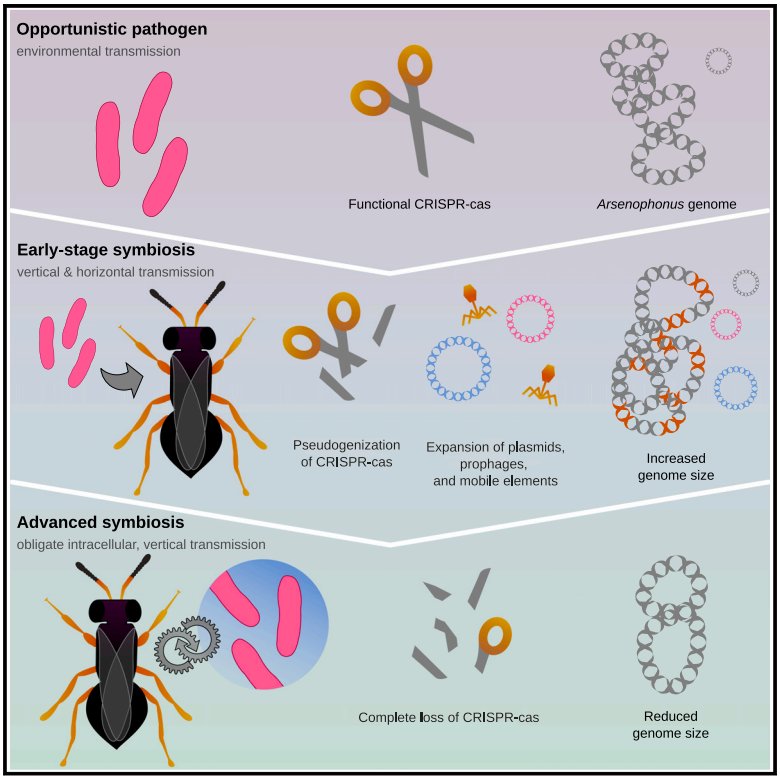
Article

Current Biology

Genome dynamics across the evolutionary transition to endosymbiosis

Graphical abstract



Highlights

- Genome expansion is observed after bacteria initially evolve vertical transmission
- Expansion is driven by phage/plasmid accumulation after loss of CRISPR-Cas defense
- The invasion of prophages is associated with the enrichment of T3SS effectors
- Genome expansion fuels initial adaptation to symbiosis and precedes genome reduction

Authors

Stefanos Siozios, Pol Nadal-Jimenez, Tal Azagi, ..., Eva Novakova, Alistair C. Darby, Gregory D.D. Hurst

Correspondence

stefanos.siozios@zoologie.uni-halle.de (S.S.), g.hurst@liverpool.ac.uk (G.D.D.H.)

In brief

Many insects carry heritable bacterial symbionts, but the evolution of vertical (parent-to-offspring) transmission is poorly understood. Using the genus *Arsenophonus*, Siozios et al. show that microbes that have recently adopted vertical transmission have expanded genomes and gene content, driven by the loss of phage defenses.

Siozios et al., 2024, Current Biology *34*, 1–12
December 16, 2024
https://doi.org/10.1016/j.cub.2024.10.044

CellPress

Current Biology

CellPress
OPEN ACCESS

Article

Genome dynamics across the evolutionary transition to endosymbiosis

Stefanos Siozios,[1,12,13,*] Pol Nadal-Jimenez,[1] Tal Azagi,[2] Hein Sprong,[2] Crystal L. Frost,[1] Steven R. Parratt,[1] Graeme Taylor,[3] Laura Brettell,[4] Kwee Chin Liew,[5] Larry Croft,[6] Kayla C. King,[1,7,8,9] Michael A. Brockhurst,[1,10] Václav Hypša,[11] Eva Novakova,[11] Alistair C. Darby,[1] and Gregory D.D. Hurst[1,14,*]

[1]Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Liverpool L59 7ZB, UK
[2]Centre for Infectious Diseases Research, National Institute for Public Health and the Environment, 3720 BA Bilthoven, the Netherlands
[3]Department of Biology, University of Victoria, Victoria, BC V8P 5C2, Canada
[4]School of Science, Engineering and Environment, University of Salford, Manchester M5 4WT, UK
[5]NSW Health Pathology Infectious Diseases Department, Wollongong Hospital, Wollongong, NSW, Australia
[6]School of Medicine, Deakin University, 75 Pigdons Road, Waurn Ponds, VIC 3216, Australia
[7]Department of Biology, University of Oxford, 11a Mansfield Road, Oxford OX1 3SZ, UK
[8]Department of Zoology, University of British Columbia, 6270 University Boulevard, Vancouver, BC V6T 1Z4, Canada
[9]Department of Microbiology & Immunology, University of British Columbia, 1365 - 2350 Health Sciences Mall, Vancouver, BC V6T 1Z3, Canada
[10]Division of Evolution, Infection and Genomics, Faculty of Biology, Medicine and Health, University of Manchester, Michael Smith Building, Dover Street, Manchester M13 9PT, UK
[11]Department of Parasitology, Faculty of Science, University of South Bohemia, Branišovská 1645/31a, 370 05 České Budějovice, Czech Republic
[12]Present address: German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Puschstraße 4, 04103 Leipzig, Germany
[13]Present address: Martin-Luther-Universität Halle-Wittenberg, Hoher Weg 8, 06120 Halle (Saale), Germany
[14]Lead contact
*Correspondence: stefanos.siozios@zoologie.uni-halle.de (S.S.), g.hurst@liverpool.ac.uk (G.D.D.H.)
https://doi.org/10.1016/j.cub.2024.10.044

SUMMARY

Endosymbiosis—where a microbe lives and replicates within a host—is an important contributor to organismal function that has accelerated evolutionary innovations and catalyzed the evolution of complex life. The evolutionary processes associated with transitions to endosymbiosis, however, are poorly understood. Here, we leverage the wide diversity of host-associated lifestyles of the genus *Arsenophonus* to reveal the complex evolutionary processes that occur during the transition to a vertically transmitted endosymbiotic lifestyle from strains maintained solely by horizontal (infectious) transmission. We compared the genomes of 38 strains spanning diverse lifestyles from horizontally transmitted pathogens to obligate interdependent endosymbionts. Among culturable strains, we observed those with vertical transmission had larger genome sizes than closely related horizontally transmitting counterparts, consistent with evolutionary innovation and the rapid gain of new functions. Increased genome size was a consequence of prophage and plasmid acquisition, including a cargo of type III effectors, alongside the concomitant loss of CRISPR-Cas genome defense systems, enabling mobile genetic element expansion. Persistent endosymbiosis was also associated with loss of type VI secretion, which we hypothesize to be a consequence of reduced microbe-microbe competition. Thereafter, the transition to endosymbiosis with strict vertical inheritance was associated with the expected relaxation of purifying selection, gene pseudogenization, metabolic degradation, and genome reduction. We argue that reduced phage predation in endosymbiotic niches drives the loss of genome defense systems driving rapid genome expansion upon the adoption of endosymbiosis and vertical transmission. This remodeling enables rapid horizontal gene transfer-mediated evolutionary innovation and precedes the reductive evolution traditionally associated with adaptation to endosymbiosis.

INTRODUCTION

Animals live in a microbial world. Their interactions with microbes range from antagonism, involving pathogenic symbionts, to mutualism, involving beneficial symbionts, with evolutionary transitions occurring commonly between these states[1,2]. Transitions in symbiotic interactions can further select for the evolution of key symbiotic traits, such as vertical transmission and the eventual integration of symbionts into host anatomy and physiology.[3] Vertical transmission relaxes selection for traits necessary for external survival while also enhancing partner fidelity[1] correlating microbe transmission with host fitness and

Please cite this article in press as: Siozios et al., Genome dynamics across the evolutionary transition to endosymbiosis, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.10.044

favoring beneficial function(s). Concurrently, population bottlenecks associated with vertical transmission limit within-host symbiont diversity, selecting for lower virulence.[4]

In endosymbiosis, symbionts live within the body or cells of the host organism. Herein, multiple transitions from horizontal to vertical endosymbiont transmission have occurred across the tree of microbial diversity, including within microeukaryotic, fungal, plant, and animal hosts.[2] We have a well-developed understanding of the evolutionary processes that impact established vertically-transmitting endosymbionts. Clonal population bottlenecks associated with vertical transmission intensify genetic drift,[5] thus reducing the efficiency of purifying selection for function. Additionally, symbionts may lose components of DNA repair systems that accelerate mutation accumulation. These processes, which are also observed for obligate extracellular gut symbionts with host-enabled vertical transmission,[6] act in concert with a deletion bias in mutation[7] and collectively drive genome degradation through pseudogenization, genome reduction, loss of repair systems, and lowered %GC.[8] The evolutionary trajectory leads, ultimately, to obligate symbionts with highly reduced genomes with a low %GC.[3]

Less well-understood are the evolutionary processes occurring in microbial endosymbionts at the origins of host association.[3] Recently, synthetic biology and experimental evolution have been deployed to identify the genes and systems enabling a microbe to establish symbiosis.[9,10] However, these studies have examined a particular case where the symbiosis is established in a host with an existing unmet requirement for a symbiont. In these cases, the host has pre-existing physiological, anatomical, and metabolic adaptations to carry and transmit symbionts onward. These studies represent valuable case studies that reflect the evolutionary and genetic processes that occur during symbiont replacement, a common occurrence, but do not address the processes that occur when a host does not have a requirement for a symbiont. In these cases, we have relied on comparative genomics, where differences between the genomes of related microbes with different patterns of transmission between hosts provide insights into the changes that occur during the transition to vertical transmission.[3] Historically, the genetic distance between many well-studied symbionts and their free-living ancestors has inhibited our capacity to understand the evolution of microbes at incipient establishment of symbiosis. More recently, comparative genomics in the clades *Sodalis* and *Serratia*, and in the *Pantoea*/stinkbug symbiosis, have been anchored to strains and species that are not vertically transmitted, enabling less coarse-grained comparisons.[11–13] Nevertheless, inference remains limited by the paucity of horizontally transmitting strains.

To gain a more precise view of the tempo and mode of evolution during the transition to persistent endosymbiosis, we leveraged the wide diversity of host-associated lifestyles in the gammaproteobacterial genus *Arsenophonus*.[14] *Arsenophonus* is found widely in parasitic and social Hymenoptera, in both blood- and plant-sucking hemipteran insects, as well as Lepidoptera. Within the *Arsenophonus* clade, there are horizontally transmitting pathogenic strains, extracellular endosymbionts with mixed modes of transmission, intracellular facultative endosymbionts with vertical transmission, and obligate vertically transmitted endosymbionts where the partners are co-dependent.[15–19] Our data reveal a new, dynamic phase of genome remodeling that occur on the initial establishment to vertical transmission and we argue the gain of function that occurs in this phase is key to establishing a vertically transmitted symbiotic lifestyle.

RESULTS AND DISCUSSION

Genome completion and estimation of phylogenetic relationships in the genus *Arsenophonus* reveals three major clades

To gain a high-resolution view of the evolutionary transition to endosymbiosis in an insect-associated bacterial endosymbiont, we first completed closed genomes for seven new strains: three *Arsenophonus nasoniae* from *Nasonia vitripennis*, one from each of the parasitic wasps *Pachycrepoideus vindemmiae* and *Ixodiphagus hookeri*, one from the blue butterfly *Polyommatus bellargus*, and a strain of *Arsenophonus apicola* isolated from Australian *Apis mellifera*. We also completed a draft genome for *Ca. A. triatominarum*. In addition, novel draft genomes for a further 17 *Arsenophonus* strains were assembled from Sequence Read Archive (SRA) deposits from a variety of insect genome sequencing projects. Genome assembly data, alongside sample details and current understanding of transmission mode and nature of symbiosis, are given in Table S1.

We then estimated the phylogenetic affiliation of the strains through core genome analysis (Figure 1), maximizing the common gene set for phylogenetic inference by excluding strains with highly reduced genomes. This analysis revealed three main clades according to lifestyle: the *nasoniae* clade, where characterized members have mixed modes of transmission or have recently become facultative vertically transmitted symbionts; the *apicola* clade, where characterized strains are horizontally transmitted; and the *triatominarum* clade, where characterized members are vertically transmitted. Importantly, strains with different transmission modes are often closely related. For instance, strains with horizontal transmission, endosymbionts with mixed modes of transmission, and strains that are intracellular facultative endosymbionts with vertical transmission show complete or near complete amino acid sequence identity at housekeeping genes. The recency of vertical transmission emergence in the clade is also evidenced by the retained *in vitro* culturability of many of these strains.[17,18] When obligate co-dependent symbionts are included in the phylogeny (Figure S1), they do not form a monophyletic clade as was previously shown,[20] indicating independent evolutionary transitions to obligate host dependence. Culturable representatives are common in the *apicola*/*nasoniae* clades, likely reflecting a current or recent requirement to live outside the endosymbiotic environment. Members of the bee genus *Bombus* have members from both *apicola* and *nasoniae* clades, indicating recurrent colonization of this host group by the endosymbiont.

Strains that have recently evolved vertical transmission have larger genomes than strains that transmit through the environment

Arsenophonus strains varied in genome size from 663,125 to 5,080,918 bp. It is generally considered that vertically transmitted symbiont genome size is an inverse function of the dependence of the host on the symbiont,[21] and that gene loss is observed rapidly following transition to vertical transmission.[3]

Please cite this article in press as: Siozios et al., Genome dynamics across the evolutionary transition to endosymbiosis, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.10.044

CellPress
OPEN ACCESS

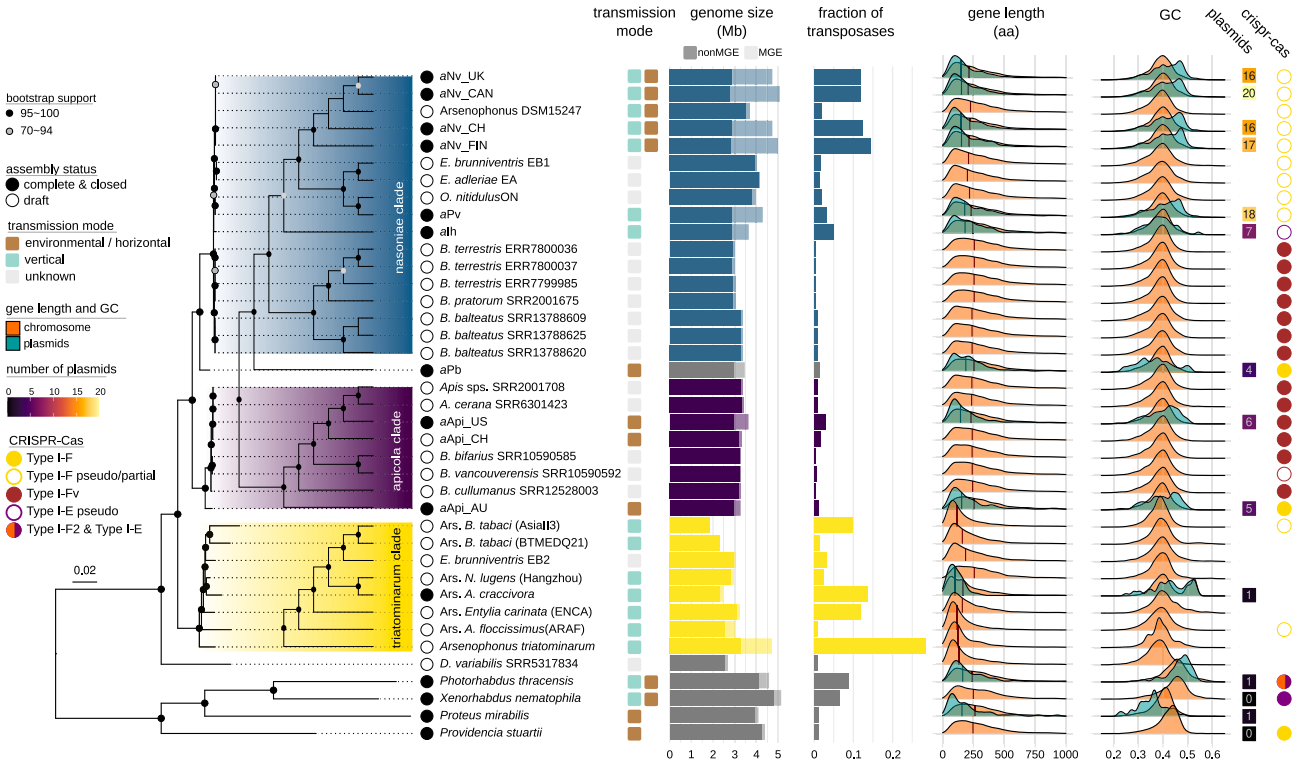


Figure 1. Core genome phylogeny and genome features of the *Arsenophonus* clade

The phylogenetic relationships between *Arsenophonus* strains were inferred using maximum likelihood on the concatenated set of 230 single-copy core protein sequences in IQ-TREE v2.1.4 under the JTTDCMut+F+R3 model. Only bootstrap support values ≥70 are shown. Inset cladograms are used to improve tree readability. The nasoniae, apicola, and triatominarum clades are highlighted in blue, magenta, and yellow respectively, and complete genomes are indicated by black circles in front of the tip labels. The transmission mode of each strain is indicated with the colored squares (green: vertical, brown: horizontal, and gray: currently unknown). The left horizontal bar plot shows the genome size in Mb with the light shading representing the fraction of the genome corresponding to mobile genetic elements (MGEs), including phages and plasmids. The right horizontal bar plot shows the fraction of transposases (insertion sequence elements) in each genome (for incomplete genome assemblies, these numbers may be an underestimate and should be treated with caution). The ridgeline plots show the distribution of CDS length (in amino acids) and GC content fraction for genes of chromosomal (orange) or extrachromosomal (green) origin. The vertical lines in the gene length ridge plot represent median values. The heatmap shows the number of plasmids for complete and closed genomes, whereas the colored circles indicate the type and intactness of the CRISPR-Cas system (filled circle: intact, non-filled circle: pseudogenized, no circle: not present). A Bayesian phylogenetic analysis including the four obligate *Arsenophonus* genomes (*Arsenophonus* of *Lipoptena fortisetosa*, *Arsenophonus* of *Aleurodicus dispersus*, *Arsenophonus* of *Ceratovacuna japonica*, and *Arsenophonus* of *Melophagus ovinus*) and *Ca.* Riesia is shown in the Figure S1. The scale bar represents substitutions per site. See also Figures S1, S5, S6, and Table S1.

As expected, the smallest genomes in the clade were indeed from interdependent obligate endosymbionts and members of the *triatominarum* clade, where members are vertically transmitted and most require a host for replication (Figure 2A). Unexpectedly, the five strains with the largest closed genomes were not those with horizontal transmission but related strains with either mixed modes of transmission or vertical transmission. This pattern reflected a larger genome size in the nasoniae clade (where characterized members have protracted host association and vertical transmission and genome size ranges from 3.65–5.1 Mb) compared with the apicola clade (where characterized members are environmentally acquired pathogens and genome size is 3.27–3.63 Mb) (Figure 2A). Examining strains with both closed genomes and known transmission mode, the cluster of nasoniae clade strains with vertical transmission have significantly larger genomes than the strains from honey bees and the strain from a butterfly that do not have vertical transmission (Mann Whitney U = 0; $n1 = 3$, $n2 = 6$; $p < 0.05$).

Ancestral reconstruction of gene content reveals a notable increase in gene family acquisitions within the nasoniae clade, particularly pronounced in the wasp subclade, which is characterized by a mixed mode or vertical transmission, in contrast to the horizontally transmitted strains (Figure 3). This observation suggests that the expansion of genome size within the nasoniae clade relative to the apicola clade is predominantly due to gene gains in the former rather than gene losses in the latter. Consistent with expectations, gene loss manifests as a prominent trend within the triatominarum clade and among clades containing obligate strains.

Genome size increase is associated with mobile genetic element accumulation and acquisition of type III effectors

The increase in genome size in early-stage endosymbionts is largely driven by an increase in mobile genetic elements, notably prophage and plasmid content, in strains that have recently

Please cite this article in press as: Siozios et al., Genome dynamics across the evolutionary transition to endosymbiosis, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.10.044

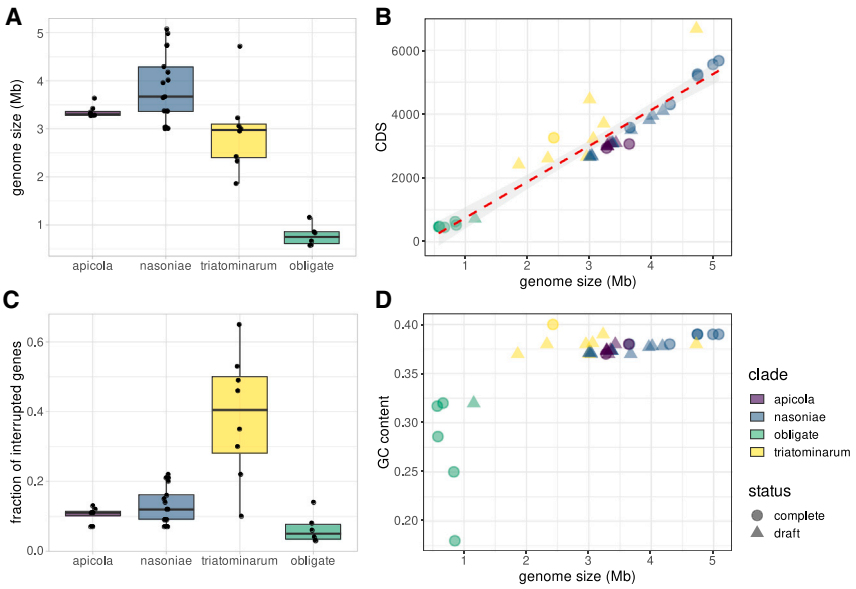


Figure 2. Genomic characteristics of the *Arsenophonus*/Riesia clades

(A) Genome size distribution across *Arsenophonus* clades.

(B) Association between genome size and the number of coding sequences (CDS). The red dashed line in (B) represents a fitted linear trend line with confidence intervals shown as gray shading.

(C) The fraction of interrupted genes across *Arsenophonus* clades. This metric was estimated by calculating the fraction of proteins with length <80% of the length of their top hit in the Swiss-Prot database.

(D) Association between genome size and the fraction of GC content. Although the number of predicted protein-coding genes shows, as expected, a linear relationship with the genome size across the *Arsenophonus*/Riesia clades, this association does not hold for GC content contrary to the classical observations between free-living and symbiotic microbes. Boxplots: center line, median; box limits, 25th and 75th percentiles; whiskers, ±1.5× interquartile range; data points are shown with the black dots.

See also Figures S7–S9, and Table S1.

entered endosymbiosis and lost environmental transmission (Figure 1). Some of these elements are shared between different *Arsenophonus* clades, suggesting that they were present in their last common ancestor. However, our results indicate an independent adoption of plasmids and phages by different clades and sub-clades or even strains (Figure 4). This expansion was greatly enhanced in the nasoniae subclade, which infects different wasp species. Notably, most of the *Arsenophonus* mobile genetic elements showed little similarity to elements previously identified in other taxa, with some exceptions such as the *Acyrthosiphon pisum* secondary endosymbiont (APSE) bacteriophage from the aphid symbiont *Hamiltonella defensa*,

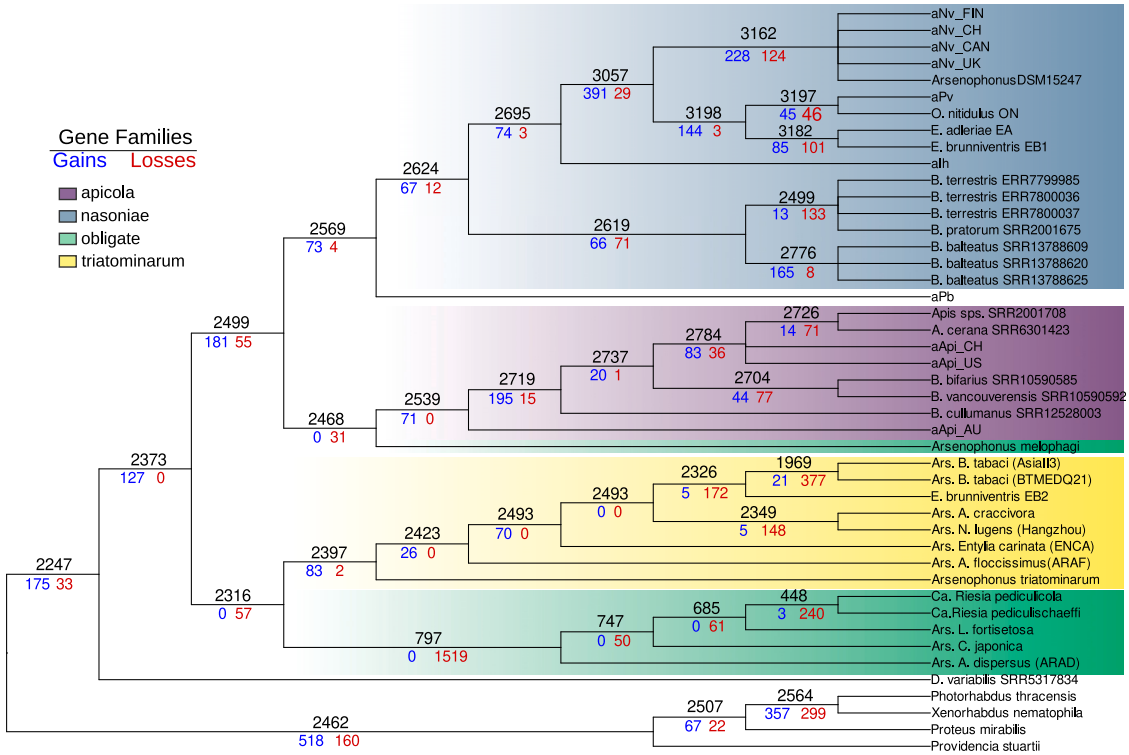


Figure 3. Ancestral reconstruction of gene content across the *Arsenophonus* phylogeny

The ancestral reconstruction was performed using Count v10.04 (Csűrös)[22] on the gain-loss-duplication model based on Bayesian inference shown in Figure S1. The estimated number of gene families (black), gene gains (blue), and gene losses (red) are shown at the ancestral nodes in the phylogeny.

Please cite this article in press as: Siozios et al., Genome dynamics across the evolutionary transition to endosymbiosis, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.10.044

CellPress
OPEN ACCESS

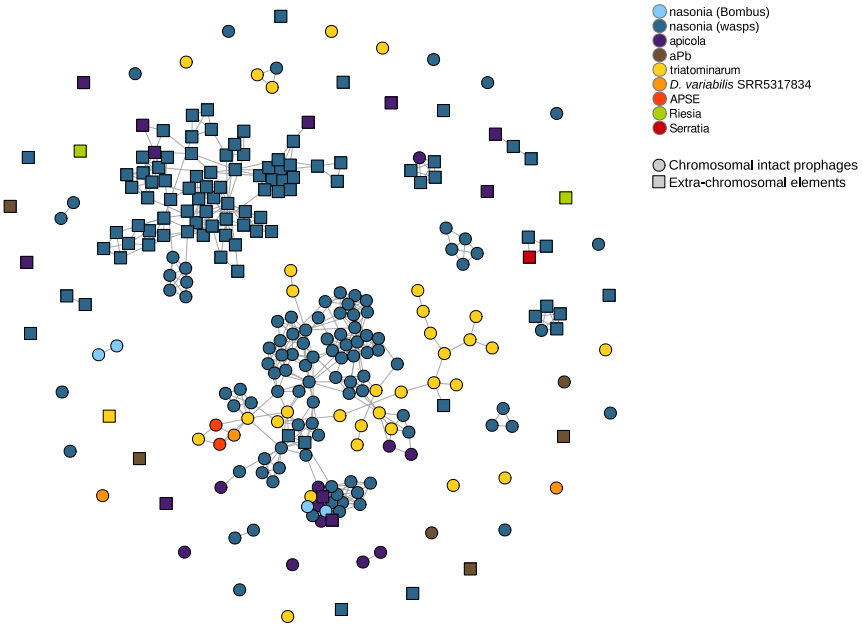


Figure 4. Relatedness between *Arsenophonus* mobile genetic elements

The relationships between all *Arsenophonus* mobile genetic elements (complete chromosomal phages and extrachromosomal elements) and previously described phages and plasmids were assessed by computing the average nucleotide identity (ANI) between all sequence pairs using FastANI v1.33. A kmer size of 16, fragment length of 1,000, and minimum fraction of shorter genome coverage of 50% were used for the analysis. Results were considered significant if at least 50% of the query fragment was shared with the reference sequence. Edges represent ANI values $\geq$ 80%. Different *Arsenophonus* clades and taxonomic groups are represented by different colors. Chromosomal prophages and extrachromosomal elements are shown as filled circles and squares, respectively. The network graph was generated using the igraph package in R.
See also Table S4.

as previously reported.[23] Consistent with the pattern of genome expansion associated with prophage, a large (4.9 Mb) high-quality draft genome of *Arsenophonus* from the triatominarum group has been recently recovered from the delphacid bug *Peregrinus maidis*,[24] and it has a very high mobile element representation (38% prophage and 19.5% IS elements).

It is notable that this increase in genome size following the adoption of an endosymbiotic lifestyle can be found more widely. For instance, *Spiroplasma poulsonii*, the male-killing symbiont of *D. melanogaster*, has the largest recorded genome size in the genus,[25] and is notably bigger than those of closely related strains that are environmentally acquired or vectored. Likewise, culturable strains of symbiotic *Pantoea* found in stink bugs have larger genome than purely environmental relatives.[13] Ancestral genome reconstruction also indicated an increase in genome size in early host-adapted Legionellaceae compared with their last common free-living ancestors.[26] In an allied study of a eukaryote parasite group, vertically transmitted microsporidia had enlarged genomes compared with horizontally transmitted ones, due to the accumulation of transposable elements.[27] Thus, this pattern does not appear to be unique to the *Arsenophonus* clade but it is a commonly observed process that occurs following the evolution of vertical transmission.

Acquisition of mobile genetic elements, like prophages and plasmids in bacteria, is often accompanied by horizontal gene transfer of accessory genes that are important in microbial virulence and adaptation.[28,29] Consistent with this pattern, we observed a gain in type III secretion (T3SS) associated effectors in nasoniae group strains that have recently become endosymbionts, compared with environmentally acquired strains (Figure 5). These data support previous work in *Sodalis*, arguing that symbionts repurpose the T3SS on transition from pathogenesis to enable persistent intracellular host association.[30] Further supporting this hypothesis, our data indicate that the process may, in some cases, involve acquisition of new effectors. Contrastingly, T3SS are either absent or heavily pseudogenized in the triatominarum clade, where vertical transmission is well established. The loss of T3SS systems in highly derived vertically inherited endosymbiont genomes suggests that these traits become redundant or costly once the host and endosymbiont have become tightly coevolved.

Genome expansion through mobile elements is correlated with loss of CRISPR-Cas and other defense systems

We next investigated potential drivers of prophage and plasmid accumulation. Past analyses across broad microbial diversity has indicated that genome defense systems are less commonly found in symbiotic microbes than in free-living ones.[31] For instance, CRISPR-Cas systems, which protect the genome from mobile DNA, are more commonly observed in non-symbionts. We observed that the identity, distribution, and completeness of CRISPR-Cas genome defense systems vary extensively among *Arsenophonus* genomes (Figure 6). We identified three types of CRISPR-Cas systems (types I-F, I-Fv and I-E) with variable gene content across all strains, suggesting multiple gain and loss processes of these important genome defense systems (Figure 6A). This turnover is likely mediated by horizontal gene transfer and recombination (Figure S2), consistent with mobile genetic element-mediated selection for genome defense varying between strains, likely according to exposure in their local environment. Variable mobile element exposure is further supported by the distinct lack of spacer matches between different *Arsenophonus* clades (Figure 6B). Furthermore, our data show that spacers from one clade are more likely to have a target against a mobile genetic element found within the same clade, with few exceptions, reflecting the exposure of these clades to different mobile genetic element communities (Figure S3).

Intact CRISPR-Cas systems were found in horizontally transmitting strains and strains of unknown transmission

Please cite this article in press as: Siozios et al., Genome dynamics across the evolutionary transition to endosymbiosis, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.10.044

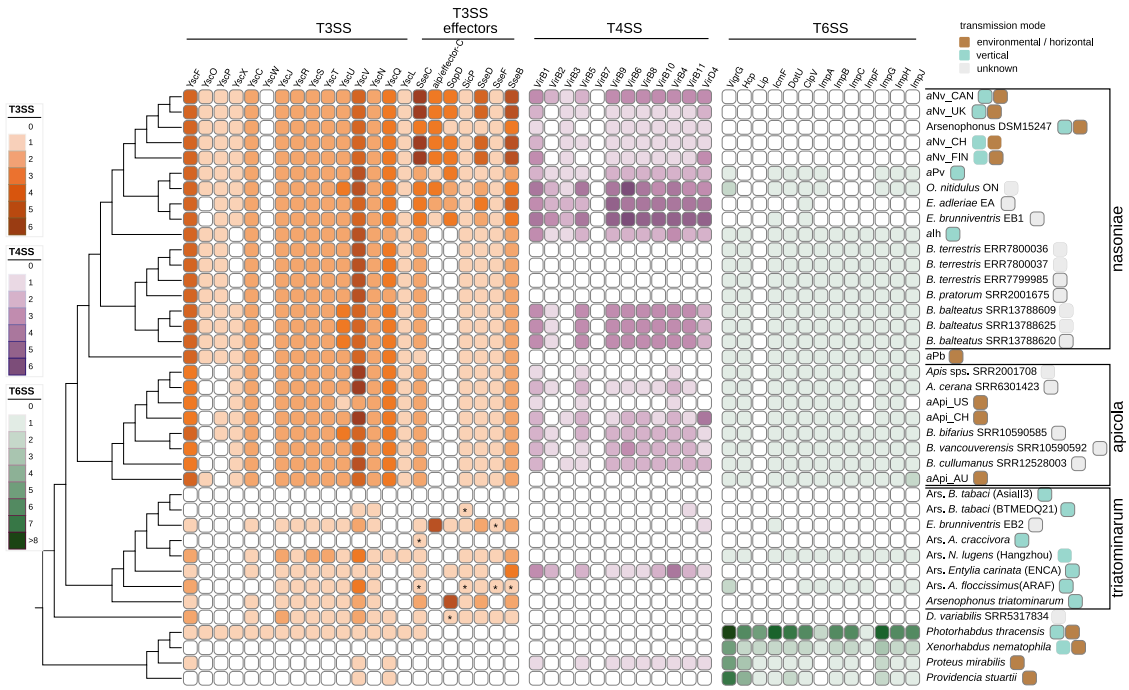


Figure 5. Comparative analysis of secretion systems across the *Arsenophonus* clades

Core components of the type III (orange), type IV (magenta), and type VI (green) systems as predicted by BlastKOALA are shown. Absence of genes is indicated by empty squares. Identified type III effectors are also shown. The relationship of the *Arsenophonus* strains is shown with the cladogram based on the core genome phylogeny (Figure 1). The asterisks indicate potential pseudogenes. Not shown in the figure: secretion systems are absent in the obligate Arsenophonus lineages (*Arsenophonus* of *Lipoptena fortisetosa*, *Arsenophonus* of *Aleurodicus dispersus*, *Arsenophonus* of *Ceratovacuna japonica*, and *Arsenophonus* of *Melophagus ovinus*) and *Ca.* Riesia.

mode. Notably, vertically inherited strains closely related to these contained recently pseudogenized CRISPR-Cas systems (evidenced by intact and shorter CRISPR arrays but insertions or frameshifts within the Cas systems), whereas obligate intracellular symbionts carried either highly degraded fragmentary CRISPR-Cas systems or none (Figures 1 and 6A). Similar to closely related members of the triatominarum clade, the recently published prophage-rich genome of *Arsenophonus* from *Peregrinus maidis*[24] carries a fragmentary CRISPR-Cas system (with similarities to type 1-Fv system), reflecting an independent event where prophage-mediated genome expansion is associated with the recent loss of CRISPR-Cas functionality.

CRISPR-Cas is known to have metabolic and autoreactive costs[32]. Theoretical and experimental data suggest that rapid loss of CRISPR-Cas function is an adaptation to maintain horizontally transferred genetic elements that are beneficial for microbial adaptation[33,34]. Thus, the extensive pseudogenization and loss of CRISPR-Cas systems outside of horizontally transmitting strains likely reflects selection against maintaining this genome defense systems following the transition to vertical transmission that renders opportunities for horizontal transfers rare. Selection against CRISPR-Cas within host-associated environments could reflect the selection to retain beneficial mobile genetic elements while avoiding autoimmunity or it may alternatively be a loss of benefit from the system associated with lower rates of phage attack in the endosymbiotic environment. Indeed, a closer look at the "wasp" subclade of nasoniae reveals that CRISPR-Cas function has been independently lost at least four times in otherwise very closely related strains (Figure S4). This convergence in loss-of-function mutations supports an adaptive basis to functional loss. These systems are fragmentary in strains exhibiting vertical transmission (Figure S4; Table S2).

The total diversity of predicted intact anti-phage system (including CRISPR-Cas) is also greater in the horizontally transmitting strains compared with strains with mixed modes of transmission or vertical transmission. In strains with only vertical transmission, there is evidence of 0–4 types of systems, those with mixed modes of transmission evidence of 1 to 2 types, and those without vertical transmission 4 to 5 types of defense systems, all excluding CRISPR systems identified as pseudogenized (Figure S5).

Within intracellular and other endosymbiotic host-associated niches, bacteria are likely to encounter fewer competitors. To test whether this resulted in the loss of anticompetitor weapon systems, we examined the presence and integrity of type VI secretion systems (T6SS), a contact-dependent system for killing competitor microbes.[35] Whereas T6SS were present in the environmentally acquired strains, these were incomplete in nasoniae clade strains with either mixed modes or vertical transmission and absent or fragmentary in all but one member of the triatominarum clade, where all characterized members show vertical transmission (Figure 5). T6SS are multiprotein complexes and thus likely to be costly to produce and use (Septer et al.[36] but see Zhang et al.[37]). Their loss in

Please cite this article in press as: Siozios et al., Genome dynamics across the evolutionary transition to endosymbiosis, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.10.044

CellPress
OPEN ACCESS

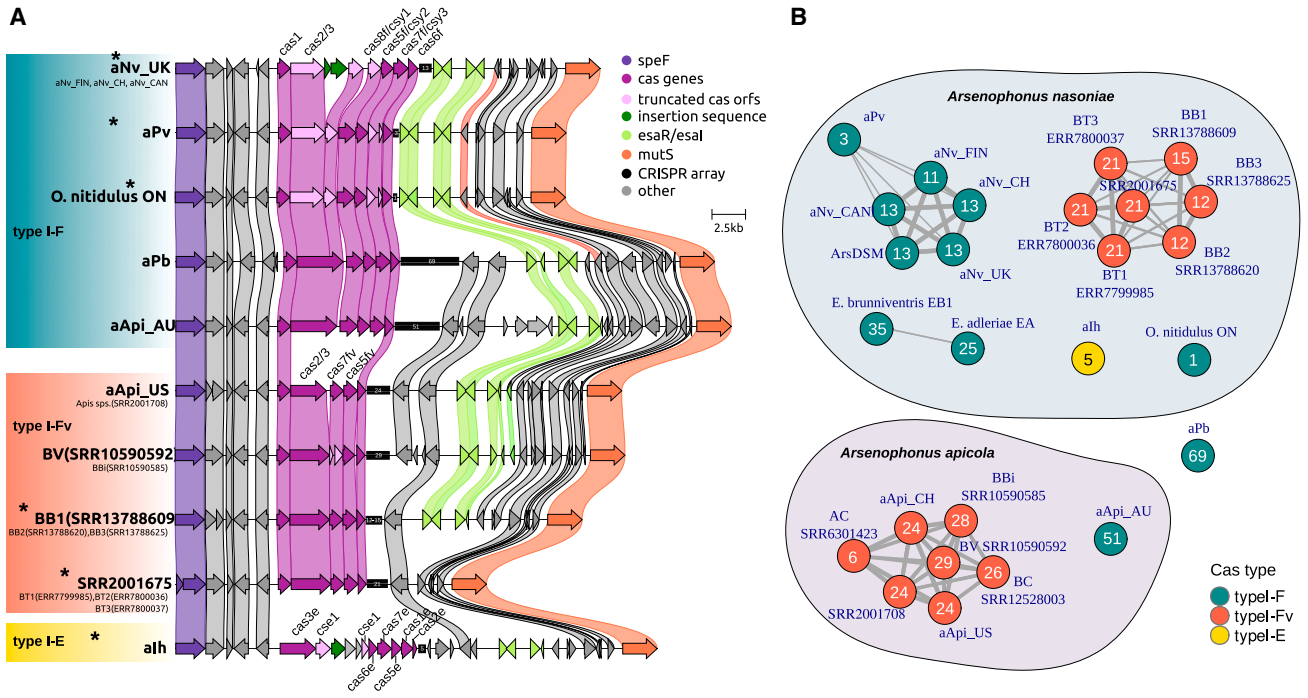


Figure 6. The CRISPR-Cas phage defense system in *Arsenophonus* and early signs of pseudogenization within the nasoniae clade

(A) Gene order and genomic context of the *Arsenophonus* CRISPR-Cas systems plotted with clinker software. Groups of homologous genes are connected by colored ribbons. *Arsenophonus* strains belonging to the nasoniae clade are highlighted with an asterisk.

(B) Networks showing the relatedness of *Arsenophonus* genomes based on shared CRISPR spacers. Nodes correspond to genomes and the edges are scaled based on the spacer repertoire relatedness (see details in STAR Methods section). Absence of edge corresponds to no shared spacers between the genomes. Nodes are colored according to the Cas type identified in each genome. The numbers in the nodes represent the number of spacers identified by CRISPRCasFinder tool.

See also Figures S2–S5, and Table S2.

endosymbionts supports a model where reduced competition within the endosymbiotic niche reducing the benefits of maintaining T6SS.

Metabolic capacity is not gained during genome expansion

Classically, endosymbionts have a reduced metabolic capability compared with free-living relatives, as the stable nutritional environment within host cells reduces the need for metabolic plasticity. Contrastingly, a recent study of Chlamydiae concluded that some lineages of intracellular symbiont had acquired metabolic capability.[38] In the *Arsenophonus* clade, metabolic capabilities are not markedly distinct between vertically transmitted endosymbionts and their closely related horizontally transmitted strains. The completeness of metabolic pathways broadly reflects the abilities of these strains to grow in *in vitro* cell-free culture.[17] The hosts of vertically transmitted strains most commonly live on other insects as parasites. Thus, anabolic provision is not expected to be of key importance in the maintenance of the symbiosis. One notable difference in predicted metabolism is beta oxidation of fatty acids, which was only intact in horizontally transmitting strains. This result was sustained when draft genomes for the triatominarum group were additionally examined (Figure S6). Aspects of cofactor synthesis, amino acid synthesis, and carbohydrate metabolism were degraded in strains that live intracellularly, and loss of function in these pathways was most pronounced for obligate co-dependent endosymbiont strains.

Gene loss and molecular evolution across varying degrees of symbiosis integration

We further utilized our data set to examine gene loss across different degrees of symbiosis integration. As expected, the number of predicted coding sequences (CDS) was an approximately linear function of genome size (Figure 2B). The number of pseudogenes was low in both horizontally transmitting strains and in interdependent obligate symbionts, at intermediate levels in the nasoniae group containing vertically transmitted strains that retain horizontal transfer and culturability, and at highest levels in unculturable facultative endosymbiont strains (Figures 2C and S7). These data are consistent with the currently accepted model, where pseudogenization accelerates upon entering into endosymbiosis, the number of pseudogenes then increases over time until the pseudogenized material is purged, resulting in a reduction in both genome size and the total number of genes encoded.

A key hypothesis in symbiont evolution is that vertical transmission leads to bottlenecks in symbiont population size. This process is expected to increase the importance of drift over selection and thus weaken the capacity of purifying selection to maintain function.[39] We analyzed the pattern of molecular evolution of highly conserved single-copy genes that are critical for

Please cite this article in press as: Siozios et al., Genome dynamics across the evolutionary transition to endosymbiosis, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.10.044

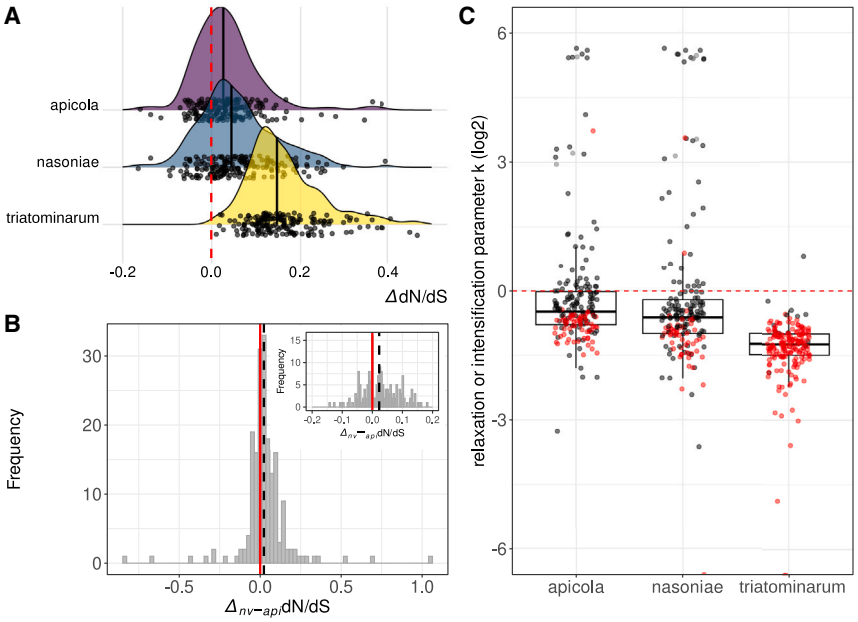


Figure 7. Evidence of relaxation of selection between *Arsenophonus* clades with contrasting modes of transmission

(A) Distribution of gene-wise dN/dS ratios in the three main *Arsenophonus* clades as compared with the outgroup clade comprised of free-living species (*Providencia stuartii*, *Proteus mirabilis*, *Morganella morganii,* and *Moellerella wisconsensis*) (dN/dStest clade−dN/dSoutgroup) for 188 highly conserved single-copy BUSCO marker genes. Individual values are shown as jitter points. Black solid lines represent the median of the distribution. In all clades the median is shifted to the right with the triatominarum clade (vertical transmission) showing the largest shift followed by the nasoniae clade (mixed mode of transmission) and last the apicola clade (environmental transmission) suggesting a gradual increase in the dN/dS ratios as we progress toward protract symbiosis.

(B) Distribution of gene-wise differences in dN/dS ratios between nasoniae and apicola clades for the same 188 highly conserved BUSCO marker genes. The black dotted vertical line represents the median of the distribution, which is shifted to the right (median = 0.0228146, Wilcoxon signed rank test V = 11,373, p= 4.018e−06), indicating that nasoniae clade has higher dN/dS ratios compared with apicola clade, which is mostly characterized by environmental mode of transmission. A narrower range of the same data between values −0.2 and 0.2 is shown in the inset plot on the top right corner.

(C) Distribution of relaxation or intensification parameter k (log_2) per gene as calculated by the RELAX method in HyPhy package (v2.3) compared with the outgroup clade. Values below zero indicate that selection strength has been relaxed while values above zero indicate an intensification of the selection strength. Genes with statistically significant k values (FDR; q < 0.1) are shown as red jitter dots.

See also Figure S10 for definition of clades in the above analysis.

microbial function (see STAR Methods for details). All three clades (nasoniae, apicola, and triatominarum) showed evidence of relaxed selection compared with the autonomous free-living outgroup. Relaxed selection was most pronounced in the triatominarum clade comprising vertically transmitted intracellular symbionts with a long history of endosymbiosis. Relaxation of selection was also observed in the other two clades. Notably, it was more pronounced in the nasoniae clade, where the strains have mixed modes of transmission or vertical transmission, than in the apicola clade, where environmental transmission is common (Figure 7). These data corroborate our current thinking of evolutionary patterns in vertically transmitted symbiosis, which combines redundant gene loss through accumulation of deletions[7] with gene degradation through the fixation of mildly deleterious alleles, the latter permitted by the increased primacy of drift processes. Notably, this signature can be detected shortly following the evolution of vertical transmission.

Finally, we examined how overall genomic features vary between strains. Reductions in %GC content are a recognized feature of evolution during symbiosis, with small endosymbiont genomes commonly being highly AT rich.[3] However, the observation is not universally found within clades, with no association between genome size and %GC in the genus *Spiroplasma*,[40] and only the very reduced genomes being AT rich in *Sodalis*.[3] Mirroring *Sodalis*, we observed markedly reduced %GC only in the obligate co-dependent *Arsenophonus* endosymbionts with highly reduced genomes (Figure 2D). For the other strains, %GC content is consistent at 37%–40%. Analysis of CDS of non-obligate strains did not support a relationship between % GC and genome size (null hypothesis of no association: $F_{1,\ 33}$ = 1.141, p = 0.29; see Figure S8). The pattern in *Arsenophonus* mirrors that of *Sodalis* and indicates that the restriction of reduced %GC to obligate interdependent symbioses may be common. In *Arsenophonus*, reduced %GC is found in the obligate strains where DNA repair systems are ablated (Figure S9). Increased primacy of genetic drift associated with the pronounced bottlenecks that accompany obligate interdependent symbioses may also contribute to the pattern.

Emerging model of evolution in the early stages of transition to vertically transmitted symbiosis

Our examination of genome evolution across symbiotic lifestyles in the genus *Arsenophonus* refines our model for the evolution of endosymbiosis. Becoming a persistent vertically transmitted endosymbiont requires rapid evolutionary innovation fueled by horizontal gene transfer, notably the gain of new functions for host manipulation. It is notable that increases in genome size are observed in other bacterial symbionts that have recently evolved vertical transmission. In *Arsenophonus*, this rapid genome expansion is achieved through the acquisition of prophage and plasmid mobile genetic elements, and this itself is enabled by the loss of genome defense systems, including CRISPR-Cas. The invasion of prophage is associated with enrichment for T3SS effector toxins, which permit establishment of endosymbiosis in a hostile host environment, but this is followed by loss of these systems in strains that become vertically transmitted and more highly adapted to the host intracellular environment.

While our model is based on data from a single genus, the centrality of CRISPR defense loss reflects recent studies of *Mycoplasma* evolution following a host switch event,[41] and the presence of intact or recently pseudogenized CRISPR in culturable, but not unculturable, aphid-associated *Serratia*.[12] In

Mycoplasma, the authors argued that the loss of CRISPR systems enabled adaptation to the novel host species. We argue that the evolutionary transition from free-living to heritable endosymbiosis may commonly be associated with more complex genome remodeling than previously reported. In our new model, genome expansion and the associated increase in opportunities for functional innovation precede the processes of reductive evolution traditionally associated with vertically transmitted endosymbiosis.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Greg Hurst (g.hurst@liverpool.ac.uk).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Sequence data and genomes generated in this study are deposited in GenBank database under BioProject accession number GenBank: PRJNA956975 and are publicly available as of the date of publication. The genome for Ca. *Arsenophonus triatominarum* can be found in GenBank under the BioProject accession number GenBank: PRJNA311587. Accession numbers are listed in the key resources table. This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- The scripts and source data for the various analyses in this study have been deposited at Zenodo under the https://doi.org/10.5281/zenodo.12530786. It can be also found on GitHub (https://github.com/SioStef/Arsenophonus-comparative-genomics). This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We wish to thank the NERC (grant NE/I01067X/1 to G.H., K.K., and M.B.), BBSRC (grant BB/S017534/1 to G.H./A.D.), Wellcome Trust ISSF (to G.H./S.S.), and NERC NEOF for sequencing support. We would also like to thank Emily Dovydaitis of the German Centre for Integrative Biodiversity Research (iDiv) for her invaluable help in designing the graphical summary for this article, Edze Westra for discussion of CRISPR-Cas function and evolution, and the three anonymous reviewers for their insights.

AUTHOR CONTRIBUTIONS

Project design: S.S., G.D.D.H., A.C.D., E.N., V.H., M.A.B., and K.C.K. Isolation, sequencing, and assembly of strains: S.S., P.N.-J., T.A., H.S., C.L.F., S.R.P., G.T., L.B., K.C.L., L.C., E.N., and A.C.D. Isolation and assembly of *Arsenophonus* MAGs from SRA datasets: S.S. Analysis: S.S. and G.D.D.H. Writing the paper: S.S., G.D.D.H., M.A.B., K.C.K., V.H., E.N., A.C.D., T.A., and H.S. All authors edited and approved the final version of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2024.10.044.

Received: May 3, 2023
Revised: April 12, 2024
Accepted: October 15, 2024
Published: November 15, 2024

REFERENCES

1. Sachs, J.L., Mueller, U.G., Wilcox, T.P., and Bull, J.J. (2004). The Evolution of Cooperation. Q. Rev. Biol. *79*, 135–160. https://doi.org/10.1086/383541.
2. Drew, G.C., Stevens, E.J., and King, K.C. (2021). Microbial evolution and transitions along the parasite–mutualist continuum. Nat. Rev. Microbiol. *19*, 623–638. https://doi.org/10.1038/s41579-021-00550-7.
3. McCutcheon, J.P., Boyd, B.M., and Dale, C. (2019). The Life of an Insect Endosymbiont from the Cradle to the Grave. Curr. Biol. *29*, R485–R495. https://doi.org/10.1016/j.cub.2019.03.032.
4. Leeks, A., dos Santos, M., and West, S.A. (2019). Transmission, relatedness, and the evolution of cooperative symbionts. J. Evol. Biol. *32*, 1036–1045. https://doi.org/10.1111/jeb.13505.
5. Bennett, G.M., and Moran, N.A. (2015). Heritable symbiosis: the advantages and perils of an evolutionary rabbit hole. Proc. Natl. Acad. Sci. USA *112*, 10169–10176. https://doi.org/10.1073/pnas.1421388112.
6. Salem, H., Bauer, E., Kirsch, R., Berasategui, A., Cripps, M., Weiss, B., Koga, R., Fukumori, K., Vogel, H., Fukatsu, T., and Kaltenpoth, M. (2017). Drastic Genome Reduction in an Herbivore's Pectinolytic Symbiont. Cell *171*, 1520–1531.e13. https://doi.org/10.1016/j.cell.2017.10.029.
7. Kuo, C.-H., and Ochman, H. (2009). Deletional Bias across the Three Domains of Life. Genome Biol. Evol. *1*, 145–152. https://doi.org/10.1093/gbe/evp016.
8. Toft, C., and Andersson, S.G.E. (2010). Evolutionary microbial genomics: insights into bacterial host adaptation. Nat. Rev. Genet. *11*, 465–475. https://doi.org/10.1038/nrg2798.
9. Su, Y., Lin, H.-C., Teh, L.S., Chevance, F., James, I., Mayfield, C., Golic, K.G., Gagnon, J.A., Rog, O., and Dale, C. (2022). Rational engineering of a synthetic insect-bacterial mutualism. Curr. Biol. *32*, 3925–3938.e6. https://doi.org/10.1016/j.cub.2022.07.036.
10. Koga, R., Moriyama, M., Onodera-Tanifuji, N., Ishii, Y., Takai, H., Mizutani, M., Oguchi, K., Okura, R., Suzuki, S., Gotoh, Y., et al. (2022). Single mutation makes Escherichia coli an insect mutualist. Nat. Microbiol. *7*, 1141–1150. https://doi.org/10.1038/s41564-022-01179-9.
11. Santos-Garcia, D., Silva, F.J., Morin, S., Dettner, K., and Kuechler, S.M. (2017). The All-Rounder Sodalis: A New Bacteriome-Associated Endosymbiont of the Lygaeoid Bug Henestaris halophilus (Heteroptera: Henestarinae) and a Critical Examination of Its Evolution. Genome Biol. Evol. *9*, 2893–2910. https://doi.org/10.1093/gbe/evx202.

12. Renoz, F., Foray, V., Ambroise, J., Baa-Puyoulet, P., Bearzatto, B., Mendez, G.L., Grigorescu, A.S., Mahillon, J., Mardulyn, P., Gala, J.-L., et al. (2021). At the Gate of Mutualism: Identification of Genomic Traits Predisposing to Insect-Bacterial Symbiosis in Pathogenic Strains of the Aphid Symbiont Serratia symbiotica. Front. Cell. Infect. Microbiol. *11*, 660007. https://doi.org/10.3389/fcimb.2021.660007.

13. Hosokawa, T., Ishii, Y., Nikoh, N., Fujie, M., Satoh, N., and Fukatsu, T. (2016). Obligate bacterial mutualists evolving from environmental bacteria in natural insect populations. Nat. Microbiol. *1*, 15011. https://doi.org/10.1038/nmicrobiol.2015.11.

14. Wilkes, T.E., Duron, O., Darby, A.C., Hypša, V., Novakova, E., and Hurst, G.D.D. (2012). The Genus Arsenophonus. In Manipulative Tenants, E. Zchori-Fein, and K. Bourtzis, eds. (CRC Press).

15. Gherna, R.L., Werren, J.H., Weisburg, W., Cote, R., Woese, C.R., Mandelco, L., and Brenner, D.J. (1991). NOTES: Arsenophonus nasoniae gen. nov., sp. nov., the Causative Agent of the Son-Killer Trait in the Parasitic Wasp Nasonia vitripennis. Int. J. Syst. Evol. Microbiol. *41*, 563–565. https://doi.org/10.1099/00207713-41-4-563.

16. Nováková, E., Hypša, V., and Moran, N.A. (2009). Arsenophonus, an emerging clade of intracellular symbionts with a broad host distribution. BMC Microbiol. *9*, 143. https://doi.org/10.1186/1471-2180-9-143.

17. Nadal-Jimenez, P., Parratt, S.R., Siozios, S., and Hurst, G.D.D. (2023). Isolation, culture and characterization of Arsenophonus symbionts from two insect species reveal loss of infectious transmission and extended host range. Front. Microbiol. *14*, 1089143. https://doi.org/10.3389/fmicb.2023.1089143.

18. Nadal-Jimenez, P., Siozios, S., Frost, C.L., Court, R., Chrostek, E., Drew, G.C., Evans, J.D., Hawthorne, D.J., Burritt, J.B., and Hurst, G.D.D. (2022). Arsenophonus apicola sp. nov., isolated from the honeybee Apis mellifera. Int. J. Syst. Evol. Microbiol. *72*, 005469. https://doi.org/10.1099/ijsem.0.005469.

19. Parratt, S.R., Frost, C.L., Schenkel, M.A., Rice, A., Hurst, G.D.D., and King, K.C. (2016). Superparasitism Drives Heritable Symbiont Epidemiology and Host Sex Ratio in a Wasp. PLoS Pathog. *12*, e1005629. https://doi.org/10.1371/journal.ppat.1005629.

20. Yorimoto, S., Hattori, M., Kondo, M., and Shigenobu, S. (2022). Complex host/symbiont integration of a multi-partner symbiotic system in the eusocial aphid *Ceratovacuna japonica*. iScience *25*, 105478. https://doi.org/10.1016/j.isci.2022.105478.

21. Fisher, R.M., Henry, L.M., Cornwallis, C.K., Kiers, E.T., and West, S.A. (2017). The evolution of host-symbiont dependence. Nat. Commun. *8*, 15973. https://doi.org/10.1038/ncomms15973.

22. Csurös, M. (2010). Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics *26*, 1910–1912. https://doi.org/10.1093/bioinformatics/btq315.

23. Duron, O. (2014). Arsenophonus insect symbionts are commonly infected with APSE, a bacteriophage involved in protective symbiosis. FEMS Microbiol. Ecol. *90*, 184–194. https://doi.org/10.1111/1574-6941.12381.

24. Wang, Y.-H., Mikaelyan, A., Coates, B.S., and Lorenzen, M. (2024). The Genome of Arsenophonus sp. and Its Potential Contribution in the Corn Planthopper, Peregrinus maidis. Insects *15*, 113. https://doi.org/10.3390/insects15020113.

25. Gerth, M., Martinez-Montoya, H., Ramirez, P., Masson, F., Griffin, J.S., Aramayo, R., Siozios, S., Lemaitre, B., Mateos, M., and Hurst, G.D.D. (2021). Rapid molecular evolution of Spiroplasma symbionts of Drosophila. Microb. Genomics *7*, 000503. https://doi.org/10.1099/mgen.0.000503.

26. Hugoson, E., Guliaev, A., Ammunét, T., and Guy, L. (2022). Host Adaptation in Legionellales Is 1.9 Ga, Coincident with Eukaryogenesis. Mol. Biol. Evol. *39*, msac037. https://doi.org/10.1093/molbev/msac037.

27. de Albuquerque, N.R.M., Ebert, D., and Haag, K.L. (2020). Transposable element abundance correlates with mode of transmission in microsporidian parasites. Mobile DNA *11*, 19. https://doi.org/10.1186/s13100-020-00218-8.

28. Fortier, L.-C., and Sekulovic, O. (2013). Importance of prophages to evolution and virulence of bacterial pathogens. Virulence *4*, 354–365. https://doi.org/10.4161/viru.24498.

29. Dragoš, A., Andersen, A.J.C., Lozano-Andrade, C.N., Kempen, P.J., Kovács, Á.T., and Strube, M.L. (2021). Phages carry interbacterial weapons encoded by biosynthetic gene clusters. Curr. Biol. *31*, 3479–3489.e5. https://doi.org/10.1016/j.cub.2021.05.046.

30. Dale, C., Young, S.A., Haydon, D.T., and Welburn, S.C. (2001). The insect endosymbiont Sodalis glossinidius utilizes a type III secretion system for cell invasion. Proc. Natl. Acad. Sci. USA *98*, 1883–1888. https://doi.org/10.1073/pnas.98.4.1883.

31. Burstein, D., Sun, C.L., Brown, C.T., Sharon, I., Anantharaman, K., Probst, A.J., Thomas, B.C., and Banfield, J.F. (2016). Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. Nat. Commun. *7*, 10613. https://doi.org/10.1038/ncomms10613.

32. Zaayman, M., and Wheatley, R.M. (2022). Fitness costs of CRISPR-Cas systems in bacteria. Microbiology *168*, 001209. https://doi.org/10.1099/mic.0.001209.

33. Gandon, S., and Vale, P.F. (2014). The evolution of resistance against good and bad infections. J. Evol. Biol. *27*, 303–312. https://doi.org/10.1111/jeb.12291.

34. Jiang, W., Maniv, I., Arain, F., Wang, Y., Levin, B.R., and Marraffini, L.A. (2013). Dealing with the Evolutionary Downside of CRISPR Immunity: Bacteria and Beneficial Plasmids. PLoS Genet. *9*, e1003844. https://doi.org/10.1371/journal.pgen.1003844.

35. Unni, R., Pintor, K.L., Diepold, A., and Unterweger, D. (2022). Presence and absence of type VI secretion systems in bacteria. Microbiology *168*, 001151. https://doi.org/10.1099/mic.0.001151.

36. Septer, A.N., Sharpe, G., and Shook, E.A. (2023). The Vibrio fischeri type VI secretion system incurs a fitness cost under host-like conditions. Preprint at bioRxiv. https://doi.org/10.1101/2023.03.07.529561.

37. Zhang, C., Datta, S., Ratcliff, W.C., and Hammer, B.K. (2024). Constitutive expression of the Type VI Secretion System carries no measurable fitness cost in Vibrio cholerae. Ecol. Evol. *14*, e11081. https://doi.org/10.1002/ece3.11081.

38. Dharamshi, J.E., Köstlbacher, S., Schön, M.E., Collingro, A., Ettema, T.J.G., and Horn, M. (2023). Gene gain facilitated endosymbiotic evolution of Chlamydiae. Nat. Microbiol. *8*, 40–54. https://doi.org/10.1038/s41564-022-01284-9.

39. Moran, N.A. (1996). Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc. Natl. Acad. Sci. USA *93*, 2873–2878. https://doi.org/10.1073/pnas.93.7.2873.

40. Lo, W.-S., Huang, Y.-Y., and Kuo, C.-H. (2016). Winding paths to simplicity: genome evolution in facultative insect symbionts. FEMS Microbiol. Rev. *40*, 855–874. https://doi.org/10.1093/femsre/fuw028.

41. Ipoutcha, T., Tsarmpopoulos, I., Gourgues, G., Baby, V., Dubos, P., Hill, G.E., Arfi, Y., Lartigue, C., Thebault, P., Bonneaud, C., and Sirand-Pugnet, P. (2024). Evolution of the CRISPR-Cas9 defence system in Mycoplasma gallisepticum following colonization of a novel bird host. Preprint at bioRxiv. https://doi.org/10.1101/2023.03.14.532377.

42. Taylor, G.P., Coghlin, P.C., Floate, K.D., and Perlman, S.J. (2011). The host range of the male-killing symbiont *Arsenophonus nasoniae* in filth fly parasitioids. J. Invertebr. Pathol. *106*, 371–379. https://doi.org/10.1016/j.jip.2010.12.004.

43. Nadal-Jimenez, P., Frost, C.L., Cláudia Norte, A., Garrido-Bautista, J., Wilkes, T.E., Connell, R., Rice, A., Krams, I., Eeva, T., Christe, P., et al. (2023). The son-killer microbe *Arsenophonus nasoniae* is a widespread associate of the parasitic wasp *Nasonia vitripennis* in Europe. J. Invertebr. Pathol. *199*, 107947. https://doi.org/10.1016/j.jip.2023.107947.

44. Liew, K.C., Graves, S., Croft, L., Brettell, L.E., Cook, J., Botes, J., and Newton, P. (2022). First human case of infection with Arsenophonus nasoniae, the male killer insect pathogen. Pathology *54*, 664–666. https://doi.org/10.1016/j.pathol.2021.08.011.

45. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

46. Wick, R.R., Judd, L.M., Gorrie, C.L., and Holt, K.E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput. Biol. *13*, e1005595. https://doi.org/10.1371/journal.pcbi.1005595.

47. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26. https://doi.org/10.1038/nbt.1754.

48. Wick, R.R., and Holt, K.E. (2022). Polypolish: Short-read polishing of long-read bacterial genome assemblies. PLoS Comput. Biol. *18*, e1009802. https://doi.org/10.1371/journal.pcbi.1009802.

49. Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P.A. (2019). Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. *37*, 540–546. https://doi.org/10.1038/s41587-019-0072-8.

50. Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., and Earl, A.M. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS One *9*, e112963. https://doi.org/10.1371/journal.pone.0112963.

51. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

52. Chin, C.-S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods *10*, 563–569. https://doi.org/10.1038/nmeth.2474.

53. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. *17*, 132. https://doi.org/10.1186/s13059-016-0997-x.

54. Ondov, B.D., Starrett, G.J., Sappington, A., Kostic, A., Koren, S., Buck, C.B., and Phillippy, A.M. (2019). Mash Screen: high-throughput sequence containment estimation for genome discovery. Genome Biol. *20*, 232. https://doi.org/10.1186/s13059-019-1841-x.

55. Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics *31*, 1674–1676. https://doi.org/10.1093/bioinformatics/btv033.

56. Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ *7*, e7359. https://doi.org/10.7717/peerj.7359.

57. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. *25*, 1043–1055. https://doi.org/10.1101/gr.186072.114.

58. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

59. Eren, A.M., Kiefl, E., Shaiber, A., Veseli, I., Miller, S.E., Schechter, M.S., Fink, I., Pan, J.N., Yousef, M., Fogarty, E.C., et al. (2021). Community-led, integrated, reproducible multi-omics with anvi'o. Nat. Microbiol. *6*, 3–6. https://doi.org/10.1038/s41564-020-00834-3.

60. Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., and Zdobnov, E.M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol. Biol. Evol. *38*, 4647–4654. https://doi.org/10.1093/molbev/msab199.

61. Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. *20*, 238. https://doi.org/10.1186/s13059-019-1832-y.

62. Bruen, T.C., Philippe, H., and Bryant, D. (2006). A Simple and Robust Statistical Test for Detecting the Presence of Recombination. Genetics *172*, 2665–2681. https://doi.org/10.1534/genetics.105.048975.

63. Steenwyk, J.L., iii, Buida, T.J., Li, Y., Shen, X.-X., and Rokas, A. (2020). ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. PLoS Biol. *18*, e3001007. https://doi.org/10.1371/journal.pbio.3001007.

64. Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLoS One *11*, e0163962. https://doi.org/10.1371/journal.pone.0163962.

65. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol. Biol. Evol. *32*, 268–274. https://doi.org/10.1093/molbev/msu300.

66. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods *14*, 587–589. https://doi.org/10.1038/nmeth.4285.

67. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069. https://doi.org/10.1093/bioinformatics/btu153.

68. Lartillot, N., Rodrigue, N., Stubbs, D., and Richer, J. (2013). PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. Syst. Biol. *62*, 611–615. https://doi.org/10.1093/sysbio/syt022.

69. Arndt, D., Grant, J.R., Marcu, A., Sajed, T., Pon, A., Liang, Y., and Wishart, D.S. (2016). PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. *44*, W16–W21. https://doi.org/10.1093/nar/gkw387.

70. Xie, Z., and Tang, H. (2017). ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. Bioinformatics *33*, 3340–3347. https://doi.org/10.1093/bioinformatics/btx433.

71. Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., Rocha, E.P.C., Vergnaud, G., Gautheret, D., and Pourcel, C. (2018). CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res. *46*, W246–W251. https://doi.org/10.1093/nar/gky425.

72. Katoh, K., and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment, software version 7: Improvements in Performance and Usability. Mol. Biol. Evol. *30*, 772–780. https://doi.org/10.1093/molbev/mst010.

73. Huson, D.H., and Bryant, D. (2006). Application of Phylogenetic Networks in Evolutionary Studies. Mol. Biol. Evol. *23*, 254–267. https://doi.org/10.1093/molbev/msj030.

74. Tesson, F., Hervé, A., Mordret, E., Touchon, M., d'Humières, C., Cury, J., and Bernheim, A. (2022). Systematic and quantitative view of the antiviral arsenal of prokaryotes. Nat. Commun. *13*, 2561. https://doi.org/10.1038/s41467-022-30269-9.

75. Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat. Commun. *9*, 5114. https://doi.org/10.1038/s41467-018-07641-9.

76. Csárdi, G., Nepusz, T., Traag, V., Horvát, S., Zanini, F., Noom, D., and Müller, K. (2024). igraph: Network Analysis and Visualization in Rx. R package version 2.0. 2. https://doi.org/10.5281/zenodo.7682609.

77. Kosakovsky Pond, S.L., Poon, A.F.Y., Velazquez, R., Weaver, S., Hepler, N.L., Murrell, B., Shank, S.D., Magalis, B.R., Bouvier, D., Nekrutenko, A., et al. (2020). HyPhy 2.5—A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. Mol. Biol. Evol. *37*, 295–299. https://doi.org/10.1093/molbev/msz197.

78. Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. *34*, W609–W612. https://doi.org/10.1093/nar/gkl315.

79. R Core Team (2020). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).
80. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag New York).
81. Pedersen, T.L. (2024). patchwork: The Composer of Plots.
82. Yu, G., Smith, D.K., Zhu, H., Guan, Y., and 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. https://doi.org/10.1111/2041-210X.12628.
83. Gilchrist, C.L.M., and Chooi, Y.-H. (2021). Clinker & clustermap.js: automatic generation of gene cluster comparison figures. Bioinformatics *37*, 2473–2475. https://doi.org/10.1093/bioinformatics/btab007.
84. Hillyard, P.D. (1996). Ticks of North-West Europe: Keys and Notes for Identification of the Species (Published for the Linnean Society of London and the Estuarine and Coastal Sciences Association by Field Studies Council).
85. Milhano, N., Palma, M., Marcili, A., Núncio, M.S., de Carvalho, I.L., and de Sousa, R. (2014). *Rickettsia lusitaniae* sp. nov. isolated from the soft tick *Ornithodoros erraticus* (Acarina: Argasidae). Comp. Immunol. Microbiol. Infect. Dis. *37*, 189–193. https://doi.org/10.1016/j.cimid.2014.01.006.
86. Ammerman, N.C., Beier-Sexton, M., and Azad, A.F. (2008). Laboratory Maintenance of Rickettsia rickettsii. Curr. Protoc. Microbiol. *11*, 3A.5. https://doi.org/10.1002/9780471729259.mc03a05s11.
87. Frost, C.L., Siozios, S., Nadal-Jimenez, P., Brockhurst, M.A., King, K.C., Darby, A.C., and Hurst, G.D.D. (2020). The Hypercomplex Genome of an Insect Reproductive Parasite Highlights the Importance of Lateral Gene Transfer in Symbiont Biology. mBio *11*, e02590-19. https://doi.org/10.1128/mBio.02590-19.
88. Quick, J. (2018). Ultra-long read sequencing protocol for RAD004.
89. Muto, A., Kotera, M., Tokimatsu, T., Nakagawa, Z., Goto, S., and Kanehisa, M. (2013). Modular Architecture of Metabolic Pathways Revealed by Conserved Sequences of Reactions. J. Chem. Inf. Model. *53*, 613–622. https://doi.org/10.1021/ci3005379.
90. Touchon, M., Cury, J., Yoon, E.-J., Krizova, L., Cerqueira, G.C., Murphy, C., Feldgarden, M., Wortman, J., Clermont, D., Lambert, T., et al. (2014). The Genomic Diversification of the Whole Acinetobacter Genus: Origins, Mechanisms, and Consequences. Genome Biol. Evol. *6*, 2866–2882. https://doi.org/10.1093/gbe/evu225.
91. Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol. Biol. Evol. *35*, 518–522. https://doi.org/10.1093/molbev/msx281.
92. Shimodaira, H. (2002). An Approximately Unbiased Test of Phylogenetic Tree Selection. Syst. Biol. *51*, 492–508. https://doi.org/10.1080/10635150290069913.
93. Schmartz, G.P., Hartung, A., Hirsch, P., Kern, F., Fehlmann, T., Müller, R., and Keller, A. (2022). PLSDB: advancing a comprehensive database of bacterial plasmids. Nucleic Acids Res. *50*, D273–D278. https://doi.org/10.1093/nar/gkab1111.
94. Wertheim, J.O., Murrell, B., Smith, M.D., Kosakovsky Pond, S.L., and Scheffler, K. (2015). RELAX: Detecting Relaxed Selection in a Phylogenetic Framework. Mol. Biol. Evol. *32*, 820–832. https://doi.org/10.1093/molbev/msu400.

Please cite this article in press as: Siozios et al., Genome dynamics across the evolutionary transition to endosymbiosis, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.10.044

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and Virus Strains		
Arsenophonus nasoniae strain aNv_CAN	This study and Taylor et al.[42] Nadal-Jimenez et al.[43]	NCBI Taxonomy ID: 638
Arsenophonus nasoniae strain aNv_CH	This study and Nadal-Jimenez et al.[43]	NCBI Taxonomy ID: 638
Arsenophonus nasoniae strain aNv_UK	This study and Nadal-Jimenez et al.[43]	NCBI Taxonomy ID: 638
Arsenophonus nasoniae strain aPv	This study and Nadal-Jimenez et al.[17] Available from culture collection LMG 32964	NCBI Taxonomy ID: 638
Arsenophonus nasoniae strain aIh	This study	NCBI Taxonomy ID: 638
Arsenophonus apicola strain aApi_AU	Liew et al.[44]	NCBI: Taxonomy ID: 2879119
Ca. Arsenophonus triatominarum	Maintained in the lab of Guenter Schaub	NCBI Taxonomy ID: 57911
Arsenophonus sp. aPb isolated from the butterfly *Polyommatus bellargus*	This study and Nadal-Jimenez et al.[17] Available from culture collection LMG 32963	NCBI Taxonomy ID: 3041619
Critical Commercial Assays		
QIAGEN DNeasy Blood & Tissue Kit	Qiagen	Cat#69504
QIAGEN Genomic-tip 20/G	Qiagen	Cat#10223
QIAGEN Genomic DNA Buffer Set	Qiagen	Cat#19060
Oxford Nanopore Ligation Sequencing kit	Oxford Nanopore	Cat#SQK-LSK109
Oxford Nanopore Rapid Sequencing kit	Oxford Nanopore	Cat#SQK-RBK004
Oxford Nanopore MinION flow cell	Oxford Nanopore	FLO-MIN106 R9.4
Oxford Nanopore MinION Flongle Flow Cell	Oxford Nanopore	FLO-FLGOP1
Deposited Data		
Genomic data	This study	GenBank: PRJNA956975
Arsenophonus nasoniae aNv_CAN	This study	GenBank: GCF_029873515.1
Arsenophonus nasoniae aNv_CH	This study	GenBank: GCF_029873535.1
Arsenophonus nasoniae aNv_UK	This study	GenBank: GCF_029873555.1
Arsenophonus nasoniae aPv	This study	GenBank: GCF_029873495.1
Arsenophonus nasoniae aIh	This study	GenBank: GCF_029873455.1
Arsenophonus sp. aPb	This study	GenBank: GCF_029873475.1
Arsenophonus apicola aApi_AU	This study	GenBank: GCF_029906405.1
Arsenophonus triatominarum Ati-2015	This study	GenBank: GCA_001640365.1
Arsenophonus nasoniae aNv_FIN	NCBI	GenBank: GCF_004768525.1
Arsenophonus nasoniae DSM15247	NCBI	GenBank: AUCC00000000.1
Arsenophonus apicola aApi_US	NCBI	GenBank: GCF_020268605.1
Arsenophonus apicola aApi_CH	NCBI	GenBank: GCF_903968575.1
Arsenophonus of *Entylia carinata* ENCA	NCBI	GenBank: GCA_002287155.1
Arsenophonus of *Aleurodicus floccissimus* (ARAF)	NCBI	GenBank: GCA_900343025.1
Arsenophonus of *Nilaparvata lugens* (Hangzhou)	NCBI	GenBank: JRLH00000000.1
Arsenophonus of *Aphis craccivora*	NCBI	GenBank: NZ_CP038155.1
Arsenophonus of *Bemisia tabaci* Asia II 3	NCBI	GenBank: NZ_MASH00000000.1
Arsenophonus of *Bemisia tabaci* ArsBTMEDQ21	NCBI	GenBank: GCA_902713415.1
Arsenophonus melophagi	NCBI or http://users.prf.jcu.cz/novake01/	GenBank: JAVYJS000000000.1
Arsenophonus of *Ceratovacuna japonica* (ArsCjap)	NCBI	GenBank: GCA_024349725.1
Arsenophonus of *Lipoptena fortisetosa*	NCBI	GenBank: GCA_001534665.1
Arsenophonus of *Aleurodicus dispersus* (ARAD)	NCBI	GenBank: GCA_900343015.1
Ca. Riesia pediculicola	NCBI	GenBank: GCA_000093065.1

(*Continued on next page*)

Please cite this article in press as: Siozios et al., Genome dynamics across the evolutionary transition to endosymbiosis, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.10.044

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Ca. Riesia pediculischaeffi	NCBI	GenBank: GCA_002073895.1
Proteus mirabilis strain HI4320	NCBI	GenBank: NC_010554
Providencia stuartii strain MRSN 2154	NCBI	GenBank: NC_017731
Photorhabdus thracensis strain DSM 15199	NCBI	GenBank: NZ_CP011104
Xenorhabdus nematophila ATCC 19061	NCBI	GenBank: NC_014228
Apidae (bulk)	NCBI	SRA: SRR2001708
Bombus pratorum	NCBI	SRA: SRR2001675
Bombus terrestris	NCBI	SRA: ERR7799985
Bombus terrestris	NCBI	SRA: ERR7800036
Bombus terrestris	NCBI	SRA: ERR7800037
Bombus balteatus	NCBI	SRA: SRR13788625
Bombus balteatus	NCBI	SRA: SRR13788609
Bombus balteatus	NCBI	SRA: SRR13788620
Eurytoma adleriae	NCBI	SRA: ERR1981844, ERR1981845, ERR1981846
Eurytoma brunniventris	NCBI	SRA: ERR1981895, ERR1981896, ERR1981897
Eurytoma brunniventris	NCBI	SRA: ERR1981889, ERR1981890, ERR1981891
Ormyrus nitidulus	NCBI	SRA: ERR1982147, ERR1982148, ERR1982149
Bombus bifarius	NCBI	SRA: SRR10590585
Bombus vancouverensis	NCBI	SRA: SRR10590592
Bombus cullumanus	NCBI	SRA: SRR12528003
Apis cerana	NCBI	SRA: SRR6301423
Dermacentor variabilis	NCBI	SRA: SRR5317834
Software and Algorithms		
MinKNOW software v18.01.6	Oxford Nanopore	https://nanoporetech.com/
Trimmomatic 0.30	Bolger et al.[45]	http://www.usadellab.org/cms/?page=trimmomatic
Unicycler v0.4.5	Wick et al.[46]	https://github.com/rrwick/Unicycler
Integrative Genomics Viewer (IGV) v2.8.9	Robinson et al.[47]	https://igv.org/
Polypolish v0.5.0	Wick et al.[48]	https://github.com/rrwick/Polypolish
Flye v2.8 and Flye v2.9.1	Kolmogorov et al.[49]	https://github.com/fenderglass/Flye
Pilon v1.22	Walker et al.[50]	https://github.com/broadinstitute/pilon
Guppy v4.2.2	Oxford Nanopore	https://nanoporetech.com/
minimap2 v2.17-r941	Li[51]	https://github.com/lh3/minimap2
HGAP	Chin et al.[52]	https://www.pacb.com
Mash v2.3	Ondov et al.[53,54]	https://mash.readthedocs.io/en/latest/
MEGAHIT v1.2.8	Li et al.[55]	https://github.com/voutcn/megahit
MetaBAT2 v2.12.1	Kang et al.[56]	https://bitbucket.org/berkeleylab/metabat/src/master/
CheckM v1.0.18	Parks et al.[57]	https://github.com/Ecogenomics/CheckM
BLAST 2.12.0+	NCBI[58]	https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
anvio v7	Eren et al.[59]	https://github.com/merenlab/anvio/releases

(*Continued on next page*)

Please cite this article in press as: Siozios et al., Genome dynamics across the evolutionary transition to endosymbiosis, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.10.044

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
BUSCO v4.1.4	Manni et al.[60]	https://busco.ezlab.org/
Orthofinder v2.3.11	Emms and Kelly[61]	https://github.com/davidemms/OrthoFinder
Phipack	Bruen et al.[62]	https://www.maths.otago.ac.nz/~dbryant/software.html
ClipKIT v1.3.0	Steenwyk et al.[63]	https://github.com/JLSteenwyk/ClipKIT
seqkit	Shen et al.[64]	https://bioinf.shenwei.me/seqkit/
IQ-TREE v1.6.12	Nguyen et al.[65]	http://www.iqtree.org/
ModelFinder	Kalyaanamoorthy et al.[66]	http://www.iqtree.org/
prokka v1.13	Seemann[67]	https://github.com/tseemann/prokka
PhyloBayes-MPI v1.9	Lartillot[68]	https://pbil.univ-lyon1.fr/software/phylobayes/
Count v10.04	Csűös[22]	http://www.iro.umontreal.ca/~csuros/gene_content/count.html
Ideel	N/A	https://github.com/mw55309/ideel
PHAge Search Tool Enhanced Release (PHASTER)	Arndt[69]	https://phaster.ca/
ISEScan v1.7.2.3	Xie and Tang[70]	https://github.com/xiezhq/ISEScan
CRISPRCasFinder	Couvin[71]	https://crisprcas.i2bc.paris-saclay.fr/CrisprCasFinder/Index
MAFFT	Katoh and Standley[72]	https://mafft.cbrc.jp/alignment/software/
SplitsTree v4.19.0	Huson and Bryant[73]	https://uni-tuebingen.de/fakultaeten/mathematisch-naturwissenschaftliche-fakultaet/fachbereiche/informatik/lehrstuehle/algorithms-in-bioinformatics/software/splitstree/
DefenseFinder	Tesson et al.[74]	https://defensefinder.mdmlab.fr/
FastANI v1.33	Jain et al.[75]	https://github.com/ParBLiSS/FastANI
igraph v1.6.0 R package	Csárdi[76]	https://igraph.org/
HyPhy v2.5.8	Kosakovsky et al.[77]	http://hyphy.org/
pal2nal	Suyama et al.[78]	https://www.bork.embl.de/pal2nal/
R v4.2.2	R core development team[79]	https://www.r-project.org/
ggplot2 R package	Wickham[80]	https://ggplot2.tidyverse.org/
patchwork R package	Pedersen[81]	https://patchwork.data-imaginist.com/
pheatmap v1.0.12 R package	N/A	https://rdrr.io/cran/pheatmap/
ggtree v4.2 R package	Yu et al.[82]	https://bioconductor.org/packages/release/bioc/html/ggtree.html
clinker v0.0.24	Gilchrist and Chooi[83]	https://github.com/gamcil/clinker

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Arsenophonus isolates and cultivation

Arsenophonus nasoniae isolates aNv_UK, aNv_CH and aNv_CAN derived from *Nasonia vitripennis* from the UK, Switzerland and Canada respectively (isolation as previously described in Nadal-Jimenez et al.[43] and Taylor et al.[42]) The isolation and culture of *A. nasoniae* aPv from *Pachycrepoideus vindemmiae* and the *Arsenophonus* strain aPb from the butterfly *Polyommatus bellargus* is described in Nadal-Jimenez et al.[17] The *Arsenophonus nasoniae* strain ArsIxoH (aIh) previously identified in the parasitoid wasp *Ixodiphagus hookeri* was isolated from questing *Ixodes ricinus* nymphs collected in September 2020 by blanket dragging in De Buunderkamp, the Netherlands (52° 00′ N, 5° 44′ E). These nymphs were morphologically identified to species level using an identification key.[84] Ticks were kept live until processing. Isolation and culture of *A. nasoniae* aIh was performed as described in Milhano et al.[85] and harvested as described in Ammerman et al.[86] with modifications. Specifically, the needle and syringe protocol was implemented using 26 gauge needles and a 0.45μm syringe-driven membrane filter. Isolation and culture of the *Arsenophonus apicola* strain aApi_AU from Australian honey bees is described in Liew et al.[44]

Ca. A. triatominarum was isolated from the host species *Triatoma infestans* maintained in the lab of Guenter Schaub (origin: colony 37, collected Bolivia 2005). The bacteria were isolated from surface sterilized *T. infestans* by collecting haemolymph. Two microliters of haemolymph was added to a 90% confluent cell monolayer of the *Drosophila melanogaster* S2 cell line (Drosophila Genetic Resource Centre stock number 6) grown in a 24-well plate with 1 ml of MMI plus 20% FBS. The preparation was centrifuged at 1,000x g at room temperature for 10 min to bring the bacteria into contact with the cells and incubated at 26.5°C for 16 h. At 10-day intervals, medium was taken from the cell layer and passaged onto a new 90% confluent S2 cell monolayer. The insect cell cultures were also tested routinely for microorganisms cultivable on 5% sheep blood agar plates incubated at 26.5°C, and they were inspected daily for 10 days at a magnification of x600 for microbial growth using an inverted M100 microscope (Swift-Microtec, Oxford, United Kingdom). Each symbiont-cell line was then cloned by limiting dilution, replicated three times and maintained as described above on S2 cell cultures. Cultures for genome sequencing were bulked up in 200 ml tissue flask. Total DNA was extracted from insects and 10-day-old insect cell cultures using a DNeasy tissue kit (QIAGEN, United Kingdom) following the manufacturer's protocol for cultured animal cells.

METHOD DETAILS

Targeted sequencing, assembly, and annotation of focal *Arsenophonus* strains

All targeted genomes were sequenced using a combination of short (Illumina) and long (Nanopore) reads as described below, with the exception of *Ca.* A. triatominarum, which was completed solely with PacBio reads.

Genome sequencing of *Arsenophonus nasoniae* strain aNv_UK. The aNv_UK strain was sequenced following the same procedure as described in Frost et al.[87] Briefly, high molecular weight (HMW) gDNA was prepared from a 50ml culture using a modified CTAB and phenol/chloroform extraction protocol.[88] Nanopore sequencing was performed with the Rapid Sequencing Kit (SQK-RAD004) (Oxford Nanopore, UK) on a FLO-MIN106 R9.4 MinION flow cell using 3ug of HMW gDNA and omitting the library loading beads to avoid blocking the sample port. The raw Nanopore reads were live basecalled in MinKNOW software v18.01.6 (Oxford Nanopore, UK). Low quality reads (quality score < 7) or small reads (<1kb) were discarded. Illumina sequencing was performed by MicrobesNG (Birmingham, AL) using the Nextera XT library prep protocol on a MiSeq platform (Illumina, San Diego, CA, USA). Reads were adapter trimmed using Trimmomatic 0.30, with a sliding window quality cutoff of Q15.[45] A hybrid assembly based on short and long reads was generated using the Unicycler pipeline version 0.4.5 under the normal mode.[46] The quality of the assembly was assessed by mapping the long reads back to it and manually inspecting in the Integrative Genomics Viewer (IGV) v2.8.9 for inconsistencies.[47] A final round of polishing using the Illumina reads was performed with Polypolish v0.5.0.[48]

Genome sequencing of the *Arsenophonus nasoniae* strains (aNv_CAN, aNv_CH and aPv). These genomes were sequenced by MicrobesNG (Birmingham, UK) using their enhanced genome service. Briefly, long-read gDNA libraries were prepared with Oxford Nanopore SQK-RBK004 kit (Oxford Nanopore, UK) using 400–500 ng high molecular weight DNA and sequenced in a FLO-MIN106 (R.9.4.1) flow cell in a GridION (Oxford Nanopore, UK). Short-read Illumina sequencing was performed with the Nextera XT library prep protocol on a HiSeq platform (Illumina) using a 250 bp paired-end protocol. Reads were adapter trimmed using Trimmomatic 0.30, with a sliding window quality cutoff of Q15.[50] An initial assembly of the long reads was performed using Flye assembler v2.8[49] under the uneven coverage mode ("-meta" option) and the "-plasmid" option enabled. Subsequently, the long reads were mapped back to each assembly and manually inspected for inconsistencies. Short read polishing of the assemblies was performed using five rounds of polishing with Pilon v1.22[50] followed by a round of polishing with Polypolish v0.5.0.[48]

Genome sequencing of the *Arsenophonus* strain aPb identified in the butterfly *Polyommatus bellargus*. High molecular weight gDNA was extracted using a Qiagen genomic-tip 20/g and the Qiagen Genomic DNA protocol for Gram-negative bacteria (QIAGEN, UK) from about 1ml of liquid culture in Brain Heart Infusion (BHI) medium. Long-read gDNA libraries were prepared with the Oxford Nanopore SQK-LSK109 kit (Oxford Nanopore, UK) using 500 ng high molecular weight DNA and sequenced on a FLO-FLGOP1 flow cell and the Flongle-MinION adapter (Oxford Nanopore, UK). Raw Nanopore reads were subsequently basecalled using Guppy v4.2.2 (Oxford Nanopore, UK) under the high accuracy model. A preliminary assembly was performed using Flye v2.8 under the uneven coverage mode ("-meta" option). Short Illumina reads obtained from Nadal-Jimenez et al.[17] were mapped to the long-read assembly using minimap2 v2.17-r941[51] to identify and extract the *Arsenophonus* reads. Short and long *Arsenophonus* reads were used to prepare the final assembly using the Unicycler pipeline version 0.4.5 under the normal mode. Like previously, the quality of the final assembly was assessed by mapping the long reads back to it and manually inspecting for inconsistencies. A final round of polishing using the Illumina reads was performed with Polypolish v0.5.0.[48]

Genome sequencing of *Arsenophonus nasoniae* strain alh. Short Illumina reads were generated on the Illumina NovaSeq 6000 system (Illumina, San Diego, CA, USA) at Baseclear (Leiden, the Netherlands) from genomic DNA extracted using the ZymoBIOMICS™ 96 MagBead DNA Kit (Zymo Research, Orange, CA). High molecular weight gDNA was extracted using a Qiagen genomic-tip as described above for *Arsenophonus* aPb. Long-read gDNA libraries were prepared with the Oxford Nanopore SQK-LSK109 kit (Oxford Nanopore, UK) using 1500 ng high molecular weight DNA and sequenced in a FLO-MIN106D (R.9.4.1) MinION flow cell (Oxford Nanopore, UK). Raw Nanopore reads were subsequently basecalled using Guppy v4.2.2 (Oxford Nanopore, UK) under the high accuracy model. Low quality and short reads were removed and the remaining reads were assembled using the Flye assembler v2.8 under the uneven coverage mode ("-meta" option) and the "-plasmid" option enabled. The quality of the

Please cite this article in press as: Siozios et al., Genome dynamics across the evolutionary transition to endosymbiosis, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.10.044

assembly was assessed as aforementioned by mapping the long reads back to the assembly and manually inspected for inconsistencies. Short read polishing was performed using five rounds of polishing with Pilon v1.22[50] followed by a round of polishing with Polypolish v0.5.0.[48]

Genome sequencing of *Arsenophonus apicola* strain aApi_AU. This genome was sequenced by Charles River Laboratories (Australia). In brief, high molecular weight DNA was extracted and Long-read gDNA libraries were prepared with the Oxford Nanopore SQK-LSK110 kit (Oxford Nanopore, UK) which were then sequenced on a single FLO-MIN106D (R.9.4.1) MinION flow cell (Oxford Nanopore, UK). Long reads >1kb were used for genome assembly using the Flye assembler v2.9.1 under the uneven coverage mode ("-meta" option). Assembly QC and short read polishing was performed as described above; Illumina reads were derived from the previous study.[44]

Genome sequencing of *Ca.* A. triatominarum. The high-quality draft genome of *Ca.* Arsenophonus triatominarum was generated using PACBIO long reads from four SMRT cells (C2, P4 chemistry). The reads were filtered using BLAST to remove Drosophila reads and assembled with HGAP using default parameters.[52] The assembly yielded 115 contigs corresponding to the Arsenophonus genome, spanning the total length of 4,721,517 bp with 70x fold average coverage.

Arsenophonus genomes assembled from publicly available SRA deposits

We screened publicly available SRA datasets (Source: DNA, Platform: Illumina, Strategy: genome) originated from the Apoidea superfamily (containing bees and bumblebees) as well as Parasitoida infraorder and Ixodida order for the presence of *Arsenophonus* reads. We performed a "*Mash screen*" using Mash v2.3[53,54] to measure the containment of a local database of reference Arsenophonus genomes within the unassembled SRA read sets. SRA datasets with at least 80% containment were taken for downstream processing. An initial metagenomic assembly of the short reads from the identified SRA datasets were performed using MEGAHIT v1.2.8[55] and the assembled contigs >= 1.5kb were binned based on their deferential tetranucleotide frequencies using MetaBAT2 v2.12.1 under default parameters.[56] The *Arsenophonus* bins were identified and completeness was assessed using CheckM v1.0.18.[57] To identify *Arsenophonus* contigs potentially missed from the initial binning process the original contigs from the metagenomic assembly were screened using blastn (-task megablast) against a local database consisted of all available and complete *Arsenophonus* genomes using BLAST 2.12.0+.[58] Contigs >1kb with significant matches (e-value < 1e-25) to *Arsenophonus* genomes were extracted and included in the metabat bins. The augmented bins were quality inspected and further refined in anvio v7[59] by identifying and removing potential contaminant contigs based on atypical coverage and gene-level taxonomic classification. The original BioProject and SRA accessions from which the draft Arsenophonus genomes were obtained are shown in Table S3.

Comparative analysis of the metabolic potential across the *Arsenophonus* clades

To avoid inconsistencies stemming from draft and incomplete genomes, only the metabolic potential of complete *Arsenophonus* genomes was estimated. To these we included for comparison the genomes of the closely related and obligate symbionts *Ca.* Riesia pediculicola and *Ca.* Riesia pediculischaeffi as well as the genomes of the four outgroup species used in the phylogenetic analysis. All genomes were annotated for functions and metabolic pathways on the basis of the KEGG database using the anvi-run-kegg-kofams command in anvio v7. KEGG MODULE metabolism was finally estimated using the anvi-estimate-metabolism pipeline as described in Muto et al.[89]. A KEGG module was considered complete in a given genome when at least 75% of the steps involved were present.

Phylogenomic analysis and ancestral reconstruction

For the maximum likelihood phylogenomic analysis we excluded the highly divergent genomes from the obligate Arsenophonus strains (Ca. *Arsenophonus lipoptenae*, *Arsenophonus* of *Aleurodicus dispersus*, *Ca.* Arsenophonus melophagi and *Arsenophonus* of *Ceratovacuna japonica*) including *Ca.* Riesia pediculola, as their placement can be affected by strong compositional heterogeneity and long branch attraction. The phylogenetic relationship of the remaining 35 *Arsenophonus* genomes was estimated on the concatenated set of 230 single-copy core protein sequences representing highly conserved gammaproteobacterial BUSCO v4.1.4 markers.[60] These were identified through Orthofinder v2.3.11.[61] Alignments of individual protein orthologs were performed using mafft program[72] as implemented in Orthofinder and screened for recombination based on the Pairwise Homoplasy Index (PHI) test using the Phipack package[62] revealing no significant evidence. The alignments were quality trimmed using ClipKIT alignment trimming tool v1.3.0[63] under the *smart-gap* mode before concatenated into a super matrix using seqkit.[64] Best protein model (JTTDCMut+F+R3) was identified using ModelFinder[66] and a ML phylogenetic tree was reconstructed in IQ-TREE v1.6.12.[65] The genomes from the related species *Proteus mirabilis* strain HI4320 (NC_010554), *Providencia stuartii* strain MRSN 2154 (NC_017731), *Photorhabdus thracensis* strain DSM 15199 (NZ_CP011104) and *Xenorhabdus nematophila* ATCC 19061 (NC_014228) were used as outgroups. All genomes were pre-annotated using prokka v1.13[67] for consistency.

To more precisely estimate the relationships between the *Arsenophonus* strains including the obligate and highly diverse lineages a separate Bayesian phylogenetic analysis was conducted based on the concatenated set of 77 manually curated single-copy core protein clusters using PhyloBayes-MPI v1.9[68] and the CAT-Poisson model. Briefly, the alignments of all single copy protein clusters (101) were manually inspected to minimize alignments with high gap content before concatenation. Two independent chains were run in parallel for at least 30,000 cycles until convergence was observed (rel diff < 0.1 and minimum effective size > 300 for all trace file metrics) assessed by running bpcomp and tracecomp commands in PhyloBayes.

Ancestral reconstruction of gene content across the *Arsenophonus* phylogeny was performed using Count v10.04[22] on the gain-loss-duplication model. As input data we used the Bayesian phylogeny and the OrthoFinder gene families as described above. The initial optimisation of rates was performed on the OrthoFinder results, allowing for different rates of gains, losses and duplications on the tree branches and a Poisson family size distribution at the root (default parameters). A total of 100 optimisation rounds were performed with a convergence threshold of 0.1 (default behaviour).

Annotation and analysis of genomic features

The proportion of pseudogenised genes were estimated by calculating the fraction of interrupted proteins using the Ideel method against the UniProt/Swiss-Prot database (https://github.com/mw55309/ideel). Prophage regions were annotated using the PHAge Search Tool Enhanced Release (PHASTER) web server.[69] The completeness of the prophage-like regions was estimated based on the detailed annotation of phage regions and the identification of genes encoding essential phage functions, such as phage structure, DNA regulation, insertion, and lysis, including the presence of attachment sites. The proportion of insertion sequences in each genome was estimated using ISEScan software v1.7.2.3.[70] Annotation of CRISPR arrays and cas genes was performed with the CRISPRCasFinder program using the default parameters.[71] CRISPR spacer relatedness was calculated as the number of shared spacers by two genomes divided by the number of spacers in the smallest array.[90] Shared spacers were identified based on an all-vs-all blastn search allowing for at least 98% similarity and 97% coverage between individual spacers. ML phylogenetic analyses of Cas protein sequences (Cas1 & Cas6) were performed using IQ-TREE v1.6.12.[65] The protein sequences were previously aligned using mafft.[72] Branch support was assessed using the ultrafast bootstrap approximation method as implemented in IQ-TREE with 1,000 replicates.[91] Congruence between tree topologies was assessed using the approximately unbiased (AU) test[92] as implemented in IQ-TREE. Finally, a phylogenetic network was reconstructed from the protein alignment of Cas1 homologs using the split decomposition method with SplitsTree version 4.19.0.[73] Apart from CRISPR-Cas systems we screened for additional phage defense systems using DefenseFinder.[74] Finally, we assessed the relatedness between the *Arsenophonus* MGEs (intact chromosomal phages and extrachromosomal elements) and their relationship with previously described phages and plasmids by calculating the average nucleotide identity (ANI) using FastANI v1.33.[75] To this end a collection of 4568 NCBI phage genomes and 59781 plasmids genomes from the PLSDB plasmid database v. 2023_11_03_v2[93] was used. Run parameters were as follows: kmer size of 16, fragment length of 1000 and minimum fraction of shorter genome coverage of 50%. Results were considered significant only if at least 50% of the query fragment was shared with the reference sequence. The raw results of the FastANI analysis are shown in Table S4. The network graph was generated in R using the igraph package v1.6.0[76] and the graphopt layout algorithm (niter=10000, charge = 0.02).

Analysis of relaxation of selection strength between *Arsenophonus* clades

We searched for signatures of relaxation of selection on a set of 188 highly conserved single-copy BUSCO orthologs across 39 *Arsenophonus* strains using the HyPhy (Hypothesis Testing using Phylogenies) software package version v2.5.8.[77] The obligate and highly diverged *Arsenophonus* strains including *Ca.* Riesia were excluded from the analyses. Four, mostly environmental, Morganellaceae (Enterobacterales) species (*Proteus mirabilis*, *Providencia stuartii*, *Morganella morganii* and *Moellerella wisconsensis*) were used as reference/outgroup since *Photorhabdus thracensis* and *Xenorhabdus nematophila* that were used in the phylogenetic analyses above have complex modes of transmission involving both vertical and horizontal transmission. To this end, single-copy protein clusters were identified as before using Orthofinder v2.3.11 and 188 clusters representing highly conserved gammaproteobacterial BUSCO v4.1.4 markers were selected for downstream analyses. Alignment of protein sequences was performed with mafft using the "-auto" option and back translated to nucleotide alignments using pal2nal.[78] A phylogenetic tree was estimated using the JTTDCMut+F+I+G4 model in IQ-TREE v1.6.12 on the concatenated data of the same set of 188 orthologues protein clusters. This tree was then used to identify genes with significant evidences of relaxation or intensification of selection using the RELAX hypothesis testing framework in HyPhy package.[94] Three sets of branches were selected as the "test branches" (nasoniae clade, apicola clade and the triatominarum clade, see Figure S10) and compared to the reference/outgroup set of branches to estimate the selection intensity parameter k for each gene. A value of $k > 1$ indicates that selection on the test branches is intensified compared to the reference branches while a value < 1 indicates a relaxation of selection. Statistically significant values were assessed through a likelihood ratio test (LRT) followed by a false discovery rate (fdr) correction to account for multiple comparisons. Branch specific dN/dS ratios were estimated for each individual gene using the partitioned MG94xREV model which fits a single dN/dS value to each branch partition as implemented in RELAX method in HyPhy. The differences of dN/dS ratios between branches (ΔdN/dS) were statistically assessed using a Wilcoxon signed-rank test in R v4.2.2.[79]

Data visualization

Tools used for data visualization: R v4.2.2,[79] ggplot2,[80] patchwork,[81] igraph v1.6.0.[76] A presence/absence heatmap for metabolic pathways was generated using the pheatmap v1.0.12 (https://rdrr.io/cran/pheatmap/) package in R v4.2.2. Phylogenetic trees were drawn and annotated using the ggtree package v4.2.[82] The gene order comparison of CRISPR-Cas systems was visualized in clinker v0.0.24.[83]

Please cite this article in press as: Siozios et al., Genome dynamics across the evolutionary transition to endosymbiosis, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.10.044

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analyses were carried out in R v4.2.2.[79] A non-parametric Mann-Whitney U test [function: *wilcox.test()*] was employed to assess the significance of the genome size difference observed between the cluster of nasoniae clade strains with vertical transmission and the strains from honey bees and the strain from a butterfly which do not have vertical transmission. A regression was carried out to assess the relationship between genome size and GC content by fitting a linear model using the function *lm()*. Further statistical details for each test can be found in the main text and figure legends. For every statistical analysis significance was defined at $p<0.05$.