Extreme genome reduction in symbiotic bacteria

John P. McCutcheon[1] and Nancy A. Moran[2]

Abstract | Since 2006, numerous cases of bacterial symbionts with extraordinarily small genomes have been reported. These organisms represent independent lineages from diverse bacterial groups. They have diminutive gene sets that rival some mitochondria and chloroplasts in terms of gene numbers and lack genes that are considered to be essential in other bacteria. These symbionts have numerous features in common, such as extraordinarily fast protein evolution and a high abundance of chaperones. Together, these features point to highly degenerate genomes that retain only the most essential functions, often including a considerable fraction of genes that serve the hosts. These discoveries have implications for the concept of minimal genomes, the origins of cellular organelles, and studies of symbiosis and host-associated microbiota.

[1]*University of Montana, Division of Biological Sciences, 32 Campus Drive, HS104, Missoula, Montana 59812, USA.*
[2]*Yale University, Department of Ecology and Evolutionary Biology, 300 Heffernan Drive, West Haven, Connecticut 06516, USA.*
e-mails: john.mccutcheon@umontana.edu; nancy.moran@yale.edu
doi:10.1038/nrmicro2670
Published online
8 November 2011

During the first decade of whole-genome sequencing, the size of the bacterial genomes that were sequenced, including those from several obligate intracellular bacterial symbionts, appeared to reach a lower limit of about 500 kb, corresponding to about 500 genes. Unexpectedly, the recently sequenced genomes of symbiotic bacteria have revealed the existence of far smaller genomes, some of which include other extreme features, such as extremely rapid sequence evolution, codon reassignments and extreme biases in nucleotide composition. The main drivers of these changes are mutation and genetic drift in the context of both small genetic population sizes and asexuality. Not only do these organisms encode few proteins, but also their highly derived proteins appear to be susceptible to misfolding and require a large investment in chaperones to preserve protein functionality. Small genomes experience ongoing erosion through loss of particular genes, although certain genes, including some that underlie mutualistic contributions to the host, remain essential and are retained. Co-adaptation involving the genes of both symbionts and hosts probably facilitates some gene losses but, unlike the transfer that has occurred from mitochondria and plastids to the nucleus, transfer of bacterial genes to host genomes does not seem to play a part. Although these genomes are extremely tiny, they are not models for replication efficiency and are in fact probably among the least efficient or robust bacterial genomes.

In this Review, we describe the history of the minimal-genome concept and summarize the processes and outcomes of genome reduction in bacteria that are obligately host associated. We next turn to our central focus, describing recent discoveries of extremely reduced genomes in nutrient-providing bacterial symbionts of sap-feeding insects. These recent results reveal five independent examples of extreme genome reduction from four distantly related bacterial groups, each example with a genome less than half the size of that of *Mycoplasma genitalium* (and in some cases less than one-quarter the size) (FIG. 1; TABLE 1). Finally, we summarize the arguments for and against the classification of these organisms as extremely degenerate bacterial symbionts, organelles or something in between.

Small genomes and the minimal-genome concept

Independent evolutionary origins of tiny genomes. It has long been appreciated that some species of the genus *Mycoplasma* have considerably smaller genomes than most other bacteria[1–3]. Although some initially argued that these small genomes represent ancestrally primitive organisms[3], by the 1980s phylogenetic work based on 16S ribosomal RNA sequence fragments had firmly established that *Mycoplasma* spp.[4] — as well as other intracellular bacteria with reduced genomes, such as some members of the family Rickettsiaceae[5] — are derived from bacteria with larger genomes. By 2000, results from early genome-sequencing projects had confirmed these findings, as cases of independent reductive genome evolution had been identified in bacterial groups as diverse as the classes Mollicutes (*M. genitalium*[6]), Alphaproteobacteria (*Rickettsia prowazekii*[7]), Gammaproteobacteria (*Buchnera aphidicola*[8]) and

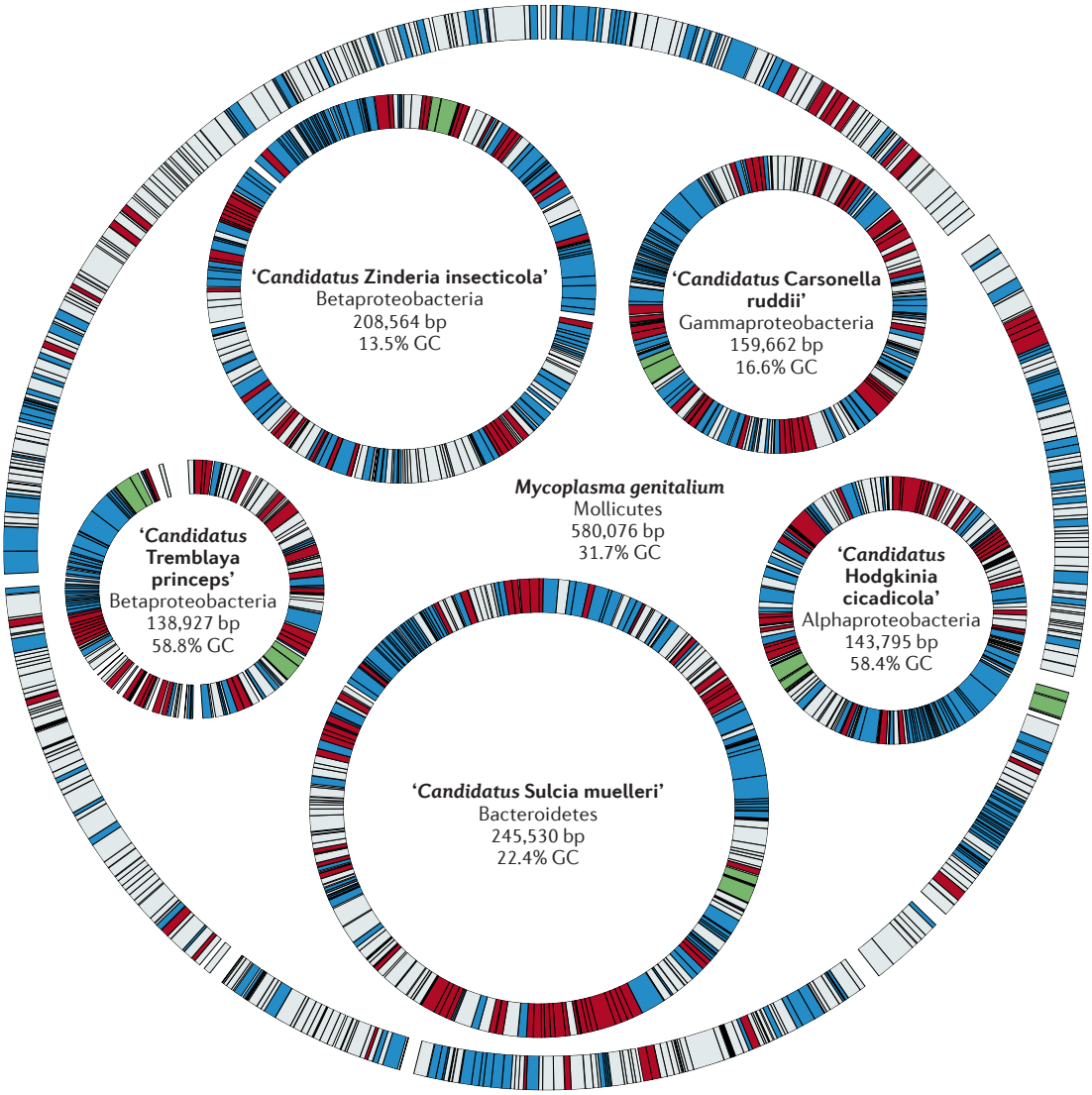


Figure 1 | **Comparison of the smallest genomes for free-living and symbiotic organisms.** The genome of *Mycoplasma genitalium*, the free-living organism with the smallest genome, is two to four times as large as the genomes of five symbionts recently shown to have tiny genomes (that is, smaller than 300 kb): '*Candidatus* Sulcia muelleri', '*Candidatus* Zinderia insecticola', '*Candidatus* Carsonella ruddii', '*Candidatus* Hodgkinia cicadicola' and '*Candidatus* Tremblaya princeps'. Genes involved in informational processesing are in blue, those involved in vitamin or amino acid biosynthesis are in maroon, ribosomal RNA genes are in green, other genes are light grey and breaks are non-coding regions.

Spirochaetes (*Borrelia burgdorferi*[9]). Although each of these genomes retains genes that strongly reflect the beneficial or pathogenic role of the microorganism in the host, collectively they show that diverse bacterial groups have the 'ability' to lose many genes given the right conditions, indicating that the mechanisms and forces governing genome reduction might be generalized to most bacteria.

Concept of the minimal genome. The minimal genome is defined as the gene set that is sufficient for life under lenient (nutrient-rich and stress-free) conditions. Many experimental studies further define it as the gene set that supports axenic culture in rich media, recognizing that most organisms require additional genes under realistic ecological conditions and that the set of required genes will vary with environmental conditions. The minimal-genome concept originated and developed in parallel with the era of complete-genome sequencing for bacteria, starting in the mid 1990s[10]. The predicted gene complements of minimal genomes have been developed mainly through computational analyses of the gene repertoires of different organisms[10–16] and through mutagenesis experiments based on the disruption of genes in organisms grown in axenic culture[17–20]. Such mutagenesis experiments are restricted to those organisms that can be grown in axenic laboratory culture, such as *M. genitalium*, which has the smallest genome of any organism yet grown axenically.

These studies have revealed that the set of universal genes is small and excludes many genes that had been experimentally identified as essential in

Axenic culture
A culture of a bacterium or other organism that is independent of any other living organism.

Table 1 | **Comparisons of representative genomes of bacteria, symbionts, viruse and organelles**

Organism	Taxonomy*	Genome size (bp)	GC content (%)	Number of coding sequences	Number of M and I COGs‡	Cell shape	Presence of rod shape-determining genes			
							ftsZ	*ispA*	*mreB*	*rodA*
Free-living bacteria										
Escherichia coli	Gammaproteobacteria	4,639,675	50.8	4145	323	Rod[112]	+	+	+	+
Bacillus subtilis	Mollicutes	4,215,606	43.5	4176	307	Rod[112]	+	+	+	+
Rickettsia prowazekii	Alphaproteobacteria	1,111,523	29.0	835	100	Rod[113]	+	–	+	+
Mycobacterium genitalium	Mollicutes	580,076	31.7	475	17	'Flask' (REF. 114)	+	–	–	–
Symbionts with reduced genomes										
'*Candidatus* Blochmannia floridanus'	Gammaproteobacteria	705,557	27.4	583	77	Rod[115]	+	–	+	+
Wigglesworthia glossinidia	Gammaproteobacteria	697,724	22.5	611	85	Rod[116]	+	+	+	+
'*Candidatus* Baumannia cicadellinicola'	Gammaproteobacteria	686,194	33.2	595	62	Sphere[117]	+	+	+	+
Buchnera aphidicola str. APS	Gammaproteobacteria	640,681	26.3	564	41	Sphere[118]	+	+	–	–
'*Candidatus* Moranella endobia'	Gammaproteobacteria	538,294	43.5	406	36	Sphere[119]	+	–	+	+
Buchnera aphidicola str. Cc	Gammaproteobacteria	416,380	20.2	357	10	Sphere[120]	+	–	–	–
Symbionts with extremely reduced genomes										
'*Candidatus* Sulcia muelleri str. GWSS'	Flavobacteria	245,530	22.4	227	5	Pleomorphic tube[71]	–	–	–	–
'*Candidatus* Zinderia insecticola'	Betaproteobacteria	208,564	13.5	202	1	Pleomorphic blob[52]	–	–	–	–
'*Candidatus* Carsonella ruddii'	Gammaproteobacteria	159,662	16.6	182	1	Pleomorphic tube[53]	–	–	–	–
'*Candidatus* Hodgkinia cicadicola'	Alphaproteobacteria	143,795	58.4	169	0	Pleomorphic tube[57]	–	–	–	–
'*Candidatus* Tremblaya princeps'	Betaproteobacteria	138,927	58.8	121	0	Pleomorphic blob[119]	–	–	–	–
Organelles with large genomes										
Cucurbita pepo	Mitochondrion	982,833	42.8	38	NA	NA	–	–	–	–
Floydiella terrestris	Chloroplast	521,168	34.5	74	NA	NA	–	–	–	–
Porphyra purpurea	Chloroplast	191,028	33.0	209	NA	NA	–	–	–	–
Reclinomonas americana	Mitochondrion	69,034	26.1	67	NA	NA	–	–	–	–
Viruses with large genomes										
Acanthamoeba polyphaga mimivirus	Mimiviridae	1,181,404	28.0	1,262	NA	NA	NA	NA	NA	NA
Cafeteria roenbergensis virus (CroV)	Mimiviridae	617,453	23.4	544	NA	NA	NA	NA	NA	NA
Coccolithovirus	Phycodnaviridae	407,339	40.2	472	NA	NA	NA	NA	NA	NA

NA, not applicable. *Taxonomy refers to bacterial class, type of organelle or viral family. ‡Total combined number of genes in the M (cell envelope biogenesis, outer membrane) and I (lipid metabolism) clusters of orthologous groups (COG) categories.

particular organisms. This reflects the fact that even fundamental functions are sometimes accomplished in entirely different ways by different organisms — for example, acquiring needed compounds via transport versus *de novo* biosynthesis, or using unrelated genes to charge certain tRNAs with the correct amino acid by different pathways[21]. Conversely, some universal or near-universal genes can be disrupted without lethality[16]. Although different studies predict different minimal gene sets, they all include genes involved in essential cellular functions that are widely distributed, if not universal. Furthermore, functions have been identified for the majority of genes that are either near-universal in their distribution or found to be essential in

Box 1 | **Muller's ratchet in host-restricted lineages**

In all organisms, genomes constantly acquire mutations in the form of deletions, base substitutions and other chromosomal changes. The majority of these changes are deleterious or have a neutral effect on fitness, and are commonly selected against and thus removed from the population. However, as intracellular bacteria reproduce asexually, have small effective population sizes, are limited by existing only in a host cytoplasm and experience frequent bottlenecks (at transmission), such purifying selection is less effective. Instead, host-restricted bacteria are subject to Muller's ratchet, a process that results in the accumulation of slightly deleterious mutations[28]. These mutations can be base substitutions that result in suboptimal amino acids within proteins (thus lowering the stability of the folded protein), insertions or deletions that disrupt an open reading frame (thus inactivating a gene that is beneficial but not essential) or even larger deletions that remove non-essential chromosomal regions. The fitness effect of each mutation is too small to prevent its fixation, but cumulatively these mutations result in a lowered overall efficiency of the cell. Consequences are rapid evolution of protein sequences, gene loss and lower predicted thermal stability of proteins.

mutagenesis experiments. These functions involve central cellular processes, primarily translation (including tRNA charging and biogenesis of the ribosome), replication and transcription. Genes underlying metabolism are not universal, although some are widespread and many are crucial for particular organisms in natural circumstances.

Small genomes, 1995–2006. Highly reduced genomes of less than 1 Mb have consistently been recovered from host-dependent microorganisms, primarily bacteria. In part, this reflects the rich source of metabolites that host tissues provide, thereby permitting the elimination of many genes that underlie the biosynthesis of small molecules and the degradation of complex substrates. However, eliminated genes fall into every functional category; for example, DNA repair and recombination genes are typically depleted in genomes of host-restricted symbionts and pathogens, but each case of genome reduction retains a distinct set of repair genes[22]. In addition to genes underlying central cellular processes, organisms with small genomes typically possess a subset of genes devoted to the interaction with the host; these can be either genes underlying mechanisms for invading and persisting in host tissues (in the case of pathogens) or genes underlying contributions to host fitness (in the case of mutualists). This set of genes depends on the biology of the particular association. For example, *B. aphidicola* and other symbionts of insects that feed on plant sap provide essential amino acids that are lacking in the host diet[8,23,24], and *Wigglesworthia glossinidia*, the intracellular symbiont of blood-feeding tsetse flies, provides B vitamins that are lacking in vertebrate blood[25]. But beyond these differences that are attributable to biologically obvious causes, gene repertoires vary dramatically for different instances of genome reduction. For instance, although *W. glossinidia*, '*Candidatus* Baumannia cicadellinicola', '*Candidatus* Blochmannia floridanus' and *B. aphidicola* (which represent distinct lineages of gammaproteobacterial symbionts of insects) have genomes of about 500–700 kb that encode 500–700 genes, their gene repertoires differ dramatically not only in genes related to host nutrition, but also in genes underlying functions such as cell envelope biogenesis, replication initiation, and DNA repair and recombination pathways[8,25–27].

Genetic drift and genome reduction. The link between host restriction and genome reduction presents one of the clearest cases of evolutionary convergence of genome features and raises the question of causation. Numerous studies implicate the small population size and asexuality of the relevant species[28–30], combined with an inherent deletional bias in bacteria[29,31], as the major causes. Such a population structure, imposed by the restriction to specific hosts and by the lack of recombination between strains in different hosts, leads to high levels of genetic drift and to inactivation and deletion of genes that are mildly beneficial but not essential (BOX 1). Another consequence of this same population structure is rapid sequence evolution affecting all genes in the genome, which was originally noticed in the small genomes of *B. aphidicola* and *M. genitalium*. This rapid evolution results in greatly altered proteins that are predicted to have destabilized secondary structures[24,32,33]. The early stages of genome reduction are represented in some more recently evolved symbionts and are characterized by proliferation of mobile elements, formation of pseudogenes, multiple genomic rearrangements and deletion of chromosome fragments[34–36]. However, in more anciently evolved symbionts such as *B. aphidicola*, mobile elements and most pseudogenes have been eliminated.

Features of tiny genomes in symbiotic bacteria

Since 2006, drastically small genomes have been recovered from several independent lineages of obligate symbionts of insects (FIG. 1; TABLE 1). These findings challenge previous views of minimal genomes and potentially blur the boundaries between cellular organisms, organelles and viruses. Here, we consider the genomic features of '*Candidatus* Sulcia muelleri', '*Candidatus* Zinderia insecticola', '*Candidatus* Carsonella ruddii', '*Candidatus* Hodgkinia cicadicola' and '*Candidatus* Tremblaya princeps', which each contain a tiny genome smaller than 300 kb.

Comparison to organellar and viral genome sizes. With some exceptions[34,37], the relationship between the size of a bacterial genome and the number of protein-coding genes it contains is linear, with an average of approximately one protein-coding gene per kilobase of sequence[35,38]. This pattern is largely conserved in the smallest symbiont genomes, which encode far fewer genes than even *M. genitalium* or *B. aphidicola* and which overlap the largest organellar and viral genomes in size and number of genes (although the types of retained genes differ[39]) (TABLE 1). The same linear relationship is generally not upheld in organellar genomes, in which huge genome expansions are not correlated with increased coding capacity[40–42]. Therefore, although organellar genomes overlap reduced bacterial genomes in terms of size, they are not comparable in terms of coding capacity. An exception to these generalizations

is '*Ca.* Tremblaya princeps', a bacterial symbiont with a tiny genome that is unusually gene sparse, with a coding density of only 73%[43]. The atypical properties of '*Ca.* Tremblaya princeps' make it an outlier in many respects, even when compared with other similarly tiny genomes (BOX 2).

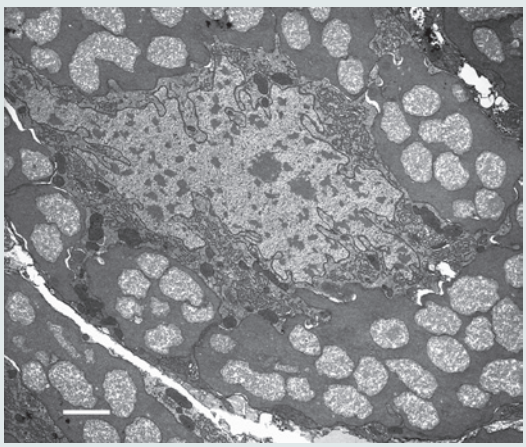
Like bacterial genomes, most viral genomes remain gene dense over their entire size range; in particular, the largest known viral genomes — of mimiviral strains — have a considerable overlap with the reduced genomes of several intracellular bacteria in terms of coding density and gene count[44,45]. Although these mimiviral genomes have large numbers of genes, including several involved in DNA repair and transcription, they lack many genes involved in translation, encode homologues primarily of eukaryotic origin and in general are not thought to be autonomous entities.

Base-compositional biases in reduced bacterial genomes. Teasing apart the different effects of mutation and selection on bacterial GC content has been difficult. In 1962, a model was proposed to describe the average GC content in a genome as a function of strictly neutral mutational processes driven by differences in (G or C)→(A or T) and (A or T)→(G or C) base substitution rates[46]. Later, the variation in genomic GC content among groups of bacteria was ascribed to lineage-specific mutational patterns[46–48] and to selection on various genome-wide properties[49]. Two recent papers support the existence of an inherent, universal mutational bias consistently from (G or C)→(A or T) and indicate that selection favouring a higher GC content is a major force determining the base composition of a bacterial genome[50,51].

Reduced bacterial genomes tend to have an increased AT content, sometimes drastically so (FIG. 2a). Indeed, the most highly AT-biased cellular genomes are from the symbionts '*Ca.* Zinderia insecticola', a betaproteobacterium with a 209 kb genome and a GC content of 13.5%[52], and '*Ca.* Carsonella ruddii', a gammaproteobacterium with a GC content of 16.6%[53] (FIG. 2a). Several hypotheses have been put forward to explain the AT bias in strict endosymbionts, invoking either selection[54] or a combination of population-genetic forces and changes in mutational patterns[28,54,55] as the major force. Small genomes tend to lose many of the genes involved in DNA repair[22], which might result in more A or T mutations, as DNA damage such as cytosine deamination and guanine oxidation often leads to (G or C)→(A or T) changes[56]. The effects of Muller's ratchet and the associated relaxation in purifying selection allow more of the slightly deleterious mutations, which tend to be biased towards A or T mutations, to be fixed in the population. This combination of forces shifts the average genomic GC content towards a new equilibrium with higher AT content.

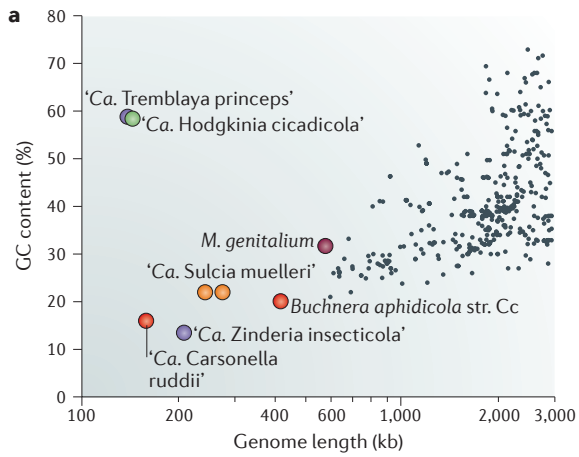
Remarkably, the two smallest bacterial genomes — of '*Ca.* Tremblaya princeps' (REF. 43), with a genomic GC content of 58.8% and a GC content of 66.6% for the third position of four-fold-degenerate codons, and '*Ca.* Hodgkinia cicadicola' (REF. 57), an alphaproteobacterium with a 144 kb genome, a genomic GC content of 58.4% and a GC content of 62.5% for the third position of four-fold-degenerate codons — break the otherwise universal relationship between extreme genome reduction and low genomic GC content (FIG. 2a). In '*Ca.* Hodgkinia cicadicola', this was assumed to be indicative of a GC-biased mutational pressure, as the third positions of

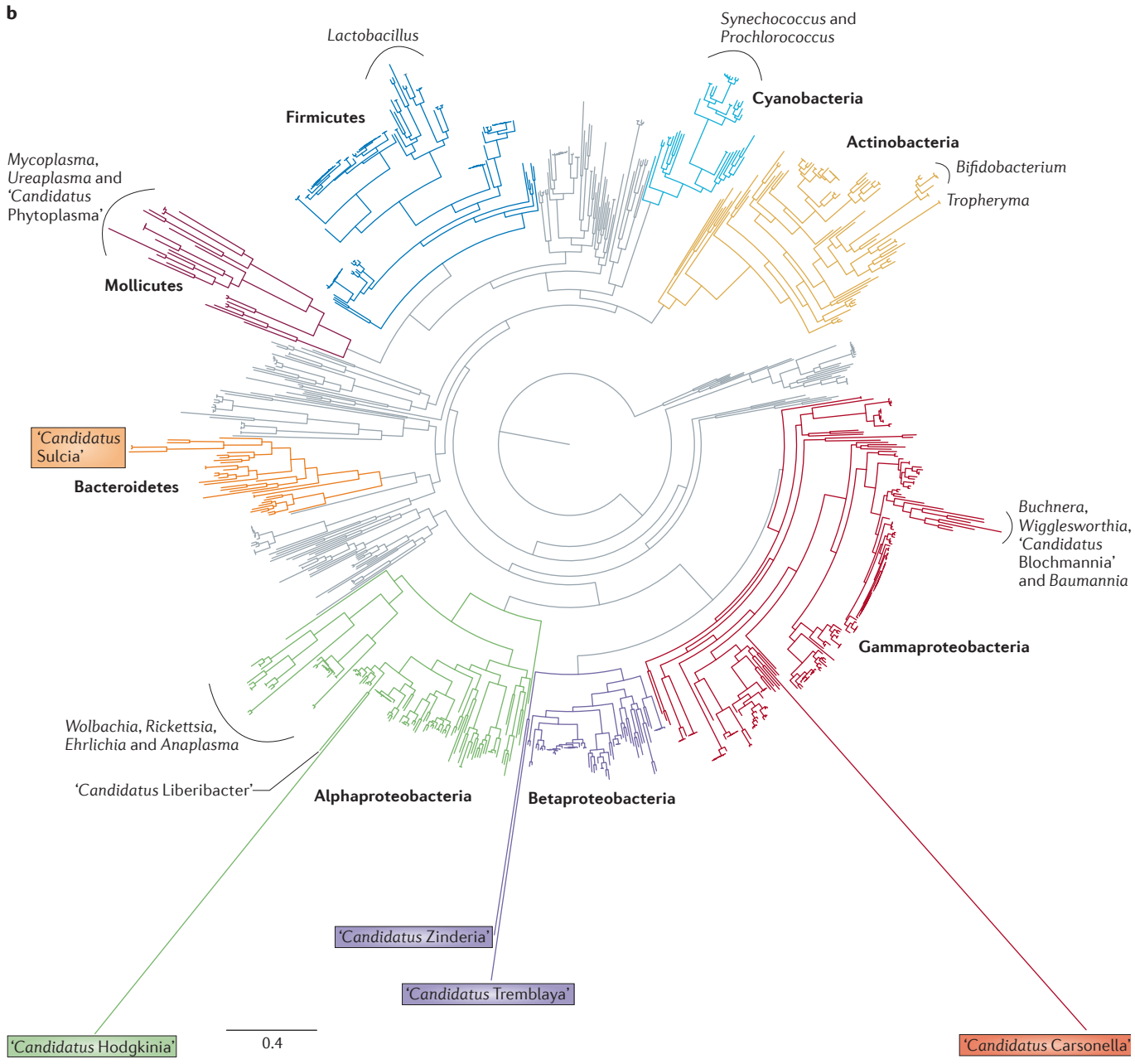
Box 2 | **The unique cell biology and genome of '*Candidatus* Tremblaya princeps'**



As is the case for many other sap-feeding insects, the mealybug *Planococcus citri* has developed a stable relationship with two bacterial symbionts: '*Candidatus* Tremblaya princeps', a betaproteobacterium, and '*Candidatus* Moranella endobia', a gammaproteobacterium. These symbionts are thought to provide the host with essential amino acids that are lacking in its diet[104–108]. Remarkably, however, '*Ca.* Moranella endobia' lives inside '*Ca.* Tremblaya princeps' (see the figure; '*Ca.* Moranella endobia' cells are the small, lightly coloured, punctate cells approximately the size of the 2.3 µm scale bar, '*Ca.* Tremblaya princeps' cells are the larger, darker, irregularly shaped cells enveloping '*Ca.* Moranella endobia', and the large, irregularly shaped structure in the centre of the image is the insect cell nucleus), forming the only known bacteria-within-a-bacterium symbiosis[43]. Recent genomic work confirmed the nutritional role of these symbionts and also revealed that the '*Ca.* Tremblaya princeps' genome is surprisingly degenerate and gene sparse[43]. '*Ca.* Tremblaya princeps' has no functional tRNA synthetase genes, lacks several other genes found even in the smallest symbiont genomes and has a low coding density (73%), in part owing to the presence of several pseudogenes[43]. These features are probably attributable to the presence of a '*Ca.* Moranella endobia', which has a relatively rich gene set. It is hypothesized that '*Ca.* Tremblaya princeps' had a compact, reduced genome before the acquisition of '*Ca.* Moranella endobia', and that this event has further relaxed the selective pressure on '*Ca.* Tremblaya princeps' genes that are now redundant in the presence of '*Ca.* Moranella endobia', resulting in the formation of pseudogenes and gene loss[43]. The pattern of pseudogene formation in '*Ca.* Tremblaya princeps' supports this idea, as almost all these genes are present and intact in '*Ca.* Moranella endobia' (REF. 43). Image courtesy of C. von Dohlen, Utah State University, USA.

Endosymbionts
Symbionts that reside inside the cells of the host.





◀ Figure 2 | **Extreme genomic features in symbionts with tiny genomes. a** | The relationship between genome size and GC content. The data for symbionts with tiny genomes and for *Mycoplasma genitalium*, the free-living organism with the smallest genome, are indicated. Other bacteria with genomes smaller than 3,000 kb are shown as black dots. **b** | Protein-based phylogeny of bacteria with highly reduced genomes in the context of all bacterial diversity. The tree was generated with RAxML version 7.2.8 (REF. 109) (relevant parameters: -f a -x 12345 -p 12345 -# 100 -m PROTGAMMAJTT). Alignments were individually made from 11 protein-coding sequences (genes for elongation factor EF-G (*fusA*), translation initiation factor 3 (*infC*), EF-TuA (*tufA*), RNA polymerase β-subunit (*rpoB*), 30S ribosomal subunits S2 (*rpsB*), S4 (*rpsD*) and S5 (*rpsE*), and 50S ribosomal subunits L2 (*rplB*), L3 (*rplC*), L4 (*rplD*) and L6 (*rplF*)) from 883 bacterial genomes using the linsi module of MAFFT[110], and then concatenated. All columns with >30% gaps were removed, leaving 4,207 positions in the alignment. The order Aquificales was constrained as the outgroup. Because of the extremely long branch lengths and highly biased amino acid composition of '*Candidatus* Carsonella ruddii', '*Candidatus* Zinderia insecticola' and '*Candidatus* Hodgkinia cicadicola', the classes Alphaproteobacteria, Betaproteobacteria, and Gammaproteobacteria were constrained to be monophyletic (in order to prevent these sequences from grouping together owing to long-branch attraction). '*Ca.* Carsonella ruddii' was further constrained to the gammaproteobacterial order Oceanospirillales, on the basis of previous phylogenetic work using small- and large-subunit ribosomal RNA sequences[111]. This tree should therefore be regarded as a generally correct grouping of these bacterial groups, but not as a definitive guide to their exact phylogenetic placement. Species are represented by genus names, for reasons of space constraints.

four-fold-degenerate codons were expected to be under little or no selection at the level of protein sequence[57]. These facts led to the hypothesis that the mutational bias present in free-living alphaproteobacteria, which tend to have a high GC content, has somehow been maintained by '*Ca.* Hodgkinia cicadicola' during the process of genome reduction[57]. It will be of interest to examine the direction of mutation in both '*Ca.* Hodgkinia cicadicola' and '*Ca.* Tremblaya princeps' to determine whether their GC mutational bias is unique or whether they, like most (or all) other bacteria, have a mutational bias in the AT direction[50,51], thereby leaving selection on the genome-wide GC content as the most obvious explanation for their unique genomic composition.

Codon reassignment in small genomes. The genetic code has independently diverged from the 'universal' code several times in both cellular and organellar genomes[58], and the reassignment of UGA from a stop codon to a tryptophan codon is one of the most common code changes. In bacteria, it is found in one lineage from the Mollicutes[59], and until recently this was the only coding reassignment event reported in bacteria[58]. Sequencing of the tiny genomes from bacterial insect symbionts has uncovered two new cases of this stop-to-tryptophan reassignment: in '*Ca.* Hodgkinia cicadicola' (REF. 57) and '*Ca.* Zinderia insecticola' (REF. 52). These cases are the only examples of codon reassignment in the Bacteria outside of the Mollicutes lineage[52,57], and '*Ca.* Hodgkinia cicadicola' is the only known case of a UGA codon reassignment in a genome with a high GC content[57]. This finding, along with a critical analysis of possible codon reassignment mechanisms based on several mitochondrial genomes[60], seems to refute the widely cited idea that low GC content is a prerequisite for some codon reassignments (the 'codon capture' hypothesis)[61–63]. We imagine a mechanism based on the 'ambiguous translation' hypothesis[60], in which mutations in $tRNA^{Trp}$ allow promiscuous decoding of both tryptophan (UGG) and stop (UGA) codons; this situation permits the loss of peptide chain release factor 2 (RF-2; encoded by *prfB*) (which recognizes UGA codons) through ongoing genome erosion[57]. According to this scenario, the UGA codon reassignment is a genetic co-adaptation to a deleterious loss of an important gene, *prfB,* not an adaptive event leading to genome (or translational) streamlining[64] or a completely neutral event involving the loss and reassignment of a codon though changes in GC content bias[61,62].

Cell envelope biosynthesis and morphology in symbionts with reduced genomes. As the genomes of obligately intracellular bacteria shrink, more and more genes involved in the production of fatty acids, phospholipids and peptidoglycan are lost (TABLE 1). '*Ca.* Tremblaya princeps' and '*Ca.* Hodgkinia cicadicola', with the two smallest known bacterial genomes, retain no genes involved in cell envelope biosynthesis; '*Ca.* Carsonella ruddii' and '*Ca.* Zinderia insecticola' each retain at most one gene involved in these biosynthetic processes (TABLE 1). Therefore, all four organisms appear to lack a cell wall and depend completely on host-derived membranes or, in some cases, possibly on membranes derived from a symbiotic partner within the same host cell. For intracellular symbionts that have not undergone such extreme genome reduction, the contribution of the host to the bacterial cell envelope is less clear. For example, in *B. aphidicola* str. APS, some genes for the production of fatty acids and peptidoglycan are present, but nearly all genes for the production of phospholipids and lipopolysaccharides are missing[8]. Likewise, most ATP-dependent transporters are missing from *B. aphidicola* and other small-genome symbionts, even for molecules such as amino acids that are known to move between host and symbiont compartments. Host-derived products may take over these functions, as suggested by the finding that genes encoding amino acid transporters are diversified in the aphid genome and specifically upregulated within bacteriocytes[65].

The rod shape that is common to many bacterial groups is usually, but not always[66], controlled by certain proteins, including the MreBCD complex[67,68], the membrane protein RodA[69] and the polyisoprenoid biosynthesis pathway protein IspA[70]. Mutations of any one of the genes that encode these proteins in rod-shaped cells can result in the formation of spherical cells, which are superficially similar in shape to intracellular symbionts with intermediate genome sizes, such as '*Ca.* Baumannia cicadellinicola' and *B. aphidicola*. Bacteria with more severely reduced genomes tend to lose this spherical shape and become highly irregular 'blobs', as is the case for '*Ca.* Zinderia insecticola' (REF. 52) and '*Ca.* Tremblaya princeps' (REF. 43) (BOX 2), or irregular elongated tubes, as is the case for '*Ca.* Sulcia muelleri' (REF. 71), '*Ca.* Carsonella ruddii' (REF. 53) and '*Ca.* Hodgkinia cicadicola' (REF. 57) (TABLE 1).

Rapid sequence evolution in tiny genomes. Bacteria that have adopted strict intracellular lifestyles have small populations that undergo frequent bottlenecks

Bacteriocytes
Specialized eukaryotic cells that contain symbionts within the cytosol.

and strictly asexual reproduction, resulting in a diminution in selection efficacy (BOX 1). As a result, they carry a substantial load of deleterious substitutions that noticeably increase the rate of sequence evolution at both the nucleotide and amino acid levels[28]. This increased rate of sequence evolution has been extensively documented in bacteria with small genome, including *B. aphidicola* and *M. genitalium*, but it is most extreme in the tiniest genomes. These rapid evolutionary rates are evident in phylogenetic trees based on protein sequences (FIG. 2b); the longest branches, reflecting the fastest sequence evolution, are associated with the four smallest genomes, those of '*Ca.* Hodgkinia cicadicola', '*Ca.* Carsonella ruddii', '*Ca.* Tremblaya princeps' and '*Ca.* Zinderia insecticola'. Other lineages with long branches mostly correspond to intracellular bacteria with reduced genomes, including *M. genitalium*, *R. prowazekii*, '*Candidatus* Liberibacter asiaticus', *B. aphidicola*, '*Ca.* Sulcia muelleri' and *Tropheryma whipplei*, or free-living bacteria with reduced genomes and elevated rates of sequence evolution, such as *Prochlorococcus marinus*[72].

Another consequence of inefficient selection caused by small effective population sizes in strictly intracellular bacteria is the loss of genes that are beneficial but not required, including genes that are involved in DNA recombination and repair[7,8,22,73]. The most extreme examples of this phenomenon are found in the smallest bacterial genomes; '*Ca.* Tremblaya princeps', '*Ca.* Hodgkinia cicadicola', '*Ca.* Carsonella ruddii' and '*Ca.* Zinderia insecticola' have lost all DNA repair pathways and retain only a handful of genes associated with the DNA polymerase III holoenzyme. In an extreme case, '*Ca.* Hodgkinia cicadicola' has retained only the DNA polymerase III α-subunit (encoded by *dnaE*) and ε-subunit (encoded by *dnaQ*)[39] (TABLE 2). Loss of genes that encode repair proteins is expected to result in a higher mutation rate, which further accelerates the rate of sequence evolution throughout the genome.

Gene packing and genome compactness in tiny genomes. The average coding density for all sequenced bacterial genomes is about 87%, with most falling within a narrow range of 85–90%[38]. Although the coding density for the genomes of known nutritional symbionts of insects varies widely, the smallest genomes are relatively gene dense, with an average coding density of 88%. '*Ca.* Hodgkinia cicadicola', '*Ca.* Carsonella ruddii', '*Ca.* Zinderia insecticola' and all four strains of '*Ca.* Sulcia muelleri' have gene densities of greater than 92%. '*Ca.* Tremblaya princeps' is a surprising exception: it has the smallest cellular genome described, but has a coding density of only 73% (BOX 2).

High levels of heat shock protein expression in small genomes. Most of the genes that are maintained in all of the smallest bacterial genomes are involved in the core processes of replication, transcription and translation[39] (FIG. 1; TABLE 2). It is therefore striking that genes related to protein folding and stability are the only other class of genes that is retained in all of the most reduced genomes. Specifically, the GroES–GroEL chaperonin complex and DnaK, the main component of the bacterial heat shock protein 70 (Hsp70) chaperone complex, are all retained in the five smallest genomes, suggesting that these enzymes have critical roles in the biology of these organisms. Studies measuring gene expression and protein abundance in symbionts with reduced genomes corroborate this view. GroEL accounts for ~10% of the total protein in the cell in *B. aphidicola*[74–76] and is the most abundant protein in *W. glossinidia*[77]. In addition, shotgun proteomics has shown that GroEL and DnaK are the most abundant proteins in '*Ca.* Hodgkinia cicadicola' (REF. 78) and are among the most abundant in '*Ca.* Sulcia muelleri' (REF. 78).

The abundance of chaperones may buffer the harmful effects of deleterious substitutions that accumulate in symbionts[28,32,79] and that can affect protein folding and stability[24,32,33]. For example, GroEL can rescue the function of the degenerate anthranilate synthase of *B. aphidicola*[80].

Gene repertoires in tiny genomes. Intracellular symbionts have the smallest known bacterial genomes, but genes are not lost randomly. In total, there are only 66 protein-coding genes conserved among '*Ca.* Sulcia muelleri', '*Ca.* Zinderia insecticola', '*Ca.* Carsonella ruddii' and '*Ca.* Hodgkinia cicadicola'; '*Ca.* Tremblaya princeps' contains only 46 of these (TABLE 2). The retained genes follow a strong pattern: in addition to genes involved in informational processing (translation, transcription and replication; see TABLE 2), these bacteria primarily contain genes related to the provision of nutrients to the host. For example, in '*Ca.* Tremblaya princeps', at least 82% of the genes that are classifiable into clusters of orthologous groups (COG) categories[81] are devoted to amino acid biosynthesis and informational processing, compared with only 32% in *Escherichia coli*.

Of the replication functions, only the replicative 5′-to-3′ polymerase subunit, DNA polymerase III α-subunit, and its tightly associated 3′-to-5′ proofreading exonuclease subunit, DNA polymerase III ε-subunit, are retained in these bacteria (because of its highly unusual nature, '*Ca.* Tremblaya princeps' is not included in these analyses). Of transcription functions, only *rpoA*, *rpoB*, *rpoC* and *rpoD*, which encode the three subunits of the RNA polymerase core enzyme along with its most tightly associated σ-factor, are conserved in symbionts with highly reduced genomes. Similarly, of the translation functions, all three ribosomal RNAs are universally retained, along with 17 of the 21 small-subunit ribosomal proteins, 16 of the 32 large-subunit ribosomal proteins, translation initiation factors IF-1, IF-2 and IF-3 (encoded by *infA*, *infB* and *infC*, respectively), and elongation factors EF-G and EF-Ts (encoded by *fusA* and *tsf*, respectively) (TABLE 2). In addition, the tRNA modification enzymes MnmA, MnmE and MnmG (which catalyse modifications at the wobble position of various tRNAs and thus normally affect codon choice and translational fidelity[82]) are conserved, although their roles in bacteria with highly reduced genomes are unclear. The presence of MnmA, MnmE and MnmG may have led in turn to the

retention of the SufS and/or IscS pathways for cysteine desulpherase activities. The relationship between the SufS and IscS pathways is complex; their activities seem to be at least somewhat redundant, and some organisms have only one of these proteins, whereas others have both[83]. Bacterial pathways in which IscS and SufS act include thiamine and biotin production, molybdenum cofactor synthesis, Fe–S cluster assembly and thionucleoside production[83], but it seems that MnmA-mediated production of 2-thiouridine is likely to be the activity for which these pathways are retained in the smallest genomes, as no genes related to the other activities are present (although several Fe–S-containing proteins are encoded by the smallest genomes). '*Ca.* Sulcia muelleri', '*Ca.* Carsonella ruddii' and '*Ca.* Hodgkinia cicadicola' all retain *sufS*, *sufB* and *sufC*, whereas '*Ca.* Zinderia insecticola' retains the *sufS* homologue *iscS* along with the accessory genes *iscA* and *iscU*. Thus, although IscS is typically thought to be involved with 2-thiouridine formation[84], it may be that '*Ca.* Sulcia muelleri', '*Ca.* Carsonella ruddii' and '*Ca.* Hodgkinia cicadicola' use SufS for this function, as has been predicted in the protozoan *Theileria parva*[85].

Genes lost. Particularly striking is the almost complete loss of genes for cell envelope biogenesis, regulation of gene expression, and DNA repair and recombination in certain species (TABLE 1). In the case of the cell envelope, the most likely explanation for this loss is co-adaptation with the host, involving modification of the host-derived membrane that surrounds symbiont cells. As symbionts lose functionality, including efficient transporters and the ability to produce phospholidpids and a robust cell wall, the host may take over these roles. This co-adaptation potentially occurs gradually, such that host-derived molecules begin to stabilize and control exchange at the host–symbiont interface, allowing symbiont-derived components to further degenerate and culminating in a complete loss of symbiont-produced materials, as observed in '*Ca.* Tremblaya princeps', '*Ca.* Hodgkinia cicadicola', '*Ca.* Carsonella ruddii' and '*Ca.* Zinderia insecticola'. The loss of regulatory genes may reflect both the constant environment within the specialized host cells in which the symbionts reside and host regulation of the host–symbiont interaction.

What are the limits of genome erosion?

Ongoing erosion in already tiny genomes. Despite their highly reduced size, there is abundant evidence that genes continue to form pseudogenes and are lost in these highly reduced genomes. This raises the question of how small a genome can become. Using the 82 protein- and RNA-coding genes that are conserved in all the smallest insect-symbiont genomes, except '*Ca.* Tremblaya princeps' (TABLE 2), we calculated a theoretical minimal genome size for an intracellular symbiont using the gene lengths from *E. coli* for each gene. These genes cover approximately 73 kb. Furthermore, we can assume that several genes would be present to provide the host with nutrients, as a complete loss of these genes would probably result in loss of the symbiotic association. The lowest numbers of genes involved in essential amino acid biosynthesis are 23 in '*Ca.* Hodgkinia cicadicola' (REF. 57) (which encodes the cobalamin-dependent version of methionine synthase; if this organism encoded the cobalamin-independent version of this enzyme, the cobalamin synthesis genes could be lost, which would reduce the number of amino acid biosynthesis genes to 11 (REF. 78)), 25 in '*Ca.* Zinderia insecticola' (REF. 52) and 30 in '*Ca.* Carsonella ruddii' (REF. 53). Adding 11 amino acid biosynthesis genes to the minimal genome would add another 11 kb (assuming 1,000 bp per gene[35]). Assuming a coding density of 95%, which is the average for tiny symbiont genomes, another 4 kb of intergenic space would be added, bringing the total genome size to about 88 kb. This rough analysis suggests that genomes even smaller than those of '*Ca.* Tremblaya princeps' and '*Ca.* Hodgkinia cicadicola' exist, but that none of these is likely to be smaller than 70–80 kb (outside of unusual cases similar to '*Ca.* Tremblaya princeps', with associated intracellular co-symbionts (BOX 2)).

Assessing evidence for gene transfer to hosts. The genomes of plastids and mitochondria encode far fewer genes than those of their free-living bacterial relatives and likely ancestors[86,87], and in some cases have been eliminated entirely[88]. Many of the missing genes were lost during the process of genome reduction, but in many cases the genes have been transferred to the host genome and the gene products are now imported across the organellar envelope by host-encoded transporters[89,90]. Although the proteins in organelles are a mixture of proteins produced in the organelle and in the host cytoplasm, the majority of organellar RNAs and proteins are host encoded. As genomes from some symbionts of insects have similar numbers of protein-coding genes to some organelles, it seems possible that similar gene transfer events have taken place in the evolution of these tiny symbiont genomes. Indeed, this mechanism has been proposed as a possible solution to the levels of genome reduction observed in both '*Ca.* Carsonella ruddii' and *B. aphidicola*[53,91].

The recent sequencing of the genomes from two insect hosts of intracellular symbionts — the pea aphid[92], which has *B. aphidicola* as a symbiont, and the human body louse[93], which has '*Candidatus* Riesia pediculicola' as a symbiont — has enabled the question of symbiont–host gene transfer to be definitively addressed. In the pea aphid genome, 12 genes or gene fragments of bacterial origin were identified[94]. Most of the intact genes are of alphaproteobacterial origin, a result that excludes *B. aphidicola*, which belongs to the Gammaproteobacteria, and implicates an endosymbiont of the genus *Wolbachia*, which is known to transfer DNA to its hosts[95,96]. Two non-functional pseudogenes originating from *B. aphidicola* were identified in the pea aphid genome[94] based on their sequence similarity to functional homologues in the *B. aphidicola* genome[94]. Genes originating from the alphaproteobacteria are intact and highly expressed in the aphid bacteriocyte, suggesting a role in symbiosis[94]; however, none appears to complement a critical activity that is missing from

Table 2 | **Genes conserved in four of the tiny-genome symbionts***

Gene	Product	Present in '*Candidatus* Tremblaya princeps'?
Replication		
dnaE	DNAP III α-subunit	Yes
dnaQ	DNAP III ε-subunit	Yes
Transcription		
rpoA	RNAP α-subunit	Yes
rpoB	RNAP β-subunit	Yes
rpoC	RNAP β′-subunit	Yes
rpoD	RNAP factor σ^{70}	No (gene present is a pseudogene)
Protein folding or stability		
groL	GroEL (chaperone Hsp60 family member)	Yes
groS	GroES (chaperone Hsp60 regulator)	Yes
dnaK	The main component of the chaperone Hsp70	Yes
tRNA modification		
mnmA	tRNA-specific 2-thiouridylase	No (gene present is a pseudogene)
mnmE	A GTP-binding protein with a role in tRNA modification	No
mnmG	A protein involved in tRNA modification	No (gene present is a pseudogene)
Sulphur metabolism		
sufS or *iscS*	Cysteine desulfurase	Yes (*iscS*)
sufBC or *iscAU*	Cysteine desulfurase accessory proteins	Yes (*iscU*, but not *iscA*)
RNA modification		
rlu genes	Ribosomal large-subunit pseudo-uridine synthase genes	No
Translation		
infA	IF-1	Yes
infB	IF-2	Yes
infC	IF-3	Yes
fusA	EF-G	Yes
tsf	EF-Ts	No
prfA	RF-1	No
prfB	RF-2‡	No
frr	Ribosome-recycling factor	No
def	Peptide deformylase	No
alaS	Alanyl-tRNA synthetase	No
gltX	Glutamyl-tRNA synthetase	No
glyQ	Glycyl-tRNA synthetase α-subunit	No
ileS	Isoleucyl-tRNA synthetase	No
metG	Methionyl-tRNA synthetase	No
pheS	Phenylalanyl-tRNA synthetase, α-subunit	No
trpS	Tryptophanyl-tRNA synthetase	No
valS	Valyl-tRNA synthetase	No
rps genes	30S ribosomal subunit proteins S1, S2, S3, S4, S5, S7, S8, S9, S10, S11, S12, S13, S14, S16, S17, S18 and S19 (17 of the 21 possible subunits)	Yes
rpl and *rpm* genes	50S ribosomal subunit proteins L2, L3, L4, L5, L6, L11, L13, L14, L15, L16, L20, L22, L27, L28, L33 and L36 (16 of the 32 possible subunits)	Yes (except *rplE* and *rplV*, encoding L5 and L22, for which the genes present are pseudogenes)

Table 2 (Cont.) | **Genes conserved in four of the tiny-genome symbionts***

Gene	Product	Present in '*Candidatus* Tremblaya princeps'?
rrsA	16S rRNA	Yes
rrlA	23S rRNA	Yes
rrfA	5S rRNA	Yes
tRNA genes	tRNAs recognizing codons for Met (three), Gly (two), Cys, Phe, Lys, Ala, Glu, Pro, Gln and Ile	No (except for those for Met, Lys and Ala)

DNAP, DNA polymerase; EF, elongation factor; IF, translation initiation factor; RF, peptide chain release factor; RNAP, RNA polymerase; rRNA, ribosomal RNA. *Genes conserved in '*Candidatus* Sulcia muelleri', '*Candidatus* Zinderia insecticola', '*Candidatus* Carsonella ruddii' and '*Candidatus* Hodgkinia cicadicola'. ‡RF-2 is missing in '*Ca.* Hodgkinia cicadicola' and '*Ca.* Zinderia insecticola'; they have reassigned UGA from a stop codon to a tryptophan codon, so the protein is no longer needed.

B. aphidicola. Likewise, no genes of bacterial origin were found in the body louse genome[93]. Together, these results suggest that gene transfer to hosts is not common in symbionts of insects and is thus not a factor that enables these organisms to survive with so few genes. These finding do not rule out the importation of host proteins that may complement lost symbiont genes, although a recent large-scale proteomics screen of pea aphid bacteriocytes did not detect aphid proteins in *B. aphidicola* cells[76]. Potentially, gene transfer, protein importation or both will be found to occur in systems in which the symbiont (or symbionts) has experienced greater levels of genome erosion than those observed in *B. aphidicola*.

The stages and end point of genome reduction. Tiny symbiont genomes represent the most reduced stages of a progression of genomic changes that result from the small population size and asexuality that can occur in bacteria which acquire a host-restricted lifestyle (FIG. 3). Although these compact genomes, which are free of mobile DNA and small intergenic spacers, are one long-term outcome of this progression, the initial stages, represented by some recently evolved symbionts and pathogens, are characterized by proliferation of mobile elements, chromosome rearrangements, gene inactivation, pseudogene accumulation and deletions, all of which reflect increased fixation of deleterious mutations[34,36,37,97]. Over time, ongoing deletion removes pseudogene fragments and mobile elements, and gene loss continues, resulting in small, compact genomes such as those of *B. aphidicola* and *M. genitalium*. As we have described, recent discoveries have revealed that some symbiotic lineages have proceeded further down the path of genome reduction than was once considered possible, losing many genes that are considered essential for cellular life.

Co-adaptation by the host is crucial in enabling the loss of additional symbiont genes, such as those underlying the production of cell envelope components. Therefore, restriction of extremely tiny genomes can only occur in beneficial symbionts, as host co-adaptation would not occur in a pathogenic association. As described above, host co-adaptation and gene loss do not require transfer of symbiont genes to the host genome, but gene loss can be further enabled by the acquisition of additional symbionts that complement and replace the biosynthetic capabilities of the original symbiont with a tiny genome, as in the case of '*Ca.* Sulcia muelleri' and its co-symbionts[52,78,98]. The genes that are lost owing to the presence of a co-symbiont appear to be involved primarily in the provision of nutrients to the host; genes involved in DNA replication, transcription or translation do not show co-dependent complementary patterns[78], with the possible exception of the genes of '*Ca.* Tremblaya princeps' in the mealybug *Planococcus citri*[43]. '*Ca.* Carsonella ruddii', with a genome of 160 kb, is the sole symbiont in its psyllid host[53], indicating that extreme genome reduction does not require the presence of a co-symbiont. By contrast, the presence of a co-symbiont appears to have been key in the extreme case of '*Ca.* Tremblaya princeps', which probably depends on the host for cell envelope components and on its intracellular co-symbiont, '*Candidatus* Moranella endobia', for parts of the translation machinery[43]. Clearly these tiny genomes should not be regarded as highly efficient. Instead, current evidence suggests that they are more correctly viewed as decrepit and barely managing to persist, even with external support from hosts or partner symbionts.

Are these symbionts, organelles or something in between? Although mitochondria and chloroplasts are derived from symbiotic bacteria[86,87], research on the evolution and functioning of organelles and bacterial symbionts has, by and large, proceeded independently[99]. This lack of integration partly reflects the seemingly clear genomic distinction between organelles and bacterial symbionts, as mitochondria and plastids have undergone high levels of genome reduction and now import most of their functional proteins from the eukaryotic cells in which they reside, whereas most bacterial symbionts retain more robust gene sets that are considered complete enough to support autonomous life.

Although the line separating organelle from endosymbiont is not clear[100–103] and depends largely on definitions, mitochondria and chloroplasts have evolved to a level of integration with the host cell that renders them clearly distinct from most endosymbiotic bacteria. The eukaryotic cells in which the organelles reside have evolved specialized machinery to import a vast array of proteins into the organelles, allowing organellar genomes to undergo extreme gene loss and, in some cases, to lose their genomes entirely[88]. In contrast to endosymbionts, true organelles are found in nearly every eukaryotic cell, are genetically intimately integrated with their host and

Psyllid
A type of small insect that feeds on plant phloem sap.

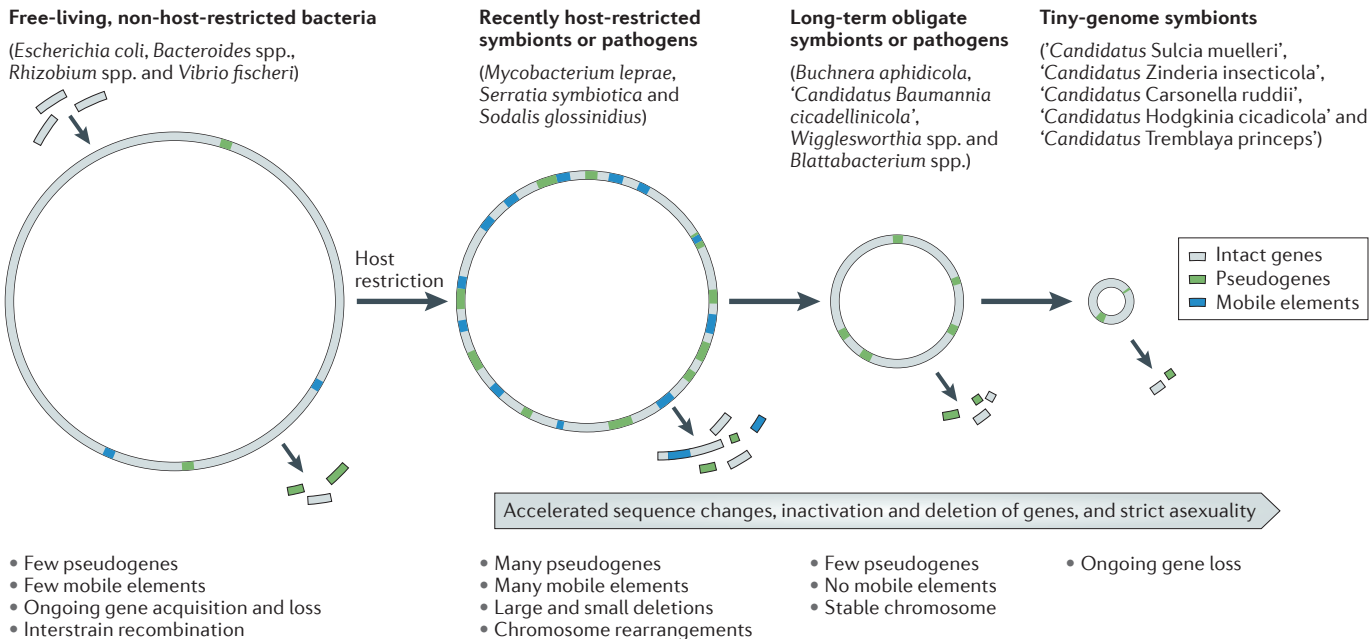


Figure 3 | **Stages of genome reduction in host-restricted bacteria for which small population sizes and an asexual life cycle result in mutation fixation.** Note that some symbionts and pathogens (such as *Rhizobium* spp. and *Vibrio fischeri*) that can persist in the outside environment and that re-infect hosts frequently do not undergo these genomic changes.

no longer encode most of the machinery required to perform the core genetic processes of DNA replication, transcription and translation; furthermore, their division and replication are controlled by the nuclear genome.

Concluding remarks

The organisms described here present a conundrum of biological classification. They have smaller genomics encoding fewer proteins than those found in some organelles and viruses, but they differ from these entities in that they retain many genes enabling the core processes for cellular life. They encode far fewer genes than most bacteria but represent one end of a continuum with no clear points of differentiation; known endosymbiont genome sizes range from 139 kb to more than 1,000 kb (FIG. 2). They are restricted to specialized host cells rather than being found in every cell in the host. Nevertheless, many of these symbiont genomes are missing genes that would widely be considered 'essential'. For example, both '*Ca.* Sulcia muelleri' and '*Ca.* Hodgkinia cicadicola' from the cicada host encode only 17 of 20 amino acyl-tRNA synthetases[78]; '*Ca.* Tremblaya princeps' and '*Ca.* Hodgkinia cicadicola' seem to have incomplete sets of tRNAs[43,78]; the '*Ca.* Hodgkinia cicadicola' genome encodes only two genes involved in DNA replication; and all five of the smallest symbiont genomes are incapable of synthesizing a complete cell envelope (TABLE 1). It is tempting to speculate that these genes have been transferred to the host genome or that proteins of host origin have been recruited to perform these functions, but no evidence is yet available to support these hypotheses. Such evidence, especially evidence that host-encoded factors are involved in symbiont DNA replication, transcription or translation, would favour the view that these organisms approach the status of organelles. Clearly, neither relying on a host nor an inability to be cultured in axenic conditions implies organelle status, as these criteria would apply to a large proportion of bacterial species.

No matter what name is given to the bacteria in these symbioses, the results reviewed here represent new and surprising examples of the intimate integration of cells from different lineages. Future work on these systems promises to yield fundamental insights into the limits of cellular evolution, the nature of organelles and what it means to be an autonomous cellular entity.

1. Bak, A. L., Black, F. T., Christiansen, C. & Freundt, E. A. Genome size of mycoplasmal DNA. *Nature* **224**, 1209–1210 (1969).
2. Maniloff, J. & Morowitz, H. J. Cell biology of the mycoplasmas. *Bacteriol. Rev.* **36**, 263–290 (1972).
3. Wallace, D. C. & Morowitz, H. J. Genome size and evolution. *Chromosoma* **40**, 121–126 (1973).
4. Woese, C. R., Maniloff, J. & Zablen, L. B. Phylogenetic analysis of the mycoplasmas. *Proc. Natl Acad. Sci. USA* **77**, 494–498 (1980).
5. Weisburg, W. G., Woese, C. R., Dobson, M. E. & Weiss, E. A common origin of rickettsiae and certain plant pathogens. *Science* **230**, 556–558 (1985).
6. Fraser, C. M. *et al.* The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403 (1995).
7. Andersson, S. G. *et al.* The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**, 133–140 (1998).
8. Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y. & Ishikawa, H. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* **407**, 81–86 (2000).
9. Fraser, C. M. *et al.* Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390**, 580–586 (1997).
10. Mushegian, A. R. & Koonin, E. V. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl Acad. Sci. USA* **93**, 10268–10273 (1996).
11. Itaya, M. An estimation of minimal genome size required for life. *FEBS Lett.* **362**, 257–260 (1995).
12. Mushegian, A. The minimal genome concept. *Curr. Opin. Genet. Dev.* **9**, 709–714 (1999).
13. Koonin, E. V. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Rev. Microbiol.* **1**, 127–136 (2003).

14. Harris, J. K., Kelley, S. T., Spiegelman, G. B. & Pace, N. R. The genetic core of the universal ancestor. *Genome Res.* **13**, 407–412 (2003).
15. Charlebois, R. L. & Doolittle, W. F. Computing prokaryotic gene ubiquity: rescuing the core from extinction. *Genome Res.* **14**, 2469–2477 (2004).
16. Koonin, E. V. How many genes can make a cell: The minimal-gene-set concept. *Annu. Rev. Genomics Hum. Genet.* **1**, 99–116 (2000).
17. Hutchison, C. A. *et al.* Global transposon mutagenesis and a minimal *Mycoplasma genome*. *Science* **286**, 2165–2169 (1999).
18. Akerley, B. J. *et al.* A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *Proc. Natl Acad. Sci. USA* **99**, 966–971 (2002).
19. Kobayashi, K. *et al.* Essential *Bacillus subtilis* genes. *Proc. Natl Acad. Sci. USA* **100**, 4678–4683 (2003).
20. Glass, J. I. *et al.* Essential genes of a minimal bacterium. *Proc. Natl Acad. Sci. USA* **103**, 425–430 (2006).
21. Curnow, A. W. *et al.* Glu-tRNAGln amidotransferase: a novel heterotrimeric enzyme required for correct decoding of glutamine codons during translation. *Proc. Natl Acad. Sci. USA* **94**, 11819–11826 (1997).
22. Moran, N. A., McCutcheon, J. P. & Nakabachi, A. Genomics and evolution of heritable bacterial symbionts. *Annu. Rev. Genet.* **42**, 165–190 (2008).
23. Tamas, I. *et al.* 50 million years of genomic stasis in endosymbiotic bacteria. *Science* **296**, 2376–2379 (2002).
24. van Ham, R. C. *et al.* Reductive genome evolution in *Buchnera aphidicola*. *Proc. Natl Acad. Sci. USA* **100**, 581–586 (2003).
25. Akman, L. *et al.* Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. *Nature Genet.* **32**, 402–407 (2002).
26. Gil, R. *et al.* The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. *Proc. Natl Acad. Sci. USA* **100**, 9388–9393 (2003).
27. Wu, D. *et al.* Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. *PLoS Biol.* **4**, e188 (2006).
28. Moran, N. A. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc. Natl Acad. Sci. USA* **93**, 2873–2878 (1996).
29. Mira, A., Ochman, H. & Moran, N. A. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **17**, 589–596 (2001).
30. Nilsson, A. I. *et al.* Bacterial genome size reduction by experimental evolution. *Proc. Natl Acad. Sci. USA* **102**, 12112–12116 (2005).
Experimental support for the hypothesis that bacteria which are subject to frequent population bottlenecks can rapidly delete large amounts of DNA from their genomes.
31. Kuo, C. H. & Ochman, H. Deletional bias across the three domains of life. *Genome Biol. Evol.* **1**, 145–152 (2009).
32. Fares, M. A., Ruiz-Gonzalez, M. X., Moya, A., Elena, S. F. & Barrio, E. Endosymbiotic bacteria: groEL buffers against deleterious mutations. *Nature* **417**, 398 (2002).
A study showing that high levels of chaperonin, as observed repeatedly in symbiotic bacteria, can ameliorate the effects of deleterious mutations, thus supporting the hypothesis that the rapid protein evolution which is characteristic of small genomes reflects largely deleterious evolution and that elevated expression of heat shock proteins represents a compensatory adaptation.
33. Fernandez, A. & Lynch, M. Non-adaptive origins of interactome complexity. *Nature* **474**, 502–505 (2011).
34. Toh, H. *et al.* Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host. *Genome Res.* **16**, 149–156 (2006).
35. Ochman, H. & Davalos, L. M. The nature and dynamics of bacterial genomes. *Science* **311**, 1730–1733 (2006).
36. 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).
37. Cole, S. T. *et al.* Massive gene decay in the leprosy bacillus. *Nature* **409**, 1007–1011 (2001).
38. Kuo, C. H., Moran, N. A. & Ochman, H. The consequences of genetic drift for bacterial genome complexity. *Genome Res.* **19**, 1450–1454 (2009).
39. McCutcheon, J. P. The bacterial essence of tiny symbiont genomes. *Curr. Opin. Microbiol.* **13**, 73–78 (2010).
40. Burger, G., Gray, M. W. & Lang, B. F. Mitochondrial genomes: anything goes. *Trends Genet.* **19**, 709–716 (2003).
41. Brouard, J. S., Otis, C., Lemieux, C. & Turmel, M. The exceptionally large chloroplast genome of the green alga *Floydiella terrestris* illuminates the evolutionary history of the Chlorophyceae. *Genome Biol. Evol.* **2**, 240–256 (2010).
42. Alverson, A. J. *et al.* Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). *Mol. Biol. Evol.* **27**, 1436–1448 (2010).
43. McCutcheon, J. P. & von Dohlen, C. D. An interdependent metabolic patchwork in the nested symbiosis of mealybugs. *Curr. Biol.* **21**, 1366–1372 (2011).
A description of the smallest reported bacterial genome, that of '*Ca.* Tremblaya princeps', and of the unusually integrated metabolic complementarity of a bacteria-within-a-bacterium symbiosis.
44. Raoult, D. *et al.* The 1.2-megabase genome sequence of Mimivirus. *Science* **306**, 1344–1350 (2004).
45. Fischer, M. G., Allen, M. J., Wilson, W. H. & Suttle, C. A. Giant virus with a remarkable complement of genes infects marine zooplankton. *Proc. Natl Acad. Sci. USA* **107**, 19508–19513 (2010).
46. Sueoka, N. On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl Acad. Sci. USA* **48**, 582–592 (1962).
47. Muto, A. & Osawa, S. The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl Acad. Sci. USA* **84**, 166–169 (1987).
48. Cox, E. C. & Yanofsky, C. Altered base ratios in the DNA of an *Escherichia coli* mutator strain. *Proc. Natl Acad. Sci. USA* **58**, 1895–1902 (1967).
49. Rocha, E. P. & Feil, E. J. Mutational patterns cannot explain genome composition: Are there any neutral sites in the genomes of bacteria? *PLoS Genet.* **6**, e1001104 (2010).
50. Hildebrand, F., Meyer, A. & Eyre-Walker, A. Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet.* **6**, e1001107 (2010).
51. Hershberg, R. & Petrov, D. A. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet.* **6**, e1001115 (2010).
Along with reference 50, provides evidence of a universal (G or C)→(A or T) mutational bias in bacteria.
52. McCutcheon, J. P. & Moran, N. A. Functional convergence in reduced genomes of bacterial symbionts spanning 200 My of evolution. *Genome Biol. Evol.* **2**, 708–718 (2010).
53. Nakabachi, A. *et al.* The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science* **314**, 267 (2006).
A report of the first discovery of a tiny cellular genome that is only about one-third the size of the smallest previously reported bacterial genome but retains some genes that are devoted to nutrition of the host insect.
54. Rocha, E. P. & Danchin, A. Base composition bias might result from competition for metabolic resources. *Trends Genet.* **18**, 291–294 (2002).
55. Bentley, S. D. & Parkhill, J. Comparative genomic structure of prokaryotes. *Annu. Rev. Genet.* **38**, 771–792 (2004).
56. Lind, P. A. & Andersson, D. I. Whole-genome mutational biases in bacteria. *Proc. Natl Acad. Sci. USA* **105**, 17878–17883 (2008).
Experimental support for the role of DNA repair enzymes and small effective population sizes in the decreased GC content seen in most endosymbiont genomes.
57. McCutcheon, J. P., McDonald, B. R. & Moran, N. A. Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. *PLoS Genet.* **5**, e1000565 (2009).
58. Knight, R. D., Freeland, S. J. & Landweber, L. F. Rewiring the keyboard: evolvability of the genetic code. *Nature Rev. Genet.* **2**, 49–58 (2001).
59. Maniloff, J. in *Molecular Biology and Pathogenicity of Mycoplasmas* (eds Razin, S. & Herrmann, R.) 31–44 (Kluwer Academic Publishers, New York, 2002).
60. Knight, R. D., Landweber, L. F. & Yarus, M. How mitochondria redefine the code. *J. Mol. Evol.* **53**, 299–313 (2001).
A good overview of the many hypotheses to explain codon reassignments in mitochondria.
61. Osawa, S. & Jukes, T. H. Evolution of the genetic code as affected by anticodon content. *Trends Genet.* **4**, 191–198 (1988).
62. Osawa, S., Jukes, T. H., Watanabe, K. & Muto, A. Recent evidence for evolution of the genetic code. *Microbiol. Rev.* **56**, 229–264 (1992).
63. Santos, M. A., Moura, G., Massey, S. E. & Tuite, M. F. Driving change: the evolution of alternative genetic codes. *Trends Genet.* **20**, 95–102 (2004).
64. Andersson, S. G. & Kurland, C. G. Genomic evolution drives the evolution of the translation system. *Biochem. Cell Biol.* **73**, 775–787 (1995).
65. Hansen, A. K. & Moran, N. A. Aphid genome expression reveals host-symbiont cooperation in the production of amino acids. *Proc. Natl Acad. Sci. USA* **108**, 2849–2854 (2011).
Work showing a high level of coordination between gene expression in the aphid host and the *B. aphidicola* symbiont, and highlighting the types of host co-adaptations that allow genome reduction in mutualistic endosymbionts.
66. Daniel, R. A. & Errington, J. Control of cell morphogenesis in bacteria: two distinct ways to make a rod-shaped cell. *Cell* **113**, 767–776 (2003).
67. Wachi, M. *et al.* Mutant isolation and molecular cloning of *mre* genes, which determine cell shape, sensitivity to mecillinam, and amount of penicillin-binding proteins in *Escherichia coli*. *J. Bacteriol.* **169**, 4935–4940 (1987).
68. Wachi, M., Doi, M., Okada, Y. & Matsuhashi, M. New *mre* genes *mreC* and *mreD*, responsible for formation of the rod shape of *Escherichia coli* cells. *J. Bacteriol.* **171**, 6511–6516 (1989).
69. Henriques, A. O., Glaser, P., Piggot, P. J. & Moran, C. P., Jr. Control of cell shape and elongation by the *rodA* gene in *Bacillus subtilis*. *Mol. Microbiol.* **28**, 235–247 (1998).
70. Leaver, M., Dominguez-Cuevas, P., Coxhead, J. M., Daniel, R. A. & Errington, J. Life without a wall or division machine in *Bacillus subtilis*. *Nature* **457**, 849–853 (2009).
The demonstration that few steps are required to form cell wall-less 'L-forms' of *Bacillus subtilis*, which become polymorphic spheres and divide by an unusual, FtsZ-independent extrusion–resolution mechanism. This work highlights the problem in defining a universal set of essential genes, as a single point mutation renders the 'essential' *ftsZ* gene non-essential.
71. Moran, N. A., Tran, P. & Gerardo, N. M. Symbiosis and insect diversification: an ancient symbiont of sap-feeding insects from the Bacterial phylum Bacteroidetes. *Appl. Environ. Microbiol.* **71**, 8802–8810 (2005).
72. Dufresne, A., Garczarek, L. & Partensky, F. Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biol.* **6**, R14 (2005).
73. Moran, N. A. Microbial minimalism: genome reduction in bacterial pathogens. *Cell* **108**, 583–586 (2002).
74. Hara, E. *et al.* The predominant protein in an aphid endosymbiont is homologous to an *E. coli* heat shock protein. *Symbiosis* **8**, 271–283 (1990).
75. Baumann, P., Baumann, L. & Clark, M. A. Levels of *Buchnera aphidicola* chaperonin GroEL during growth of the Aphid *Schizaphis graminum*. *Curr. Microbiol.* **32**, 279–285 (1996).
76. Poliakov, A. *et al.* Large-scale label-free quantitative proteomics of the pea aphid-Buchnera symbiosis. *Mol. Cell. Proteomics* **10**, M110.007039 (2011).
77. Haines, L. R., Haddow, J. D., Aksoy, S., Gooding, R. H. & Pearson, T. W. The major protein in the midgut of teneral *Glossina morsitans* morsitans is a molecular chaperone from the endosymbiotic bacterium *Wigglesworthia glossinidia*. *Insect Biochem. Mol. Biol.* **32**, 1429–1438 (2002).
78. McCutcheon, J. P., McDonald, B. R. & Moran, N. A. Convergent evolution of metabolic roles in bacterial co-symbionts of insects. *Proc. Natl Acad. Sci. USA* **106**, 15394–15399 (2009).
79. Tokuriki, N. & Tawfik, D. S. Chaperonin overexpression promotes genetic variation and enzyme evolution. *Nature* **459**, 668–673 (2009).
80. Huang, C. Y., Lee, C. Y., Wu, H. C., Kuo, M. H. & Lai, C. Y. Interactions of chaperonin with a weakly active anthranilate synthase from the aphid endosymbiont *Buchnera aphidicola*. *Microb. Ecol.* **56**, 696–703 (2008).
81. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36 (2000).
82. Bjork, G. R. *et al.* Transfer RNA modification. *Annu. Rev. Biochem.* **56**, 263–287 (1987).

83. Kessler, D. Enzymatic activation of sulfur for incorporation into biomolecules in prokaryotes. *FEMS Microbiol. Rev.* **30**, 825–840 (2006).
84. Kambampati, R. & Lauhon, C. T. IscS is a sulfurtransferase for the *in vitro* biosynthesis of 4-thiouridine in *Escherichia coli* tRNA. *Biochemistry* **38**, 16561–16568 (1999).
85. Gardner, M. J. *et al.* Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. *Science* **309**, 134–137 (2005).
86. Gray, M. W., Burger, G. & Lang, B. F. Mitochondrial evolution. *Science* **283**, 1476–1481 (1999).
87. Timmis, J. N., Ayliffe, M. A., Huang, C. Y. & Martin, W. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nature Rev. Genet.* **5**, 123–135 (2004).
88. Palmer, J. D. Organelle genomes: going, going, gone! *Science* **275**, 790–791 (1997).
89. Truscott, K. N., Brandner, K. & Pfanner, N. Mechanisms of protein import into mitochondria. *Curr. Biol.* **13**, R326–R337 (2003).
90. Schleiff, E. & Soll, J. Travelling of proteins through membranes: translocation into chloroplasts. *Planta* **211**, 449–456 (2000).
91. Andersson, J. O. Evolutionary genomics: is *Buchnera* a bacterium or an organelle? *Curr. Biol.* **10**, R866–R868 (2000).
92. Consortium, T. I. A. G. Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLoS Biol.* **8**, e1000313 (2010).
93. Kirkness, E. F. *et al.* Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proc. Natl Acad. Sci. USA* **107**, 12168–12173 (2010).
The complete louse and endosymbiont genomes reveal that no bacterial genes have been transferred to the insect genome and that genome reduction in '*Ca*. Riesia pediculicola' has not been associated with gene transfer to the host, as is common in organelles.
94. Nikoh, N. *et al.* Bacterial genes in the aphid genome: absence of functional gene transfer from *Buchnera* to its host. *PLoS Genet.* **6**, e1000827 (2010).
An exhaustive search of the aphid genome for bacterial genes, showing that the endosymbiont *B. aphidicola* has not achieved its small genome via a process of transfer of functional genes to the nuclear genome of its hosts. In this case at least, this process of gene transfer can be ruled out, distinguishing *B. aphidicola* from organelles.
95. Kondo, N., Nikoh, N., Ijichi, N., Shimada, M. & Fukatsu, T. Genome fragment of *Wolbachia* endosymbiont transferred to X chromosome of host insect. *Proc. Natl Acad. Sci. USA* **99**, 14280–14285 (2002).
96. Hotopp, J. C. *et al.* Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* **317**, 1753–1756 (2007).
97. Andersson, J. O. & Andersson, S. G. Genome degradation is an ongoing process in *Rickettsia*. *Mol. Biol. Evol.* **16**, 1178–1191 (1999).
98. McCutcheon, J. P. & Moran, N. A. Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. *Proc. Natl Acad. Sci. USA* **104**, 19392–19397 (2007).
99. Keeling, P. J. Endosymbiosis: bacteria sharing the load. *Curr. Biol.* **21**, R623–R624 (2011).
100. Keeling, P. J. & Archibald, J. M. Organelle evolution: what's in a name? *Curr. Biol.* **18**, R345–R347 (2008).
A good overview of the problems in classifying bacteria with reduced genomes as endosymbionts or organelles.
101. Theissen, U. & Martin, W. The difference between organelles and endosymbionts. *Curr. Biol.* **16**, R1016–R1017 (2006).
102. Bhattacharya, D. & Archibald, J. M. The difference between organelles and endosymbionts: response to Theissen and Martin. *Curr. Biol.* **16**, R1017–R1018 (2006).
103. Bhattacharya, D., Archibald, J. M., Weber, A. P. M. & Reyes-Prieto, A. How do endosymbionts become organelles? Understanding early events in plastid evolution. *Bioessays* **29**, 1239–1246 (2007).
104. Buchner, P. *Endosymbiosis of animals with plant microorganisms.* (Interscience, New York, 1965).
105. Baumann, L. & Baumann, P. Cospeciation between the primary endosymbionts of mealybugs and their hosts. *Curr. Microbiol.* **50**, 84–87 (2005).
106. Baumann, L., Thao, M. L., Hess, J. M., Johnson, M. W. & Baumann, P. The genetic properties of the primary endosymbionts of mealybugs differ from those of other endosymbionts of plant sap-sucking insects. *Appl. Environ. Microbiol.* **68**, 3198–3205 (2002).
107. Thao, M. L., Gullan, P. J. & Baumann, P. Secondary (gamma-Proteobacteria) endosymbionts infect the primary (beta-Proteobacteria) endosymbionts of mealybugs multiple times and coevolve with their hosts. *Appl. Environ. Microbiol.* **68**, 3190–3197 (2002).
108. Kono, M., Koga, R., Shimada, M. & Fukatsu, T. Infection dynamics of coexisting beta- and gammaproteobacteria in the nested endosymbiotic system of mealybugs. *Appl. Environ. Microbiol.* **74**, 4175–4184 (2008).
109. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
110. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005).
111. Thao, M. L. & Baumann, P. Evolutionary relationships of primary prokaryotic endosymbionts of whiteflies and their hosts. *Appl. Environ. Microbiol.* **70**, 3401–3406 (2004).
112. Stewart, G. C. Taking shape: control of bacterial cell wall biosynthesis. *Mol. Microbiol.* **57**, 1177–1181 (2005).
113. Silverman, D. J., Wisseman, C. L., Jr & Waddell, A. *In vitro* studies of *Rickettsia*-host cell interactions: ultrastructural study of *Rickettsia prowazekii*-infected chicken embryo fibroblasts. *Infect. Immun.* **29**, 778–790 (1980).
114. Tully, J. G., Taylor-Robinson, D., Cole, R. M. & Rose, D. L. A newly discovered mycoplasma in the human urogenital tract. *Lancet* **1**, 1288–1291 (1981).
115. Schroder, D. *et al.* Intracellular endosymbiotic bacteria of *Camponotus* species (carpenter ants): systematics, evolution and ultrastructural characterization. *Mol. Microbiol.* **21**, 479–489 (1996).
116. Aksoy, S. *Wigglesworthia* gen. nov. and *Wigglesworthia glossinidia* sp. nov., taxa consisting of the mycetocyte-associated, primary endosymbionts of tsetse flies. *Int. J. Syst. Bacteriol.* **45**, 848–851 (1995).
117. Moran, N. A., Dale, C., Dunbar, H., Smith, W. A. & Ochman, H. Intracellular symbionts of sharpshooters (Insecta: Hemiptera: Cicadellinae) form a distinct clade with a small genome. *Environ. Microbiol.* **5**, 116–126 (2003).
118. Griffiths, G. W. & Beck, S. D. Effects of antibiotics on intracellular symbiotes in the pea aphid, *Acyrthosiphon pisum*. *Cell Tissue Res.* **148**, 287–300 (1974).
119. von Dohlen, C. D., Kohler, S., Alsop, S. T. & McManus, W. R. Mealybug beta-proteobacterial endosymbionts contain gamma-proteobacterial symbionts. *Nature* **412**, 433–436 (2001).
120. Gomez-Valero, L. *et al.* Coexistence of *Wolbachia* with *Buchnera aphidicola* and a secondary symbiont in the aphid *Cinara cedri*. *J. Bacteriol.* **186**, 6626–6633 (2004).

Acknowledgements
Part of the work leading to this Review was supported by US National Science Foundation (NSF) awards 0626716 and 1062363 to N.A.M., and J.P.M. was supported by the NSF Montana Experimental Program to Stimulate Competitive Research grant EPS-0701906 during the writing of this Review.

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
John P. McCutcheon's homepage:
http://mccutcheonlab.dbs.umt.edu/
Nancy A. Moran's homepage:
http://www.yale.edu/moran/

ALL LINKS ARE ACTIVE IN THE ONLINE PDF