Molecular Phylogenetics and Evolution 217 (2026) 108532

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Large-scale phylogenomics reveals convergent genome evolution across repeated transitions to endosymbiosis in Enterobacterales

Giobbe Forni [a,1], Jacopo Martelossi [a,b,1], Benoit Morel [c,d], Dario Pistone [e], Claudio Bandi [e], Matteo Montagna [f,g,*]

[a] Department of Biological, Geological and Environmental Sciences (BiGeA), University of Bologna, Bologna 40126, Italy
[b] Senckenberg Research Institute and Natural History Museum, Frankfurt am Main 60325, Germany
[c] Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies, Heidelberg 69118, Germany
[d] Institute for Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe 76131, Germany
[e] Department of Biosciences, University of Milan, Milan 20133, Italy
[f] Department of Agricultural Sciences, University of Naples Federico II, Portici 80055, Italy
[g] Interuniversity Center for Studies on Bioinspired Agro-Environmental Technology (BAT Center), University of Naples Federico II, Portici 80055, Italy

ARTICLE INFO

Keywords:
Symbiosis
Comparative genomics
Phylogenetics
Gene family evolution
Enterobacterales

ABSTRACT

Symbiogenesis stands among the major transitions in the history of life on Earth. Over the past three decades, extensive research has focused on specific host-symbiont associations to investigate their genome evolution. However, the idiosyncratic sequence evolution of endosymbionts has made it challenging to establish a robust phylogenetic framework for identifying broad-scale evolutionary patterns. Here, we establish the first genome-scale phylogenomic resolution for the Enterobacterales order, encompassing both free-living and endosymbiont species, and provide an analysis of gene loss and acquisition dynamics at scale. By examining over 200 genomes, we show remarkable consistency in phenomena previously known from scattered observations: a spike in gene loss invariably accompanies the shift to endosymbiosis, followed by a slower but continuous rate of gene erosion; gene acquisition processes are reduced after the lifestyle shift. Furthermore, convergence in gene family loss across independent and distantly related symbiotic lineages is observed, with genes having conserved functions and evolving under strong constraints lost at lower rates. Our results unify scattered observations into a broad-scale view of the consequences of endosymbiont genome evolution and highlight the roles of gene essentiality and dispensability in shaping convergent evolutionary trajectories.

1. Introduction

Symbiogenesis stands among the major transitions in the history of life on Earth, with the origin of the eukaryotic cells representing the most iconic case (Margulis 1981; Sapp 1994; Raval et al. 2022). Associations with prokaryotic endosymbionts pervade the evolution of most eukaryotic lineages and represent a major force for their diversification, allowing the exploitation of new resources and colonization of novel ecological niches (Kaiwa et al. 2014; Muñoz-Gómez et al. 2021; Maire et al. 2021; Cornwallis et al. 2023; Bennett et al. 2024). At the same time, this shift in lifestyle profoundly altered prokaryotes' genome evolution (Moran and Bennett, 2014), with dynamics similar to other host-dependent lifestyles, like those of pathogens and obligate parasites (Hershberg et al. 2007; Wernegreen 2015; Weinert and Welch 2017). Besides rapid rates of sequence evolution and AT-rich base composition, the most striking consequence of a host-dependent lifestyle is reductive genome evolution (McCutcheon and Moran 2012). The latter is clearly evident in bacterial endosymbionts, a heterogeneous collection of microorganisms where the smallest known bacterial genomes have been recorded (McCutcheon and Moran 2012). Starting from a size of ~3 to ~6 Mb typical of free-living bacteria, endosymbionts' genomes underwent a dramatic reduction, as exemplified by "*Candidatus* Nasuia deltocephalincola" (~112 kb; Moran and Bennett 2014), a symbiont of phloem-feeding leafhoppers, or "*Candidatus* Pinguicoccus supinus" (~158 kb; Williams et al. 2021), found in the ciliate *Euplotes vanleeuwenhoeki*.

* Corresponding author at: Department of Agricultural Sciences, University of Naples Federico II, Portici 80055, Italy.
E-mail address: matteo.montagna@unina.it (M. Montagna).
[1] Equally contributing authors.

https://doi.org/10.1016/j.ympev.2026.108532
Received 24 January 2025; Received in revised form 12 December 2025; Accepted 5 January 2026
Available online 5 January 2026

Evidence based on specific transitions to endosymbiosis (*e.g.,* in *Buchnera aphidicola*; Chong et al. 2019) suggests that endosymbiont genome reduction is associated with large-scale gene loss (Latorre and Manzano-Marín 2017) and that it takes place in two main phases: an initial drastic reduction followed by an extended and slower process of erosion (Wernegreen 2015; McCutcheon et al. 2019). These two phases are likely associated with distinct phenomena: the early one may result from diminished functional necessity due to the safe and metabolically rich environment ensured by the host (Moya et al. 2008; Castelli et al. 2024), while the subsequent one is caused by a slower accumulation of deleterious mutations and pseudogenization due to minimal recombination rates and recurrent population bottlenecks (Dale et al. 2003; McCutcheon and Moran 2012). Nonetheless, genome reduction follows a partially deterministic path and results in the removal of functional redundancy (Mendonça et al. 2011; Boyd et al. 2024). Regarding the dynamics underlying gene acquisition – specifically horizontal transfer and duplication, which play a significant role in shaping bacterial gene content (Puigbò et al. 2014) – it is known that these processes are diminished in endosymbionts (Lerat et al. 2005). This process is relevant to the hypothesis that sexual reproduction and recombination had an original and primary function related to processes of genome repair or compartmentalization of deleterious mutations (*e.g.,* Bernstein et al. 1985; Kondrashov 1998). According to these theories, in the absence of sexual reproduction or recombination, deleterious mutations are expected to accumulate, thus leading to pseudogenization and gene loss, as originally proposed by the Muller's ratchet hypothesis (Muller 1964; Andersson and Hughes 1996).

However, a comprehensive and integrated understanding of the processes that shape gene content is hampered by the difficulty in providing a reliable and large-scale phylogenetic framework encompassing free-living and endosymbiotic species, due to the AT-rich genomes and accelerated evolutionary rates of endosymbionts (Husník et al. 2011). Such features violate the assumptions of traditional phylogenetic models, eventually leading to systematic error and artifactual clustering of endosymbiont species (Naser-Khdour et al. 2019).

This study utilizes the Enterobacterales clade as a model to understand how the opposing forces of gene acquisition and loss shaped bacterial genome content throughout the evolution of endosymbiosis. This diverse clade, whose typical representatives are rod-shaped, facultatively anaerobic, and non-spore-forming bacteria, exhibits multiple independent transitions to endosymbiosis, mostly in association with insect hosts, but also with nematodes and leeches (Husník et al. 2011; Manzano-Marin et al. 2015), typically providing nutritional supplementation to their hosts (*e.g.*, blood and phloem feeders; Wilson and Duncan 2015; Duron and Gottlieb 2020). These endosymbionts present a high disparity in the strength of interaction they established: they include intracellular (*Buchnera* spp.; Gil et al. 2002, Moran 2021) or extracellular species ("*Candidatus* Pantoea edessiphila"; Otero-Bravo et al. 2018), facultative ("*Candidatus* Regiella insecticola"; Vorburger et al. 2010) or obligate associations (*Symbiopectobacterium* spp.; Martinson et al. 2020). Enterobacterales include also numerous pathogens of medical and veterinary relevance (*e.g.*, *Yersinia pestis* and *Escherichia coli*) alongside plant pathogens (*e.g.*, *Erwinia amylovora* and *Pectobacterium carotovorum*) and free-living bacteria inhabiting diverse terrestrial and aquatic ecosystems.

By analyzing over 200 Enterobacterales genomes, representing a diverse sampling of lineages spanning a wide range of lifestyles and host associations, we established a novel phylogenomic framework to assess whether phenomena observed in isolated lineages are consistent at a broader scale. This approach allowed us to revisit several long-standing observations from a broader perspective: Is a significant spike in gene loss a universal feature of the establishment of endosymbiosis? Is this initial phase of rapid gene loss typically followed by a less intense erosion of gene content? Do gene acquisition processes consistently diminish after shifts to an endosymbiotic lifestyle? And to what extent are gene family losses convergent across distinct endosymbiont clades? Our findings provide a unified view of the evolutionary mechanisms shaping endosymbiont genome evolution, highlighting the roles of functional constraints in driving parallel evolutionary trajectories.

2. Results

Enterobacterales phylogenomic inferences and reconstructions of endosymbiosis shifts – The dataset generated consists of 207 Enterobacterales species, including representatives of the seven described families and taxa characterized by different lifestyles, alongside three outgroups (Table S1). This resulted in a total of 103 multiple sequence alignments (23,531 alignment positions) for concatenation-based inferences and 10,564 (2,869,736 alignment positions) for gene family inferences. A total of 30 and 711 multiple sequence alignments, respectively, reject the phylogenetic assumptions of stationarity and homogeneity. When mixture models are included in model selection for concatenation-based inferences, different variants of the C60 model are found to be the best fit, according to the Akaike information criterion (AICc). The AICc scores for all tested models are provided as supplementary files in the Zenodo repository linked in the Data Availability section.

Across the eleven phylogenetic inferences generated, a marginal effect on topology of the use of mixture models and the removal of non-stationary and non-homogeneous genes across our phylogenetic inferences is observed, as shown by the tree-based visualization of normalized Robinson–Foulds distances (Fig. S1). All phylogenetic inferences leveraging concatenation in combination with the removal of non-stationary and non-homogeneous genes and/or with mixture models cluster together. These inferences support the smallest number of shifts to endosymbiosis (nine shifts; Table S2) and show a strong signature of the expected phylogenetic biases, with many endosymbionts clustering together in a single group with long branches. Conversely, all phylogenetic inferences based on gene families and/or SR4 recoding are topologically more similar (Fig. S1). Inference 4 (concatenation + mixture models + SR4 recoding) and inference 11 (gene family inference + SR4 recoding) retrieved the highest number of inferred shifts to endosymbiosis among concatenation-based and gene families-based inferences (respectively eleven and twelve independent shifts; Table S2). The backbone of both inferences is well supported (IQTREE2 ultrafast bootstrap > 95 % and SpeciesRax EQPIC score > 0.2, as recommended in Hoang et al. 2018; Morel et al. 2022) and they retrieved shifts to endosymbiosis that are entirely compatible with those previously inferred in Husník et al. 2011 (Fig. 1; Fig. S2). These two inferences are almost identical (normalized Robinson–Foulds = 0.08; Fig. S2), with the only two major differences among them: (1) the position of "*Candidatus* Pantoea carbekii", which is found within group A in inference 11 and as an independent shift to endosymbiosis in inference 4; (2) the position of the endosymbionts of *F. virgata* and *H. cubana*, which are retrieved as sister to *Sodalis glossinidius* in inference 11 (group E1), while they cluster with group D (*Moranella* + *Hoaglandella* + *Baumannia* + *Westeberhardia* + *Wigglesworthia* + *Blochmannia*) in inference 4. Throughout the different analytical approaches leveraged, some endosymbiont lineages appear inconsistent in their placement, such as *Pantoea carbekii*, *Riesia* spp. and "*Candidatus* Providencia siddallii". Indeed, *Pantoea* resulted paraphyletic, with endosymbiosis that evolved independently at least twice within the genus (groups B and A). On the contrary, most of the endosymbiont groups are phylogenetically consistent throughout the different inferences (Fig. 1; Table S2). These included groups B (the symbionts of *Plautia stali*), C ("*Candidatus* Erwinia dacicola"), F (*Symbiopectobacterium* spp.), I (*Serratia* symbionts), and L (*Hamiltonella* + *Regiella* + *Fukatsuia*).

Gene duplication, transfer, and loss patterns in endosymbionts and free-living species – Contextually with the establishment of the endosymbiosis event (*i.e.,* along the same branches where ancestral state reconstruction analyses inferred the establishment of endosymbiosis) a sharp increase in gene loss is found, with a loss rate over threefolds greater on the branches where endosymbiosis was established (median ± sd =

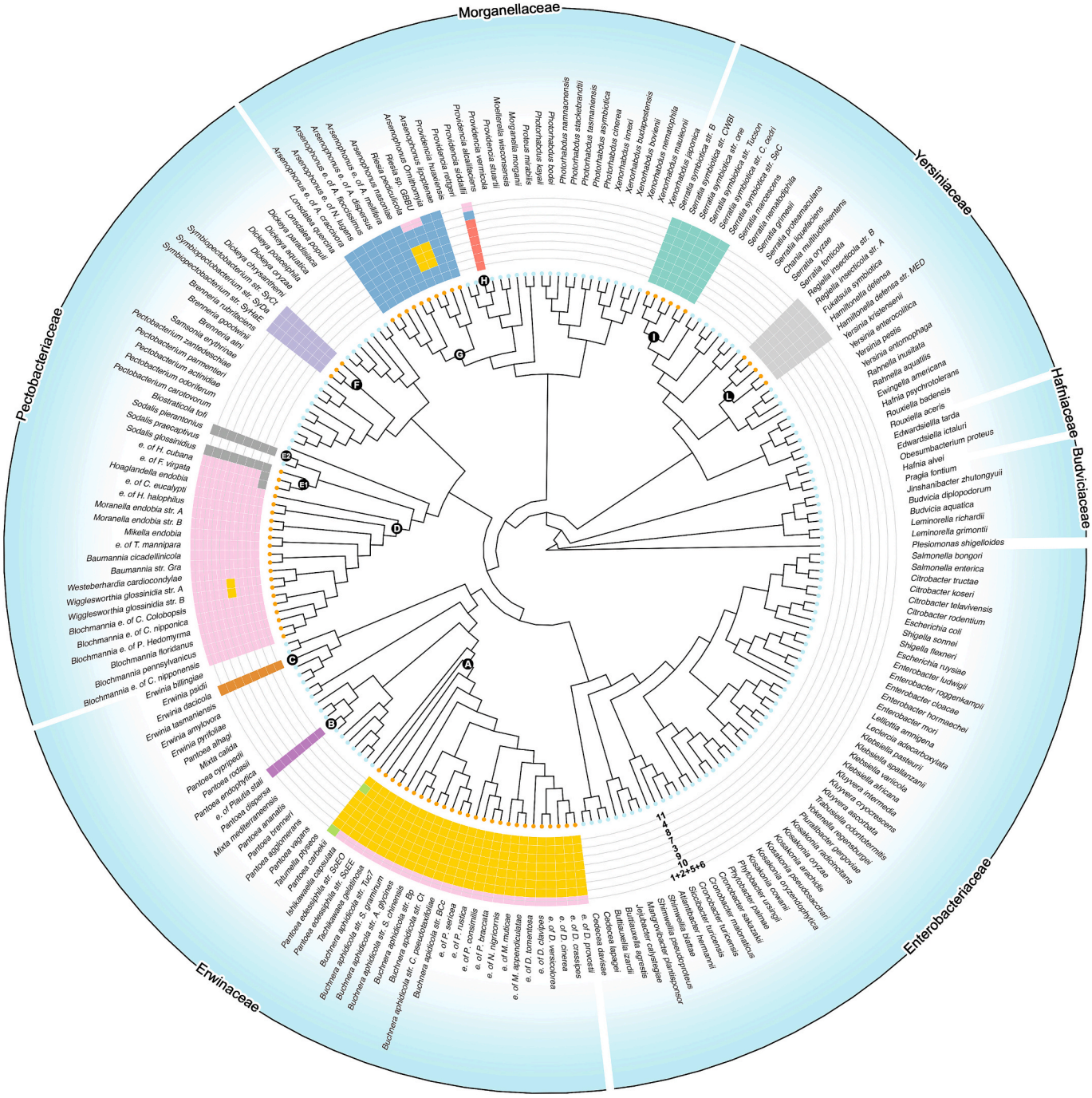

Fig. 1. Phylogenetic reconstruction of the shifts to endosymbiosis along the evolutionary history of Enterobacterales. The phylogeny was obtained using a gene family inference + SR4 recoding (inference 11); leaf color is orange for endosymbiotic taxa and turquoise for free-living ones. Inferred shifts to endosymbiosis are highlighted with letters A to L. The concentric bands of colored tiles, from the innermost to the outermost, correspond to the eleven alternative phylogenetic inferences considered. Within each band, tiles of the same colour indicate a monophyletic cluster of endosymbiotic taxa recovered in that inference; different colours distinguish distinct endosymbiont groupings, and the recurrence of the same colour across bands marks the same grouping across phylogenetic resolutions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

0.592 ± 0.21) than the rate observed on previous and subsequent branches (±3 branches; 0.185 ± 0.123; Figs. 2A and 2B). This pattern is consistently observed across all eleven independent events of endosymbiosis establishment (Fig. S3). Considering only the branches following the transition to endosymbiosis, loss rates are also found to be statistically greater on endosymbiotic branches compared to free-living ones (Fig. 2C; Tab. S3). Gene duplication and transfer rates are both significantly reduced in endosymbionts (Fig. 3B and Fig. 3D; Tab. S3). When the rates of gene loss, transfer and duplication on the branches following the shift to endosymbiosis are considered separately based on the strength of the endosymbiotic association, a significantly higher gene loss rate was observed in endosymbiont lineages characterized by mixed or strict associations with their host, compared to free-living species; loosely associated endosymbionts did not show any significant difference. On the other hand, loosely associated endosymbionts displayed significantly higher duplication rates than free-living species, while both strict and mixed groups exhibited significantly lower duplication rates (Fig. S4A-C; Tab. S3). In addition, considering each

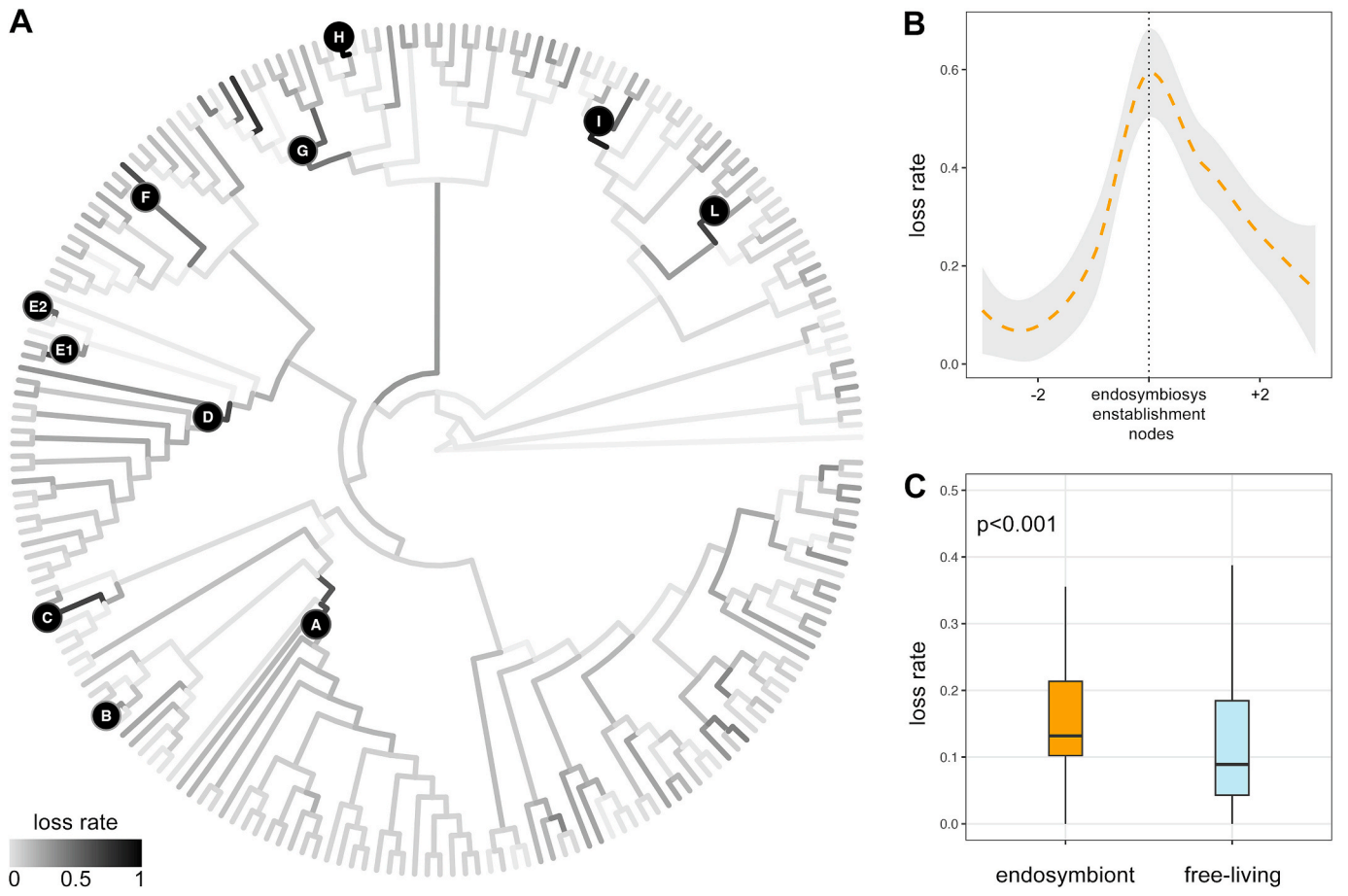

Fig. 2. Two-phase scenario for gene loss dynamics across the shift to endosymbiosis. (A) Rates of inferred gene loss throughout the Enterobacterales phylogeny; letters A to L highlight the eleven inferred shifts to endosymbiosis, as in Fig. 1. (B) Cumulative trend of gene loss rates concurrently to the establishment of endosymbiosis; on the x-axis, 0 represent the node in which the shift to endosymbiosis was inferred to have happened, negative values represent branches preceding that node and positive value represent subsequent branches. Plots for each independent shift are found in Fig. S3. (C) Gene loss rates in free-living and endosymbiotic branches throughout the genome erosion following the shift in lifestyle (*i.e.,* excluding those where the shift to endosymbiosis has been inferred). Plots for each independent shift are found in Fig. S4D.

endosymbiont lineage separately, losses are generally higher, and transfers are generally lower in endosymbionts compared with free-living (Fig. S4D-F).

For the 4,241 orthogroups inferred to have been present in at least one free-living most recent common ancestor of endosymbionts, we estimated the per-family mean gene loss frequency for both free-living species and endosymbionts (*i.e.,* the mean of the loss frequencies of each family in branches inferred in either state). Endosymbionts exhibit significantly higher loss frequencies compared to free-living ones, across all functional categories (Fig. S5; Wilcoxon $p < 0.001$); however, the extent of this increase varied. When considering the Δ mean loss frequency on endosymbiont branches and on free-living branches for each functional category, essential processes such as translation (COG J), cell cycle control (COG D), posttranslational modifications and chaperones (COG O), recombination and repair (COG L) mechanisms are found to have undergone a less intense mean loss frequency acceleration (Δ mean loss frequency < 0.2; Fig. S5). Conversely, many functional categories that are expected to be more dispensable in the new stable environment provided by the host are found to have undergone a marked mean loss frequency acceleration (Δ mean loss frequency > 0.2; Fig. S5), including signal transduction (COG T), and defense (COG V) mechanisms. Regarding the metabolism of macromolecules, nucleotides and coenzymes were found to have undergone a less intense acceleration (COG F and H), while secondary metabolites, inorganic ions, carbohydrates, and amino acids underwent a more intense one (COG Q, P, G, and E) (Fig. S5).

For endosymbionts, per-family mean loss frequencies were also inferred separately for each lineage, and they resulted in being correlated across all eleven endosymbiont lineages (Fig. 4A). There is a significant correlation between mean loss frequencies in endosymbionts (nine out of eleven lineages) and their free-living relatives, although the strength of this correlation is generally lower. When considering the ranking of functional categories based on the mean loss frequencies of their associated families across the eleven independent shifts to endosymbiosis, the categories most consistently lost in endosymbionts are largely the same as those exhibiting the strongest increase in the Δ mean loss frequency (Fig. 4B and S6).

The relationship between evolutionary constraints acting on gene families and their mean loss frequencies was then assessed. A positive Δ mean dN/dS between endosymbionts and free-living species is retrieved across all gene functional categories ($\Delta dN/dS = 0.084 \pm 0.069$), highlighting a consistent signature of relaxation of purifying selection in endosymbionts, regardless of specific gene function (Fig. S7). Gene families retained in all endosymbiotic lineages exhibited lower dN/dS values in free-living species compared to those completely lost in endosymbionts (Kolmogorov-Smirnov test $p < 0.001$; Fig. S8A). This pattern held true even for gene families present only in some endosymbiont species; families retained in more than half of the endosymbiont species had lower dN/dS in free-living bacteria compared to those retained in less than half (Kolmogorov-Smirnov $p < 0.001$; Fig. S8A).

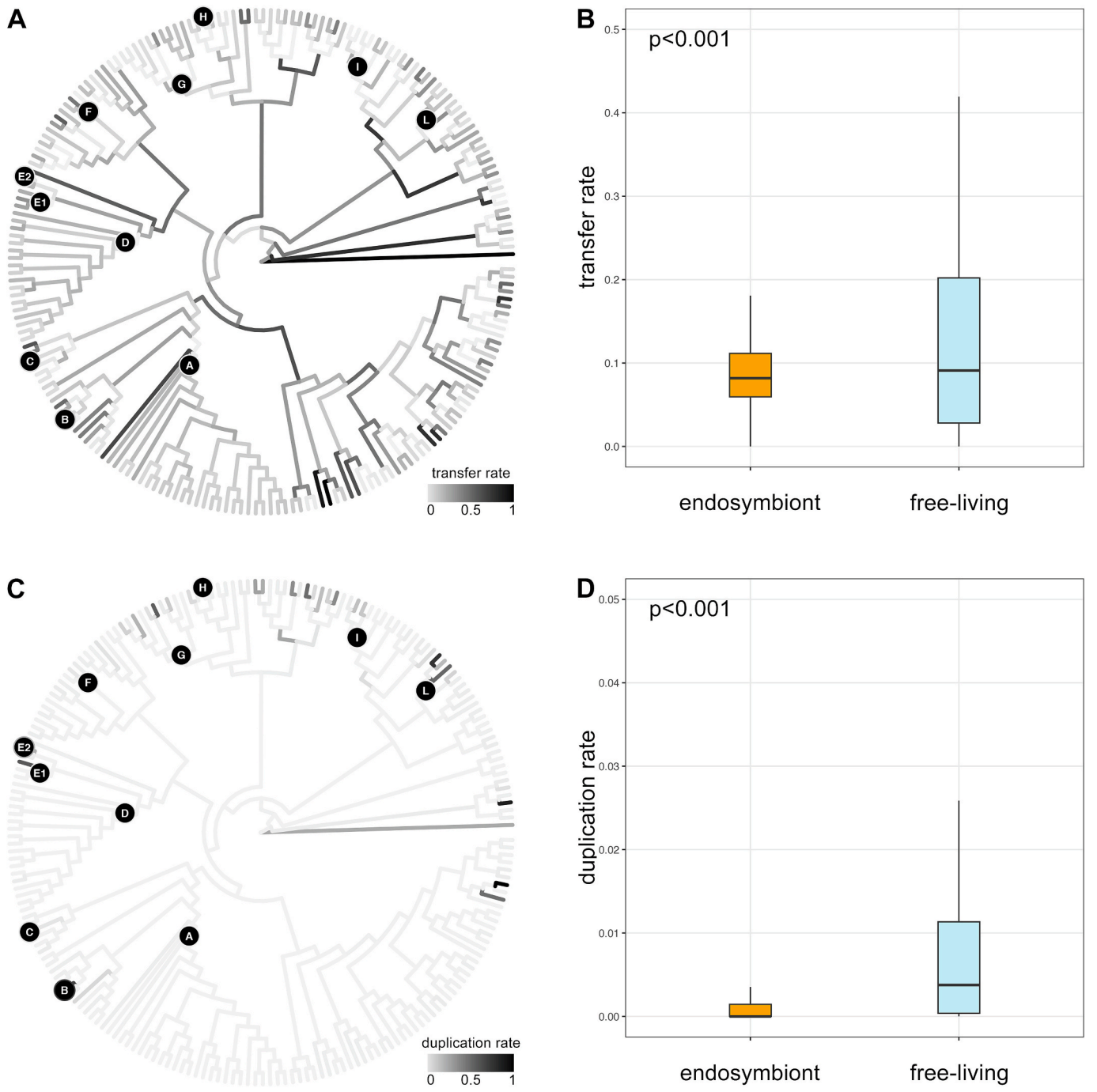

Fig. 3. Reduction of gene acquisition processes after the transition to endosymbiosis. Panels (A) and (C) show the rates of inferred gene transfer and duplication throughout Enterobacterales phylogeny, respectively; letters A to L highlight the eleven inferred shifts to symbiosis, as in Fig. 1. Panels (B) and (D) show the statistical difference of gene transfer rate and gene duplication rate among free-living and endosymbiotic branches. The branches where the shift to endosymbiosis was inferred were excluded; the duplication rates in panel (D) were square-root transformed for visualization. Plots for each independent shift are found in Figs. S4E and F.

Furthermore, a significant positive correlation between gene families' mean loss frequency in endosymbionts and dN/dS in free-living species is observed (Fig. 4C). This trend is consistent across all endosymbiont groups (Spearman $p < 0.001$, $\rho = 0.4$; Fig. S8B).

Results presented in the previous paragraphs were obtained leveraging inference 11 (gene family inference + SR4 recoding; Fig. 1). As a sensitivity test for phylogenetic uncertainty, the same analyses were performed on inference 4 (concatenation + mixture models + SR4 recoding; Fig. S2). The results of analyses using either phylogenetic inference closely match each other: (1) a sharp increase in gene loss rates is concurrent to the shift to endosymbiosis (Fig. 2B and S9A); (2) subsequently to the shift in lifestyle, genome erosion is associated with a marginal increase in loss rates (Fig. 2C and S9B) and a decrease of both duplication (Fig. 3B and S9C) and transfer rates (Fig. 3D and S9D); (3) families' mean loss frequencies are correlated across all endosymbiont lineages (Fig. 4A and S9E); (4) families with more essential functions tend to have lower mean loss frequency ranks (Fig. 4B and S9G); (5) families' mean loss frequency in endosymbionts is correlated with the dN/dS in free-living species (Fig. 4C and S9F).

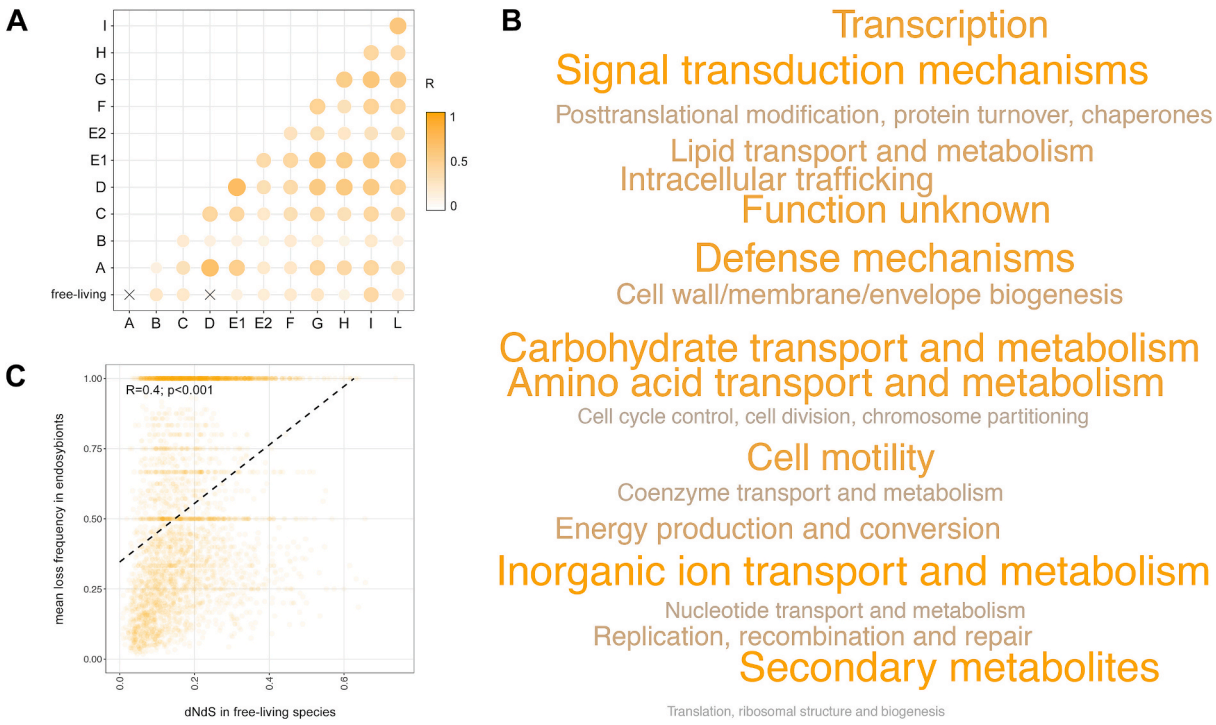

Fig. 4. Endosymbionts' gene loss frequency is correlated across independent shifts in lifestyle. (A) Correlation of orthogroup gene loss frequencies between the eleven endosymbiont clades and free-living species; non-significant tests ($p > 0.001$) are highlighted by a cross symbol, while the correlation strength (r) is reflected by bubble size and color intensity. (B) Word cloud representation of the COG categories whose genes are more consistently lost across independent endosymbiosis shifts. Each COG category was ranked based on their mean loss frequency for each independent endosymbiotic lineage and free-living species; the size in the word cloud plot is proportional to the cumulative sum of all mean loss frequency COG ranks. Ranks are plotted for each independent shift in Fig. S6. (C) Spearman rank correlations between orthogroup dN/dS rates in free-living species and gene loss frequencies in free-living species (blue line) and endosymbionts (orange line). Correlations are plotted for each independent shift in Fig. S8. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3. Discussion

Our phylogenetic inferences consistently retrieved all seven previously established Enterobacterales families as monophyletic (Enterobacteriaceae, Erwiniaceae, Pectobacteriaceae, Yersiniaceae, Hafniaceae, Morganellaceae, and Budviciaceae; Adeolu et al. 2016; Soutar and Stavrinides 2020). However, reconstructed relationships among these families displayed only partial coherence with those presented in previous studies. According to past findings, Budviciaceae was the first lineage to diverge, while Enterobacteriaceae and Erwiniaceae have consistently been retrieved as a monophyletic lineage. However, published studies are not coherent regarding the relationships among other families, and these relationships are often characterized by weak support. In contrast, our findings are supported by two very different inference frameworks and datasets, which yield a consistent topology with strong nodal support (Fig. S2). Among the major lineages of Enterobacterales, the families of Budviciaceae, Hafniaceae, and Enterobacteriaceae appear to lack any transition to an endosymbiotic lifestyle, while others (*e.g.,* Pectobacteriaceae) present multiple transitions to endosymbiosis (Fig. 1). The phylogenetic approaches expected to be more apt to accommodate the biases associated with endosymbiont sequence evolution (Table S2) retrieved shifts to endosymbiosis that are coherent with those reported by Husník et al. (2011) (Fig. 1). The more complex phylogenetic approaches used here retrieved a high number of endosymbiosis transitions proposed so far for the Enterobacterales, but even this high number could represent an underestimation. Leveraging a more taxonomically dense sampling could potentially reveal additional shifts to endosymbiosis occurring at a shallower taxonomic level. This is the case of *Sodalis*, where two independent shifts to endosymbiosis are inferred, which is coherent with the marked disparity in transitional steps – from free-living (*Sodalis praecaptivus*) to facultative (*Sodalis glossinidius*) and obligate ("*Candidatus* Sodalis pierantonii") intracellular mutualist – and is also reflected in the inferred gene loss rates (Fig. 2; Fig. S3). Interestingly, inferred endosymbiont lineages are all nested within, or have a sister relationship with, lineages of animal or plant pathogens. For example, group L (*Regiella*, *Fukatsuia*, *Hamiltonella*) has a sister relationship with *Yersinia* pathogens (as in Degnan et al. 2009), and group F (*Symbiopectobacterium*) is closely related to "soft rot" plant pathogens of the genera *Dickeya*, *Brenneria* and *Pectobacterium* (as in Martinson et al. 2020). These results suggest, once again, that the mechanisms and adaptations underlying pathogenicity may represent a step toward the evolution of much tighter, non-conflictual associations (Ochman and Moran 2001; McCutcheon et al. 2019). This hypothesis is further corroborated even at the genomic level, as pathogens also undergo instances of genome reduction, similar to endosymbionts (Hershberg et al. 2007).

In our analyses, gene loss in endosymbionts was found to consistently follow a two-phase dynamics: i) a massive spike in loss rate concurrent with the transition to endosymbiosis (*i.e.,* on the same node where the shift occurs); ii) a reduced but ongoing loss, consistent with long-term genomic erosion (Fig. 2). These results are coherent with previous findings (Wernegreen 2015; Chong et al. 2019). The spike in gene loss rates occurring concurrently with the establishment of symbiosis is consistent across all shifts to endosymbiosis and appears to be independent of the current strength of the association with the host, as previously observed (Salem et al. 2017). On the other hand, rates of gene loss during genome erosion phase (*i.e.,* after the spike) present some variability across different endosymbiont lineages: lineages displaying a strict association with the host, an obligate intracellular lifestyle, and vertical transmission tend to have higher rates of gene loss compared to endosymbionts with a less intimate association with their host (Fig. S4).

Gene loss is not the only process shaping the evolution of gene

content, as it can be balanced by the acquisition of new genes, primarily via horizontal gene transfer (Puigbò et al. 2014). Gene duplication events have been proposed to play a relatively marginal role in the expansion of bacterial genomes compared to gene transfers (Treangen and Rocha 2011), as also observed here throughout Enterobacterales evolution. Inferences presented here highlight a reduction in gene transfer events in endosymbionts compared to free-living species (Fig. 3). Therefore, genome erosion associated with endosymbiosis appears to be coupled with a reduction of the main mechanism that can counteract the accumulation of deleterious mutations and offset gene losses in free-living species (Muller 1964). While some endosymbionts do occur in physical proximity with other bacteria (Koga et al. 2012) and horizontal gene transfer has indeed been observed in endosymbionts (Nikoh et al. 2014), our findings are consistent with reduced opportunities for gene transfer in the host environment, where endosymbionts have indeed fewer chances of encountering other bacteria compared to free-living taxa (Wernegreen 2015). With gene exchange limited in endosymbionts, gene losses may become effectively irreversible, as is the endosymbiotic lifestyle itself (Husník and Keeling 2019).

A consistent signature of convergence in gene family loss frequencies is observed among independent endosymbiont lineages. This phenomenon is also present in free-living bacteria, yet to a lesser extent (Fig. 4A). Some authors have associated endosymbiont gene loss with multiple, non-exclusive mechanisms, including shifts in mutation rates (Bourguignon et al. 2020; Kinjo et al. 2021) and the relaxation of purifying selection (O'Fallon 2008; Boscaro et al. 2017; Sabater-Muñoz et al. 2017). In all scenarios, genes are expected to be lost when the forces driving their loss outweigh their essentiality (Korona 2011; Bolotin and Hershberg 2016). In our results endosymbionts appear to be more prone to losing genes associated with interactions with a changing environment (*e.g.,* secondary metabolism or signal transduction), which are largely dispensable in the stable niche provided by the host (Fig. 4B). Conversely, genes encoding fundamental and conserved cellular processes (*i.e.,* DNA replication and cell cycle control; Koonin 2003; Chong et al. 2019) show reduced loss frequencies. Interestingly, this difference in loss frequencies among gene families for functional categories is not observed when considering the selection regime (Fig. S7; Pérez-Brocal et al. 2006). Analyses presented here reveal a significant positive correlation between families' mean loss frequencies and their selective constraints before the shift to endosymbiosis (approximated by dN/dS in free-living species; Fig. 4C). Both in silico inferences and experimental knockouts support that purifying selection is more stringent for essential genes than for genes that are more functionally dispensable or redundant (Hirsh and Fraser 2001; Jordan et al. 2002). While previous studies do not exclude a role for context-specific shifts in the intensity or directionality of selection acting on specific genes (Albalat and Cañestro 2016; Kinjo et al. 2021), our results suggest that differences in gene essentiality and dispensability contribute to the observed parallelism.

The genome-scale phylogeny inferred in this study confirms that genome evolutionary dynamics during the evolution of Enterobacterales endosymbiosis are highly consistent, particularly regarding gene loss. Gene loss follows a two-phase dynamic, with an initial massive spike followed by a slower erosion. Gene acquisition processes (duplications and horizontal transfers) are more heterogeneous, but their decline also appears associated with genomic erosion. A marked convergence in gene family loss frequencies across independent and distantly related endosymbiont lineages was observed. Genes under stronger sequence constraints and associated with core cellular functions are lost less frequently, indicating that functional constraints shape these patterns. Overall, these results show that independent lineages repeatedly lose the same dispensable environmental functions while retaining essential cellular processes.

4. Methods

Phylogenetic inferences and endosymbiosis ancestral state reconstructions

– The genomic data of 207 Enterobacterales representatives (including 75 endosymbiont and 132 free-living taxa) were obtained from the NCBI database (Table S1). Starting from the 50 taxa of Husník et al. (2011) the taxa sampling was improved, relying on the phylogenetic resolutions available for each lineage to maximize taxonomic diversity and representativeness of endosymbionts and free-living species. The rationale behind the coding of endosymbiont and free-living species has been whether the bacterium is localized in the body cavity or internal organs of the host (Douglas 2020); as such, the species of the two genera *Photorhabdus* and *Xenorhabdus* were not coded as endosymbionts (Fukruksa et al. 2017). For each inferred clade/lineage of endosymbionts, information on the strength of association with their respective host was retrieved from the literature and reported in Table S1. Briefly, we categorized the endosymbionts clades/lineages into three groups (strict, loose, and mixed) based on several criteria: genome features (including genome size and AT-rich base composition), mode of transmission to the progeny, possibility to be cultured in vitro, and localization within the host tissues and organs (*e.g.,* intracellular, harbored by gut caeca).

Nucleotide sequences corresponding to all CDS features annotated on the assembly were downloaded and translated into amino acids using *transeq* from the EMBOSS package (Rice et al. 2000), and clustering of orthologous gene families was carried out using Orthofinder v. 1.0.6 (Emms and Kelly 2019) with default parameters. Subsequent analyses considered only sequences without non-terminal stop codons. Eleven approaches based on gene concatenation or gene family trees were adopted to obtain a solid phylogenetic framework.

For phylogenetic inferences based on a concatenation approach, orthologous genes (OGs) consisting of single-copy genes with one representative for at least 95% of the taxa were aligned using mafft v.7 (*--auto* option; Katoh and Standley 2013). Trimal v1.4 (Capella-Gutiérrez et al. 2009) was used to remove positions with gaps (*-gappyout* mode) and spurious sequences (*-resoverlap* 0.6 and *-seqoverlap* 50) from multiple sequence alignments (MSAs). 103 MSAs were concatenated using AMAS (Borowiec 2016). Phylogenetic inferences based on concatenation were performed by IQ-TREE 2 (Minh et al. 2020), using the model selection of ModelFinder (*-m* TESTMERGE and *-rclust* 10; Kalyaanamoorthy et al. 2017) and performing 1,000 UFBoot2 bootstrap replicates. Alternative substitution models (including mixture models) were compared using the corrected Akaike information criterion (AICc), which IQ-TREE computes from the maximum log-likelihood and the number of free parameters to approximate model fit while penalizing over-parameterization (lower AICc indicates better fit).

Two inferences were carried out on the complete concatenation, leveraging: (1) gene partition models (*-spp* and *−mset* JTT, WAG, LG), and (2) mixture models (*-mset* JTT, WAG, LG, LG4M, LG4X, CF4, C10-60 and *-madd* JTT + C10-60, WAG + C10-60, LG + C10-60). Two other inferences leveraged instead SR4 recoding of amino acids (amino acids AGNPST as A, CHWY as C, DEKQR as G, and FILMV as T) using gene partition models (3) and mixture models (4), as described by Redmond and McLysaght (2021). Subsequently, gene partitions that rejected the assumption of stationarity and homogeneity were excluded, leveraging the three matched-pairs tests of symmetry (Naser-Khdour et al. 2019).

On this smaller dataset, we conducted four additional analyses that mirrored those carried out on the complete concatenation: using the amino acid alphabet and gene partition models (5); using the amino acid alphabet and mixture models (6); using SR4 recoding and gene partition models (7); and using SR4 recoding and mixture models (8).

For phylogenetic tree inference based on a gene family trees approach, all OGs with more than 3 sequences were aligned and filtered as described above. Then, the ParGene pipeline (Morel et al. 2019) was used to generate gene family trees for each OG. In this pipeline, ModelTest-NG (Darriba et al. 2020) is used to identify the best-fit model of evolution, and RAxML-NG (Kozlov et al. 2019) is used to infer a Maximum Likelihood gene family tree for each MSA. Subsequently, a species phylogeny was inferred from the gene family trees (9) using SpeciesRax (Morel et al. 2022) under a duplication/transfer/loss (DTL)

model, allowing taxa-tree pruning (*prune-species-tree*; recommended for species tree inference in the presence of missing data), estimating both branch lengths in expected number of substitutions per site (*--si-estimate-bl*) and branch support values via paralogy-aware quartets (*--si-quartet-support*). Furthermore, two additional inferences using SpeciesRax with the same parameters described above were carried out: using the amino acid alphabet but including only the gene families that accepted the assumption of stationarity and homogeneity in the tests of symmetry (10) and using all gene families in combination with SR4 recoding (11).

To quantify and summarize the topological similarity of our inferences, a distance matrix using information-based generalized Robinson–Foulds distances was generated using the function *TreeDist* in the TreeDistance (Smith 2022) and a neighbor-joining tree inferred with the *nj* function of ape (Paradis et al. 2004). Topological differences of the inferred eleven phylogenies were visualized using the *cophylo* function implemented in the *phytools* R package (Revell 2012), and a summary of the different groupings of endosymbionts has been plotted using the *ggtree* R package (Yu et al. 2017). The inferred topologies have been then linearized using the penalized likelihood approach implemented in *chronos* function from the *ape* R package (Paradis et al. 2004), using a relaxed model and lambda of 1; the latter has been chosen by testing values from 1 to 10 and selecting the one with the best likelihood score. Ancestral state reconstruction of the shifts to endosymbiosis was performed using a custom continuous-time Markov chain model implemented in the *ace* function of the *phytools* R package (Revell 2012) by leveraging a model allowing only transitions from free-living to endosymbiosis, due to the irreversible association with the host (Wernegreen 2015).

Inference of gene families duplication, transfer, and loss – Analyses on duplication, transfer, and loss (DTL) rates, representing the non-normalized probabilities of genes to either duplicate, transfer or get lost along each branch of the species tree; Morel et al. 2020) were performed on inferences 4 and 11 (one based on gene families and another based on the concatenation of single-copy orthologs) to assess the impact of phylogenetic uncertainty on the results, similarly to Kinjo et al. (2021). After correcting the root position to *Plesiomonas shigelloides*, species trees coming from analyses 4 (concatenation + all genes + mixture model + SR4 recoding) and 11 (gene family + all genes + SR4 recoding) were selected for gene tree correction with a maximum radius of 3 (--max-*spr-radius 3*) and reconciliation optimizing the DTL rates individually for each branch of the tree (*--per-species-rates*), using GeneRax (Morel et al. 2020). All phylogenetic tree branches were categorized as endosymbiosis-establishment, endosymbiosis-maintenance, or free-living; each endosymbiotic clade inferred was categorized as strict, mixed, or loose, depending on whether the species within a clade exhibited tight associations with their host, loose associations, or both. The per-branch DTL rates were separately plotted on the phylogeny using the *ggtree* R package, and statistical comparisons were performed using Spearman rank correlation in R. For each orthogroup (*i.e.,* the set of genes derived from a single gene in the last common ancestor of all the species under consideration; Emms and Kelly 2019) the number of genes lost, and the number of genes inherited from the ancestor for each species tree node were extracted from GeneRax results and used to calculate *per* branch and *per* family loss frequencies with the following formula: genes lost / (genes inherited from the ancestor − genes lost). Then, per family mean loss frequencies were calculated for all branches inferred as free-living, all branches inferred as endosymbiont, and separately for each of the eleven endosymbiont clades retrieved. Subsequent analyses considered only gene families that were present in at least one free-living ancestor of endosymbionts; this conservative approach excluded all families that originated after the establishment of endosymbiosis, allowing us to compare families whose evolutionary history of acquisition and loss occurred across similar timespans. A Spearman correlation matrix of the orthogroup mean loss frequencies between each independent endosymbiotic lineage and free-living species was calculated and plotted.

Orthogroup dN/dS rates in free-living and endosymbiont lineages – Inferred OGs were split into two groups containing free-living and symbiont species only. These OGs were then processed as previously described: they were aligned as proteins using mafft v.7 (Katoh and Standley 2013) and *retro*-translated into nucleotides. The ParGene pipeline was adopted to infer a gene family phylogenetic tree for each OG as described above. Then, each OG dN/dS ratio was inferred using *codeml* from the PAML package (v. 4.8; Yang 2007) for a single omega class (*m0*) and leveraging the gene tree. Subsequent analyses were restricted to OGs with dN/dS values < 1, because here the focus was on selective constraints (*i.e.,* the strength of purifying selection), hence we excluded any instance of positive selection. For each OG, the shift in selection regime between endosymbionts and free-living was obtained (Δ dN/dS = dN/dS in endosymbionts − dN/dS in free-living species). Spearman correlations between *per* branch and *per* family loss frequencies and dN/dS were calculated using R across all OGs including more than four species.

Orthogroups COG annotation – All the sequences included in each OG were annotated with Clusters of Orthologous Groups (COG) terms, using eggNOG-mapper v2 (Cantalapiedra et al. 2021). The most represented term among those assigned to the sequence included in each OG was assigned to it. Mean loss frequencies for all OGs assigned to each COG functional category were calculated for free-living and endosymbiont species. To identify the functional categories whose genes have been more consistently lost during the transition to endosymbiosis, each COG category was ranked based on its mean loss frequency for each independent endosymbiotic lineage (*i.e.,* considering all branches following an inferred shift of endosymbiosis as resulting from ancestral state reconstruction analyses) and free-living species. A word cloud plot was generated using R, with the word size proportional to the cumulative sum of all COG ranks across the twelve independent endosymbiosis establishment events. For each COG, the Δ mean dN/dS and Δ mean loss frequency between endosymbiont and free-living branches were calculated across all OGs associated with that specific COG.

CRediT authorship contribution statement

Giobbe Forni: Writing – review & editing, Writing – original draft, Visualization, Formal analysis, Conceptualization. **Jacopo Martelossi:** Writing – review & editing, Writing – original draft, Formal analysis, Conceptualization. **Benoit Morel:** Software, Formal analysis. **Dario Pistone:** Writing – review & editing. **Claudio Bandi:** Writing – review & editing, Supervision. **Matteo Montagna:** Writing – review & editing, Writing – original draft, Formal analysis, Supervision, Conceptualization.

Acknowledgments

The authors would like to thank: Michele Castelli and Davide Sassera for the initial discussion on the manuscript. Marco Gebiola, Giovanni Piccinini, Fabrizio Ghiselli, and Andrea Luchetti for the comments on the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2026.108532.

Data availability

A markdown describing all the code used for the paper is available at https://github.com/for-giobbe/enterobacterales and at https://github.com/MontagnaLab/enterobacterales. All intermediate files of the analyses are available at https://zenodo.org/records/11611738.

References

Adeolu, M., Alnajar, S., Naushad, S., Gupta, R.S., 2016. Genome-based phylogeny and taxonomy of the 'Enterobacteriales': proposal for Enterobacterales ord. nov. divided into the families Enterobacteriaceae, Erwiniaceae fam. nov., Pectobacteriaceae fam. nov., Yersiniaceae fam. nov., Hafniaceae fam. nov., Morganellaceae fam. nov., and Budviciaceae fam. Int. J. Syst. Evol. Microbiol. 66, 5575–5599.

Albalat, R., Cañestro, C., 2016. Evolution by gene loss. Nat. Rev. Genet. 17, 379–391.

Andersson, D.I., Hughes, D., 1996. Muller's ratchet decreases fitness of a DNA-based microbe. PNAS 93, 906–907.

Bennett, G.M., Kwak, Y., Maynard, R., 2024. Endosymbioses have shaped the evolution of biological diversity and complexity time and time again. Genome Biol. Evol., evae112

Bernstein, H., Byerly, H.C., Hopf, F.A., Michod, R.E., 1985. Genetic damage, mutation, and the evolution of sex. Science 229, 1277–1281.

Bolotin, E., Hershberg, R., 2016. Bacterial intra-species gene loss occurs in a largely clocklike manner mostly within a pool of less conserved and constrained genes. Sci. Rep. 6, 1–9.

Borowiec, M.L., 2016. AMAS: a fast tool for alignment manipulation and computing of summary statistics. PeerJ 4, e1660.

Boscaro, V., Kolisko, M., Felletti, M., Vannini, C., Lynn, D.H., Keeling, P.J., 2017. Parallel genome reduction in symbionts descended from closely related free-living bacteria. Nat. Ecol. Evol. 1, 1160–1167.

Bourguignon, T., Kinjo, Y., Villa-Martin, P., Coleman, N.V., Tang, Q., Arab, D.A., Wang, Z., Tokuda, G., Hongoh, Y., Ohkuma, M., Ho, S.Y.W., Pigolotti, S., Lo, N., 2020. Increased mutation rate is linked to genome reduction in prokaryotes. Curr. Biol. 30, 3848–3855.

Boyd, B.M., James, I., Johnson, K.P., Weiss, R.B., Bush, S.E., Clayton, D.H., Dale, C., 2024. Stochasticity, determinism, and contingency shape genome evolution of endosymbiotic bacteria. Nat. Commun. 15 (1), 4571.

Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., Huerta-Cepas, J., 2021. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol. Biol. Evol. 38, 5825–5829.

Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973.

Castelli, M., Nardi, T., Gammuto, L., Bellinzona, G., Sabaneyeva, E., Potekhin, A., Serra, V., Petroni, G., Sassera, D., 2024. Host association and intracellularity evolved multiple times independently in the Rickettsiales. Nat. Commun. 15 (1), 1093.

Dale, C., Wang, B., Moran, N., Ochman, H., 2003. Loss of DNA recombinational repair enzymes in the initial stages of genome degeneration. Mol. Biol. Evol. 20 (8), 1188–1194.

Douglas, A.E., 2020. Housing microbial symbionts: evolutionary origins and diversification of symbiotic organs in animals. Philos. Trans. R. Soc. B 375 (1808), 20190603.

Chong, R.A., Park, H., Moran, N.A., 2019. Genome evolution of the obligate endosymbiont Buchnera aphidicola. Mol. Biol. Evol. 36, 1481–1489.

Cornwallis, C.K., van't Padje, A., Ellers, J., Klein, M., Jackson, R., Kiers, E.T., West, S.A., Henry, L.M., 2023. Symbioses shape feeding niches and diversification across insects. Nat. Ecol. Evol. 7 (7), 1022–1044.

Darriba, D., Posada, D., Kozlov, A.M., Stamatakis, A., Morel, B., Flouri, T., 2020. ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. Mol. Biol. Evol. 37 (1), 291–294.

Degnan, P.H., Yu, Y., Sisneros, N., Wing, R.A., Moran, N.A., 2009. *Hamiltonella defensa*, genome evolution of protective bacterial endosymbiont from pathogenic ancestors. PNAS 106, 9063–9068.

Duron, O., Gottlieb, Y., 2020. Convergence of nutritional symbioses in obligate blood feeders. Trends Parasitol. 36, 816–825.

Emms, D.M., Kelly, S., 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20, 1–14.

Fukruksa, C., Yimthin, T., Suwannaroj, M., Muangpat, P., Tandhavanant, S., Thanwisai, A., Vitta, A., 2017. Isolation and identification of *Xenorhabdus* and *Photorhabdus* bacteria associated with entomopathogenic nematodes and their larvicidal activity against Aedes aegypti. Parasit. Vectors 10 (1), 440.

Gil, R., Sabater-Muñoz, B., Latorre, A., Silva, F.J., Moya, A., 2002. Extreme genome reduction in *Buchnera* spp.: toward the minimal genome needed for symbiotic life. PNAS 99, 4454–4458.

Hirsh, A.E., Fraser, H.B., 2001. Protein dispensability and rate of evolution. Nature 411, 1046–1049.

Hershberg, R., Tang, H., Petrov, D.A., 2007. Reduced selection leads to accelerated gene loss in *Shigella*. Genome Biol. 8, 1–11.

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., Vinh, L.S., 2018. UFBoot2: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35, 518–522.

Husník, F., Chrudimský, T., Hypša, V., 2011. Multiple origins of endosymbiosis within the Enterobacteriaceae (γ-Proteobacteria): convergence of complex phylogenetic approaches. BMC Biol. 9, 1–14.

Husník, F., Keeling, P.J., 2019. The fate of obligate endosymbionts: reduction, integration, or extinction. Curr. Opin. Genet. Dev. 58, 1–8.

Jordan, I.K., Rogozin, I.B., Wolf, Y.I., Koonin, E.V., 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. 12, 962–968.

Kaiwa, N., Hosokawa, T., Nikoh, N., Tanahashi, M., Moriyama, M., Meng, X.Y., Maeda, T., Yamaguchi, K., Shigenobu, S., Ito, M., Fukatsu, T., 2014. Symbiont-supplemented maternal investment underpinning host's ecological adaptation. Curr. Biol. 24, 2465–2470.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K., von Haeseler, A., Jermiin, L.S., 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14, 587–589.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.

Kondrashov, A.S., 1998. Deleterious mutations and the evolution of sexual reproduction. Nature 336, 435–440.

Koga, R., Meng, X.Y., Tsuchida, T., Fukatsu, T., 2012. Cellular mechanism for selective vertical transmission of an obligate insect symbiont at the bacteriocyte–embryo interface. PNAS 109, E1230–E1237.

Korona, R., 2011. Gene dispensability. Curr. Opin. Biotechnol. 22, 547–551.

Kozlov, A.M., Darriba, D., Flouri, T., Morel, B., Stamatakis, A., 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 35, 4453–4455.

Kinjo, Y., Lo, N., Martín, P.V., Tokuda, G., Pigolotti, S., Bourguignon, T., 2021. Enhanced mutation rate, relaxed selection, and the "domino effect" are associated with gene loss in *Blattabacterium*, a cockroach endosymbiont. Mol. Biol. Evol. 38, 3820–3831.

Koonin, E.V., 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat. Rev. Microbiol. 1, 127–136.

Latorre, A., Manzano-Marín, A., 2017. Dissecting genome reduction and trait loss in insect endosymbionts. Ann. N.Y. Acad. Sci. 1389, 52–75.

Lerat, E., Daubin, V., Ochman, H., Moran, N.A., 2005. Evolutionary origins of genomic repertoires in bacteria. PLoS Biol. 3 (5), e130.

Manzano-Marin, A., Oceguera-Figueroa, A., Latorre, A., Jimenez-Garcia, L.F., Moya, A., 2015. Solving a bloody mess: B-vitamin independent metabolic convergence among gammaproteobacterial obligate endosymbionts from blood-feeding arthropods and the leech Haementeria officinalis. Genome Biol. Evol. 7, 2871–2884.

Martinson, V.G., Gawryluk, R.M., Gowen, B.E., Curtis, C.I., Jaenike, J., Perlman, S.J., 2020. Multiple origins of obligate nematode and insect symbionts by a clade of bacteria closely related to plant pathogens. PNAS 117, 31979–31986.

Mendonça, A.G., Alves, R.J., Pereira-Leal, J.B., 2011. Loss of genetic redundancy in reductive genome evolution. PLoS Comput. Biol. 7 (2), e1001082.

McCutcheon, J.P., Moran, N.A., 2012. Extreme genome reduction in symbiotic bacteria. Nat. Rev. Microbiol. 10, 13–26.

McCutcheon, J.P., Boyd, B.M., Dale, C., 2019. The life of an insect endosymbiont from the cradle to the grave. Curr. Biol. 29, R485–R495.

Moya, A., Peretó, J., Gil, R., Latorre, A., 2008. Learning how to live together: genomic insights into prokaryote-animal symbioses. Nat. Rev. Genet. 9, 218–229.

Moran, N.A., 2021. Microbe profile: *Buchnera aphidicola*: ancient aphid accomplice and endosymbiont exemplar. Microbiology 167, 001127.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Von Haeseler, A., Lanfear, R., 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37, 1530–1534.

Maire, J., Girvan, S.K., Barkla, S.E., Perez-Gonzalez, A., Suggett, D.J., Blackall, L.L., van Oppen, M.J., 2021. Intracellular bacteria are common and taxonomically diverse in cultured and in hospite algal endosymbionts of coral reefs. ISME J. 15, 2028–2042.

Margulis, L. 1981. Symbiosis in cell evolution: Life and its environment on the early earth.

Moran, N.A., Bennett, G.M., 2014. The tiniest tiny genomes. Annu. Rev. Microbiol. 68, 195–215.

Morel, B., Kozlov, A.M., Stamatakis, A., 2019. ParGenes: a tool for massively parallel model selection and phylogenetic tree inference on thousands of genes. Bioinformatics 35, 1771–1773.

Morel, B., Kozlov, A.M., Stamatakis, A., Szöllősi, G.J., 2020. GeneRax: a tool for species-tree-aware maximum likelihood-based gene family tree inference under gene duplication, transfer, and loss. Mol. Biol. Evol. 37, 2763–2774.

Morel, B., Schade, P., Lutteropp, S., Williams, T.A., Szöllősi, G.J., Stamatakis, A., 2022. SpeciesRax: a tool for maximum likelihood species tree inference from gene family trees under duplication, transfer, and loss. Mol. Biol. Evol. 39, msab365.

Muñoz-Gómez, S.A., Kreutz, M., Hess, S., 2021. A microbial eukaryote with a unique combination of purple bacteria and green algae as endosymbionts. Sci. Adv. 7, eabg4102.

Muller, H.J., 1964. The relation of recombination to mutational advance. Mutat. Res. 1, 2–9.

Naser-Khdour, S., Minh, B.Q., Zhang, W., Stone, E.A., Lanfear, R., 2019. The prevalence and impact of model violations in phylogenetic analysis. Genome Biol. Evol. 11, 3341–3352.

Nikoh, N., Hosokawa, T., Moriyama, M., Oshima, K., Hattori, M., Fukatsu, T., 2014. Evolutionary origin of insect–Wolbachia nutritional mutualism. PNAS 111, 10257–10262.

Ochman, H., Moran, N.A., 2001. Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. Science 292, 1096–1099.

O'Fallon, B., 2008. Population structure, levels of selection, and the evolution of intracellular symbionts. Evolution 62, 361–373.

Otero-Bravo, A., Goffredi, S., Sabree, Z.L., 2018. Cladogenesis and genomic streamlining in extracellular endosymbionts of tropical stink bugs. Genome Biol. Evol. 10, 680–693.

Paradis, E., Claude, J., Strimmer, K., 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20, 289–290.

Pérez-Brocal, V., Gil, R., Ramos, S., Lamelas, A., Postigo, M., Michelena, J.M., Silva, F.J., Moya, A., Latorre, A., 2006. A small microbial genome: the end of a long symbiotic relationship? Science 314, 312–313.

Puigbò, P., Lobkovsky, A.E., Kristensen, D.M., Wolf, Y.I., Koonin, E.V., 2014. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. BMC Biol. 12, 1–19.

Raval, P.K., Garg, S.G., Gould, S.B., 2022. Endosymbiotic selective pressure at the origin of eukaryotic cell biology. eLife 11, e81033.
Redmond, A.K., McLysaght, A., 2021. Evidence for sponges as sister to all other animals from partitioned phylogenomics with mixture models and recoding. Nat. Commun. 12, 1–14.
Revell, L.J., 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. 2, 217–223.
Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16, 276–277.
Sabater-Muñoz, B., Toft, C., Alvarez-Ponce, D., Fares, M.A., 2017. Chance and necessity in the genome evolution of endosymbiotic bacteria of insects. ISME J. 11, 1291–1304.
Sapp, J., 1994. Evolution by association: a history of symbiosis. Oxford University Press, New York.
Salem, H., Bauer, E., Kirsch, R., Berasategui, A., Cripps, M., Weiss, B., Koga, R., Fukumori, K., Vogel, H., Fukatsu, T., Kaltenpoth, M., 2017. Drastic genome reduction in an herbivore's pectinolytic symbiont. Cell 171, 1520–1531.
Smith, M.R., 2022. Robust analysis of phylogenetic tree space. Syst. Biol. 71 (5), 1255–1270.
Soutar, C.D., Stavrinides, J., 2020. Phylogenetic analysis supporting the taxonomic revision of eight genera within the bacterial order Enterobacterales. Int. J. Syst. Evol. Microbiol. 70, 6524–6530.
Treangen, T.J., Rocha, E.P., 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet. 7, e1001284.
Vorburger, C., Gehrer, L., Rodriguez, P., 2010. A strain of the bacterial symbiont Regiella insecticola protects aphids against parasitoids. Biol. Lett. 6, 109–111.
Weinert, L.A., Welch, J.J., 2017. Why might bacterial pathogens have small genomes? Trends Ecol. Evol. 32, 936–947.
Wernegreen, J.J., 2015. Endosymbiont evolution: predictions from theory and surprises from genomes. Ann. N. Y. Acad. Sci. 1360, 16–35.
Williams, T.J., Allen, M.A., Ivanova, N., Huntemann, M., Haque, S., Hancock, A.M., Brazendale, S., Cavicchioli, R., 2021. Genome analysis of a verrucomicrobial endosymbiont with a tiny genome discovered in an Antarctic lake. Front. Microbiol. 12, 674758.
Wilson, A.C., Duncan, R.P., 2015. Signatures of host/symbiont genome coevolution in insect nutritional endosymbioses. PNAS 112, 10255–10261.
Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591.
Yu, G., Smith, D.K., Zhu, H., Guan, Y., Lam, T.T.Y., 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. 8, 28–36.
Zhou, X., Lutteropp, S., Czech, L., Stamatakis, A., Looz, M.V., Rokas, A., 2020. Quartet-based computations of internode certainty provide robust measures of phylogenetic incongruence. Syst. Biol. 69 (2), 308–324.