Formal Recognition and Classification of Gene Transfer Agents as Viriforms

Roman Kogay[1], Sonja Koppenhöfer[2], J. Thomas Beatty[3], Jens H. Kuhn[4], Andrew S. Lang[2], and Olga Zhaxybayeva[1,#]

[1]Department of Biological Sciences, Dartmouth College, Hanover, NH, USA; [2]Department of Biology, Memorial University of Newfoundland, St. John's, NL, Canada; [3]Department of Microbiology & Immunology, University of British Columbia, Vancouver, BC, Canada; [4]Integrated Research Facility at Fort Detrick, Division of Clinical Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Frederick, MD, USA

[#]Correspondence to Olga Zhaxybayeva, olga.zhaxybayeva@dartmouth.edu

Keywords: GTA; polydnaviriformid; *Brachyspira*; *Bartonella*; *Rhodobacterales*; nomenclature

ABSTRACT

Morphological and genetic features strongly suggest that gene transfer agents (GTAs) are caudoviricete-derived entities that have evolved in concert with cellular genomes to such a degree that they should not be considered viruses. Indeed, GTA particles resemble caudoviricete virions but, in contrast to caudoviricetes (or any viruses), GTAs can encapsidate at best only part of their own genomes, are induced solely in small subpopulations of prokaryotic host cells and are transmitted vertically as part of cellular genomes during replication and division. Therefore, the lifecycles of GTAs are analogous to virus-derived entities found in parasitoid wasps, which have recently been recognized as non-virus entities and therefore reclassified as viriforms. We evaluated three distinct, independently exapted GTA groups for which the genetic basis for GTA particle production has been established. Based on the evidence, we outline a classification scheme for these viriforms.

26 **INTRODUCTION**

27 In 2021, the International Committee on Taxonomy of Viruses (ICTV) ratified a taxonomic
28 proposal to formally accept a new operational definition of the term "virus" (Koonin et al. 2021;
29 Kuhn et al. 2020; Walker et al. 2021). Consequently, the most current version of the
30 International Code of Virus Classification and Nomenclature (ICVCN) states that viruses are
31
32 " … a type of MGEs [mobile genetic elements] that encode at least one protein that is a
33 major component of the virion encasing the nucleic acid of the respective MGE and
34 therefore the gene encoding the major virion protein itself; or MGEs that are clearly
35 demonstrable to be members of a line of evolutionary descent of such major virion
36 protein-encoding entities" (ICVCN Rule 3.3) (International Committee on Taxonomy of
37 Viruses 2022; Kuhn et al. 2020).
38
39 This definition also formalized the postulate that some MGEs, long understood by the
40 general virology community to be distinct from viruses, are indeed distinct. At the time, the
41 ICTV had already classified viroids and satellite nucleic acids in taxa separate from viral taxa (in
42 families/genera with names that end with suffixes -*viroidae*/-*viroid* and -*satellitidae*/-*satellite*,
43 respectively, as opposed to -*viridae*/-*virus*) (International Committee on Taxonomy of Viruses
44 2022), and these elements were logically placed into the perivirosphere rather than the
45 orthovirosphere (Koonin et al. 2021; Kuhn et al. 2020).
46 The adoption of the new virus definition brought into question the taxonomic standing of
47 one official virus family, *Polydnaviridae*. Indeed, entities classified into this polyphyletic family
48 fundamentally deviate from MGEs fulfilling the virus definition because "polydna" particles

encapsidate multiple segments of circular double-stranded DNAs that, however, do not encode the entire "polydna" genomes. Instead, the genomes are permanently endogenized into the "polydna" host (i.e., parasitoid wasp) genomes and inherited vertically. The resultant non-mobile nonviral entities are used by the wasps to deliver immunomodulatory genes into insects that serve as prey for the wasps (Drezen et al. 2017; Herniou et al. 2013). "Polydna" entities are likely evolutionarily derived from various groups of insect viruses, including nudivirids (Darboux et al. 2019; Drezen et al. 2017; Gauthier et al. 2018; Petersen et al. 2022; Strand and Burke 2020; Thézé et al. 2011), but, because they have lost the ability to replicate and instead have been fully exapted by their wasp hosts, they have left the virosphere altogether (Koonin and Krupovic 2018; Koonin et al. 2021). Consequently, in 2021, the ICTV recognized "polydna" entities as representatives of a new MGE category distinct from viruses called "viriforms" (Koonin et al. 2021; Kuhn et al. 2020; Walker et al. 2021), and reclassified *Polydnaviridae* as (still polyphyletic) *Polydnaviriformidae* (Kuhn et al. 2021; Walker et al. 2022). In the ICVCN, viriforms are defined operationally as

> " … a type of virus-derived MGEs that have been exapted by their organismal (cellular) hosts to fulfill functions important for the host life cycle; or MGEs that are derived from such entities in the course of evolution" (ICVCN Rule 3.3) (International Committee on Taxonomy of Viruses 2022; Kuhn et al. 2020).

Importantly, the following comment was added to the Rule 3.3:

71 "Gene transfer agents (GTAs) and the MGEs previously classified in the family
72 *Polydnaviridae* are considered to be viriforms in classification and nomenclature"
73 (International Committee on Taxonomy of Viruses 2022; Kuhn et al. 2020).
74
75 Notably, there are no discernible evolutionary relationships between GTAs and
76 polydnaviriformids. The term "viriform", similar to the term "virus", is an umbrella term for
77 certain MGEs with comparable lifecycles and properties; it is currently applied to six realms of
78 MGEs that are not evolutionary related to each other.
79 Based on the properties of entities referred to as "GTAs" in the literature (reviewed in
80 (Lang et al. 2012; Lang et al. 2017) we define GTAs as viriforms with the following features:
81 1. GTAs use caudoviricete ancestor-derived proteins (established either via significant
82 similarity of at least some GTA proteins to caudeviricete proteins or by image-based
83 evidence of caudovirion-like particles) to form caudovirion-like particles;
84 2. GTAs encapsidate mostly random pieces of host DNA (established experimentally);
85 3. GTA genomes are fully endogenized in host genomes, often across multiple loci
86 (established experimentally and via genomic examination);
87 4. GTA genomes are not/cannot be fully packaged into particles due to limited particle
88 head size (established via comparison of the packaged DNA length and size of GTA
89 loci);
90 5. GTA genomes are mostly vertically inherited and GTAs co-diversify with their hosts
91 (established via congruence between phylogenies of host and GTA genes); and
92 6. DNA encapsidated in GTA particles is delivered to other cells (established
93 experimentally).

94 Having these attributes, GTAs have lost the ability to replicate and have become fully exapted by
95 their cellular hosts. They are produced under specific conditions (e.g., nutrient depletion
96 [Westbye et al. 2017b]) and mediate horizontal gene transfer (HGT), typically among cells of the
97 same species.

98 The first GTA discovered, of the alphaproteobacterium *Rhodobacter capsulatus*, was
99 described in 1974 by Barry Marrs (Marrs 1974). Since that time, distinct functional GTAs have
100 been described in other alphaproteobacteria, a sulfate-reducing deltaproteobacterium, a
101 methanogenic archaeon, and a spirochete that infects domestic pigs (Bertani 1999; Guy et al.
102 2013; Humphrey et al. 1997; Rapp and Wall 1987). Additionally, clusters of genes homologous
103 to those encoding the *R. capsulatus* GTA are found in many alphaproteobacterial genomes,
104 suggesting a wider prevalence of GTA production than presently appreciated (Kogay et al. 2019;
105 Lang et al. 2002; Lang and Beatty 2007; Shakya et al. 2017). Indeed, some of these bacteria
106 produce functional GTAs (Biers et al. 2008; Nagao et al. 2015; Tomasch et al. 2018).

107 The recent ICTV recognition of viriforms and the formal establishment of
108 *Polydnaviriformidae* provides an opportunity to initiate a systematic classification of GTAs.
109 Here we outline initial steps to establish such a formal taxonomic scheme for GTA viriforms,
110 focusing specifically on GTAs experimentally documented as being produced by cells and
111 performing gene transfer—and for which the genetic basis of particle production has been
112 established. Simultaneously, we have also officially proposed this taxonomic scheme to the
113 ICTV for the 2022–2023 proposal cycle.
114

115 **NOMENCLATURE OF GENE TRANSFER AGENTS AND ASSOCIATED TAXA**

116 Per ICTV rules, virus names are written in lower case (except if a name component is a proper
117 noun), without italics in any part of the name (even if a host species name is part of the name),
118 and ending in the term "virus", which in virus name abbreviations is "V". Examples are measles
119 virus (MeV) and Ebola virus (EBOV). The nomenclature of already classified viriforms
120 (polydnaviriformids) follows these rules, with "virus" being replaced by "viriform" and the
121 abbreviation "V" being replaced with "Vf" (e.g., "Glyptapanteles liparidis bracoviriform" is
122 abbreviated "GlBVf"). We suggest applying these general rules to GTAs, but with "viriform"
123 being replaced by "gene transfer agent" due to the long-established use of this phrase and "Vf"
124 being replaced with "GTA". Therefore, the gene transfer agent produced by *Rhodobacter*
125 *capsulatus* would be called "Rhodobacter capsulatus gene transfer agent" and abbreviated as
126 "RcGTA", consistent with the established use of this abbreviation in the literature.

127

128 Rules for viriform taxon naming have been established by the ICVCN. Specifically,

129

130 "[t]he formal endings for taxon names of viriforms are the suffixes "-*viriformia*" for
131 realms, "-*viriforma*" for subrealms, "-*viriformae*" for kingdoms, "-*viriformites*" for
132 subkingdoms, "-*viriformicota*" for phyla, "-*viriformicotina*" for subphyla,
133 "-*viriformicetes*" for classes, "-*viriformicetidae*" for subclasses, "-*viriformales*" for
134 orders, "-*viriformineae*" for suborders, "-*viriformidae*" for families, "-*viriforminae*" for
135 subfamilies, and "-*viriform*" for genera and subgenera" (ICVCN Rule 3.26) (International
136 Committee on Taxonomy of Viruses 2022; Kuhn et al. 2020)

137

138 and

139

140 "[a] species name shall consist of only two distinct word components separated by a
141 space. The first word component shall begin with a capital letter and be identical in
142 spelling to the name of the genus to which the species belongs. The second word
143 component shall not contain any suffixes specific for taxa of higher ranks. The entire
144 species name (both word components) shall be italicized" (International Committee on
145 Taxonomy of Viruses 2022).

146

147 We suggest adding the infix -*gta*- prior to the taxon-specific suffixes for immediate recognition
148 of GTA-specific taxa (e.g., -*gtaviriform*).

149

150 **GENE TRANSFER AGENTS CAN BE ASSIGNED TO AT LEAST THREE MAJOR**
151 **CLADES**

152 Based on functionally and genetically characterized GTAs, at least three major GTA clades can
153 be delineated.

154

155 **Alphaproteobacterial type I GTAs**

156 The best characterized GTA of this clade is RcGTA, produced by *R. capsulatu*s
157 (*Pseudomonadota*: *Alphaproteobacteria*: *Rhodobacterales*: *Rhodobacteraceae*). We designate
158 RcGTA here as the founding member of one major GTA clade, the alphaproteobacterial type I
159 GTAs. For many years since its discovery (Marrs 1974), RcGTA was the only known GTA.
160 Now we know that homologous GTAs are produced by other bacteria from the order

161 *Rhodobacterales*: *Dinoroseobacter shibae* (Dinoroseobacter shibae gene transfer agent
162 [DsGTA]) (Tomasch et al. 2018), *Ruegeria pomeroyi* (Ruegeria pomeroyi gene transfer agent
163 [RpGTA]) (Biers et al. 2008), and *Rhodovulum sulfidophilum* (Rhodovulum sulfidophilum gene
164 transfer agent [RsGTA]) (Nagao et al. 2015). Additionally, genes encoding RcGTA-like GTAs
165 are conserved in most genomes in the order *Rhodobacterales* and in many genomes of the
166 alphaproteobacterial orders *Caulobacterales*, *Sphingomonadales*, *Parvibaculales,* and
167 *Hyphomicrobiales* (formerly *Rhizobiales*) (Kogay et al. 2019; Lang et al. 2002; Lang and Beatty
168 2007; Shakya et al. 2017).

169 RcGTA and RcGTA-like GTA genes are similar in sequence to those of viruses classified
170 in the uroviricot class *Caudoviricetes* (*Duplodnaviria*: *Heunggongvirae*) (Shakya et al. 2017).
171 These GTAs are transmitted vertically from a bacterial parent to progeny during cell division
172 (Lang and Beatty 2007; Shakya et al. 2017), similar to propagation of temperate viruses
173 ("prophages"). However, in contrast to temperate virus genomes, the set of genes required for
174 production of the GTA particle (the GTA "genome") is not necessarily localized in one region of
175 the host genome. In the case of RcGTA, known structural and regulatory genes are scattered
176 across five loci in the *R. capsulatus* genome (Hynes et al. 2016), cumulatively spanning
177 approximately 20 kilobases (kb) (**Figure 1** and **Supplementary Table S1**). Moreover, cellular
178 regulatory genes are involved in controlling GTA particle production (Westbye et al. 2017a),
179 adding another factor that makes the GTA genome difficult to differentiate from its host's
180 genome.

181 RcGTA particles resemble virions of caudoviricetes (Yen et al. 1979) and have been
182 structurally characterized at high resolution (Bárdy et al. 2020). RcGTA particles have head
183 diameters of 38 nm and tail lengths of 49 nm. A small percentage of RcGTA particles have T = 3

184 quasi-icosahedral heads, but the capsid shape of most particles is oblate, as they lack the five
185 hexamers of capsid protein needed to form genuine icosahedral heads. Because of the small head
186 size, RcGTA particles can only package double-stranded DNA of approximately 4 kb in length
187 (Yen et al. 1979). The DNA is also encapsidated at 10–25% lower density than typical
188 caudoviricetes (Bárdy et al. 2020). Both RcGTA particle production and acquisition of the GTA-
189 packaged DNA by other host cells in the population are controlled by the same cellular
190 regulatory systems (Westbye et al. 2017a). Only 0.1–3.0% of cells produce GTA particles (Fogg
191 et al. 2012; Hynes et al. 2012), whereas the remaining cells produce a GTA receptor
192 (Brimacombe et al. 2013).

193 Compositionally, structural proteins encoded by RcGTA and RcGTA-like GTAs are
194 biased towards amino acids that are energetically cheaper to produce (Kogay et al. 2020). To
195 date, such a bias has not yet been associated with viruses. Based on this difference in amino-acid
196 composition, GTA proteins can be distinguished from their viral homologs using a machine-
197 learning approach, which is implemented in the publicly available GTA-Hunter program (Kogay
198 et al. 2019).

199 In a comprehensive evolutionary analysis of homologs of the large subunit of the DNA
200 packaging terminase enzyme (TerL, encoded by the *g2* gene in the RcGTA genome), RcGTA
201 and RcGTA-like GTAs form a clade closely related to, but distinct from, duplodnavirians
202 (Esterman et al. 2021). To illustrate the relationships of alphaproteobacterial type I GTAs to each
203 other and to their closest viral homologs, we reconstructed evolutionary histories of their TerL
204 proteins and the HK97-like major capsid proteins (HK97-MCP, encoded by the *g5* gene in the
205 GTA genome, is the hallmark protein that defines the virus realm *Duplodnaviria* [Koonin et al.
206 2020]). Consistent with an earlier analysis (Esterman et al. 2021), RcGTA and RcGTA-like

207 GTAs formed a clade closely related to, but distinct from, caudoviricetes (**Figure 2**), with a few
208 exceptions that are likely artefacts of phylogenetic reconstruction.
209 Specifically, in the TerL phylogeny (**Figure 2A**), all viral homologs except one
210 (Caulobacter virus Sansa) are separated from GTA proteins (with a solid bootstrap support of
211 81%). Caulobacter virus Sansa groups with one GTA sequence from a bacterium of the order
212 *Sphingomonadales* (with a low bootstrap support of 50%), whereas all other GTAs of
213 *Sphingomonadales* bacteria group together (with a strong bootstrap support of 96%). We
214 hypothesize that the phylogenetic placement of the Caulobacter virus Sansa TerL is due to the
215 long-branch attraction artefact (Felsenstein 1978). We searched for a maximum-likelihood tree in
216 which caudoviricete- and GTA-derived TerLs were required to group separately from each other
217 and compared that tree to the tree depicted in **Figure 2A**. We found that the likelihoods of the
218 two trees are not significantly different (approximately unbiased [AU] test; p-value = 0.555),
219 confirming that the placement of the Caulobacter virus Sansa sequence within the GTA
220 sequences is unreliable.
221 In the HK97-MCP phylogeny (**Figure 2B**), GTAs and most caudoviricetes are separated
222 by a branch with 63% bootstrap support. Several caudoviricetes that group within GTAs are
223 located on long branches, are situated outside of well-supported groups of GTAs from several
224 alphaproteobacterial orders and have very low bootstrap support for their placements. It is
225 therefore likely that the positions of these viral homologs are unreliable. To test this hypothesis,
226 we identified a maximum-likelihood phylogeny among trees in which GTAs and caudoviricetes
227 were required to be separated by a branch. The likelihoods of this tree and the phylogeny shown
228 in **Figure 2B** are not significantly different (AU test; p-value = 0.534). Therefore, these viruses
229 are likely positioned in different places in trees reconstructed from different bootstrap replicates,

230 which would lead to their artificial (and poorly supported) basal positions with the GTA
231 homologs on the tree shown in **Figure 2B**.
232 In the **Figure 2** trees, GTA branches have shorter lengths than their caudoviricete
233 counterparts, conforming with the reported slower evolutionary rate of GTAs compared to
234 viruses (Shakya et al. 2017). Additionally, on both phylogenetic trees, GTAs from
235 alphaproteobacteria of different orders form separate groups with very high support,
236 corroborating vertical inheritance of most GTA genes (Lang et al. 2002; Lang and Beatty 2007;
237 Shakya et al. 2017).
238 Together, these results justify the classification of RcGTA and three RcGTA-like GTAs
239 in a common viriform taxon: family *Rhodogtaviriformidae* (from *Rhodobacterales*, infix -*gta*-,
240 and family-specific suffix -*viriformidae*). Given limited dataset size (i.e., just four GTAs), it is
241 challenging to establish quantifiable criteria for demarcating taxonomic relationships among the
242 four GTAs. In the future, when more GTA sequences become available for analyses, a criterion
243 based on percent sequence similarity among shared genes should be considered. For now, based
244 on the evidence of co-evolution of these GTAs and their specific hosts, we argue that at least
245 four rhodogtaviriformid genera, each for GTAs of bacteria classified in distinct genera included
246 in *Rhodobacterales*, ought to be established:
247 • *Dinogtaviriform* (named after DsGTA host genus *Dinoroseobacter*, infix -*gta*-, and genus-
248 specific suffix -*viriform*) to include one new species, *Dinogtaviriform tomaschi* (species
249 epithet to honor GTA researcher Jürgen Tomasch, who was instrumental in the discovery of
250 DsGTA) for DsGTA (**Supplementary Table S2**);
251 • *Rhodobactegtaviriform* (named after RcGTA host genus Rhodobacter, infix -*gta*-, and genus-
252 specific suffix -*viriform*) to include one new species, *Rhodobactegtaviriform marrsi* (species

253 epithet to honor GTA researcher Barry Marrs, who first discovered GTAs and coined the
254 term "gene transfer agent") for RcGTA (**Supplementary Table S1**);

255 • *Rhodovulugtaviriform* (named after RsGTA host genus *Rhodovulum*, infix *-gta-*, and genus-
256 specific suffix *-viriform*) to include one new species, *Rhodovulugtaviriform kikuchii* (species
257 epithet to honor GTA researcher Yo Kikuchi, who was instrumental in the discovery of
258 RsGTA) for RsGTA (**Supplementary Table S3**); and

259 • *Ruegerigtaviriform* (named after RpGTA host genus *Ruegeria*, infix *-gta-*, and genus-specific
260 suffix *-viriform*) to include one new species, *Ruegerigtaviriform cheni* (species epithet to
261 honor GTA researcher Feng Chen, who was instrumental in the discovery of RpGTA) for
262 RpGTA (**Supplementary Table S4**).

263

264 **Alphaproteobacterial type II GTAs**

265 There was a lag between discovery of these elements and their recognitions as bona fide GTAs.
266 Phage-like particles, originally referred to as bacteriophage-like particles (BLPs), that contained
267 heterogenous DNA from *Bartonella* host genomes were first characterized in *B. henselae*
268 (Anderson et al. 1994), and noted to be similar in structure to particles produced by *B.*
269 *bacilliformis* (Umemori et al. 1992). These *B. bacilliformis* particles were subsequently shown to
270 also contain heterogeneous genomic DNA fragments, but attempts to demonstrate their gene
271 transfer ability were not successful (Barbian and Minnick 2000). Functionality of the particles
272 produced by *Bartonella* for gene transfer (Bartonella gene transfer agent [BaGTA]) was
273 eventually demonstrated by work on *B. henselae* (*Pseudomonadota*: *Alphaproteobacteria*:
274 *Hyphomicrobiales*: *Bartonellaceae*) (Guy et al. 2013). BaGTA genes were initially proposed to
275 be located within a single cluster of 11–13 genes spanning approximately 14 kb (Guy et al.

276 2013). However, a subsequent screen for genes essential for BaGTA functionality identified a
277 total of 29 genes located within a larger (approximately 79-kb-long) region (Québatte et al.
278 2017) (**Figure 3** and **Supplementary Table S5**). Homologs of BaGTA genes (BaGTA-like
279 GTAs) were found in the genomes of multiple species of *Bartonella* (Berglund et al. 2009; Guy
280 et al. 2013; Tamarit et al. 2018). BaGTA genes are located near an active virus-derived origin of
281 replication and next to genes encoding secretion systems (Guy et al. 2013). As a result, the
282 region of the genome containing BaGTA and these secretion-system genes are amplified and
283 packaged more often than other genomic regions (Guy et al. 2013; Québatte et al. 2017). These
284 findings led to the hypothesis that BaGTA and BaGTA-like GTAs have been maintained due to
285 their mediation of HGT of secretion-system and toxin genes, thereby enabling *Bartonella*
286 bacteria to adapt to diverse hosts (Guy et al. 2013). However, actual GTA-mediated DNA
287 transfer among bacterial cells has only been demonstrated for *B. henselae* (Guy et al. 2013).
288 There, BaGTA production is restricted to a distinct subpopulation of fast-growing cells, which
289 comprise about 6% of the total population (Québatte et al. 2017), and the uptake of BaGTA-
290 packaged DNA was proposed to be limited to cells undergoing division (Québatte et al. 2017).
291 There are some discrepancies in the literature regarding the structure of BaGTA particles,
292 suggesting some bacteria might release additional phage-like particles. The *B. henselae* particles
293 were originally reported as particles without tails or with short non-contractile tails with a head
294 diameter of 40 nm (Anderson et al. 1994). The head diameter of the *B. bacilliformis* particles
295 was originally measured at 40 nm (Umemori et al. 1992) and subsequently 80 nm (Barbian and
296 Minnick 2000). Those of *B. grahamii* were reported as possessing long non-contractile tails and
297 icosahedral heads of 50–70 nm or 80 nm and tails of 100 nm (Berglund et al. 2009). Although
298 BaGTA particles are potentially able to package the entire main structural gene cluster of 11–13

299 genes, they cannot package all 29 genes required for BaGTA production due to a capacity of 14
300 kb (Anderson et al. 1994; Guy et al. 2013; Lang et al. 2017).
301 In the TerL phylogeny, BaGTA-like homologs are separated from almost all
302 caudoviricetes by longer branches (with 100% bootstrap support; **Figure 4A**). Two caudoviricete
303 homologs (Sulfitobacter phage pCB2047-C and Sulfitobacter phage NYA-2014a) group together
304 and are nested within the BaGTA-like group (with 84% bootstrap support). We hypothesize that
305 the *terL* gene was horizontally transferred from GTAs to these caudoviricetes, with similar HGT
306 events documented between RcGTA-like GTAs and caudoviricetes infecting bacteria of the
307 *Rhodobacterales* (Zhan et al. 2016). In the HK97-MCP phylogeny, BaGTA homologs are
308 located on shorter branches than their caudoviricete counterparts and are separated from
309 caudoviricete homologs with 100% bootstrap support (**Figure 4B**). Phylogenomic analyses
310 suggest that *Bartonella* GTAs have co-evolved with their hosts (Tamarit et al. 2018).
311 Together, these results justify the classification of BaGTA and BaGTA-like GTAs in a
312 common viriform taxon, family *Bartogtaviriformidae* (from *Bartonella*, infix -*gta*-, and family-
313 specific suffix -*viriformidae*). For now, we argue that at least one bartogtaviriformid genus ought
314 to be established: *Bartonegtaviriform* (named after BaGTA host genus *Bartonella*, infix -*gta*-,
315 and genus-specific suffix -*viriform*) including one new species, *Bartonegtaviriform andersoni*
316 (species epithet to honor GTA researcher Burt Anderson, who first discovered BaGTA particles
317 [Anderson et al. 1994]) for BaGTA.
318
319 **GTAs of spirochaetes**
320 A GTA originally called virus of *Serpulina hyodysenteriae* 1 (VSH-1) was identified in
321 *Brachyspira* (formerly *Serpulina*) *hyodysenteriae* (*Spirochaetota*: *Spirochaetia*: *Brachyspirales*:

322 *Brachyspiraceae*) (Humphrey et al. 1997). In accordance with the nomenclature rules established
323 here, we suggest renaming this GTA to Brachyspira hyodysenteriae gene transfer agent
324 (BhGTA). The structural gene cluster responsible for production of BhGTA particles—i.e., the
325 BhGTA "genome"—is 16.3 kb in length (Matson et al. 2005) (**Figure 5** and **Supplementary**
326 **Table S6**).

327 BhGTA particles have a head diameter of 45 nm and a flexible non-contractile tail of 65
328 nm (Humphrey et al. 1997). Like other GTAs, BhGTA is unable to package and transfer its
329 entire genome, given the limiting capacity of 7.5 kb (Humphrey et al. 1997; Matson et al. 2005).
330 Restriction enzyme digests of the packaged DNA and the range of marker genes that can be
331 transferred by BhGTA particles suggest that they package any region of the *B. hyodysenteriae*
332 genome (Humphrey et al. 1997) without an obvious bias for the genomic region that encodes
333 BhGTA. The induction of BhGTA particle production by DNA-damaging agents, such as
334 mitomycin C and antibiotics, results in large-scale lysis of cells (Stanton et al. 2008). However,
335 the proportion of *B. hyodysenteriae* cells in a population that naturally produce and release
336 BhGTA particles has not been quantified. BhGTA particles are capable of transferring
337 antimicrobial resistance genes within the bacterial population (Stanton et al. 2008), pointing at
338 possible selective advantages of maintaining the capability of BhGTA particle production.

339 Homologs of genes in the BhGTA genome were found in the genomes of other members
340 of the genus *Brachyspira*, but there is no gene synteny in their organization (Motro et al. 2009).
341 Unlike in rhodogtaviriformids and bartogtaviriformids, an endolysin-encoding gene is the only
342 gene in the BhGTA genome that has a significant sequence similarity to caudoviricete genes in
343 the National Center for Biotechnology Information (NCBI) Reference Sequence (RefSeq)
344 database (accessed in May 2022). Some genes encoding the BhGTA particle proteins were

345 experimentally validated (including endolysin), and the particles structurally resemble those of
346 caudoviricetes (Matson et al. 2005). Therefore, the absence of their homologs in the viral RefSeq
347 database is likely due to the limited sampling of the virosphere.

348 In the endolysin phylogeny, the *Brachyspira* homologs group together and are separated
349 from all caudoviricetes by a long branch (with 100% bootstrap support; **Figure 6A**).
350 Additionally, the *B. hyodysenteriae* genome encodes a single copy of an identifiable *terL* gene,
351 which is located outside of the currently delineated BhGTA genome. Homologs of this *terL* gene
352 are also present in a single copy in genomes of other *Brachyspira* bacteria that encode BhGTA-
353 like MCPs. These homologs are highly conserved, with pairwise amino-acid identities of 81–
354 100%. In a phylogenetic tree, the *Brachyspira* TerLs are separated from all caudoviricete TerLs
355 by a longer branch (with 100% bootstrap support; **Figure 6B**). Although the role of this TerL
356 homolog in the BhGTA lifecycle has not been experimentally validated, the presence of the
357 encoding gene as the only identifiable *terL* in the *Brachyspira* genomes, its high degree of
358 conservation within the *Brachyspira* genus and its divergence from the related caudoviricete
359 sequences support its potential involvement in the packaging of DNA into the BhGTA particles.
360 Based on comparison of *Brachyspira* GTA and host genes, GTAs have co-diversified with
361 *Brachyspira* (Motro et al. 2009).

362 Together, these results justify the classification of BhGTA and BhGTA-like GTAs in a
363 common viriform taxon, family *Brachygtaviriformidae* (form *Brachyspira*, infix -*gta*-, and
364 family-specific suffix -*viriformidae*). For now, we argue that at least one brachygtaviriformid
365 genus ought to be established: *Brachyspigtaviriform* (named after BhGTA host genus
366 *Brachyspira*, infix -*gta*-, and genus-specific suffix -*viriform*) to include one new species,

367 *Brachyspigtaviriform stantoni* (species epithet to honor GTA researcher Thaddeus Stanton, who
368 first discovered BhGTA particles [Humphrey et al. 1997]) for BhGTA.
369
370 **Independent origins of the three GTAs**
371 Genes from the genomes of these three GTAs are either not homologous or too divergent to have
372 significant sequence similarity in BLASTP searches of the encoded proteins. For example,
373 pairwise amino-acid identity of TerLs, which is one of the most conserved GTA and
374 caudoviricete proteins, is 14–20% among RcGTA, BaGTA, and BhGTA. Nevertheless, an
375 iterative clustering-alignment-phylogeny procedure (Wolf et al. 2018) established the homology
376 among known TerL proteins that include RcGTA, BaGTA, and putative BhGTA TerLs
377 (Esterman et al. 2021). The evolutionary history of RcGTA-like, BaGTA-like, putative BhGTA-
378 like TerLs, and their closest known caudoviricete homologs (**Figure 7**) demonstrates that GTA-
379 like TerLs appear in three distinct clades within viral TerLs. Based on this phylogenetic
380 evidence, we propose that these three GTA clades are a result of three independent exaptation
381 events. Therefore, just like viruses (which are classified in at least six unrelated realms), GTA
382 viriforms are polyphyletic.
383
384 **DISCUSSION**
385 Based on the evolutionary differences between GTA and caudoviricete genes encoding well-
386 conserved proteins and on morphological differences of GTA particles, we propose three
387 families for these GTAs. The greatest number of functionally confirmed and putative GTAs are
388 in the alphaproteobacterial type I clade, which, for now, is proposed to be a family
389 *Rhodogtaviriformidae* that includes at least four genera. The members of this family are

390 currently restricted to a single cellular order (*Rhodobacterales*). The TerLs and MCPs of these
391 RcGTA-like GTAs and alphaproteobacterial type II GTAs (*Bartogtaviriformidae*) are clearly
392 distinguishable from each other and their caudoviricete homologs and evolve at a slower rate
393 (**Figure 2** and **Figure 4**) (Esterman et al. 2021; Shakya et al. 2017). The spirochaete GTAs
394 (*Brachygtaviriformidae*) are more difficult to distinguish from caudoviricetes due to a lack of
395 available viral representatives in GenBank for all but one experimentally validated BhGTA gene.
396 Nevertheless, both the experimentally validated BhGTA endolysin and the putative BhGTA
397 TerL and their *Brachyspira* homologs also form a well-supported cluster distinct from
398 caudoviricete lineages; moreover, brachygtaviriformid TerLs evolve at a slower rate than their
399 spirochete homologs (**Figure 6B**). As in the case with the experimentally validated RcGTA, the
400 "genome" of BhGTA is also likely dispersed across multiple loci.
401 Analyses of environmental samples and genome sequences suggest the existence of a
402 large number of GTAs, especially those related to the rhodogtaviriformids (Biers et al. 2008; Fu
403 et al. 2010; McDaniel et al. 2010; Zhao et al. 2009). In a genome-wide screen of 1,423
404 alphaproteobacterial genomes, 57.5% were found to encode RcGTA-like "genomes", which are
405 often annotated as either intact or incomplete prophages (Kogay et al. 2019). The great majority
406 of RcGTA-like genes in alphaproteobacterial genomes are associated with bacteria for which a
407 GTA-based gene-transfer activity has not been documented, and it is possible that some of these
408 RcGTA-like genes may not be expressed to produce functional particles. Therefore, we have
409 restricted our proposal to those GTAs that have been shown to be functional. However, we
410 speculate that at least some (and perhaps many) of these GTA-like gene clusters will be shown to
411 produce functional GTAs that will need to be classified.

412 Based on the evolutionary history of TerL proteins (**Figure 7**), it is likely that the
413 proposed three GTA families had distinct caudoviricete progenitors. Eventual deduction of the
414 relatives of these progenitors may make it possible (or necessary) to include these GTA families
415 in the virus class *Caudoviricetes*, thereby creating an overarching taxon for distinct MGEs
416 (viruses and viriforms). Since the exaptation events, however, the three families have evolved as
417 part of the host genomes (Esterman et al. 2021; Lang et al. 2002; Lang and Beatty 2007; Shakya
418 et al. 2017), in the case of the rhodogtaviriformids for hundreds of millions of years (Shakya et
419 al. 2017). As a result, GTAs effectively became a component of cellular genomes, integrated into
420 cellular regulatory circuits that also control processes such as motility, quorum sensing,
421 extracellular polysaccharide synthesis, and biofilm formation (Lang et al. 2017; Pallegar et al.
422 2020; Shimizu et al. 2022). There is also mounting evidence that GTA genes experience
423 selective pressures to be maintained in their host genomes (Kogay et al. 2020; Lang et al. 2012).
424 Although the fitness benefits associated with GTA production remain to be elucidated, the time
425 is now ripe to have the known GTAs officially recognized and classified as specific viriforms.
426 We recognize this step as the initiation of a taxonomic framework that undoubtedly will rapidly
427 expand and change in the future.
428

429 **METHODS**

430 To identify alphaproteobacterial type I GTAs, we searched for RcGTA-like sequences in 1,248
431 complete alphaproteobacterial genomes extracted from the NCBI RefSeq database (accessed in
432 October 2020) using GTA-Hunter (Kogay et al. 2019). We identified 503 genomes that
433 contained at least six RcGTA homologs in the same genetic neighborhood and had both *g2*
434 (encoding TerL) and *g5* (encoding HK97-MCP) genes. To remove redundancy, we clustered

435 genomes into the operational taxonomic units (OTUs) using an average nucleotide identity
436 threshold of 95%. From all genomes within an OTU, we selected one genome with the largest
437 number of GTA genes. This strategy resulted in 290 representative GTAs selected for further
438 analysis. We identified the closest viral homologs of the TerL and HK97-MCP proteins from
439 these GTAs by conducting a BLASTP search (Altschul et al. 1997) of the RefSeq database
440 (accessed in March 2021) (O'Leary et al. 2016), using TerL and HK97-MCP proteins from
441 representative GTAs as queries, an e-value cutoff of 0.001, and query coverage of at least 50%.
442 Retrieved viral homologs with identical amino-acid sequences were removed from further
443 analyses. For both proteins, we aligned amino-acid sequences of GTA and virus homologs using
444 MAFFT v7.455 with -linsi option (Katoh and Standley 2013). We reconstructed phylogenetic
445 trees using IQ-TREE v2 (Minh et al. 2020), identifying the best substitution models using the
446 built-in ModelFinder (Kalyaanamoorthy et al. 2017). The selected models were LG+F+R9 and
447 LG+F+R7 for TerL and HK97-MCP datasets, respectively. Branch support values were assessed
448 using 1,000 ultrafast bootstrap replicates and a hill-climbing nearest-neighbor interchange search
449 for optimal trees (Hoang et al. 2018). Additionally, for both protein phylogenies, we
450 reconstructed a phylogenetic tree in IQ-TREE v2 (Minh et al. 2020) using a tree search that was
451 constrained by requiring all GTAs and all viruses to be separated by a branch. We compared the
452 resultant trees in unconstrained and constrained searches using the AU test (Shimodaira 2002), as
453 implemented in the IQ-TREE v2 program.
454 To identify alphaproteobacterial type II GTAs, we used the BaGTA TerL and HK97-
455 MCP sequences (accession numbers WP_034448260.1 and WP_011181178.1, respectively) as
456 queries in a BLASTP search against the 57 complete *Bartonella* genomes extracted from the
457 RefSeq database (accessed in May 2022). We restricted our search only to matches for which

458 BaGTA TerL and HK97-MCP homologs are in the same genomic neighborhood (defined as
459 being within 5 kb of each other). In genomes with multiple matches to the query protein, we
460 retained only the homolog with the highest BLASTP bit score. We clustered 57 genomes using a
461 95% average nucleotide identity (ANI) threshold and randomly selected one TerL and HK97-
462 MCP representative from each cluster for phylogenetic analysis. We identified caudoviricete
463 homologs by conducting a BLASTP search (e-value cutoff of 0.001, and query coverage of at
464 least 50%) against viral RefSeq database (accessed in May 2022). We performed phylogenetic
465 reconstructions as described above for alphaproteobacterial type I GTAs. The selected best
466 substitution models were LG+R6 and LG+G4 for TerL and HK97-MCP datasets, respectively.
467 To identify GTAs of spirochaetes, we used BhGTA's MCP sequence (GenBank
468 accession number WP_012671344.1) as a query in a BLASTP search (with an e-value cutoff of
469 0.001 and query coverage of at least 50%) against the 13 complete *Brachyspira* genomes
470 extracted from the RefSeq database (accessed in May 2022). We used TerL of *B. hyodysenteriae*
471 (GenBank accession number WP_012671469.1) and endolysin protein of *B. hyodysenteriae*
472 (GenBank accession number WP_012671356.1) as queries in a BLASTP search (with an e-value
473 cutoff of 0.001 and query coverage of at least 50%) against the same set of 13 genomes. For
474 endolysins, we only retained matches that co-localized within the BhGTA region on the
475 chromosome. We clustered 13 genomes using a 95% ANI threshold and randomly chose one
476 TerL and endolysin representative from each cluster for phylogenetic analyses. We identified
477 caudoviricete homologs by doing BLASTP searches (e-value cutoff of 0.001 and query coverage
478 of at least 50%) against the viral RefSeq database (accessed in May 2022). We performed
479 phylogenetic reconstructions as described above for the alphaproteobacterial type I GTAs. The

480 selected best substitution models were VT+F+R3 and WAG+R6 for TerL and endolysin datasets,
481 respectively.
482 To reconstruct the phylogeny that includes all three clades of GTAs, we combined all
483 TerL homologs extracted in the above-described procedures into one dataset. We aligned the
484 TerL sequences using MAFFT v7.455 with -dash option (Rozewicki et al. 2019) and trimmed the
485 obtained alignment using ClipKIT with -gappy option (Steenwyk et al. 2020). We computed the
486 phylogenetic tree using IQ-TREE v2 (Minh et al. 2020) as described above with the LG+F+R10
487 substitution model selected by ModelFinder. We rooted the tree using a larger TerL phylogeny
488 presented in Esterman et al. (2021).
489 We visualized all phylogenetic trees in iTOL v6 (Letunic and Bork 2021).
490

491 **ACKNOWLEDGEMENTS**
492 We thank Anya Crane and Jiro Wada (Integrated Research Facility at Fort Detrick, Division of
493 Clinical Research, National Institute of Allergy and Infectious Diseases, National Institutes of
494 Health) for editing the manuscript and figures, respectively.
495

496 **SUPPLEMENTARY DATA**
497 Amino-acid sequence alignments in FASTA format and phylogenetic trees in Newick format for
498 trees in **Figures 2**, **4**, **6**, and **7** are available in **Supplementary_Data.zip** file.
499

500 **DATA AVAILABILITY**

501 All data used in this manuscript were retrieved from publicly available GenBank databases, as
502 described in the Methods. The accession numbers of database records used in the phylogenetic
503 analyses can be found in alignments that are included in the **Supplementary Data**.

504

505 **FUNDING**

506 This work was supported in part through Laulima Government Solutions, LLC, prime contract
507 with the U.S. National Institute of Allergy and Infectious Diseases (NIAID) under Contract No.
508 HHSN272201800013C; J.H.K. performed this work as an employee of Tunnell Government
509 Services (TGS), a subcontractor of Laulima Government Solutions, LLC, under Contract No.
510 HHSN272201800013C. This work was also supported by the Simons Foundation Investigator in
511 Mathematical Modeling of Living Systems award #327936 (O.Z.), the Canadian Natural
512 Sciences and Engineering Research Council (NSERC) award RGPIN 2018-03898 (J.T.B.),
513 NSERC award RGPIN-2017-04636 (A.S.L.). S.K. was partially supported by funding from the
514 Memorial University of Newfoundland School of Graduate Studies.

515 The views and conclusions contained in this document are those of the authors and
516 should not be interpreted as necessarily representing the official policies, either expressed or
517 implied, of the U.S. Department of Health and Human Services or of the institutions and
518 companies affiliated with the authors, nor does mention of trade names, commercial products, or
519 organizations imply endorsement by the U.S. Government.

520

521 **CONFLICTS OF INTEREST STATEMENT**

522 The authors declare no conflicts of interest.

523 **FIGURES**

524 **Figure 1. Genome of Rhodobacter capsulatus gene transfer agent (RcGTA).** Genes (arrows)
525 are depicted to scale, in their locations in the host genome (*R. capsulatus*). Exact coordinates of
526 the RcGTA genes, their locus tags, and their functional annotations are listed in **Supplementary**
527 **Table S1**.

528

529 **Figure 2. Maximum Likelihood phylogenies of (A) large terminase (TerL) subunits and (B)**
530 **HK97 major capsid protein (HK97-MCP) sequences of rhodogtaviriformids and their**
531 **closest known caudoviricete homologs.** Alphaproteobacterial type I gene transfer agent (GTA)
532 (rhodogtaviriformid) lineages are shown in orange. Caudoviricete lineages that are nested within
533 GTA lineages are shown in dashed black lines. Other caudoviricete lineages are shown in solid
534 black lines. Bootstrap support values are shown only for selected branches. Scale bars represent
535 substitutions per site. DsGTA, Dinoroseobacter shibae gene transfer agent; GTA, gene transfer
536 agent; RcGTA, Rhodobacter capsulatus gene transfer agent; RpGTA, Ruegeria pomeroyi gene
537 transfer agent; RsGTA, Rhodovolum sulfidophilum, gene transfer agent.

538

539 **Figure 3. Genome of Bartonella gene transfer agent (BaGTA).** Genes (arrows) are depicted to
540 scale, in their locations in the host genome (*B. henselae*). Exact coordinates of the BaGTA genes,
541 their locus tags, and their functional annotations are listed in **Supplementary Table S5**.

542

543 **Figure 4. Maximum Likelihood phylogenies of (A) large terminase (TerL) subunits and (B)**
544 **HK97 major capsid protein (HK97-MCP) sequences of bartogtaviriformids and their**
545 **closest known caudoviricete homologs.** Alphaproteobacterial type II gene transfer agent (GTA)

546 (bartogtaviriformid) lineages are shown in blue. Caudoviricete lineages are shown in black. Two
547 nearly identical caudoviricete lineages that are nested within GTA lineages are shown in dashed
548 black lines. A bootstrap support value is shown only for the branch separating GTA and
549 caudoviricete sequences. Scale bars indicate substitutions per site. BaGTA, Bartonella gene
550 transfer agent; GTA, gene transfer agent.
551
552 **Figure 5. Genome of Brachyspira hyodysenteriae gene transfer agent (BhGTA).** Genes
553 (arrows) are depicted to scale, in their locations in the host genome (*B. hyodysenteriae*). Exact
554 coordinates of the BhGTA genes, their locus tags, and their functional annotations are listed in
555 **Supplementary Table S6**.
556
557 **Figure 6. Maximum Likelihood phylogenies of (A) endolysin and (B) the putative large**
558 **terminase (TerL) subunits of brachygtaviriformids and their closest known caudoviricete**
559 **homologs.** *Brachyspira* gene transfer agent (GTA) lineages are shown in purple. Caudoviricete
560 lineages are shown in black. A bootstrap support value is shown only for the branch separating
561 GTA and caudoviricete sequences. Scale bar indicates substitutions per site. BhGTA,
562 Brachyspira hyodysenteriae gene transfer agent; GTA, gene transfer agent.
563
564 **Figure 7. Maximum Likelihood phylogeny of the large terminase (TerL) subunits of three**
565 **major clades of GTAs and their closest known caudoviricete homologs.** This tree includes all
566 TerL homologs from Figures 2A, 4A, and 6A, using their color coding. Bootstrap support values
567 are shown only for the branches separating three GTA clades and their closest caudoviricete
568 sequences. Scale bar indicates substitutions per site. GTA, gene transfer agent.

569 **REFERENCES**

570 Altschul, S. F., et al. (1997), 'Gapped BLAST and PSI-BLAST: a new generation of protein
571 database search programs', *Nucleic Acids Res,* 25 (17), 3389-402.

572 Anderson, B., et al. (1994), 'Bacteriophage-like particle of *Rochalimaea henselae*', *Mol*
573 *Microbiol,* 13 (1), 67-73.

574 Barbian, K. D. and Minnick, M. F. (2000), 'A bacteriophage-like particle from *Bartonella*
575 *bacilliformis*', *Microbiology (Reading),* 146 (Pt 3), 599-609.

576 Bárdy, P., et al. (2020), 'Structure and mechanism of DNA delivery of a gene transfer agent', *Nat*
577 *Commun,* 11 (1), 3034.

578 Berglund, E. C., et al. (2009), 'Run-off replication of host-adaptability genes is associated with
579 gene transfer agents in the genome of mouse-infecting *Bartonella grahamii*', *PLoS Genet,*
580 5 (7), e1000546.

581 Bertani, G. (1999), 'Transduction-like gene transfer in the methanogen *Methanococcus voltae*', *J*
582 *Bacteriol,* 181 (10), 2992-3002.

583 Biers, E. J., et al. (2008), 'Occurrence and expression of gene transfer agent genes in marine
584 bacterioplankton', *Appl Environ Microbiol,* 74 (10), 2933-9.

585 Brimacombe, C. A., et al. (2013), 'Quorum-sensing regulation of a capsular polysaccharide
586 receptor for the *Rhodobacter capsulatus* gene transfer agent (RcGTA)', *Mol Microbiol,*
587 87 (4), 802-17.

588 Darboux, I., Cusson, M., and Volkoff, A. N. (2019), 'The dual life of ichnoviruses', *Curr Opin*
589 *Insect Sci,* 32, 47-53.

590 Drezen, J.-M., et al. (2017), 'Endogenous viruses of parasitic wasps: variations on a common
591 theme', *Curr Opin Virol,* 25, 41-48.

592 Esterman, E. S., et al. (2021), 'Evolution of DNA packaging in gene transfer agents', *Virus Evol,*
593 7 (1), veab015.
594 Felsenstein, Joseph (1978), 'Cases in which parsimony or compatibility methods will be
595 positively misleading', *Syst Biol,* 27 (4), 401-10.
596 Fogg, P. C. M., Westbye, A. B., and Beatty, J. T. (2012), 'One for all or all for one:
597 heterogeneous expression and host cell lysis are key to gene transfer agent activity in
598 *Rhodobacter capsulatus*', *PLoS One,* 7 (8), e43772.
599 Fu, Yunyun, et al. (2010), 'High diversity of *Rhodobacterales* in the subarctic North Atlantic
600 Ocean and gene transfer agent protein expression in isolated strains', *Aquat Microb Ecol,*
601 59, 283–93.
602 Gauthier, J., Drezen, J.-M., and Herniou, E. A. (2018), 'The recurrent domestication of viruses:
603 major evolutionary transitions in parasitic wasps', *Parasitology,* 145 (6), 713-23.
604 Guy, L., et al. (2013), 'A gene transfer agent and a dynamic repertoire of secretion systems hold
605 the keys to the explosive radiation of the emerging pathogen *Bartonella*', *PLoS Genet,* 9
606 (3), e1003393.
607 Herniou, E. A., et al. (2013), 'When parasitic wasps hijacked viruses: genomic and functional
608 evolution of polydnaviruses', *Philos Trans R Soc Lond B Biol Sci,* 368 (1626), 20130051.
609 Hoang, D. T., et al. (2018), 'UFBoot2: improving the Ultrafast Bootstrap approximation', *Mol*
610 *Biol Evol,* 35 (2), 518-22.
611 Humphrey, S. B., et al. (1997), 'Purification and characterization of VSH-1, a generalized
612 transducing bacteriophage of *Serpulina hyodysenteriae*', *J Bacteriol,* 179 (2), 323-9.

613 Hynes, A. P., et al. (2012), 'DNA packaging bias and differential expression of gene transfer
614 agent genes within a population during production and release of the *Rhodobacter*
615 *capsulatus* gene transfer agent, RcGTA', *Mol Microbiol,* 85 (2), 314-25.
616 Hynes, A. P., et al. (2016), 'Functional and evolutionary characterization of a gene transfer
617 agent's multilocus "genome"', *Mol Biol Evol,* 33 (10), 2530-43.
618 International Committee on Taxonomy of Viruses (2022), 'ICTV Code. The International Code
619 of Virus Classification and Nomenclature (ICVCN). March 2021.
620 https://ictv.global/about/code '.
621 Kalyaanamoorthy, S., et al. (2017), 'ModelFinder: fast model selection for accurate phylogenetic
622 estimates', *Nat Methods,* 14 (6), 587-89.
623 Katoh, K. and Standley, D. M. (2013), 'MAFFT multiple sequence alignment software version 7:
624 improvements in performance and usability', *Mol Biol Evol,* 30 (4), 772-80.
625 Kogay, R., et al. (2020), 'Selection for reducing energy cost of protein production drives the GC
626 content and amino acid composition bias in gene transfer agents', *mBio,* 11 (4), e01206-
627 20.
628 Kogay, R., et al. (2019), 'Machine-learning classification suggests that many
629 alphaproteobacterial prophages may instead be gene transfer agents', *Genome Biol Evol,*
630 11 (10), 2941-53.
631 Koonin, E. V. and Krupovic, M. (2018), 'The depths of virus exaptation', *Curr Opin Virol,* 31, 1-
632 8.
633 Koonin, E. V., et al. (2021), 'Viruses defined by the position of the virosphere within the
634 replicator space', *Microbiol Mol Biol Rev,* 85 (4), e0019320.

635 Koonin, E. V., et al. (2020), 'Global organization and proposed megataxonomy of the virus
636 world', *Microbiol Mol Biol Rev,* 84 (2), e00061-19.
637 Kuhn, J. H., et al. (2020), 'Expand, amend, and emend the International Code of Virus
638 Classification and Nomenclature (ICVCN; "the Code") and the Statutes to clearly define
639 the remit of the ICTV. International Committee on Taxonomy of Viruses (ICTV)
640 TaxoProp 2020.005G.R.Code_and_Statute_Change.
641 https://ictv.global/filebrowser/download/7146 '.
642 Kuhn, J. H., et al. (2021), 'Rename the family *Polydnaviridae* (as *Polydnaviriformidae*), rename
643 the genus *Bracovirus* (as *Bracoviriform*) and rename all polydnaviriformid species to
644 comply with the newly ICTV-mandated binomial format. International Committee on
645 Taxonomy of Viruses (ICTV) TaxoProp
646 2021.006D.R.Polydnaviriformidae_1renfam_3rensp.
647 https://ictv.global/filebrowser/download/4986 '.
648 Lang, A. S. and Beatty, J. T. (2007), 'Importance of widespread gene transfer agent genes in α-
649 proteobacteria', *Trends Microbiol,* 15 (2), 54-62.
650 Lang, A. S., Taylor, T. A., and Beatty, J. T. (2002), 'Evolutionary implications of phylogenetic
651 analyses of the gene transfer agent (GTA) of *Rhodobacter capsulatus*', *J Mol Evol,* 55
652 (5), 534-43.
653 Lang, A. S., Zhaxybayeva, O., and Beatty, J. T. (2012), 'Gene transfer agents: phage-like
654 elements of genetic exchange', *Nat Rev Microbiol,* 10 (7), 472-82.
655 Lang, A. S., Westbye, A. B., and Beatty, J. T. (2017), 'The distribution, evolution, and roles of
656 Ggne transfer agents in prokaryotic genetic exchange', *Annu Rev Virol,* 4 (1), 87-104.

657 Letunic, I. and Bork, P. (2021), 'Interactive Tree Of Life (iTOL) v5: an online tool for
658 phylogenetic tree display and annotation', *Nucleic Acids Res,* 49 (W1), W293-w96.
659 Marrs, B. (1974), 'Genetic recombination in *Rhodopseudomonas capsulata*', *Proc Natl Acad Sci*
660 *U S A,* 71 (3), 971-3.
661 Matson, E. G., et al. (2005), 'Identification of genes of VSH-1, a prophage-like gene transfer
662 agent of *Brachyspira hyodysenteriae*', *J Bacteriol,* 187 (17), 5885-92.
663 McDaniel, L. D., et al. (2010), 'High frequency of horizontal gene transfer in the oceans',
664 *Science,* 330 (6000), 50.
665 Minh, B. Q., et al. (2020), 'IQ-TREE 2: new models and efficient methods for phylogenetic
666 inference in the genomic era', *Mol Biol Evol,* 37 (5), 1530-34.
667 Motro, Y., et al. (2009), 'Identification of genes associated with prophage-like gene transfer
668 agents in the pathogenic intestinal spirochaetes *Brachyspira hyodysenteriae*, *Brachyspira*
669 *pilosicoli* and *Brachyspira intermedia*', *Vet Microbiol,* 134 (3-4), 340-5.
670 Nagao, N., et al. (2015), 'The gene transfer agent-like particle of the marine phototrophic
671 bacterium *Rhodovulum sulfidophilum*', *Biochem Biophys Rep,* 4, 369-74.
672 O'Leary, N. A., et al. (2016), 'Reference sequence (RefSeq) database at NCBI: current status,
673 taxonomic expansion, and functional annotation', *Nucleic Acids Res,* 44 (D1), D733-45.
674 Pallegar, P., et al. (2020), 'Cyclic di-GMP-mediated regulation of gene transfer and motility in
675 *Rhodobacter capsulatus*', *J Bacteriol,* 202 (2).
676 Petersen, J. M., et al. (2022), 'The naked truth: An updated review on nudiviruses and their
677 relationship to bracoviruses and baculoviruses', *J Invertebr Pathol,* 189, 107718.
678 Québatte, M., et al. (2017), 'Gene transfer agent promotes evolvability within the fittest
679 subpopulation of a bacterial pathogen', *Cell Syst,* 4 (6), 611-21.e6.

680 Rapp, B. J. and Wall, J. D. (1987), 'Genetic transfer in *Desulfovibrio desulfuricans*', *Proc Natl*
681 *Acad Sci U S A,* 84 (24), 9128-30.
682 Rozewicki, J., et al. (2019), 'MAFFT-DASH: integrated protein sequence and structural
683 alignment', *Nucleic Acids Res,* 47 (W1), W5-w10.
684 Shakya, M., Soucy, S. M., and Zhaxybayeva, O. (2017), 'Insights into origin and evolution of α-
685 proteobacterial gene transfer agents', *Virus Evol,* 3 (2), vex036.
686 Shimizu, T., et al. (2022), 'Persulfide-responsive transcription factor SqrR regulates gene transfer
687 and biofilm formation via the metabolic modulation of cyclic di-GMP in *Rhodobacter*
688 *capsulatus*', *Microorganisms,* 10 (5), 908.
689 Shimodaira, H. (2002), 'An approximately unbiased test of phylogenetic tree selection', *Syst Biol,*
690 51 (3), 492-508.
691 Stanton, T. B., et al. (2008), 'Collateral effects of antibiotics: carbadox and metronidazole induce
692 VSH-1 and facilitate gene transfer among *Brachyspira hyodysenteriae* strains', *Appl*
693 *Environ Microbiol,* 74 (10), 2950-6.
694 Steenwyk, J. L., et al. (2020), 'ClipKIT: a multiple sequence alignment trimming software for
695 accurate phylogenomic inference', *PLoS Biol,* 18 (12), e3001007.
696 Strand, M. R. and Burke, G. R. (2020), 'Polydnaviruses: evolution and function', *Curr Issues Mol*
697 *Biol,* 34, 163-82.
698 Tamarit, D., et al. (2018), 'Origin and evolution of the *Bartonella* gene transfer agent', *Mol Biol*
699 *Evol,* 35 (2), 451-64.
700 Thézé, J., et al. (2011), 'Paleozoic origin of insect large dsDNA viruses', *Proc Natl Acad Sci U S*
701 *A,* 108 (38), 15931-5.

702 Tomasch, J., et al. (2018), 'Packaging of *Dinoroseobacter shibae* DNA into gene transfer agent
703 particles is not random', *Genome Biol Evol,* 10 (1), 359-69.
704 Umemori, E., et al. (1992), 'A phage in *Bartonella bacilliformis*', *Microbiol Immunol,* 36 (7),
705 731-6.
706 Walker, P. J., et al. (2021), 'Changes to virus taxonomy and to the International Code of Virus
707 Classification and Nomenclature ratified by the International Committee on Taxonomy of
708 Viruses (2021)', *Arch Virol,* 166 (9), 2633-48.
709 --- (2022), 'Changes to virus taxonomy and to the International Code of Virus Classification and
710 Nomenclature ratified by the International Committee on Taxonomy of Viruses (2022)',
711 *Arch Virol*, in press (published online August 23, 2022), DOI:
712 https://doi.org/10.1007/s00705-022-05516-5.
713 Westbye, A. B., Beatty, J. T., and Lang, A. S. (2017a), 'Guaranteeing a captive audience:
714 coordinated regulation of gene transfer agent (GTA) production and recipient capability
715 by cellular regulators', *Curr Opin Microbiol,* 38, 122-29.
716 Westbye, A. B., et al. (2017b), 'The *Rhodobacter capsulatus* gene transfer agent is induced by
717 nutrient depletion and the RNAP omega subunit', *Microbiology (Reading),* 163 (9), 1355-
718 63.
719 Wolf, Y. I., et al. (2018), 'Origins and evolution of the global RNA virome', *mBio,* 9 (6), e02329-
720 18.
721 Yen, H. C., Hu, N. T., and Marrs, B. L. (1979), 'Characterization of the gene transfer agent made
722 by an overproducer mutant of *Rhodopseudomonas capsulata*', *J Mol Biol,* 131 (2), 157-
723 68.

724 Zhan, Y., et al. (2016), 'A novel roseobacter phage possesses features of podoviruses,
725 siphoviruses, prophages and gene transfer agents', *Sci Rep,* 6, 30372.
726 Zhao, Y., et al. (2009), 'Gene transfer agent (GTA) genes reveal diverse and dynamic
727 *Roseobacter* and *Rhodobacter* populations in the Chesapeake Bay', *ISME J,* 3 (3), 364-
728 73.
729

Figure 1

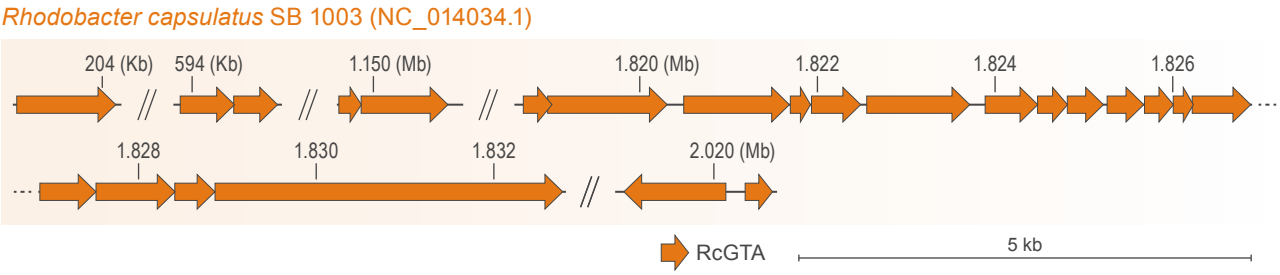

Rhodobacter capsulatus SB 1003 (NC_014034.1)
204 (Kb)
594 (Kb)
1.150 (Mb)
1.820 (Mb)
1.822
1.824
1.826
1.828
1.830
1.832
2.020 (Mb)
RcGTA
5 kb

Figure 2

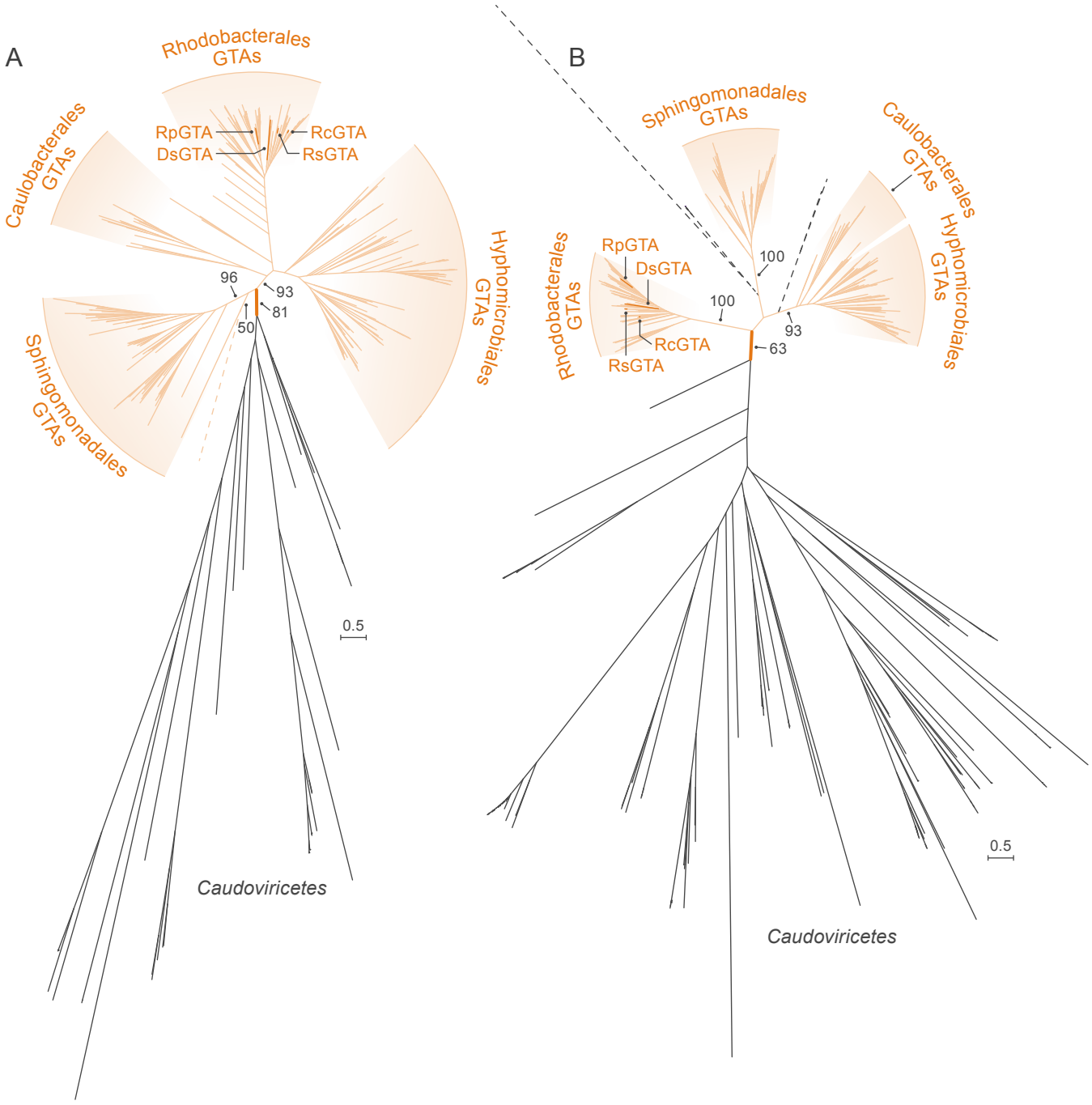

A
Rhodobacterales GTAs
Caulobacterales GTAs
RpGTA
DsGTA
RcGTA
RsGTA
96
93
81
50
Hyphomicrobiales GTAs
Sphingomonadales GTAs
0.5
Caudoviricetes
B
Sphingomonadales GTAs
Caulobacterales GTAs
Hyphomicrobiales GTAs
Rhodobacterales GTAs
RpGTA
DsGTA
RcGTA
RsGTA
100
100
93
63
0.5
Caudoviricetes

Figure 3

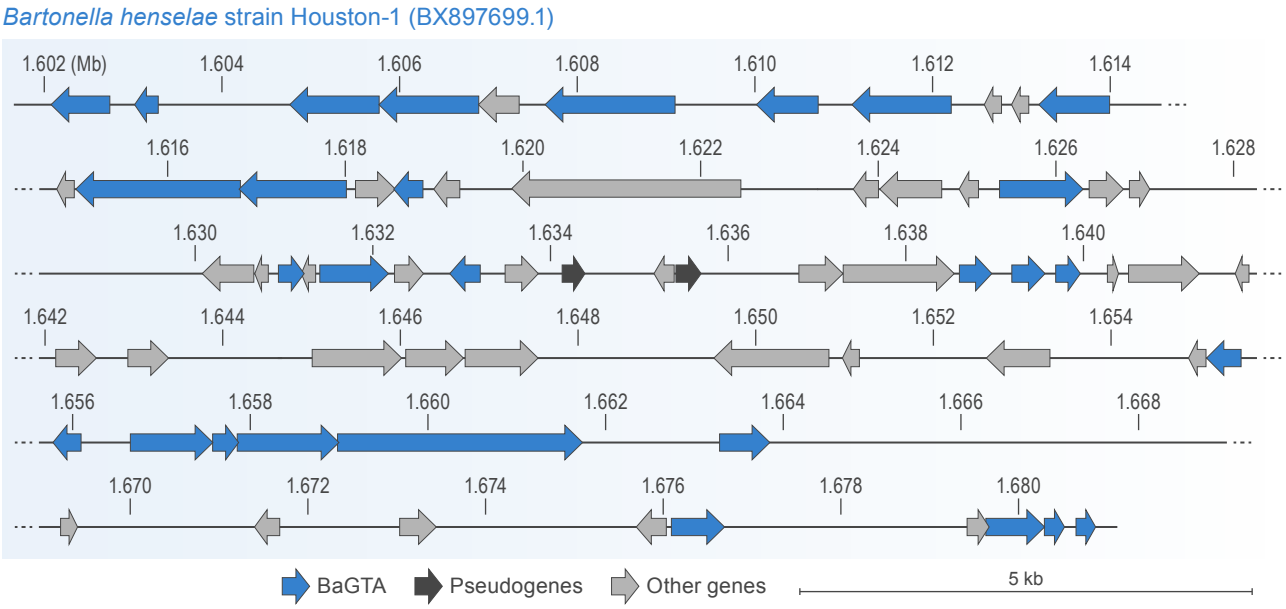

Bartonella henselae strain Houston-1 (BX897699.1)
1.602 (Mb)
1.604
1.606
1.608
1.610
1.612
1.614
1.616
1.618
1.620
1.622
1.624
1.626
1.628
1.630
1.632
1.634
1.636
1.638
1.640
1.642
1.644
1.646
1.648
1.650
1.652
1.654
1.656
1.658
1.660
1.662
1.664
1.666
1.668
1.670
1.672
1.674
1.676
1.678
1.680
BaGTA
Pseudogenes
Other genes
5 kb

Figure 4

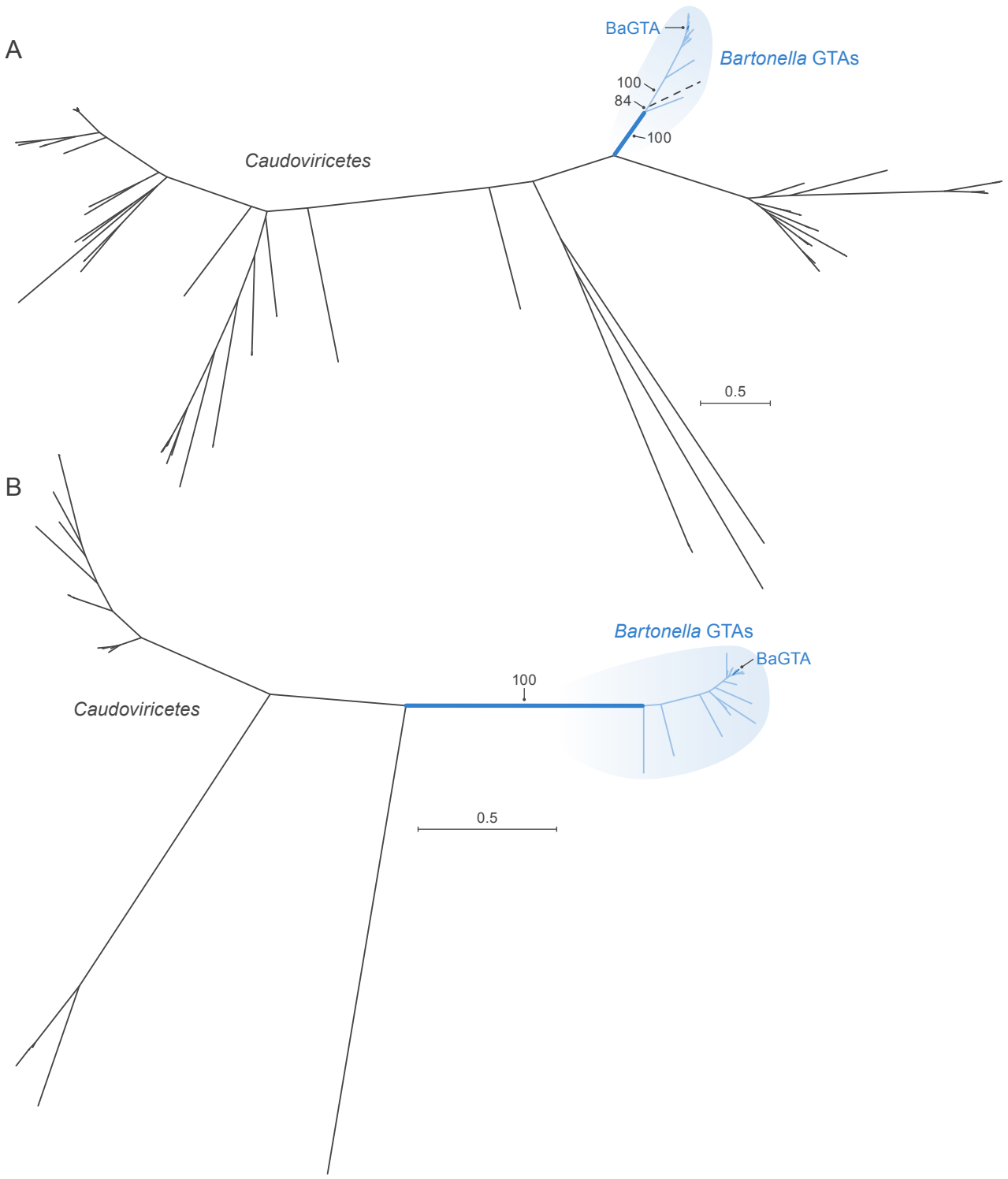

A
BaGTA
Bartonella GTAs
100
84
100
Caudoviricetes
0.5
B
Bartonella GTAs
BaGTA
100
Caudoviricetes
0.5

Figure 5

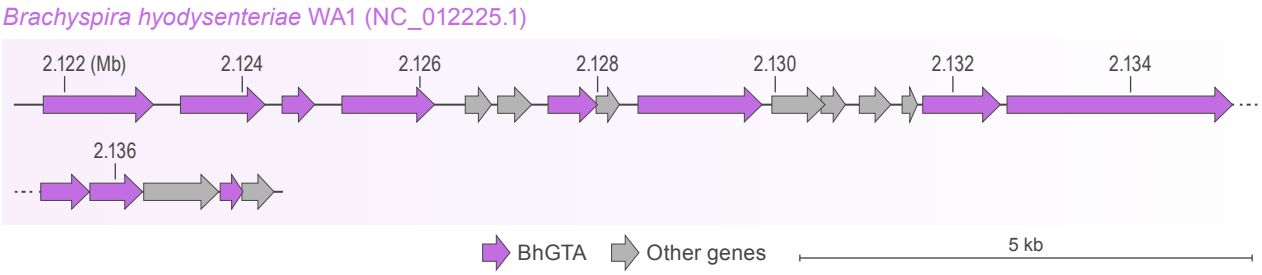

Brachyspira hyodysenteriae WA1 (NC_012225.1)
2.122 (Mb)
2.124
2.126
2.128
2.130
2.132
2.134
2.136
BhGTA
Other genes
5 kb

Figure 6

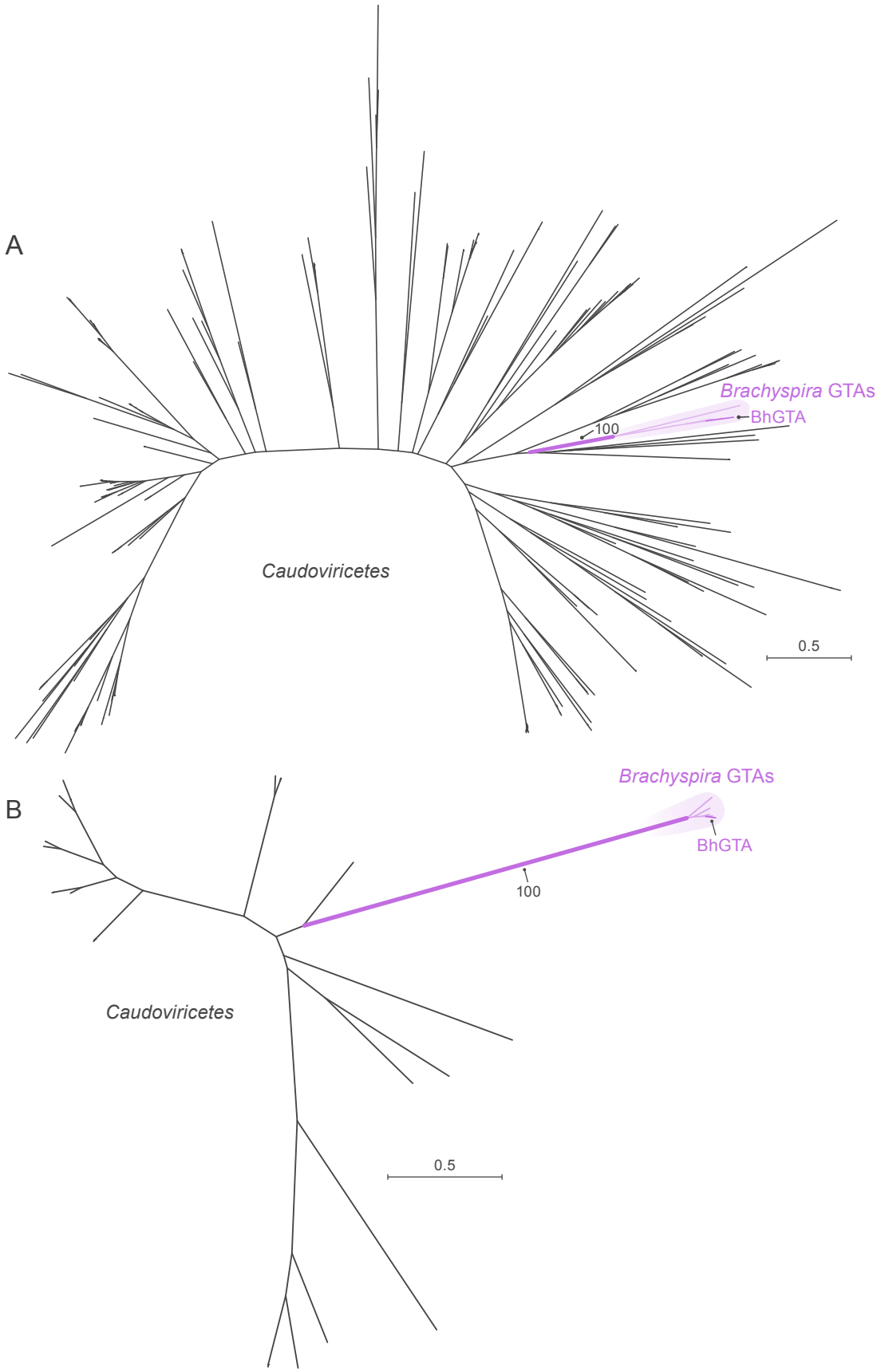
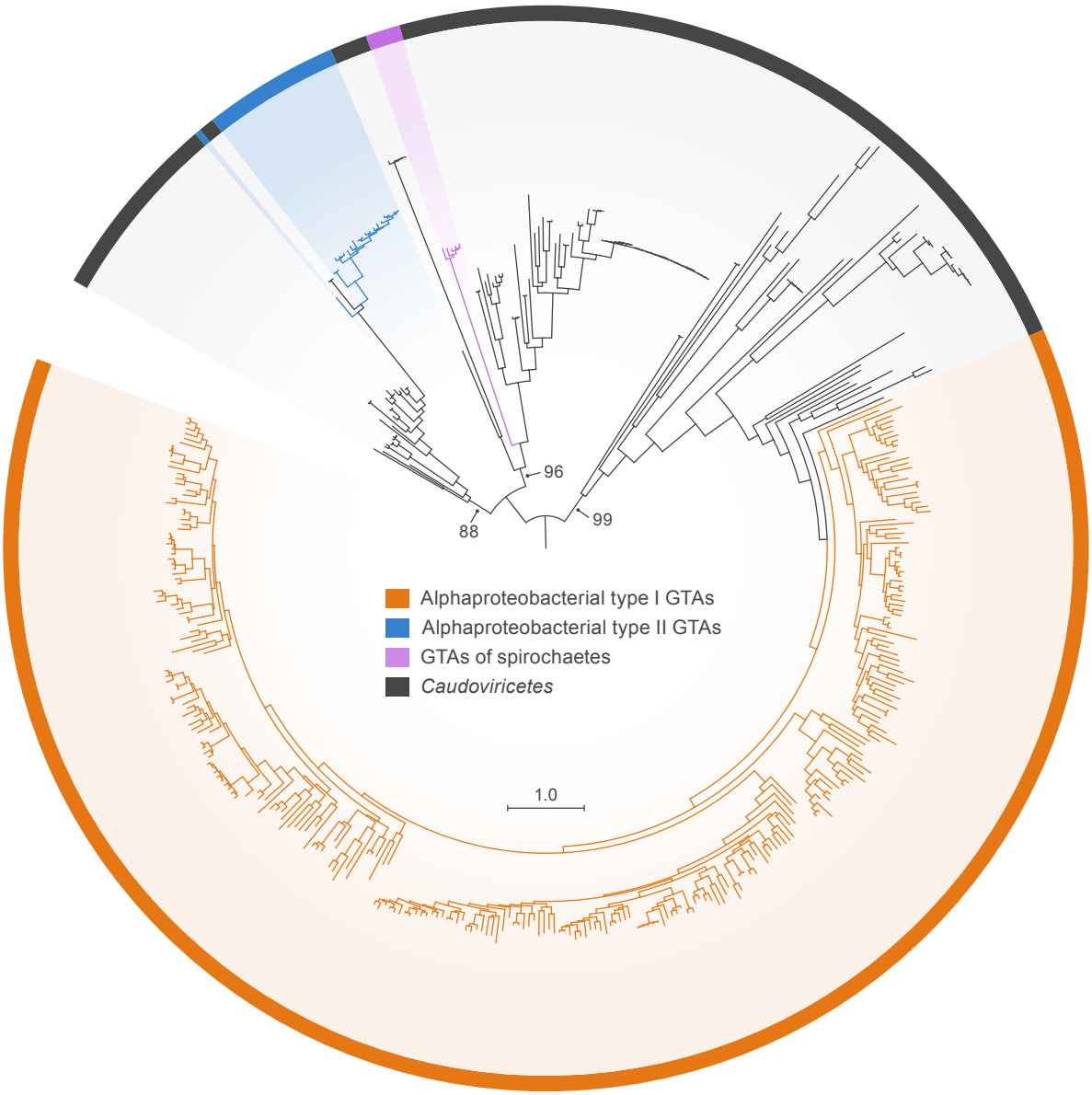

A
Brachyspira GTAs
BhGTA
100
Caudoviricetes
0.5
B
Brachyspira GTAs
BhGTA
100
Caudoviricetes
0.5

Figure 7


96
88
99
Alphaproteobacterial type I GTAs
Alphaproteobacterial type II GTAs
GTAs of spirochaetes
Caudoviricetes
1.0