatus Nanosalinarum\(^{39}\), we propose the following phyla; Diapherotrites (pMC2A384)\(^{40}\), Parvarchaeota, Aenigmarchaeota (DSEG)\(^{37}\) and Nanoarchaeota (Fig. 2 and Supplementary Table 5). The Nanoarchaeota includes the recently proposed class Nanoarchaeota that was incorrectly placed within the Euryarchaea 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, ‘Patescibacteria’ (Superphylum Patescibacteria)\(^{41}\). 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.

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 archaean 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 archaean 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 Caldserica 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 recoding 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 Gracilbacteria 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 ribonucleotide 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 archaean 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; FALK, 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 archaean. Two members of the Diapherotites (pMC2A384) and one representative of the Nanoarchaeota encode complete bacteria-like sigma factors (\(^\text{\sigma}\)s). The bacterial stringent response based on deployment of signalling molecules (ppGpp) was identified in a member of the Diapherotites and the Nanoarchaeota. A bacterial-like lytic murein transglycosylase was identified in two members of the Nanoarchaeota. \(\alpha\)C-terminal domain; \(\alpha\)N-terminal domain; ADP, adenosine di-phosphate; GDP, guanosine di-phosphate; GTP, guanosine tri-phosphate.
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 an 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 Archaea46. Here we report the first instance of complete bacteria-like sigma factors (分辨率?) 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 (分辨率?) groups (3 and 4) 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 (Supplementary Fig. 17 and Supplementary Table 9). Although putative single domain alarmones (synthases and hydrolases) have been found in a number of Euryarchaeota48, 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. 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 classification50. 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 (473) 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), Aminocrenantes (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 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
METHODS SUMMARY

Nine sites were sampled for single-cell sorting, whole-genome amplification and SSU RNA sequencing. 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 RNA maximum likelihood tree. All steps are detailed in the Supplementary Information.

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. 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 CoBiTeC 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 B.P.H. by a generous contribution from G. Fullmer through the UNLV Foundation. S.M.S. was supported by NSF grants DEB-841933, EF-821374, OCE-1232982, OCE-821374 and OCE-1136488, and the Deep Life Foundation. P.H. was supported by a Discovery Outstanding Researcher Award (DORA) from the Australian Research Council, grant DP120103498. B.P.H. and J.A.D. provided the samples, C.R. 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., B.P.H. and T.W. wrote the manuscript with significant input from A.S., T.W., 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, AQR00000000–AQRZ00000000, AQSA00000000–AQYZ00000000, AQTA00000000–AQTF00000000, AQYX00000000–ARTZ00000000, ARW00000000–ASKL00000000, ASMA00000000–ASMF00000000, ASN00000000–ASN20000000, ASO00000000–ASPF00000000, ASPH00000000–ASZL00000000, ASZK00000000–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