doi:10.1038/nature14486

Supplementary Figures (separate file)

Supplementary Figure 1 | Phylogenetic analyses (see also Supplementary Data File 2). Sequences from this study are labeled GW[A-F][1-2] depending on which sample they originated from (e.g GWA1 is the sample taken at time point A from the 0.1 μm filter). Unless otherwise specified, maximum-likelihood phylogenies were inferred using RAxML with 100 bootstrap re-samplings. **(a)** Phylogeny inferred from 16S rRNA gene sequences aligned using SSU-Align after having removed insertions ≥10 bp long (sequences with ≥800 aligned bp). Reference sequences from previously assembled genomes from the CPR were included, along with sequences associated with the CPR in the Silva database (clustered at 90% sequence identity), sequences from Silva that were the best-hits of sequences assembled here, and a set of reference sequences representative of major clades within domain bacteria. The 75% sequence identity threshold cluster that each sequence was assigned to is indicated with a unique identifier. Based on the 16S rRNA gene phylogeny, the SR1 are not part of the CPR, although they have previously been associated with this group. A subset of this phylogeny was used in **Fig. 1** and **Extended Data Fig. 8**. **(b)** Approximate maximum-likelihood phylogeny inferred using FastTree2 for a concatenation of aligned ribosomal protein sequences (rp16) assembled here and from reference genomes (sequences with ≥1,000 aligned aa residues are included). **(c)** Maximum-likelihood phylogeny inferred for subset of sequences represented in (b). **(d)** Phylogeny for rpS3 sequences identified here and from reference genomes. **(e)** Phylogeny for rpL9 sequences associated with CPR genomes.

Supplementary Tables (separate file)

Supplementary Table 1 | Geochemical measurements from acetate amendment field experiment conducted in aquifer well CD-01 at the Rifle IFRC site.

Supplementary Table 2 | Metagenomics and metatranscriptomics sequencing and assembly statistics.

Supplementary Table 3 | Candidate phyla and Candidate Phyla Radiation (CPR) genomes reconstructed from groundwater metagenomics. The columns for 16S rRNA gene, rp16, and rpS3 phylogeny designate whether or not the specified phylogenetic marker was used to confirm the taxonomy for the organism. Complete genomes are circular and have been manually curated. Draft and partial genome status was determined based on the percent of single copy genes (SCGs) that could be identified. The inventory of these SCGs makes up the right-most columns of the table (rp is an abbreviation for ribosomal protein). Draft-quality genomes have ≥50% of these single copy genes with ≤1.125 copies overall, and must encode at least one of the specified phylogenetic marker genes.

Supplementary Table 4. CPR genomes from previous studies. Previously described CPR genomes[3-6,64-66] were assessed in order to make comparisons with genomes reconstructed in the current study. Genome status was determined in the same way as was done for the genomes presented in **Supplementary Table 3**. The 16S rRNA gene column specifies whether or not a

16S rRNA gene was sequenced for the genome, and if so how it was determined. The rp16 column designates whether or not the genes that make up rp16 were assembled. The inventory of these SCGs makes up the right-most columns of the table (rp is an abbreviation for ribosomal protein).

Supplementary Table 5 | 16S and 23S rRNA gene insertions in sequences reconstructed from groundwater-associated bacteria. The 97% ID centroid column refers to whether or not a given rRNA gene sequence was the representative sequence for a group of rRNA gene sequences clustered based on a 97% sequence identity threshold (USEARCH –cluster_smallmem –query_cov 0.50 –target_cov 0.50 –id 0.75). When identified, sequences of insertion-encoded open reading frames (ORFs) and catalytic RNA introns (group I or group II introns) are provided.

Supplementary Table 6 | 16S rRNA gene insertions in the Silva database. When identified, sequences of insertion-encoded open reading frames (ORFs) and catalytic RNA introns (group I or group II introns) are provided.

Supplementary Table 7 | 16S rRNA gene copy number. 16S rRNA gene copy number was estimated for all draft CPR genomes and genome bins for organisms found outside of the CPR. Relative 16S rRNA gene copy number was calculated as: (16S rRNA gene coverage)/(genome coverage). Estimated 16S rRNA gene copy number was determined for each genome based on whichever value was greatest, the number of assembled genes or relative copy number. Shown in red are the few CPR genomes with discrepant copy number estimates, as discussed elsewhere (see section on 16S rRNA gene copy number in Methods and **Extended Data Fig. 6**).

Supplementary Table 8 | 16S and 23S rRNA gene and insertion transcript analysis. Coverage of rRNA gene and insertion regions by metatranscriptomic sequences were separately determined and compared. Coverage was calculated as (number of mapped bases)/(length of the region). Length coverage refers to the percent of bases in a given region with coverage >0.

Supplementary Table 9 | Comparison of 16S and 23S rRNA gene-encoded insertions and ORFs. All insertions identified in 16S rRNA genes were compared with insertions in 23S rRNA genes (reciprocal BLASTn). Likewise, ORFs encoded within these rRNA genes were compared with one another (reciprocal BLASTp). Reported here are the top 10 hits for each search.

Supplementary Table 10 | HMM identification of ribosomal proteins in CPR genomes. Six-frame translations of all CPR genomes were searched against Pfam ribosomal protein HMM profiles. Identifiers for the profiles searched for each ribosomal protein are shown above the name of the protein. Shown here is the number of genes assigned to each ribosomal protein for all complete and draft-quality genomes. Ribosomal proteins discussed in the text are highlighted in red.

Supplementary Data (separate files)

Supplementary Data File 1-a. Curated 16S rRNA gene sequences in FASTA format.

Supplementary Data File 1-b. Curated 23S rRNA gene sequences in FASTA format.

Supplementary Data File 1-c. Curated rp16 encoding contigs in FASTA format.

Supplementary Data File 2-a. 16S rRNA gene phylogeny (RAxML) in Nexus format.

Supplementary Data File 2-b. rp16 phylogeny (FastTree) in Nexus format.

Supplementary Data File 2-c. rp16 phylogeny (RAxML) in Nexus format.

Supplementary Data File 2-d. rpS3 phylogeny (RAxML) in Nexus format.

Supplementary Data File 2-e. rpL9 phylogeny (RAxML) in Nexus format.

Supplementary Data File 3. 23S rRNA HMM in Infernal format.

Supplementary Data File 4. ggKbase summary of draft genomes in SVG format.