environmental microbiology

Environmental Microbiology (2016) **18**(12), 5072–5081 doi:10.1111/1462-2920.13477

Bipartite graph analyses reveal interdomain LGT involving ultrasmall prokaryotes and their divergent, membrane-related proteins

Alexander L. Jaffe,† Eduardo Corel,† Jananan Sylvestre Pathmanathan, Philippe Lopez and Eric Bapteste*

Equipe AIRE, UMR 7138, Laboratoire Evolution Paris-Seine, Université Pierre et Marie Curie, 7 Quai St. Bernard 75005, Paris, France.

Summary

Based on their small size and genomic properties, ultrasmall prokaryotic groups like the Candidate Phyla Radiation have been proposed as possible symbionts dependent on other bacteria or archaea. In this study, we use a bipartite graph analysis to examine patterns of sequence similarity between draft and complete genomes from ultrasmall bacteria and other complete prokaryotic genomes, assessing whether the former group might engage in significant gene transfer (or even endosymbioses) with other community members. Our results provide preliminary evidence for many lateral gene transfers with other prokaryotes, including members of the archaea, and report the presence of divergent, membrane-associated proteins among these ultrasmall taxa. In particular, these divergent genes were found in TM6 relatives of the intracellular parasite *Babela massiliensis*.

Introduction

Recent metagenomic analyses are revealing a wealth of new, unusual microbes that challenge current knowledge about prokaryotic diversity and microbial symbiosis. Among these groups is a cosmopolitan clade termed the Candidate Phyla Radiation (CPR; Brown *et al.*, 2015; Luef *et al.*, 2015; Hug *et al.*, 2016), mostly ultrasmall cells nearing the lower theoretical size limit for viability predicted by physical models (Velimirov, 2001). This clade comprises 15% of the described bacterial phyla and shares cell envelope characteristics with both Gram-positive bacteria and archaea (Brown *et al.*, 2015; Luef *et al.*, 2015). Based on these unusual membranes, small cellular size/genomes, and lack of certain biosynthetic pathways, it has been suggested that these bacteria are obligate fermenters dependent on other microbial community members (Brown *et al.*, 2015). This makes them prime candidates for an endosymbiotic lifestyle.

However, larger novel microbes like the hydrothermal vent organism Lokiarchaeum have also been recently described. This surprising archaeal group harbors membrane-remodeling systems compatible with rudimentary phagocytic capability, and displays a composite proteome possibly acquired by LGT (Spang *et al.*, 2015). Thus, the lineage to which Lokiarchaeum belongs has been proposed as a prime candidate host for prokaryotic endosymbionts, with a possible contribution to eukaryogenesis (Koonin, 2015; Spang *et al.*, 2015; but see Nasir *et al.*, 2015). In principle, the discovery of these novel, candidate hosts and symbionts in the environment adds to debated theoretical suggestions that (i) massive gene transfers between archaea and bacteria (Nelson-Sathi *et al.*, 2015) and (ii) prokaryote-in-prokaryote endosymbiosis (Lake, 2009; Swithers *et al.*, 2011) might have facilitated major evolutionary transitions like the origin of eukaryotes and the emergence of Gram-negative bacteria. However, prokaryote-in-prokaryote symbioses remain extremely rare, with only one described example in the mealybug (Husnik *et al.*, 2013).

New environmental datasets provide a first opportunity to examine the genomic relationships among the CPR, Lokiarchaeota, and other prokaryotic groups. Given the particular characteristics described above, we tested whether members of the CPR might have been endosymbiotic or partners in gene exchange with other bacteria or archaea. More precisely, we looked for signs of endosymbiotic gene transfer—a process by which a symbiont transfers genetic material to the host (Timmis *et al.*, 2004: Martin *et al.*, 2015)—and LGT involving organisms from the ultrasmall size fraction published by Brown *et al.* (2015). To this end, we performed a large-scale BLAST comparison of protein sequences from both draft and complete genomes of CPR (and TM6, a related phylum)

Received 8 July, 2016; accepted 27 July, 2016. *For correspondence. *E-mail: eric.bapteste@upmc.fr; Tel. +330144272164. †These authors contributed equally to this work.

dataset against all complete bacterial and archaeal genomes on NCBI (February 1, 2016). Subsequently, we used a bipartite graph analysis (BGA) to examine the resulting patterns of gene sharing across these genomes. This approach to environmental sequence data allowed us to identify specific processes of transfer or diversity, like those involving membrane-related proteins.

Results/discussion

We began by BLASTing a large dataset of predicted proteins from a set of binned and curated CPR/TM6 genomes ($n = 637{,}155$) against proteins from all complete bacterial and archaeal genomes on NCBI (4,600 genomes, $n = 15{,}373{,}158$ proteins). We dereplicated the CPR/TM6 sequences and partitioned them into four categories—those that showed an above-threshold BLAST hit with only archaeal genomes ($n = 2{,}236$), only bacterial genomes (BAC, $n = 124{,}022$), both (PROK, $n = 81{,}634$), or neither (CPR/TM6, $n = 158{,}245$). We first performed a BGA on the BAC subset, delineating groups of CPR/TM6 sequences with shared, exclusive similarity to a given set of prokaryotic genomes. These groupings of proteins exclusively associated with a given set of genomes are known as "twins," the detection of which is an efficient way to represent complex patterns of gene sharing among organisms (Corel *et al.*, 2016). For example, a "twin" associating a group of sequences to one or more complete CPR genomes indicates that these sequences are likely from a CPR organism. However, twins connecting a set of CPR/TM6 sequences with one or more bacteria or archaea distantly related to CPR/TM6 suggests a case of gene transfer, endosymbiotic or otherwise, between CPR/TM6 organisms and other distantly related prokaryotes (Fig. 1).

The BGA resulted in 82,953 "twins" that were sorted by decreasing number of CPR/TM6 proteins they contained. The twins containing the largest number of CPR/TM6 sequences (between 268 and 3,540) involved the complete CPR genomes from the Brown *et al.* (2015) dataset, as well as 5 strains of *Peribacter riflensis*, another phylum in CPR (Anantharaman *et al.*, 2016). This result is expected, and offers a good proof of concept for our methodology: given that most proteins in the Brown *et al.* (2015) dataset are already classified as CPR, the BGA approach should associate those proteins with complete CPR genomes. The next strongest signal (i.e., CPR/TM6 proteins exclusively associated with a particular prokaryotic genome) revealed 456 proteins showing distant (~39% mean sequence identity, Fig. 2) but exclusive similarity to *Babela massiliensis*, a gram-negative, intracellular amoeboid parasite in the candidate phylum TM6 (Pagnier *et al.*, 2015). These 456 genes were contained by 16 different

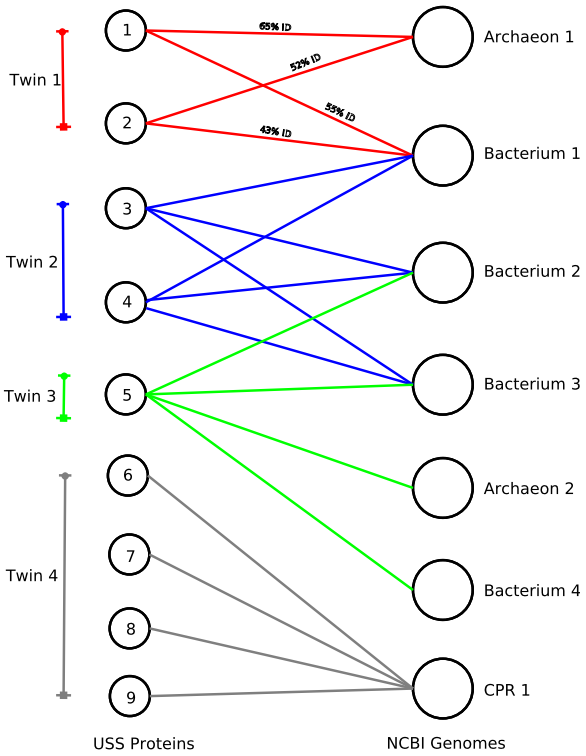

Fig. 1. The process of defining 'twins' in a BGA delineates groups of sequences with shared, exclusive similarity to a given set of genomes. For example, sequences 1 and 2 belong to a twin because they exclusively associate with the same set of genomes (Archaeon 1 and Bacterium 1). Note that genomes can be included in two or more different twins—Twin 2 also contains Bacterium 1 but involves a different set of proteins (3 and 4). In this case, sequences 2 and 3 show similarity to different genes within Bacterium 1 (i.e., they are not homologous). Twin 3 is an example of a twin where one or several CPR/TM6 sequences associate with many different genomes. Twin 4, in which multiple CPR/TM6 sequences associate exclusively with one genome, is an example of an interesting case that can allow attribution of CPR/TM6 sequences to a particular species (when the contained genome is a CPR/TM6 bacterium) or can hint at patterns of gene transfer or novel diversity (when contained genome is not a CPR/TM6 bacterium). Each edge between sequences and genomes has a corresponding weight, or percent identity (see Twin 1 for example), which were calculated from the BLAST results.

bins (Supporting Information Table S1), all but one of which were taxonomically annotated as TM6.

Patterns of gene similarity in the TM6

Our results indicate that as many as 16 ultrasmall organisms in the Brown *et al.* dataset have genes with exclusive similarity to those in *Babela massiliensis*. This is interesting for two reasons: First, these relationships may help to begin constructing more detailed phylogeny among the TM6, which to date remains mostly unstudied. Specifically, that a set of genes among novel TM6 representatives resembles *Babela* (or one of its relatives) adds to existing

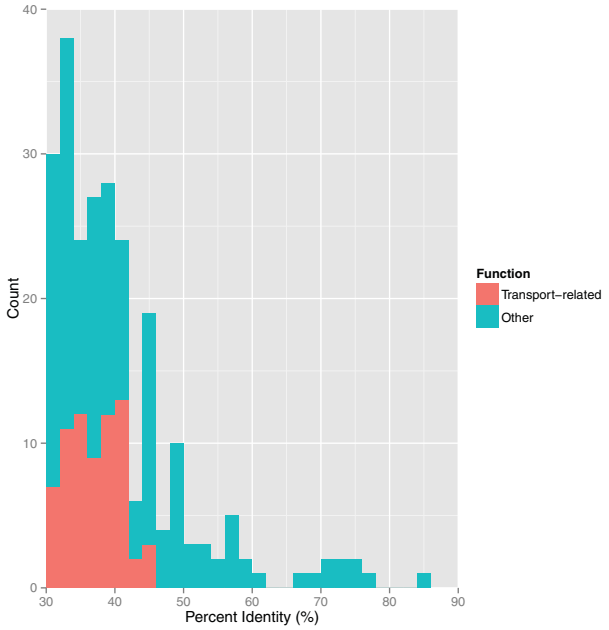

Fig. 2. The distribution of percent identity between CPR/TM6 proteins and their counterparts in *Babela massiliensis*, with highly divergent membrane, transporter, pump, and translocase-related ones highlighted in red. While the BGA twin relating the CPR/TM6 to this TM6 bacterium contained 456 proteins, only 236 could be annotated by RPS Blast.

knowledge of gene ancestry in this group. However, because only several complete TM6 genomes were included in our analysis, this relationship remains relative and may change as additional genomes in this phylum are discovered, described, and analyzed. Second, our results could help to shed light on the genomic consequences of ultrasmall size in the TM6. These organisms, based on the size fraction from which they were collected, would be 6–10 times smaller than their relative *Babela* (Pagnier *et al.*, 2015). Despite this, our twin analysis clearly confirms that small and ultrasmall TM6 have many related genes, some of which appear divergent.

Interestingly, *Babela* exhibits many of the genomic characteristics typical of intracellular symbionts, including reduction of genome size through loss of biosynthetic pathways (Pagnier *et al.*, 2015). In addition, *Babela* contains many genes related to transport, including ATP/ADP translocases, porins, an ABC-family permease, and other transporters (Pagnier *et al.*, 2015). These membrane-associated proteins could be important in the integration of metabolisms at the host-endosymbiont interface—in eukaryotes, specialized transporters currently play a role in moving small molecules across the inner envelope membrane of chloroplasts, connecting cytosolic and organellar pathways (Weber and Fisher, 2007). Interestingly, we recovered highly divergent versions of some of these same genes, among others, in the BGA twin associated with *Babela*—in particular, 17 amino acid transporters, 20 ATP/ADP and preprotein translocases, 18 multidrug pump/transporters, and several other related genes in the CPR/TM6 (Table 1). These membrane-related genes were related to that of *Babela* but with a low identity (~37% mean % ID, Fig. 2), and were contained in 14 bins also belonging to uncharacterized TM6 organisms. At any rate, further work should address the possibility that the highly divergent transport proteins recovered among the environmental TM6 play a role in adapting to a lifestyle in the ultrasmall size fraction. This lifestyle may not necessarily be parasitic, although recent work has indicated that this mode is likely both common and ancestral among the TM6 clade (Gong *et al.*, 2014; Yeoh *et al.*, 2015).

Lateral gene transfer between CPR/TM6 and other prokaryotes

We repeated the BGA for the ARC data partition, again sorting the resulting twins by decreasing number of CPR/TM6 proteins they contained. This yielded three top twins, each of which linked sequences to a single archaeal genome—Woesearchaeota AR20 and Diapherotrites AR10, two ultrasmall size-fraction archaea from the superphylum DPANN (Rinke *et al.*, 2013; Castelle *et al.*, 2015), and Lokiarchaeum (Spang *et al.*, 2015)—with which there were 230, 131, and 53 exclusively associated proteins, respectively. We subsequently created a heatmap showing the distribution of sequence similarity between CPR/TM6 sequences in the ARC subset and genes within the complete archaeal genomes from NCBI (Fig. 3). This revealed further regions of interest.

First, we did not observe a pattern of high similarity (>70% ID) between CPR/TM6 proteins in the ARC subset and any archaeal genes, indicating that recent interdomain gene transfer is an unlikely explanation for the presence of numerous CPR/TM6 homologs in Archaea. However, the heatmap did reveal a “core group” of 62 CPR genes that showed distant homology (mean ~39% ID) to a large distribution of the complete archaeal genomes (Region A in Fig. 3). Region C and D generally corresponded to the two twins identified as top results in the BGA, involving the novel archaeal genomes Diapherotrites AR10 and Woesearchaeota AR20 (39–40% ID). Interestingly, these genomes were assembled from the same sample site and size fraction as the CPR dataset as part of a larger study identifying new members of the DPANN (Brown *et al.*, 2015; Castelle *et al.*, 2015). Additionally, the two groups of CPR/TM6 sequences associated with these genomes showed similar functional profiles, containing many divergent, membrane-related proteins (Table 2).

To determine whether these archaea-exclusive signals stemmed from inaccurate binning (and may therefore reflect that some contigs belong to archaea rather than

Table 1. Functional description of 236 CPR/TM6 proteins associating exclusively with *Babela massiliensis*, as annotated by RPS Blast. Highly divergent membrane, transporter, pump, and translocase-related proteins are in marked in bold face.

Count	COG	Annotation
17	**COG0531**	**Amino acid transporters**
14	COG0612	Predicted Zn-dependent peptidases
13	**COG3202**	**ATP/ADP translocase**
12	**COG0534**	**Na+-driven multidrug efflux pump**
11	COG0265	Trypsin-like serine proteases, typ. perip. contain C-term PDZ dom.
10	COG0285	Folylpolyglutamate synthase
10	COG2932	Predicted transcriptional regulator
9	COG0544	FKBP-type peptidyl-prolyl cis-trans isomerase (trigger factor)
9	COG0592	DNA polymerase sliding clamp subunit (PCNA homolog)
9	**COG4775**	**Outer membrane protein/protective antigen OMA87**
8	COG2812	DNA polymerase III, gamma/tau subunits
6	COG0681	Signal peptidase I
6	**COG0706**	**Preprotein translocase subunit YidC**
6	**COG1132**	**ABC-type multidrug transport system, ATPase permease comps.**
6	COG1524	Uncharacterized proteins of the AP superfamily
5	COG0712	F0F1-type ATP synthase, delta sub. (mito. oligomycin sens. prot.)
4	COG3264	Small-conductance mechanosensitive channel
3	COG0333	Ribosomal protein L32
3	COG0456	Acetyltransferases
3	COG0596	Pred. hydrolases/acyltransferases (alpha/beta hydrolase superf.)
3	COG0607	Rhodanese-related sulfurtransferase
3	COG0636	F0F1-type ATP synth. sub. c/Arch./vacuolar-type H+-ATPase, sub K
3	COG1011	Predicted hydrolase (HAD superfamily)
3	COG1214	Inactive homolog of metal-dep. proteases, putative mol. Chaperone
3	COG2165	Type II secretory pathway, pseudopilin PulG
3	COG3031	Type II secretory pathway, component PulC
3	COG3283	Transcriptional regulator of aromatic amino acids metabolism
3	COG4972	Tfp pilus assembly protein, ATPase PilM
2	COG0037	Pred. ATPase of the PP-loop superf. implicated in cell cycle control
2	COG0200	Ribosomal protein L15
2	COG0224	F0F1-type ATP synthase, gamma subunit
2	COG0355	F0F1-type ATP synthase, epsilon sub. (mitochondrial delta subunit)
2	COG0360	Ribosomal protein S6
2	COG1974	SOS-response trans. repressors (RecA-mediated autopeptidases)
2	COG2204	Response reg. w/CheY-like receiver, ATPase, & DNA-bind. Doms
2	COG2267	Lysophospholipase
2	COG3688	Predicted RNA-binding protein containing a PIN domain
2	COG4564	Signal transduction histidine kinase
2	**COG4591**	**ABC-type transport sys., inv. in lipop. release, permease comp.**
1	COG0006	Xaa-Pro aminopeptidase
1	COG0204	1-acyl-sn-glycerol-3-phosphate acyltransferase
1	COG0269	3-hexulose-6-phosphate synthase and related proteins
1	COG0331	(acyl-carrier-protein) S-malonyltransferase
1	COG0356	F0F1-type ATP synthase, subunit a
1	COG0419	ATPase involved in DNA repair
1	COG0545	FKBP-type peptidyl-prolyl cis-trans isomerases 1
1	COG0666	FOG: Ankyrin repeat
1	COG0707	UDP-N-acetylglucosamine:LPS N-acetylglucosamine transferase
1	COG0758	Pred. Rossmann fold nucl.-binding protein involved in DNA uptake
1	COG0793	Periplasmic protease
1	COG0858	Ribosome-binding factor A
1	COG1221	Trans. Regs. w/AAA-type ATPase domain & DNA-binding dom
1	COG1222	ATP-dependent 26S proteasome regulatory subunit
1	**COG1297**	**Predicted membrane protein**
1	**COG1314**	**Preprotein translocase subunit SecG**
1	COG1450	Type II secretory pathway, component PulD
1	**COG1463**	**ABC-type tranp. sys. Inv. in resisting org. solvents, peripl. comp.**
1	COG1544	Ribosome-associated protein Y (PSrp-1)
1	COG1579	Zn-ribbon protein, possibly nucleic acid-binding
1	COG1723	Uncharacterized conserved protein
1	COG3027	Uncharacterized protein conserved in bacteria
1	COG3829	Trans. regulator w/PAS, AAA-type ATPase, & DNA-binding dom.
1	COG4232	Thiol:disulfide interchange protein
1	**COG4907**	**Predicted membrane protein**
1	COG4970	Tfp pilus assembly protein FimT

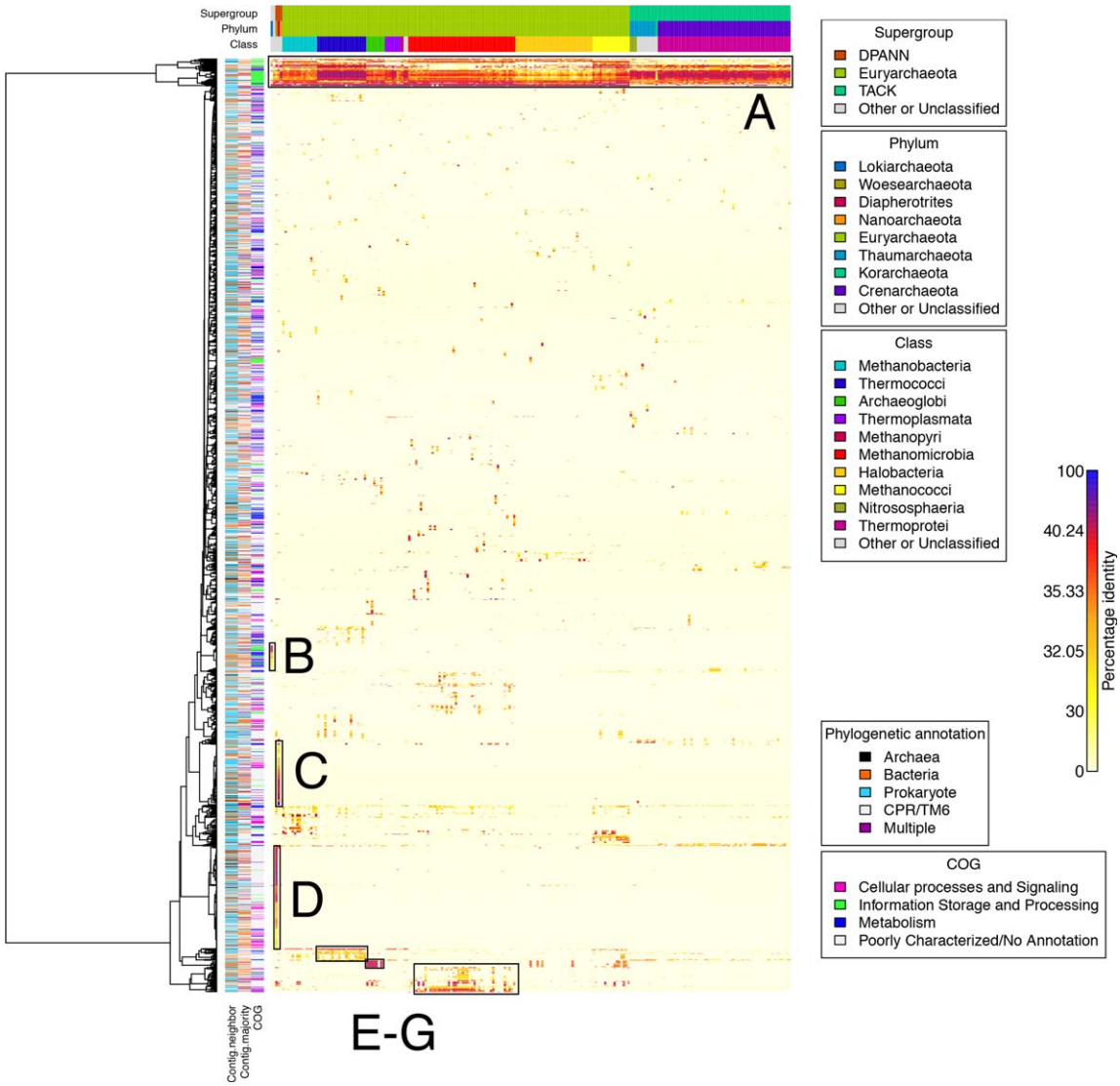

Fig. 3. A heatmap showing patterns of similarity between the CPR/TM6 proteins contained in the ARC subset and the archaeal genomes retrieved from NCBI. The percent identities shown were calculated from the BLAST hits between CPR/TM6 proteins and their corresponding proteins in the NCBI archaeal genomes. Taxonomic information for these genomes and genomic context/COG info for the CPR/TM6 proteins are shown in the heatmap sidebars (see Procedures and Supporting Information Methods).

CPR/TM6 organisms) or from interdomain gene transfer, we retrieved the number and taxonomy of genomic bins containing the genes in each twin. For the twin corresponding to Woesearchaeota AR20, 230 sequences were contained in 156 bins; for that corresponding to Diapherotrites AR10, 131 sequences were contained in 112 bins; for that corresponding to Lokiarchaeota, 53 sequences were contained in 52 bins; for that corresponding to the "core group" (Region A, Fig. 3), 62 sequences were contained in 52 bins; 90% or more of the genes in these twins were identified as belonging to the CPR phyla *Microgenomates* or *Parcubacteria*; a small number were of *Berkelbacteria*, *Peregrinibacteria*, or other origin (Supporting Information Table S1). Most were unannotated at the class level. We also examined the sequences at a contig level, retrieving the most frequent BLAST-assigned taxonomic annotations on each of the contigs containing twins with exclusive similarity to archaeal genes. These results showed that very few of these genes (<5% of each twin) were from contigs that met the majority rule for ARC placement. 72% or more of these contigs (containing genes exclusively similar to archaea) met the majority rule for bacterial (BAC) or CPR (no BLAST match to other bacteria, CPR, or archaea) origin (Supporting Information Table S1). Thus, while semiautomatic taxonomic assignments are limited, and contamination by low abundance,

Table 2. Functional description of CPR/TM6 proteins associating exclusively with singular archaeal genomes –Diapherotrites AR10, Woesearchaeota AR20, and Lokiarchaeum, respectively, as annotated by RPS Blast. COG annotations shared across genome groups ("twins") are marked in bold face.

Count	COG	Annotation
Diapherotrites AR10		
8	COG4095	Uncharacterized conserved protein
5	COG0785	Cytochrome c biogenesis protein
3	**COG1651**	**Protein-disulfide isomerase**
2	COG0526	Thiol-disulfide isomerase and thioredoxins
2	COG1378	Predicted transcriptional regulators
2	COG1522	Transcriptional regulators
2	COG2267	Lysophospholipase
1	COG0089	Ribosomal protein L23
1	COG0189	Glutathione synth/Ribo. Prot. S6 mod enzyme
1	**COG0438**	**Glycosyltransferase**
1	COG0451	Nucleoside-diphosphate-sugar epimerases
1	COG0500	SAM-dependent methyltransferases
1	COG1032	Fe-S oxidoreductase
1	COG1305	Transglutaminase-like enzymes, putative cysteine proteases
1	COG1418	Predicted HD superfamily hydrolase
1	COG1577	Mevalonate kinase
1	**COG2226**	**Methylase involved in ubiquinone/menaquinone biosynthesis**
1	COG2230	Cyclopropane fatty acid synthase and related methyltransferases
1	COG2717	Predicted membrane protein
1	COG3118	Thioredoxin domain-containing protein
1	COG4106	Transaconitate methyltransferase
1	COG5542	Predicted integral membrane protein
1	COG5650	Predicted integral membrane protein
Woesearchaeota AR20		
22	COG1215	Glycosyltransferases, probably involved in cell wall biogenesis
9	**COG0438**	**Glycosyltransferase**
6	**COG2226**	**Methylase involved in ubiquinone/menaquinone biosynthesis**
5	COG4243	Predicted membrane protein
3	COG0463	Glycosyltransferases involved in cell wall biogenesis
3	COG2244	Membrane protein involved in export of O-antigen/teichoic acid
3	COG2510	Predicted membrane protein
2	COG2511	Archaeal Glu-tRNAGln amidotrans. Sub. E (contains GAD domain)
2	COG3177	Uncharacterized conserved protein
1	COG0262	Dihydrofolate reductase
1	COG0719	ABC-type trans. sys. inv. in Fe-S cluster assem., permease comp.
1	COG1216	Predicted glycosyltransferases
1	COG1414	Transcriptional regulator
1	COG1437	Adenylate cyclase, class 2 (thermophilic)
1	**COG1651**	**Protein-disulfide isomerase**
1	COG1669	Predicted nucleotidyltransferases
1	COG1814	Uncharacterized membrane protein
1	COG1898	dTDP-4-dehydrorhamnose 3,5-epimerase and related enzymes
1	COG2129	Predicted phosphoesterases, related to the Icc protein
1	COG2259	Predicted membrane protein
1	COG2887	RecB family exonuclease
1	COG3882	Predicted enzyme involved in methoxymalonyl-ACP biosynthesis
Lokiarchaeum		
14	COG1449	Alpha-amylase/alpha-mannosidase
12	COG0064	Asp-tRNAAsn/Glu-tRNAGln amidotrans. B sub. (PET112 hom.)
6	COG0178	Excinuclease ATPase subunit
4	COG0642	Signal transduction histidine kinase
3	COG3259	Coenzyme F420-reducing hydrogenase, alpha subunit
2	COG0084	Mg-dependent DNase
1	COG0125	Thymidylate kinase
1	COG0171	NAD synthase
1	COG0183	Acetyl-CoA acetyltransferase
1	COG0322	Nuclease subunit of the excinuclease complex
1	COG0334	Glutamate dehydrogenase/leucine dehydrogenase
1	COG0674	Pyruvate:ferredoxin oxidoreductase, related oxidored., alpha sub.
1	COG0714	MoxR-like ATPases
1	COG1042	Acyl-CoA synthetase (NDP forming)
1	COG1690	Uncharacterized conserved protein

highly fragmented genomes is possible, wide-spread sequence misbinning in the CPR dataset seems unlikely. Furthermore, contigs containing genes with exclusive similarity to archaeal ones appeared to be of generally high quality—only several had a coverage below 5, with most higher (median coverage between 9 and 11, depending on the twin). Finally, only three genes in the examined twins were labeled as "Possibly archaeal contamination" in the original study. Overall, our contig majority analyses revealed that sequence binning was likely accurate.

Thus, we observed multiple twins exclusively associating CPR sequences to genes in a variety of archaea, including those from both major taxonomic groups as well as novel, ultrasmall DPANN. In other words, numerous ultrasmall bacteria presented genes exclusively similar to those of archaeal groups in their genomes. Ultimately, the results of the BGA, when combined with the binning and contig analyses, suggest that this similarity between CPR and known prokaryotic genomes may be the result of multiple interdomain LGT between these organisms. For one, an ancestor of Woesearchaeota AR20 and an ancestor of Diapherotrites AR10 could have exchanged genes with members of the CPR, helping in part to explain the observed patterns of similarity in Region C and D (Fig. 3). An RPSBlast of genes in these regions revealed that many coded for proteins relating to the membrane—integral proteins, cytochrome biogenesis, methylase involved in ubiquinone/menaquinone biosynthesis, and glycosyltransferases involved in cell wall biogenesis (Table 2). Common in certain ultrasmall archaeal genomes these glycosyltransferases are predicted to play a key role in synthesizing structural and signaling saccharides (Castelle *et al.*, 2015).

Under the hypothesis of LGT between CPR and archaea, the large "core" band of similarity seen across all groups in the ARC heatmap (Region A, Fig. 3) is surprising, as it includes more conserved archaeal genes like those in information storage and processing. Indeed, an RPS Blast analysis indicated that Region A included many genes that coded for ribosomal proteins, DNA polymerase, and tRNA synthethases (green on COG sidebar, also see Supporting Information Table S2). Several of these synthethases appeared to show higher homology with thermophilic archaeal classes, whereas some ribosomal genes were restricted to members of the phylum Euryarchaeota. This pattern may indicate ancient gene exchange involving CPR and some broad distributions of relatively large, varied archaea. However, it may also reflect an ancient phylogenetic relationship between CPR and archaea, if CPR are indeed relatively basal in the prokaryotic tree of life as suggested by a recent concatenation of 16 ribosomal proteins (Hug *et al.*, 2016). Nonetheless, there are several other regions apparent on the heatmap with exclusive above-threshold homology of CPR/TM6 genes with particular classes of Archaea, for example, Region E (Fig. 3) with varied similarity to members of Thermococci, Region F (Fig. 3), with ~40% similarity to the members of Archaeglobi, or Region G (Fig. 3), with varied similarity to members of Methanomicrobia. These other patterns of similarity strengthen the suggestion that ancient gene transfer may have occurred among members of the ultrasmall size fraction.

Lastly, we also recovered a large group of CPR/TM6 proteins ($n = 53$) that showed distant homology exclusively with Lokiarchaeum, which is already known to have a proteome nearly 30% homologous with bacteria (Region B, Fig. 3; Spang *et al.*, 2015). As above, these genes were placed in quality bins of diverse bacterial origin, and so interdomain gene transfer with a relative of Lokiarchaeota is a possible explanation for the observed pattern of similarity. However, the functional profile of these CPR/TM6 genes was largely different from that of the genes matching with AR10 and AR20. Genes shared exclusively between CPR/TM6 and Lokiarchaeota were composed mostly of amidotransferases involved in tRNA biosynthesis and a family of enzymes involved in carbohydrate metabolism, but lacking membrane-related genes (Table 2). We can only speculate that these different functional patterns may hint at different gene-capture mechanisms among archaea. Lokiarchaeota, if phagotrophic, could prey on a diversity of ultrasmall bacteria, while AR20/AR10 may be involved in symbiotic relationships with CPR. This could then lead to the convergent sharing of membrane-related genes compatible with such a lifestyle.

The observed results for the ARC subset are consistent with literature suggesting that ancient gene transfer from bacteria to archaea can play a major role in evolution of specific lineages (Nelson *et al.*, 1999; Lopez-García *et al.*, 2015; Nelson-Sathi *et al.*, 2015). In the striking case of the Haloarchaea, as many as 157 gene families coding for transporters were imported from Eubacteria (Nelson *et al.*, 1999). These transfers can facilitate colonization of new niche space, for example, Lopez-García *et al.* (2015) details the convergent acquisition of metabolism, transport, and membrane genes allowing adaptation to mesophilic conditions among three distant archaeal lineages. Ancient transfer of metabolic genes from bacteria to archaea has also been implicated in the origin of several major archaeal groups (Nelson-Sathi *et al.*, 2015). While polarity of any CPR/TM6-Archaea gene transfers in this dataset would be difficult to determine, transfer events among these domains are generally believed to be skewed towards those in which bacteria act as donors (Lopez-García *et al.*, 2015; Nelson-Sathi *et al.*, 2015). This may be due to adaptive gains made by use of new metabolic strategies and a lower fitness cost to archaea of incorporating foreign genetic material (Lopez-García *et al.*, 2015). Transfer of membrane-related genes could also be achieved endosymbiotically, where the symbiont (by lysis or another process) donates genes to the host. In fact, this scenario

has been suggested in the literature—as a possible step in the retention of the mitochondrial progenitor during early eukaryogenesis (Martin *et al*., 2015), or as a mechanism to regulate the cell wall of an intracellular bacterium (Husnik *et al*., 2013). Although LGT of membrane transporters was observed primarily between ultra-small donors and recipients (CPR, DPANN), we speculate that, in the case of a hypothetical CPR/TM6-large archaeon symbiosis, transfer and subsequent expression of the symbiont's transporters or other membrane-related genes could be critical.

Finally, graph analyses of the BAC and PROK subsets provided several other examples of more recent transfer between the CPR/TM6 and larger prokaryotes. We ran BGA analyses maintaining an 80% cover requirement between CPR/TM6 sequences and their homologs, but for the PROK subset generated gene families with more stringent percent identity (>=50, >=60, >=70, >=80, >=90% ID, see Supporting Information Methods). These gene families would later be detected as twin members. At 80% similarity, we observed that CPR/TM6 mannose isomerase genes paired with both genomes of methanogenic archaea like *Methanosarcina* and with complete CPR genomes (Supporting Information Table S3). We also observed several sets of CPR/TM6 genes associated with single deltaproteobacterial genomes, for example, a set of recombinases with *Hippea maritima* and a set of GTP-binding protein TypA/BipA with *Desulfomonile tiedjei*. Likewise, at 90% ID, a set of transcriptional regulators from the CPR/TM6 showed homology to a wide array of *Bacillus* genomes (these genes also showed similarity to archaea, just at a lower threshold). These patterns may indicate that CPR/TM6 have also exchanged genes with other bacteria, and expand upon Brown *et al*. (2015) proposal of a possible ribosomal protein transfer among members of the CPR.

Conclusion

Recent phylogenetic analyses have underscored the importance of studying ultrasmall microbial groups like CPR in expanding our knowledge of the tree of life (Hug *et al*., 2016). The patterns of genetic diversity and gene transfer reported in the present study contribute to this body of knowledge and bring forward a reticulate aspect of their evolution. Methods complementary to environmental metagenomics, like single cell genomics, could help to better elucidate relationships among organisms and their gene content (Stepanauskas, 2012) and ultimately shed additional light on patterns of transfer among these organisms. Furthermore, as we report the unusual membranes in a second domain of life (Castelle *et al*., 2015), we propose that these characteristics may be the result of a convergent evolutionary pressure. The ultrasmall niche may require underappreciated membrane adaptations, and further work should address the role of these proteins in adapting to or managing this lifestyle. Future analysis of massive environmental datasets from this size fraction, like that of TARA Oceans (Karsenti *et al*., 2011), could help to shed more light on gene transfer and phylogeny in these organisms and ultimately further our understanding of any drivers underlying their evolution.

Procedures

We downloaded the full dataset of CPR/TM6 proteins from the online repository (ggkbase.berkeley.edu/CPR-complete-draft/organisms) listed in Brown *et al*. (2015). We then removed sequences with mid-protein stop codons, leaving a final dataset of 637,155 proteins. We also downloaded all proteins from all complete archaeal and bacterial genomes on NCBI (4,600 genomes, 15,373,158 sequences, February 1, 2016). This NCBI dataset included the eight complete CPR genomes from the Brown *et al*. (2015) dataset but not the ~800 other draft genomes also reported in that study. Full taxonomy information for the complete genomes was retrieved from the NCBI taxonomy database (ncbi.nlm.nih.gov/taxonomy). We performed a BLAST analysis of all CPR/TM6 proteins against all proteins from the complete genomes on a distributed cluster (version 2.3.0+, with the following options: -seg yes, -soft_masking true, and -max_target_seqs 5000). We filtered these results for sequence hits $\geq$ 30% identity, $\geq$ 80% mutual cover, and e-value $\leq$ 1e-5 to retain only full sized homologs of CPR/TM6 proteins in complete prokaryotic genomes. We partitioned the CPR/TM6 proteins into ARC, BAC, PROK, and CPR/TM6 groups as explained above, de-replicating each set using cd-hit (version 4.6, -c 1 –s 1; Li and Godzik, 2006) to yield only unique CPR/TM6 sequences. PROK CPR/TM6 genes were further clustered into gene families (Supporting Information Methods).

We performed a BGA on the BLAST results for each subset, delineating groups of CPR/TM6 proteins with shared, exclusive similarity to a given set of prokaryotic genomes (Corel *et al*., 2016). This procedure defines "twins" composed of the CPR/TM6 sequences and the NCBI genomes hosting homologs of these sequences (Fig. 1). Twins were sorted on the number of included CPR/TM6 proteins and were filtered to retain those with low numbers of included NCBI genomes, as these allowed us to look more easily for candidate gene transfers. Recent gene transfer among the PROK and BAC subsets was detected using a BGA with higher identity thresholds (i.e., to be included in a twin, a link between a CPR/TM6 protein and a gene in an NCBI genome must be of >=50, >=60, >=70, >=80, or >=90 percent identity). For each CPR/TM6 protein in the ARC subset, we retrieved the identity of its home contig from the original sequence metadata and used this to retrieve all other sequences, regardless of

annotation, for those contigs. From these data, we created the "contig majority" parameter, or the highest frequency annotation on that contig among ARC, PROK, BAC, or CPR/TM6, and "contig neighbor," the annotations of the genes flanking the ARC gene on that contig (Supporting Information Methods). For the ARC, PROK, and *Babela*-associated CPR/TM6 genes, we performed an RPS BLAST (version 2.3.0+, with the following options: -seg yes, -soft_masking true and -max_target_seqs 5) with the NCBI COG database (ncbi.nlm.nih.gov/COG/) to obtain full gene annotations and COG categories. Finally, bin analyses were performed for the relevant gene subsets by retrieving the original sequence headers from Brown *et al*. (2015) and extracting bin/taxonomy information.

Author contributions

All authors designed the study. E.C., A.L.J., and J.S.P. performed the bioinformatic analyses; A.L.J. and E.B. wrote and revised the manuscript. All authors discussed results and commented on the manuscript.

Acknowledgements

We thank two anonymous reviewers for their insightful and constructive comments. We also thank Raphaël Méheust and Adrien Danzon for their aid in designing bioinformatic analyses. E.C., J.S.P., and E.P. were funded by the European Research Council (FP7/2007-2013 Grant Agreement 615274) and A.L.J. by the Alex G. Booth Traveling Scholarship and the Benjamin Franklin Travel Grant.

Data availability

Results of the bipartite graph analyses will be made available at http://www.evol-net.fr/index.php/en/downloads. Sequence files available upon request.

References

Anantharaman, K., Brown, C.T., Burstein, D., Castelle, C.J., Probst, A.J., Thomas, B.C., *et al*. (2016) Analysis of five complete genome sequences for members of the class Peribacteria in the recently recognized Peregrinibacteria bacterial phylum. *PeerJ* **4:** e1607.

Brown, C.T., Hug, L.A., Thomas, B.C., Sharon, I., Castelle, C.J., Singh, A., *et al*. (2015) Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523:** 208–211.

Castelle, C.J., Wrighton, K.C., Thomas, B.C., Hug, L.A., Brown, C.T., Wilkins, M.J., *et al*. (2015) Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr Biol* **25:** 690–701.

Corel, E., Lopez, P., Méheust, R., and Bapteste, E. (2016) Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution. *Trends Microbiol* **24:** 224–227.

Gong, J., Qing, Y., Guo, X., and Warren, A. (2014) "Candidatus Sonnebornia yantaiensis", a member of candidate division OD1, as intracellular bacteria of the ciliated protist Paramecium bursaria (Ciliophora, Oligohymenophorea). *Syst Appl Microbiol* **37:** 35–41.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., *et al*. (2016) A new view of the tree of life. *Nat Microbiol* 16048.

Husnik, F., Nikoh, N., Koga, R., Ross, L., Duncan, R.P., Fujie, M., *et al*. (2013) Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell* **153:** 1567–1578.

Karsenti, E., Acinas, S.G., Bork, P., Bowler, C., De Vargas, C., Raes, J., *et al*. (2011) A holistic approach to marine ecosystems biology. *PLoS Biol* **9:** e1001177.

Koonin, E.V. (2015) Archaeal ancestors of eukaryotes: Not so elusive any more. *BMC Biol* **13:** 84.

Lake, J.A. (2009) Evidence for an early prokaryotic endosymbiosis. *Nature* **460:** 967–971.

Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22:** 1658–1659.

López-García, P., Zivanovic, Y., Deschamps, P., and Moreira, D. (2015) Bacterial gene import and mesophilic adaptation in archaea. *Nat Rev Microbiol* **13:** 447–456.

Luef, B., Frischkorn, K.R., Wrighton, K.C., Holman, H.Y.N., Birarda, G., Thomas, B.C., *et al*. (2015) Diverse uncultivated ultra-small bacterial cells in groundwater. *Nat Commun* **6:** 6372.

Martin, W.F., Garg, S., and Zimorski, V. (2015) Endosymbiotic theories for eukaryote origin. *Philos Trans R Soc Lond B Biol Sci* **370:** 20140330.

Nasir, A., Kim, K.M., and Caetano-Anollés, G. (2015) Lokiarchaeota: eukaryote-like missing links from microbial dark matter? *Trends Microbiol* **23:** 448–450.

Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., *et al*. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. *Nature* **399:** 323–329.

Nelson-Sathi, S., Sousa, F.L., Roettger, M., Lozada-Chávez, N., Thiergart, T., Janssen, A., *et al*. (2015) Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* **517:** 77–80.

Pagnier, I., Yutin, N., Croce, O., Makarova, K.S., Wolf, Y.I., Benamar, S., *et al*. (2015) Babela massiliensis, a representative of a widespread bacterial phylum with unusual adaptations to parasitism in amoebae. *Biol Direct* **10:** 1–17.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.F., *et al*. (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499:** 431–437.

Spang, A., Saw, J.H., Jørgensen, S.L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A.E., *et al*. (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521:** 173–179.

Stepanauskas, R. (2012) Single cell genomics: an individual look at microbes. *Curr Opin Microbiol* **15:** 613–620.

Swithers, K.S., Fournier, G.P., Green, A.G., Gogarten, J.P., and Lapierre, P. (2011) Reassessment of the lineage fusion hypothesis for the origin of double membrane bacteria. *PloS One* **6:** e23774.

Timmis, J.N., Ayliffe, M.A., Huang, C.Y., and Martin, W. (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet* **5:** 123–135.

Velimirov, B. (2001) Nanobacteria, Ultramicrobacteria and Starvation Forms: A Search for the Smallest Metabolizing Bacterium. *Microbes Environ* **16:** 67–77.

Weber, A.P. and Fischer, K. (2007) Making the connections–the crucial role of metabolite transporters at the interface between chloroplast and cytosol. *FEBS Lett* **581:** 2215–2222.

Yeoh, Y.K., Sekiguchi, Y., Parks, D.H., and Hugenholtz, P. (2015) Comparative genomics of candidate phylum TM6 suggests that parasitism is widespread and ancestral in this lineage. *Mol Biol Evol.* **33:** 915–927.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site.

Fig. S1. Using multi-step BLAST to create 'gene families'.

Table S1. Characteristics of each 'twin' or gene grouping described in the text, including the number of genes in each twin/grouping, the number of unique genomic bins associated with the twin/grouping and their taxonomic composition, and the number of unique contigs associated with the twin/grouping and their BLAST-assigned phylogenetic annotations.

Table S2. Functional description and contig majority/neighbor information for 62 CPR/TM6 proteins in Region A of Fig. 3, matching exclusively with archaeal genomes. A contig majority marker including a '/' indicates a contig where multiple gene types are 'most frequent'.

Table S3. An exemplar 'twin' from the PROK BGA analysis at $\geq$ 80% identity. Each CPR/TM6 protein listed in the first table shows exclusive similarity at this threshold with a gene in all of the genomes listed in the second table, which includes both archaea and a member of the CPR. This pattern suggests possible inter-domain gene transfer. Also noted is the BLAST-assigned phylogenetic annotation for the neighbor of each gene, and the most frequent annotation on the contig that contains it. As above, a contig majority marker including a '/' indicates a contig where multiple gene types are 'most frequent'.