ARTICLES
DOI: 10.1038/s41559-017-0237-0
nature ecology & evolution

Parallel genome reduction in symbionts descended from closely related free-living bacteria

Vittorio Boscaro[1,2], Martin Kolisko[1,3], Michele Felletti[4], Claudia Vannini[2], Denis H. Lynn[5,6] and Patrick J. Keeling[1*]

Endosymbiosis plays an important role in ecology and evolution, but fundamental aspects of the origin of intracellular symbionts remain unclear. The extreme age of many symbiotic relationships, lack of data on free-living ancestors and uniqueness of each event hinder investigations. Here, we describe multiple strains of the bacterium *Polynucleobacter* that evolved independently and under similar conditions from closely related, free-living ancestors to become obligate endosymbionts of closely related ciliate hosts. As these genomes reduced in parallel from similar starting states, they provide unique glimpses into the mechanisms underlying genome reduction in symbionts. We found that gene loss is contingently lineage-specific, with no evidence for ordered streamlining. However, some genes in otherwise disrupted pathways are retained, possibly reflecting cryptic genetic network complexity. We also measured substitution rates between many endosymbiotic and free-living pairs for hundreds of genes, which showed that genetic drift, and not mutation pressure, is the main non-selective factor driving molecular evolution in endosymbionts.

Free-living, pleomorphic strains of the betaproteobacterium *Polynucleobacter* are abundant in the plankton of freshwater lentic habitats[1,2]. They can be cultivated under laboratory conditions and have recently been investigated at the genomic level[2,3]. Morphologically distinct strains of *Polynucleobacter* are also known to be intracellular symbionts of several species of the ciliate genus *Euplotes*[4–6]. The *Euplotes* depend on endosymbionts for survival[5] and endosymbiotic *Polynucleobacter*, which are vertically transmitted during host division, have never been isolated or cultured outside their hosts[7], suggesting that the relationship is obligate for both partners.

Despite their distinct morphology, endosymbiotic and free-living *Polynucleobacter* are so similar at the molecular level that until recently they were classified as the same species[8]. Single-marker phylogenetic inferences do not fully resolve their evolutionary relationships[7,9], but symbiotic strains do not form a monophyletic group in small subunit ribosomal RNA (rRNA) gene trees, leading to speculation that transitions to symbiosis occurred multiple times[9].

Our understanding of the process of endosymbiotic integration is often based on comparisons between distantly related symbionts that evolved from unlike ancestors under different evolutionary pressures[10,11]. Individual symbiotic lineages have also been compared with their closest free-living[12] or opportunistic[13,14] relatives. The first approach gives us insights into the universal features of endosymbioses, but its resolving power is limited by the 'noise' imposed by distantly related ancestral states. The second approach reveals how each evolutionary event unfolded, but is intrinsically non-reproducible and often limited by long spans of time separating the free-living ancestor and modern endosymbionts. If *Polynucleobacter* endosymbionts arose multiple times from similar and closely related ancestors, this natural system could represent something close to 'replaying the tape' of an evolutionary process[15]. Each symbiotic lineage would be an independent example that could be compared with multiple close relatives, both free-living and endosymbiotic. This would provide an ideal combination of the two traditional approaches to assess fundamental questions in the process of endosymbiotic reduction.

The loss of DNA and genes is widespread in obligate symbionts, but its causes have been a matter of debate. Endosymbionts usually have smaller population sizes and a reduced number of essential functions[12,16]—factors that weaken the power of natural selection and trigger the accumulation of deleterious mutations[17], ultimately leading to gene inactivation and loss of now non-functional DNA when there is no strong counterbalancing selection acting on the host[18]. Thus, the commonly accepted model, mostly based on symbioses between insects and bacteria[19,20], explains even the most extreme reductions in genome and proteome size as a result of non-adaptive mechanisms. Many of the predictions of this model have been tested in various systems[21–23], but the relative contributions of different critical processes remain elusive. Some authors consider genetic drift to be the major catalyst for genome erosion[12,24]. Others propose that elevated mutation pressure may play an additional or prevalent role[25,26]. Distinguishing between the two factors requires the measurement of non-synonymous (dN) and synonymous (dS) substitution rates in pairwise comparisons of free-living and symbiotic bacteria. Unfortunately, in systems studied so far, either synonymous sites are largely saturated[21,25] due to the divergence of the two species, or there is no suitable outgroup for a direct comparison[14]. *Polynucleobacter* provides the opportunity to examine this and other questions about the process of genomic transformation accompanying endosymbiosis.

Results

Phylogenomics confirms that symbioses originated multiple times independently in *Polynucleobacter*. We collected and cultured eight strains of *Euplotes* and obtained the genomic sequences

[1]Department of Botany, University of British Columbia, Vancouver, BC V6T 1Z4, Canada. [2]Department of Biology, University of Pisa, Pisa 56126, Italy. [3]Institute of Parasitology, Biology Centre, Czech Academy of Sciences, Prague 370 05, Czech Republic. [4]Department of Chemistry, University of Konstanz, Konstanz 78464, Germany. [5]Department of Integrative Biology, University of Guelph, Guelph, ON N1G 2W1, Canada. Present address: [6]Department of Zoology, University of British Columbia, Vancouver, BC V6T 1Z4, Canada. *e-mail: pkeeling@mail.ubc.ca

of their *Polynucleobacter* symbionts. We added to our analyses available genomes from one symbiotic strain (Stir1; ref. [27]) and seven free-living strains (three of which are complete: QLW-P1DMWA-1 (ref. [2]), CB (ref. [28]) and MWH-MoK4 (ref. [3])) (Fig. 1).

Phylogenomic trees of these taxa inferred from 612 protein-coding genes using multiple methods and models produced topologies differing in only a few nodes (Fig. 2). In the reconstruction of ancestral states, we assumed that all ancestors of free-living bacteria were also free living; that is, that no endosymbiont reverted to a free-living state. In support of this, all genomes from symbionts are characterized by a massive loss of functional genes, impaired vital pathways and an abundance of pseudogenes (see the following section). Genomes from free-living strains are consistently larger and possess a common set of related, unimpaired pathways with few pseudogenes[2,8]. For any free-living strain to have passed through an obligate symbiotic phase to regain the free-living state would require it to accomplish the following highly unlikely conditions: reacquire hundreds of vital functional genes phylogenetically related to those of other *Polynucleobacter* strains; lose remnants of as many pseudogenes; and regain a gene set and genome size similar to those of the free-living strains that never had a symbiotic stage.

Our phylogenomic analyses (Fig. 2) showed at least eight independent origins of symbiosis among the nine symbionts with sequenced genomes. Nodes that varied in different analyses or were unsupported in one analysis did not affect the minimum estimated number of symbiotic origins despite minor differences among the models and methods. In none of the inferred topologies were two symbiotic strains sister lineages except for Eae5 and Eae1. Overall, all analyses confirmed that the *Euplotes*–*Polynucleobacter* obligate relationship originated numerous times. The independent establishment of symbiosis in strains Stir1, Ewo1, Eae3, Eoc1 and Fsp1.4, which had already been proved by the phylogenetic position of the free-living strain MWH-JaK3, was further supported by their different host species.

The host *Euplotes* species forms a robust monophyletic group[29] (Supplementary Fig. 1) and, while the exact functional basis for the symbiosis is unknown, it is parsimonious to assume its origin is in the common ancestor of this clade. As the symbiosis is obligate for the host, it is likely that after the original establishment, these *Euplotes* species always harboured essential symbionts. The non-monophyly of symbiotic *Polynucleobacter* therefore implies that distinct lineages invaded *Euplotes* strains that already possessed a symbiont—either another *Polynucleobacter* or a different bacterium[9]—and replaced it. This contrasts with other systems, where ancient symbionts formed multipartite relationships with new symbionts (for example, the *Buchnera*–*Serratia* consortium in aphids[13,30]).

Gene loss precedes genome reduction in *Polynucleobacter* symbionts. Complete genomes from extant free-living *Polynucleobacter* share most genes and pathways (Supplementary Fig. 2). Most variation is restricted to strain-specific genes, about half of which encode predicted open reading frames with no assigned function. Other variation corresponds to pathways of considerable biological relevance[2,3,8] for which the associated genes are not present in any other free-living or symbiotic strain. This suggests that environmental *Polynucleobacter* are metabolically diverse and can adapt to various niches. Nevertheless, their diversity is probably short-lived

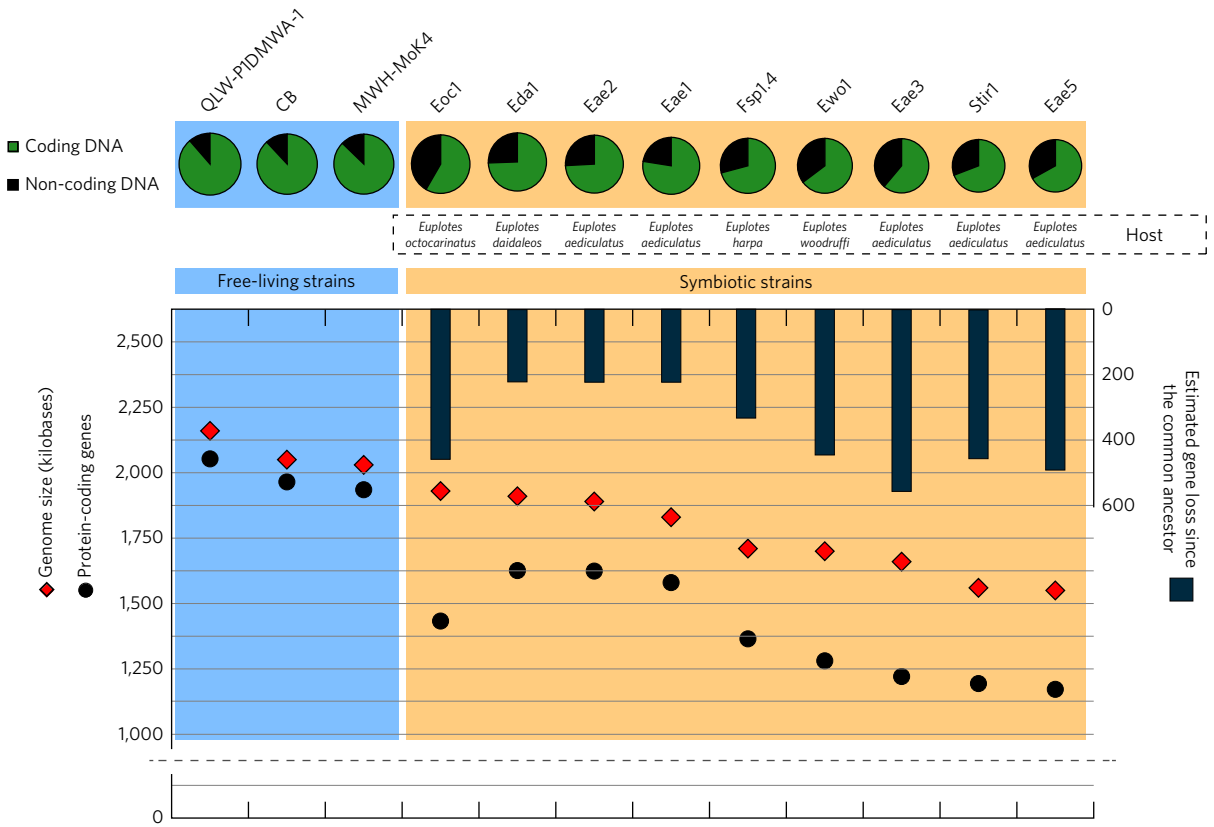

Figure 1 | Genomic reduction in symbiotic *Polynucleobacter* strains. Free-living strains are shown in blue and symbiotic strains in orange. Top: the proportion of non-coding DNA. Bottom: the genome size (red diamonds) and the number of protein-coding genes (black circles) (left axis; numbers correspond to both gene number and the number of kilobases). The black bars indicate gene loss (right axis; gene number). Symbionts showed moderately reduced genome size but extensive gene loss, which was reflected by higher percentages of non-coding DNA (mostly pseudogenes).

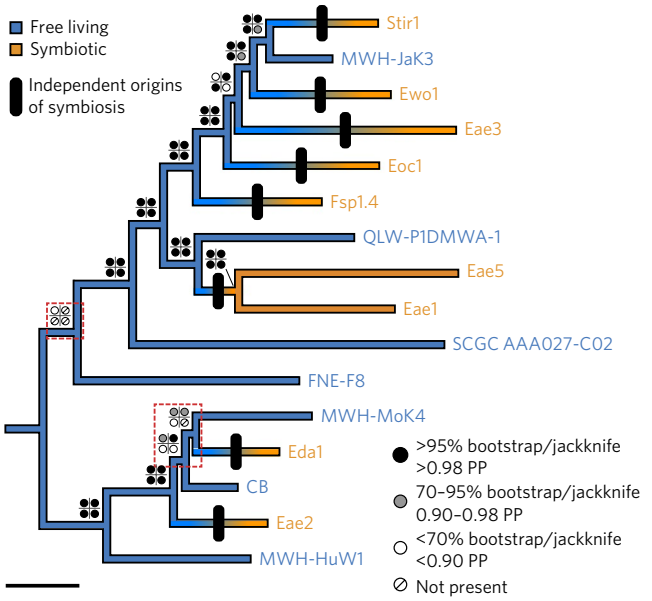

Figure 2 | Phylogenomic tree of symbiotic and free-living *Polynucleobacter*. The tree shown is the maximum likelihood topology inferred with RAxML (LG4X + GAMMA model) on 612 protein-coding genes. The four circles at each node show support values obtained with different methods (from the top left clockwise: RAxML, LG4X + GAMMA model; IQ-TREE, C60 mixed model; RAxML, jackknifed 100-gene consensus; and PhyloBayes, CAT-GTR model). The dashed red rectangles highlight less reliable nodes. Ancestral states were inferred by parsimony, assuming that the ancestors of free-living strains were also free living. The black bars mark branches where symbioses were established. PP, posterior probability. Scale bar: 2% divergence.

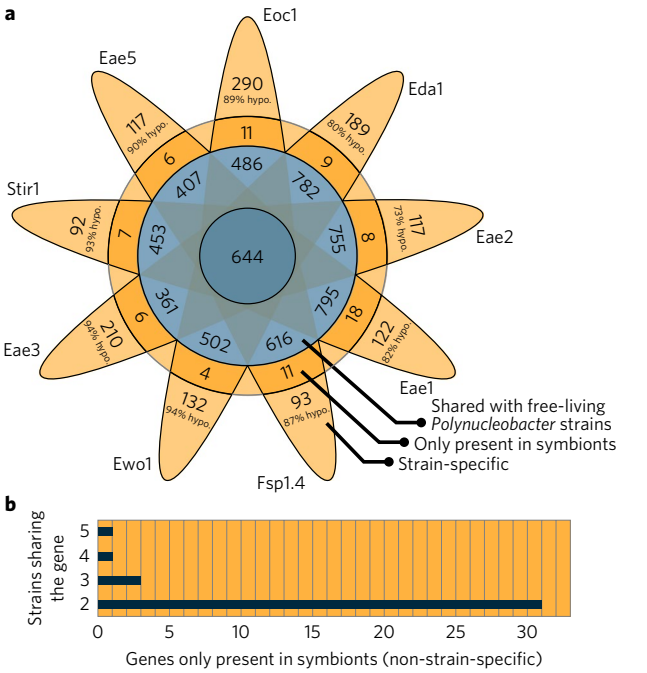

Figure 3 | Protein-coding genes shared by symbiotic *Polynucleobacter* strains. a, Genes in free-living and symbiotic strains are shown in the blue regions, while genes only found in symbiotic strains are shown in the orange regions. From the *centre* outward: genes shared by all 12 investigated *Polynucleobacter* strains with a complete genome; genes in each symbiotic strain shared with at least one free-living relative; genes in each symbiotic strain shared only with other symbionts; and strain-specific genes. Of the symbiont-specific genes, most were strain-specific (light orange) and the great majority of these had no known function. hypo., genes coding for hypothetical open reading frames. **b**, The distribution of symbiont-specific genes between multiple symbionts (dark orange ring in **a**).

on the evolutionary scale while a large fraction of core genes remain conserved[31] (which is not unusual for bacteria). The core includes about 1,488 genes, most likely present in the ancestor of all investigated *Polynucleobacter* strains. Hence, it can be used to conservatively count minimal gene losses in symbionts (Fig. 1) with a smaller margin of error than direct comparisons with extant strains, which might have diverged from a common ancestor with additional recently acquired genes.

A universally recognized feature of the genomes of endosymbiotic bacteria is a reduction in gene number and genome size[20,32], which were also reported in a previous comparison of *Polynucleobacter* genomes from one free-living and one symbiotic strain[27]. In the present study, all nine symbionts had smaller genomes with reduced gene content and an elevated proportion of non-coding DNA compared with free-living strains (Fig. 1). The non-coding fraction largely consisted of pseudogenes, and its proportion varied considerably among the strains. The correlation between genome reduction and gene loss (Pearson correlation coefficient, $r = 0.900$) and the larger extent of gene loss suggest that the two processes overlapped in time and that gene inactivation took place earlier, as has been proposed to have occurred in the initial stages of other obligate symbioses[20,33]. Conversely, the *Polynucleobacter* system differs from others at a similar stage of reduction because of the absence of transposable elements[27], which are often assumed to play an essential role in the early inactivation of genes[14,20].

Despite the extensive loss of about 200–600 genes in each symbiont, only eight genes had been lost in all nine symbiotic strains. None of these (listed in Supplementary Discussion 1) was functionally related to another, nor did they have an obvious intracellular function. Thus, it seems unlikely that the loss of any specific gene is an absolute requirement for initiating symbiosis.

Symbionts possess strain-specific genes in proportions comparable to free-living *Polynucleobacter*. However, the percentage of these as unknown open reading frames is much higher in symbionts than in free-living strains and very few complete pathways can be identified (Supplementary Discussion 1). The sequencing of additional free-living strains, especially of those more closely related to symbiotic strains, would test the hypothesis that these genes descended from ancestral strain-specific gene sets. Very few symbiont-specific genes were found in more than one symbiont, and none was shared by more than five (Fig. 3). No gain of function was therefore required for, or tied to, symbiosis.

Predominantly random pattern of gene loss and retention. Of the core ancestral genes, 644 were still present in all symbiotic strains, probably because their selective coefficients were high enough that purifying selection had prevented loss-of-function substitutions. The others had been lost in at least one lineage. To examine potential patterns, we first classified genes into broad functional categories. General trends were observed that were consistent with previous observations[27]; for example, genes related to DNA, RNA and protein metabolism or energy production were, on average, lost less frequently than transporters or genes involved in growth, sensing or regulation (Supplementary Fig. 3). However, the loss among different strains varied conspicuously within all categories, suggesting a substantial degree of heterogeneity behind these general trends, as would be expected in the initial stages of obligate symbiosis.

To inspect these differences more deeply, we identified 237 modules that grouped functionally related proteins—enzymes in

the same pathway or components of multimeric complexes. At this level, stochastic loss was prominent and consistent trends were much more difficult to recognize, as the pattern of losses differed for each strain (Supplementary Data 1). Clustering analyses confirmed that, in contrast to phylogeny, free-living strains were more similar to each other than to any symbiont in terms of modules and gene content (Supplementary Figs. 4 and 5). Moreover, clustering of the symbiotic strains according to either gene content or functional modules did not match the phylogeny of the bacteria or their hosts. Notably, five strains from the same host species (*Euplotes aediculatus*) did not cluster together, suggesting that functional variation was not tied to host physiology. Clustering order did not reflect genome size well, but showed some correlation with the number of protein-coding genes. Unsurprisingly, the three strains with the biggest proteomes were closer to the free-living strains in the clustering of functional modules, and Eae5 and Stir1, the two strains with the smallest proteomes, were associated in both graphs. Thus, phylogenetically unrelated symbionts started from a gene content similar to those of extant free-living strains, and might eventually converge at the end of the reduction process when most non-essential functions have already been lost[34].

Deeper still, at the level of individual genes, non-functional modules often retained some components and were disrupted in different ways in different strains (Fig. 4 and Supplementary Fig. 6). In some instances, the preservation of individual genes in incomplete modules did not warrant special explanation because these genes had other known, conserved functions. For example, only strain Fsp1.4 possessed the full glyoxylate cycle, but all symbionts retained the three genes that are shared between this pathway and the tricarboxylic acid cycle. Similarly, no symbiont had a complete gene set for the synthesis of thiamin, but all still possessed the two genes in the module that are also required for the biosynthesis of pyridoxine, terpenoids and iron–sulfur clusters.

In other instances, retained genes in disrupted modules had no other known function. Sometimes these were remarkably conserved among symbiotic strains that lacked the corresponding module. Examples include *moeB*, which is apparently functional in all eight strains unable to synthesize the coenzyme MoCo, and the genes *soxY* and *soxZ*, which are preserved in strains without a complete *sox* operon. Non-functional genes are thought to degenerate quickly[35–37], so the retention of the same gene in a non-functional context in multiple strains is of interest. It is possible that such genes were retained because of metabolic portioning with the host (as described in other systems[38,39]), where the host provides missing metabolites or activities to the bacteria. However, this explanation cannot be applied to all cases; for example, the sulfur-oxidizing *sox* pathway is not present in eukaryotes so the host cannot compensate for missing components. Alternatively, some of these genes may possess unrecognized secondary functions, either in other pathways or as regulators. These exceptions to the otherwise stochastic disassembling of modules may therefore reveal unrecognized interactions in cryptic genetic networks that could be investigated in the future.

Genetic drift, but not mutation pressure, drives the molecular evolution of symbionts. It is commonly observed that symbionts have higher rates of amino acid sequence evolution compared with free-living relatives[12,25]. It has been debated whether this is due to an overall increase in the total number of substitutions (indicated by an elevated dS and evidence of higher mutation pressure) or genetic drift (indicated by an elevated dN/dS ratio) or both. For most symbiotic systems these cannot be distinguished because the free-living and symbiotic strains are too distant and dS is saturated[21,25]. *Sodalis praecaptivus*, an opportunistic human pathogen, has many non-saturated synonymous sites compared with its close relatives in the *Sodalis*-allied clade of insect symbionts[14], but the absence of

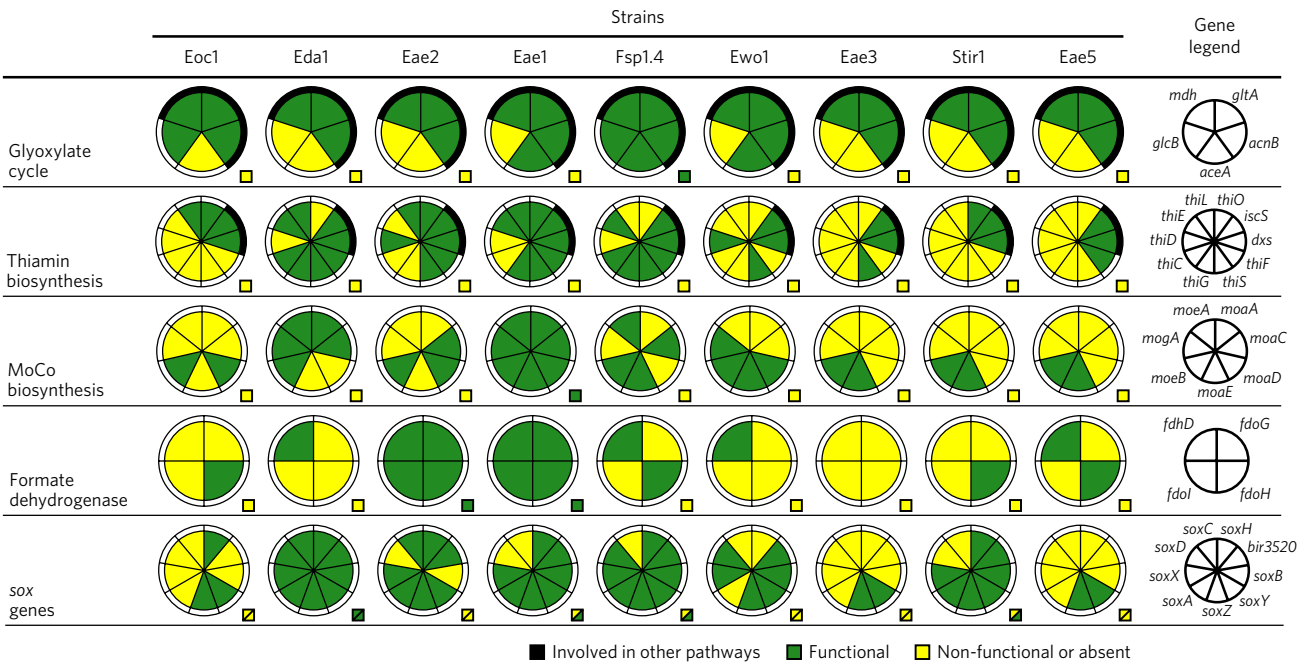

Figure 4 | Examples of within-module gene losses in symbiotic *Polynucleobacter*. Each pie chart represents a functional module (for example, a biochemical pathway or multi-subunit complex) and the sections represent the presence (green) or absence (yellow) of essential genes that make it up (the gene names are indicated in the last column). The small squares display the status (functional, nonfunctional or absent) of the module as a whole. The *sox* gene cluster can be found in two different variants: a shorter version containing *soxBXAZY* and a longer operon including *soxD* and *soxC* (see Supplementary Discussion 1). To reflect this, the squares on the bottom row are divided into two. The outer ring indicates whether additional functions are known (black) or unknown (white) for that gene. More examples are shown in Supplementary Fig. 6.

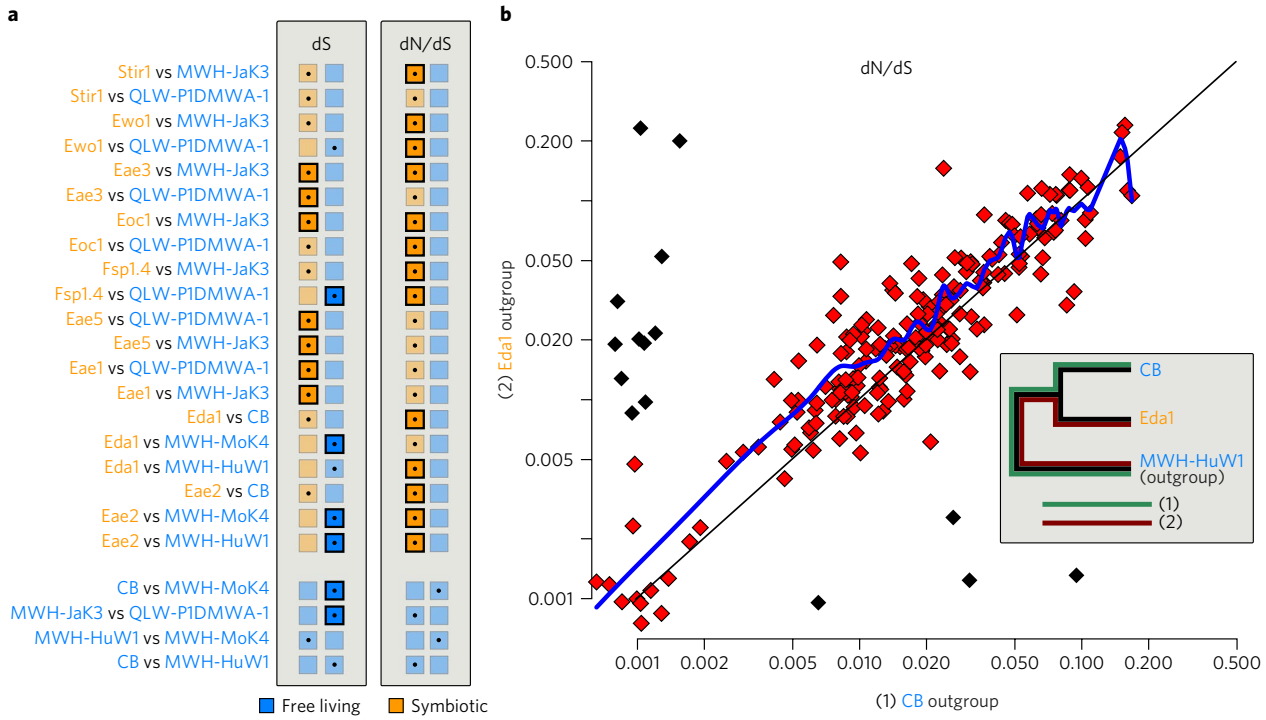

Figure 5 | Pairwise comparisons of dS and dN/dS values. a, 24 pairwise comparisons of symbiotic versus free-living *Polynucleobacter* strains and free-living pairs. The columns show the dS results (left) and dN/dS ratios (right). For each comparison, orange boxes indicate symbionts and blue boxes indicate free-living strains. A dot in the box indicates the strain with a higher mean value and bold square borders indicate statistically significant differences (*P* values < 0.05 after Bonferroni correction). The numerical data are reported in Supplementary Data 2. **b**, The graph shows one example of dN/dS values for 219 orthologous genes in the symbiotic strain Eda1 and the free-living strain CB. The inset depicts the approach used to calculate the dS and dN/dS values, comparing the same gene for each strain of the pair with the orthologous gene in the outgroup. The blue trend line, which was generated using the smooth.spline function of the R package, lies above the line of equality. The black points represent outliers (see Methods).

a suitable outgroup and the status of *S. praecaptivus* as the only characterized free-living bacterium in the clade prevent direct and repeated pairwise comparisons in that system. Free-living and endosymbiotic *Polynucleobacter* strains are so similar at the nucleotide level that substitution rates can be estimated for hundreds of genes. Knowing with confidence the phylogenetic relationships between the strains additionally permits many independent measurements.

We performed 20 pairwise comparisons between free-living and symbiotic strains, as well as four pairwise comparisons among free-living strains as controls (Fig. 5a and Supplementary Data 2). For dS values, statistically significant differences existed between some of the controls, highlighting strain-specific variability even between free-living *Polynucleobacter*. The symbiotic strains had significantly higher dS mean values than their free-living relatives in seven comparisons and significantly lower mean values in four comparisons. The remaining nine cases showed no significant difference, although mean values were sometimes higher for the symbionts (seven cases) and sometimes for the free-living strains (two cases). In contrast, there was a much more consistent pattern in the dN/dS analyses. No significant variation in the dN/dS was observed between free-living strains, while endosymbiotic *Polynucleobacter* always display higher dN/dS mean values than their free-living relatives, the difference being statistically significant in 13 of the 20 comparisons.

dN/dS values were low overall (mostly below 0.5 and never above 0.8), suggesting that if positive selection played a role at all, it was limited to a small subset of residues. Higher means in the symbionts were caused instead by an overall increase in dN across functional genes (Fig. 5b), confirming the hypothesis that genetic drift is the main non-adaptive mechanism responsible for genomic evolution in symbiotic *Polynucleobacter*. Mutation pressure may play a role, as means significantly differ among some lineages, but it appears to be a strain-specific factor not strongly correlated with endosymbiosis.

Discussion

Endosymbiotic associations have evolved for diverse reasons involving a great assortment of species, and have driven virtually all of the most extreme departures from canonical genome structure and content[20,38,40]. Despite this variability, endosymbiont genome evolution appears to share several common fundamental features, but discerning these from the tangle of resulting diversity has been challenging. The relatively recent origin of symbiotic *Polynucleobacter*, their multiple establishments, and the possibility of comparing different symbiotic lineages with closely related extant free-living bacteria provide unique glimpses into the process. We found that gene loss in independent lineages was ordered to the extent that many functions were retained, but in general genetic networks crumbled in an unpredictable pattern that was mostly stochastic, with a few interesting exceptions that may reveal cryptic molecular interactions. Multiple comparisons of *Polynucleobacter* genomes showed that endosymbiosis in this system has led to the accumulation of substitutions via genetic drift, but not necessarily mutation pressure.

Most known endosymbiotic systems arose from a unique event and seem to have evolved towards stable integration between partners. Examples have recently been found, as described here and elsewhere[23,33,41], of tight associations that were instead established through independent transitions to endosymbiosis. The *Polynucleobacter*–*Euplotes* symbiosis is an interesting example of an obligate relationship that not only originated multiple times, but involved free-living strains invading already symbiotic systems and replacing a pre-existing bacterial partner.

Methods

Collection, culturing and screening of *Euplotes*. *Euplotes* populations were collected in freshwater environments in Tuscany, Italy. Monoclonal strains were cultured starting from single cells, as previously described[9]. Ciliates were screened using fluorescence in situ hybridization[6] to assess the presence of *Polynucleobacter* bacteria in the cytoplasm (Supplementary Fig. 7). Seven positive strains (Eae1, Eae2, Eae3, Eae5, Eda1, Eoc1 and Ewo1) were selected from seven populations. Species were assigned according to in vivo observations and by sequencing the small subunit rRNA gene[42]. Supplementary Fig. 1 shows the phylogeny of the *Euplotes* strains as inferred from this marker. Sequences were aligned and trimmed with the softwares MAFFT[43] and trimAl[44], respectively; a maximum likelihood topology was obtained with IQ-TREE[45] (searching for the best-fitting evolutionary model) and a Bayesian topology was obtained with MrBayes[46] (three runs with three heated and one cold chain each, iterated for 1,000,000 generations). Small subunit rRNA gene sequences of the symbiotic bacteria were also obtained through PCR with betaproteobacterial-specific primers[6].

Genomic DNA extraction and sequencing. The seven new strains, plus the *Euplotes harpa* monoclonal strain Fsp1.4 previously characterized[6], were mass cultured for genomic DNA extraction. Filtered and starved cultures (about 2 l each) were treated with 0.2 mg ml^{-1} chloramphenicol overnight to decrease bacterial contamination in the culture medium[27]. Ciliates were then concentrated and mechanically lysed with a syringe. The homogenate was subjected to several rounds of centrifugation at increasing acceleration to remove most of the hosts' cellular debris. *Polynucleobacter* bacteria were collected during a final centrifugation at 10,000 *g* for 10 min. Light microscopy and fluorescence in situ hybridization observations on all pellets confirmed that the last precipitate contained the highest number of *Polynucleobacter* and the lowest amount of contamination from the host. Total DNA was extracted using the NucleoSpin Plant II DNA extraction kit (Macherey-Nagel). Agarose gel electrophoresis confirmed that the amount of *Euplotes* DNA, identifiable as a smear due to its fragmented genome[47] decreased with successive centrifugations.

DNA yields were measured using the Qubit 2.0 Fluorometer and the dsDNA HS Assay Kit (Life Technologies). Libraries were constructed with the Nextera XT DNA Library Preparation kit (Illumina) and sequenced by Génome Québec on a MiSeq platform (2×250 paired-end). Standard (Sanger) small subunit rRNA gene sequencing of each culture was repeated after this step to ensure that no contamination had occurred.

Genome assembly and annotation. Genome assembly was performed using the software A5 (ref. [48]). Two to four large scaffolds for each strain shared high sequence similarities with available *Polynucleobacter* genomes and were assembled manually into circular chromosomes. No removal of contaminant reads (from the hosts or bacteria in the medium) was required. For three genomes (strains Eae2, Fsp1.4 and Eda1) the final assembly was not circular but linear, potentially highlighting the existence of one gap in each. Two such gaps were present in the genome of strain Eae3. An analysis using the software BLAST of the nucleotide sequences flanking the gaps against complete genomes suggested that all the potential gaps were small (less than 1.5% of the total genome). They fell in gene-poor regions and gene synteny comparisons strongly indicated that they were unlikely to contain any gene. Sequence data were deposited in the European Nucleotide Archive (accession number: PRJEB15088).

Preliminary gene annotation and amino acid sequence prediction were performed with RAST[49]. To improve comparability, the previously sequenced genomes of the symbiotic *Polynucleobacter* strain Stir1 and the free-living strains QLW-P1DMWA-1, CB and MWH-MoK4 were annotated again with the same software. Gene homology among the strains was identified with OrthoFinder[50]. Functional annotations were then manually curated verifying the correspondence of results for orthologous genes within *Polynucleobacter* and with the first ten non-*Polynucleobacter* BLAST hits in the database. In the nine symbiotic strains, a gene was considered non-functional (a pseudogene) if its longest open reading frame was less than 80% or more than 125% of the length of its shortest and longest homolog in the free-living strains. Gene lengths among free-living strains were, on average, much more consistent. A small number of genes with higher variability were removed from the quantitative analyses.

Three nested gene sets were built. One ('CORE') contained only genes present in all 12 *Polynucleobacter* strains with a complete genome. The second ('ANCESTOR') contained genes present in all three free-living strains, regardless of their status in symbionts. These genes were assumed to be present in the (free-living) ancestor of all investigated *Polynucleobacter* strains. The third dataset ('COMPLETE') contained all genes.

Sequence retrieval from additional taxa. Orthologous amino acid sequences in four incomplete genomes of free-living *Polynucleobacter* (FNE-F8[51], SCGC AAA027-C02[52], MWH-JaK3 and MWH-HuW1[8]) and outgroup species (genera *Ralstonia* and *Cupriavidus*; accession numbers: GCA_000009285.2, GCA_000020205.1, GCA_000069785.1, GCA_000196015.1, GCA_000832305.1, GCA_000009125.1 and GCA_000954135.2) were identified by reciprocal-best BLAST of the free-living strain CB against the protein predictions of the particular taxon. The following protocol was used when gene predictions were not available (that is, for *Polynucleobacter* strains MWH-JaK3, FNE-F8 and SCGC AAA027-C02): (1) contigs from each assembly were translated in all six frames; (2) BLASTp (maximum *E* value: 10^{-5}) of strain CB's sequences were used to identify and extract amino acid sequences for each protein of interest in the translated contigs; (3) the extracted amino acid sequences were reciprocally BLASTed against proteins predicted in the CB genome; (4) only proteins with a reciprocal best hit were kept; and (5) gene nucleotide sequences were extracted from the contig according to the coordinates of the corresponding protein sequence.

Metadata associated with several deposited free-living *Polynucleobacter* genomes confirmed the homogeneity of some of their features. Estimated protein-coding percentages were: QLW-P1DMWA-1: 92.6%, CB: 91.9%, MWH-MoK4: 91.7% (confirmed by our reannotation; Fig. 1), MWH-JaK3: 90.7% and MWH-HuW1: 91.2%. Estimated pseudogenes in the same strains were more variable in number (3–46), but were always far fewer than in symbionts, and did not affect coding percentage nearly as much.

BLASTn and BLASTx analyses were conducted on all genes listed in the 'ANCESTOR' table to preliminarily assess the presence of obvious cases of extensive lateral gene transfer with other bacteria. All but 13 sequences (one in strain CB, three in strain FNE-F8, four in strain QLW-P1DMWA-1, five in strain MWH-HuW1, and none in strains MWH-JaK3 and MWH-MoK4) produced *Polynucleobacter* as a first hit and all but one of these hits were other betaproteobacteria. Lateral gene transfer with distantly related bacteria thus seems to have had little or no effect on this set of genes since the divergence of the investigated strains.

Phylogenomic analyses. A total of 612 single-copy genes from the 'CORE' table were used for the phylogenomic analyses. Each single-gene amino acid dataset was aligned by the 'linsi' algorithm from MAFFT[43] and ambiguously aligned positions were removed using BMGE[53]. Trimmed alignments were concatenated into one matrix by the script Alvert from the package Barrel-O-Monkeys (http://rogerlab.biochemistryandmolecularbiology.dal.ca/Software/Software.htm#Monkeybarrel) resulting in an alignment consisting of 23 taxa and 202,245 sites. A maximum likelihood tree was constructed with the RAxML programme[54] using the LG4X model[55] with GAMMA correction for among-site rate variation (model settings PROTGAMMALG4X). Statistical support was inferred from 1,000 bootstrap replicates (-b setting) generated in RAxML using the same model. To further assess the stability of the result, analyses using different methods, models and software were performed: the LG model[56] was also used in RAxML; the protein mixture model C60[57], as implemented in IQ-TREE[45], was employed (C60+LG, obtaining identical results); and Bayesian inference was performed on PhyloBayes[58] using the CAT-GTR model and four independent runs (two runs converged with maxdiff <0.3 on one topology, and two on a topology differing by only two nodes; conflicting nodes were assigned zero support values in Fig. 2). Additionally, to test the reliance of the analysis on the chosen gene set, we constructed 250 datasets each from 100 randomly sampled genes (jackknifed). These datasets were analysed on RAxML (LG4X) and a consensus tree was built. Trees were rooted between the *Polynucleobacter* clade and the outgroups.

dS and dN/dS analyses. Of the 1,488 genes in the 'ANCESTOR' table, 1,297 single-copy genes were used for the dS and dN/dS analyses. Each single-gene amino acid dataset was aligned with the 'linsi' algorithm from the MAFFT package and then back translated into nucleotide alignment using an in-house script. Non-*Polynucleobacter* sequences used as outgroups in the phylogenomic analyses were removed before the alignment. The incomplete genome of *Polynucleobacter* strain SCGC AAA027-C02 was also excluded from the analysis due to the absence of many genes. A total of 100 bootstrap replicates were generated for each single-gene alignment and analysed in RAxML under the LG+GAMMA model; the few highly supported (>85% bootstrap) nodes obtained in this way, which were in conflict with the phylogenomic tree (Fig. 2), were removed to avoid biases from conflicting topologies possibly due to lateral gene transfer.

Maximum likelihood estimates of dS and dN were computed using the programme codeml from the PAML package[59] (runmode -2) under the F61 model, preferred over the F3×4 model according to Akaike's information criterion. To perform each comparison between two strains, the free-living strain phylogenetically closest to the pair was chosen as the outgroup. The outputs of all phylogenomic methods were considered and choices were unaffected by alternative resolutions of weakly supported nodes. dS and dN/dS values were then calculated as pairwise differences between each target strain and the outgroup. dS was considered non-saturated if the estimated value was below 2. dS values below 0.01 were also discarded, as were dN/dS values equal to 0. The remaining values were then compared between the two taxa of interest (see the insert in Fig. 5), considering only genes where values were available for both lineages.

In total, 24 pairs were compared using this approach (Supplementary Data 2). Paired Student's *t*-tests (two-tailed) were used to evaluate the significance of differences between the means of the two lineages, applying the Bonferroni correction for multiple comparisons. The values were also plotted on a scatterplot (see Fig. 5). Visual examination of the scatterplots showed the presence of potential outliers corresponding to log-ratio values outside the 'mean ± 2 × s.d.' range.

To confirm that the difference between lineages and its significance was not driven by these extremes, an outlier removal step was implemented. Statistical testing was performed again and only differences supported by both analyses were considered statistically significant. Outlier removal, plotting and *t*-tests were all performed using the package R[60].

Functional analysis. Each gene was assigned to a broad functional category[27]: (1) biosynthesis and catabolism; (2) DNA, RNA and protein metabolism; (3) energy production; (4) cell wall; (5) membrane (non-transporters); (6) transporters; (7) growth, sensing and regulation; (8) others; and (9) hypothetical. In total, 237 functional modules were also defined, including sets of genes coding for proteins involved in well-known pathways or other functional units, such as multimeric complexes or regulatory functions (Supplementary Data 1). Modules, and the assignment of genes to them, were designed consulting online resources such as the Kyoto Encyclopedia of Genes and Genomes[61] and MetaCyc[62] for metabolic pathways and InterPro[63] and BRENDA[64] for protein function and classification. A module was considered absent in a strain if one or more genes deemed essential was missing or non-functional.

A synopsis of the genomic functional analysis on *Polynucleobacter* strains, including an update on previous conclusions drawn from a single strain[27] on topics such as DNA repair pathways and symbiont-specific features is presented in Supplementary Discussion 1.

Gene and module clustering. Binary tables for the presence or absence of 4,112 genes ('COMPLETE' dataset) and 237 modules were used to compute Euclidean distance matrices followed by clustering using the neighbour-joining algorithm (performed in the R package). Presence and absence plots were generated from the binary table using the heatmap.2 function from the gplots package.

Symbiont cultivation attempts. Many failed attempts to isolate symbionts from *Polynucleobacter*-bearing *Euplotes* have been performed in the past employing several techniques[7]. Strains Stir1 and Fsp1.4 were among those previously tested. To investigate the possibility that some of the novel strains might still be able to survive outside the host, new isolation attempts were performed on Eda1 and Eae2—the symbionts with the highest number of functional protein-coding genes. Two previously described techniques—direct plating on nutrient broth soyotone yeast extract medium and a variant of the dilution–acclimatization method optimized for free-living *Polynucleobacter*[7,65]—were employed and produced negative results.

Data availability. New genomic sequences obtained during this study were deposited in the European Nucleotide Archive with the accession codes LT606946 to LT606951, LT615227 and LT615228.

Received: 15 August 2016; Accepted: 14 June 2017;
Published online: 21 July 2017

References

1. Jezberová, J. *et al.* Ubiquity of *Polynucleobacter necessarius* ssp. *asymbioticus* in lentic freshwater habitats of a heterogeneous 2000 km^2 area. *Environ. Microbiol.* **12**, 658–669 (2010).
2. Hahn, M. W. *et al.* The passive yet successful way of planktonic life: genomic and experimental analysis of the ecology of a free-living *Polynucleobacter* population. *PLoS ONE* **7**, e32772 (2012).
3. Hahn, M. W., Jezberová, J., Koll, U., Saueressig-Beck, T. & Schmidt, J. Complete ecological isolation and cryptic diversity in *Polynucleobacter* bacteria not resolved by 16S rRNA gene sequences. *ISME J.* **10**, 1642–1655 (2016).
4. Heckmann, K. & Schmidt, H. J. *Polynucleobacter necessarius* gen. nov., sp. nov., an obligately endosymbiotic bacterium living in the cytoplasm of *Euplotes aediculatus*. *Int. J. Syst. Bacteriol.* **37**, 456–457 (1987).
5. Heckmann, K., Ten Hagen, R. & Görtz, H.-D. Freshwater *Euplotes* species with a 9 type 1 cirrus pattern depend upon endosymbionts. *J. Protozool.* **30**, 284–289 (1983).
6. Vannini, C., Petroni, G., Verni, F. & Rosati, G. *Polynucleobacter* bacteria in the brackish-water species *Euplotes harpa* (Ciliata Hypotrichia). *J. Eukaryot. Microbiol.* **52**, 116–122 (2005).
7. Vannini, C. *et al.* Endosymbiosis in statu nascendi: close phylogenetic relationship between obligately endosymbiotic and obligately free-living *Polynucleobacter* strains (Betaproteobacteria). *Environ. Microbiol.* **9**, 347–359 (2007).
8. Hahn, M. W., Schmidt, J., Pitt, A., Taipale, S. J. & Lang, E. Reclassification of four *Polynucleobacter necessarius* strains as *Polynucleobacter asymbioticus* comb. nov., *Polynucleobacter duraquae* sp. nov., *Polynucleobacter yangtzensis* sp. nov., and *Polynucleobacter sinensis* sp. nov., and emended description of the species *Polynucleobacter necessarius*. *Int. J. Syst. Evol. Microbiol.* **66**, 2883–2892 (2016).
9. Vannini, C., Ferrantini, F., Ristori, A., Verni, F. & Petroni, G. Betaproteobacterial symbionts of the ciliate *Euplotes*: origin and tangled evolutionary path of an obligate microbial association. *Environ. Microbiol.* **14**, 2553–2563 (2012).
10. Gil, R. *et al.* The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. *Proc. Natl Acad. Sci. USA* **100**, 9388–9393 (2003).
11. Rio, R. V. M., Lefevre, C., Heddi, A. & Aksoy, S. Comparative genomics of insect-symbiotic bacteria: influence of host environment on microbial genome composition. *Appl. Environ. Microbiol.* **69**, 6825–6832 (2003).
12. Moran, N. A. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc. Natl Acad. Sci. USA* **93**, 2873–2878 (1996).
13. Lamelas, A. *et al. Serratia symbiotica* from the aphid *Cinara cedri*: a missing link from facultative to obligate insect endosymbiont. *PLoS Genet.* **7**, e1002357 (2011).
14. Oakeson, K. F. *et al.* Genome degeneration and adaptation in a nascent stage of symbiosis. *Genome Biol. Evol.* **6**, 76–93 (2014).
15. Gould, S. J. *Wonderful Life* (W. W. Norton & Co., New York, 1989).
16. Wernegreen, J. J., Richardson, A. O. & Moran, N. A. Parallel acceleration of evolutionary rates in symbiont genes underlying host nutrition. *Mol. Phylogenet. Evol.* **19**, 479–485 (2001).
17. O'Fallon, B. Population structure, levels of selection, and the evolution of intracellular symbionts. *Evolution* **62**, 361–373 (2008).
18. Pettersson, M. E. & Berg, O. G. Muller's ratchet in symbiont populations. *Genetica* **130**, 199–211 (2007).
19. Moran, N. A., McLaughlin, H. J. & Sorek, R. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science* **323**, 379–382 (2009).
20. McCutcheon, J. P. & Moran, N. A. Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* **10**, 13–26 (2012).
21. Burke, G. R. & Moran, N. A. Massive genomic decay in *Serratia symbiotica*, a recently evolved symbiont of aphids. *Genome Biol. Evol.* **3**, 195–208 (2011).
22. Clayton, A. L., Jackson, D. G., Weiss, R. B. & Dale, C. Adaptation by deletogenic replication slippage in a nascent symbiont. *Mol. Biol. Evol.* **33**, 1957–1966 (2016).
23. Husnik, F. & McCutcheon, J. P. Repeated replacement of an intrabacterial symbiont in the tripartite nested mealybug symbiosis. *Proc. Natl Acad. Sci. USA* **113**, E5416–E5424 (2016).
24. Wernegreen, J. J. & Moran, N. A. Evidence for genetic drift in endosymbionts (*Buchnera*): analyses of protein-coding genes. *Mol. Biol. Evol.* **16**, 83–87 (1999).
25. Itoh, T., Martin, W. & Nei, M. Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. *Proc. Natl Acad. Sci. USA* **99**, 12944–12948 (2002).
26. Nei, M. Selectionism and neutralism in molecular evolution. *Mol. Biol. Evol.* **22**, 2318–2342 (2005).
27. Boscaro, V. *et al. Polynucleobacter necessarius*, a model for genome reduction in both free-living and symbiotic bacteria. *Proc. Natl Acad. Sci. USA* **110**, 18590–18595 (2013).
28. Hao, Z. *et al.* Genome sequence of a freshwater low-nucleic-acid-content bacterium, betaproteobacterium strain CB. *Genome Announc.* **1**, e0013513 (2013).
29. Syberg-Olsen, M. J. *et al.* Biogeography and character evolution of the ciliate genus *Euplotes* (Spirotrichea, Euplotia), with description of *Euplotes curdsi* sp. nov. *PLoS ONE* **11**, e0165442 (2016).
30. Manzano-Marín, A. & Latorre, A. Settling down: the genome of *Serratia symbiotica* from the aphid *Cinara tujafina* zooms in on the process of accommodation to a cooperative intracellular life. *Genome Biol. Evol.* **6**, 1683–1698 (2014).
31. Hoetzinger, M., Schmidt, J., Jezberová, J., Koll, U. & Hahn, M. W. Microdiversification of a pelagic *Polynucleobacter* species is mainly driven by acquisition of genomic islands from a partially interspecific gene pool. *Appl. Environ. Microbiol.* **83**, e02266-16 (2017).
32. Moran, N. A., McCutcheon, J. P. & Nakabachi, A. Genomics and evolution of heritable bacterial symbionts. *Annu. Rev. Genet.* **42**, 165–190 (2008).
33. Clayton, A. L. *et al.* A novel human-infection-derived bacterium provides insights into the evolutionary origins of mutualistic insect-bacterial symbioses. *PLoS Genet.* **8**, e1002990 (2012).
34. Bennett, G. M., McCutcheon, J. P., McDonald, B. R. & Moran, N. A. Lineage-specific patterns of genome deterioration in obligate symbionts of sharpshooter leafhoppers. *Genome Biol. Evol.* **8**, 296–301 (2016).
35. Andersson, J. O. & Andersson, S. G. E. Insights into the evolutionary process of genome degradation. *Curr. Opin. Genet. Dev.* **9**, 664–671 (1999).
36. Nilsson, A. I. *et al.* Bacterial genome size reduction by experimental evolution. *Proc. Natl Acad. Sci. USA* **102**, 12112–12116 (2005).
37. Hershberg, R., Tang, H. & Petrov, D. A. Reduced selection leads to accelerated gene loss in *Shigella*. *Genome Biol.* **8**, R164 (2007).
38. McCutcheon, J. P. & von Dohlen, C. D. An interdependent metabolic patchwork in the nested symbiosis of mealybugs. *Curr. Biol.* **21**, 1366–1372 (2011).
39. Ghignone, S. *et al.* The genome of the obligate endobacterium of the AM fungus reveals an interphylum network of nutritional interactions. *ISME J.* **6**, 136–145 (2012).
40. Nakabachi, A. *et al.* The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science* **314**, 267 (2006).

41. Smith, W. A. *et al.* Phylogenetic analysis of symbionts in feather-feeding lice of the genus *Columbicola*: evidence for repeated symbiont replacements. *BMC Evol. Biol.* **13**, 109 (2013).
42. Rosati, G., Modeo, L., Melai, M., Petroni, G. & Verni, F. A multidisciplinary approach to describe protists: a morphological, ultrastructural, and molecular study on *Peritromus kahli* Villeneuve-Brachon, 1940 (Ciliophora, Heterotrichea). *J. Eukaryot. Microbiol.* **51**, 49–59 (2004).
43. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
44. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
45. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
46. Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542 (2012).
47. Prescott, D. M. Evolution of DNA organization in hypotrichous ciliates. *Ann. NY Acad. Sci.* **870**, 301–313 (1999).
48. Tritt, A., Eisen, J. A., Facciotti, M. T. & Darling, A. E. An integrated pipeline for *de novo* assembly of microbial genomes. *PLoS ONE* **7**, e42304 (2012).
49. Aziz, R. K. *et al.* The RAST server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).
50. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
51. Garcia, S. L. *et al.* Metabolic potential of a single cell belonging to one of the most abundant lineages in freshwater bacterioplankton. *ISME J.* **7**, 137–147 (2013).
52. Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K. & Tolstoy, I. RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res.* **42**, D553–D559 (2014).
53. Criscuolo, A. & Gribaldo, S. BMGE (block mapping and gathering with entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).
54. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
55. Le, S. Q., Dang, C. C. & Gascuel, O. Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Mol. Biol. Evol.* **29**, 2921–2936 (2012).
56. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).
57. Le, S. Q., Gascuel, O. & Lartillot, N. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* **24**, 2317–2323 (2008).
58. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109 (2004).
59. Yang, Z. PAML4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
60. R Core Team *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna, 2013).
61. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).
62. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome databases. *Nucleic Acids Res.* **42**, D459–D471 (2014).
63. Finn, R. D. *et al.* InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Res.* **45**, D190–D199 (2017).
64. Placzek, S. *et al.* BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Res.* **45**, D380–D388 (2017).
65. Hahn, M. W., Stadler, P., Wu, Q. L. & Pöckl, M. The filtration-acclimatization method for isolation of an important fraction of the not readily cultivable bacteria. *J. Microbiol. Methods* **57**, 379–390 (2004).

Acknowledgements

We thank S. Gabrielli for helping with the artwork. This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada (227301 and 6544-2013 awarded to P.J.K. and D.H.L, respectively). V.B. and M.K. were supported by fellowships from the Tula Foundation to the Centre for Microbial Diversity and Evolution.

Author contributions

V.B., D.H.L. and P.J.K. designed the study. V.B. sampled and isolated the ciliates. V.B. and C.V. cultured, screened and identified the *Euplotes* strains and *Polynucleobacter* symbionts. C.V. performed the isolation experiments on the symbionts. V.B. and C.V. optimized and performed the genomic DNA extractions. D.H.L. prepared the libraries. V.B. assembled and annotated the genomes. V.B. and M.F. conducted the functional analysis. M.K. performed the phylogenomic inference, clustering analysis and dN/dS calculations. V.B., M.K. and P.J.K. wrote the paper. All authors participated in the drafting process.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at doi:10.1038/s41559-017-0237-0.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to P.J.K.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.