Origin and Evolution of the *Bartonella* Gene Transfer Agent

Daniel Tamarit,[1] Minna-Maria Neuvonen,[1] Philipp Engel,[2] Lionel Guy,[†,1] and Siv G.E. Andersson[*,1]

[1]Department of Molecular Evolution, Cell and Molecular Biology, Science for Life Laboratory, Biomedical Centre, Uppsala University, Uppsala, Sweden

[2]Department of Fundamental Microbiology, University of Lausanne, Lausanne, Switzerland

[†]Present address: Department of Medical Biochemistry and Microbiology, Biomedical Centre, Uppsala University, Uppsala, Sweden

[*]**Corresponding author:** E-mail: siv.andersson@icm.uu.se.

Associate editor: Eduardo Rocha

Abstract

Gene transfer agents (GTAs) are domesticated bacteriophages that have evolved into molecular machines for the transfer of bacterial DNA. Despite their widespread nature and their biological implications, the mechanisms and selective forces that drive the emergence of GTAs are still poorly understood. Two GTAs have been identified in the Alphaproteobacteria: the RcGTA, which is widely distributed in a broad range of species; and the BaGTA, which has a restricted host range that includes vector-borne intracellular bacteria of the genus *Bartonella*. The RcGTA packages chromosomal DNA randomly, whereas the BaGTA particles contain a relatively higher fraction of genes for host interaction factors that are amplified from a nearby phage-derived origin of replication. In this study, we compare the BaGTA genes with homologous bacteriophage genes identified in the genomes of *Bartonella* species and close relatives. Unlike the BaGTA, the prophage genes are neither present in all species, nor inserted into homologous genomic sites. Phylogenetic inferences and substitution frequency analyses confirm codivergence of the BaGTA with the host genome, as opposed to multiple integration and recombination events in the prophages. Furthermore, the organization of segments flanking the BaGTA differs from that of the prophages by a few rearrangement events, which have abolished the normal coordination between phage genome replication and phage gene expression. Based on the results of our comparative analysis, we propose a model for how a prophage may be transformed into a GTA that transfers amplified bacterial DNA segments.

Key words: **gene transfer agent, bacteriophage, horizontal gene transfer, *Bartonella*.**

Introduction

Nearly half of all sequenced bacterial genomes contain integrated prophages, with lysogeny being most frequent in small and slowly growing bacterial cells (Touchon et al. 2016). The size of 300 complete and defective prophage segments in enterobacterial genomes follows a bimodal distribution, with one peak represented by intact, functional prophages in the 30–70 kb range, and another by defective prophages in the 5–30 kb range (Bobay et al. 2014). Some of the defective phages provide beneficial functions to the bacterium, for example protecting the bacterium against further phage infections, or by increasing its tolerance to antibiotics (Waldor and Friedman 2005; Wang et al. 2010; Rabinovich et al. 2012).

Another adopted function that has recently gained attention due to its widespread occurrence is the transfer of genomic DNA by gene transfer agents (GTAs), which are bacteriophage-like particles that mediate the transfer of chromosomal DNA between bacterial cells (Lang et al. 2012; Soucy et al. 2015). Given their biological implications, the origin and evolution of GTAs constitute important evolutionary questions for which little evidence is available. It is generally admitted that GTAs originate from bacteriophages, but no close phage relatives have been identified that could help understand the mechanisms and selective forces driving the transformation of a prophage into a GTA.

The best-studied GTA (RcGTA) is encoded by the genome of *Rhodobacter capsulatus*, a member of the Rhodobacterales in the Alphaproteobacteria. An unusual mechanism of genetic exchange via a novel type of vector was discovered in *R. capsulatus* already four decades ago (Marrs 1974; Rapp and Wall 1987; Humphrey et al. 1997). Since then, it has been shown that genes encoding the RcGTA are located in one large structural gene cluster and several smaller clusters that are regulated by quorum sensing (Schaefer et al. 2002; Brimacombe et al. 2013; Hynes et al. 2016). The RcGTA particles are produced by only a few percent of the bacterial cells, and these die after expression and release of the RcGTA (Fogg et al. 2012; Hynes et al. 2012). Genes that are homologous to the RcGTA have been detected in several alphaproteobacterial lineages, consistent with a long coevolutionary history between the RcGTAs and their hosts (Lang and Beatty 2007).

A novel GTA (BaGTA) was discovered in *Bartonella*, a genus in another order of the Alphaproteobacteria (Rhizobiales) that do not contain the RcGTA. *Bartonella* infect endothelial cells and erythrocytes of mammals with the aid of type IV and type V secretion systems (Engel et al. 2011; Eicher and Dehio 2012). The genus *Bartonella* has undergone an explosive radiation and colonizes a broad range of mammalian hosts, possibly aided by adaptive evolution of the secretion systems and their effector molecules to match a divergent set of host cells (Chomel et al. 2009; Engel et al. 2011; Guy et al. 2013).

Article

Intriguingly, the BaGTA gene cluster is highly conserved, it is present in the same genomic location in all *Bartonella* spp., and is flanked by various gene clusters for type IV and type V secretion systems (Alsmark et al. 2004; Berglund et al. 2009; Guy et al. 2013).

The chromosomal region that contains the genes for the BaGTA and the secretion systems was shown to be present in higher copy numbers than genes located elsewhere in the genomes of the cat-associated species *Bartonella henselae*, with the peak of the amplification located at a putative phage replication initiation site (Lindroos et al. 2006). It was thus suggested that the higher copy numbers were due to replication from an alternative origin derived from a defective prophage, in a process referred to as run-off replication. Experimental studies performed in the rodent-associated species *Bartonella grahamii* confirmed that the BaGTA particle contains random fragments of bacterial DNA, with an overrepresentation of sequences flanking the phage replication initiation site (Berglund et al. 2009). Subsequently, it was demonstrated that the chromosomal DNA that is packaged into the GTA could be transferred from donor to recipient *B. henselae* strains (Guy et al. 2013). More recently, it was shown that expression of the BaGTA is normally repressed by ppGpp and only induced in a subset of cells under rapid growth conditions when the levels of ppGpp are low (Quebatte et al. 2017). Uptake of the particles is restricted to actively dividing cells and dependent on the competence and recombination machinery of the bacterial host cell (Quebatte et al. 2017). It has been proposed that the BaGTA increases the rate of transfer and recombination of beneficial mutations in genes for host-adaptability factors that accumulate in the vicinity of the *Bartonella* run-off replication (BaROR) gene cassette (Berglund et al. 2009; Guy et al. 2013).

The recent emergence of the BaGTA within a restricted set of species in the Rhizobiales lacking the RcGTA, combined with the identification of homologous bacteriophage genes, provides an opportunity to study the mutational events associated with the recent conversion of a bacteriophage into a GTA that is beneficial for the bacterial population. The transfer of phage-replicated DNA segments by the BaGTA has been regarded as a primitive feature since replication and transfer of short DNA fragments are typical bacteriophage properties. However, in *Bartonella* it is bacterial genes rather than phage genes that are amplified prior to transfer by the GTA. Given the implications of GTAs for bacterial population models, it is of general interest to learn more about their emergence. In this study, we compare the BaGTA with homologous bacteriophage genes in *Bartonella* and propose a model for the evolution of GTAs with specialized bacterial functions.

Results and Discussion

Phyletic Distribution Patterns of the RcGTA and BaGTA

The genes for the RcGTA in *R. capsulatus* code for structural phage proteins such as the capsid and tail proteins as well as enzymes involved in the assembly of the particle (fig. 1A). Likewise, the BaGTA gene cluster in *Bartonella australis* NH1 (here used as a representative species of the eubartonellae—the radiating *Bartonella* clade; Zhu et al. 2014) codes for structural phage proteins as well as enzymes involved in the assembly of the phage particle (fig. 1B).

To assess the potential overlap in the phyletic distribution profiles of these two GTAs, we searched for RcGTA and BaGTA homologs in all species in the Rhizobiales, Rhodobacterales, and Caulobacterales for which complete genome data are available using two iterative PSI-BLAST searches (supplementary tables S1 and S2, Supplementary Material online). As also shown previously (Lang and Beatty 2007), the RcGTA gene cluster (more than ten consecutive genes) is widely distributed in these bacterial orders. Previous single-gene phylogenies inferred from fewer taxa have indicated that the RcGTA has codiversified with the bacterial genome (Lang et al. 2002; Lang and Beatty 2007). Consistently, a maximum likelihood phylogeny inferred from a concatenated alignment of seven RcGTA genes yielded a tree topology that matched the bacterial phylogeny (supplementary fig. S1, Supplementary Material online). Thus, the occurrence pattern of the RcGTA indicates vertical inheritance with multiple independent losses (fig. 1C and supplementary table S1, Supplementary Material online).

In contrast to the wide distribution of the RcGTA in the Alphaproteobacteria, the BaGTA gene cluster (more than ten consecutive genes) was only identified in the Bartonellaceae family, and shorter fragments were found in a very small number of genomes in other families (fig. 1C and supplementary table S2, Supplementary Material online). Although all members of the Bartonellaceae family with sequenced genomes contained clusters with hits to over five BaGTA genes, no trace of the RcGTA could be identified in any of them. Clusters of eight to ten homologous genes syntenic with the BaGTA were detected in *Rhodopseudomonas palustris* TIE-1, *Azorhizobium caulinodans*, and *Methylobacterium radiotolerans* (fig. 1C and supplementary fig. S2, Supplementary Material online), all three of which also contain the RcGTA. The identified BaGTA homologs in these three species were embedded in regions that also contained integrases, phage replication genes, and other phage-associated genes. Since the genes were not conserved among close relatives, and given their genetic context, we hypothesize that their behaviour is representative of integrated prophages, rather than GTAs.

In order to gain a deeper understanding of the evolution of the BaGTA structure within Bartonellaceae, we performed additional PSI-BLAST searches in a group of genomes that contained incomplete and recently published genomes (supplementary table S3, Supplementary Material online). We identified intact BaGTA gene clusters in all eubartonellae (fig. 2A and supplementary table S4, Supplementary Material online) as well as in the earlier diverging human pathogen *Bartonella tamiae* strain Th307 (Kosoy et al. 2008) and in the honeybee gut symbiont *Bartonella apis* (Kesnerova et al. 2016). Additionally, we identified homologs of seven BaGTA genes in the genome of the ant-associated species *Candidatus* Tokpelaia hoelldoblerii

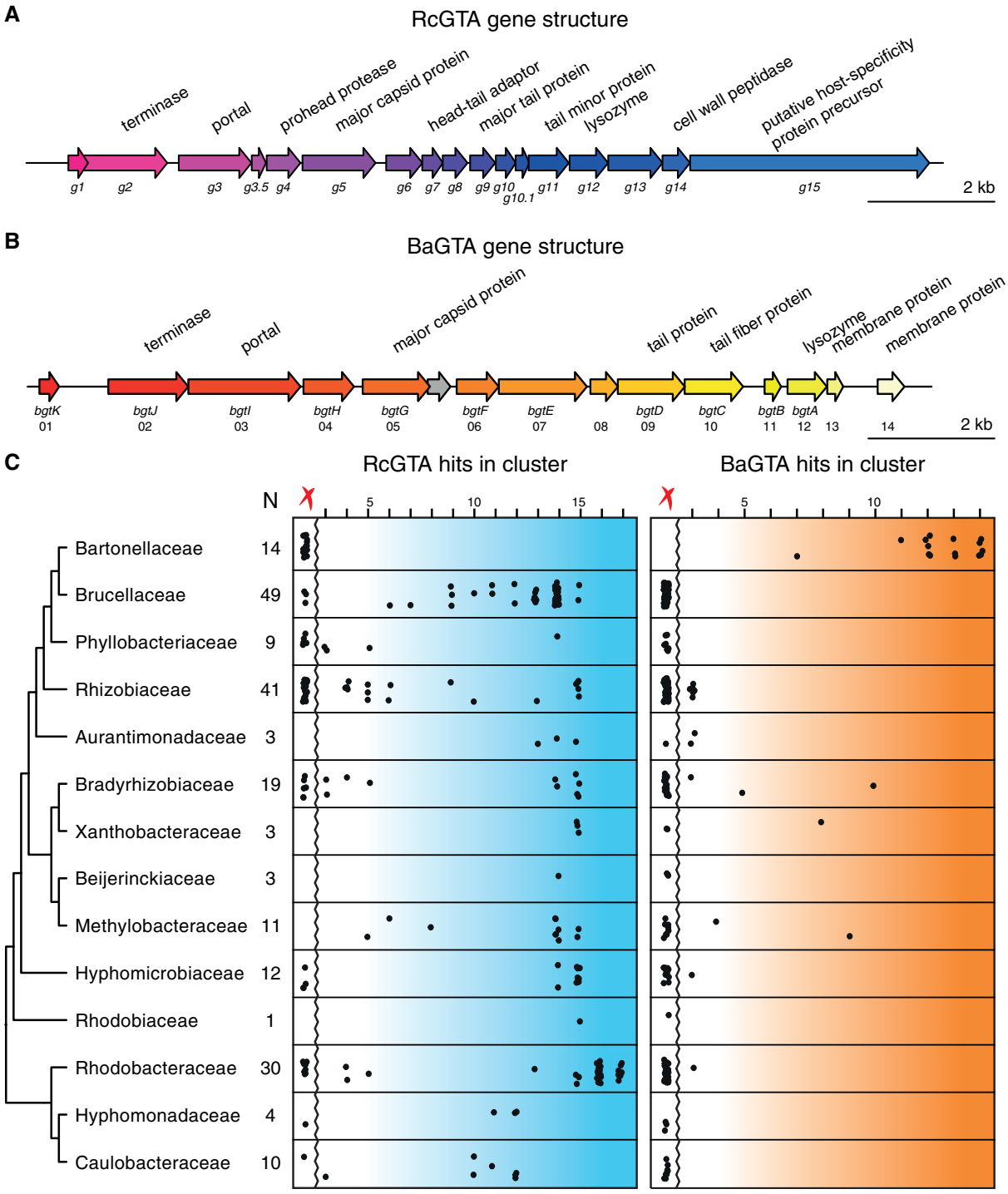

Fig. 1. Phyletic distribution pattern of the RcGTA and the BaGTA. Gene order structure of the (*A*) RcGTA in *Rhodobacter capsulatus* (Lang and Beatty 2001; Lang et al. 2012) and (*B*) BaGTA in *Bartonella australis* NH1 (Guy et al. 2013), drawn using GenoPlotR (Guy et al. 2010). Arrows represent genes, which in (*B*) are numbered and named as in supplementary table S4, Supplementary Material online. (*C*) Number of genes within a region (three or more syntenic hits at < 25 kb distance) showing similarity to the RcGTA and BaGTA gene products in complete genomes of the orders Rhizobiales, Rhodobacteriales, and Caulobacterales (supplementary tables S1 and S2, Supplementary Material online). Only the region with most hits per genome is shown. The leftmost boxes ("X" mark) represent genomes with no identified GTA region, defined as above. The column next to the phylogeny shows the number of analyzed genomes per family (N). The tree topology was taken from Viklund et al. (2012).

(Neuvonen et al. 2016). This shorter gene cluster contained genes for capsid proteins but no genes for tail proteins (fig. 2A).

We also examined the phyletic distribution pattern of the BaROR gene cluster, which contains six contiguous genes in *B. australis* (fig. 2B) including BaROR-04 (*brrC*), which is thought to contain the phage-derived origin of replication. Homologs to the first (*brrH*) and the fourth (*brrC*) gene in the BaROR gene cluster were identified in eubartonellae species

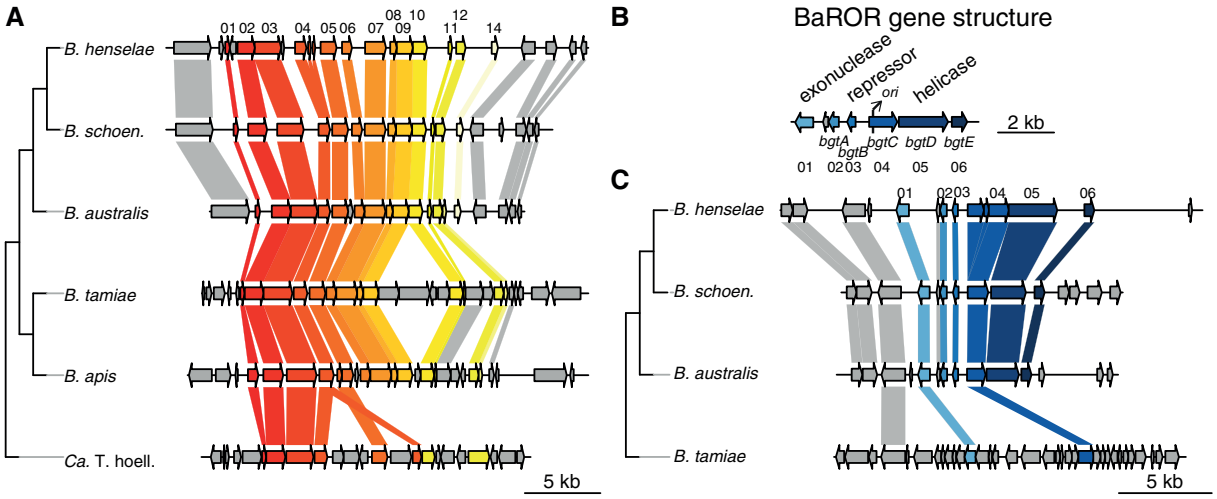

FIG. 2. Comparison of the BaGTA and BaROR gene clusters. (*A*) Comparison of the BaGTA gene cluster. (*B*) Gene order structure of the BaROR region in *Bartonella australis*. (*C*) Comparison of the BaROR gene cluster. Arrows represent genes numbered and named as in supplementary table S4, Supplementary Material online. Genes in (*A*) are colored as in figure 1*B*. Connecting lines represent reciprocal PSI-BLAST hits after two iterations with an E-value <0.001. Species abbreviations: *B. schoen.*, *Bartonella schoenbuchensis*; T. hoell., *Candidatus* Tokpelaia hoelldoblerii. All gene maps were plotted using genoPlotR (Guy et al. 2010).

as well as in *B. tamiae* Th307, but not in *B. apis* or *Ca.* T. hoelldoblerii (fig. 2C). In eubartonellae, the BaROR genes are located in close proximity to the BaGTA genes, at a distance of <60 kb. The two homologs in *B. tamiae* Th307 were also located in the vicinity of the BaGTA genes within a genomic segment that showed overall synteny with *B. australis* and the other *Bartonella* spp. (supplementary fig. S3, Supplementary Material online). In comparison, the genome of *Ca.* T. hoelldoblerii is highly rearranged (Neuvonen et al. 2016), and the genes flanking the BaGTA have homologs elsewhere in the genomes of the eubartonellae (supplementary fig. S4, Supplementary Material online).

The BaGTA Has Codiversified with the *Bartonella* Genome

To test the hypothesis that the BaGTA genes have codiversified with their host bacterial genome, like the RcGTA, we performed a phylogenetic analysis from a concatenated alignment of the gene products of *bgtJIHG* (BaGTA02-05) (fig. 3A) as well as from single protein alignments of all BaGTA genes obtained from the selected genomes (supplementary table S3, Supplementary Material online) and public databases (supplementary fig. S5, Supplementary Material online; summarized in fig. 3B). The four genes selected for the phylogenetic analysis were conserved in both presence and order in all species including *Ca.* T. hoelldoblerii, consistent with the hypothesis of a shared evolutionary history. The concatenated protein tree confirmed that the BaGTA homologs form a distinct monophyletic clade with 100% bootstrap support in the maximum likelihood analysis and with a posterior probability of 1 in the Bayesian analysis (fig. 3A), consistent with a shared, common origin. Furthermore, the tree indicated a series of expansion events in the ancestor of the eubartonellae, resulting in three paralogous BaGTA-like lineages each of which is highly supported: the BaGTA clade itself, plus two additional groups that we have here named BaGTA-like 1 (BGL1) and BaGTA-like 2 (BGL2). Consistently, despite reduced phylogenetic signal, the large majority of the single protein trees supported the presence of one BaGTA (10 out of 14 trees) and two BaGTA-like clades (9 out of 12 trees for both clades) (fig. 3B and supplementary fig. S5, Supplementary Material online).

The diversification pattern within the BaGTA clade (fig. 3A) matched the species topology, especially regarding the monophyly and inner topology of the major phylogroups (Guy et al. 2013), suggesting that the BaGTA genes have codiversified with the host bacterial genome. Moreover, the general congruency of the deeper branches of the BaGTA phylogeny with recent phylogenomic reconstructions of bacterial sequences (Neuvonen et al. 2016; Segers et al. 2017) indicates that the association between the genes that originated the BaGTA and the bacterial host goes back to, at least, the Bartonellaceae common ancestor.

The tree further indicated that the two BGL clades are sister groups with full posterior probability and >80% bootstrap support in the concatenated protein tree (fig. 3A), as also indicated by 4 of the 12 single protein trees (fig. 3B and supplementary fig. S5, Supplementary Material online). However, the BGL1 and BGL2 clades contained only subsets of the *Bartonella* species and, in the BGL1 clade, *B. tribocorum* clustered with *B. henselae* instead of with its closest relative *B. grahamii*, as in the BaGTA clade (fig. 3A and supplementary fig. S5, Supplementary Material online). The BGL2 genes could only be identified in *B. grahamii* and *B. tribocorum*. There were two copies of BGL2 in *B. tribocorum*, representing the only case in our data set of two coresident copies of a single type, probably generated by a recent duplication or integration event. Notably, both BGL1 and BGL2 were absent from *B. quintana*, *B. bacilliformis*, *B. vinsonii*, and *B. australis*. These patterns indicate duplication and diversification of the BGL genes followed by independent losses and occasional horizontal transfers of the BGL gene clusters. The BGL2

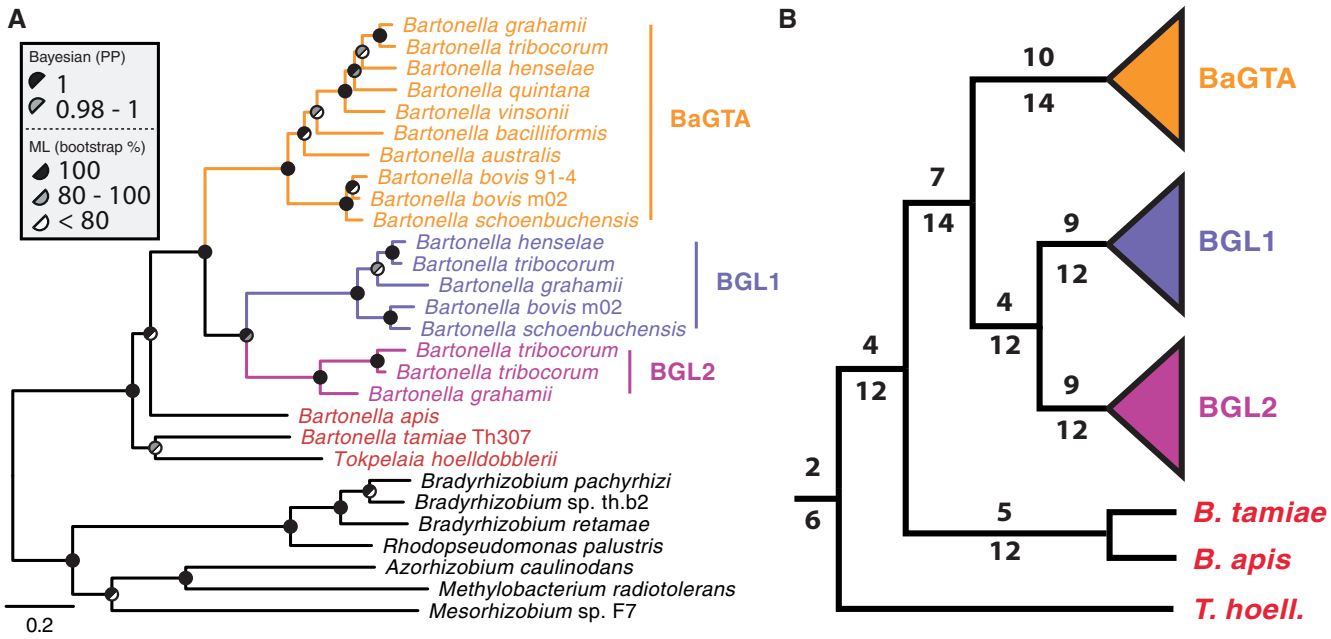

FIG. 3. Phylogeny of the BaGTA genes. (*A*) Phylogenetic analysis based on a concatenated alignment of the products of genes *bgtJIHG* (BaGTA-02 to BaGTA-05) and homologs thereof, using Phylobayes and RAxML. Taxa were selected from the local genomes (supplementary table S3, Supplementary Material online) and the outgroups containing hits for all the selected proteins (supplementary fig. S5, Supplementary Material online). The topology of the tree is taken from the Phylobayes reconstruction. Circles in nodes represent posterior probabilities (higher half) and bootstrap support (lower half) with values coded as shown in the inset. (*B*) Summary of the phylogenetic analyses performed on the BaGTA genes and their homologs. The given topology represents the consensus of well-supported groups taken from the single-gene trees. Above each branch is shown the number of single-gene trees with over 70% bootstrap support for the defined monophyletic external group; and below each branch, the total number of trees that can contribute to the given group (i.e., that include taxa belonging to that group). BGL, BaGTA-like; T. hoell., *Candidatus* Tokpelaia hoelldoblerii.

occurrence pattern could also be explained by a separate integration event of the BGL ancestor in the lineage leading to *B. grahamii* and *B. tribocorum*. Moreover, contrary to the conserved location of the BaGTA, the insertion sites for the BGL1 and BGL2 regions differed among the genomes where they were detected (see below). Therefore, we conclude that neither BGL1 nor BGL2 have codiversified with the host bacterial genome.

Consistent with long-term coevolutionary processes for the BaGTA genes, the GC content at third codon synonymous sites (GC3s) was more similar to the GC content of the bacterial genome in which they are located (ca. 30% for *Bartonella* and ca. 60% for *Ca.* T. hoelldoblerii) than they were to each other (supplementary table S5, Supplementary Material online). However, the three extra genes in *Ca.* T. hoelldoblerii, which are inserted between the fourth and fifth genes in the BaGTA cluster (fig. 2*A*), displayed lower GC3s (between 25% and 40%) than the BaGTA gene homologs, possibly indicating more recent acquisitions. Likewise, a gene for a lysozyme at the 3′-end of the cluster is more AT-rich than the core structural genes, suggesting that also this gene may have been added to the cluster subsequent to its formation.

Surprisingly, no BaGTA region could be identified in *B. tamiae* strain Th239. However, a closer inspection of the positional homologs in this strain using PSI-BLAST against a data set of alphaproteobacterial genomes showed a weak similarity to the RcGTA (supplementary fig. S6, Supplementary Material online). Phylogenies inferred from the putative terminase and portal proteins revealed exceptionally long branch lengths for the *B. tamiae* Th239 sequences, with no affiliation to any of the two GTA groups (supplementary fig. S7, Supplementary Material online). However, manual inspection of the portal protein alignment showed a few short stretches (<10 aa) with sequence identity to the RcGTA (supplementary fig. S8, Supplementary Material online). The phage portal and phage prohead protease protein sequences belong to families of conserved structural proteins that are widely used across tailed phages. The *B. tamiae* genes may thus represent a distantly related or rapidly evolving prophage. Although the functionality and evolutionary origin of this region is unclear, a possible GTA replacement event is likely to have occurred.

Vertical Inheritance of the BaROR

Next, we examined the occurrence and genomic location of the region containing the phage-derived origin of replication (BaROR). The BaROR region had been found to contain six conserved genes (Guy et al. 2013), which here we refer to as BaROR-01 to -06 (supplementary table S4, Supplementary Material online). We used PSI-BLAST searches to detect *Bartonella* regions with homology to these six genes, and found several such cases. A comparison of the gene neighborhoods of the BaROR region and its newly found homologs

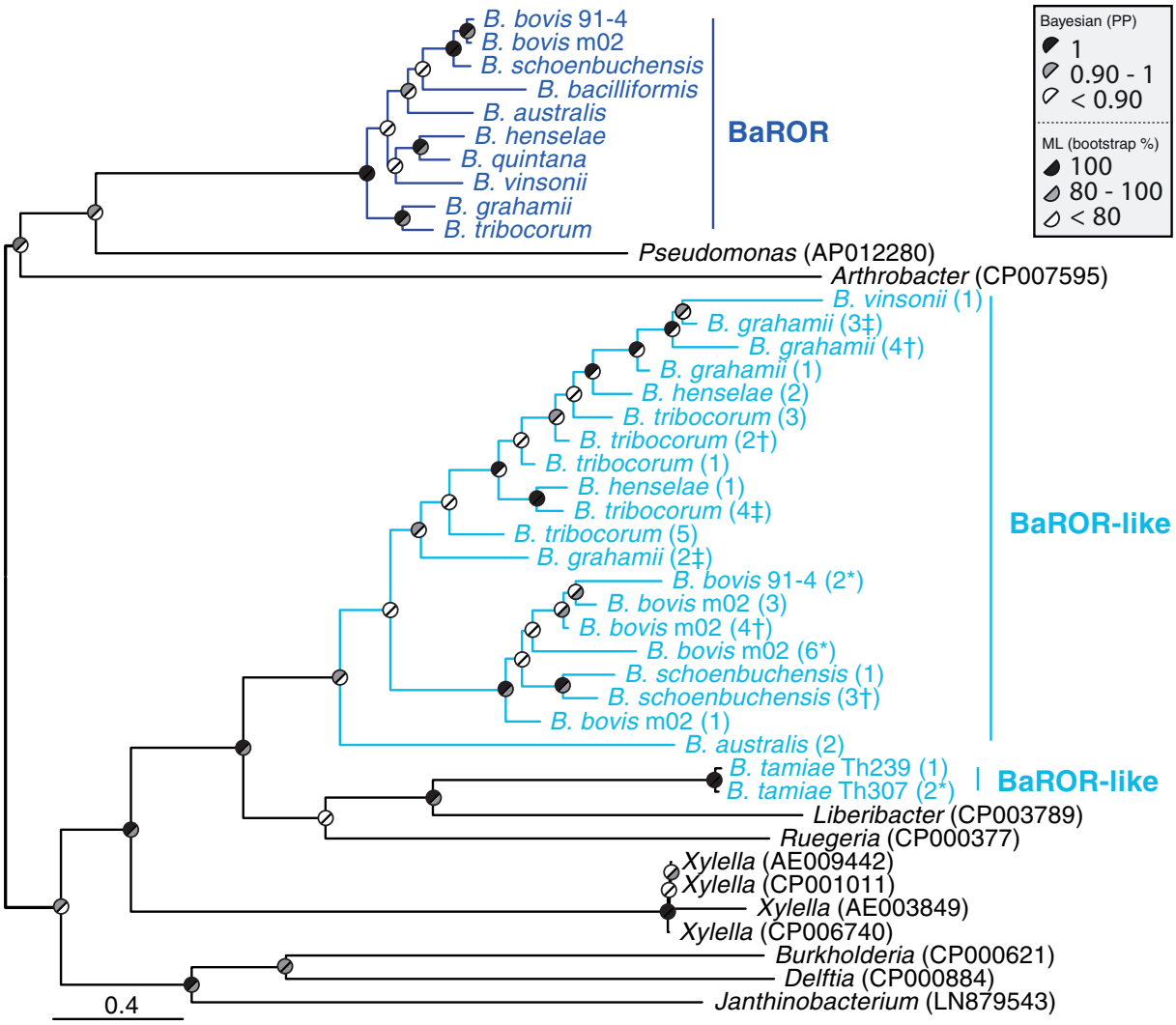

Fig. 4. Phylogeny of the BaROR and BaROR-like gene clusters showing their distinct origins. Phylogenetic tree constructed from a concatenated alignment of the proteins BaROR-01 and BaROR-04 and homologs thereof, using Mr. Bayes and RAxML. The topology of the tree is as taken from Mr. Bayes. Circles in nodes represent posterior probabilities (higher half) and bootstrap support (lower half) with values as shown in the inset. Next to the BaROR-like external nodes, numbers represent the order in which that sequence appears in the genome, as showed in supplementary figure S9, Supplementary Material online. Symbols indicate proximity to a BaGTA (*), BGL1 (†), or BGL2 (‡) region.

revealed frequent linkage of BaROR-01, BaROR-02, and BaROR-04, with tandem duplications of the latter gene in several cases (supplementary table S6 and fig. S9, Supplementary Material online). A short seventh gene was found to be present in all surveyed *Bartonella* genomes between BaROR-01 and BaROR-02, thus possibly constituting an additional BaROR gene. However, the phylogenetic signal in this and all the other BaROR genes was insufficient to yield conclusive single-gene phylogenies (supplementary fig. S10, Supplementary Material online).

A phylogenetic inference based on a concatenated alignment of BaROR-01 and BaROR-04 provided support for two distinct lineages: the BaROR clade containing genes located in the vicinity of the BaGTA in the canonical *Bartonella* species, and another paraphyletic clade, here called BaROR-like, which consisted of all the homologous gene copies located elsewhere in these genomes (fig. 4). The average branch length from the common ancestor of the eubartonellae to the tips of the branches was ~3-fold higher for the BaROR-like clade compared with the BaROR clade (distances = 0.85 vs. 0.30, Standard Deviations 0.23 and 0.08), suggesting relaxed selective constraints in the former. The diversification pattern within the BaROR-like clade did not match the species divergence patterns, indicative of multiple integration events. For example, the phylogeny showed that the single copy of BaROR-like region identified in *B. tamiae* strain Th307 did not cluster with the BaROR clade, but with the BaROR-like clade and with phage sequences from *Liberibacter* (>80% bootstrap support in the maximum likelihood tree) (fig. 4). This suggests that the BaROR homologs might have either been transferred to other alpha-proteobacteria from *Bartonella*, or that the BaROR homologs in *B. tamiae* have a different origin than both the BaROR and the BaROR-like genes in the other *Bartonella* species.

Previous studies have shown that the higher recombination rates caused by the local amplification near the BaROR regions causes GC-biased gene conversion (Berglund et al. 2009; Guy et al. 2013). However, to the contrary of what is seen in other *Bartonella* genomes, no general increase in GC3s values was observed in the genomic region flanking the BaROR-like genes in *B. tamiae* (supplementary fig. S11, Supplementary Material online). This suggests that the surrounding segments may not be amplified and thereby not be recombining more frequently than the rest of the genome. Thus, if the GTA-like region in *B. tamiae* strain Th307 does have a function that transfers host DNA, it is most likely a generalist GTA that transfers host DNA randomly, without prior amplification of specific regions in the genome. However, it still remains to be experimentally shown that the GTA-like region in *B. tamiae* acts as a GTA.

BGL1 and BGL2 Segments

Although the genomic location of BGL1 and BGL2 was not conserved, these were in most cases in the vicinity of BaROR-like genes (fig. 5). We compared the organization of the segments flanking BGL1 and BGL2 in *B. grahamii* and *B. schoenbuchensis*, which indicated the presence of five homologous gene blocks, which are structured in the same order and orientation in BGL1 and BGL2 in the two species (fig. 6). Based on the conserved order of the five gene blocks, we suggest that the gene neighborhoods of BGL1 and BGL2 represent the order of genes in their shared bacteriophage ancestor. In comparison, the segments flanking the BaGTA have been rearranged, potentially abolishing a coordinated regulation of the initiation of phage replication and the expression of phage structural genes.

We noted that the genes in region BGL2 in *B. grahamii* are separated by multiple short insertions, including two genomic islands that code for the hemolysin activator protein HecB and the filamentous hemagglutinin protein FhaC, which is a two-component type V secretion system, together with phage proteins and integrases. One of the islands has been inserted into the homolog of *bgtF* (BaGTA-06) for the major capsid protein and the second has been inserted into the intergenic region between the homologs of *bgtE* (BaGTA-07) and BaGTA-08. A third genomic island with genes for filamentous hemagglutinin, has been inserted inside a cassette of unknown function immediately upstream of BGL2 (fig. 6).

The genes for HecB/FhaC have been identified as highly variable in their chromosomal presence/absence patterns, being absent from about one-third of 27 *B. grahamii* strains isolated at sites located within 30 km from each other (Berglund et al. 2009). Interestingly, 8 of 11 strains that lack the *hecB/fhaC* genes show a higher hybridization signal over the BGL2 region, indicating that this region and the other surrounding phage genes may be replicating and reintegrating (Berglund et al. 2009). In contrast, strains that contain the *hecB/fhaC* insertion showed no indications of any gene copy number variation (Berglund et al. 2009). Genes for secretion systems may have been inserted into these phage-like regions due to the lack of counter-selection, or driven by selection to inactivate the prophage-induced lytic cycle.

Nucleotide Sequence Divergence

Prophage genes are normally less conserved than bacterial core genes, due to recombination and high rates of nucleotide substitutions and insertion–deletions. To test the hypothesis that the BaGTA genes evolve under purifying selection similar to bacterial core genes, whereas BGL1 and BGL2 evolve more like prophages, we estimated the nonsynonymous (dN) and synonymous (dS) substitution frequencies for the most closely related pairs of genomes for which the dS values were not saturated ($dS < 1$). These include pairwise comparisons of *B. henselae*, *B. grahamii*, and *B. tribocorum* and of *B. schoenbuchensis* and *B. bovis*. The dS values were similar for all genes of the BaGTA and not significantly different to the dS values of single-copy core genes present in all *Bartonella* species (Mann–Whitney test, $P > 0.15$ for all pairwise comparisons including *B. grahamii*, *B. tribocorum*, and *B. henselae*) (supplementary fig. S12*A*, Supplementary Material online). Only the comparison between the GTA and panorthologs dS values in *B. schoenbuchensis* and *B. bovis* yielded a significant difference (Mann–Whitney test, $P = 0.039$), with the GTA values being slightly lower than those of the panorthologs, although this difference was not found after correcting for multiple testing (Holm–Bonferroni-corrected $P = 0.156$).

Furthermore, the dN/dS ratios for the BaGTA genes were also comparable with those of the single-copy core genes (Mann–Whitney test, *P* values for the four comparisons >0.25) (supplementary fig. S12*B*, Supplementary Material online). The BaGTA genes with higher dN/dS values corresponded to shorter genes, and no gene longer than 750 bp (over half of the BaGTA genes) showed a dN/dS >0.12. This negative correlation between dN/dS and gene length ($P = 1.4\times10^{-4}$ and 1.9×10^{-6} for Kendall's and Spearman's rank correlations, respectively) may be caused by uncertainty in the calculation of the higher dN/dS values (0.11–0.30) for the short genes. Thus, in each pairwise comparison the substitution frequencies for the BaGTA genes were as expected for genes that have coevolved with the bacterial genome since the two species diverged.

In comparison, the dS values of the genes in the BGL1 region showed much more variability (supplementary fig. S12*A* and table S7, Supplementary Material online). For example, the unsaturated dS values between *B. grahamii* and *B. tribocorum* ranged from 0.3 for the capsid gene (*bgtG*—BaGTA-05) up to 0.9 substitutions per synonymous site for the portal gene, with the dS values for four genes being well above ($dS > 0.6$) the median genomic dS value of 0.35 substitutions per synonymous site. On the other extreme, the dS value was only 0.19 for the portal gene in BGL1 in *B. henselae* and *B. tribocorum*, and four out of five dS values were well below the median genomic dS value of 0.8 substitutions per synonymous site for this pair of strains. Furthermore, the comparisons between *B. henselae* and *B. tribocorum* (Mann–Whitney test, $P = 0.040$), and *B. tribocorum* and *B. grahamii* (Mann–Whitney test, $P = 0.014$) showed significantly lower and higher dS values, respectively, than the

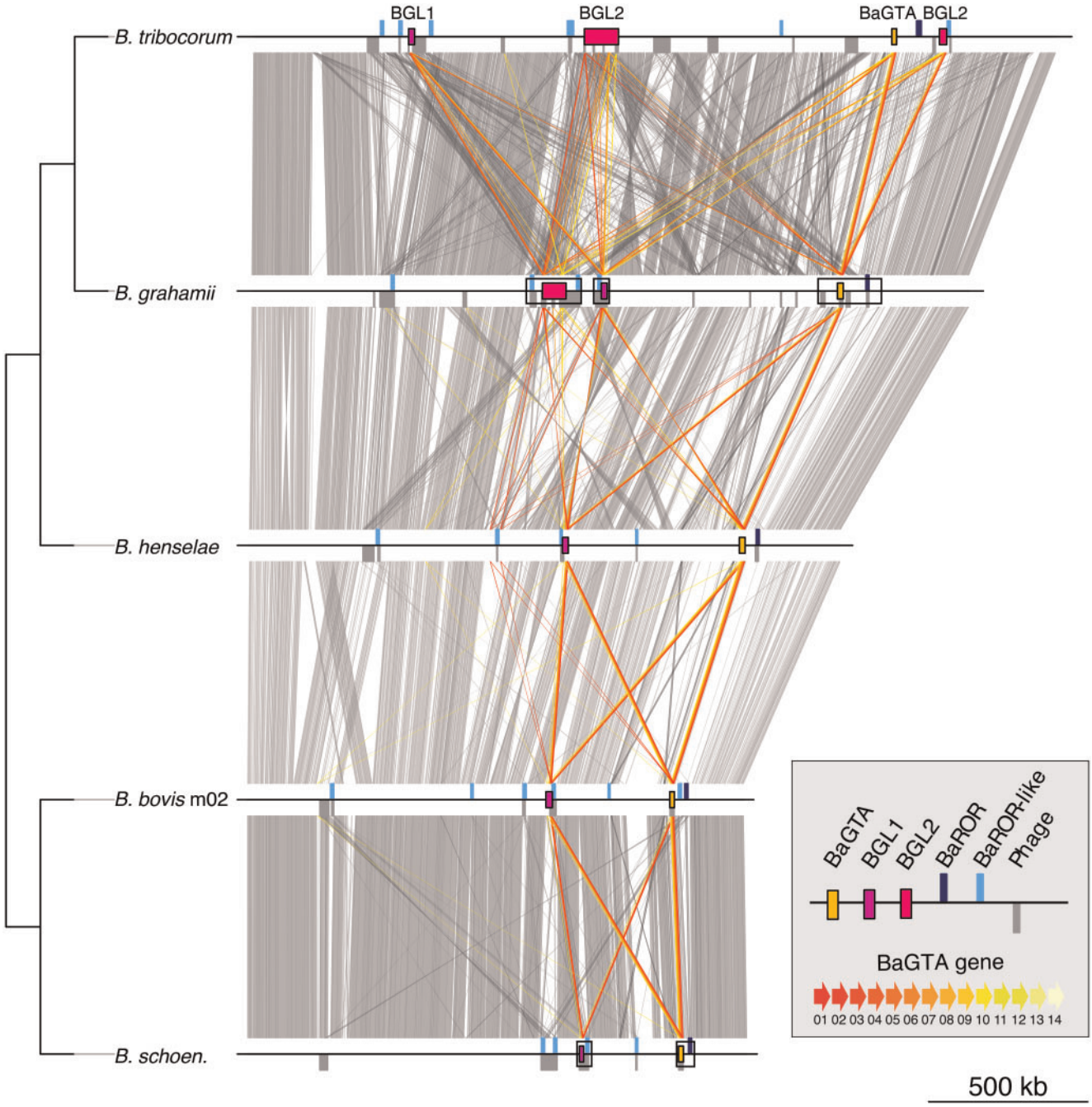

Fig. 5. Genomic location of BaGTA, BaROR, and homologous sequences in *Bartonella* and close relatives. Colored boxes represent BaGTA, BaROR, and homologous regions, as defined in the upper part of the inset legend. Gray boxes below the genome lines represent regions encoding genes annotated as "phage" at distances shorter than 20 kb. Homology lines represent BLASTn hits of at least 600 nucleotides and 30% identity (gray), overlaid with reciprocal PSI-BLAST hits for BaGTA gene homologs (following colors in fig. 1*B*, shown in the lower part of the inset legend). The tree topology was taken from Guy et al. (2013). *B. schoen.*, *Bartonella schoenbuchensis*.

core genes although the differences were not found to be significant after correcting for multiple testing due to the small BGL1 sample sizes (Holm–Bonferroni-corrected *P* values of 0.123 and 0.055, respectively).

The phage genes represented by BGL2 could only be identified in *B. grahamii* and *B. tribocorum*, with two copies in *B. tribocorum*. However, the dS values approached saturation (>0.7 substitutions per site) for all genes in the BGL2 segment, and were thus higher than the dS values of both the BaGTA and the BGL1 genes. We conclude that the BaGTA genes have mostly diverged by vertical descent, whereas the BGL1 and BGL2 genes show signs of rapid sequence evolution, recombination between strains and/or multiple, independent phage integration events. However, the effect is species-specific, probably due to the stochastic nature of such events and the time elapsed since the species diverged.

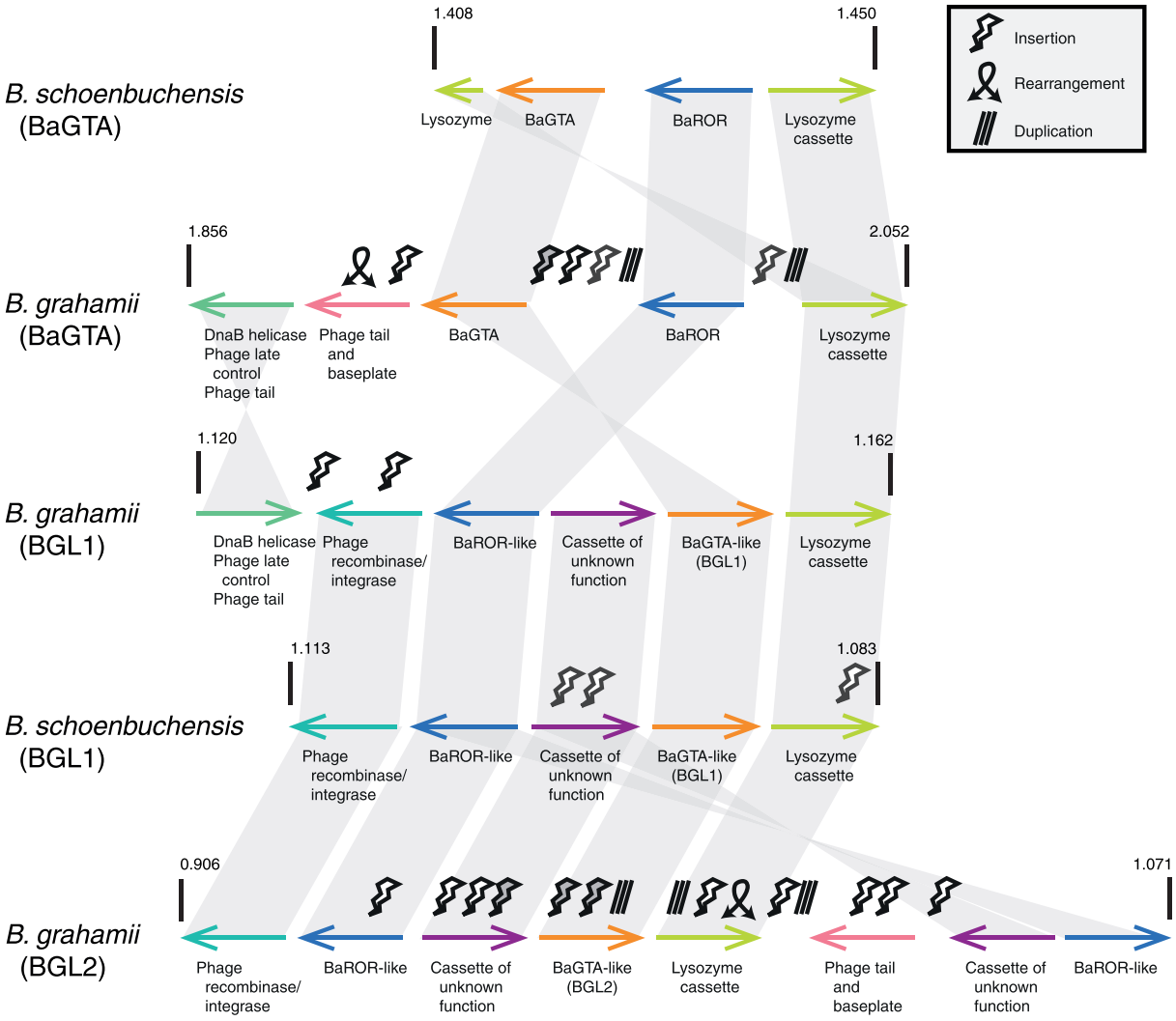

Fig. 6. Gene order conservation around BaGTA and BaGTA-like gene clusters in other *Bartonella* spp. Plot showing similarity in the arrangement of gene blocks surrounding the BaGTA and BaGTA-like regions in *B. grahamii* and *B. schoenbuchensis*, as an example of a canonical *Bartonella* species with a BGL1 gene cluster. Horizontal arrows represent groups of genes with general gene order conservation. Colors and connecting lines indicate homology between genes. Events of gene insertion, rearrangements, and duplications are shown as indicated in the inset. Insertions are shaded in gray when they represent a repeated mobile element containing genes for hemolysin activator and filamentous hemagglutinin.

A Model for the Origin and Evolution of the BaGTA

In this study, we have traced the evolution of the BaGTA and suggest that it has evolved from a bacteriophage that was acquired in the last common ancestor of *Bartonella* and *Ca.* Tokpelaia. A series of observations suggest that the BaGTA is coevolving with the bacterial genome: the BaGTA tree topology matches the species topology, the gene order structure and genomic location of the BaGTA is conserved across species and the base composition patterns are similar to those of other genes in the bacterial host genome. However, we could not find any trace of an ortholog copy of the phage-derived origin of replication in a genome outside of the eubartonellae. Furthermore, the phylogenetic analysis indicates that the only BaROR homologs present in these earlier diverging genomes belong to a separate group that is associated with the BGL rather than with the BaGTA regions, and they do not show the peak in the GC3s values normally observed in the BaROR region in the eubatonellae. Therefore, the BaROR is likely to have a distinct, more recent origin than the GTA itself. Here, we propose that the BaGTA has evolved from an integrated prophage that has been connected with the BaROR through a series of events, as briefly outlined below.

Bacteriophage Integration

The first step in the evolution of the BaGTA was the integration of a bacteriophage into the bacterial genome (fig. 7A). The primordial bacteriophage probably belonged to a diverse lineage, with variants infecting a variety of hosts, because homologous phage sequences are present in a few species outside the genus *Bartonella*. Moreover, the multiplicity and relatedness of the BaGTA homologs in *Bartonella* spp. suggests extensive proliferation of these sequences, with rates and patterns of sequence evolution rather resembling a phage-like mobile element than a GTA. On the other hand,

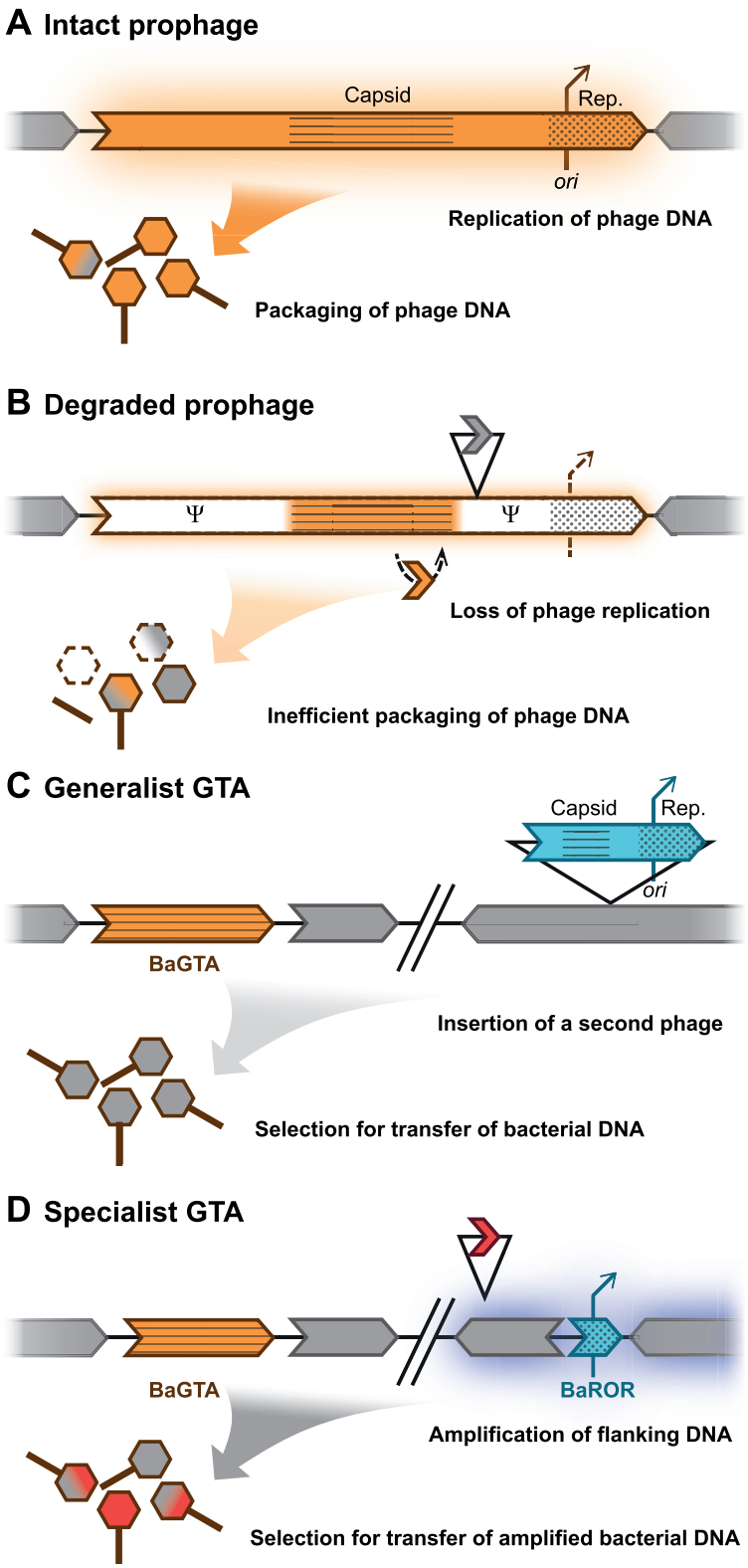

Fig. 7. Model for the evolution of the *Bartonella* GTA. Throughout the scheme, colored, and gray horizontal arrows represent phage and bacterial genes, respectively; vertical arrows toward capsids represent the control over capsid formation and DNA packaging processes; and colored glow represents DNA amplification by a phage-derived origin of replication (*ori*). (*A*) A prophage with basic modules encoding functions such as regulation, capsid formation (lined), bacterial lysis, and replication (dotted) inserts in the chromosome. The prophage regulates the replication and expression of its own genes; upon induction, phage particles are produced containing bacteriophage DNA, with the rare inclusion of bacterial DNA. (*B*) The prophage accumulates mutations and deletions, which reduces the efficiency of phage DNA replication and capsid production. (*C*) The loss of the phage-derived replication genes causes the structural genes to incorporate random bacterial DNA into their capsids, as an early stage generalist GTA. A subsequent insertion of a different phage (not to scale), including replication machinery and origin, causes independent

the overall congruency between phylogenies of BaGTA homologs and trees reconstructed from genomic data of the hosts (Neuvonen et al. 2016; Segers et al. 2017) suggests that the integration event that would generate the BaGTA was at least as old as the Bartonellaceae lineage. The gene segment represented by BGL1 in *B. grahamii* provides an example of what the content and order of genes may have looked like in the ancestral bacteriophage.

The Initiation of Phage Replication Became Uncoupled from the Genes for Phage Capsid Proteins

We hypothesize that the inserted bacteriophage started to accumulate mutations, deletions, insertions, and rearrangements, which affected the integrity and regulation of the prophage, and thereby prevented the lytic cycle (fig. 7*B*). We further suggest that the uncoupling of genes for the phage capsid proteins from genes involved in the initiation and termination of phage replication was a key event in the evolution of the GTA (fig. 7*B* and *C*). The BGL2 segment in *B. grahamii* provides an example of how the inserted phage genes may have become disconnected from each other through the insertion of novel genes, such as the *fha/hec* repeat genes. Consistently, a previous gene content study of 27 *B. grahamii* strains by microarray comparative genome hybridization indicated an inverse correlation between the presence of the *fha/hec* repeat genes and an increased copy number of the flanking phage genes (Berglund et al. 2009). The simplest interpretation is that the insertion of the *fha/hec* genes abolished replication of the integrated phage genes. However, complete genome sequence data and experimental studies from multiple *B. grahamii* strains with and without the *fha/hec* insertion would be needed to test this hypothesis.

Phage Particles Started Packaging and Transferring Random Fragments of Bacterial DNA

The loss of coordination between the initiation of phage replication and expression of phage particles enabled random fragments of chromosomal DNA to be packaged into the phage particle (fig. 7*B* and *C*). We hypothesize that noneubartonellae *Bartonella* species, such as *B. apis* and *B. tamiae*, which contain homologs to the BaGTA genes but not to the BaROR genes, encode generalist GTAs that package genomic DNA randomly, thus resembling the RcGTA.

Run-off Replication by a Phage Replication Initiation Site Amplified Flanking Bacterial DNA

Our analysis revealed multiple copies of genes for a putative phage-derived origin of replication in most of the *Bartonella* genomes, often but not always associated with other phage genes. Any of these may incidentally have started to replicate the surrounding genes, including in some cases bacterial genes, in a process referred to as run-off replication (fig. 7*C* and *D*). The BaROR genes, which contain such a replication initiation site, may have been derived from the same or from a different bacteriophage than the one that evolved into the BaGTA. We hypothesize that the amplification of bacterial genes from the phage replication initiation site may have conferred a selective advantage due to higher gene expression levels or a higher propensity for gene transfer, recombination, and diversification, thereby facilitating adaptation to diverse host species (fig. 7*D*).

The Bacterium Controls the Amplification of the Bacterial Genes Near the Phage Replication Initiation Site and the Expression of the GTA

Finally, the bacterial host may have started to regulate the expression of the BaGTA and the phage-derived origin of replication (fig. 7*D*). It has been shown in *R. capsulatus* that both the production of the RcGTA particle and flagellar motility is under regulatory control of the host cell via the CtrA phosphorelay system, and induced by quorum sensing during stationary phase (Brimacombe et al. 2013; Mercer and Lang 2014). It is interesting to note that flagellar genes present in some *Bartonella* species are located in the secretion system cassette near the genes for the BaGTA and the BaROR, and thus may be regulated by similar control mechanisms. However, bacteriophage-like particles that package chromosomal DNA were produced during both exponential and stationary phase in *B. grahamii*, suggesting that the regulatory systems that activate GTA-mediated transfer of bacterial DNA might differ (Berglund et al. 2009). Furthermore, it has been recently found that high levels of ppGpp inhibit the activity of the BaGTA, thus suggesting that gene transfer via the BaGTA is restricted to actively replicating bacterial cells (Quebatte et al. 2017).

Although there are many similarities between the RcGTA and the BaGTA, a main difference is that the RcGTA packages bacterial DNA randomly and there are no indications of specific regions in the bacterial genome are transferred more frequently than the genome overall (Hynes et al. 2012). Our results suggest that the BaGTA replaced the RcGTA; however, exactly how this happened is still unclear. Once the replacement occurred, the BaGTA precursor integrated within the host physiological processes, probably co-opting preadapted regulatory networks, and behaved as a generalist GTA. Subsequently, it evolved a specialist behavior by coupling its action with a phage-derived origin of replication that amplified the surrounding DNA consisting of potentially adaptive genes. Although the BaGTA would still package DNA randomly, the effective result of this ternary system would result in the frequent transfer of DNA from the amplified regions, thereby increasing the evolvability of the eubartonellae and facilitating their radiation to multiple mammalian hosts. These results highlight the complex evolutionary patterns

Fig. 7. Continued
regulation and amplification of separate gene repertoires. (*D*) The bacterial genome takes over the regulation of the replication from the phage-derived origin of replication as well as the production of phage particles through optimization of the GTA by streamlining and modification. The more likely encapsidation of amplified DNA generates an advantage for adaptive genes to relocate to the amplified region.

dictating the replacement of GTA systems and their ensuing acquisition of more sophisticated behaviors. The finding that one strain of *B. tamiae* contains the BaGTA while a closely related strain contains a highly divergent putative GTA at the exact same site with a few short matches to the RcGTA indicates that GTA replacements may be more common than previously thought. Further analyses of closely related strains of *B. tamiae* may provide insights into the process of GTA replacement and divergence, as well as into the functionality and dynamics of early stage GTA regions.

Materials and Methods

BaGTA Sequence Retrieval and Searches

The genomes of the radiating *Bartonella* species, *B. tamiae*, *B. apis*, and *Ca*. T. hoelldoblerii strains were retrieved from the public databases (supplementary table S3, Supplementary Material online). In order to find homologs from BaGTA and BaROR in *Bartonella* species and relatives, all 14 genes comprising the canonical BaGTA and 6 genes comprising the BaROR region were selected from 8 canonical *Bartonella* genomes (supplementary table S4, Supplementary Material online). Each gene was aligned using Mafft-linsi v6.857b (Katoh et al. 2002; Katoh and Toh 2008), and the alignments were used to perform PSI-BLAST searches (Altschul et al. 1997) against NCBI's database nr and the local genomes (supplementary table S3, Supplementary Material online) and using an E-value cutoff of 10^{-3} and two search iterations.

BaGTA Identification and Phylogenetic Analyses

The hits were retrieved, aligned with Mafft-linsi v6.857b, sites with over 50% gaps were trimmed with TrimAl v1.2rev59 (Capella-Gutierrez et al. 2009), and the alignments were used for phylogenetic inference using FastTree v2.1.7 (Price et al. 2009, 2010). The obtained trees were manually checked and sequences belonging to overrepresented groups and well-supported groups placed far from the *Tokpelaia-Bartonella* ingroup were removed from the alignment. The remaining sequences were aligned this time using ProbCons v1.12 (Do et al. 2005) and trimmed as before. Duplicated sequences were removed, and the resulting alignment was used for phylogenetic inference with RAxML v7.2.6 (Stamatakis 2006) under the PROTGAMMALG model and performing 100 bootstrap pseudoreplicates. If necessary, the last iteration was repeated following the removal of still overrepresented groups. Given the lack of several BaGTA homologs in *Tokpelaia*, in contrast with the presence and synteny conservation of genes *bgtJIHG* (BaGTA-02 to BaGTA-05), the final alignments obtained for these four genes were concatenated and two trees were constructed, one using RAxML as before, and another using Phylobayes v4.4f (Lartillot et al. 2009) using the CAT model until convergence with a MaxDiff value <0.3 and an Effective size >50. The sequences from the newly published *B. apis* genomes (Segers et al. 2017) were added to the concatenated alignment of genes *bgtJIHG* (BaGTA-02 to -05), and were used for a maximum-likelihood phylogenetic reconstruction as above, confirming the monophyly of the *B. apis* clade (supplementary fig. S13, Supplementary Material online) and justifying the use of *B. apis* PEB0122 as representative for the rest of the analyses.

BaGTA Sequence and Synteny Analyses

GC3s values were calculated using the CodonW package (Peden 1999); and plotted using R (R Core Team 2017). Calculation of dN and dS values was performed using the Yang and Nielsen (2000) method implemented in PAML v.4.9a (Yang 1997, 2007). Statistical tests and plotting were performed using R. The identification of single-copy panorthologs for the used genomes was taken from a previous study (Neuvonen et al. 2016), with the addition of the orthologs of *B. bovis* m02, identified by a best reciprocal BLAST hit analysis with *B. schoenbuchensis* m07a. PSI-BLAST with the above-mentioned settings was used to search for homology and gene order conservation in genes surrounding the BaGTA and BaGTA-like regions in *Ca.* T. hoelldoblerii and the different *Bartonella* spp., and comparative figures were drawn using the R package genoPlotR (Guy et al. 2010).

Phyletic Comparison of BaGTA and RcGTA

In order to identify GTA regions in other Alphaproteobacteria, the complete genomes from Rhizobiales, Caulobacterales, and Rhodobacterales were downloaded from NCBI in April 2016, adding the genome of *Ca.* T. hoelldoblerii (Neuvonen et al. 2016). PSI-BLAST searches against these genomes were performed as above for the 14 BaGTA genes. The 17 genes that comprise the RcGTA (Lang and Beatty 2001) were taken from *Rhodobacter capsulatus* SB 100 (CP001312), and from *Rhodobacter sphaeroides* strains ATCC 1702 (CP000577), KD13 (CP001150), MTBLJ (CP012960), and WS8 (CM001161). Homolog detection in these genomes was done as for the BaGTA genes above. For the purpose of parsing and plotting the hits, a GTA region was defined as a single region, at most 50 kb long, containing at least three hits to the GTA, with a distance of at most 25 kb between each hit. Regions were discarded if they contained one or more synteny breaks when compared with the original GTA. Genes for the concatenated RcGTA phylogenetic tree were chosen to maximize the number of genes and genomes included, and comprised genes *g2*, *g3*, *g4*, *g5*, *g6*, *g9*, and *g12*. The genes were first aligned with Mafft-linsi and a tree reconstructed with FastTree to search and remove recent paralogs; the resulting sequences were then aligned with Probcons and a tree was reconstructed with RAxML as explained earlier. The portal and terminase sequences used for this tree, together with the portal and terminase sequences used in the concatenated BaGTA alignment, were joined with the B. tamiae Th239 homologs, were used for a phylogenetic reconstruction as above,

BaROR Region Identification and Synteny Analysis

PSI-BLAST hits to the local *Bartonella* spp. and *Tokpelaia* genomes (supplementary table S3, Supplementary Material online) for the six ROR genes were used to identify *Bartonella* regions with homology to the BaROR. These regions were defined by extending ten genes at either side of the ROR homologs, and regions were joined if they overlapped.

Thus, 173 BaROR hits yielded 46 BaROR regions. One of them was found to contain two contiguous inverted regions, and two were found to have one and two additional inserted regions, yielding a total of 50 gene neighborhoods to study. Gene homologs were defined as all genes in a network formed by genes as nodes, and edges being PSI-BLAST reciprocal hits between each pairs of genomes, with PSI-BLAST parameters as before. The synteny of BaROR regions was analyzed visually using custom perl and R scripts. For visualization purposes, the regions were ordered based on hierarchical clustering, which was applied to the logarithm of the number of shared reciprocal PSI-BLAST hits between each region, and performed as implemented by default in the R function heatmap.2 in the package gplots (Warnes et al. 2013).

BaROR Phylogenetic Analyses

Phylogenies for each of the six BaROR genes were constructed following the same pipeline as described earlier for the BaGTA genes. Additional PSI-BLAST searches were performed for each BaROR gene and homologs thereof, against a local database containing all complete alphaproteobacterial genomes in NCBI (as of February 2016). Given that BaROR-01, BaROR-02, and BaROR-04 homologs were found in several BaROR-like regions in *Bartonella* spp., we selected those genomes where homologs of BaROR-01 and BaROR-04 were found at a distance of <20 kb, because BaROR-02 could not be found in combination with these. We then joined these hits to the ones from *Bartonella* spp. local genomes and, as above, we aligned the sequences obtained for each gene with ProbCons and trimmed them with TrimAl. Then we concatenated them and obtained phylogenetic trees using RAxML and MrBayes (Huelsenbeck and Ronquist 2001; Ronquist et al. 2012), with the LG model with four categories under Gamma distribution. MrBayes ran for 50 milliion generations until SD was lower than 0.04. The tree was drawn in FigTree (Andrew Rambaut, available at: http://tree.bio.ed.ac.uk/software/figtree/; last accessed November 26, 2017) and edited in Adobe Illustrator.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

This work was supported by the Swedish Research Council (349-2007-8732, 621-2014-4460 to S.G.E.A.); the Knut and Alice Wallenberg Foundation (2011.0148, 2012.0075 to S.G.E.A.); the European Union from the Marie Curie ITN SYMBIOMICS (264774 to D.T.); and the Swiss National Science Foundation (PA00P3_131491 to L.G., 31003A_160345 to P.E.). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

References

Alsmark CM, Frank AC, Karlberg EO, Legault BA, Ardell DH, Canback B, Eriksson AS, Naslund AK, Handley SA, Huvet M, et al. 2004. The louse-borne human pathogen *Bartonella quintana* is a genomic derivative of the zoonotic agent *Bartonella henselae*. *Proc Natl Acad Sci U S A*. 101(26):9716–9721.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25(17):3389–3402.

Berglund EC, Frank AC, Calteau A, Vinnere Pettersson O, Granberg F, Eriksson A-S, Näslund K, Holmberg M, Lindroos H, Andersson SGE, et al. 2009. Run-off replication of host-adaptability genes is associated with gene transfer agents in the genome of mouse-infecting *Bartonella grahamii*. *PLoS Genet*. 5(7):e1000546.

Bobay LM, Touchon M, Rocha EP. 2014. Pervasive domestication of defective prophages by bacteria. *Proc Natl Acad Sci U S A*. 111(33):12127–12132.

Brimacombe CA, Stevens A, Jun D, Mercer R, Lang AS, Beatty JT. 2013. Quorum-sensing regulation of a capsular polysaccharide receptor for the *Rhodobacter capsulatus* gene transfer agent (RcGTA). *Mol Microbiol*. 87(4):802–817.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15):1972–1973.

Chomel BB, Boulouis HJ, Breitschwerdt EB, Kasten RW, Vayssier-Taussat M, Birtles RJ, Koehler JE, Dehio C. 2009. Ecological fitness and strategies of adaptation of *Bartonella* species to their hosts and vectors. *Vet Res*. 40(2):29.

Do CB, Mahabhashyam MS, Brudno M, Batzoglou S. 2005. ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res*. 15(2):330–340.

Eicher SC, Dehio C. 2012. *Bartonella* entry mechanisms into mammalian host cells. *Cell Microbiol*. 14(8):1166–1173.

Engel P, Salzburger W, Liesch M, Chang C-C, Maruyama S, Lanz C, Calteau A, Lajus A, Médigue C, Schuster SC, et al. 2011. Parallel evolution of a type IV secretion system in radiating lineages of the host-restricted bacterial pathogen *Bartonella*. *PLoS Genet*. 7(2):e1001296.

Fogg PC, Westbye AB, Beatty JT. 2012. One for all or all for one: heterogeneous expression and host cell lysis are key to gene transfer agent activity in *Rhodobacter capsulatus*. *PLoS One* 7(8):e43772.

Guy L, Kultima JR, Andersson SG. 2010. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* 26(18):2334–2335.

Guy L, Nystedt B, Toft C, Zaremba-Niedzwiedzka K, Berglund EC, Granberg F, Näslund K, Eriksson A-S, Andersson SGE, Casadesús J. 2013. A gene transfer agent and a dynamic repertoire of secretion systems hold the keys to the explosive radiation of the emerging pathogen *Bartonella*. *PLoS Genet*. 9(3):e1003393.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17(8):754–755.

Humphrey SB, Stanton TB, Jensen NS, Zuerner RL. 1997. Purification and characterization of VSH-1, a generalized transducing bacteriophage of *Serpulina hyodysenteriae*. *J Bacteriol*. 179(2):323–329.

Hynes AP, Mercer RG, Watton DE, Buckley CB, Lang AS. 2012. DNA packaging bias and differential expression of gene transfer agent genes within a population during production and release of the *Rhodobacter capsulatus* gene transfer agent, RcGTA. *Mol Microbiol*. 85(2):314–325.

Hynes AP, Shakya M, Mercer RG, Grull MP, Bown L, Davidson F, Steffen E, Matchem H, Peach ME, Berger T. 2016. Functional and evolutionary characterization of a gene transfer agent's multi-locus "genome". *Mol Biol Evol*. 33(10): 2530–2543.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 30(14):3059–3066.

Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform*. 9(4):286–298.

Kesnerova L, Moritz R, Engel P. 2016. *Bartonella apis* sp. nov., a honey bee gut symbiont of the class Alphaproteobacteria. *Int J Syst Evol Microbiol*. 66(1):414–421.

Kosoy M, Morway C, Sheff KW, Bai Y, Colborn J, Chalcraft L, Dowell SF, Peruski LF, Maloney SA, Baggett H, et al. 2008. *Bartonella tamiae* sp. nov., a newly recognized pathogen isolated from three human patients from Thailand. *J Clin Microbiol*. 46(2):772–775.

Lang AS, Beatty JT. 2001. The gene transfer agent of *Rhodobacter capsulatus* and "constitutive transduction" in prokaryotes. *Arch Microbiol*. 175(4):241–249.

Lang AS, Beatty JT. 2007. Importance of widespread gene transfer agent genes in alpha-proteobacteria. *Trends Microbiol*. 15(2):54–62.

Lang AS, Taylor TA, Beatty JT. 2002. Evolutionary implications of phylogenetic analyses of the gene transfer agent (GTA) of *Rhodobacter capsulatus*. *J Mol Evol*. 55(5):534–543.

Lang AS, Zhaxybayeva O, Beatty JT. 2012. Gene transfer agents: phage-like elements of genetic exchange. *Nat Rev Microbiol*. 10(7):472–482.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25(17):2286–2288.

Lindroos H, Vinnere O, Mira A, Repsilber D, Naslund K, Andersson SG. 2006. Genome rearrangements, deletions, and amplifications in the natural population of *Bartonella henselae*. *J Bacteriol*. 188(21):7426–7439.

Marrs B. 1974. Genetic recombination in *Rhodopseudomonas capsulata*. *Proc Natl Acad Sci U S A*. 71(3):971–973.

Mercer RG, Lang AS. 2014. Identification of a predicted partner-switching system that affects production of the gene transfer agent RcGTA and stationary phase viability in *Rhodobacter capsulatus*. *BMC Microbiol*. 14:71.

Neuvonen MM, Tamarit D, Näslund K, Liebig J, Feldhaar H, Moran NA, Guy L, Andersson SGE. 2016. The genome of Rhizobiales bacteria in predatory ants reveals urease gene functions but no genes for nitrogen fixation. *Sci Rep*. 6(1):39197.

Peden JF. 1999. Analysis of codon usage [doctoral thesis]. United Kingdom: University of Notthingham.

Price MN, Dehal PS, Arkin AP. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol*. 26(7):1641–1650.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* 5(3):e9490.

Quebatte M, Christen M, Harms A, Korner J, Christen B, Dehio C. 2017. Gene transfer agent promotes evolvability within the fittest subpopulation of a bacterial pathogen. *Cell Syst*. 4:611–621 e616.

R Core Team. 2017. R: a language and environment for statistical computing [Internet]. Vienna (Austria): R Foundation for Statistical Computing. Available from: http://www.R-project.org/, last accessed November 26, 2017

Rabinovich L, Sigal N, Borovok I, Nir-Paz R, Herskovits AA. 2012. Prophage excision activates *Listeria* competence genes that promote phagosomal escape and virulence. *Cell* 150(4):792–802.

Rapp BJ, Wall JD. 1987. Genetic transfer in *Desulfovibrio desulfuricans*. *Proc Natl Acad Sci U S A*. 84(24):9128–9130.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*. 61(3):539–542.

Schaefer AL, Taylor TA, Beatty JT, Greenberg EP. 2002. Long-chain acyl-homoserine lactone quorum-sensing regulation of *Rhodobacter capsulatus* gene transfer agent production. *J Bacteriol*. 184(23):6515–6521.

Segers FH, Kesnerova L, Kosoy M, Engel P. 2017. Genomic changes associated with the evolutionary transition of an insect gut symbiont into a blood-borne pathogen. *isme J*. 11(5):1232–1244.

Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. *Nat Rev Genet*. 16(8):472–482.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21):2688–2690.

Touchon M, Bernheim A, Rocha EP. 2016. Genetic and life-history traits associated with the distribution of prophages in bacteria. *isme J*. 10(11):2744–2754.

Viklund J, Ettema TJ, Andersson SG. 2012. Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. *Mol Biol Evol*. 29(2):599–615.

Waldor MK, Friedman DI. 2005. Phage regulatory circuits and virulence gene expression. *Curr Opin Microbiol*. 8(4):459–465.

Wang X, Kim Y, Ma Q, Hong SH, Pokusaeva K, Sturino JM, Wood TK. 2010. Cryptic prophages help bacteria cope with adverse environments. *Nat Commun*. 1:147.

Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, Maechler M, Magnusson A, Moeller S, Schwartz M, et al. 2013. gplots: various R programming tools for plotting data [Internet]. Available from: http://CRAN.R-project.org/package=gplots, last accessed November 26, 2017

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*. 13(5):555–556.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24(8):1586–1591.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol*. 17(1):32–43.

Zhu Q, Kosoy M, Olival KJ, Dittmar K. 2014. Horizontal transfers and gene losses in the phospholipid pathway of bartonella reveal clues about early ecological niches. *Genome Biol Evol*. 6(8):2156–2169.