UC Berkeley

UC Berkeley Previously Published Works

Title

Unusual biology across a group comprising more than 15% of domain Bacteria.

Permalink

https://escholarship.org/uc/item/9ks7v8nv

Journal

Nature, 523(7559)

ISSN

0028-0836

Authors

Brown, Christopher T
Hug, Laura A
Thomas, Brian C
et al.

Publication Date

2015-07-01

DOI

10.1038/nature14486

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Unusual biology across a group comprising more than 15% of domain Bacteria

Christopher T. Brown, Laura A. Hug, Brian C. Thomas, Itai Sharon, Cindy J. Castelle, Andrea Singh, Michael J. Wilkins, Kelly C. Wrighton, Kenneth H. Williams, & Jillian F. Banfield

Abstract

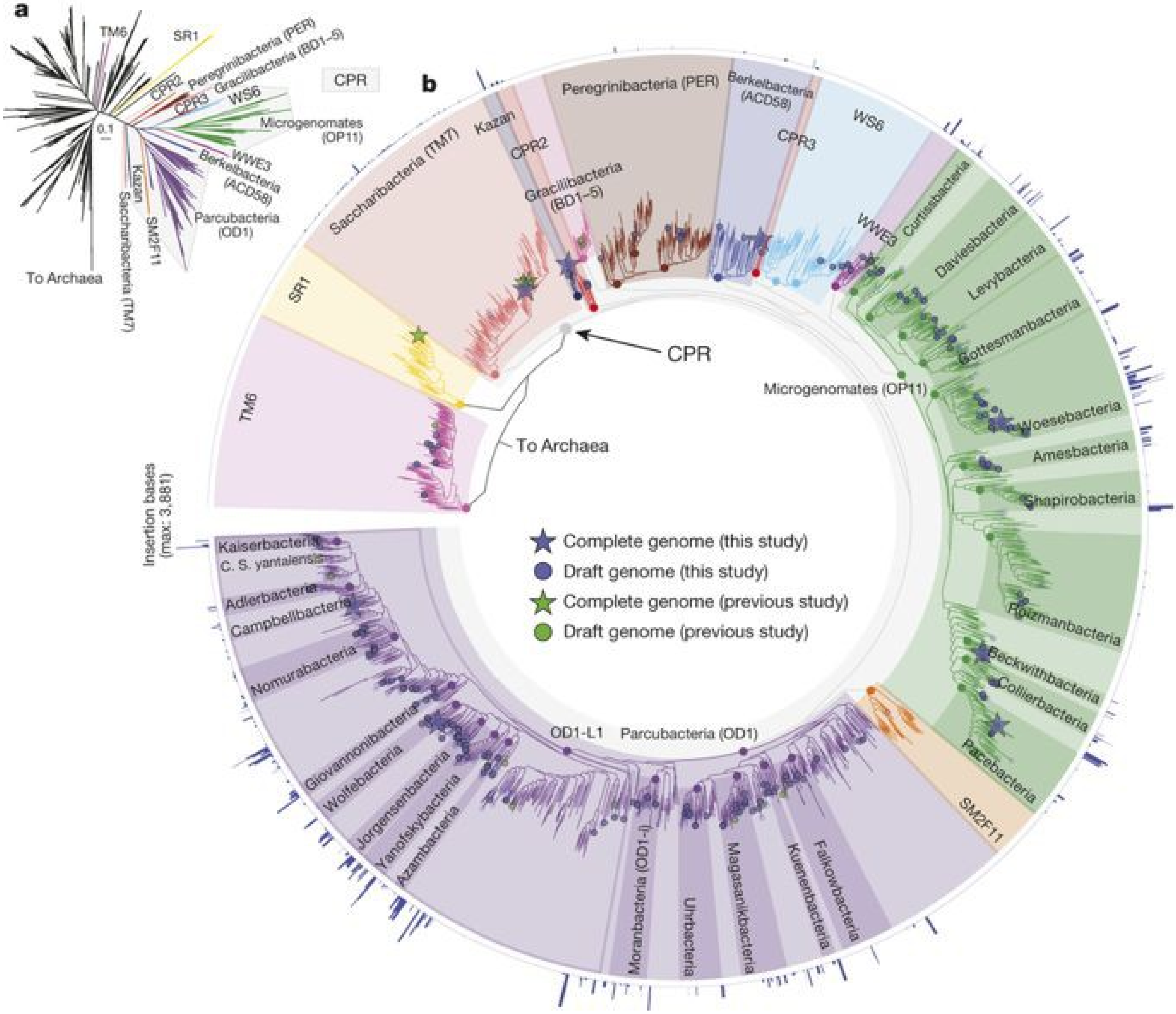
A prominent feature of the bacterial domain is a radiation of major lineages that are defined as candidate phyla because they lack isolated representatives. Bacteria from these phyla occur in diverse environments[1] and are thought to mediate carbon and hydrogen cycles[2]. Genomic analyses of a few representatives suggested that metabolic limitations have prevented their cultivation[2,3,4,5,6]. Here we reconstructed 8 complete and 789 draft genomes from bacteria representing >35 phyla and documented features that consistently distinguish these organisms from other bacteria. We infer that this group, which may comprise >15% of the bacterial domain, has shared evolutionary history, and describe it as the candidate phyla radiation (CPR). All CPR genomes are small and most lack numerous biosynthetic pathways. Owing to divergent 16S ribosomal RNA (rRNA) gene sequences, 50–100% of organisms sampled from specific phyla would evade detection in typical cultivation-independent surveys. CPR organisms often have self-splicing introns and proteins encoded within their rRNA genes, a feature rarely reported in bacteria. Furthermore, they have unusual ribosome compositions. All are missing a ribosomal protein often absent in symbionts, and specific lineages are missing ribosomal proteins and biogenesis factors considered universal in bacteria. This implies different ribosome structures and biogenesis mechanisms, and underlines unusual biology across a large part of the bacterial domain.

Main

We sampled microbial communities from an aquifer adjacent to the Colorado River near the town of Rifle, Colorado, USA in 2011. Groundwater was filtered through a 1.2 µm pre-filter and cells were collected on serial 0.2 and 0.1 µm filters (Extended Data Fig. 1). Post-0.2 µm filtrates were targeted because CPR bacteria were predicted to have ultra-small cells on the basis of their small genomes[2]. Groundwater was sampled before and during an acetate amendment experiment that reproduced conditions that generated the first genomes from CPR bacteria[2,4,7,8] (Supplementary Table 1). Total DNA and RNA were extracted from filters and sequenced. We obtained 224 gigabase pairs (Gb) of paired-end metagenomic sequence from 12 samples (150 bp reads, 6 time points, 0.2 and 0.1 µm filters; Supplementary Table 2). Sequence assembly generated 3.9 Gb of contiguous sequences ≥5 kb. We also obtained 181 Gb of metatranscriptomic sequence from six samples (50 bp reads, 0.2 µm filters).

Assembled scaffolds were binned into genomes on the basis of their GC content, DNA sequence coverage, abundance pattern across samples, and taxonomic affiliation (binning was validated with a tetranucleotide sequence signature method; Extended Data Fig. 2). Overall, we reconstructed >1,750 genome bins from microbial community sequence data. Here, we focus on genomes from CPR bacteria and TM6, which represented >60% of bins. Included in our analyses of the CPR are members of the Parcubacteria (OD1), Microgenomates (OP11), WWE3, Berkelbacteria (ACD58), Saccharibacteria (TM7), WS6, Peregrinibacteria (PER), and Kazan phyla, in addition to previously unrecognized lineages (CPR1–3; Fig. 1). In total, 789 draft-quality (≥50% complete) genomes were reconstructed (Table 1). We manually curated eight genomes to completion: the first three from Microgenomates, two from Parcubacteria, one each from Kazan and Berkelbacteria, and an additional genome from Saccharibacteria. All complete and draft genomes are small and most are <1 Mb in length (Supplementary Tables 3 and 4).

Figure 1: Phylogeny and genomic sampling of the CPR.



a, **b**, Subsets of a maximum-likelihood 16S rRNA gene phylogeny (Supplementary Fig. 1) showing the CPR, a monophyletic radiation of candidate phyla (**a**), and genomic sampling of candidate phyla (**b**). Proposed names for phyla within the superphyla Parcubacteria and Microgenomates are explained in Extended Data Table 1. Many CPR 16S rRNA genes encode insertions (length shown by blue bars, combined length for multiple insertions).

Full size image

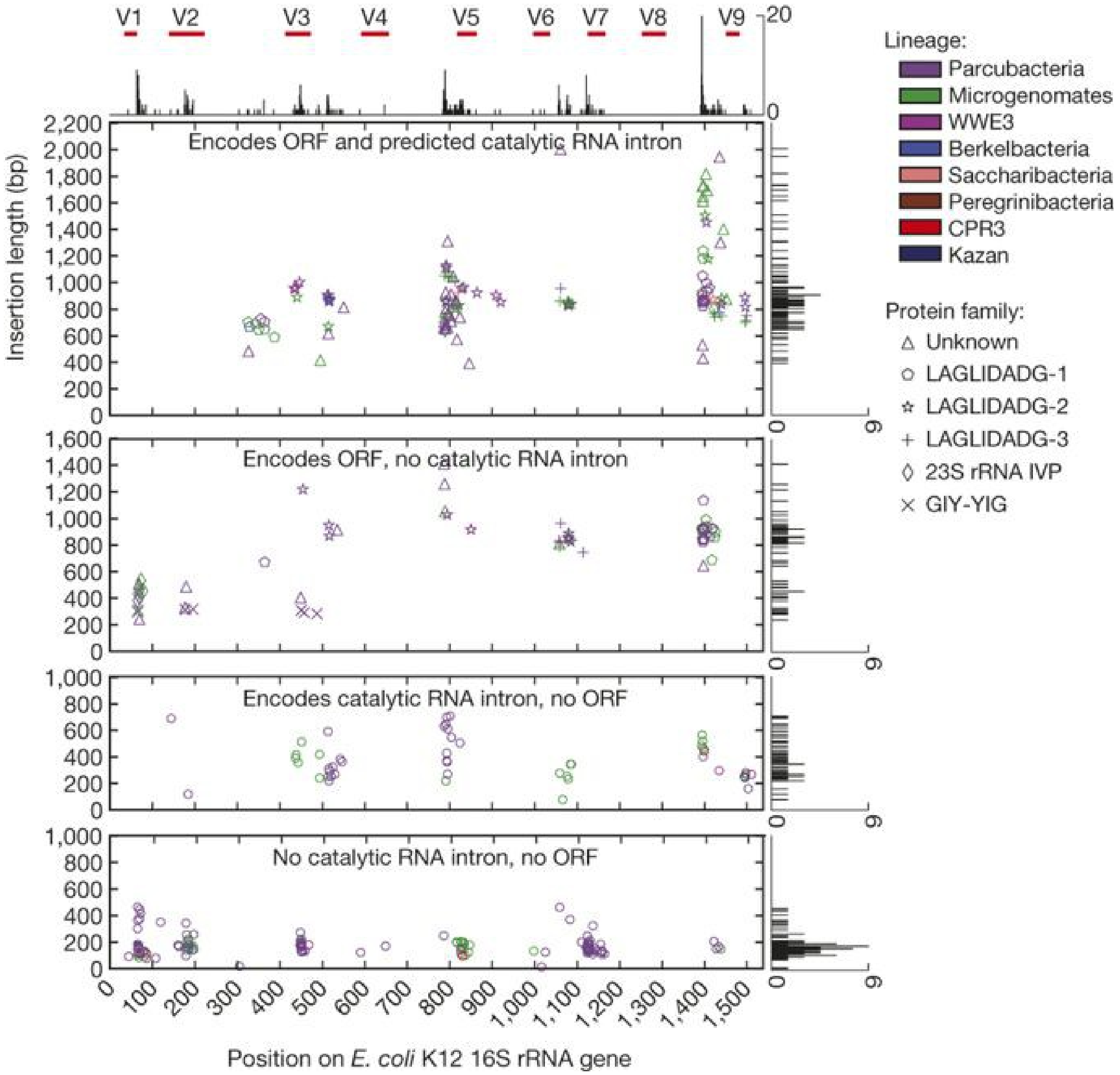
Download PowerPoint slide

Table 1: Genomes from candidate phyla bacteria

Full size table

In total, 1,543 bacterial 16S rRNA genes ≥800 bp were assembled and curated to eliminate assembly errors (713 sequences clustered at 97% identity; Supplementary Data 1). Relative abundance measurements show enrichment of CPR organisms in small-cell filtrates, suggesting that they have ultra-small cells (Extended Data Fig. 3). This finding is supported by a recent microscopy study[8]. Surprisingly, 31% of 16S rRNA genes encoded a large (≥10 bp) insertion sequence (maximum 2,004 bp; mean 519 bp; standard deviation (s.d.) 372 bp; Supplementary Table 5). Insertions are found in phylogenetically diverse members of CPR phyla (Fig. 1, Supplementary Fig. 1 and Supplementary Data 2). Insertion sites are clustered in several distinct locations on the 16S rRNA gene, both in variable and conserved regions (Fig. 2). Most insertions ≥500 bp encode a catalytic RNA intron (group I or II) and/or an open reading frame (ORF), suggesting that they are self-splicing. Encoded proteins frequently belong to families of homing endonucleases (LAGLIDAG 1–3 and GIY-YIG). However, 25% are not similar to known protein families or to each other. These may represent novel endonucleases or may no longer be functional, since loss of function is common in homing endonucleases[9].

Figure 2: Features of insertions encoded within CPR 16S rRNA genes.



Insertions identified in assembled, unique bacterial 16S rRNA genes occur in conserved and variable (V; red bars) regions (Supplementary Table 5). Histograms show the frequency of insertions. Insertions are of several types distinguishable by catalytic RNA introns and/or ORFs. IVP, intervening sequence protein.

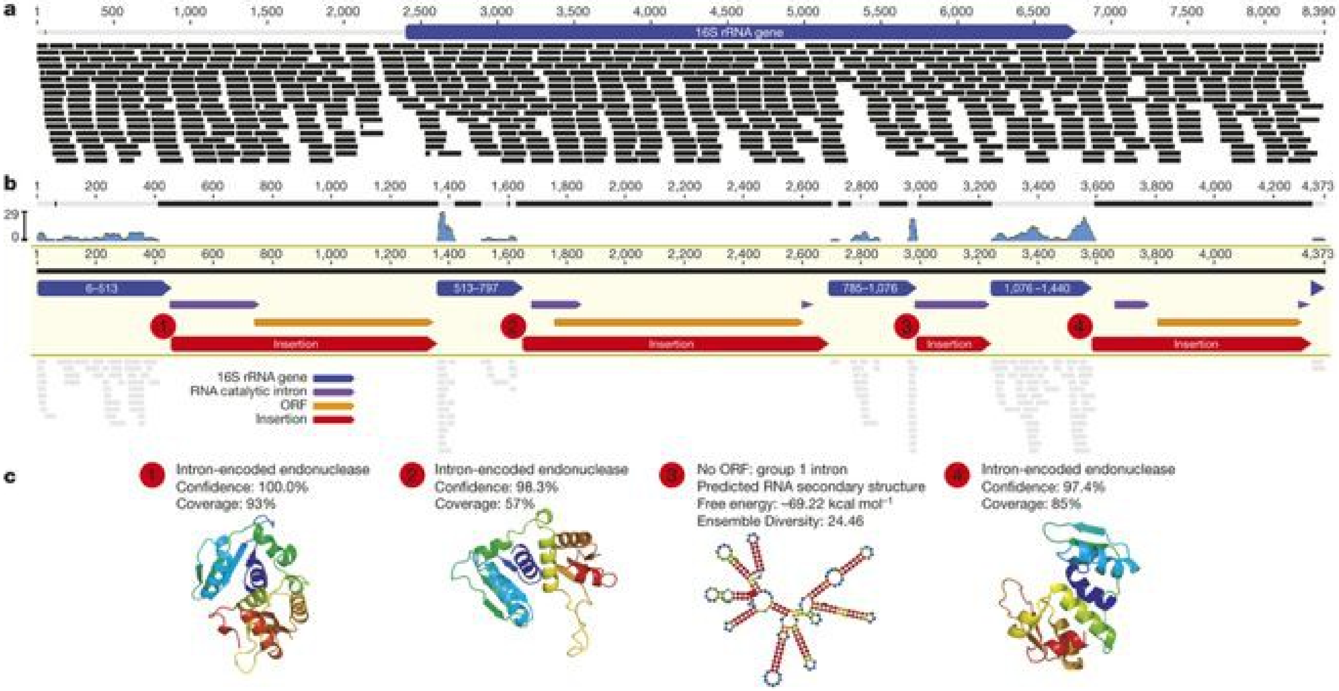
Full size image

Download PowerPoint slide

Four members of the Thiotrichaceae family are the only bacteria known to have self-splicing introns within their 16S rRNA genes[10]. An extensive search for insertions in genes from our study and the Silva database[11] suggests their rarity in bacteria outside the CPR (Extended Data Fig. 4 and Supplementary Table 6). Especially rare are insertions encoding predicted self-splicing introns and/or ORFs. However, these genes need not be functional if the genome encodes additional, insertion-free copies. Importantly, all complete CPR genomes have only one copy of the 16S rRNA gene (this study and others[3,6]). Sequencing coverage analysis of draft genomes further indicates that a single copy is typical for these lineages (Extended Data Fig. 5 and Supplementary Table 7).

Mapping metatranscriptomic sequences to assembled 16S rRNA genes showed that insertions are not retained in transcribed RNAs and are probably rapidly degraded (Supplementary Table 8). However, it is possible that spliced sequences are rendered inaccessible to sequencing after hybridizing, circularizing, or, in some cases, due to their small size. Regardless of their fate, splicing establishes these insertions as introns. Self-splicing is expected if insertions encode a catalytic RNA intron; however, splicing could also occur via an RNase-III-mediated mechanism[12]. Several genes contain multiple introns. For example, one of the complete genomes we obtained encodes a 16S rRNA gene with four introns (Fig. 3).

Figure 3: Intron-encoding 16S rRNA gene from complete Microgenomates genome.



a, Stringent mapping of paired-read metagenome sequences confirms the assembly. **b**, 16S rRNA encoding regions, but not insertions, are covered by perfectly matched metatranscriptome sequences. The absence of RNA sequences for insertions indicates that they are introns. Shown are regions corresponding to *Escherichia coli* K12 gene positions, RNA catalytic introns, ORFs and insertions. **c**, Structural models of encoded proteins (1, 2 and 4: coloured by the colours of the rainbow from the amino to the carboxy terminus) and predicted structure for a catalytic RNA intron (3: coloured by base-pairing probability; red is high, green is moderate, and blue is low). Protein Data Bank structures were used as templates for structural modelling (1: accession 1R7M; 2: 1B24; 4: 1B24).

Full size image

Download PowerPoint slide

CPR bacteria frequently encode introns in 23S rRNA genes with features similar to those in 16S rRNA genes (Extended Data Fig. 6, Supplementary Tables 5, 8 and Supplementary Data 3). However, these introns and encoded proteins share little sequence similarity with one another (Supplementary Table 9). It remains a puzzle why introns in critical, highly transcribed rRNA genes do not make these organisms uncompetitive, as their transcription is costly, even though formation of nonfunctional ribosomes is avoided by splicing.

Insertions in rRNA genes are found in *Coxiella* and Rickettsiales-lineage endosymbionts[13,14]. Interestingly, one member of the Parcubacteria, '*Candidatus* Sonnebornia yantaiensis', is intracellular[15], but does not contain an insertion in its 16S rRNA gene (Fig. 1). However, there is no evidence that an intercellular lifestyle is typical across CPR lineages, although a strong dependence on other community members is likely[3,4].

Metagenomic analyses are polymerase chain reaction (PCR)-independent and, therefore, not biased by primers designed on the basis of expectations of sequence conservation. As a consequence, our sampling indicated that many CPR organisms would evade detection by 16S rRNA gene amplicon surveys. Primer binding analysis showed that primers extensively used in microbial surveys (515F and 806R[16]) would probably not bind to 16S rRNA genes of ∼50% of Microgenomates, ∼50% of Saccharibacteria, 60% of WWE3, and 100% of WS6 sequences sampled here (Extended Data Fig. 7). In fact, these primers would probably miss ∼20% of all bacteria detected in this study, including organisms outside the CPR. Furthermore, introns in these genes would interfere with amplification, both because they occur in regions targeted by primers and because they increase the length of the target sequence. In addition to being excluded during size-selection of amplicons, intron-containing genes are less likely to amplify compared with shorter, intron-free genes[10]. Thus, several barriers have prevented identification of many CPR bacteria.

Removal of introns from 16S rRNA gene sequences, followed by structural alignment[17], was critical to establishing a reliable phylogeny. The new phylogenetic analysis shows that the CPR is monophyletic (Fig. 1), a result also evident in concatenated ribosomal protein trees (Supplementary Fig. 1), and seen in previous analyses[2,3,4,5,11,18]. Phylogenetic analysis defined 35 phyla within the CPR (see later), which encompasses a proposed superphylum, 'Patescibacteria', previously suggested to include just three phyla[5].

The existence of ∼1,500 bacterial phyla was recently suggested[19] using a 75% 16S rRNA gene sequence identity threshold. This contrasts with the current view, which includes 29 established phyla and ∼60 candidate phyla. Using this recent definition[19], we estimate that the CPR consists of >250 phyla (Fig. 1 and Supplementary Fig. 1). With the addition of >550 Mb of CPR genome sequence, there is sufficient sampling to clearly resolve 14 phyla within the Parcubacteria and 11 phyla within the Microgenomates, which have sufficient sequence divergence to account for >120 and >60 phyla, respectively. We propose that these 25 phyla be recognized because (1) complete and/or draft genomes are available, (2) they are monophyletic lineages in both 16S rRNA gene and concatenated ribosomal protein trees, and (3) they pass an approximate 75% 16S rRNA gene sequence identity threshold. Importantly, regardless of whether previous phyla designations or new criteria[19] are used, the CPR comprises >15% of domain Bacteria.

A striking finding from analysis of complete and draft genomes (see statistical assessment in Methods) is unusual ribosome composition in CPR bacteria. All CPR and TM6 bacteria lack ribosomal protein L30 (rpL30; Table 1, Extended Data Fig. 8, Supplementary Table

10 and Supplementary Data 4). Apparently non-essential in bacteria[20], this protein is commonly present except in some symbionts, parasites, Cyanobacteria, and throughout the Planctomycetes–Verrucomicrobia–Chlamydiae (PVC) superphylum[21,22,23]. Although loss of ribosomal protein L25 is often seen in conjunction with absence of rpL30 (ref. 21), TM6 (not within the CPR) is the only candidate phylum studied here for which this is the case. This suggests different trajectories of ribosome evolution between the CPR and other lineages without rpL30.

WS6, WWE3, Saccharibacteria and almost all Microgenomates are missing ribosomal protein L9 (rpL9; Table 1). rpL9 is thought to be universal in bacteria[23], and is involved in both initiation of ribosome assembly[24] and maintaining translation fidelity[25], yet culture-based studies suggest it does not contribute to fitness[20]. Of the three complete Microgenomates genomes, one encodes rpL9. This rpL9 sequence is phylogenetically related to Parcubacteria sequences (Supplementary Fig. 1), suggesting acquisition by lateral gene transfer.

Ribosomal protein L1 (rpL1) is absent from a group within the Parcubacteria that potentially includes >90 phyla. We refer to this group as OD1-L1 (Fig. 1). No other organisms are known to lack rpL1, a large protein that forms a prominent feature of the large subunit[26]. This ribosome initiator protein[24] controls its own expression[27], and loss of rpL1 results in severe growth defects[20]. Absence of rpL1 in this diverse clade suggests alternative mechanisms of ribosome regulation, possibly involving an analogous protein and/or an alternative ribosome structure.

The ribosomal protein biogenesis factor GTPase Der is missing from almost all organisms lacking either rpL9 or rpL1 (Extended Data Fig. 8). Der is essential for ribosome production and is conserved throughout bacteria[28]. Thus, in addition to having unusual ribosome composition, many CPR bacteria probably employ alternative ribosome assembly methods. Although some CPR bacteria have both atypical ribosomes and rRNA introns, these features are not directly linked and thus are not compensatory.

Typically, bacteria within a phylum have widely varying genome sizes and metabolic capacities. In contrast, organisms throughout the CPR have consistently small genomes and similar metabolic limitations. Specifically, all have incomplete tricarboxylic acid cycles and lack electron transport chain complexes, including terminal oxidases and reductases; some lack ATP synthase (Extended Data Fig. 8). With the notable exception of the Peregrinibacteria, most have incomplete nucleotide and amino acid biosynthesis pathways. CPR bacteria are probably obligate fermenters dependent on other organisms for survival, although they could support respiring organisms by excreting fermentation end products. Overall, these characteristics, in addition to unusual ribosomes, a high frequency of rRNA introns and a distinct phylogeny, establish the CPR as a subdivision within domain Bacteria.

Methods

Groundwater sampling and geochemical measurements

We studied groundwater microbial communities from an aquifer adjacent to the Colorado River near Rifle, Colorado, USA at the Rifle Integrated Field Research Challenge (IFRC) site. Aquifer well CD-01 (39° 31′ 44.69″ N, 107° 46′ 19.71″ W; 1,617.5 m above mean sea level) was observed from 23 August to 22 December, 2011, during which a 79-day acetate amendment experiment was conducted (Extended Data Fig. 1 and see refs 7, 8). This well had been subjected to an acetate stimulation experiment during the previous year[29,30]. Acetate (15 μM target concentration within the aquifer) was administered to the alluvial aquifer through a series of injection wells, and microbial biomass was sampled from groundwater pumped from a down gradient monitoring well. Approximately 100 l of groundwater was sampled from a depth of 5 m below ground surface through a 1.2 μm pre-filter, and cells were collected on serial 0.2 and 0.1 μm filters (Supor disc filters; Pall Corporation), with the specific objective of enriching for organisms with small cell sizes. Filters were immediately frozen after collection in a dry ice and ethanol bath. See Supplementary Table 1 for sampling dates and the amount of groundwater filtered over the course of the experiment.

Geochemical measurements were made on samples collected 5 m below ground surface (Supplementary Table 1). The Hach phenanthroline assay and sulphide reagent kits were used to measure ferrous iron and sulfide concentrations, respectively. Acetate and sulfate concentrations were measured by ion chromatography, as previously described[30]. Briefly, acetate and sulfate concentrations were measured with a Dionex ICS-2100 fitted with an AS-18 guard and analytical column.

Metagenome and metatranscriptome sequencing

Six time points spanning a range of geochemical conditions were chosen for metagenomic and metatranscriptomic analysis (Extended Data Fig. 1 and Supplementary Table 1). DNA was extracted from ∼1.5 g of each frozen filter using the PowerSoil DNA Isolation Kit (MO-BIO Labs) with the following modifications: DNA was concentrated by sodium acetate/ethanol precipitation with glycogen, and DNA was eluted in 50 μl Tris buffer. DNA library preparation and sequencing was conducted at the Joint Genome Institute. Total DNA was sequenced on an Illumina HiSeq, producing 150 bp paired reads with a targeted insert size of 500 bp. Sequence data were processed using version 1.8 of the Illumina CASAVA pipeline, and all reads were trimmed based on quality scores using Sickle (https://github.com/najoshi/sickle; default parameters; Supplementary Table 2).

RNA was extracted from the 0.2 μm filters using the Invitrogen TRIzol reagent, followed by genomic DNA removal and cleaning using the Qiagen RNase-Free DNase Set kit and the Qiagen Mini RNeasy kit. An Agilent 2100 Bioanalyzer (Agilent Technologies) was used to assess the integrity of the RNA samples. The Applied Biosystems SOLiD Total RNA-Seq kit was used to generate the cDNA template library. The SOLiD EZ Bead system (Life Technologies) was used to perform emulsion clonal bead amplification to generate bead templates for SOLiD platform sequencing. Samples were sequenced at Pacific Northwest National Laboratory on the 5500XL SOLiD platform. The 50 bp single reads were trimmed using Sickle (default parameters; Supplementary Table 2).

Metagenome assembly, annotation and genome binning

Total community DNA was assembled individually for each sample using IDBA_UD[31] with default parameters (Supplementary Table 2). 16S and 23S rRNA gene sequences were identified from all assembled sequences and curated using an automated method (see later). Scaffold coverage was calculated by mapping reads back to the assembly using Bowtie2 (ref. 32) with default parameters for paired reads. All scaffolds ≥5 kb were included when binning genomes from the metagenome assembly. These scaffolds were annotated by first predicting ORFs using the metagenome implementation of Prodigal[33], and then using USEARCH (-ublast; http://drive5.com/usearch/)[34] to search protein sequences against UniRef90 (ref. 35), KEGG[36,37], and an in-house database composed of ORFs predicted from genomes of candidate phyla organisms. The in-house database includes previously published genomes[2,3,4,38,39] and genomes from ongoing work. Scaffolds were binned on the basis of their GC content, DNA sequence coverage, abundance pattern across samples and taxonomic affiliation, both automatically with the ABAWACA algorithm (see later) and manually using ggKbase (http://ggkbase.berkeley.edu/). Bins generated by ABAWACA were manually inspected within ggKbase. Reported here are genomes binned for organisms associated with the CPR (Fig. 1and Supplementary Table 3) and TM6 (a phylum of organisms with similar characteristics).

To test the accuracy of this binning method, 20 draft-quality genomes were randomly selected from a sample with a high proportion of CPR genomes (GWA2). These genomes were fragmented and then re-clustered on the basis of tetranucleotide signatures using an emergent self-organizing map (ESOM), as previously described[40]. Tetranucleotide frequencies were calculated for 5–10 kb scaffold fragments. The number of occurrences of each tetranucleotide in each fragment was normalized on the basis of the total number of times the tetranucleotide was observed across all fragments, and then these values were log-transformed, standardized so they would follow a normal distribution, and then scaled from 0–1. Normalized tetranucleotide values for each fragment were standardized so that they would also follow a normal distribution. The resulting matrix was used to train an ESOM for 100 epochs using esom_train.pl (https://github.com/micronorman/bantools) (downloaded October 2014). The ESOM was visualized using the Databionic ESOM Tools software[62] (http://databionic-esom.sourceforge.net/). Colouring fragments (data points) in the ESOM on the basis of the genome each fragment originated from enabled validation of these genome bins (Extended Data Fig. 2).

ABAWACA genome binning

ABAWACA was used to generate preliminary genome bins for each sample. This algorithm assesses different characteristics of assembled scaffolds to bin them into genomes. Here, we used a combination of mono-, di- and tri- nucleotide frequencies and coverage values calculated by mapping DNA sequences from all samples to the scaffolds from the sample being binned. This algorithm uses the given information in a hierarchical clustering fashion as follows. First, all scaffolds are broken into 5 kb segments called data points, and the properties of each data point are computed. The binning process begins with a single bin that contains all scaffolds and proceeds by iteratively splitting this and subsequent bins. All non-final bins are evaluated during each iteration. The algorithm searches for a single value for one of the characteristics that will result in the best separation of the scaffolds into two bins. Separation quality is calculated based on the number of data points that were assigned correctly given the separation of the scaffolds. Once a split has been made, scaffolds are separated into the bin with the majority of the data points representing the scaffold. Bins are approved if the quality score exceeds a predefined threshold, and both bins consist of at least 50 data points. A bin is considered final if no separation can be made; otherwise, it undergoes further rounds of binning.

Genome assessment and finishing

Genome bins were associated with CPR lineages on the basis of phylogenetic analysis of 16S rRNA genes and/or ribosomal proteins (see later). When these phylogenetic markers were not present for a particular genome bin, taxonomic placement was achieved based on a consensus of the taxonomic assignments given to ORFs on the basis of their similarity to ORFs from CPR representatives in the candidate phyla database described earlier. Genome completeness was assessed using a modified version of a previously reported list of universal single copy genes (SCGs) for bacteria[41](Supplementary Table 3). Several SCGs were not included as they were found to be unsuitable for the CPR, either because these genes were too divergent in CPR genomes to be reliably detected, or because members of the CPR do not encode these genes. For example, the genes for ribosomal proteins L1 and L9 are not encoded in the genomes of many CPR organisms (see main text). SCGs were identified based on a reciprocal best BLAST[42] hit procedure using a database of SCG protein sequences from a representative set of genomes. First, SCG proteins from the database were searched against all protein sequences in a given genome to identity SCG candidates (blastall -p blastp -F F -e 1e-2). Then, these candidate proteins were searched against the SCG protein sequence database to confirm the assignment (blastall -p blastp -F F -e 1e-5 -b 1 -v 1). SCGs were considered to be present if they were identified by the reciprocal hit method, and the best alignment with a database sequence covered ≥50% of the protein sequence.

To be included in this study as a draft genome, a bin must have contained at least 50% of these SCGs with fewer than 1.125 copies of the genes (indicating that the bin does not contain appreciable contamination from other genomes). To make consistent comparisons with previously sequenced genomes from the CPR, all available genomes were re-assessed using these methods (Supplementary Table 4)[43,44,45].

Several high-quality genome bins were selected for manual curation and genome finishing. Binned scaffolds were connected with one another by extending scaffolds and searching for overlaps. Scaffold extension was achieved by assembling reads mapped to the ends of scaffolds. Assembly errors were detected by manually inspecting the read mapping for these genomes. Genomes were only considered to be complete if they were circular, did not contain gaps, and were, based on complete visual inspection of mapped reads, free of assembly errors. Assembly errors can be identified as regions that do not have read support (that is, reads may map but with mismatches, or regions may not be supported by paired reads). These regions can be manually corrected. Genomes were also checked for the presence of ‘orphaned pairs’, which could indicate alternative assembly paths. The complete genome for GWB1_sub10_OD1-complete was obtained by first assembling 1/10 of the sequence data for sample GWB1, binning scaffolds on the basis of GC content, coverage and taxonomic affiliation, and then genome finishing as described earlier.

Identification of rRNA genes and insertions

16S and 23S rRNA gene sequences were identified based on hidden Markov model (HMM) searches using the cmsearch program from the Infernal package[46] (cmsearch -hmmonly -acc -noali -T -1). Importantly, all identified gene sequences were curated to remove assembly errors before any analysis was conducted (see later). To identify 16S rRNA gene sequences, all assembled contigs were searched against the manually curated structural alignment of the 16S rRNA provided with SSU-Align[17]. Since the SSU-Align 16S rRNA gene covariance model did not include sequences with insertions, large gaps in the alignment between each sequence and the model revealed the boundaries of insertions. Because no equivalent model existed for the 23S rRNA gene, we built a sequence-only model from the manually curated seed alignment maintained by the Comparative RNA Web[47](Supplementary Data 3). While this model did not contain secondary structure information, it was appropriate for identifying 23S rRNA genes, and the boundaries of insertion sequences, from sequence-based HMM alignments, as was done for 16S rRNA genes. To identify the location of rRNA gene insertions with respect to well-studied *Escherichia coli*sequences, all bacterial rRNA gene sequences found to encode insertions were aligned against models consisting of only the respective rRNA from *E. coli* strain K12 substrain DH10B (Fig. 2, Extended Data Fig. 6 and Supplementary Table 5).

Similarity of rRNA insertions to previously studied structural RNA families (for example, group I and group II catalytic RNAs) was determined by searching full rRNA sequences against Rfam[48] using cmscan (also from Infernal; Supplementary Table 5). Regions of the rRNA with significant alignments to a structural RNA family (passed model inclusion threshold) were considered as positive hits if at least 25% of the alignment overlapped with an insertion. These rRNA structural families were of particular interest for determining whether or not insertions encode catalytic RNAs potentially capable of self-splicing from containing RNA sequences (Fig. 2 and Extended Data Fig. 6). RNA secondary structure was predicted for selected intervening sequences using the Andronescu 2007 model[49] implemented in Geneious v. 7.1.5 (ref. 50) (Fig. 3).

ORFs encoded within rRNA insertion sequences were identified by first predicting ORFs across full rRNA genes, and then selecting ORFs encoded within insertion regions. To exclude false ORF predictions, at least 90% of the ORF had to overlap with an insertion. Insertion-encoded ORFs were searched against Pfam[51] to associate encoded proteins with known families (Fig. 2, Extended Data Fig. 6 and Supplementary Table 5). In some cases, Phyre2 (ref. 52) was used to model protein sequences and provide further support for identified homing endonucleases (Fig. 3). Insertions and ORFs identified within 16S and 23S rRNA genes were compared with one another using BLAST (Supplementary Table 9). To assess the prevalence and types of intervening sequences previously sampled in 16S rRNA genes from bacteria, version 115 of non-redundant SILVA[11] was analysed using the same methods (Extended Data Fig. 4 and Supplementary Table 6). Importantly, all insertions ≥10 bp were removed before multiple sequence alignment and phylogenetic analysis of 16S rRNA gene sequences.

Bacterial community composition based on assembled 16S rRNA genes

The composition of the bacterial community was determined on the basis of assembled and curated 16S rRNA gene sequences. Each sequence was given a taxonomic assignment based on the phylogenetic analysis described later. Coverage of all assembled 16S rRNA gene sequences was determined for each sample by stringently mapping reads using Bowtie2 (no mismatches allowed). For each sample, the coverage of all sequences belonging to each lineage of interest was summed, and then converted to a percent relative abundance to observe the composition of each filtrate and shifts in the community across the time series (Extended Data Fig. 3).

16S rRNA gene copy number

16S rRNA gene copy number was estimated for all complete and draft genomes based on two assessments. First, the number of assembled 16S rRNA gene sequences was determined. Second, coverage of 16S rRNA gene regions was compared with the coverage of the rest of the genome to determine relative copy number. Relative copy number was calculated because of the likeliness of assembling only one 16S rRNA gene for organisms with multiple, identical copies of the gene. Owing to the conserved nature of the 16S rRNA gene, it is common for these regions to have inflated coverage values based on default mapping parameters due to inaccurate assignment of reads to sequences from other organisms. To avoid this, both genome and 16S rRNA gene coverage values were calculated based on reads that mapped with zero mismatches. Relative copy number was calculated as: (16S rRNA gene coverage)/(genome coverage). Copy number for each genome was determined by whichever value was greatest, the number of assembled genes or relative copy number (Extended Data Fig. 5 and Supplementary Table 7). Only ten CPR genomes were found to encode more than one copy of the 16S rRNA gene; however, since these genes were not similar to one another, it is more likely that these rare cases were binning errors.

rRNA gene transcript analysis

To determine the fate of rRNA insertion sequences, RNA transcript sequences recovered from 0.2 μm filters were stringently mapped to assembled, curated rRNA genes. To prevent short reads from erroneously matching to either rRNA genes or insertions, zero mismatches were allowed between reads and assemblies. Coverage was calculated separately for 16S rRNA gene and predicted insertion regions, and then the values were compared with one another (Supplementary Table 8). Most insertions were found to have zero coverage. However, in some cases very low coverage of insertion regions was found. In almost all cases these low coverage values were the result of a small portion of the insertion region being covered by RNA sequence, probably the result of a small difference between predicted and actual insertion regions, but possibly the result of partial recovery of spliced insertion sequences.

16S rRNA gene primer binding analysis

The level of sequence divergence of the 16S rRNA genes assembled here from metagenome data, compared with sequences from existing databases, suggests that they would elude PCR-based analysis. We assessed the binding affinity of commonly used 16S rRNA gene survey primers 515F and 806R[16,53]. Assembled 16S rRNA gene sequences were clustered at 97% sequence identity using USEARCH (-cluster_smallmem -query_cov 0.50 -target_cov 0.50 -id 0.97) to remove redundant sequences from the analysis. Because some of the sequences are not complete, only those spanning the 515–806 region of the *E. coli* 16S rRNA gene were included. Primer binding was assessed with PrimerProspector[54] using default parameters (Extended Data Fig. 7).

Phylogenetic analysis

Phylogenetic analysis was carried out using several different marker sequences in order to best survey the diversity within the groundwater microbial community, and to robustly assign taxonomy to complete and draft genomes. Markers included the 16S rRNA gene, ribosomal proteins encoded by a syntenic block of genes, and ribosomal protein S3 (rpS3). The syntenic block encodes the genes for ribosomal proteins L2, 3, 4, 5, 6, 14, 15, 16, 18, 22, 24 and S3, 8, 10, 17, 19, hereafter referred to as rp16. In the rp16 analysis, individual protein sequence alignments were concatenated for phylogenetic inference. Unlike in previous metagenomic studies, near-complete 16S rRNA gene sequences were assembled commonly enough to be able to infer phylogeny for many community members. However, rp16 was also used for phylogenetic analysis because (1) it is encoded in genomes as a syntenic block and is found in only one copy, and thus can be used as a proxy for a particular genotype independent of binning, (2) it encodes ribosomal proteins that provide a robust phylogenetic signal, and (3) it is assembled more frequently from metagenome sequence data compared with the 16S rRNA gene[38]. rpS3 was also independently used as a phylogenetic marker because of its strong phylogenetic signal, despite having a relatively short protein sequence. In cases where a genome did not contain any of these markers (Supplementary Table 3), taxonomic assignment was made based on whole genome comparisons to the database of reference genomes described earlier. In all cases, metagenome assembly was necessary for providing a robust phylogenetic analysis.

After removing insertions ≥10 bp from 16S rRNA gene sequences from this and previous studies, sequences were aligned with SSU-Align. SSU-Align classifies sequences as bacteria, archaea or eukarya, and then generates separate alignments for sequences from each domain. The resulting Stockholm-formatted bacterial multiple sequence alignment was converted to FASTA, and all alignment insert columns were removed. This resulted in a 1,582 bp alignment. All sequences with ≥800 bp of aligned sequence were used for phylogenetic analysis. Several archaeal reference sequences were chosen for the phylogenetic root, aligned to the bacterial 16S rRNA gene model provided with SSU-Align, and concatenated with the bacterial multiple sequence alignment. A maximum-likelihood phylogeny was inferred using RAxML[55] with the GTRCAT model of evolution and 100 bootstrap re-samplings (Supplementary Fig. 1 and Supplementary Data 2). A subset of the tree was annotated using GraPhlAn (http://huttenhower.sph.harvard.edu/graphlan) (Fig. 1).

rp16 ORFs were identified by searching all ORFs encoded on scaffolds ≥5 kb against databases of each of these ribosomal proteins. Searches were carried out with USEARCH (-ublast). Syntenic groups of ORFs were selected if at least three of the ribosomal proteins in rp16 could be identified with an *E*-value $\leq 1 \times 10^{-6}$. This allowed for identification of all instances of each ribosomal protein in rp16 encoded within assembled scaffolds. For each ribosomal protein, all identified protein sequences along with reference sequences were aligned to their respective Pfam HMM profile using hmmalign from the HMMER 3.0 package[56]. Protein sequence alignments were converted from Stockholm format to FASTA, alignment insert columns were removed, and the 16 protein alignments concatenated. This resulted in a 1,935 amino acid alignment. All sequences with ≥1,000 aligned residues were kept for phylogenetic analysis. Because of the size of the multiple sequence alignment, phylogenetic analysis was carried out in two steps. First, FastTree2 (ref. 57) was used to infer the phylogeny of the entire sequence set using the Jones–Taylor–Thornton model of amino acid evolution (JTT) and by assuming a single rate of evolution for each site, the ‘CAT’ approximation (additional options: -spr 4 -mlacc 2 -slownni). Then, sequences associated with the CPR and TM6 were selected, along with representatives of the Archaea and Chloroflexi, to infer a maximum-likelihood phylogeny using RAxML with the LG + alpha + gamma model of evolution and 100 bootstrap re-samplings (see ref. 38 for choice of evolutionary model). Archaea were included as a root for the tree, and Chloroflexi as a root for the CPR. Notably, the CPR is evident as a monophyletic group in both of these analyses, and in the 16S rRNA gene phylogeny (Fig. 1 and Supplementary Fig. 1).

Phylogenies were inferred from individual protein sequences for rpS3 and ribosomal protein L9 (rpL9). All rpS3 protein sequences were identified from metagenome ORFs by searching protein annotation descriptions. The same was done for rpL9, except only sequences associated with CPR genome bins were included. Erroneously annotated sequences were excluded based on the alignment score inclusion threshold for their respective Pfam HMM profiles (aligned using hmmalign), followed by manual removal of non-rpS3 or rpL9 sequences. Sequences were combined with reference sequences and aligned. rpS3 sequences were aligned to Pfam HMM profile PF00189 using the same procedure as was described for the rp16 protein sequences (see earlier). rpL9 was aligned using MUSCLE[58]. All sequences with ≥50 aligned amino acid residues were used for phylogenetic analysis using RAxML with 100 bootstrap re-samplings and an evolutionary model chosen using ProtTest[59] (Supplementary Fig. 1). The ProtTest 2.4 server[59] was run on the Pfam seed alignment for rpS3 and on a random subset of the rpL9 alignment, indicating that the LG + gamma, and the LG + gamma (with fixed base frequencies) evolutionary models should be used for rpS3 and rpL9, respectively.

All phylogenetic trees were visualized using Dendroscope[60].

Identification of novel phyla

The number of phyla within the CPR, Parcubacteria (OD1), and Microgenomates (OP11) was estimated by counting 16S rRNA gene sequence clusters created based on a 75% sequence identity threshold. After removing insertions ≥10 bp, sequences were clustered using USEARCH (-cluster_smallmem -query_cov 0.50 -target_cov 0.50 -id 0.75). This threshold and method for estimating the number of phyla were proposed previously[19]. These authors proposed that phyla could be identified as monophyletic lineages composed of members distinguished by approximately this level of sequence divergence. We classified new phyla based on this and additional, strict criteria. Clusters of 16S rRNA genes that share ≥75% sequence identity were used to assess the divergence and coherence of deep branches of the phylogenetic tree (Supplementary Fig. 1). Bootstrap support values were often higher for lineages primarily composed of one or few clusters, validating the use of this threshold. Lineages were proposed as phyla if (1) they formed a monophyletic group in the 16S rRNA gene phylogeny, (2) 16S rRNA genes were approximately 25% divergent from other lineages, (3) they were also supported by the rp16 concatenated ribosomal protein phylogeny, and (4) representative complete and/or draft genomes were available. Names for these phyla were proposed based on the names of lifetime achievement award recipients in microbiology (Fig. 1, Extended Data Table 1 and Supplementary Fig. 1). Genomes were associated with these phyla using the 16S rRNA gene and/or rp16 phylogenies (Supplementary Table 3).

Sequence curation

Assembled 16S rRNA genes, 23S rRNA genes, and scaffold regions encoding rp16 genes were curated to identify and fix assembly errors before assessment of insertions in rRNA genes and/or phylogenetic analysis. For curation, these genes were extracted along with 2 kb of sequence from each side. Assembly errors, typically short regions of misassembled sequence associated with scaffolding contigs with one another, were identified as regions with zero coverage by stringently mapped paired-end reads. Only one mismatch per read was permitted and only paired reads were included in the analysis. Regions with 1× coverage were only allowed if at least 3 bp on either side of the read overlapped with other reads, with zero mismatches in the overlap region. When an assembly error was detected, read pairs mapped (Bowtie2) to a 1 kb region surrounding the error were collected and reassembled using Velvet[61]. Reads were collected for reassembly as long as at least one read in the pair mapped with two or fewer mismatches. Velvet was run by iterating from kmer 21 to 71, increasing by 10 in each iteration. Reassembled fragments were then merged with the original assembly based on overlap of ≥10 bp. All assembly modifications were verified with a subsequent round of error detection. If an error could not be corrected, the original scaffold was split at the position of the error. In addition to error correction, reads mapped to the ends of scaffolds were reassembled and used to extend scaffolds, or the ends of broken scaffolds, when possible. After curation, genes of interest were re-identified on curated scaffolds using the methods described earlier (Supplementary Data 1). On average, 1.5 assembly errors were corrected for each scaffold region containing a 16S rRNA gene.

Ribosomal protein inventory and metabolic potential of CPR genomes

Metabolic potential of CPR genomes was assessed using ggKbase. In ggKbase, lists related to different proteins or metabolic pathways were generated by searching for specific keywords in gene annotations. Here, lists were created to assess ribosomal protein composition and metabolic potential across the CPR (Extended Data Fig. 8). Genomes were compared with one another by creating ggKbase genome summaries based on a selection of these lists. This allowed for the simultaneous assessment and comparison of the 8 complete and 789 draft-quality genomes assembled here.

To compare genomes on the basis of both their phylogenetic associations and metabolic capacity, and to get the clearest picture of the metabolic potential of the CPR, an additional analysis was conducted with only complete and near-complete genomes (≥75% of single copy genes and ≤1.125 copies, including an assembled 16S rRNA gene). Since similar genotypes were assembled independently from different samples, this set of complete and near-complete genomes was de-replicated by choosing a representative genome for all flat branches on the 16S rRNA gene tree (Supplementary Fig. 1). The genome summary was then ordered based on the 16S rRNA gene phylogeny, a step that was critical for identifying lineages missing specific ribosomal proteins (Extended Data Fig. 8). To find ribosomal proteins that may have evaded detection due to sequence divergence, six-frame translations (bacterial translation table 11) of all complete and draft CPR genomes were searched against Pfam ribosomal protein HMM profiles using hmmscan; however, this confirmed the initial finding of missing ribosomal proteins in organisms from CPR lineages (Supplementary Table 10).

Although complete genomes are invaluable for metabolic analyses, this extensive inventory of draft-quality genomes from organisms representing diverse lineages, and assembled from different samples, enabled confident assessment of gene absence. For example, there are no reported complete WS6 genomes, but the 16 reconstructed draft-quality genomes from this study (median estimated completeness of 91%) showed that this lineage is missing rpL9. The probability of the gene being present, but missing in all 16 genome reconstructions, is $(1 - 0.91)^{16}$, that is, $\sim 2 \times 10^{-17}$. Even if we lower the completion requirement to a very conservative value of 35% complete, 16 such genomes would yield a confidence value of 0.001 for the gene being absent. For lineages where we have hundreds of genomes the probability of missing the gene due to chance is effectively zero.

Code availability

ABAWACA is maintained under https://github.com/CK7/abawaca (version 1.00 used in this analysis: https://github.com/CK7/abawaca/releases/tag/v1.00) and the script used for curating scaffolds, re_assemble_errors.py, is maintained under https://github.com/christophertbrown/fix_assembly_errors(version 1.00 used in this analysis: https://github.com/christophertbrown/fix_assembly_errors/releases/tag/1.00).

Change history

- **29 January 2016**
  Extended Data Table 1 was corrected on 25 January 2016

Accessions

PRIMARY ACCESSIONS

BioProject

- PRJNA273161

Sequence Read Archive

- SRP050083

Data deposits

DNA and RNA sequences have been deposited in the NCBI Sequence Read Archive under accession number SRP050083, and genome sequences have been deposited in NCBI BioProject under accession number PRJNA273161 (first versions described here). Genomes are also available through ggKbase: http://ggkbase.berkeley.edu/CPR-complete-draft/organisms. ggKbase is a ‘live data’ site, thus annotations and genomes may be improved after publication.

References

1.
Harris, J. K., Kelley, S. T. & Pace, N. R. New perspective on uncultured bacterial phylogenetic division OP11. Appl. Environ. Microbiol. 70, 845–849 (2004).
Show contextfor reference 1

CAS
PubMed
Article
Google Scholar
2.
Wrighton, K. C. et al. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. Science 337, 1661–1665 (2012).
Show contextfor reference 2

CAS
PubMed
Article
Google Scholar
3.
Kantor, R. S. et al. Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. MBio 4, e00708–e00713 (2013).
Show contextfor reference 3

CAS
PubMed
Article
Google Scholar
4.
Wrighton, K. C. et al. Metabolic interdependencies between phylogenetically novel fermenters and respiratory organisms in an unconfined aquifer. ISME J.8, 1452–1463 (2014).
Show contextfor reference 4

CAS
PubMed
Article
Google Scholar
5.
Rinke, C. et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499, 431–437 (2013).
Show contextfor reference 5

CAS

PubMed

Article

Google Scholar

6.

Albertsen, M. et al. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nature Biotechnol. 31, 533–538 (2013).

Show contextfor reference 6

Google Scholar

7.

Castelle, C. J. et al. Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. Curr. Biol. 25, 690–701 (2015).

Show contextfor reference 7

CAS

PubMed

Article

Google Scholar

8.

Luef, B. et al. Diverse, uncultivated ultra-small bacterial cells in groundwater. Nature Commun. 6, 6372 (2015).

Show contextfor reference 8

CAS

Article

Google Scholar

9.

Burt, A. & Koufopanou, V. Homing endonuclease genes: the rise and fall and rise again of a selfish element. Curr. Opin. Genet. Dev. 14, 609–615 (2004).

Show contextfor reference 9

CAS

PubMed

Article

Google Scholar

10.

Salman, V., Amann, R., Shub, D. A. & Schulz-Vogt, H. N.Multiple self-splicing introns in the 16S rRNA genes of giant sulfur bacteria. Proc. Natl Acad. Sci. USA 109, 4203–4208 (2012).

Show contextfor reference 10

PubMed

Article

Google Scholar

11.

Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41, D590–D596 (2013).

Show contextfor reference 11

CAS
PubMed
Article
Google Scholar
12.
Evguenieva-Hackenberg, E. Bacterial ribosomal RNA in pieces. Mol. Microbiol. 57, 318–325 (2005).
Show contextfor reference 12

CAS
PubMed
Article
Google Scholar
13.
Raghavan, R., Hicks, L. D. & Minnick, M. F. Toxic introns and parasitic intein in Coxiella burnetii: legacies of a promiscuous past. J. Bacteriol. 190, 5934–5943 (2008).
Show contextfor reference 13

CAS
PubMed
Article
Google Scholar
14.
Baker, B. J., Hugenholtz, P., Dawson, S. C. & Banfield, J. F.Extremely acidophilic protists from acid mine drainage host Rickettsiales-lineage endosymbionts that have intervening sequences in their 16S rRNA genes. Appl. Environ. Microbiol. 69, 5512–5518 (2003).
Show contextfor reference 14

CAS
PubMed
Article
Google Scholar
15.
Gong, J., Qing, Y., Guo, X. & Warren, A. ‘CandidatusSonnebornia yantaiensis’, a member of candidate division OD1, as intracellular bacteria of the ciliated protist Paramecium bursaria (Ciliophora, Oligohymenophorea). Syst. Appl. Microbiol. 37, 35–41 (2014).
Show contextfor reference 15

CAS
PubMed
Article
Google Scholar
16.
Caporaso, J. G. et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J. 6, 1621–1624 (2012).

Show contextfor reference 16

CAS
PubMed
Article
Google Scholar
17.
Nawrocki, E. P. in Structural RNA Homology Search and Alignment using Covariance Models (ed. Eddy, S. R. et al.) (Washington Univ. in Saint Louis, 2009).
Show contextfor reference 17

Google Scholar
18.
Baker, B. J. & Dick, G. J. Omic approaches in microbial ecology: charting the unknown. Microbe 8, 353–360 (2013).
Show contextfor reference 18

Google Scholar
19.
Yarza, P. et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nature Rev. Microbiol. 12, 635–645 (2014).
Show contextfor reference 19

CAS
Article
Google Scholar
20.
Akanuma, G. et al. Inactivation of ribosomal protein genes in Bacillus subtilis reveals importance of each ribosomal protein for cell proliferation and cell differentiation. J. Bacteriol. 194, 6282–6291 (2012).
Show contextfor reference 20

CAS
PubMed
Article
Google Scholar
21.
Lecompte, O. Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. Nucleic Acids Res. 30, 5382–5390 (2002).
Show contextfor reference 21

CAS
PubMed
Article
Google Scholar
22.
Lagkouvardos, I., Jehl, M.-A., Rattei, T. & Horn, M.Signature protein of the PVC superphylum. Appl. Environ. Microbiol. 80, 440–445 (2014).

Show contextfor reference 22

CAS
PubMed
Article
Google Scholar
23.
Yutin, N., Puigbò, P., Koonin, E. V. & Wolf, Y. I.Phylogenomics of prokaryotic ribosomal proteins. PLoS ONE 7, e36972 (2012).
Show contextfor reference 23

CAS
PubMed
Article
Google Scholar
24.
Nowotny, V. & Nierhaus, K. H. Initiator proteins for the assembly of the 50S subunit from Escherichia coliribosomes. Proc. Natl Acad. Sci. USA 79, 7238–7242 (1982).
Show contextfor reference 24

CAS
PubMed
Article
Google Scholar
25.
Atkins, J. F. & Björk, G. R. A gripping tale of ribosomal frameshifting: extragenic suppressors of frameshift mutations spotlight P-site realignment. Microbiol. Mol. Biol. Rev. 73, 178–210 (2009).
Show contextfor reference 25

CAS
PubMed
Article
Google Scholar
26.
Schuwirth, B. S. Structures of the bacterial ribosome at 3.5 Å resolution. Science 310, 827–834 (2005).
Show contextfor reference 26

CAS
PubMed
Article
Google Scholar
27.
Nevskaya, N. Ribosomal protein L1 recognizes the same specific structural motif in its target sites on the autoregulatory mRNA and 23S rRNA. Nucleic Acids Res.33, 478–485 (2005).
Show contextfor reference 27

CAS
PubMed
Article
Google Scholar
28.
Shajani, Z., Sykes, M. T. & Williamson, J. R. Assembly of bacterial ribosomes. Annu. Rev. Biochem. 80, 501–526 (2011).
Show contextfor reference 28

CAS
PubMed
Article
Google Scholar
29.
Luef, B. et al. Iron-reducing bacteria accumulate ferric oxyhydroxide nanoparticle aggregates that may support planktonic growth. ISME J. 7, 338–350 (2013).
Show contextfor reference 29

CAS
PubMed
Article
Google Scholar
30.
Williams, K. H. et al. Acetate availability and its influence on sustainable bioremediation of uranium-contaminated groundwater. Geomicrobiol. J. 28, 519–539 (2011).
Show contextfor reference 30

CAS
Article
Google Scholar
31.
Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28, 1420–1428 (2012).
Show contextfor reference 31

CAS
PubMed
Article
Google Scholar
32.
Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nature Methods 9, 357–359 (2012).
Show contextfor reference 32

CAS
PubMed
Article
Google Scholar

33.
Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119 (2010).
Show contextfor reference 33

CAS
PubMed
Article
Google Scholar
34.
Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461 (2010).
Show contextfor reference 34

CAS
PubMed
Article
Google Scholar
35.
Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23, 1282–1288 (2007).
Show contextfor reference 35

CAS
PubMed
Article
Google Scholar
36.
Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 40, D109–D114 (2012).
Show contextfor reference 36

CAS
PubMed
Article
Google Scholar
37.
Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27 (2000).
Show contextfor reference 37

CAS
PubMed
Article
Google Scholar
38.
Hug, L. A. et al. Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. Microbiome 1, 22 (2013).

Show contextfor reference 38

PubMed
Article
Google Scholar
39.
Castelle, C. J. et al. Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. Nature Commun. 4, 2120 (2013).
Show contextfor reference 39

Article
Google Scholar
40.
Dick, G. J. et al. Community-wide analysis of microbial genome sequence signatures. Genome Biol. 10, R85 (2009).
Show contextfor reference 40

CAS
PubMed
Article
Google Scholar
41.
Raes, J., Korbel, J. O., Lercher, M. J., von Mering, C. & Bork, P. Prediction of effective genome size in metagenomic samples. Genome Biol. 8, R10 (2007).
Show contextfor reference 41

CAS
PubMed
Article
Google Scholar
42.
Altschul, S. F., Gish, W., Miller, W., Meyers, E. W. & Lipman, D. J. Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990).
Show contextfor reference 42

CAS
PubMed
Article
Google Scholar
43.
McLean, J. S. et al. Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. Proc. Natl Acad. Sci. USA 110, E2390–E2399 (2013).
Show contextfor reference 43

PubMed
Article
Google Scholar

44.
Podar, M. et al. Targeted access to the genomes of low-abundance organisms in complex microbial communities. Appl. Environ. Microbiol. 73, 3205–3214 (2007).
Show contextfor reference 44

CAS
PubMed
Article
Google Scholar
45.
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).
Show contextfor reference 45

CAS
PubMed
Article
Google Scholar
46.
Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. Bioinformatics 25, 1335–1337 (2009).
Show contextfor reference 46

CAS
PubMed
Article
Google Scholar
47.
Cannone, J. J. et al. The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics 3, 2 (2002).
Show contextfor reference 47

PubMed
Article
Google Scholar
48.
Burge, S. W. et al. Rfam 11.0: 10 years of RNA families. Nucleic Acids Res. 41, D226–D232 (2013).
Show contextfor reference 48

CAS
PubMed
Article
Google Scholar
49.
Andronescu, M., Condon, A., Hoos, H. H., Mathews, D. H.& Murphy, K. P. Efficient parameter estimation for RNA secondary structure prediction. Bioinformatics 23, i19–i28 (2007).
Show contextfor reference 49

CAS
PubMed
Article
Google Scholar
50.
Kearse, M. et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647–1649 (2012).
Show contextfor reference 50

PubMed
Article
Google Scholar
51.
Finn, R. D. et al. Pfam: the protein families database. Nucleic Acids Res. 42, D222–D230 (2014).
Show contextfor reference 51

CAS
PubMed
Article
Google Scholar
52.
Kelley, L. A. & Sternberg, M. J. E. Protein structure prediction on the Web: a case study using the Phyre server. Nature Protocols 4, 363–371 (2009).
Show contextfor reference 52

CAS
PubMed
Article
Google Scholar
53.
Gilbert, J. A. et al. Meeting report: the terabase metagenomics workshop and the vision of an Earth microbiome project. Stand. Genomic Sci. 3, 243–248 (2010).
Show contextfor reference 53

PubMed
Article
Google Scholar
54.
Walters, W. A. et al. PrimerProspector: de novo design and taxonomic analysis of barcoded polymerase chain reaction primers. Bioinformatics 27, 1159–1161 (2011).
Show contextfor reference 54

CAS
PubMed
Article

Google Scholar

55.

Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313 (2014).

Show contextfor reference 55

CAS
PubMed
Article
Google Scholar

56.

Eddy, S. R. Accelerated profile HMM searches. PLOS Comput. Biol. 7, e1002195 (2011).

Show contextfor reference 56

CAS
PubMed
Article
Google Scholar

57.

Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS ONE 5, e9490 (2010).

Show contextfor reference 57

CAS
PubMed
Article
Google Scholar

58.

Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res.32, 1792–1797 (2004).

Show contextfor reference 58

CAS
PubMed
Article
Google Scholar

59.

Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21, 2104–2105 (2005).

Show contextfor reference 59

CAS
PubMed
Article
Google Scholar

60.

Huson, D. H. & Scornavacca, C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst. Biol. 61, 1061–1067 (2012).
Show contextfor reference 60

PubMed
Article
Google Scholar
61.
Zerbino, D. R. & Birney, E. Velvet: algorithms for de novoshort read assembly using de Bruijn graphs. Genome Res.18, 821–829 (2008).
Show contextfor reference 61

CAS
PubMed
Article
Google Scholar
62.
Ultsch, A. & Moerchen, F. ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM. Technical Report no. 46 (Dept. of Mathematics and Computer Science, University of Marburg, Germany, 2005).
Show contextfor reference 62

Google Scholar

Download references

Acknowledgements

We thank J. Cate and S. Moore for input into the ribosomal protein analysis, J. Doudna and E. Nawrocki for suggestions on the rRNA insertion analysis, and M. Markillie and R. Taylor for assistance with RNA sequencing. Research was supported by the US Department of Energy (DOE), Office of Science, Office of Biological and Environmental Research under award number DE-AC02-05CH11231 (Sustainable Systems Scientific Focus Area and DOE-JGI) and award number DE-SC0004918 (Systems Biology Knowledge Base Focus Area). L.A.H. was partially supported by a Natural Sciences and Engineering Research Council postdoctoral fellowship. DNA sequencing was conducted at the DOE Joint Genome Institute, a DOE Office of Science User Facility, via the Community Science Program. RNA sequencing was performed at the DOE-supported Environmental Molecular Sciences Laboratory at Pacific Northwest National Laboratory.

Author information

Affiliations

Department of Plant and Microbial Biology, University of California, Berkeley, California 94720, USA
Christopher T. Brown

Department of Earth and Planetary Science, University of California, Berkeley, California 94720, USA

Laura A. Hug, Brian C. Thomas, Itai Sharon, Cindy J. Castelle, Andrea Singh, & Jillian F. Banfield

School of Earth Sciences, The Ohio State University, Columbus, Ohio 43210, USA
Michael J. Wilkins

Department of Microbiology, The Ohio State University, Columbus, Ohio 43210, USA
Michael J. Wilkins & Kelly C. Wrighton

Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA
Kenneth H. Williams & Jillian F. Banfield
Department of Environmental Science, Policy, and Management, University of California, Berkeley, California 94720, USA
Jillian F. Banfield

Contributions

Samples and geochemical measurements were taken by M.J.W., K.C.W. and K.H.W. B.C.T. assembled the metagenome data. I.S. implemented the ABAWACA algorithm. C.T.B. and J.F.B. binned the data and carried out the ESOM binning validation. J.F.B. closed and curated the complete genomes. C.T.B., L.A.H. and B.C.T. conducted the rRNA gene insertion analysis. C.T.B. and L.A.H. performed phylogenetic analyses. M.J.W. and K.C.W. conducted the RNA sequencing. C.T.B. carried out the 16S rRNA gene copy number, primer binding and transcript analyses. C.T.B. and J.F.B. carried out the ribosomal protein analyses. C.T.B., L.A.H., C.J.C. and J.F.B. conducted the metabolic analysis. A.S. and B.C.T. provided bioinformatics support. C.T.B. and J.F.B. drafted the manuscript. All authors reviewed the results and approved the manuscript.

Competing interests

The authors declare no competing financial interests.

Corresponding author

Correspondence to Jillian F. Banfield.

Extended data

Extended data figures

Features of insertion sequences encoded within 16S rRNA genes from the Silva database.

5.

16S rRNA gene copy number estimations for genomes reconstructed from groundwater metagenomics.

6.

Features of insertion sequences encoded within 23S rRNA genes recovered from groundwater-associated bacteria.

7.

Analysis of the ability of PCR primers 515F and 806R to bind to recovered groundwater-associated 16S rRNA gene sequences.

8.

Metabolic potential and ribosomal protein analysis of genomes from CPR and TM6 organisms.

Extended data tables

1.

Proposed names for CPR phyla based on microbiology lifetime achievement award recipients

Supplementary information

PDF files

1.

Supplementary Information

This file contains a guide to Supplementary Figure 1, Supplementary Tables 1-10 and the Supplementary Data (see separate files).

2.

Supplementary Figure

This file contains Supplementary Figure 1 (see the Supplementary Information file for details).

Excel files

1.

Supplementary Tables

This file contains Supplementary Tables 1-10 (see the Supplementary Information file for details).

Zip files

1.

Supplementary Data

This zipped file contains the Supplementary Data (see the Supplementary Information file for details).

About this article

Published
15 June 2015

DOI

https://doi.org/10.1038/nature14486