Analysis of five complete genome sequences for members of the class Peribacteria in the recently recognized Peregrinibacteria bacterial phylum

Karthik Anantharaman[1], Christopher T. Brown[2], David Burstein[1], Cindy J. Castelle[1], Alexander J. Probst[1], Brian C. Thomas[1], Kenneth H. Williams[3] and Jillian F. Banfield[1,3]

[1] Department of Earth and Planetary Sciences, University of California, Berkeley, California, United States
[2] Department of Plant and Microbial Biology, University of California, Berkeley, California, United States
[3] Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, United States

Submitted 19 October 2015
Accepted 28 December 2015
Published 28 January 2016

Corresponding author
Jillian F. Banfield,
jbanfield@berkeley.edu

Academic editor
David Baltrus

Additional Information and Declarations can be found on page 11

DOI 10.7717/peerj.1607

ABSTRACT

Five closely related populations of bacteria from the Candidate Phylum (CP) Peregrinibacteria, part of the bacterial Candidate Phyla Radiation (CPR), were sampled from filtered groundwater obtained from an aquifer adjacent to the Colorado River near the town of Rifle, CO, USA. Here, we present the first complete genome sequences for organisms from this phylum. These bacteria have small genomes and, unlike most organisms from other lineages in the CPR, have the capacity for nucleotide synthesis. They invest significantly in biosynthesis of cell wall and cell envelope components, including peptidoglycan, isoprenoids via the mevalonate pathway, and a variety of amino sugars including perosamine and rhamnose. The genomes encode an intriguing set of large extracellular proteins, some of which are very cysteine-rich and may function in attachment, possibly to other cells. Strain variation in these proteins is an important source of genotypic variety. Overall, the cell envelope features, combined with the lack of biosynthesis capacities for many required cofactors, fatty acids, and most amino acids point to a symbiotic lifestyle. Phylogenetic analyses indicate that these bacteria likely represent a new class within the Peregrinibacteria phylum, although they ultimately may be recognized as members of a separate phylum. We propose the provisional taxonomic assignment as '*Candidatus* Peribacter riflensis', Genus Peribacter, Family Peribacteraceae, Order Peribacterales, Class Peribacteria in the phylum Peregrinibacteria.

Subjects Bioinformatics, Environmental sciences, Genomics, Microbiology
Keywords Peregrinibacteria, Candidate phyla radiation, Metagenomics, Complete genomes, Strain variation

INTRODUCTION

Metagenomic analyses of microbial communities in natural environments have revealed numerous previously unrecognized microorganisms, including those now described to be part of the phylum Peregrinibacteria. The Peregrinibacteria are a lineage within the bacterial CPR (*Wrighton et al., 2012*; *Brown et al., 2015*), a monophyletic radiation primarily of candidate phyla. Members of the Peregrinibacteria phylum have been genomically sampled from a contaminated sediment-hosted aquifer (*Wrighton et al., 2012*; *Brown et al., 2015*) and a deep subsurface high CO_2 environment (*Emerson et al., 2015*). However, ribosomal proteins and 16S rRNA gene sequences from this group have been detected in a variety of environments, including sediments and soil (*Hug et al., 2015*). Presently, there are only 18 partial genomes from this phylum. Prior analyses (*Wrighton et al., 2012*) predicted the lack of a complete respiratory chain and most TCA cycle components, consistent with a fermentation-based lifestyle. However, lack of complete, closed genomes from this phylum have limited confidence about predicted metabolic potential and genome size.

Here, we report the first complete genomes from the Peregrinibacteria bacterial phylum. These genomes represent five very closely related strains of a single species.

MATERIALS AND METHODS

Sampling

Groundwater from an alluvial aquifer near the town of Rifle, Colorado (Latitude 39.52927920, Longitude −107.77162320, altitude 1618 m above sea level (*Williams et al., 2011*)) was pumped from a depth of 5 m below ground surface. Samples were collected prior to injection of oxygen-saturated water up gradient from the sampling wells (August 2, 2012) and on three dates after the injection (September 12, 2012, October 24, 2012 and December 12, 2012) and filtered through sequential 1.2 μm, 0.2 μm and 0.1 μm filters.

DNA sequencing, assembly, binning and annotation

DNA was extracted from the filters, as described previously (*Brown et al., 2015*), and paired-end libraries with a read length of 150 bp were sequenced using Illumina HiSeq 2000 at the Joint Genome Institute. Sequence data were processed using the Illumina CASAVA pipeline (version 1.8). Reads were trimmed using Sickle (https://github.com/najoshi/sickle) and assembled using IDBA-UD (-step 20, -maxk 100, -mink 40) (*Peng et al., 2012*). All samples were assembled separately (Supplementary Table S1). Binning of assembled scaffolds was performed on the basis of GC content and differential read coverage with ABAWACA (*Brown et al., 2015*), and tetranucleotide composition with Emergent Self-Organizing Maps (ESOM) (*Dick et al., 2009*). The five genomes described in this study were recovered from five separate samples (A2, B2, C2, D1, D2) across all four sampling time points (A, B, C, D) (Table 1).

Subsequently, individual bins were inspected for scaffolding errors and identified scaffolds were reassembled using Velvet (*Zerbino, 2010*) as described previously (*Brown et al., 2015*). Read-mapping was performed using Bowtie2 (*Langmead & Salzberg, 2012*) with default parameters. Visualization of mapped reads and manual curation of the genomes was conducted in Geneious (*Kearse et al., 2012*) to check the integrity of the five assembled genomes. Complete circularization and filling of internal gaps was achieved using mapped reads from the same sample in three cases, and in two

Table 1 **Summary of the five *Ca.* P. riflensis genomes.**

Genome	Source sample	Length (bp)	GC (%)	Coverage (x)	ORFs
RIFOXYA2_FULL_PER-ii_58_14	A2	1,248,026	58.48	14	1,102
RIFOXYB2_FULL_PER-ii_58_17	B2	1,248,181	58.48	23	1,104
RIFOXYC2_FULL_PER-ii_58_32	C2	1,248,112	58.48	32	1,101
RIFOXYD1_FULL_PER-ii_59_16	D1	1,247,854	58.48	18	1,101
RIFOXYD2_FULL_PER-ii_51_23	D2	1,248,180	58.48	27	1,105

cases, from a different sample. Protein-coding genes were predicted using Prodigal (*Hyatt et al., 2010*), and rRNA genes using Rfam (*Nawrocki et al., 2015*). The genomes were annotated by searching the predicted proteins against Uniref90 (*Suzek et al., 2007*), KEGG (*Kanehisa & Goto, 2000*) and a custom in-house database (*Brown et al., 2015*) using Usearch (-ublast) (*Edgar, 2010*). Functional predictions for proteins of interest were further analyzed using Interpro (*Jones et al., 2014*), psortb (*Yu et al., 2010*), phylogenetic analysis and HMM models constructed from KEGG proteins assigned the same KO annotation (9), which were further sub-clustered using MCL (I = 1.1) (*Van Dongen, 2008*) based on full-length identity percent.

Analysis of Single-Copy Genes (SCG) was performed using 43 universal marker genes (*Raes et al., 2007*; *Brown et al., 2015*).

Phylogenetic analysis

Phylogenetic analysis for validating the taxonomy of the 5 Peregrinibacteria strains was performed with the 16S ribosomal RNA (SSU) gene as well as a set of 16 syntenic universal Ribosomal Proteins (RP) (L2, L3, L4, L5, L6, L14, L15, L16, L18, L22, L24 and S3, S8, S10, S17, S19).

For SSU analysis, 69 SSU genes from the phylum Peregrinibacteria were aligned with MUSCLE (*Edgar, 2004*) with default parameters. All introns in 16S rRNA genes were removed as described previously (*Brown et al., 2015*). Sequences from the phylum Berkelbacteria were used as an outgroup. All columns with >97% gaps were removed and the final alignment spanned 1,649 nucleotides. Phylogenetic analysis of SSU genes was inferred by RAxML using the GTRGAMMA algorithm with 1000 bootstrapped replicates (*Stamatakis, 2014*).

RAxML was called as follows: raxmlHPC-PTHREADS -# 1000 -s input -n output -m GAMMAGTR -p 777 -x 3333 -f a.

For the RP analysis, each ribosomal protein was aligned individually along with reference sequences using MUSCLE (*Edgar, 2004*) with default parameters. The Reference sequence database comprised proteins from 11 previously identified Peregrinibacteria, and Berkelbacteria (*Wrighton et al., 2012*; *Brown et al., 2015*) that served as an outgroup. Individual ribosomal protein alignments were concatenated in Geneious version 7 (*Kearse et al., 2012*). All columns with >97% gaps were removed prior to further analyses. In total, the alignment spanned 3085 columns. Phylogenetic analysis of RP was inferred by

RAxML using the PROTGAMMAGTR algorithm with 100 bootstrapped replicates (*Stamatakis, 2014*). RAxML was called as follows: raxmlHPC-PTHREADS -# 100 -s input -n output -m PROTGAMMAGTR -p 777 -x 777 -f a.

Phylogenetic trees were visualized with Figtree v1.2.2 (http://tree.bio.ed.ac.uk/software/figtree/).

RESULTS AND DISCUSSION

Three complete, circularized genomes were reconstructed from samples D2, B2 and C2. For two other genomes (from samples A2 and D1), very small gaps were filled using mapped reads from other samples. All genomes contain one direct repeat region of a length similar to that of the read insert size. The correct assembly path was chosen because the alternative would have excluded an ~3 kbp sequence that must be part of the genome because paired reads placed there.

The average length of the complete genomes was 1,248,071 ± 217 bp, a result that confirms the previous prediction of small genome size that was made based on genome completeness estimates and draft genome lengths (*Wrighton et al., 2012*). The average GC content was 58.5% and average read coverage varied from 14× to 32× (Table 1). A plot of GC skew across the genome has the expected form, providing confidence in the accuracy of the assembly and an indication of the locations of the origin and terminus of replication (which are not otherwise apparent from the gene predictions). We predicted between 1,101 and 1,105 coding sequences. 52% of the genes had confident functional assignments, based on HMM analysis (Supplementary Table S2), 29% additional proteins had weak similarity to the HMM database (below predetermined noise cutoff), and the others had no detectable similarity to any HMM (19% of all genes in the genome). Notably, 29% of the genes had no significant BLAST hit against NCBI's nr database (E-value <10^{-5}, percent identity >20%). We identified 46 tRNA genes, single copies of each of the 5S rRNA, 16S rRNA and 23S rRNA genes, and all 43 universal SCG in the five genomes. No insertion sequences, such as reported in many CPR bacteria, were identified in the 16S rRNA genes (*Brown et al., 2015*). The 16S rRNA gene sequences of all five genomes are identical, suggesting they derive from strain variants of a single species. The 16S rRNA gene sequence is only 89% similar to that of the closest genomically-sampled relative, Peregrinibacteria bacterium GW2011_GWC2_54_8 (LCRS00000000) (Supplementary Table S3) (*Brown et al., 2015*). The sequences from the genomes reported here, those of GWC2_PER_54_8 and a highly similar genotype GWB1_PER_54_8, and clone sequences identified in the SILVA database (*Pruesse et al., 2007*), branch separately from the other Peregrinibacteria for which partial genomes are available. Thus, the newly reported genomes likely represent a new class within the Peregrinibacteria (Fig. 1). It remains to be seen whether this lineage should instead be recognized as a separate phylum sibling to the Peregrinibacteria. We propose the name Peribacteria for this class. We designate the species as '*Candidatus* Peribacter riflensis' and suggest this lineage be placed as Genus Peribacter, Family Peribacteraceae, Order Peribacteria, Class Peribacteria, within the phylum Peregrinibacteria.

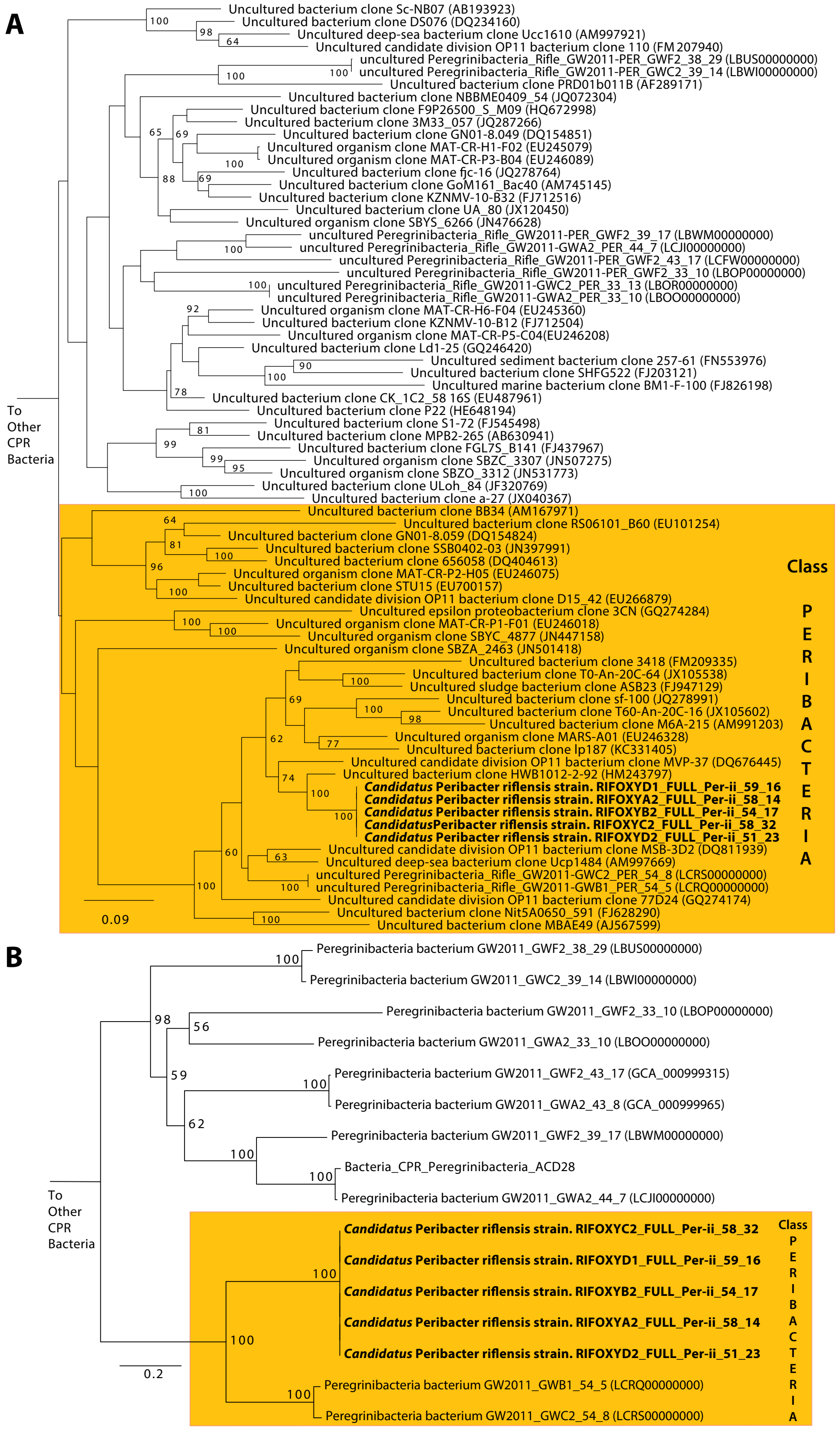
A
Uncultured bacterium clone Sc-NB07 (AB193923)
Uncultured bacterium clone DS076 (DQ234160)
Uncultured deep-sea bacterium clone Ucc1610 (AM997921)
Uncultured candidate division OP11 bacterium clone 110 (FM207940)
uncultured Peregrinibacteria_Rifle_GW2011-PER_GWF2_38_29 (LBUS00000000)
uncultured Peregrinibacteria_Rifle_GW2011-PER_GWC2_39_14 (LBWI00000000)
Uncultured bacterium clone PRD01b011B (AF289171)
Uncultured bacterium clone NBBME0409_54 (JQ072304)
Uncultured bacterium clone F9P26500_S_M09 (HQ672998)
Uncultured bacterium clone 3M33_057 (JQ287266)
Uncultured bacterium clone GN01-8.049 (DQ154851)
Uncultured organism clone MAT-CR-H1-F02 (EU245079)
Uncultured organism clone MAT-CR-P3-B04 (EU246089)
Uncultured bacterium clone fjc-16 (JQ278764)
Uncultured bacterium clone GoM161_Bac40 (AM745145)
Uncultured bacterium clone KZNMV-10-B32 (FJ712516)
Uncultured bacterium clone UA_80 (JX120450)
Uncultured organism clone SBYS_6266 (JN476628)
uncultured Peregrinibacteria_Rifle_GW2011-PER_GWF2_39_17 (LBWM00000000)
uncultured Peregrinibacteria_Rifle_GW2011-GWA2_PER_44_7 (LCJI00000000)
uncultured Peregrinibacteria_Rifle_GW2011-PER_GWF2_43_17 (LCFW00000000)
uncultured Peregrinibacteria_Rifle_GW2011-PER_GWF2_33_10 (LBOP00000000)
uncultured Peregrinibacteria_Rifle_GW2011-GWC2_PER_33_13 (LBOR00000000)
uncultured Peregrinibacteria_Rifle_GW2011-GWA2_PER_33_10 (LBOO00000000)
Uncultured organism clone MAT-CR-H6-F04 (EU245360)
Uncultured bacterium clone KZNMV-10-B12 (FJ712504)
Uncultured organism clone MAT-CR-P5-C04(EU246208)
Uncultured bacterium clone Ld1-25 (GQ246420)
Uncultured sediment bacterium clone 257-61 (FN553976)
Uncultured bacterium clone SHFG522 (FJ203121)
Uncultured marine bacterium clone BM1-F-100 (FJ826198)
Uncultured bacterium clone CK_1C2_58 16S (EU487961)
Uncultured bacterium clone P22 (HE648194)
Uncultured bacterium clone S1-72 (FJ545498)
Uncultured bacterium clone MPB2-265 (AB630941)
Uncultured bacterium clone FGL7S_B141 (FJ437967)
Uncultured organism clone SBZC_3307 (JN507275)
Uncultured organism clone SBZO_3312 (JN531773)
Uncultured bacterium clone ULoh_84 (JF320769)
Uncultured bacterium clone a-27 (JX040367)
To
Other
CPR
Bacteria
Uncultured bacterium clone BB34 (AM167971)
Uncultured bacterium clone RS06101_B60 (EU101254)
Uncultured bacterium clone GN01-8.059 (DQ154824)
Uncultured bacterium clone SSB0402-03 (JN397991)
Uncultured bacterium clone 656058 (DQ404613)
Uncultured organism clone MAT-CR-P2-H05 (EU246075)
Uncultured bacterium clone STU15 (EU700157)
Uncultured candidate division OP11 bacterium clone D15_42 (EU266879)
Uncultured epsilon proteobacterium clone 3CN (GQ274284)
Uncultured organism clone MAT-CR-P1-F01 (EU246018)
Uncultured organism clone SBYC_4877 (JN447158)
Uncultured organism clone SBZA_2463 (JN501418)
Uncultured bacterium clone 3418 (FM209335)
Uncultured bacterium clone T0-An-20C-64 (JX105538)
Uncultured sludge bacterium clone ASB23 (FJ947129)
Uncultured bacterium clone sf-100 (JQ278991)
Uncultured bacterium clone T60-An-20C-16 (JX105602)
Uncultured bacterium clone M6A-215 (AM991203)
Uncultured organism clone MARS-A01 (EU246328)
Uncultured bacterium clone Ip187 (KC331405)
Uncultured candidate division OP11 bacterium clone MVP-37 (DQ676445)
Uncultured bacterium clone HWB1012-2-92 (HM243797)
Candidatus Peribacter riflensis strain. RIFOXYD1_FULL_Per-ii_59_16
Candidatus Peribacter riflensis strain. RIFOXYA2_FULL_Per-ii_58_14
Candidatus Peribacter riflensis strain. RIFOXYB2_FULL_Per-ii_54_17
CandidatusPeribacter riflensis strain. RIFOXYC2_FULL_Per-ii_58_32
Candidatus Peribacter riflensis strain. RIFOXYD2_FULL_Per-ii_51_23
Uncultured candidate division OP11 bacterium clone MSB-3D2 (DQ811939)
Uncultured deep-sea bacterium clone Ucp1484 (AM997669)
uncultured Peregrinibacteria_Rifle_GW2011-GWC2_PER_54_8 (LCRS00000000)
uncultured Peregrinibacteria_Rifle_GW2011-GWB1_PER_54_5 (LCRQ00000000)
Uncultured candidate division OP11 bacterium clone 77D24 (GQ274174)
Uncultured bacterium clone Nit5A0650_591 (FJ628290)
Uncultured bacterium clone MBAE49 (AJ567599)
Class
PERIBACTERIA
0.09
B
Peregrinibacteria bacterium GW2011_GWF2_38_29 (LBUS00000000)
Peregrinibacteria bacterium GW2011_GWC2_39_14 (LBWI00000000)
Peregrinibacteria bacterium GW2011_GWF2_33_10 (LBOP00000000)
Peregrinibacteria bacterium GW2011_GWA2_33_10 (LBOO00000000)
Peregrinibacteria bacterium GW2011_GWF2_43_17 (GCA_000999315)
Peregrinibacteria bacterium GW2011_GWA2_43_8 (GCA_000999965)
Peregrinibacteria bacterium GW2011_GWF2_39_17 (LBWM00000000)
Bacteria_CPR_Peregrinibacteria_ACD28
Peregrinibacteria bacterium GW2011_GWA2_44_7 (LCJI00000000)
To
Other
CPR
Bacteria
Candidatus Peribacter riflensis strain. RIFOXYC2_FULL_Per-ii_58_32
Candidatus Peribacter riflensis strain. RIFOXYD1_FULL_Per-ii_59_16
Candidatus Peribacter riflensis strain. RIFOXYB2_FULL_Per-ii_54_17
Candidatus Peribacter riflensis strain. RIFOXYA2_FULL_Per-ii_58_14
Candidatus Peribacter riflensis strain. RIFOXYD2_FULL_Per-ii_51_23
Peregrinibacteria bacterium GW2011_GWB1_54_5 (LCRQ00000000)
Peregrinibacteria bacterium GW2011_GWC2_54_8 (LCRS00000000)
Class
PERIBACTERIA
0.2
100
98
56
59
100
62
100
100
100
100
100

Figure 1 **Phylogenetic placement of the *Ca.* P. riflensis genomes.** (A) 16S rRNA and (B) concatenated ribosomal protein trees showing the phylogenetic placement of the *Ca.* P. riflensis genomes inferred by maximum likelihood. Bootstrap values (percent of 1000 bootstrap repetitions) are marked on internal nodes. Only bootstrap values ≥50 are shown.

Variation among the five *Ca.* P. riflensis genomes primarily involves insertions and deletions (tens of bps) in twelve proteins (Fig. 2). Of these, seven proteins were very large (between 1,219 and 3,861 amino acids) and several were predicted to have an extracellular localization.

A metabolic reconstruction for the *Ca.* P. riflensis species was established based on analysis of annotated genes (Fig. 3). Examination of the genomes revealed few glycosyl hydrolases that would support growth by utilization of externally derived sugar polymers, a lifestyle suggested for many other CPR bacteria (*Wrighton et al., 2012*; *Kantor et al., 2013*). We identified pathways for both formation and utilization of glycogen/starch and dextrin. Externally-derived glycogen/starch and dextrin-like compounds may be broken down using these enzymes to provide inputs to glycolysis. In addition, the genomes encodes seven enzymes lacking confident functional predictions, but with weak similarity to dextrinases involved in the hydrolysis of starch-derived dextrin. The genomes also encode for synthesis of the compatible solute trehalose.

The genomes lack genes encoding for a complete TCA cycle and most components of respiratory chains, although those for the subunits of the ATP synthase complex were identified, along with an inorganic pyrophosphatase, providing a route to make ATP. The genomes encodes a protein annotated as cytochrome oxidase subunit II, as do those of many other CPR (*Brown et al., 2015*), but the catalytic subunit I that is critical for O_2 utilization is not present. Thus, the ability to metabolize O_2 is not supported.

We predict that the conversion of alpha-D-glucose to pyruvate occurs via the upper and lower glycolytic pathway, with the lack of 6-phosphofructokinase compensated for by enzymes of the Pentose Phosphate Pathway (PPP). This workaround is a common strategy in many members of the CPR that lack 6-phosphofructokinase (*Wrighton et al., 2012*). Lacking are genes required for gluconeogenesis (fructose-1,6-bisphosphatase, pyruvate carboxylase and PEP carboxykinase were not identified). Pyruvate can be converted to oxaloacetate (via phosphoenolpyruvate carboxylase) and to acetyl-CoA (via the two subunit Pyruvate:Ferredoxin Oxidoreductase Complex (PFOR)). The genes for ferredoxin and Fe-S cluster biosynthesis are encoded, as is a gene involved in synthesis of flavodoxin. The redox recycling of ferredoxin is unclear, but one subunit of a Fe-hydrogenase (ferredoxin hydrogenase large subunit) may be involved (however other subunits were not identified). A pathway involving electron bifurcation may be required for production of reduced ferredoxin, possibly involving an electron transferring flavoprotein, but candidate mechanisms are not apparent. One potential mechanism may involve the oxidation of alanine through a Stickland reaction, which could result in oxidative phosphorylation to form ATP and NADH. In turn, the generated NADH could serve as the substrate for electron bifurcation to produce reduced ferredoxin.

We identified genes for production of cofactors needed in nucleotide and other metabolism. Although the genes for biosynthesis of folate itself were not identified, enzymes for tetrahydrofolate compounds (required for one carbon transfer reactions during biosynthesis of serine, methionine, glycine, and purine nucleotides, for example) are present. Folate may be acquired from other organisms, as is often the case for this cofactor (*Graber & Breznak, 2005*). We identified a complete biosynthesis pathway for

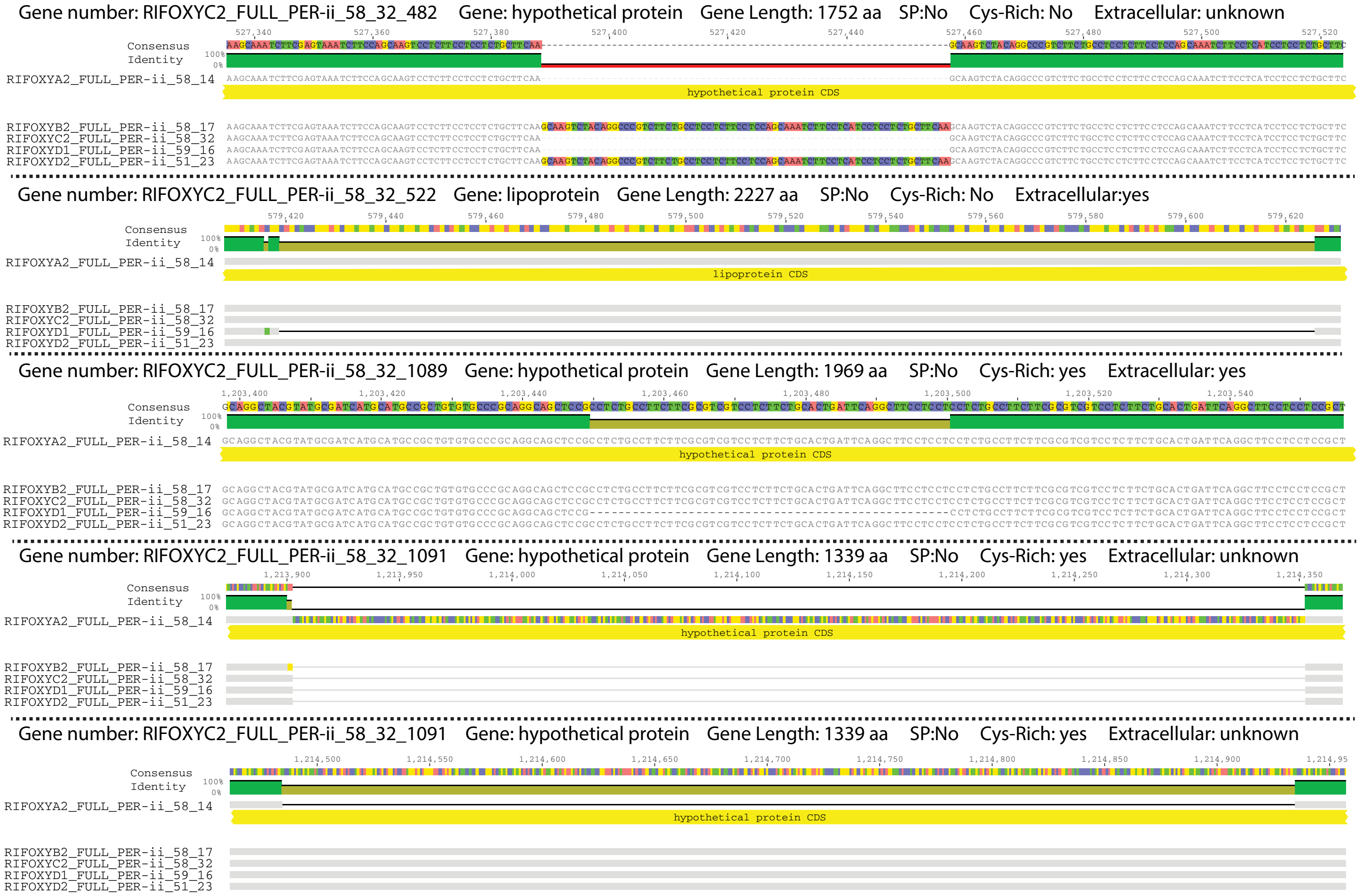
Gene number: RIFOXYC2_FULL_PER-ii_58_32_482 Gene: hypothetical protein Gene Length: 1752 aa SP:No Cys-Rich: No Extracellular: unknown
527,340
527,360
527,380
527,400
527,420
527,440
527,460
527,480
527,500
527,520
Consensus
Identity
100%
0%
RIFOXYA2_FULL_PER-ii_58_14
hypothetical protein CDS
RIFOXYB2_FULL_PER-ii_58_17
RIFOXYC2_FULL_PER-ii_58_32
RIFOXYD1_FULL_PER-ii_59_16
RIFOXYD2_FULL_PER-ii_51_23
Gene number: RIFOXYC2_FULL_PER-ii_58_32_522 Gene: lipoprotein Gene Length: 2227 aa SP:No Cys-Rich: No Extracellular:yes
579,420
579,440
579,460
579,480
579,500
579,520
579,540
579,560
579,580
579,600
579,620
Consensus
Identity
100%
0%
RIFOXYA2_FULL_PER-ii_58_14
lipoprotein CDS
RIFOXYB2_FULL_PER-ii_58_17
RIFOXYC2_FULL_PER-ii_58_32
RIFOXYD1_FULL_PER-ii_59_16
RIFOXYD2_FULL_PER-ii_51_23
Gene number: RIFOXYC2_FULL_PER-ii_58_32_1089 Gene: hypothetical protein Gene Length: 1969 aa SP:No Cys-Rich: yes Extracellular: yes
1,203,400
1,203,420
1,203,440
1,203,460
1,203,480
1,203,500
1,203,520
1,203,540
Consensus
Identity
100%
0%
RIFOXYA2_FULL_PER-ii_58_14
hypothetical protein CDS
RIFOXYB2_FULL_PER-ii_58_17
RIFOXYC2_FULL_PER-ii_58_32
RIFOXYD1_FULL_PER-ii_59_16
RIFOXYD2_FULL_PER-ii_51_23
Gene number: RIFOXYC2_FULL_PER-ii_58_32_1091 Gene: hypothetical protein Gene Length: 1339 aa SP:No Cys-Rich: yes Extracellular: unknown
1,213,900
1,213,950
1,214,000
1,214,050
1,214,100
1,214,150
1,214,200
1,214,250
1,214,300
1,214,350
Consensus
Identity
100%
0%
RIFOXYA2_FULL_PER-ii_58_14
hypothetical protein CDS
RIFOXYB2_FULL_PER-ii_58_17
RIFOXYC2_FULL_PER-ii_58_32
RIFOXYD1_FULL_PER-ii_59_16
RIFOXYD2_FULL_PER-ii_51_23
Gene number: RIFOXYC2_FULL_PER-ii_58_32_1091 Gene: hypothetical protein Gene Length: 1339 aa SP:No Cys-Rich: yes Extracellular: unknown
1,214,500
1,214,550
1,214,600
1,214,650
1,214,700
1,214,750
1,214,800
1,214,850
1,214,900
1,214,95
Consensus
Identity
100%
0%
RIFOXYA2_FULL_PER-ii_58_14
hypothetical protein CDS
RIFOXYB2_FULL_PER-ii_58_17
RIFOXYC2_FULL_PER-ii_58_32
RIFOXYD1_FULL_PER-ii_59_16
RIFOXYD2_FULL_PER-ii_51_23

Figure 2 Diagram showing the form of variation that distinguishes the five closely related genomes. Coding Region on the genome is shown in yellow. Colors on the identity bar reflect similarity between the five genomes: Green, 100%; Olive, 80%; Red, 40%; Black, 20%.

riboflavin and its conversion to FMN and possibly to FAD. Genes for biotin and thiamine biosynthesis were not identified, and are probably not needed due to the absence of pathways in which these enzymes are required (e.g., fatty acid biosynthesis and pyruvate carboxylase of the gluconeogenesis pathway). The source of NAD, a cofactor in several proteins in these genomes (e.g., NAD^+-dependent epimerase), is unclear. However, NAD^+ kinase for interconverting NAD^+ and $NADP^+$ and ADP-dependent NAD(P)H-hydrate dehydratase for dehydration of NADHX and NADPHX were identified.

Genes of the PPP also encode enzymes to form D-ribose 5-phosphate and 5-phospho-alpha-D-ribose 1-diphosphate that are required for synthesis of nucleotides for use as energy carriers and for production of purine, pyrimidine and RNA. In fact, essentially complete pathways for nucleotide synthesis are encoded in the *Ca.* P. riflensis genomes. This capacity distinguishes these organisms from most other CPR bacteria, but it may be a shared feature of some Peregrinibacteria. Genes for homologous recombination (Holliday junction Ruv pathway) and *recA* were also identified, as were genes for cell division and DNA repair. The origin and terminus of replication were identified using a GC skew plot

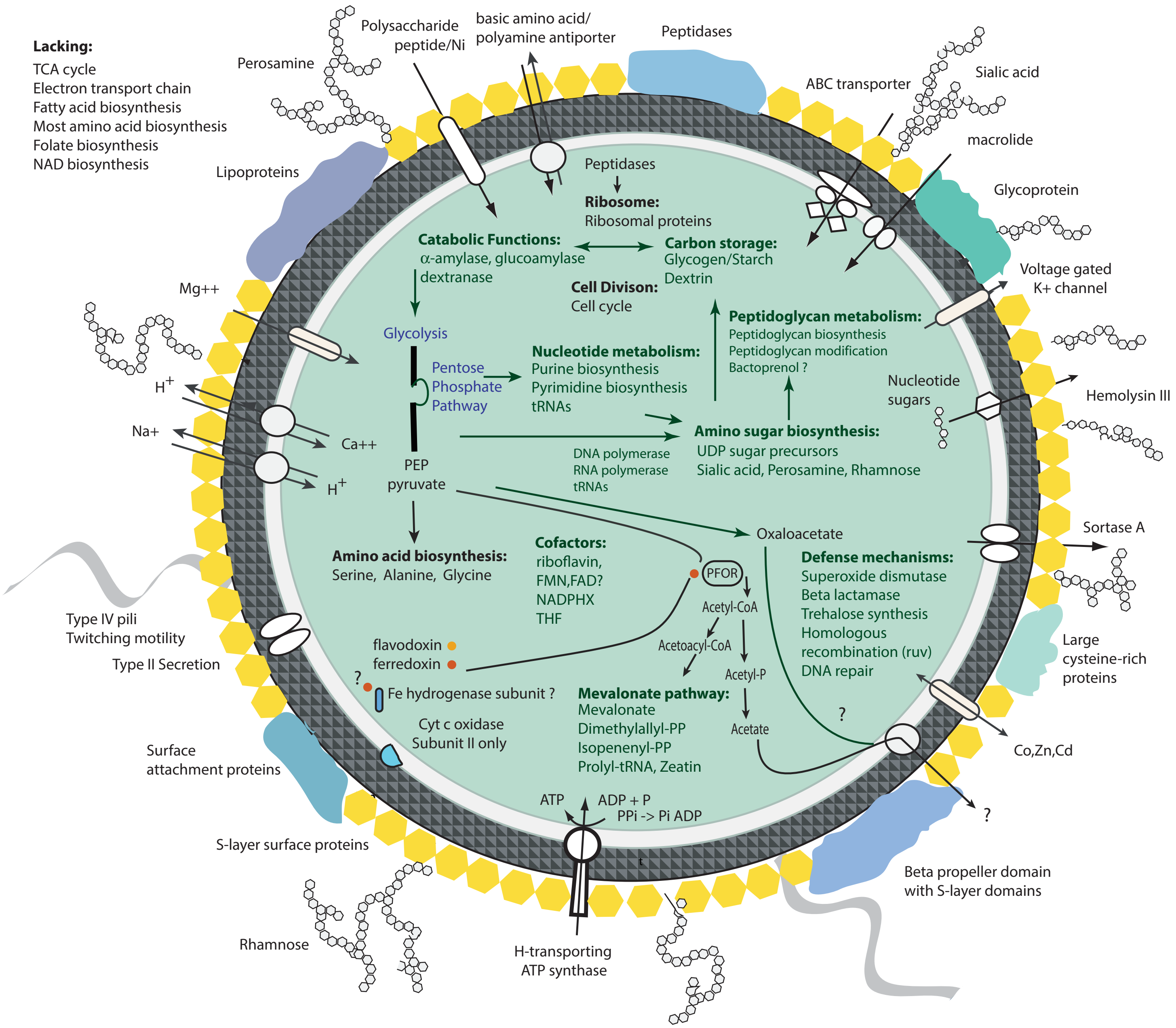

Figure 3 Overview of the protein investments in biosynthesis in *Ca.* P. riflensis. TCA, Tricarboxylic acid cycle; NADPHX, Reduced nicotinamide adenine dinucleotide phosphate; FMN, Flavin mononucleotide; FAD, flavin adenine dinucleotide; PEP, phosphoenolpyruvate; THF, Tetrahydrofolic acid; PPi, Pyrophosphate; PFOR, pyruvate:ferredoxin oxidoreductase.

(Fig. 4A), which shows the form expected for a correctly assembled genome. Short repeats were localized close to the replication origin, and DNaA box-like sequences were present in the vicinity (Fig. 4B).

Unlike some other Peregrinibacteria, *Ca.* P. riflensis lack form II/III RuBisCO, which has been implicated in CO_2 incorporation using a nucleotide salvage pathway (*Wrighton et al., 2012*). Also lacking are hydrogenases (other than the above mentioned single subunit) and CRISPR-Cas phage defense.

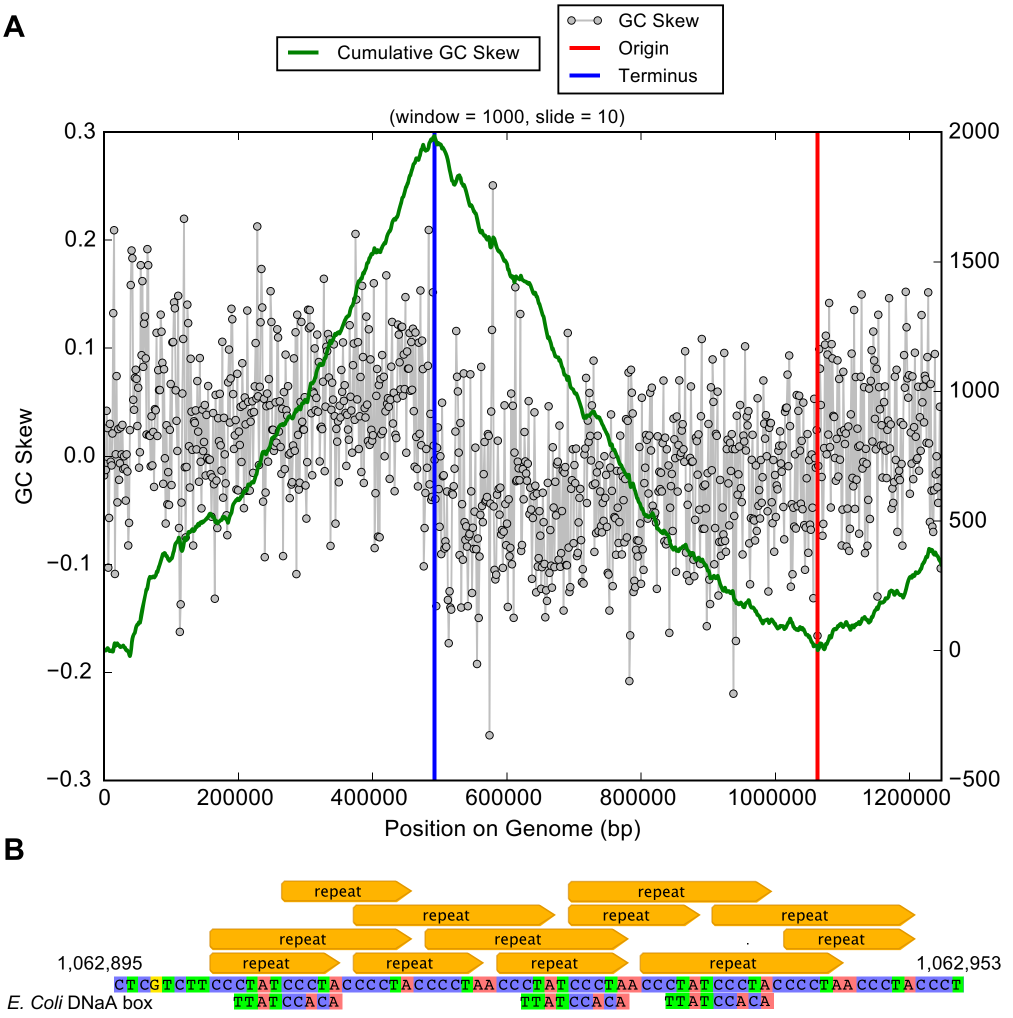

Figure 4 Plot of GC skew over the *Ca.* P. riflensis str.RIFOXYC2_FULL_PER-ii_58_32 genome. (A) Plot of GC skew indicating the locations of the origin/terminus of replication. The plot has the expected form, providing added confidence about the accuracy of the assembly. (B) Localization of short repeats in the identified origin of replication. *E. Coli* DNaA box sequences are shown for reference.

There are only few genes predicted to be involved in amino acid biosynthesis. Synthesis of alanine, serine and glycine is possible. The genomes encode a wide variety of peptidases (including methionyl aminopeptidase, carboxyl-terminal processing protease, Xaa-Pro dipeptidase, ATP-dependent Clp protease). Thus, we infer that these bacteria obtain amino acids via degradation of externally-derived proteins. We identified relatively few genes for degradation of amino acids (e.g., serine hydroxymethyl transferase, L-asparaginase and aspartate amino transferase) to ammonia, oxaloactetate or pyruvate. Thus, we conclude that externally derived amino acids (rather than their breakdown products) are key to protein biosynthesis.

We did not identify a gene for production of acetate from acetyl-CoA. This function is often encoded in CPR genomes as a way to dispose of acetyl-CoA and to produce ATP via the mechanism of substrate-level phosphorylation as part of a fermentation-based metabolism. Acetate production may occur via a pathway involving a protein identified as having a domain that has been implicated in conversion of acetyl-CoA to acetyl-P_i (an acetyltransferase), and a confidently identified enzyme identified to convert acetyl-P_i to acetate (with production of ATP).

Conversion of acetyl-CoA to acetoacyl-CoA is possible. This is likely used in isoprenoid biosynthesis as the genes of the mevalonate pathway, including those that form mevalonate and mevalonate-5P and those to convert mevalonate-5PP to isopentenyl-PP and dimethylallyl-PP were identified. This pathway is relatively rare in bacteria (*Lombard & Moreira, 2010*) outside of the CPR. The purpose for isoprenoids formed by this pathway is uncertain. Intriguingly, however, one gene encoding for an enzyme required to convert dimethylallyl-PP to a zeatin (a plant hormone) precursor was identified (tRNA dimethylallyltransferase).

The source of bacterial fatty acids required for construction of the cytoplasmic membrane is also unclear, as only β-ketoacyl-Acyl Carrier Protein (ACP) reductase (*fabG*) was identified. The many genes for biosynthesis of isoprenoid-based compounds suggest the importance of these molecules in the cell envelope, although there is no indication that they are the predominant structural constituents (*Luef et al., 2015*).

All genes for biosynthesis and modification of peptidoglycan were identified, including genes for production of the amino sugar precursors. Genes for synthesis of cell surface-associated sugar-containing molecules including polysaccharides and a variety of amino sugars were identified, including those for production of sialic acid. In addition, we identified genes encoding for aquaporin Z. The presence of hydrophilic polymers and aquaporin may be important for life in groundwater, where these cells are disproportionately highly abundant (relative to sediment) (*Hug et al., 2015*). Also present are genes for production of rhamnose (dTDP-l-rhamnose synthase is encoded by *rfbABCD*) and perosamine, a mannose-derived 4-aminodeoxysugar. Notably, the genomes encode three copies of perosamine synthetase (*rfbE*) and two copies of glucose-1-phosphate thymidylyltransferase (*rfbA*) suggesting the importance of cell surface polymers to these bacteria. Also present in at least four copies is phosphatidylglycerol: prolipoprotein diacylglycerol transferase. The function of these enzymes in *Ca.* P. riflensis is unclear, but they may be involved in modification of surface proteins, including lipoproteins. Also identified in multiple copies (possibly 8) are proteins with the annotation of dichol-phosphate mannosyltransferase. These may be involved in sugar modification of polyprenols (possibly formed by the mevalonate pathway).

An interesting feature of the Peribacteria genomes is the relatively high cysteine content (1.67%) of the predicted proteins. This value is significantly larger than those reported for bacteria from aquatic and terrestrial environments (average values of 0.05% and 0.15%, although these previously published data indicate substantial within-environment variation) (*Moura, Savageau & Alves, 2013*). Notably, 19% of these cysteines are incorporated into twenty proteins including a 14-gene cluster (7.2 ± 2.4% cysteine) (Supplementary Table S4). These C-rich proteins are large (average length 1,655 ± 751 aa), most are predicted to have a signal peptide and many have predicted extracellular localization, although some may be associated with the cytoplasmic membrane. Two are annotated as integrin alpha FG-GAP repeat containing protein 1, one as a lactonase;

the remainder lack functional predictions, but several show very weak similarity to *Myxococcus* cysteine-rich repeat proteins (and to pappalysin-1). One protein for disulfide bond formation was present. The very high cysteine content may contribute to protein rigidity and/or to surface attachment (*Nakayama, Kurokawa & Lee, 2012*). Genes for production of S-layer proteins were also identified. Other predicted exported proteins include some with lipoprotein annotations (e.g., *NlpD*), and a sortase A. Sortases occur in Gram-positive bacteria, where they function in export and attachment of surface proteins to the cell wall (*Mazmanian, Ton-That & Schneewind, 2001*). Also encoded are genes for twitching motility, genes involved in sensing and response (chemotaxis), a Type II secretion system and genes for biosynthesis of type IV pili, possibly involved in inter-organism interactions (*Luef et al., 2015*).

In summary, these genomes are notable because of the high incidence of genes that encode large exported and often cysteine-rich proteins, substantial investment in production of amino sugar-based polymers, a near-complete mevalonate pathway for production of isoprenoids, and multi-copy genes possibly involved in modification of exported proteins with lipids and sugars. These features point to a prioritization towards synthesis of cell surface-associated compounds by these bacteria. The metabolic predictions are consistent with, at best, a semi-independent lifestyle. Degradation pathways funnel carbon into acetyl-CoA, oxaloacetate, and possibly acetate. Given scant opportunities for use of oxaloacetate by these bacteria, it is possible that this important carbon currency is exported for use by host bacteria. If *Ca.* P. riflensis are symbionts, the localization of sequence variation in large exported proteins may reflect adaption of the different strains to distinct host species.

ACKNOWLEDGEMENTS

DNA sequencing was conducted at the DOE Joint Genome Institute, a DOE Office of Science User Facility, via the Community Science Program.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by Lawrence Berkeley National Laboratory's Sustainable Systems Scientific Focus Area funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under contract DE-AC02-05CH11231. DB was supported by a long-term EMBO fellowship. DNA sequencing was conducted at the DOE Joint Genome Institute, a DOE Office of Science User Facility, via the Community Science Program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research: DE-AC02-05CH11231.

Competing Interests

The authors declare no conflict of interest.

Author Contributions

- Karthik Anantharaman conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Christopher T. Brown performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, reviewed drafts of the paper.
- David Burstein analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, reviewed drafts of the paper.
- Cindy J. Castelle reviewed drafts of the paper.
- Alexander J. Probst reviewed drafts of the paper.
- Brian C. Thomas contributed reagents/materials/analysis tools, reviewed drafts of the paper.
- Kenneth H. Williams reviewed drafts of the paper.
- Jillian F. Banfield conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

The five Peregrinibacteria genome sequences have been deposited at DDBJ/EMBL/GenBank under the accession numbers CP013062–CP013066.

Data Deposition

The following information was supplied regarding data availability:

NCBI Sequence Read Accession (SRA).

SRX1085362

SRX1085365

SRX1085368

SRX1085370

SRX1085371

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.1607#supplemental-information.

REFERENCES

Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF. 2015. Unusual biology across a group comprising more than 15% of domain bacteria. *Nature* **523**:208–211 DOI 10.1038/nature14486.

Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, Banfield JF. 2009. Community-wide analysis of microbial genome sequence signatures. *Genome Biology* **10(8)**:R85 DOI 10.1186/gb-2009-10-8-r85.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32(5):**1792–1797 DOI 10.1093/nar/gkh340.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26(19):**2460–2461 DOI 10.1093/bioinformatics/btq461.

Emerson JB, Thomas BC, Alvarez W, Banfield JF. 2015. Metagenomic analysis of a high carbon dioxide subsurface microbial community populated by chemolithoautotrophs and bacteria and archaea from candidate phyla. *Environmental Microbiology* Epub ahead of print 8 April 2015 DOI 10.1111/1462-2920.12817.

Graber JR, Breznak JA. 2005. Folate cross-feeding supports symbiotic homoacetogenic spirochetes. *Applied and Environmental Microbiology* **71(4):**1883–1889 DOI 10.1128/AEM.71.4.1883-1889.2005.

Hug LA, Thomas BC, Brown CT, Frischkorn KR, Williams KH, Tringe SG, Banfield JF. 2015. Aquifer environment selects for microbial species cohorts in sediment and groundwater. *ISME Journal* **9**:1846–1856 DOI 10.1038/ismej.2015.2.

Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**:119 DOI 10.1186/1471-2105-11-119.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30(9):**1236–1240 DOI 10.1093/bioinformatics/btu031.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28(1):**27–30 DOI 10.1093/nar/28.1.27.

Kantor RS, Wrighton KC, Handley KM, Sharon I, Hug LA, Castelle CJ, Thomas BC, Banfield JF. 2013. Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. *mBio* **4(5):**e00708–e00713 DOI 10.1128/mBio.00708-13.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28(12):**1647–1649 DOI 10.1093/bioinformatics/bts199.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**:357–359 DOI 10.1038/nmeth.1923.

Lombard J, Moreira D. 2010. Origins and early evolution of the mevalonate pathway of isoprenoid biosynthesis in the three domains of life. *Molecular Biology and Evolution* **28(1):**87–99 DOI 10.1093/molbev/msq177.

Luef B, Frischkorn KR, Wrighton KC, Holman H-YN, Birarda G, Thomas BC, Singh A, Williams KH, Siegerist CE, Tringe SG, Downing KH, Comolli LR, Banfield JF. 2015. Diverse uncultivated ultra-small bacterial cells in groundwater. *Nature Communications* **6**:6372 DOI 10.1038/ncomms7372.

Mazmanian SK, Ton-That H, Schneewind O. 2001. Sortase-catalysed anchoring of surface proteins to the cell wall of Staphylococcus aureus. *Molecular Microbiology* **40(5):**1049–1057 DOI 10.1046/j.1365-2958.2001.02411.x.

Moura A, Savageau MA, Alves R. 2013. Relative amino acid composition signatures of organisms and environments. *PLoS ONE* **8(10):**e77319 DOI 10.1371/journal.pone.0077319.

Nakayama H, Kurokawa K, Lee BL. 2012. Lipoproteins in bacteria: structures and biosynthetic pathways. *FEBS Journal* **279(23):**4247–4268 DOI 10.1111/febs.12041.

Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, Finn RD. 2015. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Research* **43(D1):**D130–D137 DOI 10.1093/nar/gku1063.

Peng Y, Leung HCM, Yiu SM, Chin FYL. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28(11):**1420–1428 DOI 10.1093/bioinformatics/bts174.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO. 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research* **35(21):**7188–7196 DOI 10.1093/nar/gkm864.

Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P. 2007. Prediction of effective genome size in metagenomic samples. *Genome Biology* **8(1):**R10 DOI 10.1186/gb-2007-8-1-r10.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30(9):**1312–1313 DOI 10.1093/bioinformatics/btu033.

Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23(10):**1282–1288 DOI 10.1093/bioinformatics/btm098.

Van Dongen S. 2008. Graph clustering via a discrete uncoupling process. *SIAM Journal on Matrix Analysis and Applications* **30(1):**121–141 DOI 10.1137/040608635.

Williams KH, Long PE, Davis JA, Wilkins MJ, N'Guessan AL, Steefel CI, Yang L, Newcomer D, Spane FA, Kerkhof LJ, McGuinness L, Dayvault R, Lovley DR. 2011. Acetate availability and its influence on sustainable bioremediation of uranium-contaminated groundwater. *Geomicrobiology Journal* **28(5–6):**519–539 DOI 10.1080/01490451.2010.520074.

Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, Wilkins MJ, Hettich RL, Lipton MS, Williams KH, Long PE, Banfield JF. 2012. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* **337(6102):**1661–1665 DOI 10.1126/science.1224041.

Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ, Brinkman FSL. 2010. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* **26(13):**1608–1615 DOI 10.1093/bioinformatics/btq249.

Zerbino DR. 2010. Using the velvet de novo assembler for short-read sequencing technologies. *Current Protocols in Bioinformatics*. Chapter 11:Unit 11.5 DOI 10.1002/0471250953.bi1105s31.