nature biotechnology

ARTICLES

Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes

Mads Albertsen[1], Philip Hugenholtz[2,3], Adam Skarshewski[2], Kåre L Nielsen[1], Gene W Tyson[2,4] & Per H Nielsen[1]

Reference genomes are required to understand the diverse roles of microorganisms in ecology, evolution, human and animal health, but most species remain uncultured. Here we present a sequence composition–independent approach to recover high-quality microbial genomes from deeply sequenced metagenomes. Multiple metagenomes of the same community, which differ in relative population abundances, were used to assemble 31 bacterial genomes, including rare (<1% relative abundance) species, from an activated sludge bioreactor. Twelve genomes were assembled into complete or near-complete chromosomes. Four belong to the candidate bacterial phylum TM7 and represent the most complete genomes for this phylum to date (relative abundances, 0.06–1.58%). Reanalysis of published metagenomes reveals that differential coverage binning facilitates recovery of more complete and higher fidelity genome bins than other currently used methods, which are primarily based on sequence composition. This approach will be an important addition to the standard metagenome toolbox and greatly improve access to genomes of uncultured microorganisms.

A grand challenge in biology today is to obtain representative genomic coverage of the tree of life to improve our understanding of evolution and ecology. Initiatives such as the Genomic Encyclopedia of Bacteria and Archaea have begun to address this challenge by systematically sequencing genomes from pure cultures[1]. However, culture-independent molecular surveys using the 16S rRNA gene have revealed that only a small fraction of the phylogenetic diversity of Bacteria and Archaea is represented by cultivated organisms[2]. Conspicuous among the uncultured and unsequenced microbial majority are the numerous 'candidate phyla' that constitute major unstudied evolutionary lines of descent[3]. Members of candidate phyla play important roles in the environment, such as in anaerobic methane oxidation[4], and in human disease, such as oral and gut inflammation[5–8].

The cultivation bottleneck can be bypassed using metagenomics, in which bulk DNA is directly extracted from environmental samples and shotgun sequenced[9]. Near-complete genomes of dominant populations from relatively low-complexity communities have been assembled from shotgun-sequenced metagenomes, showing the potential of this method for obtaining genomes of uncultivated microbial species[10,11]. However, it has proven difficult to assemble genomes for populations below 1% relative abundance using current metagenome sequencing and assembly approaches, owing to insufficient sequencing depth or difficulty in binning (classification) and assembly of individual genomes from complex metagenomes[9,12]. One solution is single-cell genomics, but obtaining complete genomes from single cells is technically challenging owing to amplification bias, which can result in uneven coverage and fragmented assemblies[13,14].

Rapid improvements in sequencing throughput, read length and quality make it theoretically possible to assemble genomes from low-abundance populations in 'deep' (tens of Gbp) metagenomic data sets, provided strain heterogeneity is limited[15]. Recovery of individual genomes from a background of many other genomes in complex samples has been addressed primarily using methods dependent on the sequence composition of the genomes[16]. Tetranucleotide frequency–based emergent self-organizing maps (ESOMs) have been applied to obtain 49 mostly low-abundance population genomes from a deep metagenome of an acetate-amended aquifer[17]. However, many of the obtained genomes were not cleanly separated using ESOM-based binning as indicated by the presence of multiple copies of single-copy genes within the population bins (Fig. S8 in ref. 17), which might lead to inaccurate understanding of their evolution, physiology and ecology.

To assemble high-quality, near-complete draft genomes of rare species from deeply sequenced metagenomes, we used sequence composition–independent information as the primary binning approach, complemented by post-binning refinement of population genomes using sequence composition–dependent methods and visualization tools. We generated two related metagenome data sets (29 and 57 Gbp) and obtained a total of 31 population genome bins including rare (<1% relative abundance) species. Twelve of these were further refined into complete or near-complete chromosomes of which four belong to the candidate phylum TM7.

[1]Department of Biotechnology, Chemistry and Environmental Engineering, Aalborg University, Aalborg, Denmark. [2]Australian Centre for Ecogenomics, School of Chemistry & Molecular Biosciences, The University of Queensland, St. Lucia, Queensland, Australia. [3]Institute for Molecular Bioscience, The University of Queensland, St. Lucia, Queensland, Australia. [4]Advanced Water Management Centre, The University of Queensland, St. Lucia, Queensland, Australia. Correspondence should be addressed to P.H.N. (phn@bio.aau.dk).

Received 12 September 2012; accepted 10 April 2013; published online 26 May 2013; doi:10.1038/nbt.2579

RESULTS

Differential binning of deep metagenomes

We deeply sequenced DNA from an activated sludge bioreactor obtained using two different extraction methods (with and without hot phenol, HP^+ and HP^-, respectively), producing a total of 29 and 57 Gbp of high-quality, paired-end sequence data for the HP^+ and HP^- samples, respectively, with an average read length of 124 bp. The larger data set (HP^-) was assembled into a total of 423 Mbp of scaffolds ranging in size from 1 kbp to 3.6 Mbp (**Supplementary Table 1**). Mapping reads from each data set to the assembled scaffolds indicated that populations were differentially represented (that is, different coverage of scaffolds in each data set) due to species-specific extraction efficiencies. The differences in extraction efficiency between the two extraction methods were also reflected by differences in the associated V4 16S rRNA gene amplicon data (**Supplementary Fig. 1**).

Binning of scaffolds into population genomes was facilitated by plotting the two coverage estimates against each other for all scaffolds (**Fig. 1**). Scaffolds clustering together represent putative population genome bins. As multiple species can be present in the same coverage-defined subset, the selected scaffold subset was further refined using principal component analysis of tetranucleotide frequencies (**Fig. 2**). In total, 31 population bins were identified representing wide bacterial diversity (five phyla) and GC content range (32–71%) (**Fig. 1**). These population bins included many rare members of the community with the lowest being 0.02% relative abundance (**Supplementary Table 2**). The population bins capture 41% of all sequenced reads and 65% of all reads that could be assembled into scaffolds >1 kbp. The 13 most-complete population genomes were further improved by tracking paired-end reads in network graphs. This allowed association of additional scaffolds and repeat regions (for example, rRNA genes) with the correct bin, and removal of scaffolds wrongly included in bins (**Fig. 2**). In addition, the graph visualization allowed us to track and estimate the strain heterogeneity of each population (**Supplementary Fig. 2**).

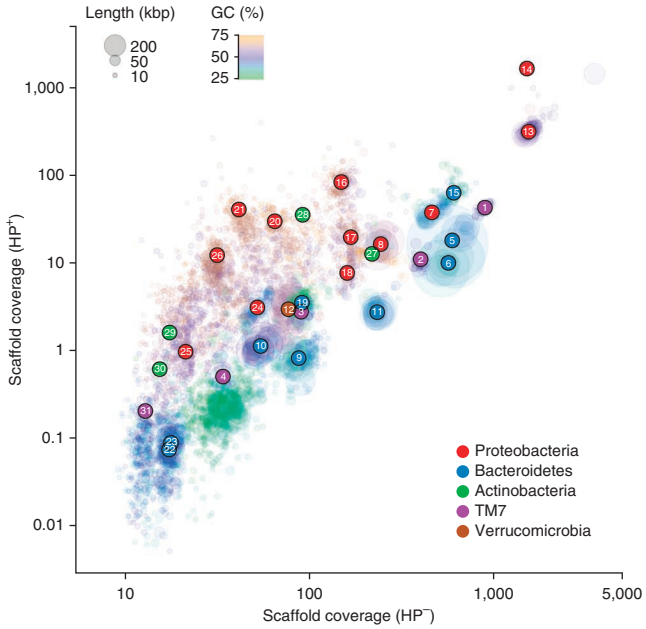

Figure 1 Sequence composition–independent binning of metagenome scaffolds from the lab-scale bioreactor using differential coverage (HP^+, HP^-). Circles represent scaffolds, scaled by the square root of their length and colored by GC content. Only scaffolds ≥5 kbp are shown. Clusters of similarly colored circles represent potential genome bins, the centroids of which are indicated by numbered circles and colored according to phylum-level taxonomic affiliation (**Table 1** and **Supplementary Table 2**). This differential coverage plot provides the starting point for secondary refinement and finishing of genome assemblies (**Fig. 2**).

Assembly of individual genomes

Once the initial scaffold membership of each population bin had been refined, that is, additional correct scaffolds had been included and contaminant scaffolds removed, all reads mapping to those scaffolds and any associated paired-end reads were *de novo* assembled using Velvet[18] to produce individual draft population genomes. Finally, by tracking and visualizing all paired-end reads mapping to the reassembled scaffolds through Cytoscape[19] and Circos[20] (**Fig. 2** and **Supplementary Fig. 3**), we manually curated the assemblies by correcting ends of scaffolds, small mis-assemblies and manual scaffolding. In 5 of 12 cases we were able to assemble the genomes to the theoretical limit, given the estimated repeat content. The repeat content was estimated as the number of repetitive sequences that cannot be spanned by paired-end reads, in this case ~500 bp.

Validation of population genome assemblies

Currently, there is no best practice for validation of population genome assemblies from metagenomic data in terms of completeness and potential contamination with other species. In general, validation has been done using essential single-copy genes, either universally conserved genes[17], for example, ribosomal genes or conserved core genes within related organisms[21]. However, the small set of ribosomal genes makes it difficult to assess the true completeness and contamination of assembled genomes because ribosomal genes are located in a restricted part of the genome, whereas using core genes is problematic because of the difficulty in defining the core genes, especially when no closely related organisms are available. Here, the completeness of each population bin was investigated using a suite of hidden Markov models (HMM) covering 107 proteins conserved in 95% of all sequenced bacteria[22]. The number of expected essential single-copy genes was further investigated on the phylum level for all sequenced bacteria to improve the estimation of completeness and contamination (**Supplementary Fig. 4**).

Although an overall completeness and contamination estimate can be obtained using essential single-copy genes, the complexity of metagenome assemblies necessitates manual inspection of all scaffolds in the population genome bins to identify potential mis-assemblies and chimeric scaffolds. To facilitate visual inspections of the population genome bins, we wrote a script that takes a FASTA file and a SAM file of the associated metagenome read mapping and generates all relevant data for visual inspection through Circos[20] and also integrates the reference-free assembly validation statistics generated by FRCbam[23]. The use of Circos enables both a complete overview of the population genome assembly and detailed inspection of specific regions.

A complete overview of the binning, assembly and validation process is shown in **Figure 2** and a detailed step-by-step user guide is available on GitHub (https://github.com/MadsAlbertsen/multi-metagenome). After refinement of the 13 most-complete population genome bins, we recovered 12 as essentially complete genomes (≥99% of essential genes identified), of which two were obtained in single, circular chromosomes (**Table 1**).

Differential binning compared with ESOM

To directly compare the efficiency of our differential coverage binning approach to a widely used, sequence composition–based binning method (ESOM), we reanalyzed a published data set[17]. In that work,

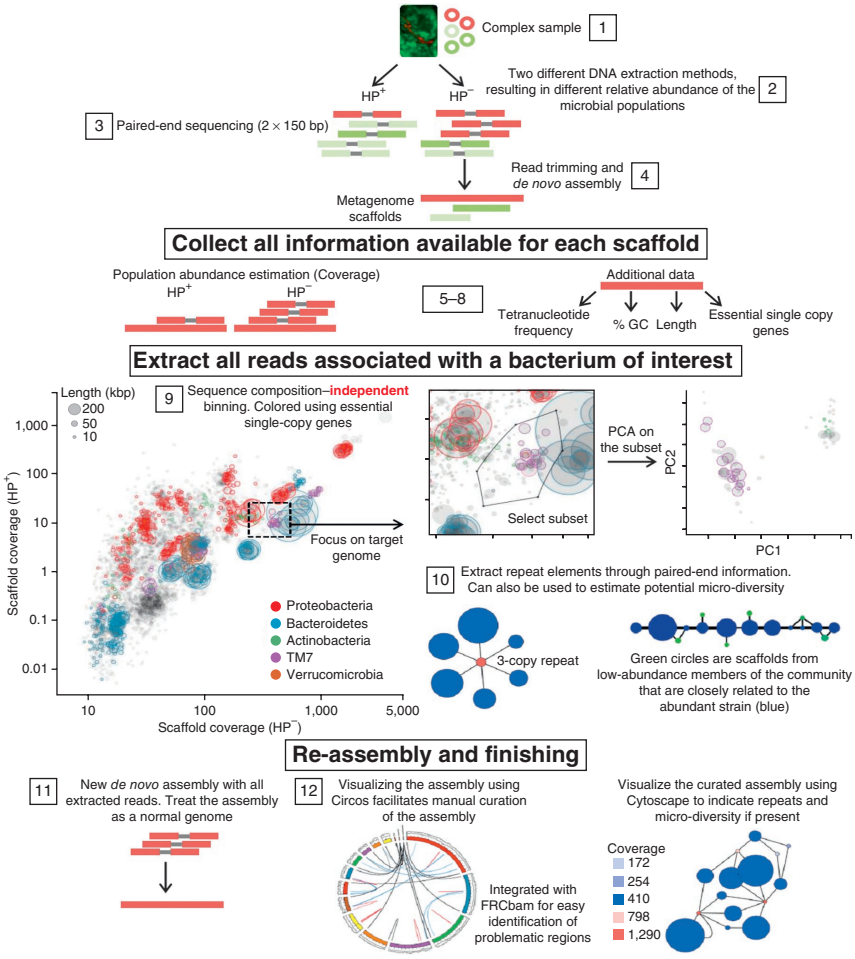

Figure 2 Overview of the pipeline to obtain high-quality population genomes from multiple deep metagenomes using differential coverage as the primary binning method, illustrated using the population genome TM7-AAU-ii. Numbers refer to subsections in Online Methods and in the detailed step-by-step guide on GitHub. Steps 1–4: DNA was extracted using two different methods (HP^+, HP^-), which produced different population abundances. Each sample (HP^+, HP^-) was then shotgun-sequenced (150 bp paired-end, average 124 bp after trimming) followed by independent scaffold assembly. Only the HP^- scaffolds were used to extract population genome bins. Steps 5–8: preparation of data for the subsequent binning steps. Differential coverage was estimated by independently mapping the reads from each metagenome to the scaffolds from the HP^- assembly, to produce two abundance estimates (coverage) per scaffold. In addition, for each scaffold the GC content and tetranucleotide frequency was calculated, and conserved essential single-copy marker genes identified. Step 9: binning (clustering) of scaffolds into population genomes was done by plotting the two coverage estimates (one from each metagenome) against each other for all HP^- scaffolds (**Fig. 1**). Scaffold subsets clustering together represent putative population genomes and were extracted as initial bins. As multiple species could be present in the same coverage-defined subset, the selected scaffold subset was further refined using principal component analysis of tetranucleotide frequencies. Step 10: as some genes are present in multiple copies (for example, 16S rRNA or transposases) they will not be included in the initial coverage-defined subset. Instead paired-end read information is used to associate multiple copy genes with the appropriate genome bin (**Supplementary Fig. 2**). Steps 11, 12: all reads associated with a genome bin of interest are extracted and re-assembled using parameters optimized for each genome as the bins can now be treated as standard single genomes. Population genome assemblies were validated using conserved single-copy gene analysis, and through Circos (a visualization tool) in which all relevant assembly metrics, including FRCbam statistics[23], are integrated to identify mis-assemblies and other structural problems. All data generation and integration are automated and can be carried out using a FASTA file of the assembled scaffolds and SAM files of the read mappings to the scaffolds.

three metagenomic data sets were obtained from an acetate-amended aquifer (time series, day 5, 7 and 10), enabling differential coverage analysis[17]. Although we used different DNA extraction methods for differential binning, different data sets from the same sample obtained over time are also suitable. The initial differential coverage binning plot that we obtained from reanalysis of the three data sets (**Fig. 3a**) revealed a high degree of structure in the data, with distinct clusters representing potential genome bins. When we compared the individual sequence composition–derived population genomes reported in the original study[17] with the data we reanalyzed using differential-coverage binning, we found the latter method resulted in more complete genome bins (for example, genome ACD7, 101 versus 89 essential genes) with less contamination from other populations (usually closely related ones, for example, ACD7, 0 versus 11 duplicated single-copy genes; **Fig. 3b**,**c**, **Supplementary Table 3**, and **Supplementary Figs. 5** and **6**).

Complete genomes for candidate phylum TM7

Four of the refined genomes from the activated sludge bioreactor were recovered from rare populations of candidate phylum TM7. These represent the first high-quality genomes reported for this phylum as only four partial (<50% complete) single-cell genomes are currently available[7,24]. The presence of multiple TM7 populations was confirmed in the activated sludge reactor using both 16S rRNA gene amplicon community profiling (**Supplementary Fig. 1**) and fluorescence *in situ* hybridization (FISH) with a range of TM7-specific probes (**Fig. 4** and **Supplementary Fig. 7**). Candidate phylum TM7 is often found as a minor but persistent constituent of microbial communities. TM7 bacteria are widespread in natural and engineered ecosystems[25–27], and are found in humans where they have been implicated in oral and gut inflammation[5–8]. Furthermore, TM7 appear to have a monoderm (single membrane, typically Gram-positive, for example, Actinobacteria and Firmicutes) cell envelope[27] and the addition of complete TM7 genomes may further inform cell envelope evolution[28]. Conserved gene analysis revealed that 7 of 107 marker genes were absent from all four TM7 genomes assembled in this study (**Supplementary Table 4**) despite assembly metrics indicating that three of the four genomes were essentially complete (**Table 1**). This demonstrates the power of using multiple genomes belonging to the same phylogenetic lineage to infer absence of even highly conserved genes.

The average TM7 genome size from this study is 1 Mbp, which is close to the currently observed lower limit for free-living bacteria[29,30].

Table 1 Assembly statistics for the 13 high-quality genomes

Phylogenetic affiliation	No. contigs	Total length (bp)	GC (%)	HP$^-$ coverage	HP$^+$ coverage	No. essential genes	No. duplicated single-copy genes	Relative metagenome abundance (%)	Figure ID
Candidate phylum TM7	1[a]	1,013,781	49	890	43	100/100	0	1.58	1 AAU-TM7-i
Bacteriodetes	1[a]	4,935,720	39	568	10	105/105	0	4.92	6
Alphaproteobacteria	9	2,824,578	41	463	38	105/105	0	2.29	7
Bacteriodetes	20	4,205,643	38	88	1	105/105	0	0.65	9
Bacteriodetes	35	4,864,786	50	55	1	105/105	0	0.47	10
Bacteriodetes	34	3,986,094	42	234	3	105/105	0	1.64	11
Gammaproteobacteria	99	3,892,057	57	1,550	314	105/105	0	10.59	13
Bacteriodetes	1	4,606,165	42	596	18	104/105	0	4.82	5
Candidate phylum TM7	15	974,669	53	402	11	99/100	0	0.69	2 AAU-TM7-ii
Verrucomicrobia	68	6,714,619	61	77	3	103/104	0	0.91	12
Betaproteobacteria	13	3,265,081	61	244	16	106/106	1	1.40	8
Candidate phylum TM7	19	953,918	48	34	1	100/100	2	0.06	4 AAU-TM7-vi
Candidate phylum TM7	131	925,229	51	94	3	88/100	8	0.15	3 AAU-TM7-iii

The statistics rival those obtained by conventional sequencing of pure cultures in terms of completeness and the number of contigs. The number of essential genes was estimated using 107 HMMs of protein coding–essential, single-copy genes that are conserved in 95% of all bacteria. The number is given in relation to the average number of essential genes in all sequenced finished genomes in the given phyla (**Supplementary Fig. 4**). For candidate phylum TM7 this was estimated using the four genomes obtained in the present study. The relative metagenome abundance was calculated as the percentage of reads contained in the genome bin compared to the total number of sequenced metagenome reads. "ID" refers to the identifier used in **Figure 1** and in the text. Feature response curves[23] for all 13 population genomes before and after finishing is shown in **Supplementary Figure 10**.

[a]Obtained as a single circular chromosome.

We estimated a similar genome size using the conserved single-copy gene analysis for the four previously published partial TM7 single-cell genomes from an oral[7] and soil[24] habitat (**Supplementary Notes** and **Supplementary Fig. 8**). Given the phylogenetic breadth of these genomes (up to 10% divergence in 16S rRNA genes), we hypothesize that compact genomes are a unifying feature of the TM7 phylum.

Metabolism and evolution of TM7

Absence of key genes for the Embden-Meyerhof (phosphofructokinase) and Entner-Doudoroff (KDPG aldolase) pathways suggest that TM7 can use only the pentose phosphate and heterolactic fermentation pathways for which all key genes were identified (**Fig. 4a** and **Supplementary Table 5**). Fermentation of glucose to lactate and acetate by these pathways is further supported by the presence of polysaccharases that could help to provide the necessary sugar precursors, and by *in situ* uptake of glucose using FISH-microautoradiography[31]. Genes for the electron transport chain and tricarboxylic acid cycle are absent from the TM7 genomes with the exception of succinyl-CoA synthase[7,24] (**Supplementary Table 5**), consistent with an obligately fermentative metabolism. The presence of superoxide dismutase and glutathione peroxidase genes provides mechanisms for oxygen tolerance under aerobic conditions experienced in the bioreactor. Furthermore, microscopy shows TM7 cells forming small microcolonies often buried deeply in floc material, which would promote microaerophilic conditions (**Supplementary Fig. 7**).

As previously identified in partial TM7 single-cell genomes, the sludge TM7 bacteria also encode the genes necessary for production of Type IV pili. These pili play a role in twitching motility and adhesion to surfaces, and are known virulence factors in some pathogenic bacteria[32]. Type IV pili may be of importance for TM7 species implicated in clinical disorders such as Crohn's disease[8], colitis[5] and periodontitis[6].

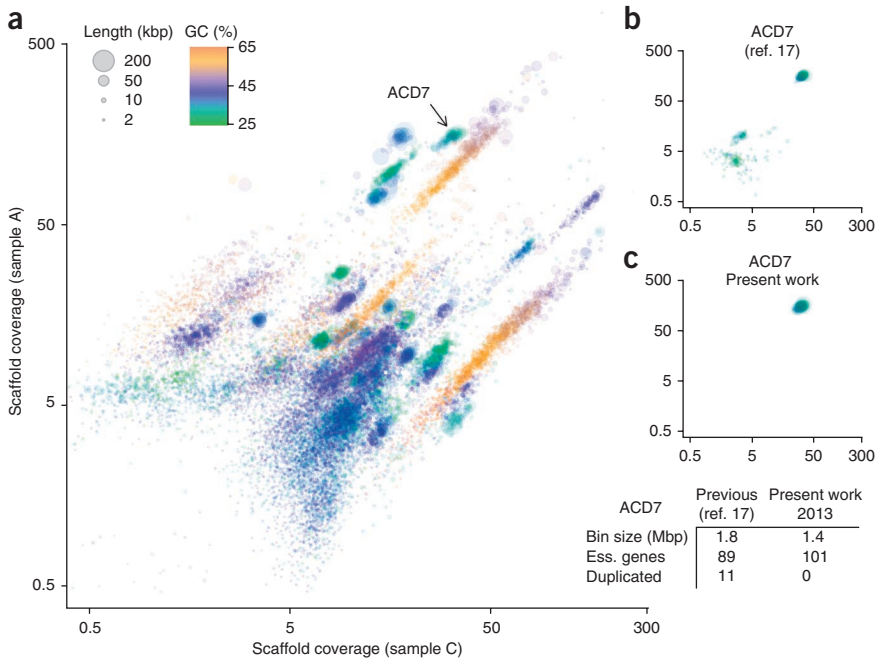

Figure 3 Reanalysis of published metagenomes[17] using the differential coverage approach. (**a**) Sequence composition–independent binning using metagenome coverage of two samples, A and C. All circles represent scaffolds, scaled by the square root of their length and colored by GC content. Only scaffolds >2 kbp are shown for consistency with the original study. Clusters of scaffolds represent putative genome bins. (**b**) Coverage analysis of the scaffolds in the genome bin ACD7 in which ESOM was used for primary binning[17]. (**c**) Primary binning by differential coverage improved genome completeness (101 versus 89 essential genes) and removed non-target scaffolds from closely related populations (0 versus 11 duplicated genes and no low-coverage contamination). Ess., essential.

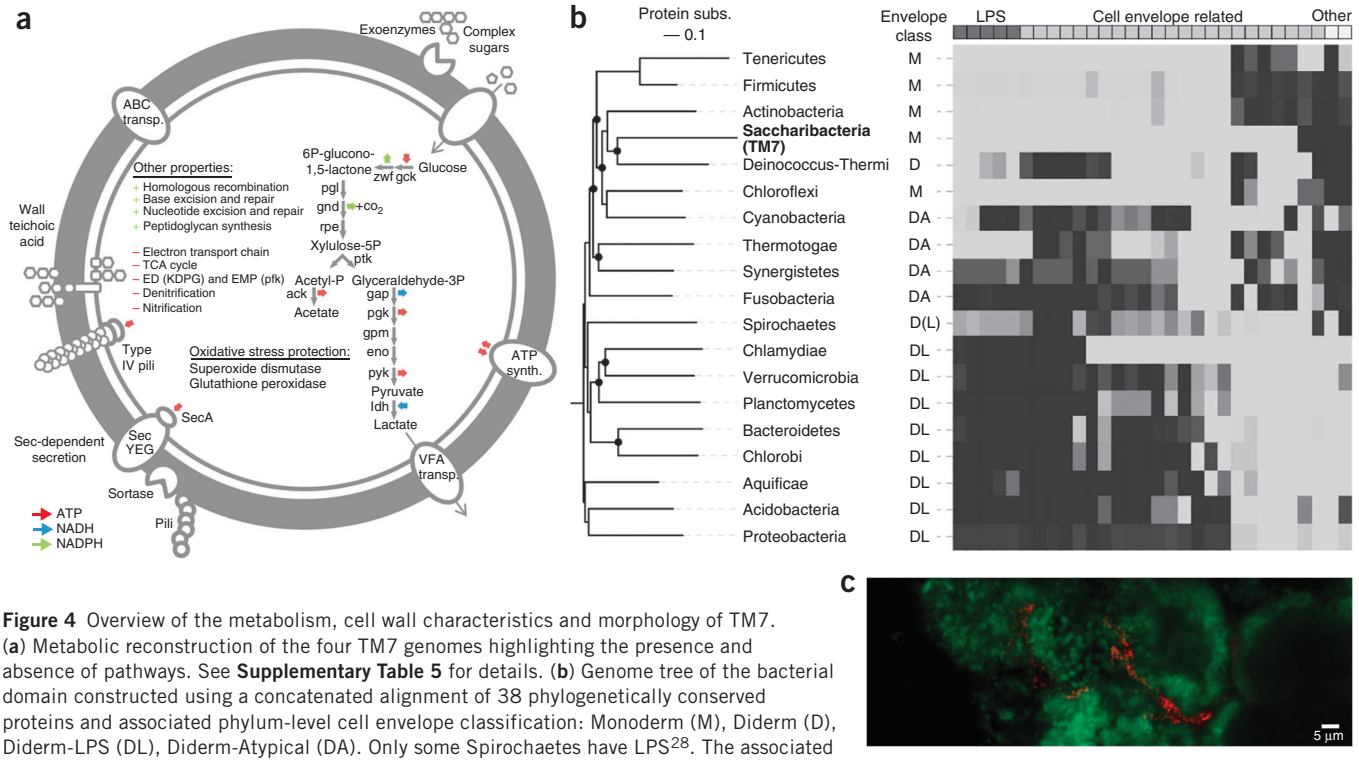

Figure 4 Overview of the metabolism, cell wall characteristics and morphology of TM7. (**a**) Metabolic reconstruction of the four TM7 genomes highlighting the presence and absence of pathways. See **Supplementary Table 5** for details. (**b**) Genome tree of the bacterial domain constructed using a concatenated alignment of 38 phylogenetically conserved proteins and associated phylum-level cell envelope classification: Monoderm (M), Diderm (D), Diderm-LPS (DL), Diderm-Atypical (DA). Only some Spirochaetes have LPS[28]. The associated heat map shows protein families substantially enriched (black) or depleted in archetypal monoderm lineages (Actinobacteria and Firmicutes) relative to an archetypal diderm lineage (Proteobacteria), most of which have known roles in cell envelope biosynthesis. Black dots in the genome tree represents branches with ≥75% bootstrap support. (**c**) FISH micrographs of TM7 (red) cells showing coccus morphology with a size of ~0.7 µm in diameter. The images show that they are embedded in flocs and confirm they are in low abundance.

An early ultrastructural investigation of a filamentous TM7 morphotype indicated a monoderm cell envelope[27]. Our set of near-complete TM7 genomes support this observation, as genes necessary for a typical diderm (two-membrane, typically Gram-negative) cell envelope are absent, including those for production of lipopolysaccharides and diderm-specific outer membrane assembly and secretion[33]. Also, several gene families specific to monoderms were identified including sortases that facilitate covalent attachment of proteins to the peptidoglycan layer[34] and genes involved in production and incorporation of teichoic acid in the cell wall[35] (**Fig. 4a** and **Supplementary Table 5**). Phylogenetically, the TM7 genomes cluster with the other recognized monoderm phyla, namely the Actinobacteria, Firmicutes, Tenericutes and Chloroflexi (**Fig. 4b**) as well as atypical diderm lineages such as the Thermi (*Deinococcus-Thermus*) and Cyanobacteria[36], which collectively places TM7 in the proposed higher-level grouping, the Terrabacteria[37]. This finding lends further weight to the hypothesis of a common monoderm ancestor in the bacterial domain[38]. We propose the following names for the TM7 phylum and complete TM7 genome, AAU-TM7-i (**Table 1**), respectively:

"*Candidatus* Saccharibacteria" phyl. nov.
"*Candidatus* Saccharimonas aalborgensis" gen. et sp. nov.

Etymology. Saccharimonas (Latin noun): sugar; aalborgensis (Latin verb/name): from Aalborg City. The name alludes to bacteria found in the city of Aalborg that consume primarily sugar compounds.

Locality. The bacteria were grown in an activated sludge bioreactor at Aalborg University.

Diagnosis. "*Candidatus* Saccharimonas aalborgensis" has a coccus morphology with a size of ≈0.7 µm in diameter. It has obligate fermentative metabolism, fermenting glucose and other sugars, and produces lactate. It has a monoderm cell envelope and stains Gram-positive.

DISCUSSION

The addition of complete Saccharibacteria (TM7) genomes to the tree of life adds an important chapter to the story of microbial evolution. However, numerous phylum-level lineages remain to be explored at the level of genome sequences, and multiple representatives are required from each lineage to fully address important questions in the ecology and evolution of environmental communities as well as microbial communities associated with higher organisms. The differential coverage binning approach that we report here will greatly accelerate our ability to obtain high-quality genomes from candidate and other underrepresented or rare phyla. Populating the tree of life with genomes will provide the necessary basis to resolve many present scientific conflicts arising from a chronic undersampling of microbial lineages[1].

Deep metagenome sequencing augmented by sequencing multiple samples with varying relative abundance of identical target community members, whether arising from methodology (as in this study) or spatial or temporal[17] sampling, can be used on any microbial system from any environment, provided DNA can be extracted and sequenced. The approach we describe should enable reconstruction of population genomes from any species, provided that there is adequate sequencing depth (≈50 ×, according to simulations; see **Supplementary Notes** and **Supplementary Fig. 9**) and limited strain heterogeneity, which could compromise assembly of the target population genome. Differential-coverage binning, therefore, holds great

potential for providing sample-specific genome catalogs necessary for omics-based systems biology.

Continued improvements in sequencing throughput will enable access to increasingly lower abundance populations, whereas improvements in read length and quality will reduce the impact of co-assembly of closely related strains (strain heterogeneity) on the initial *de novo* assembly. These improvements will strengthen the value of differential coverage binning, which benefits from and scales well with increasing coverage, while potentially complicating sequence composition–dependent approaches applied directly to whole metagenomic data sets, which scale poorly with increasing data set size. We also note that increasing the number of metagenomes per habitat improves differential-coverage binning resolution and genome recovery (data not shown). We are presently using differential-coverage binning successfully with samples from permafrost, hindgut and fecal communities, anaerobic digesters and full-scale wastewater treatment plants and anticipate that such binning will become a standard part of the metagenomics toolkit.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Raw sequence data reported in this study have been submitted to the National Center for Biotechnology Information Sequence Read Archive under accessions SRX247688 (HP^-) and SRX206471 (HP^+).

This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession APMI00000000. The version described in this paper is the first version, APMI01000000.

The genome of *Candidatus* Saccharobacterium alaburgensis has been deposited in GenBank under the accession number CP005957. All scripts used to analyze the data are available at https://github.com/MadsAlbertsen/multi-metagenome.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

This study was funded by Aalborg University and the Danish Research Council for Strategic Research via the Centre "EcoDesign-MBR" and the Obelske Family foundation. P.H. was supported by a Discovery Outstanding Researcher Award (DORA) from the Australian Research Council, grant DP120103498 and G.W.T. was supported by a QEII fellowship from the Australian Research Council, grant DP1093175. We thank S. McIlroy and P. Larsen for assistance with FISH analyses, A.M. Saunders for 16S rRNA data generation and J.P. Euzeby for suggesting the new genus name.

AUTHOR CONTRIBUTIONS

M.A., experimental design, data analysis and manuscript; P.H., data analysis and manuscript; A.S., data analysis; K.L.N., sequencing; G.W.T., data analysis and manuscript; P.H.N., experimental design and manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Wu, D. *et al.* A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**, 1056–1060 (2009).
2. Hugenholtz, P. Exploring prokaryotic diversity in the genomic era. *Genome biology* **3**, S0003 (2002).
3. Achtman, M. & Wagner, M. Microbial diversity and the genetic nature of microbial species. *Nat. Rev. Microbiol.* **6**, 431–440 (2008).
4. Knittel, K. & Boetius, A. Anaerobic oxidation of methane: progress with an unknown process. *Annu. Rev. Microbiol.* **63**, 311–334 (2009).
5. Elinav, E. *et al.* NLRP6 inflammasome regulates colonic microbial ecology and risk for colitis. *Cell* **145**, 745–757 (2011).
6. Brinig, M.M., Lepp, P.W., Ouverney, C.C., Armitage, G.C. & Relman, D.A. Prevalence of bacteria of division TM7 in human subgingival plaque and their association with disease. *Appl. Environ. Microbiol.* **69**, 1687–1694 (2003).
7. 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).
8. Kuehbacher, T. *et al.* Intestinal TM7 bacterial phylogenies in active inflammatory bowel disease. *J. Med. Microbiol.* **57**, 1569–1576 (2008).
9. Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K. & Hugenholtz, P. A bioinformatician's guide to metagenomics. *Microbiol. Mol. Biol. Rev.* **72**, 557–578 (2008).
10. Tyson, G.W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
11. García Martín, H. *et al.* Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat. Biotechnol.* **24**, 1263–1269 (2006).
12. Tringe, S.G. *et al.* Comparative metagenomics of microbial communities. *Science* **308**, 554–557 (2005).
13. Rodrigue, S. *et al.* Whole genome amplification and *de novo* assembly of single bacterial cells. *PLoS ONE* **4**, e6864 (2009).
14. Lasken, R.S. Genomic sequencing of uncultured microorganisms from single cells. *Nat. Rev. Microbiol.* **10**, 631–640 (2012).
15. Luo, C., Tsementzi, D., Kyrpides, N.C. & Konstantinidis, K.T. Individual genome assembly from complex community short-read metagenomic datasets. *ISME J.* **6**, 898–901 (2012).
16. Mande, S.S., Mohammed, M.H. & Ghosh, T.S. Classification of metagenomic sequences: methods and challenges. *Brief. Bioinform.* **13**, 669–681 (2012).
17. Wrighton, K.C. *et al.* Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* **337**, 1661–1665 (2012).
18. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
19. Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.-L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**, 431–432 (2011).
20. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
21. Hess, M. *et al.* Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331**, 463–467 (2011).
22. Dupont, C.L. *et al.* Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J.* **6**, 1186–1199 (2012).
23. Vezzi, F., Narzisi, G. & Mishra, B. Reevaluating assembly evaluations with feature response curves: GAGE and assemblathons. *PLoS ONE* **7**, e52210 (2012).
24. Podar, M. *et al.* Targeted access to the genomes of low-abundance organisms in complex microbial communities. *Appl. Environ. Microbiol.* **73**, 3205–3214 (2007).
25. Nielsen, P.H., Saunders, A.M., Hansen, A.A., Larsen, P. & Nielsen, J.L. Microbial communities involved in enhanced biological phosphorus removal from wastewater-a model system in environmental biotechnology. *Curr. Opin. Biotechnol.* **23**, 452–459 (2012).
26. Luo, C., Xie, S., Sun, W., Li, X. & Cupples, A.M. Identification of a novel toluene-degrading bacterium from the candidate phylum TM7, as determined by DNA stable isotope probing. *Appl. Environ. Microbiol.* **75**, 4644–4647 (2009).
27. Hugenholtz, P., Tyson, G.W., Webb, R.I., Wagner, A.M. & Blackall, L.L. Investigation of candidate division TM7, a recently recognized major lineage of the domain Bacteria with no known pure-culture representatives. *Appl. Environ. Microbiol.* **67**, 411–419 (2001).
28. Sutcliffe, I.C. A phylum level perspective on bacterial cell envelope architecture. *Trends Microbiol.* **18**, 464–470 (2010).
29. McCutcheon, J.P. & Moran, N.A. Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* **10**, 13–26 (2012).
30. Baker, B.J. *et al.* Enigmatic, ultrasmall, uncultivated Archaea. *Proc. Natl. Acad. Sci. USA* **107**, 8806–8811 (2010).
31. Thomsen, T.R., Kjellerup, B.V., Nielsen, J.L., Hugenholtz, P. & Nielsen, P.H. *In situ* studies of the phylogeny and physiology of filamentous bacteria with attached growth. *Environ. Microbiol.* **4**, 383–391 (2002).
32. Mandlik, A., Swierczynski, A., Das, A. & Ton-That, H. Pili in Gram-positive bacteria: assembly, involvement in colonization and biofilm development. *Trends Microbiol.* **16**, 33–40 (2008).
33. Sutcliffe, I.C. Cell envelope architecture in the Chloroflexi: a shifting frontline in a phylogenetic turf war. *Environ. Microbiol.* **13**, 279–282 (2011).
34. Schneewind, O. & Missiakas, D.M. Protein secretion and surface display in Gram-positive bacteria. *Phil. Trans. R. Soc. Lond. B* **367**, 1123–1139 (2012).
35. Weidenmaier, C. & Peschel, A. Teichoic acids and related cell-wall glycopolymers in Gram-positive physiology and host interactions. *Nat. Rev. Microbiol.* **6**, 276–287 (2008).
36. Hoiczyk, E. & Hansel, A. Cyanobacterial cell walls: news from an unusual prokaryotic envelope. *J. Bacteriol.* **182**, 1191–1199 (2000).
37. 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).
38. Gupta, R.S. Origin of diderm (Gram-negative) bacteria: antibiotic selection pressure rather than endosymbiosis likely led to the evolution of bacterial cells with two membranes. *Antonie van Leeuwenhoek* **100**, 171–182 (2011).

ONLINE METHODS

Biological sample. An activated sludge sample was collected from Ejby Mølle wastewater treatment plant, Odense, Denmark (55.398487, 10.420596) and used for inoculation of a 5-l sequencing batch reactor fed with acetate as the sole carbon source. The reactor was run in a sequential mode with 2.5 h anoxic conditions with acetate feed the first 5 min followed by 2.5 h aeration without feed. The reactor was well mixed. Subsequently, a 1-h phase without mixing and aeration allowed the biomass to settle. After each cycle a part of the water was withdrawn. Samples were continuously taken for DNA extractions (stored at −80 °C) and fixed for FISH investigations. Samples from day 39 were used for the metagenomic sequencing presented in this paper.

Fluorescence *in situ* hybridization. TM7 genome–specific FISH probes were designed in ARB v 5.2 (ref. 39) using the TM7 genomes and TM7 sequences available in the current SILVA 16S rRNA database (SSU Parc v108 (ref. 40)). In order to maximize the specificity of the probes we designed competitor and helper probes (**Supplementary Table 6**). All probes were hybridized overnight to improve the specific probe signal, while no increase in signal intensity was seen in the non-EUB probe (negative control).

FISH was conducted as described[41] using the probes listed in **Supplementary Tables 6** and **7** and results can be seen in **Supplementary Figures 7** and **11**. Images were recorded using a confocal laser scanning microscope (LSM 510 META; Carl Zeiss) equipped with an Ar-ion laser (458 and 488 nm) and two HeNe lasers (543 and 633 nm) and subsequently processed using ImageJ (http://rsb.info.nih.gov/ij/).

DNA extraction. DNA was extracted from 1-ml aliquots of the day 39 reactor sample using two different methods to obtain different relative abundance of the component microbial populations. The first extraction was performed using the FastDNA spin kit for soil (MP Biomedicals) according to the manufacturer's instructions (named: Hot Phenol minus: HP^-).

The second extraction method involved an initial incubation for 3 min in 90 °C phenol before proceeding with the FastDNA spin kit for soil (named: HP^+). Briefly, a 1-ml aliquot was centrifuged at 13,000 r.p.m. for 5 min and the supernatant discarded. The pellet was redissolved in 250 µl phosphate buffer (included in the FastDNA kit) and transferred to the FastDNA bead beating tubes, after which 750 µl preheated phenol (Biological grade, pH 8, 0.1M EDTA, Sigma-Aldrich) was added and the sample incubated for 3 min at 90 °C with occasional shaking. The sample was bead beaten using the manufacturer's instructions and afterward centrifuged at 13,000 r.p.m. for 10 min to separate the phenol phase from the top liquid phase containing the DNA. The top phase was extracted and the DNA further purified using the FastDNA kit according to the manufacturer's instructions. DNA concentrations were measured using Qubit (Life technologies) and DNA integrity evaluated using gel electrophoresis.

V4 16S rRNA gene amplicon sequencing. The following samples were selected for amplicon sequencing; day 0, 6, 14, 28 and 39. 16S rRNA gene amplicon sequencing was conducted with the primer pair 515F and 806R adapted for use on the Illumina platform[42] with the following modifications. The PCR amplifications were carried out with Platinum High Fidelity (6 × proofreading) Taq Polymerase (Invitrogen) in two steps: an initial 20-cycle PCR with the native primers (515F_Pcr1 and 806R_Pcr1) followed by a second 7 cycles of PCR with fusion primers (515F_Pcr2 and 806R_Pcr2) containing the 8-nt barcode and Illumina adaptor sequences (**Supplementary Table 8**). Two-step PCR was done to avoid barcode bias[43]. PCR products were purified using Agencourt AMPure XP (Beckman Coulter) and the DNA concentrations measured using the QuantIT kit (Molecular Probes). Barcoded amplicons were pooled in equimolar amounts and paired-end sequenced (2 × 150 bp) on an Illumina HiSeq2000 (Illumina Inc.). The 515F_Seq, 806R_Seq and 515_Index primers were spiked into the normal TruSeq SBS kit v.3-HS sequencing kits (Illumina Inc.) in a final concentration of 0.5 µM. This allowed sequencing of 16S rRNA gene amplicon samples with shotgun metagenomic samples on the Illumina HiSeq2000, thereby also circumventing phasing issues of low-complexity samples. We designed 12 separate barcodes, which can be spiked into all lanes on the Illumina HiSeq2000 allowing sequencing of 12 × 8 samples per flow cell.

16S rRNA gene amplicon analysis. The 150 bp paired-end reads were merged using pandaseq v.2.0 (ref. 44) with the following parameters: -N -o 15 -l 245 -L 260. After merging, 118 k–230 k amplicons were left per sample. Sequences seen less than 10 times were removed and all libraries were sub-sampled to the size of the smallest library (41 k sequences) and formatted for direct use in QIIME[45] using pandaseq.to.qiime.pl. QIIME v1.5.0 was used to cluster the sequences into OTUs (97%), classify sequences, summarize the data, and calculate alpha and beta diversity.

Metagenome sequencing. Samples were prepared for sequencing using TruSeq DNA Sample Preparation Kits v2 with 2 µg of DNA following the manufacturer's instructions with nebulizer fragmentation. Library DNA concentration was measured using the QuantIT kit (Molecular Probes) and paired-end sequenced (2 × 150 bp) on an Illumina HiSeq2000 using the TruSeq PE Cluster Kit v3-cBot-HS and TruSeq SBS kit v.3-HS sequencing kit (Illumina Inc.). A single HiSeq2000 lane was allocated to the HP^- metagenome library and 0.5 lane to the HP^+ metagenome library.

Metagenome data analysis strategy. The analysis strategy used to extract near-complete genomes from metagenomes is detailed in **Figure 2**. The individual steps are described in more details below. Numbers in headings refer to **Figure 2**. All scripts used in the analysis are in **Supplementary Data Set 1** and are available at https://github.com/MadsAlbertsen/multi-metagenome along with example data sets and a detailed step-by-step guide. R (http://www.r-project.org) was used to visualize and extract the initial genome bins.

Read trimming and *de novo* metagenome assembly [4]. Metagenome reads in FASTQ format were imported to CLC Genomics Workbench v. 5.1 (CLC Bio) and trimmed using a minimum phred score of 20, a minimum length of 50 bp, allowing no ambiguous nucleotides and trimming off Illumina sequencing adaptors if found. The trimmed metagenome reads were assembled using CLC's *de novo* assembly algorithm, using a k-mer of 63 and a minimum scaffold length of 1 kbp. The two metagenomes were assembled independently. However, because fewer reads were sequenced and there was a skewed species abundance toward few organisms in the HP^+ metagenome, only the HP^- metagenome scaffolds were used for extraction of full genomes.

Scaffold coverage [5]. Reads were mapped to scaffolds using CLC's map reads to reference algorithm with a minimum similarity of 95% over 100% of the read length. The relative metagenome abundance of each genome bin was calculated as the percentage of metagenome reads mapping to the individual genomes compared to the total number of metagenome reads.

Tetranucleotide binning [6]. Principal component analysis of tetranucleotide frequencies was used when several species were present in the same coverage-defined bin. Tetranucleotide frequencies were calculated using calc.kmerfreq.pl. A principal coordinate analysis of the resulting frequencies was conducted using the vegan package[46] in R. Often only a few species were present in each coverage-defined bin and could easily be separated by their tetranucleotide frequency patterns. To distinguish genome bins we used the combination of principal components that gave most resolution between the genomes.

GC content [7]. GC content of all assembled scaffolds was calculated using calc.gc.pl.

Identification of conserved marker genes [8]. Open reading frames were predicted in the assembled scaffolds using the metagenome version of Prodigal[47]. A set of 107 HMMs of essential single-copy genes[22] were searched against the predicted open reading frames using HMMER3 (http://hmmer.janelia.org/) with the default settings, except the trusted cutoff was used (-cut_tc). The identified proteins were taxonomic classified using BLASTP against the RefSeq (version 52) protein database with a maximum e-value cutoff of 1e-5. MEGAN[48] was used to extract class level taxonomic assignments from the BLAST .xml output file.

Sequence composition-independent binning [9]. Sequence composition–independent binning was facilitated by the use of two metagenomes generated

from the same sample, but using different DNA extraction protocols. Differences in species-specific DNA extraction efficiency between the two protocols resulted in differing relative abundance of the populations in the two metagenomes. Distinct groupings were identified by independently mapping the reads from each metagenome to the assembled scaffolds (calculating coverage) and plotting the two coverages against one another in R. The groupings represent scaffolds with similar coverage in both metagenomes, indicating that they are from the same species. Initial bin selection was facilitated by coloring scaffolds according to their taxonomic affiliation as determined using a suite of 107 single-copy marker genes[22]. A step-by-step guide is available on https://github.com/MadsAlbertsen/multi-metagenome that document how to integrate all data and perform the binning in R.

Tracking and visualization of paired-end connections between scaffolds [10]. In order to include repeat regions (for example, multi-copy rRNA genes) and small scaffolds not included in the initial bins, we determined paired-end connections between ends of scaffolds using cytoscapeviz.pl with the settings -c -f 2 -m 3000 using a SAM file of the read mapping to scaffolds. The script tracks the mapping of each paired-end read and outputs a file indicating paired-end connected scaffolds and the number of connections for direct use in Cytoscape v.2.8.1 (ref. 19). In addition, Cytoscape attribute files are generated with coverage and length information for each scaffold.

Re-assembly [11]. All reads mapping to scaffolds of a particular bin were extracted using CLC genomics workbench and used to extract all associated paired-end reads using extract.fasta.pe.reads.using.single.pl. The extracted reads were used for a new *de novo* assembly using velvet[18] v.1.1.04 with a k-mer of 99. The high k-mer value was chosen in an effort to reduce the impact of micro-heterogeneity on the assemblies (CLC has a maximum k-mer size of 64). The expected coverage and coverage cutoff parameters were adjusted individually by running multiple assemblies. Only genome bins with a coverage of >100 were reassembled.

Finishing [12]. Reads were mapped to scaffolds after re-assembly using CLC and the resulting mapping exported in SAM format. The assembly was visualized using Circos[20] with data generated by circosviz.pl. The script generates all data used in **Supplementary Figure 3** from a SAM file and a FASTA file of the associated scaffolds. In addition, the assembly statistics generated by the reference-free validation tool FRCbam[23] can be integrated in the Circos analysis. The visualization was used as a starting point for manual inspection of the assembly. The most common error was integration of fragments of repeat elements in the ends of scaffolds, which could easily be resolved manually.

In order to close gaps (Ns) within scaffolds, reads were mapped to scaffolds using CLCs map reads to reference function with 95% similarity over 50% of the read length. The 150-bp reads meant that each gap had reads mapping on each side, with a large part of the read extending through the gap area. By aligning reads from each side of the gap region most gaps could be resolved. Many gaps were due to low-level micro-diversity where the consensus could be resolved easily. Other examples were overextension of the contigs flanking the gap, resulting in identical regions on each side of the gap. Gaps were also seen in very low-complexity areas (for example, extreme %GC) where more indels seemed present. By manual overlapping the reads, most gaps could be easily closed. A few rare cases involved multiple small repeat regions.

Identification of rRNA genes. The current Greengenes 16S database[49] (downloaded 2012/07/01) and SILVA LSUref 108 rRNA database[40] were individually clustered at 90% similarity using uclust[50]. The reduced databases were used in a BLASTN search against the assembled scaffolds with a maximum e-value of 1e-5. The longest BLASTN hit (≥300 bp) for each scaffold was extracted using extract.long.hits.from.blast.pl and subsequently taxonomically classified using SINA online[51].

Genome analysis. The four assembled TM7 genomes were annotated using the Integrated Microbial Genomes - Expert Review[52] (IMG/ER). Other genome bins were automatically annotated using the RAST server[53].

Essential single-copy gene analysis. Completeness and potential contamination of each genome bin was evaluated using HMMs of 107 essential single-copy genes conserved in 95% of all bacteria[22]. In the case of TM7, all missing genes were manually investigated by BLASTP and afterwards using BLASTN to rule out that they had been missed by the gene-calling programs.

The script img.matrix.pl was used to investigate the phylum level distribution of the 107 essential single-copy genes. The script uses a list of PFAMs, COGs or TIGRFAMs and a list of genomes with associated metadata to search through the IMG annotations of the selected genomes. The output can be condensed using the metadata in the genome list file. We downloaded the full IMG database (release 3.5) from the IMG ftp server containing all genomes in IMG and their annotations (for example, PFAMs and TIGRFAMs). In addition, a list of all finished bacterial genomes and associated metadata (for example, taxonomic affiliation and genome size) was obtained though IMG. The genome list was inspected to remove genomes with an unusual small number of annotations. In this case we grouped the data at phylum level, except for Proteobacteria which were grouped at class level. The resulting abundance matrix was clustered and visualized using the heat map function in R (**Supplementary Fig. 4**).

Genome size estimation of publicly available TM7 single-cell genomes. To investigate if the small genome size of TM7 identified in this study is a general trait of the phylum we revisited the four publicly available single-cell TM7 genomes[7,24]: TM7a (IMG/M: 2004247010), TM7b (IMG/M: 2005503000), TM7c (IMG/M: 2004000001) and TM7_GTL1 (NZ_AAXS01000001-132). The 100 essential single-copy marker genes identified in TM7 were used as an indicator of genome completeness. However, owing to multiple copies of single-copy genes in all single-cell genomes, we introduced an initial binning step using canonical correspondence analysis (cca) of tetranucleotide frequency of scaffolds >1 kbp through the vegan package in R. In addition, we predicted proteins in all scaffolds >1 kbp and taxonomically classified the scaffolds using BLASTP against RefSeq (version 52) with the four refined TM7 genomes from this study included. MEGAN[48] was used to retrieve lowest common ancestor taxonomic classification of each protein at phylum level. Only scaffolds with ≥2 proteins taxonomically classified were used as an indicator for taxonomic affiliation. The taxonomic assignment was overlayed on the cca plot to indicate distinct bins. Only the TM7c single-cell genome seemed useable for an estimate of the TM7 genome size. TM7a was heavily contaminated with other species and TM7b contained only 60 kbp of sequence. TM7_GTL1 originated from a mix of five single cells with identical 16S rRNA sequences, however, the tetranucleotide frequency and single-copy gene analysis indicated that two closely related strains were present. The genome size was estimated by extrapolating the percentage of essential single-copy genes to the fraction of the genome that was recovered (assumes that the marker genes are independently distributed throughout the genome, which they are not).

Comparative analysis of TM7 genomes. Orthologous proteins between the four TM7 genomes were identified using pairwise bi-directional best hit BLASTP searches with a minimum identity of 30%. To extend the pairwise comparison to all TM7 genomes, we visualized the BLASTP outputs in Cytoscape[19] v.2.8.1, which allowed easy identification of orthologous proteins between all four TM7 genomes. All proteins in the TM7 genomes were used as nodes (four different shapes) and the percent identity between the proteins as edges (derived from the bi-directional best hit BLASTP searches). The results are summarized in **Supplementary Figure 12**.

Full genome tree. FASTA files of finished bacterial genomes were downloaded from IMG (release 3.5) and processed using PhyloSift (A. Darling, H. Bik, G. Jospin, J.A. Eisen, manuscript in preparation), downloaded from https://github.com/gjospin/PhyloSift (commit: 060143a), to produce HMMer (HMMER 3.0; http://hmmer.janelia.org/) alignments of 38 conserved proteins per genome (**Supplementary Table 9**). Aligned protein sequences from each genome were concatenated and combined to form a multiple alignment. This alignment was masked with Gblocks[54] using default parameters except that a conserved position was not allowed to have gaps in more than half of the sequences. A phylogenetic tree was produced from the masked alignment using

doi:10.1038/nbt.2579

FastTree[55] v.2.1.3 with default parameters. PHYLIPs SEQBOOT module[56] was used to generate 100 resampled alignments and FastTree was used to analyze the resampled alignments (-n 100) with the original tree as starting tree for each resampling (-intree1). CompareToBootstrap.pl included in FastTree was used to compare the original tree to the resampled trees and generate bootstrap values.

Enriched gene analysis. The script img.matrix.pl (see prior description) was used to identify protein families (through PFAMs[57]) substantially enriched or depleted in archetypal monoderm lineages (Actinobacteria and Firmicutes) relative to an archetypal diderm lineage (Proteobacteria). A PFAM was classified as enriched in for example, diderms if it was seen in >80% of Proteobacteria and <20% of Actinobacteria and Firmicutes. The identified PFAMs were used to make a phylum level percentage prevalence profile. Only phyla with at least four finished genomes were included (**Supplementary Table 10**).

To investigate putative functions of PFAMs with poor functional description the COG[58] equivalent (if one could be identified) was used to search the string database for putative interactions[59]. Most PFAMs with poor functional description were found to have interactions with other known cell envelope–related proteins (**Supplementary Fig. 13**).

The resulting matrix was combined with the full genome tree (described above) using the heatmap function of iToL[60] and edited using Inkscape (http://inkscape.org).

39. Ludwig, W. *et al.* ARB: a software environment for sequence data. *Nucleic Acids Res.* **32**, 1363–1371 (2004).
40. Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **35**, 7188–7196 (2007).
41. Nielsen, P.H. *et al.* A conceptual ecosystem model of microbial communities in enhanced biological phosphorus removal plants. *Water Res.* **44**, 5070–5088 (2010).
42. Caporaso, J. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. USA* **108**, 4516–4522 (2011).
43. Berry, D., Ben Mahfoudh, K., Wagner, M. & Loy, A. Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. *Appl. Environ. Microbiol.* **77**, 7846–7849 (2011).
44. Masella, A.P., Bartram, A.K., Truszkowski, J.M., Brown, D.G. & Neufeld, J.D. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* **13**, 31 (2012).
45. Caporaso, J.G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010).
46. Oksanen, J. *et al.* Vegan: Community Ecology Package. R package version 2.0–5 (2011).
47. Hyatt, D., LoCascio, P.F., Hauser, L.J. & Uberbacher, E.C. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* **28**, 2223–2230 (2012).
48. Huson, D.H., Mitra, S., Ruscheweyh, H.-J., Weber, N. & Schuster, S.C. Integrative analysis of environmental sequences using MEGAN4. *Genome Res.* **21**, 1552–1560 (2011).
49. 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).
50. Edgar, R.C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
51. Pruesse, E., Peplies, J. & Glöckner, F.O. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**, 1823–1829 (2012).
52. Markowitz, V.M. *et al.* IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics* **25**, 2271–2278 (2009).
53. Aziz, R.K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).
54. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).
55. Price, M.N., Dehal, P.S. & Arkin, A.P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
56. Felsenstein, J. PHYLIP - Phylogeny inference package (version 3.2). *Cladistics* **5**, 164–166 (1989).
57. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **40**, D290–D301 (2012).
58. Tatusov, R.L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
59. Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* **39**, D561–D568 (2011).
60. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* **39**, W475–W478 (2011).