Species boundaries in the *Agrobacterium tumefaciens* complex and multi-level modular evolution of their antibacterial type VI secretion system and tumor-inducing plasmids

Running Title: ***Agrobacterium* species boundaries and modular evolution**

Lin Chou[a#], Yu-Chen Lin[a#], Mindia Haryono[a], Mary Nia M. Santos[a,b,c], Shu-Ting Cho[a], Alexandra J. Weisberg[d], Chih-Feng Wu[a,d], Jeff H. Chang[d], Erh-Min Lai[a,b,e], Chih-Horng Kuo[a,b,e*]

[a] Institute of Plant and Microbial Biology, Academia Sinica, Taipei, Taiwan

[b] Molecular and Biological Agricultural Sciences Program, Taiwan International Graduate Program, National Chung-Hsing University and Academia Sinica, Taipei, Taiwan

[c] Graduate Institute of Biotechnology, National Chung-Hsing University, Taichung, Taiwan

[d] Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon, USA

[e] Biotechnology Center, National Chung-Hsing University, Taichung, Taiwan

[#]Equal contribution; ordered alphabetically

*Correspondence: Chih-Horng Kuo, chk@gate.sinica.edu.tw

Abstract

Many pathogenic bacteria are recognized as species complexes and uncertainties regarding the organization of their genetic diversity are challenges for research efforts. Within *Agrobacterium tumefaciens*, multiple genomospecies have been identified; however, the exact species boundaries are unclear, which causes chaos in nomenclature and hampers communication. In this work, we conducted targeted genome sequencing to achieve a comprehensive and balanced taxon sampling within this complex. Our results from genome-wide sequence identity, core genome phylogeny, and gene content not only supported that those recognized genomospecies are distinct biological entities but also identified novel genomospecies. Based on the fully resolved phylogeny, we further investigated the evolution of genes critical in *Agrobacterium* fitness and ecology. For the type VI secretion system (T6SS) involved in interbacterial competition, multiple losses and one horizontal gene transfer (HGT) event were inferred. For the tumor-inducing plasmids (pTi) and the pTi-encoded type IV secretion system (T4SS) that determine *Agrobacterium* phytopathogenicity, the evolution of these accessory replicons was decoupled from the chromosomes, thus contributing to another level of complexity. Intriguingly, for both T6SS and T4SS, genes that encode the structural components are highly conserved, whereas extensive diversity exists at multiple levels (i.e., between-species, within-species, intra-genome, and intra-gene) for genes that encode effectors and associated proteins. These findings suggest that opposite modes of selection may act on components conferring different functions within a system. In conclusion, this work provides insights into the genomic

diversification of these bacteria and sheds light on the modularity of their molecular evolution.

Importance

The alphaproteobacterium *Agrobacterium tumefaciens* is important as a plant pathogen and a critical biotechnology tool for genetic engineering. Based on comprehensive sampling within this species complex, this study utilized multiple approaches of genome analysis to delineate the species boundaries and resolved the phylogenetic relationships. The results shed light on the patterns of genomic divergence among a group of closely related lineages, particularly regarding how these lineages may be partitioned into distinct biological entities at the species level and above. Additionally, targeted in-depth analysis of those genetic elements involved in critical aspects of microbial fitness, including the secretion system for interbacterial competition and the oncogenic plasmid for pathogenicity against eukaryotic hosts, provides novel insights into the multi-level modularity and diverse modes of selection that govern the evolution of these components. This work improves our understanding of this study system and demonstrates a highly effective strategy applicable to other bacterial species complexes.

Keywords

Agrobacterium, genome, secretion system, virulence, plasmid, molecular evolution

69 **Introduction**

70 Investigation of species boundaries in bacteria is fundamentally important to
71 basic and applied microbiology because species is the basic unit for formal description
72 of biological entities and the key identifier for application of microbial resources (1, 2).
73 However, many bacterial groups, including several notable pathogens such as
74 *Pseudomonas* spp. (3) and *Ralstonia solanacearum* (4), exist in the form of species
75 complexes. Uncertainties regarding species boundaries within such groups could
76 hamper research, communication, and policy-making (e.g., quarantine regulations). In a
77 survey of > 20,000 bacterial genome sequences from 91 species belonging to 13 phyla,
78 21 were found to comprise multiple biological species based on barriers to homologous
79 recombination (5). This finding highlighted the importance of utilizing comparative
80 genomics to infer if certain species boundaries exist within a group of closely related
81 bacteria, and possibly also the need towards a standardized taxonomy (6, 7). Improved
82 understanding of the species boundaries can refine gene content analysis to provide
83 insights into their biology and facilitate the interpretation of experimental works on
84 microbial physiology or ecology.

85 In this study, we utilized the *Agrobacterium tumefaciens* species complex, also
86 commonly known as *Agrobacterium* biovar 1 (8), as the study system. These bacteria
87 are noted for their phytopathogenicity as the causative agent of crown gall disease that
88 affects over 90 plant families (9). More importantly, the development of *Agrobacterium*-
89 mediated transformation has provided a critical tool for genetic manipulation in plant
90 sciences and agricultural biotechnology (10, 11).

91 The extensive genetic diversity of *A. tumefaciens* has been noted prior to the
92 availability of genome sequences. Based on DNA-DNA hybridization and biochemical
93 characteristics (12), amplified fragment length polymorphism (AFLP) analysis (13–15),
94 and recombination frequencies (16), 10 genomospecies (i.e., G1-G9 and G13; also
95 known as genomic species, genomic groups, or genomovars) have been identified.
96 Later on, based on microarray hybridization for comparisons with the reference strain
97 C58, 196 G8-specific genes were found to be mostly clustered into seven islands, which
98 provided the basis for the description of G8 as a distinct species *Agrobacterium fabrum*
99 (17). However, because C58 in G8 is the standard strain for most *A. tumefaciens*
100 studies and the progenitor for nearly all of the disarmed strains used in *Agrobacterium*-
101 mediated transformation (11, 18, 19), the continuing mixed usage of these two species
102 names causes confusion. Additionally, the name *Agrobacterium radiobacter* refers to *A.*
103 *tumefaciens* G4 (15, 17), the entire *A. tumefaciens* complex (20), or the more divergent
104 *Agrobacterium* biovar 2 (21). Using *A. radiobacter* for referring to *Agrobacterium* biovar
105 2 is particularly confusing because *Agrobacterium* biovar 2 is also known as *Rhizobium*
106 *rhizogenes* (22) and differs from the *A. tumefaciens* complex (i.e., *Agrobacterium* biovar
107 1) in their chromosomal organization (21, 23). Finally, within Rhizobiaceae,
108 *Agrobacterium* was designated as either clade B (24) or C (25), and these designations
109 were also used for *Neorhizobium* (25) or *Rhizobium* (24), respectively. All these issues
110 have caused further confusion and remain to be resolved.
111 As genome sequences became available from more *A. tumefaciens*
112 genomospecies, the understanding of their divergence and evolution was improved. For
113 example, with ~5,200 protein-coding genes per genome, pairwise comparisons between

114 different genomospecies typically found only ~4,000 shared genes and their core-
115 genome contains < 3,600 genes (26, 27). Strikingly, > 32,000 horizontal gene transfer
116 (HGT) events were inferred for the history of *A. tumefaciens* evolution (28). Because the
117 HGT patterns indicated co-transfers of genes that encode coherent biochemical
118 pathways, it was hypothesized that purifying selection on those acquired gene clusters
119 and overall gene content drove the ecological diversification among genomospecies
120 (28). Also, their oncogenic plasmids exhibit complex modularity and transmission
121 patterns, which further contributes to the evolution of these bacteria (29).
122 Despite these progresses, several key issues have remained unresolved. First,
123 although previous studies suggested that each of the *A. tumefaciens* genomospecies
124 should be considered as a distinct species (12, 16, 27–29), quantitative and definitive
125 guidelines for establishing the species boundaries are lacking. Second, the phylogenetic
126 relationships among these genomospecies needs to be resolved with confidence,
127 particularly for the deeper branches, to provide a concrete foundation for evolutionary
128 analysis. Third, previous studies were limited by the taxon sampling and/or the
129 assembly quality of available genome sequences, hence those better-characterized
130 genomospecies (e.g., G1 and G8) were overrepresented while others (e.g., G2, G3, G5,
131 G6, and G9) were underrepresented. Such biases may have affected the conclusions
132 obtained from previous evolutionary genomics studies (27–29).
133 To address these issues, we performed targeted genome sequencing for those
134 underrepresented genomospecies to achieve a comprehensive and balanced taxon
135 sampling of *A. tumefaciens* diversity. By limiting the analysis to high-quality assemblies,
136 this dataset enabled us to conduct detailed examinations of replicon-level synteny and

137 confident inferences of gene presence/absence. In addition to providing a global view of
138 their genomic diversity and resolving the phylogeny, we also performed focused
139 analysis of genetic elements involved in key aspects of *A. tumefaciens* fitness and
140 ecology, namely the type VI secretion system (T6SS) for interbacterial competition (27,
141 30) and the tumor-inducing plasmids (pTi) for phytopathogenicity (9–11). Taken
142 together, the investigations conducted at the levels of whole-genome, whole-replicon,
143 gene clusters, individual genes, and protein domains provide novel and detailed
144 information on the evolution and genetic diversity of these bacteria that are important in
145 plant pathology and biotechnology.
146

147 **Results**

148 Molecular Phylogeny and Genome Divergence

149 Whole-genome sequencing was conducted for 12 strains belonging to the *A.*
150 *tumefaciens* species complex to improve taxon sampling and two for the sister species
151 *Agrobacterium larrymoorei* to provide an ideal outgroup. Twenty-one additional *A.*
152 *tumefaciens* genome sequences were obtained from GenBank (31) (Table 1). In
153 consideration of a balanced sampling, each of the 10 known *A. tumefaciens*
154 genomospecies was represented by at least two but no more than five genome
155 sequences. Multiple rounds of preliminary phylogenetic analysis were performed to
156 select representative genome sequences that yielded a dataset with maximal within-
157 genomospecies diversity.

158 A total of 2,093 single-copy genes are shared by the 35 genomes analyzed,
159 corresponding to ~40% of the genes in each individual genome. Compared to previous

160 studies that conducted genome-based phylogenetic analysis for Rhizobiaceae or
161 beyond (22, 28, 29), the more focused sampling of this work increased the core gene
162 count by one-to-two orders of magnitude and produced a well-resolved maximum
163 likelihood phylogeny for *A. tumefaciens* (Figure 1A). All 10 recognized genomospecies
164 form distinct monophyletic clades with > 80% bootstrap support. Two novel
165 genomospecies were identified, each represented by a single strain (i.e., KCJK1736
166 and MAFF210266), and are sister to G2 and G3, respectively. The patterns of genome
167 sequence divergence were entirely consistent with the 95% average nucleotide identity
168 (ANI) cutoff suggested for delineating bacterial species (32). For pairwise comparisons
169 within each genomospecies, 73.5%-97.9% of the genomic segments were mapped and
170 those segments have 95.1-99.9% ANI (Figure 1B). For between-genomospecies
171 comparisons, 65.2-82.6% of the genomes were mapped with 85.0-92.9% ANI. For *A.*
172 *tumefaciens-A. larrymoorei* comparisons, only ~50% of the genomes were mapped with
173 < 82% ANI.
174 The 12 *A. tumefaciens* genomospecies could be classified into seven groups
175 based on 90% ANI, and may be further classified into three supergroups according to
176 the phylogeny (Figure 1A). Based on a time-calibrated phylogeny (29), the most recent
177 common ancestor (MRCA) of *A. tumefaciens* emerged ~48 million years ago (Mya),
178 those three supergroups emerged ~40 Mya, those seven groups emerged ~30 Mya,
179 and most of the recognized genomospecies emerged ~2-10 Mya. The rapid radiation in
180 the early history of *A. tumefaciens* had made the inference of those deeper relationships
181 difficult (22, 28, 29). Nevertheless, those early nodes were supported by ~70%
182 bootstrap in this work (Figure 1A).

183 For gene content comparisons, the principal coordinate analysis and hierarchical
184 clustering results indicated that all *A. tumefaciens* genomospecies are similar to one
185 another while distinct from *A. larrymoorei* (Supplementary Figure S1A and S1C).
186 Nevertheless, with the exception of G4 and G7, these genomospecies could be
187 distinguished based on gene content (Supplementary Figure S1B and S1D). This
188 finding conflicts with a previous work that stated it is inappropriate to use gene content
189 to distinguish these genomospecies (28). This conflict may be explained by the
190 improvements in taxon sampling and assembly quality.
191
192 The Type VI Secretion System (T6SS) Genes
193 Using the fully resolved organismal phylogeny as a framework, next we
194 examined the molecular evolution of their T6SS genes. The T6SS is a phage tail-like
195 contractile nanomachine for injecting effectors (i.e., toxins) into eukaryotic or bacterial
196 cells, and plays a major role in pathogenesis, symbiosis, or interbacterial competition
197 (33, 34). For *A. tumefaciens*, the T6SS is a key weapon for *in planta* competition
198 between different genomospecies (27) and against other bacteria (30). Thus,
199 investigation on the T6SS genes may shed light on the genetic basis for their
200 competition against rivals in the same niche.
201 In *A. tumefaciens*, the T6SS genes are located on the linear chromosome and
202 mostly form a cluster of ~20 genes that are organized into two adjacent operons (i.e.,
203 *imp* and *hcp*) with opposite orientations (27, 35) (Figure 2). Some strains harbor
204 accessary T6SS loci located elsewhere on the same chromosome (27, 35, 36)
205 (Supplementary Figure S2). Among the strains examined, a patchy distribution of the

206 T6SS genes was observed (Figure 1). The T6SS gene phylogeny is largely congruent
207 with the species tree (Figure 2). The only incongruence is that the MRCA of G1 appears
208 to have acquired this gene cluster from G8. Consistent with this inference, the
209 chromosomal location of this gene cluster is conserved except in G1 strains
210 (Supplementary Figure S2). Based on these findings, it is likely that this T6SS gene
211 cluster was present in the MRCA of G8-G6-G2-G4-G7-G9-G3 and at least two
212 independent losses have occurred (i.e., G6 and G2). It is unclear if the MRCA of *A.*
213 *tumefaciens* harbored the T6SS gene cluster or not, as presence followed by a loss in
214 the MRCA of G13-G5-G1 or absence followed by a gain in the MRCA of G8-G6-G2-G4-
215 G7-G9-G3 are equally parsimonious. Although the outgroup *A. larrymoorei* harbors the
216 T6SS genes, the T6SS genes are known to have a patchy distribution among
217 Rhizobiaceae (36, 37), indicating that these genes are not essential for these bacteria
218 and have high rates of gains and losses. Consistent with this, multiple pseudogenes
219 were found (e.g., *tssH* in S56 and *tssA*/*vgrG* in ATCC31749) (Figure 2), suggesting that
220 these gene clusters may be in the process of degradation and possibly will be lost
221 eventually. The presence or absence of the T6SS genes do not have obvious links to
222 the isolation sources or geographic origins (Table 1), so the determinants for
223 maintaining these genes are unknown.
224 Examination of the synteny revealed that the *imp* operon has near-perfect
225 conservation, whereas the *hcp* operon often has different genes downstream of *vgrG*
226 (Figure 2). Importantly, because the genes downstream of *vgrG* include those that
227 encode effector and immunity (EI) protein pairs, this genetic diversity may play a key
228 role in interbacterial competition (27). The rapid evolution of these genes is illustrated by

229 three examples. First, despite the low levels of sequence divergence among G1 strains
230 (i.e., > 98% ANI in all pairwise comparisons), different genes are found downstream of
231 *vgrG* and these variations are not consistent with either the species phylogeny or the
232 T6SS gene phylogeny. Second, CFBP4996 of G7 shares the same EI genes with 12D1
233 and 183 of G4, rather than with other G7 strains. Third, for G3 strains, *vgrG* and the
234 associated EI genes are located elsewhere on the linear chromosome, rather than a
235 part of the T6SS-*hcp* operon (Figure 2 and Supplementary Figure S2). These results
236 suggest that recombination involving gene modules has contributed to the genetic
237 diversity of these *A. tumefaciens* T6SS EI genes.

238 The modularity of *vgrG* and its downstream genes is expected based on the
239 knowledge that *vgrG* homologs encode proteins with distinct C-terminal domains, which
240 are responsible for effector-binding specificities (36, 37). For in-depth investigation, we
241 examined the VgrG protein domain architecture among these homologs. The 44 *vgrG*
242 homologs, including 17 associated with the main T6SS gene cluster and 27 accessory
243 homologs located elsewhere, could be classified into six major types and nine subtypes
244 (Figure 3). Among the eight protein domains identified, only three are present in all
245 homologs. The N-terminal domain 1 was the most conserved domain (Supplementary
246 Figure S3) and the only domain that could be found in current databases. This domain
247 corresponds to TIGR03361, which accounts for ~66-77% of the protein length and the
248 bulk of the predicted 3D structures that presumably form a trimeric complex analogous
249 to phage tail spike (38, 39) (Supplementary Figure S4). The variable C-terminal
250 domains are difficult for protein structure prediction programs due to sequence
251 divergence from the reference and thus were often omitted. Nevertheless, when the

252 structure predictions are available for these variable domains (e.g., domain 5 of E1 and
253 domain 7 of F1), some differences could be observed (Supplementary Figure S4),
254 which may provide information on the determinants of effector-binding specificities. For
255 domain 5 that was found in all homologs belonging to subtypes A1-A3 and E1 (Figure
256 3), its presence is perfectly correlated with a downstream DUF4123-domain-containing
257 gene (Figure 4). Because this DUF4123 domain acts as an adaptor/chaperone for
258 effector loading onto VgrG in *A. tumefaciens* (37) and *Vibrio cholerae* (40, 41), this
259 strong co-occurrence suggests specific interactions between VgrG domain 5 and
260 DUF4123. Similar strategies based on co-occurrence patterns may be used to infer the
261 interacting domains of VgrG with other T6SS components.

262 Intriguingly, despite the conservation of domain architecture within each subtype
263 (Figure 3), the phylogenies inferred from the three domains present in all *vgrG*
264 homologs do not share the same topology (Supplementary Figure S5). For domain 1,
265 sequences from the same subtype do not always form monophyletic groups. For
266 domains 2 and 3, the short sequence lengths limited the phylogenetic resolution;
267 nonetheless, low divergence within the same subtype and high divergence between
268 subtypes were observed. These patterns suggest that each domain-encoding region
269 may have evolved independently and can recombine between subtypes. At the level of
270 gene cluster organization, some *vgrG* homologs belonging to the same subtype have
271 distinct downstream genes (e.g., A1 and B1), regardless of whether these *vgrG*
272 homologs are associated with the main T6SS gene cluster or not (Figure 4 and
273 Supplementary Data Set S1). These findings suggest that in addition to domain

274 shuffling among *vgrG* homologs, recombination also facilitated novel *vgrG*-effector
275 pairings in the evolution of these T6SS genes.
276 When T6SS diversity was examined in a phylogenetic context, the numbers and
277 the types of *vgrG* homologs, as well as their linked EI genes, lack any strong correlation
278 with the species phylogeny (Figure 5). For genomes that harbor multiple *vgrG*
279 homologs, typically those *vgrG* homologs belong to different subtypes. Strain 1D1609 is
280 a notable exception by having the highest number of *vgrG* homologs and all four
281 belonging to subtype B1. However, our experimental characterization of this strain
282 demonstrated that while these four homologs were functionally redundant in mediating
283 T6SS secretion, only three of these homologs contributed to antibacterial activity (36).
284 Among the putative EI genes, EI15 (i.e., the peptidoglycan amidase gene *tae* and
285 its cognate immunity gene *tai*) were found in nearly all T6SS-harboring strains (Figure
286 5). The conservation may be attributed to their locations, as these are the only EI genes
287 located upstream of *vgrG*, which is a more central part of the T6SS-*hcp* operon.
288 Additionally, or alternatively, the conservation may be explained by purifying selection
289 on maintaining these genes for their functions, as Tae could provide competitive growth
290 advantage over other *A. tumefaciens* or *Escherichia coli* that lack *tai* (42). Based on
291 functional classifications of the 63 putative effector genes, peptidoglycan-targeting
292 toxins and nucleases are the two most commonly found categories with 21 each
293 (Supplementary Data Set S1). This finding is consistent with an investigation of the
294 T6SS-harboring Proteobacteria (43), which also found these two as the dominant
295 categories and suggested that they play complementary roles in T6SS-mediated
296 competition. Finally, we found 10 putative effector genes families without an identifiable

297 cognate immunity gene (Figure 5). These findings may be explained by the lack of
298 known domains for immunity genes, as well as our conservative approaches of
299 annotation. It is likely that the cognate immunity genes are those immediately
300 downstream of the putative effector genes. Alternatively, some of these orphan effector
301 genes may not encode bacterial toxins.
302
303 The Tumor-inducing Plasmids (pTi) and Associated Genes
304 The virulence plasmids pTi are an important component of *A. tumefaciens*
305 genomes. These large accessory replicons harbor the virulence (*vir*) regulon genes for
306 delivering a transfer DNA (T-DNA) into plant cells via the type IV secretion system
307 (T4SS), and are the determinant of *A. tumefaciens* phytopathogenicity (9–11). Our
308 recent work on the oncogenic plasmids in Rhizobiaceae, including pTi and root-inducing
309 plasmids (pRi), established a k-mer based classification scheme for these plasmids and
310 revealed complex patterns in their evolution and transmission (29). To gain a
311 comprehensive and detailed look into the pTi diversity and evolution within the *A.*
312 *tumefaciens* species complex, we adopted this classification scheme on the available
313 sequences for in-depth characterization. Among the 35 genome sequences analyzed in
314 this work (Table 1), 15 were found to harbor a putative pTi (Figure 1). Additionally,
315 exhaustive GenBank searches identified five other key representatives (Table 2).
316 Within the *A. tumefaciens* species complex, type I pTi is the most abundant type,
317 followed by types III and II (Figure 1). This observation is consistent with our previous
318 study that examined 57 pTi-harboring *A. tumefaciens* strains without considering the
319 balance of taxon sampling (29). The type assignments based on k-mer profiles were

320 consistent with the patterns observed in global alignment (Figure 6), molecular
321 phylogeny of shared genes (Supplementary Figure S6), and gene content comparison
322 (Supplementary Figure S7). No notable congruence was found between the
323 chromosomal and plasmid phylogenies (Figure 6), indicating frequent transfers of pTi
324 across genomospecies in the evolutionary history.

325 Based on phylogeny and gene content, types II and III pTi are more similar to
326 each other than to type I pTi. Within type I, the two subtypes (i.e., I.a and I.b) are
327 distinguishable by gene content (Supplementary Figure S7) but do not form mutually
328 exclusive clades in the core-gene phylogeny (Supplementary Figure S6). All putative
329 pTi, including the outgroup pTiS4 (i.e., type V from *Agrobacterium vitis*) and the
330 unclassified pTiCFBP4996/pTiCFBP5473, contain genes for the T4SS that mediates T-
331 DNA transfer into plant cells (i.e., *virB1-B11* and *virD4*) and the corresponding two-
332 component regulatory system (i.e., *virA* and *virG*) (Figure 7). Strain 1D1609 is a notable
333 case because its *virA* and *virJ* are located on another plasmid, rather than the pTi (26).

334 For the other *vir* regulon genes, several differentiations among pTi types were
335 observed (Figure 7). For example, while all pTi harbor a conserved *virE3a* that
336 facilitates T-DNA protection and entry into host (44), type I pTi harbor one or two
337 additional copies of *virE3* that belong to different sequence types (i.e., sharing the same
338 annotation but classified as distinct homologs due to sequence divergence). Similarly,
339 *virF*, which encodes an F-box protein that is a putative host-range determinant (45), can
340 be classified into three sequence types and each is associated with a specific pTi type.
341 Those less well-characterized *vir* genes, such as *virD3* (46, 47), *virJ* (48), and *virP*, are
342 also distributed differently among types I-III pTi. Finally, in addition to the

343 presence/absence of individual genes, the overall organization of *vir* regulons also differ
344 among these pTi (Supplementary Figure S8). All type I pTi are conserved in sharing a
345 ~40 kb region that contains all *vir* genes. In comparison, locations of *vir* genes are more
346 variable among type II/III pTi; *virF/P* (and *virQ/H* if present) are located ~5-50 kb away
347 from the main *vir* gene cluster. Other than *vir* genes, the gene content and organization
348 of T-DNA are also different (Figure 8). All type I pTi have one single T-DNA region,
349 while types II and III have two and type V have four, respectively (29) (Figure 6). Within
350 the T-DNA regions, the plant hormone synthesis genes (i.e., *tms1*/*iaaM*, *tms2/iaaH*, and
351 *ipt*) are the most conserved ones, while others are more variable (Figure 8). Taken
352 together, these genetic variations may contribute to the host range differences observed
353 among strains harboring different types of pTi (49).

354 Two novel putative pTi (i.e., pTiCFBP4996 and pTiCFBP5473) were found from
355 the 14 newly sequenced strains. This efficiency of discovering novel pTi is surprising,
356 given that our previous study of defining those six pTi types was comprehensive in
357 sampling historical collections of 162 *Agrobacterium* strains that spanned the past 90
358 years and > 50 host species from six continents (29), and demonstrated the usefulness
359 of a phylogeny-guided approach for investigating genetic diversity. These two novel pTi
360 are distinctive in their large sizes (Table 2), gene organization (Figure 6), phylogenetic
361 placements (Supplementary Figure S6), and gene content (Supplementary Figure S7).
362 Moreover, their T-DNA regions also differ from the typical sizes of ~18-26 kb observed
363 in types I-III pTi (Figure 6). For pTiCFBP4996, its 7-kb T-DNA contains only four genes
364 (i.e., two for opine synthesis and two encode hypothetical proteins). Consistent with the
365 lack of identifiable plant hormone synthesis genes, CFBP4996 did not induce tumor

366 formation on tomato (Supplementary Figure S9). It is unclear if this plasmid represents
367 a primitive pTi that has not acquired the plant hormone synthesis genes yet, or it
368 represents a degraded pTi that has lost these genes. For the tumorigenic strain
369 CFBP5473, its T-DNA is ~93 kb in size and contains all of the *vir* regulon genes. The
370 extremely large T-DNA may have resulted from a translocation of the T-DNA border
371 sequence. Alternatively, there may be some non-canonical T-DNA border sequences
372 and the T-DNA of pTiCFBP5473 is smaller than predicted.
373

374 **Discussion**

375 Biological Entities at the Species Level and Above

376 The new genome sequences reported in this work greatly improved the taxon
377 sampling for a group of bacteria with importance in plant pathology and biotechnology.
378 Based on the divergence of core gene sequences (Figure 1) and gene content
379 (Supplementary Figure S2), 95% ANI is a reliable approach for defining species within
380 the *A. tumefaciens* complex, as is the case for most other bacteria (32). The discrete
381 multimodal distribution of genome similarities (Supplementary Figure S1) suggested
382 that there are genetic barriers between different genomospecies, which may be
383 explained by neutral processes and/or selection (28, 32). Regardless of the exact
384 mechanisms, these patterns supported application of the biological species concept,
385 which is based on genetic barriers, to these *A. tumefaciens* genomospecies and other
386 bacteria (5). With the continuing drop in sequencing cost, it is foreseeable that the ANI
387 analysis may become a practical or even a standard approach for accurate
388 classification of additional strains, which in turn could facilitate research and

389 communication, and ideally leads to improvements in bacterial taxonomy for basic
390 works and applications.

391 At the above-species level, *A. tumefaciens* genomospecies exhibited some
392 intriguing patterns of genome divergence. In a previous study that compared ~90,000
393 prokaryotic genomes, it was extremely rare to find ANI values in the range of 82-96%
394 (32). In other words, strains either belong to the same natural biological entity at the
395 species level and have > 95% ANI, or belong to different species and have <82% ANI.
396 This observation could be due to biases in the sampling of available genomes, or the
397 all-against-all pairwise comparisons included mostly distantly-related species. In our set
398 of strains that provided a more detailed look at a group of closely-related species, the
399 ~85-93% ANI observed between *A. tumefaciens* genomospecies (Supplementary
400 Figure S1) indicated that the *A. tumefaciens* complex is indeed a coherent entity with
401 high divergence from its closest sister lineage within the same genus (i.e., *A.*
402 *larrymoorei*). The driving forces for maintaining species complexes and the prevalence
403 of such above-species level entities are interesting questions that require further
404 investigations. For *A. tumefaciens*, although the nomenclature originated from its
405 phytopathogenicity, it is well-established that this group contains both pathogenic and
406 non-pathogenic strains that differ in the possession of an oncogenic plasmid (i.e., pTi)
407 or not (8). The promiscuous nature of their pTi (29, 50–52) suggested that lineages
408 within this complex may experience frequent transitions between pathogenic and non-
409 pathogenic lifestyles, and such shared ecological niches may be the force that
410 maintains the coherence of this species complex. Compared to the sister lineages (e.g.,

411 *A. larrymoorei* and *A. rubi*), the more diverse host range of *A. tumefaciens* (8, 9) may
412 have facilitated the divergence of this complex into multiple genomospecies.
413 At the genus level and above, genome-based classification is more challenging.
414 Although it may be desirable to establish a standardized taxonomy with a defined range
415 of genomic divergence for each taxonomic work, large variations in the divergence
416 values at a given rank were observed among different taxonomic groups in previous
417 attempts (6, 7, 53). These variations created situations where some families contain
418 higher divergence levels than some orders or lower divergence levels than some
419 genera, even after normalization and a full revision of the current taxonomy (6). Such
420 situations demonstrated the challenges of establishing a new standardized taxonomy
421 even when the practical issues of transitioning from the current taxonomy are not
422 considered, and perhaps is to be expected given the highly variable evolutionary rates
423 across different lineages (54). Given these considerations, other aspects of biology
424 (e.g., physiology, ecology, etc) may play more important roles in defining those higher
425 taxonomic ranks. Thus, the current system used by the International Committee on
426 Systematics of Prokaryotes that assigns one subcommittee to each major taxonomic
427 group is helpful in incorporating the organism-specific knowledge.
428
429 Units and Modularity of Molecular Evolution
430 For evolutionary studies, the levels at which selection and other processes
431 operate on have been a topic that received much attention (55). For prokaryotes, levels
432 from the entire genome to individual functional domains within genes are of particular

433 interest. Based on our results, all of these levels must be considered to comprehend the
434 complex patterns.

435 At the whole-genome level, the clear species boundaries based on overall
436 similarity (Figure 1 and Supplementary Figure S1) suggested that the entire genome
437 largely evolves as a single coherent unit. This result is consistent with previous findings
438 that at the global level HGT has very little impact on the reconstruction of organismal
439 phylogeny (56, 57), despite rampant HGT in bacterial evolution (57–60) and the
440 importance of HGT in adaptation (61–63). A possible explanation for these seemingly
441 conflicting observations is that most of the acquired genes are lost quickly (62),
442 presumably due to the strong mutational bias towards deletions observed in bacteria
443 (64, 65). Additionally, acquired genes are subjected to the selection that drives species
444 diversification (28), which should act on all genes in a genome together.

445 At the level of individual replicons, chromosomes and plasmids certainly have
446 distinct evolutionary histories (Figure 6). Because novel chromosome/plasmid
447 combinations may lead to speciation (51), and the spread of plasmids have important
448 implications on the evolution of virulence (29, 66) and antimicrobial resistance (67),
449 further investigations on the evolution of plasmids and their compatibilities with
450 chromosomes are important (29, 68, 69).

451 At the levels of gene clusters and below, several interesting observations were
452 made based on the two secretion systems investigated in this work. First, although
453 these two systems may provide some fitness advantages (e.g., T6SS for interbacterial
454 competitions and T4SS for host exploitation), complex patterns of gains and losses
455 were observed (Figures 2 and 6). These patterns suggested that there may not be a

456 strong selective pressure to maintain these genes, and non-adaptive stochastic
457 processes may be important. Furthermore, there is a certain degree of modularity
458 regarding their evolution, such that the presence patterns are all-or-none and no partial
459 cluster was found for the chromosomal T6SS genes or the plasmid-encoded T4SS
460 genes. This is particularly evident for the T6SS genes, as when the main cluster was
461 lost (e.g., G2 and G6), no accessory *vgrG* loci located elsewhere was found
462 (Supplementary Figure S3). Second, for both systems, genes for the structural
463 components are conserved, while those for the effectors and others are not (Figures 2,
464 5, and 7). This is consistent with the expectation that opposite selective forces may act
465 on these two categories of genes (i.e., purifying selection against changes to preserve
466 the apparatus of a functional secretion system versus positive selection for more
467 diverse effectors and other accessory components). Third, modularity that may reflect
468 functional constraints could be observed at finer scales. For example, each *vgrG* is
469 linked to its cognate effector/chaperone genes, and each effector is linked to its cognate
470 immunity gene. Such modularity is expected to be maintained by selection, similar to
471 the observations regarding co-transfers of genes involved in associated biochemical
472 pathways (28). However, the linkage could be broken down by recombination at within-
473 or between-species levels, as evident in the diversity of *vgrG* gene neighborhoods, even
474 for those homologs belonging to the same type (Figure 4). Fourth, at the level of
475 individual genes, the within-genome diversity of *vgrG* homologs (Figure 5) provides
476 further support to the hypothesis that HGT is more important than duplications in driving
477 gene family expansions in bacteria (70). Finally, at the intra-gene level, within- or
478 between-species recombination may be important in generating novel combinations of

479 domains, thus promoting the diversification of homologs (Figure 3 and Supplementary
480 Figure S6).
481 Taken together, these observations illustrated the complexity of biological
482 systems. While it is difficult to draw up generalized rules or to estimate the relative
483 importance of each evolutionary process at different levels, it is important to consider
484 and examine these complexities in order to better understand the organisms of interest.
485
486 Conclusions
487 In summary, using a group of important bacteria as the study system, this work
488 utilized a strategy of phylogeny-conscious genome sampling for systematic
489 investigations of a species complex. This approach is not only cost-effective but also
490 critical in obtaining an unbiased picture that does not over-emphasize certain
491 subgroups. Our examination of biological boundaries at the species level and above
492 improves the understanding of how natural biodiversity is organized, and provides a
493 strong foundation for future taxonomic revisions that could facilitate communication. The
494 targeted analysis of those secretion system genes and oncogenic plasmids provides
495 novel insights regarding the key genetic variations involved in the fitness and ecology of
496 these soil-borne phytopathogens that need to compete in complex microbiota and
497 invade plant hosts. Moreover, the multi-level analysis of their genetic diversity from
498 whole-genome to intra-genic domains highlights the complexity of these biological
499 systems. The strategy and findings of this work provide useful guides for future studies
500 of other bacteria.
501

502 **Materials and Methods**

503 Genome Sequencing

504 A total of 14 *Agrobacterium* strains were acquired from the French International
505 Center for Microbial Resources (CIRM) Collection for Plant-associated Bacteria (CFBP)
506 (Table 1). These include 12 strains that belong to the *Agrobacterium tumefaciens*
507 species complex, and two *Agrobacterium larrymoorei* strains (i.e., CFBP5473 and
508 CFBP5477) as the outgroup.

509 The procedure for whole-genome shotgun sequencing was based on that
510 described in our previous studies (27, 71, 72). All bioinformatics tools were used with
511 the default settings unless stated otherwise. Briefly, total genomic DNA was prepared
512 using the Wizard Genomic DNA purification kit (Promega, USA). The Illumina paired-
513 end sequencing libraries were prepared using KAPA LTP Library Preparation Kits
514 (Roche Sequencing, USA) with a targeted insert size of ~550 bp. The Illumina MiSeq
515 platform was used to generate 300x2 reads with an average coverage of 306-fold per
516 strain (range: 141- to 443-fold). The raw reads were quality trimmed using a Q20 cutoff
517 and used for *de novo* assembly based on Velvet v1.2.10 (73) with the settings "-
518 exp_cov auto -min_contig_lgth 2000 -scaffolding no". The contigs were oriented by
519 mapping to those complete genomes available (Table 1) using MAUVE v2015-02-13
520 (74). Due to the difficulties of identifying appropriate reference genomes for several
521 evolutionary branches, four strains (i.e., CFBP5473, CFBP5875, CFBP5877, and
522 CFBP6623) were selected for PacBio long-read sequencing and PacBio HGAP v3
523 assembly. These PacBio-based assemblies were used as a guide for scaffolding, rather
524 than the finalized results.

525 To improve the Illumina-based draft assemblies, an iterative process was used to
526 examine the raw reads mapping results and to incorporate gap-filling results based on
527 PCR and Sanger sequencing. This process was repeated until the complete assembly
528 was obtained or the draft assembly could not be improved further. The finalized
529 assemblies were submitted to the National Center for Biotechnology Information (NCBI)
530 and annotated using the Prokaryotic Genome Annotation Pipeline (PGAP) (75).
531
532 Comparative and Evolutionary Analysis
533 The genomes analyzed are listed in Table 1. The procedures for genome
534 comparisons were based on those described in our previous studies (27, 76–78).
535 Briefly, pairwise genome similarities were calculated using FastANI v1.1 (32). For
536 comparisons of plasmids, FastANI was executed with the custom settings that reduced
537 fragment length to 1,000 bp and minimum matched fragments to 25. For global
538 alignments of chromosomes and plasmids, the syntenic regions were identified by
539 BLASTN v2.6.0 (79) and visualized using genoPlotR v0.8.9 (80). For gene content
540 comparison, BLASTP v2.6.0 (79) with e-value cutoff set to $1e^{-15}$ and OrthoMCL v1.3
541 (81) were used to infer the homologous gene clusters. The result was converted into a
542 matrix of 35 genomes by 17,058 clusters, with the value in each cell corresponding to
543 the copy number. This matrix was further converted into a Jaccard distance matrix
544 among genomes using the VEGAN package v2.5-6 in R, then processed using the
545 principal coordinates analysis function in the APE package (82) and visualized using
546 ggplot2 v3.3.2 (83). The hierarchical clustering analysis was performed using PVCLUST
547 v3.4.4 (84).

548 For phylogenetic analysis, homologous sequences were aligned using MUSCLE
549 v3.8.31 (85) for maximum likelihood inference by PhyML v.3.3.20180214 (86). The
550 proportion of invariable sites and the gamma distribution parameter were estimated
551 from the data set, the number of substitute rate categories was set to four. The
552 bootstrap supports were estimated based on 1,000 replicates.
553
554 Analysis of the Type VI Secretion System Genes
555 To identify the T6SS-associated genes, C58 (30, 37, 87) and other strains (27,
556 36) that have been characterized experimentally were used as the references. Based
557 on the known T6SS genes in these genomes, homologous genes in other genomes
558 were identified based on the OrthoMCL result. To screen for novel T6SS effector,
559 chaperone, and immunity genes, genes that are located near *vgrG* (i.e., three upstream
560 and ten downstream) were examined manually by using the NCBI conserved domain
561 database (CDD) (88) and the Phyre2 protein fold recognition server (89); the e-value
562 cutoff was set to 0.01. A few genes with a hit to the DUF4123 domain (i.e., a known
563 T6SS chaperone) but have an e-value above the cutoff were manually added back to
564 the list of T6SS-associated genes (e.g., CFBP5477_RS20350). To confirm the absence
565 of specific T6SS genes, the genome sequences were used as the subject and the
566 protein sequences of known genes were used as the query to run TBLASTN searches.
567 For classification of the *vgrG* homologs, we developed a domain-based scheme.
568 The conserved N-terminal TIGR03361 domain was first identified by the NCBI CDD. A
569 global alignment of all homologs was used to determine the exact boundaries of this
570 domain. After this TIGR03361 domain was removed, the remaining C-terminal

571 sequences were processed using MEME v5.1.1 (90) to identify conserved domains that
572 meet these criteria: (1) present in at least two sequences, (2) with zero or one
573 occurrence per sequence, (3) with a size between 30 and 300 a.a., and (4) with an e-
574 value lower than 0.05. The results were manually curated to break down large domains
575 that are composed of smaller domains. Pairwise BLASTP searches were conducted to
576 verify that each domain is unique and no two domains have a BLASTP e-value of lower
577 than 1e-05. For each domain, the consensus sequence was generated using Jalview
578 v2.10.5 (91) and sequence conservation was visualized using WebLogo server v3 (92).
579 For functional prediction, the consensus sequence of each domain was used to query
580 against NCBI CDD and Phyre2. Additionally, one representative from each subtype of
581 *vgrG* homologs was used for structure modeling using Phyre2 with the "normal" mode.
582 The chain D of PA0091 VgrG1 (PDB identifier: 4MTK) was selected as the template.
583 The predicted structures were visualized using PyMOL v1.2r3pre (Schrödinger, USA).
584 For the EI gene pairs identified, EI1 through EI11 were named based on the
585 nomenclature proposed previously (27) and novel pairs were named starting from EI12.
586 When only the putative effector (E) or the putative immunity (I) genes were found, those
587 genes were classified in the format of "E??" or "I??", respectively. For some of the EI
588 pairs that were identified previously based on adjacency to T6SS genes but lacked
589 high-confidence annotation (i.e., EI2, EI3, EI5, and EI8), we chose a more conservative
590 approach and annotated those genes as hypothetical proteins.
591

592 Analysis of the Tumor-inducing Plasmids and Type IV Secretion System Genes

593 The list of 20 putative pTi sequences analyzed is provided in Table 2. These
594 included all of the 15 complete pTi sequences determined in this study and five
595 representatives from GenBank. The pTi typing was performed based on k-mer profile
596 clustering with a reference set of 143 oncogenic plasmids in Rhizobiaceae (29). For T-
597 DNA identification, putative T-DNA borders were identified based on the motif
598 YGRCAGGATATATNNNNNKGTMAWN (93). Genes involved in opine metabolism (94)
599 and T4SS (95) were identified based on the annotation and homologous gene clustering
600 results produced by OrthoMCL. Additionally, putative T4SS effectors were identified
601 using the T4SEpre tool (96) in EffectiveDB (97) with the minimal score set to 0.8. All
602 protein sequences of pTi-encoded genes were used as the queries.

603

604 Tumorigenesis Assay

605 Tomato tumorigenesis assays (98) were performed to evaluate the virulence of
606 selected strains. The plants (cultivar Known-You 301) were maintained in growth
607 chambers with a 16-/8-h light/dark regime and a constant temperature of 22°C.
608 Inoculation was performed on three-week old seedlings. Bacterial strains were
609 transferred from stock to 5 mL 523 broth and cultured overnight at 28°C in a shaker
610 incubator (250 rpm), then sub-cultured for four hours prior to inoculation. Bacterial cells
611 were washed and resuspended in 0.9% NaCl solution with a concentration of OD_{600} 0.2.
612 The stem was punctured with a sterilized sewing needle and 5 μL of bacterial
613 suspension was added to the wounding site. The plants were collected three weeks

614 after inoculation and 1-cm stem segments centered at the wounding site were cut for
615 weighing.
616
617 **Data Availability**
618 The 14 new genome sequences have been deposited in NCBI under the
619 BioProjects PRJNA534385-PRJNA534397 and PRJNA534399.
620
621 **Author Contribution**
622 Conceptualization: CHK
623 Funding acquisition: JHC, EML, CHK
624 Investigation: LC, YCL, MH, MNS, AJW, CFW
625 Methodology: LC, YCL, MNS, STC, CHK
626 Project administration: CHK
627 Supervision: JHC, EML, CHK
628 Validation: LC, YCL, MH, MNS, AJW
629 Visualization: LC, YCL
630 Writing – original draft: LC, YCL, CHK
631 Writing – review & editing: LC, YCL, AJW, CFW, JHC, EML, CHK
632
633 **Funding**
634 Research in the Chang lab was supported by the National Institute of Food and
635 Agriculture, US Department of Agriculture awards 2014-51181-22384 and 2020-51181-
636 32154. Research in the Lai lab was supported by Academia Sinica and the Ministry of

637 Science and Technology of Taiwan (MOST 103-2321-B-001-009). Research in the Kuo
638 lab was supported by Academia Sinica and the Ministry of Science and Technology of
639 Taiwan (MOST 109-2628-B-001-012). The funders had no role in study design, data
640 collection and interpretation, or the decision to submit the work for publication.
641

642 **Acknowledgement**

643 We thank Ai-Ping Chen, Hsin-Ying Chiang, Shu-Jen Chou, Mei-Jane Fang, Ya-Yi
644 Huang, and Wen-Sui Lo for technical assistance. The bacterial strains were imported
645 under the permits 103-B-003 and 104-B-002 issued by the Council of Agriculture of
646 Taiwan. The Sanger sequencing service and the Illumina sequencing library preparation
647 service were provided by the Genomic Technology Core (Institute of Plant and Microbial
648 Biology, Academia Sinica). The Illumina MiSeq sequencing service was provided by the
649 Genomics Core (Institute of Molecular Biology, Academia Sinica). The PacBio
650 sequencing and data processing service was provided by Genomics BioSci & Tech. Co.
651 Ltd. (New Taipei City, Taiwan). The Institute of Plant and Microbial Biology (Academia
652 Sinica) and the Department of Botany and Plant Pathology (Oregon State University)
653 provided computing resources.
654

655 **References**

656 1. Rosselló-Mora R, Amann R. 2001. The species concept for prokaryotes. FEMS
657 Microbiol Rev 25:39–67.

658 2. Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. 2009. The bacterial species
659 challenge: making sense of genetic and ecological diversity. Science 323:741–
660 746.

661 3. Peix A, Ramírez-Bahena M-H, Velázquez E. 2018. The current status on the
662 taxonomy of *Pseudomonas* revisited: an update. Infect Genet Evol 57:106–116.

663 4. Genin S, Denny TP. 2012. Pathogenomics of the *Ralstonia solanacearum* species
664 complex. Annu Rev Phytopathol 50:67–89.

665 5. Bobay L-M, Ochman H. 2017. Biological species are universal across life's
666 domains. Genome Biol Evol 9:491–501.

667 6. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A,
668 Hugenholtz P. 2018. A standardized bacterial taxonomy based on genome
669 phylogeny substantially revises the tree of life. 10. Nat Biotechnol 36:996–1004.

670 7. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P.
671 2020. A complete domain-to-species taxonomy for Bacteria and Archaea. 9. Nat
672 Biotechnol 38:1079–1086.

673 8. Young JM. 2008. *Agrobacterium*—taxonomy of plant-pathogenic *Rhizobium*
674 species, p. 183–220. *In* Tzfira, T, Citovsky, V (eds.), *Agrobacterium*: From Biology
675 to Biotechnology. Springer New York.

676 9. Kado CI. 2014. Historical account on gaining insights on the mechanism of crown
677 gall tumorigenesis induced by *Agrobacterium tumefaciens*. Front Microbiol 5:340.

678 10. Nester EW. 2015. *Agrobacterium*: nature's genetic engineer. Front Plant Sci 5:730.

679 11. Hwang H-H, Yu M, Lai E-M. 2017. *Agrobacterium*-mediated plant transformation:
680 biology and applications. Arab Book 15:e0186.

681 12. Popoff MY, Kersters K, Kiredjian M, Miras I, Coynault C. 1984. Position
682 taxonomique de souches de *Agrobacterium* d'origine hospitalière. Ann Inst
683 Pasteur Microbiol 135:427–442.

684 13. Mougel C, Thioulouse J, Perrière G, Nesme X. 2002. A mathematical method for
685 determining genome divergence and species delineation using AFLP. Int J Syst
686 Evol Microbiol 52:573–586.

687 14. Portier P, Saux MF-L, Mougel C, Lerondelle C, Chapulliot D, Thioulouse J, Nesme
688 X. 2006. Identification of genomic species in *Agrobacterium* biovar 1 by AFLP
689 genomic markers. Appl Environ Microbiol 72:7123–7131.

690 15. Costechareyre D, Rhouma A, Lavire C, Portier P, Chapulliot D, Bertolla F,
691 Boubaker A, Dessaux Y, Nesme X. 2010. Rapid and efficient identification of
692 *Agrobacterium* species by *recA* allele analysis: *Agrobacterium recA* diversity.
693 Microb Ecol 60:862–872.

694 16. Costechareyre D, Bertolla F, Nesme X. 2009. Homologous recombination in
695 *Agrobacterium*: potential implications for the genomic species concept in bacteria.
696 Mol Biol Evol 26:167–176.

697 17. Lassalle F, Campillo T, Vial L, Baude J, Costechareyre D, Chapulliot D, Shams M,
698 Abrouk D, Lavire C, Oger-Desfeux C, Hommais F, Gueguen L, Daubin V, Muller D,
699 Nesme X. 2011. Genomic species are ecological species as revealed by
700 comparative genomics in *Agrobacterium tumefaciens*. Genome Biol Evol 3:762–
701 781.

702 18. Hellens R, Mullineaux P, Klee H. 2000. A guide to *Agrobacterium* binary Ti
703 vectors. Trends Plant Sci 5:446–451.

704 19. Lee L-Y, Gelvin SB. 2008. T-DNA binary vectors and systems. Plant Physiol
705 146:325–332.

706 20. Young JM, Pennycook SR, Watson DRW. 2006. Proposal that *Agrobacterium*
707 *radiobacter* has priority over *Agrobacterium tumefaciens*. Request for an Opinion.
708 Int J Syst Evol Microbiol 56:491–493.

709 21. Slater SC, Goldman BS, Goodner B, Setubal JC, Farrand SK, Nester EW, Burr TJ,
710 Banta L, Dickerman AW, Paulsen I, Otten L, Suen G, Welch R, Almeida NF,
711 Arnold F, Burton OT, Du Z, Ewing A, Godsy E, Heisel S, Houmiel KL, Jhaveri J, Lu
712 J, Miller NM, Norton S, Chen Q, Phoolcharoen W, Ohlin V, Ondrusek D, Pride N,
713 Stricklin SL, Sun J, Wheeler C, Wilson L, Zhu H, Wood DW. 2009. Genome
714 sequences of three *Agrobacterium* biovars help elucidate the evolution of
715 multichromosome genomes in bacteria. J Bacteriol 191:2501–2511.

716 22. Ormeño-Orrillo E, Servín-Garcidueñas LE, Rogel MA, González V, Peralta H,
717 Mora J, Martínez-Romero J, Martínez-Romero E. 2015. Taxonomy of rhizobia and
718 agrobacteria from the Rhizobiaceae family in light of genomics. Syst Appl Microbiol
719 38:287–291.

720 23. Slater S, Setubal JC, Goodner B, Houmiel K, Sun J, Kaul R, Goldman BS, Farrand
721 SK, Almeida N, Burr T, Nester E, Rhoads DM, Kadoi R, Ostheimer T, Pride N,
722 Sabo A, Henry E, Telepak E, Cromes L, Harkleroad A, Oliphant L, Pratt-Szegila P,
723 Welch R, Wood D. 2013. Reconciliation of sequence data and updated annotation

724 of the genome of *Agrobacterium tumefaciens* C58, and distribution of a linear
725 chromosome in the genus *Agrobacterium*. Appl Environ Microbiol 79:1414–1417.

726 24. Mousavi SA, Österman J, Wahlberg N, Nesme X, Lavire C, Vial L, Paulin L, de
727 Lajudie P, Lindström K. 2014. Phylogeny of the *Rhizobium–Allorhizobium–*
728 *Agrobacterium* clade supports the delineation of *Neorhizobium* gen. nov. Syst Appl
729 Microbiol 37:208–215.

730 25. Mousavi SA, Willems A, Nesme X, de Lajudie P, Lindström K. 2015. Revised
731 phylogeny of Rhizobiaceae: proposal of the delineation of *Pararhizobium* gen.
732 nov., and 13 new species combinations. Syst Appl Microbiol 38:84–90.

733 26. Haryono M, Cho S-T, Fang M-J, Chen A-P, Chou S-J, Lai E-M, Kuo C-H. 2019.
734 Differentiations in gene content and expression response to virulence induction
735 between two *Agrobacterium* strains. Front Microbiol 10:1554.

736 27. Wu C-F, Santos MNM, Cho S-T, Chang H-H, Tsai Y-M, Smith DA, Kuo C-H,
737 Chang JH, Lai E-M. 2019. Plant-pathogenic *Agrobacterium tumefaciens* strains
738 have diverse type VI effector-immunity pairs and vary in in-planta competitiveness.
739 Mol Plant Microbe Interact 32:961–971.

740 28. Lassalle F, Planel R, Penel S, Chapulliot D, Barbe V, Dubost A, Calteau A,
741 Vallenet D, Mornico D, Bigot T, Guéguen L, Vial L, Muller D, Daubin V, Nesme X.
742 2017. Ancestral genome estimation reveals the history of ecological diversification
743 in *Agrobacterium*. Genome Biol Evol 9:3413–3431.

744 29. Weisberg AJ, Davis EW, Tabima J, Belcher MS, Miller M, Kuo C-H, Loper JE,
745 Grünwald NJ, Putnam ML, Chang JH. 2020. Unexpected conservation and global
746 transmission of agrobacterial virulence plasmids. Science 368:eaba5256.

747 30. Ma L-S, Hachani A, Lin J-S, Filloux A, Lai E-M. 2014. *Agrobacterium tumefaciens*
748 deploys a superfamily of type VI secretion DNase effectors as weapons for
749 interbacterial competition in planta. Cell Host Microbe 16:94–104.

750 31. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Ostell J, Pruitt KD, Sayers
751 EW. 2018. GenBank. Nucleic Acids Res 46:D41–D47.

752 32. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High
753 throughput ANI analysis of 90K prokaryotic genomes reveals clear species
754 boundaries. Nat Commun 9:5114.

755 33. Hernandez RE, Gallegos-Monterrosa R, Coulthurst SJ. 2020. Type VI secretion
756 system effector proteins: Effective weapons for bacterial competitiveness. Cell
757 Microbiol 22:e13241.

758 34. Jurėnas D, Journet L. 2021. Activity, delivery, and diversity of Type VI secretion
759 effectors. Mol Microbiol 115:383–394.

760 35. Wu H-Y, Chung P-C, Shih H-W, Wen S-R, Lai E-M. 2008. Secretome analysis
761 uncovers an Hcp-family protein secreted via a type VI secretion system in
762 *Agrobacterium tumefaciens*. J Bacteriol 190:2841–2850.

763 36. Santos MNM, Cho S-T, Wu C-F, Chang C-J, Kuo C-H, Lai E-M. 2020.
764 Redundancy and specificity of type VI secretion vgrG loci in antibacterial activity of
765 *Agrobacterium tumefaciens* 1D1609 strain. Front Microbiol 10:3004.

766 37. Bondage DD, Lin J-S, Ma L-S, Kuo C-H, Lai E-M. 2016. VgrG C terminus confers
767 the type VI effector transport specificity and is required for binding with PAAR and
768 adaptor–effector complex. Proc Natl Acad Sci 113:E3931–E3940.

769 38. Pukatzki S, Ma AT, Revel AT, Sturtevant D, Mekalanos JJ. 2007. Type VI
770 secretion system translocates a phage tail spike-like protein into target cells where
771 it cross-links actin. Proc Natl Acad Sci 104:15508–15513.

772 39. Leiman PG, Basler M, Ramagopal UA, Bonanno JB, Sauder JM, Pukatzki S,
773 Burley SK, Almo SC, Mekalanos JJ. 2009. Type VI secretion apparatus and phage
774 tail-associated protein complexes share a common evolutionary origin. Proc Natl
775 Acad Sci 106:4154–4159.

776 40. Liang X, Moore R, Wilton M, Wong MJQ, Lam L, Dong TG. 2015. Identification of
777 divergent type VI secretion effectors using a conserved chaperone domain. Proc
778 Natl Acad Sci 112:9106–9111.

779 41. Unterweger D, Kostiuk B, Ötjengerdes R, Wilton A, Diaz-Satizabal L, Pukatzki S.
780 2015. Chimeric adaptor proteins translocate diverse type VI secretion system
781 effectors in *Vibrio cholerae*. EMBO J 34:2198–2210.

782 42. Yu M, Wang Y-C, Huang C-J, Ma L-S, Lai E-M. 2021. *Agrobacterium tumefaciens*
783 deploys a versatile antibacterial strategy to increase its competitiveness. J
784 Bacteriol 203:e00490-20.

785 43. Smith WPJ, Vettiger A, Winter J, Ryser T, Comstock LE, Basler M, Foster KR.
786 2020. The evolution of the type VI secretion system as a disintegration weapon.
787 PLOS Biol 18:e3000720.

788 44. Li X, Tu H, Pan SQ. 2018. *Agrobacterium* delivers anchorage protein VirE3 for
789 companion VirE2 to aggregate at host entry sites for T-DNA protection. Cell Rep
790 25:302-311.e6.

791 45. Jarchow E, Grimsley NH, Hohn B. 1991. virF, the host-range-determining virulence
792 gene of *Agrobacterium tumefaciens*, affects T-DNA transfer to *Zea mays*. Proc
793 Natl Acad Sci 88:10426–10430.

794 46. Vogel AM, Das A. 1992. The *Agrobacterium tumefaciens virD3* gene is not
795 essential for tumorigenicity on plants. J Bacteriol 174:5161–5164.

796 47. Lin T-S, Kado CI. 1993. The *virD4* gene is required for virulence while *virD3* and
797 orf5 are not required for virulence of *Agrobacterium tumefaciens*. Mol Microbiol
798 9:803–812.

799 48. Pan SQ, Jin S, Boulton MI, Hawes M, Gordon MP, Nester EW. 1995. An
800 *Agrobacterium* virulence factor encoded by a Ti plasmid gene or a chromosomal
801 gene is required for T-DNA transfer into plants. Mol Microbiol 17:259–269.

802 49. Hwang H-H, Wu ET, Liu S-Y, Chang S-C, Tzeng K-C, Kado CI. 2013.
803 Characterization and host range of five tumorigenic *Agrobacterium tumefaciens*
804 strains and possible application in plant transient transformation assays. Plant
805 Pathol 62:1384–1397.

806 50. Hooykaas PJJ, Klapwijk PM, Nuti MP, Schilperoort RA, Rörsch A. 1977. Transfer
807 of the *Agrobacterium tumefaciens* Ti plasmid to avirulent agrobacteria and to
808 *Rhizobium ex planta*. J Gen Microbiol 98:477–484.

809 51. Haryono M, Tsai Y-M, Lin C-T, Huang F-C, Ye Y-C, Deng W-L, Hwang H-H, Kuo
810 C-H. 2018. Presence of an *Agrobacterium*-type tumor-inducing plasmid in
811 *Neorhizobium* sp. NCHU2750 and the link to phytopathogenicity. Genome Biol
812 Evol 10:3188–3195.

813 52. Rathore DS, Mullins E. 2018. Alternative non-*Agrobacterium* based methods for
814 plant transformation, p. 891–908. *In* Roberts, JA (ed.), Annual Plant Reviews
815 online. John Wiley & Sons, Ltd., Hoboken, New Jersey, USA.

816 53. Barco RA, Garrity GM, Scott JJ, Amend JP, Nealson KH, Emerson D. 2020. A
817 genus definition for Bacteria and Archaea based on a standard genome
818 relatedness index. mBio 11:e02475-19.

819 54. Kuo C-H, Ochman H. 2009. Inferring clocks when lacking rocks: the variable rates
820 of molecular evolution in bacteria. Biol Direct 4:35.

821 55. Okasha S. 2006. Evolution and the Levels of Selection. Oxford University Press.

822 56. Daubin V, Moran NA, Ochman H. 2003. Phylogenetics and the cohesion of
823 bacterial genomes. Science 301:829–832.

824 57. Choi I-G, Kim S-H. 2007. Global extent of horizontal gene transfer. Proc Natl Acad
825 Sci 104:4489–4494.

826 58. Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the
827 nature of bacterial innovation. Nature 405:299–304.

828 59. Dagan T, Artzy-Randrup Y, Martin W. 2008. Modular networks and cumulative
829 impact of lateral transfer in prokaryote genome evolution. Proc Natl Acad Sci
830 105:10039–10044.

831 60. Chan CX, Beiko RG, Darling AE, Ragan MA. 2009. Lateral transfer of genes and
832 gene fragments in prokaryotes. Genome Biol Evol 1:429–438.

833 61. Pál C, Papp B, Lercher MJ. 2005. Adaptive evolution of bacterial metabolic
834 networks by horizontal gene transfer. Nat Genet 37:1372–1375.

835 62. Kuo C-H, Ochman H. 2009. The fate of new bacterial genes. FEMS Microbiol Rev
836 33:38–43.

837 63. Wiedenbeck J, Cohan FM. 2011. Origins of bacterial diversity through horizontal
838 genetic transfer and adaptation to new ecological niches. FEMS Microbiol Rev
839 35:957–976.

840 64. Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial
841 genomes. Trends Genet 17:589–596.

842 65. Kuo C-H, Ochman H. 2009. Deletional bias across the three domains of life.
843 Genome Biol Evol 1:145–152.

844 66. Sundin GW. 2007. Genomic insights into the contribution of phytopathogenic
845 bacterial plasmids to the evolutionary history of their hosts. Annu Rev Phytopathol
846 45:129–151.

847 67. Bennett PM. 2009. Plasmid encoded antibiotic resistance: acquisition and transfer
848 of antibiotic resistance genes in bacteria. Br J Pharmacol 153:S347–S357.

849 68. Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EPC, Cruz F de la. 2010.
850 Mobility of plasmids. Microbiol Mol Biol Rev 74:434–452.

851 69. Redondo-Salvo S, Fernández-López R, Ruiz R, Vielva L, de Toro M, Rocha EPC,
852 Garcillán-Barcia MP, de la Cruz F. 2020. Pathways for horizontal gene transfer in
853 bacteria revealed by a global map of their plasmids. Nat Commun 11:3602.

854 70. Treangen TJ, Rocha EPC. 2011. Horizontal transfer, not duplication, drives the
855 expansion of protein families in prokaryotes. PLOS Genet 7:e1001284.

856 71. Huang Y-Y, Cho S-T, Lo W-S, Wang Y-C, Lai E-M, Kuo C-H. 2015. Complete
857 genome sequence of *Agrobacterium tumefaciens* Ach5. Genome Announc
858 3:e00570-15.

859 72. Cho S-T, Haryono M, Chang H-H, Santos MNM, Lai E-M, Kuo C-H. 2018.
860 Complete genome sequence of *Agrobacterium tumefaciens* 1D1609. Genome
861 Announc 6:e00253-18.

862 73. Zerbino DR, Birney E. 2008. Velvet: algorithms for *de novo* short read assembly
863 using de Bruijn graphs. Genome Res 18:821–829.

864 74. Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of
865 conserved genomic sequence with rearrangements. Genome Res 14:1394–1403.

866 75. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L,
867 Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI prokaryotic genome
868 annotation pipeline. Nucleic Acids Res 44:6614–6624.

869 76. Lo W-S, Chen L-L, Chung W-C, Gasparich GE, Kuo C-H. 2013. Comparative
870 genome analysis of *Spiroplasma melliferum* IPMB4A, a honeybee-associated
871 bacterium. BMC Genomics 14:22.

872 77. Lo W-S, Gasparich GE, Kuo C-H. 2018. Convergent evolution among ruminant-
873 pathogenic *Mycoplasma* involved extensive gene content changes. Genome Biol
874 Evol 10:2130–2139.

875 78. Cho S-T, Kung H-J, Huang W, Hogenhout SA, Kuo C-H. 2020. Species
876 boundaries and molecular markers for the classification of 16SrI phytoplasmas
877 inferred by genome analysis. Front Microbiol 11:1531.

878 79. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden
879 T. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

880 80. Guy L, Roat Kultima J, Andersson SGE. 2010. genoPlotR: comparative gene and
881 genome visualization in R. Bioinformatics 26:2334–2335.

882 81. Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for
883 eukaryotic genomes. Genome Res 13:2178–2189.

884 82. Popescu A-A, Huber KT, Paradis E. 2012. ape 3.0: New tools for distance-based
885 phylogenetics and evolutionary analysis in R. Bioinformatics 28:1536–1537.

886 83. Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag
887 New York, New York, NY.

888 84. Suzuki R, Shimodaira H. 2006. Pvclust: an R package for assessing the
889 uncertainty in hierarchical clustering. Bioinformatics 22:1540–1542.

890 85. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and
891 high throughput. Nucleic Acids Res 32:1792–1797.

892 86. Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate
893 large phylogenies by maximum likelihood. Syst Biol 52:696–704.

894 87. Lin J-S, Ma L-S, Lai E-M. 2013. Systematic dissection of the *Agrobacterium* type
895 VI secretion system reveals machinery and secreted components for subcomplex
896 formation. PLOS ONE 8:e67647.

897 88. Marchler-Bauer A, Zheng C, Chitsaz F, Derbyshire MK, Geer LY, Geer RC,
898 Gonzales NR, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Lu S, Marchler GH, Song
899 JS, Thanki N, Yamashita RA, Zhang D, Bryant SH. 2013. CDD: conserved
900 domains and protein three-dimensional structure. Nucleic Acids Res 41:D348–
901 D352.

902 89. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. 2015. The Phyre2
903 web portal for protein modeling, prediction and analysis. Nat Protoc 10:845–858.

904 90. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW,
905 Noble WS. 2009. MEME Suite: tools for motif discovery and searching. Nucleic
906 Acids Res 37:W202–W208.

907 91. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview
908 version 2—a multiple sequence alignment editor and analysis workbench.
909 Bioinformatics 25:1189–1191.

910 92. Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: a sequence
911 logo generator. Genome Res 14:1188–1190.

912 93. Conner AJ, Barrell PJ, Baldwin SJ, Lokerse AS, Cooper PA, Erasmuson AK, Nap
913 J-P, Jacobs JME. 2007. Intragenic vectors for gene transfer without foreign DNA.
914 Euphytica 154:341–353.

915 94. Vladimirov IA, Matveeva TV, Lutova LA. 2015. Opine biosynthesis and catabolism
916 genes of *Agrobacterium tumefaciens* and *Agrobacterium rhizogenes*. Russ J
917 Genet 51:121–129.

918 95. Bhatty M, Laverde Gomez JA, Christie PJ. 2013. The expanding bacterial type IV
919 secretion lexicon. Res Microbiol 164:620–639.

920 96. Wang Y, Wei X, Bao H, Liu S-L. 2014. Prediction of bacterial type IV secreted
921 effectors by C-terminal features. BMC Genomics 15:50.

922 97. Eichinger V, Nussbaumer T, Platzer A, Jehl M-A, Arnold R, Rattei T. 2016.
923 EffectiveDB—updates and novel features for a better annotation of bacterial
924 secreted proteins and Type III, IV, VI secretion systems. Nucleic Acids Res
925 44:D669–D674.

926 98. Wu H-Y, Chen C-Y, Lai E-M. 2014. Expression and functional characterization of
927 the *Agrobacterium* VirB2 amino acid substitution variants in T-pilus biogenesis,
928 virulence, and transient transformation efficiency. PLOS ONE 9:e101142.

929

Table 1. List of the genome sequences used in this study. These include 14 new genomes derived from this study and 21 additional representatives from GenBank. Two *Agrobacterium larrymoorei* strains are included as the outgroup. Species name abbreviations: *At*, *Agrobacterium tumefaciens*; *Al*, *Agrobacterium larrymoorei.*

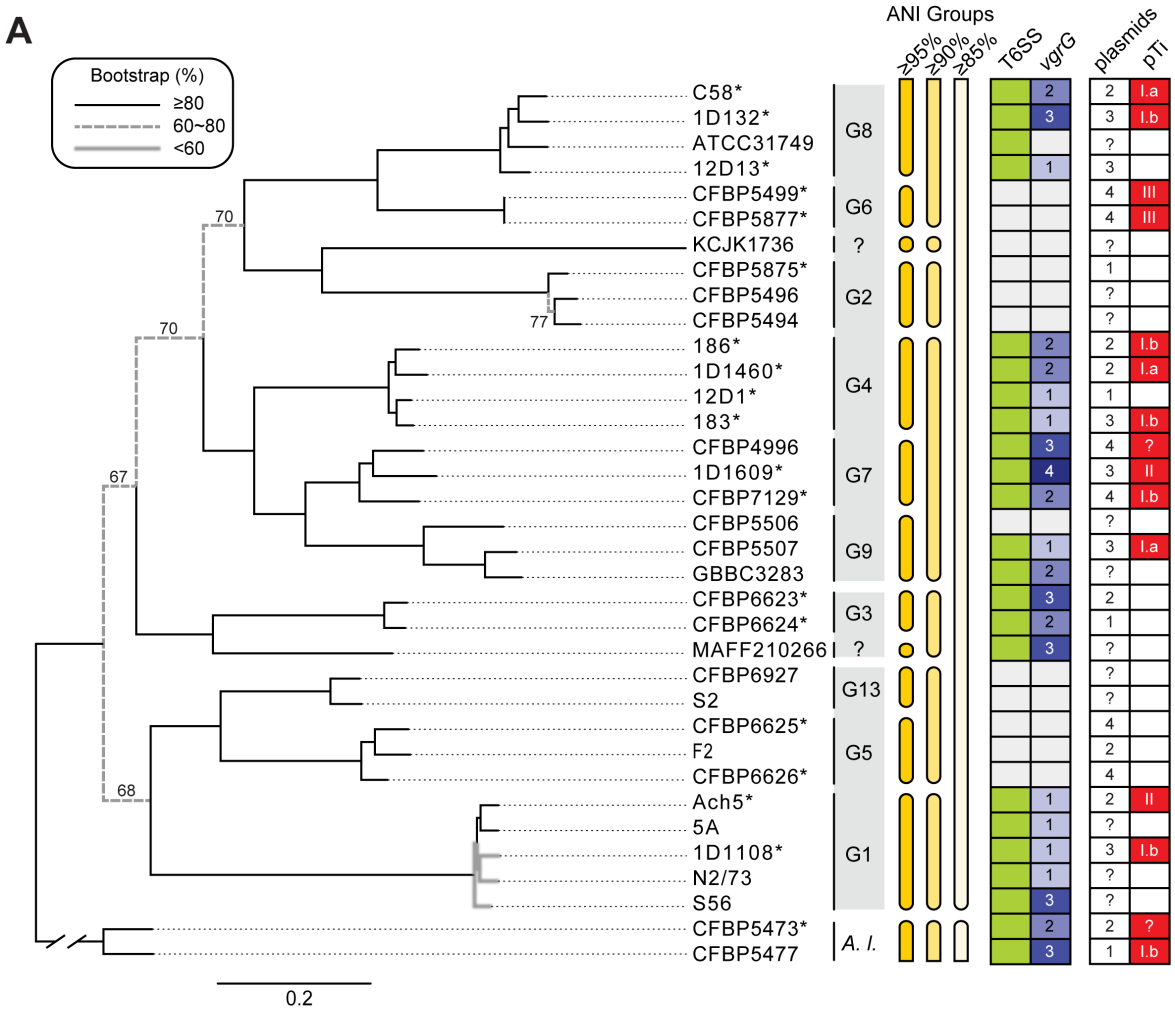
Species	Strain	Accession	Assembly	Coding Sequences	Pseudo-genes	Geographic Origin	Isolation Source
At G1	1D1108	GCF_003666425	Complete	5,312	167	MD, USA	*Euonymus* sp.
At G1	5A	GCF_000236125	50 contigs	5,343	169	MT, USA	Soil
At G1	Ach5	GCF_000971565	Complete	5,184	145	CA, USA	*Achillea ptarmica*
At G1	N2/73	GCF_001692195	29 scaffolds	5,345	157	OR, USA	Cranberry
At G1	S56	GCF_900014385	6 scaffolds	5,414	218	?	Plant
At G2	CFBP5494	GCF_900013495	5 scaffolds	5,469	233	France	Human
At G2	CFBP5496	GCF_005144405	9 contigs	5,137	223	France	Human
At G2	CFBP5875	GCF_005221365	Complete	4,570	146	Belgium	Ditch water
At G3	CFBP6623	GCF_005221385	Complete	5,081	180	France	Antiseptic flask
At G3	CFBP6624	GCF_005221425	Complete	5,157	148	France	Human
At G4	183	GCF_004023565	Complete	5,051	199	Tunisia	*Prunus dulcis*
At G4	186	GCF_002591665	Complete	5,298	207	CA, USA	*Juglans regia*
At G4	12D1	GCF_003667905	Complete	5,005	167	?	?
At G4	1D1460	GCF_003666445	Complete	5,290	269	CA, USA	*Rubus* sp.
At G5	CFBP6625	GCF_005221465	Complete	5,446	198	France	Food
At G5	CFBP6626	GCF_005221445	Complete	5,080	199	France	Human
At G5	F2	GCF_000219665	8 contigs	5,085	135	Harbin, China	Soil
At G6	CFBP5499	GCF_005221325	Complete	5,654	240	South Africa	*Dahlia* sp.
At G6	CFBP5877	GCF_005221345	Complete	5,463	236	Israel	*Dahlia* sp.

At G7	1D1609	GCF_002943835	Complete	5,539	254	CA, USA	*Medicago sativa*
At G7	CFBP4996	GCF_005144435	10 contigs	5,741	213	UK	*Flacourtia ramontchi*
At G7	CFBP7129	GCF_005221405	Complete	5,927	285	Tunisia	*Pyrus communis*
At G8	12D13	GCF_003667945	Complete	5,083	178	?	?
At G8	1D132	GCF_003667725	Complete	5,176	169	CA, USA	*Cerasus pseudocerasus*
At G8	ATCC31749	GCF_002916755	4 contigs	4,560	811	China	Plant
At G8	C58	GCF_000092025	Complete	5,355	29	NY, USA	*Cerasus pseudocerasus*
At G9	CFBP5506	GCF_005144495	15 contigs	4,236	157	Australia	Soil
At G9	CFBP5507	GCF_005144505	16 contigs	5,299	253	Australia	Soil
At G9	GBBC3283	GCF_007002765	48 contigs	4,946	203	Belgium	*Solanum lycopersicum*
At G13	CFBP6927	GCF_900012615	9 scaffolds	4,722	94	France	Rhizospheric soil from *Prunus persicae*
At G13	S2	GCF_000723345	60 contigs	5,574	167	?	?
At G?	KCJK1736	GCF_001641425	41 contigs	4,933	162	FL, USA	*Bos taurus* feces
At G?	MAFF210266	GCF_007002865	26 contigs	5,199	162	Japan	*Cucumis melo*
Al	CFBP5473	GCF_005145045	Complete	4,777	142	FL, USA	*Ficus benjamina*
Al	CFBP5477	GCF_005144425	24 contigs	4,657	95	Italy	?

Table 2. List of the pTi sequences analyzed. These include 15 from the genome data set listed in **Table 1** and five additional representatives downloaded from GenBank. The pTi type assignments are based on k-mer profiles. A type V pTi from *Agrobacterium vitis* is included as the outgroup. Species name abbreviations: *At*, *Agrobacterium tumefaciens*; *Al*, *Agrobacterium larrymoorei*; *Av*, *Agrobacterium vitis.*

Species	Strain	pTi	pTi Type	Opine Type	Accession	Size (bp)	Coding Sequences	Geographic Origin	Isolation Source
At G1	1D1108	pTi1D1108	I.b	?	NZ_CP032925	176,213	159	MD, USA	*Euonymus* sp.
At G1	Ach5	pTiAch5	II	Octopine	NZ_CP011249	194,264	153	CA, USA	*Achillea ptarmica*
At G1	183	pTi183	I.b	?	NZ_CP029048	192,674	173	Tunisia	*Prunus dulcis*
At G4	186	pTi186	I.b	?	NZ_CP042277	177,577	159	CA, USA	*Juglans regia*
At G4	1D1460	pTi1D1460	I.a	?	NZ_CP032929	214,233	198	CA, USA	*Rubus* sp.
At G4	MAFF301001	pTiSAKURA	I.b	Nopaline	NC_002147	206,479	195	Japan	*Cerasus pseudocerasus*
At G6	CFBP5499	pTiCFBP5499	III	?	NZ_CP039893	220,025	178	South Africa	*Dahlia* sp.
At G6	CFBP5877	pTiCFBP5877	III	?	NZ_CP039902	220,025	181	Israel	*Dahlia* sp.
At G7	1D1609	pTi1D1609	II	Octopine	NZ_CP026926	166,117	138	CA, USA	*Medicago sativa*
At G7	CFBP4996	pTiCFBP4996	?	?	NZ_CM016551	605,495	521	UK	*Flacourtia ramontchi*
At G7	CFBP7129	pTiCFBP7129	I.b	?	NZ_CP039927	189,955	176	Tunisia	*Pyrus communis*
At G8	1D132	pTi1D132	I.b	?	NZ_CP033026	177,577	160	CA, USA	*Cerasus pseudocerasus*
At G8	C58	pTiC58	I.a	Nopaline	NC_003065	214,233	197	NY, USA	*Cerasus pseudocerasus*
At G9	CFBP5507	pTiCFBP5507	I.a	?	NZ_CM016546	248,634	225	Australia	Soil
?	Bo542	pTiBo542	III	Agropine	NC_010929	244,978	223	Germany	*Dahlia* sp.

?	Chry5	pTiChry5	III	Chrysopine	KX388536	197,268	210	FL, USA	*Chrysanthemum* sp.
?	EU6	pTiEU6	I.b	Succinamopine	KX388535	176,375	194	CT, USA	*Euonymus* sp.
Al	CFBP5473	pTiCFBP5473	?	?	NZ_CP039694	404,101	360	FL, USA	*Ficus benjamina*
Al	CFBP5477	pTiCFBP5477	I.b	?	NZ_CM016547	190,036	169	Italy	?
Av	S4	pTiS4	V	Vitopine	NC_011982	258,824	193	Hungary	*Vitis* sp.


A
Bootstrap (%)
≥80
60~80
<60
ANI Groups
≥95%
≥90%
≥85%
T6SS
vgrG
plasmids
pTi
C58*
1D132*
ATCC31749
12D13*
CFBP5499*
CFBP5877*
KCJK1736
CFBP5875*
CFBP5496
CFBP5494
186*
1D1460*
12D1*
183*
CFBP4996
1D1609*
CFBP7129*
CFBP5506
CFBP5507
GBBC3283
CFBP6623*
CFBP6624*
MAFF210266
CFBP6927
S2
CFBP6625*
F2
CFBP6626*
Ach5*
5A
1D1108*
N2/73
S56
CFBP5473*
CFBP5477
G8
G6
?
G2
G4
G7
G9
G3
?
G13
G5
G1
A. l.
70
70
77
67
68
0.2

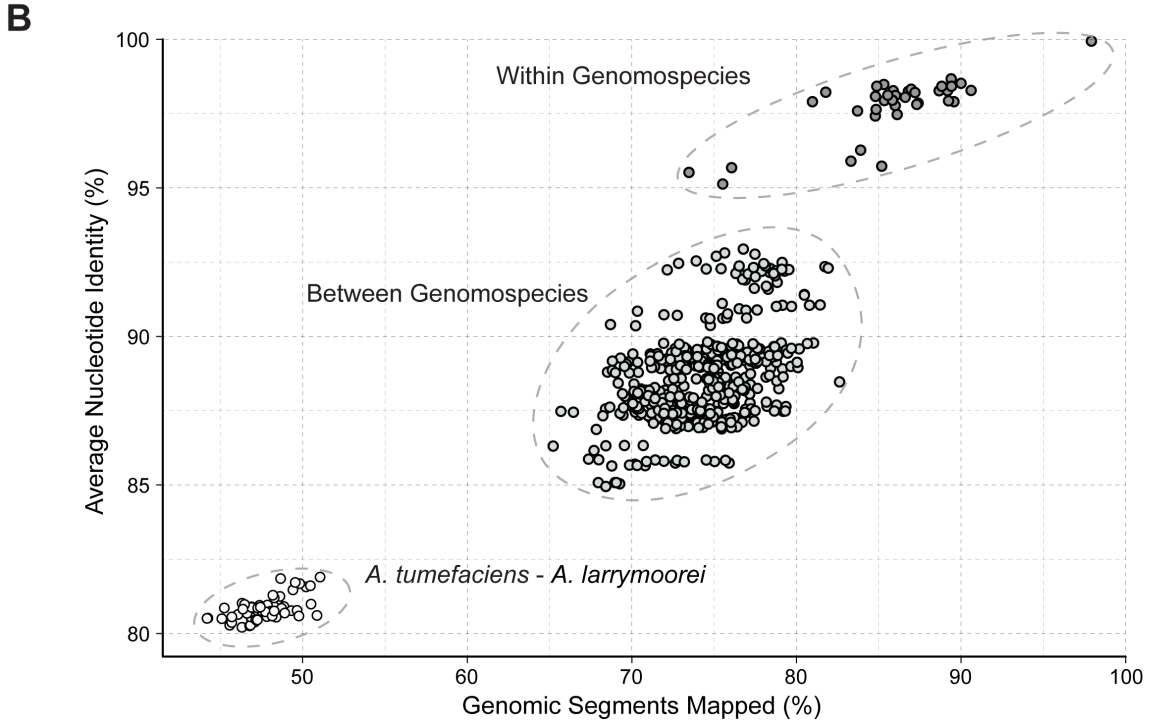

B
Within Genomospecies
Between Genomospecies
A. tumefaciens - A. larrymoorei
Average Nucleotide Identity (%)
Genomic Segments Mapped (%)
100
95
90
85
80
50
60
70
80
90
100

Figure 1. Relationships among representatives of the *Agrobacterium tumefaciens* species complex. The sister species *Agrobacterium larrymoorei* (*A. l.*) is included as the outgroup. (A) Maximum likelihood phylogeny based on a concatenated alignment of 2,093 single-copy genes shared by all 35 genomes (635,594 aligned amino acid sites). Bootstrap support values in the range of 60-80% are labeled. Strains with complete genome assemblies are highlighted with an asterisk ('*'). The genomospecies assignments (i.e., G1-G9 and G13) are labeled to the right of strain names. The three *A. tumefaciens* supergroups are indicated by the colored background of the genomospecies assignments. Information to the right of the genomospecies assignments shows the grouping of genomes according to different cutoff values of genome-wide average nucleotide identity (ANI), the presence/absence of type VI secretion system (T6SS)-encoding gene cluster (green: present; white: absent), copy number of *vgrG* (white background: absent), number of plasmids, and the type of tumor-inducing plasmid (pTi) (?: novel type; white background: absent). (B) Pairwise genome similarities based on the percentages of genomic segments mapped and the ANI values.

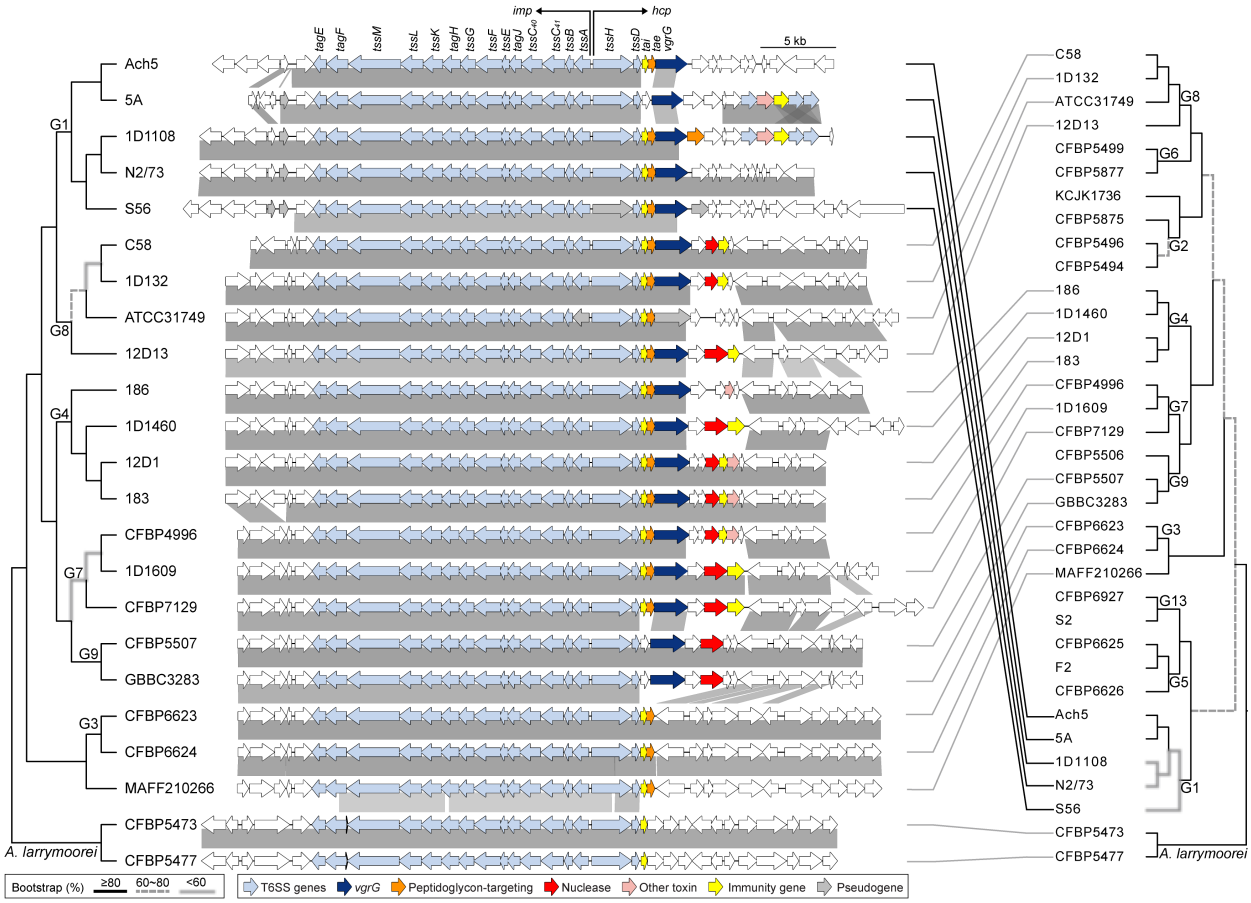


Figure 2. Phylogeny and organization of the T6SS gene cluster. The maximum likelihood phylogeny was inferred based on a concatenated alignment (5,960 aligned amino acid sites) of 14 core T6SS genes, including *tagE*, *tagF*, *tssM*, *tssL*, *tssK*, *tagH*, *tssG*, *tssF*, *tssE*, *tagJ*, *tssC40*, *tssC41*, *tssB*, and *tssD*. Two other core genes, *tssA* and *tssH*, are excluded because some homologs were pseudogenized. Genes downstream of *tssD* (e.g., *tai*, *tae*, and *vgrG*) are excluded due to variable presence. The species phylogeny on the right is based on **Figure 1**. Genes are color-coded according to annotation, syntenic regions are indicated by gray blocks.

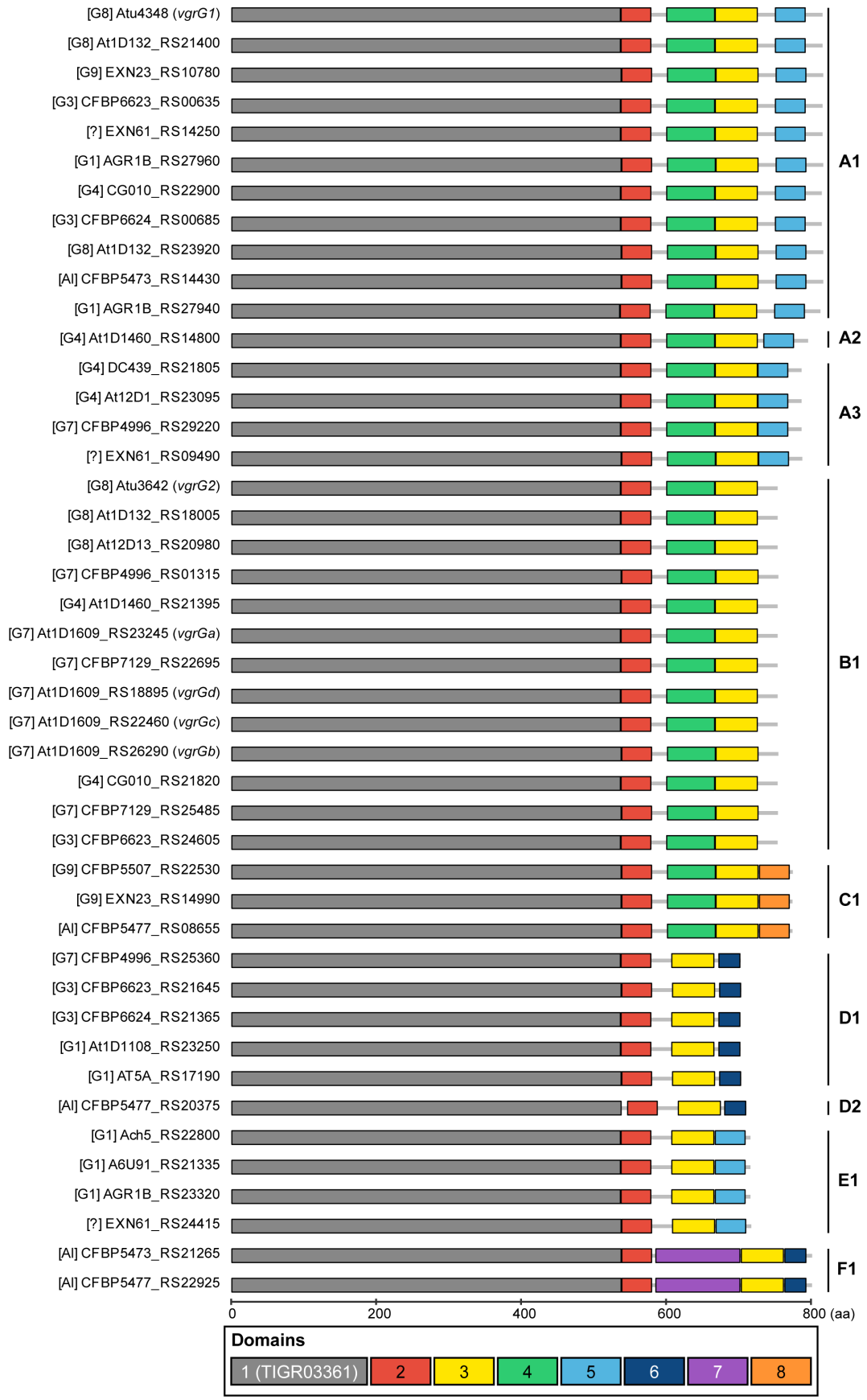

[G8] Atu4348 (vgrG1)
[G8] At1D132_RS21400
[G9] EXN23_RS10780
[G3] CFBP6623_RS00635
[?] EXN61_RS14250
[G1] AGR1B_RS27960
[G4] CG010_RS22900
[G3] CFBP6624_RS00685
[G8] At1D132_RS23920
[AI] CFBP5473_RS14430
[G1] AGR1B_RS27940
[G4] At1D1460_RS14800
[G4] DC439_RS21805
[G4] At12D1_RS23095
[G7] CFBP4996_RS29220
[?] EXN61_RS09490
[G8] Atu3642 (vgrG2)
[G8] At1D132_RS18005
[G8] At12D13_RS20980
[G7] CFBP4996_RS01315
[G4] At1D1460_RS21395
[G7] At1D1609_RS23245 (vgrGa)
[G7] CFBP7129_RS22695
[G7] At1D1609_RS18895 (vgrGd)
[G7] At1D1609_RS22460 (vgrGc)
[G7] At1D1609_RS26290 (vgrGb)
[G4] CG010_RS21820
[G7] CFBP7129_RS25485
[G3] CFBP6623_RS24605
[G9] CFBP5507_RS22530
[G9] EXN23_RS14990
[AI] CFBP5477_RS08655
[G7] CFBP4996_RS25360
[G3] CFBP6623_RS21645
[G3] CFBP6624_RS21365
[G1] At1D1108_RS23250
[G1] AT5A_RS17190
[AI] CFBP5477_RS20375
[G1] Ach5_RS22800
[G1] A6U91_RS21335
[G1] AGR1B_RS23320
[?] EXN61_RS24415
[AI] CFBP5473_RS21265
[AI] CFBP5477_RS22925
A1
A2
A3
B1
C1
D1
D2
E1
F1
0
200
400
600
800 (aa)
Domains
1 (TIGR03361)
2
3
4
5
6
7
8

Figure 3. Domain organization and classification of *vgrG* homologs. Six major types (i.e., A-F) and nine subtypes were identified and labeled on the right. For each homolog, the genomospecies assignment is provided in square brackets, followed by the locus tag. The gene names are provided in parenthesis for those functionally characterized homologs (i.e., *vgrG1-2* for C58 homologs and *vgrGa-d* for 1D1609 homologs).

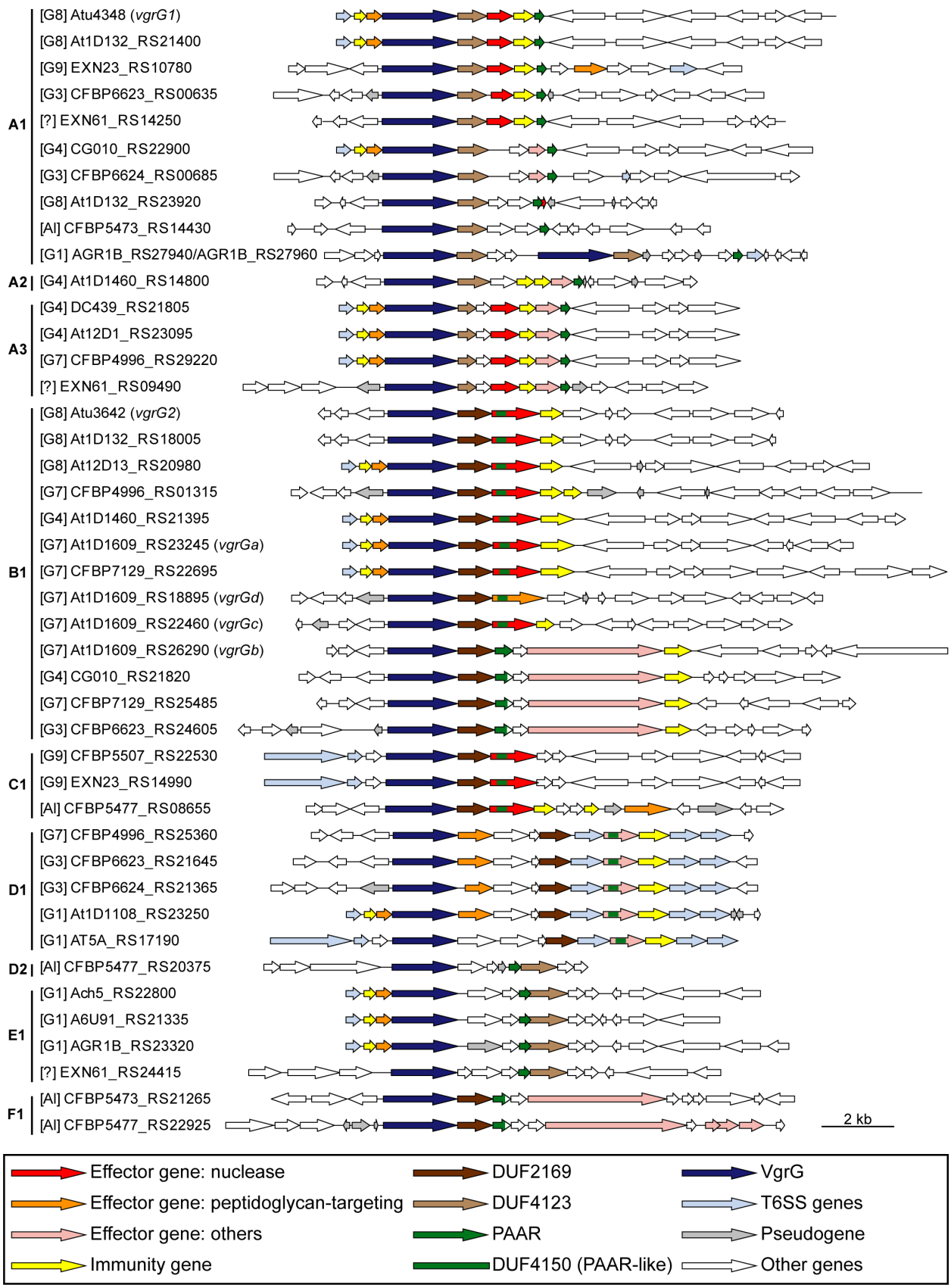

A1
[G8] Atu4348 (vgrG1)
[G8] At1D132_RS21400
[G9] EXN23_RS10780
[G3] CFBP6623_RS00635
[?] EXN61_RS14250
[G4] CG010_RS22900
[G3] CFBP6624_RS00685
[G8] At1D132_RS23920
[Al] CFBP5473_RS14430
[G1] AGR1B_RS27940/AGR1B_RS27960
A2
[G4] At1D1460_RS14800
A3
[G4] DC439_RS21805
[G4] At12D1_RS23095
[G7] CFBP4996_RS29220
[?] EXN61_RS09490
B1
[G8] Atu3642 (vgrG2)
[G8] At1D132_RS18005
[G8] At12D13_RS20980
[G7] CFBP4996_RS01315
[G4] At1D1460_RS21395
[G7] At1D1609_RS23245 (vgrGa)
[G7] CFBP7129_RS22695
[G7] At1D1609_RS18895 (vgrGd)
[G7] At1D1609_RS22460 (vgrGc)
[G7] At1D1609_RS26290 (vgrGb)
[G4] CG010_RS21820
[G7] CFBP7129_RS25485
[G3] CFBP6623_RS24605
C1
[G9] CFBP5507_RS22530
[G9] EXN23_RS14990
[Al] CFBP5477_RS08655
D1
[G7] CFBP4996_RS25360
[G3] CFBP6623_RS21645
[G3] CFBP6624_RS21365
[G1] At1D1108_RS23250
[G1] AT5A_RS17190
D2
[Al] CFBP5477_RS20375
E1
[G1] Ach5_RS22800
[G1] A6U91_RS21335
[G1] AGR1B_RS23320
[?] EXN61_RS24415
F1
[Al] CFBP5473_RS21265
[Al] CFBP5477_RS22925
2 kb
Effector gene: nuclease
Effector gene: peptidoglycan-targeting
Effector gene: others
Immunity gene
DUF2169
DUF4123
PAAR
DUF4150 (PAAR-like)
VgrG
T6SS genes
Pseudogene
Other genes

Figure 4. Gene neighborhoods of *vgrG* homologs. The grouping and labeling of *vgrG* homologs are based on the convention used in **Figure 3**. For each *vgrG* homolog, three upstream genes are plotted to illustrate if it is associated with the main T6SS gene cluster or not, and 10 downstream genes are plotted to illustrate putative effector/immunity genes. Two A1-type homologs from the strain S56 (i.e., AGR1B_RS27940 and AGR1B_RS27960) are in close association with each other and plotted together.

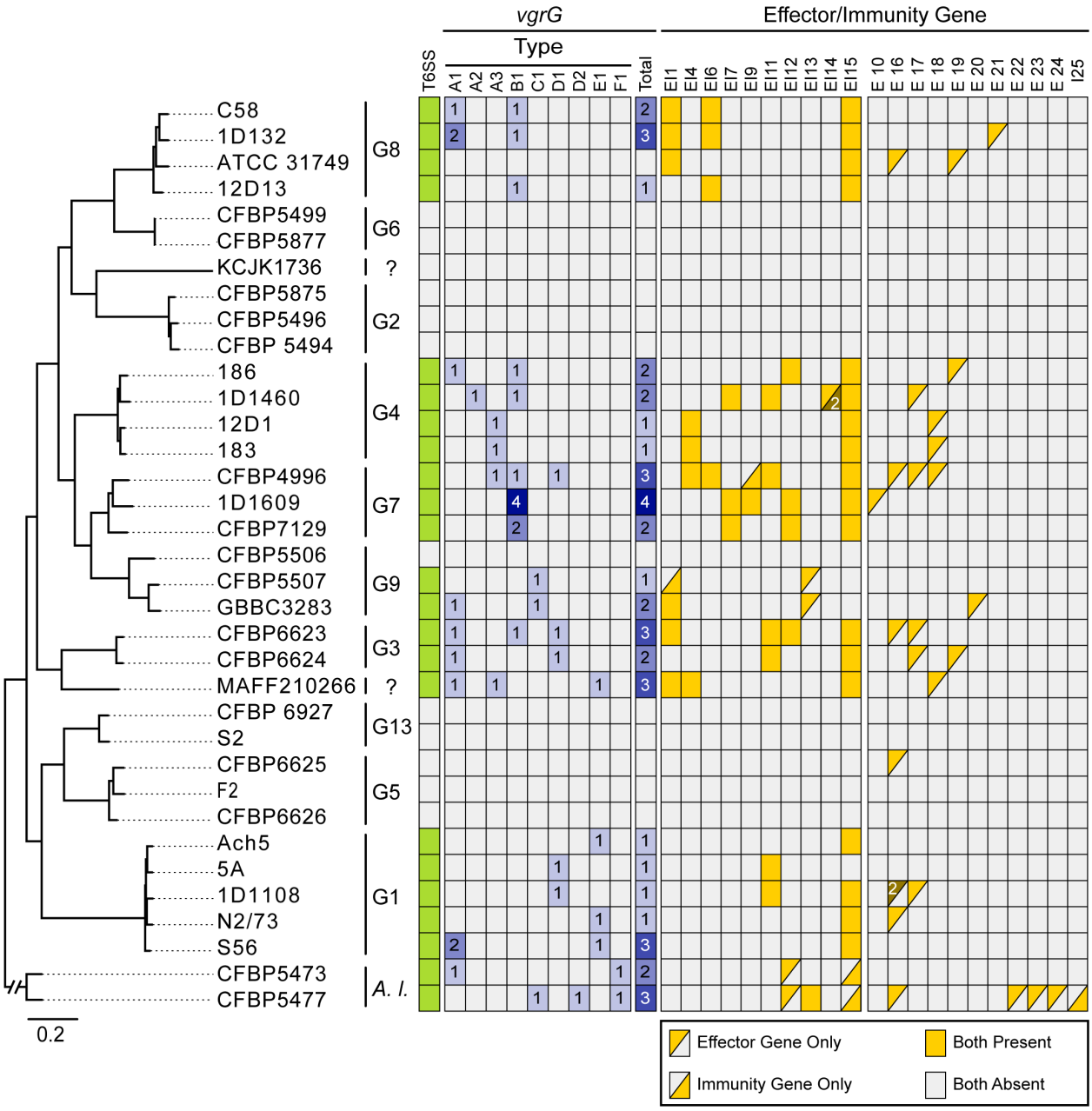


Figure 5. Phylogenetic distribution of *vgrG* homologs and T6SS effector/immunity genes. The species tree is based on **Figure 1**. Gene presence is illustrated with colored cells in the heatmap, gene copy numbers are labeled when applicable.

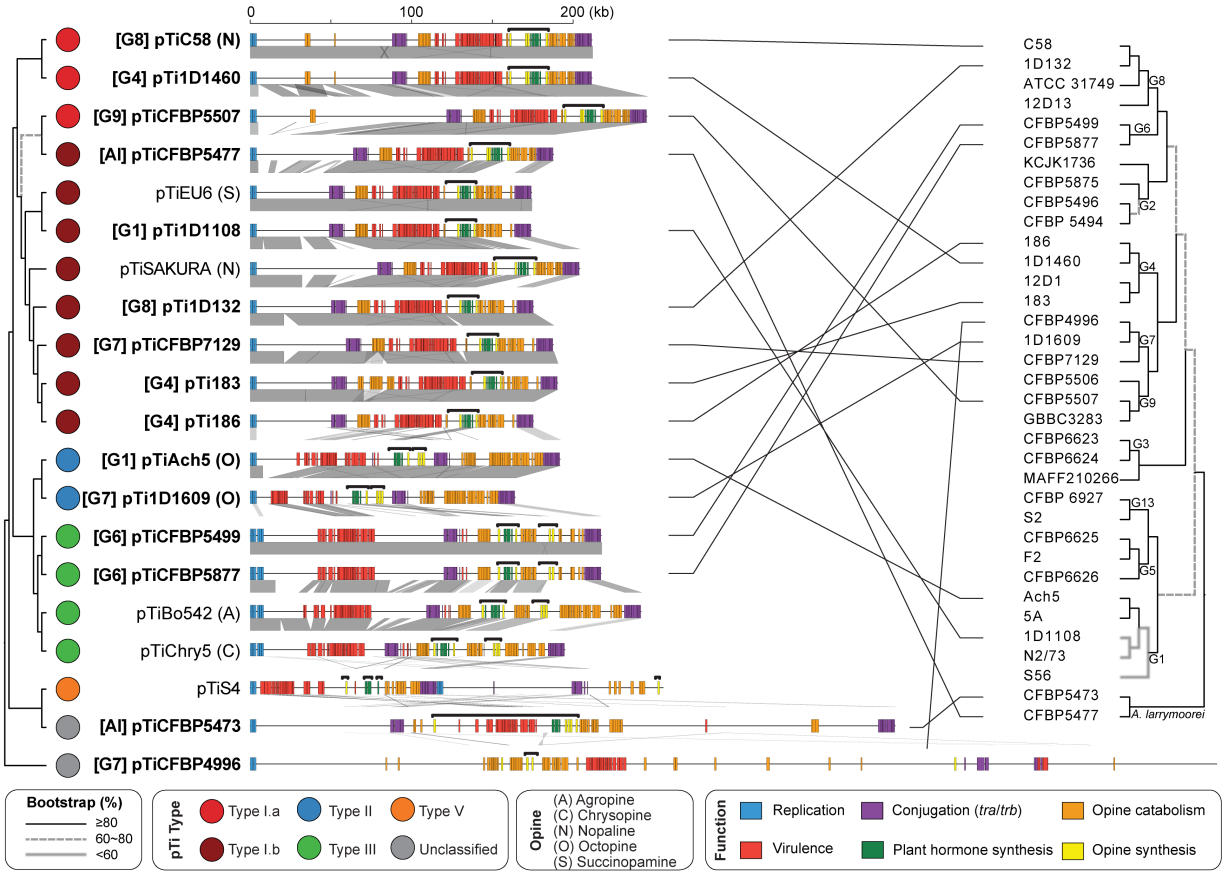


Figure 6. Molecular phylogeny and global alignment of pTi. The maximum likelihood phylogeny was inferred based on the concatenated alignment of 21 core genes and 8,534 aligned amino acid sites (**Supplementary Figure S6A**). The species phylogeny on the right is based on **Figure 1**. The pTi sequences derived from this study are highlighted in bold. For those with relevant information available, the genomospecies assignments are indicated in square brackets and the opine types are indicated in parentheses. For the alignment, all plasmids are visualized in linear form starting from the replication genes. Other gene clusters are color-coded according to functions. Syntenic regions are indicated by gray blocks.

Figure 7. Distribution of key *vir* genes among the putative pTi. The pTi sequences derived from this study are highlighted in bold. For those with relevant information available, the genomospecies assignments are indicated in square brackets and the opine types are indicated in parentheses. Gene presence and absence are indicated by filled and empty circles, respectively. The main *vir* genes include the components of the type IV secretion system (T4SS; *virB1-B11* and *virD4*). For *virE3*, the three sequence types are listed separately. For *virF* and *virD3*, the sequence types are labeled inside the circles. For 1D1609, the genes *virA* and *virJ* are located on another plasmid and plotted as absent in this figure. The locus tags are provided in **Supplementary Data Set S2**.

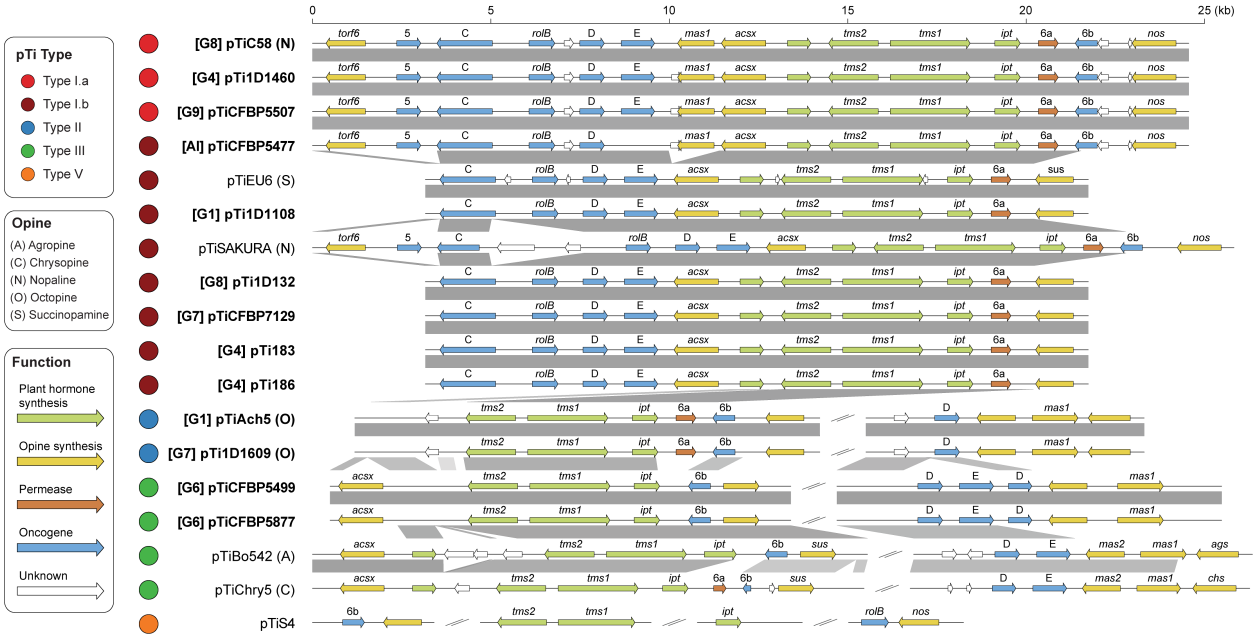
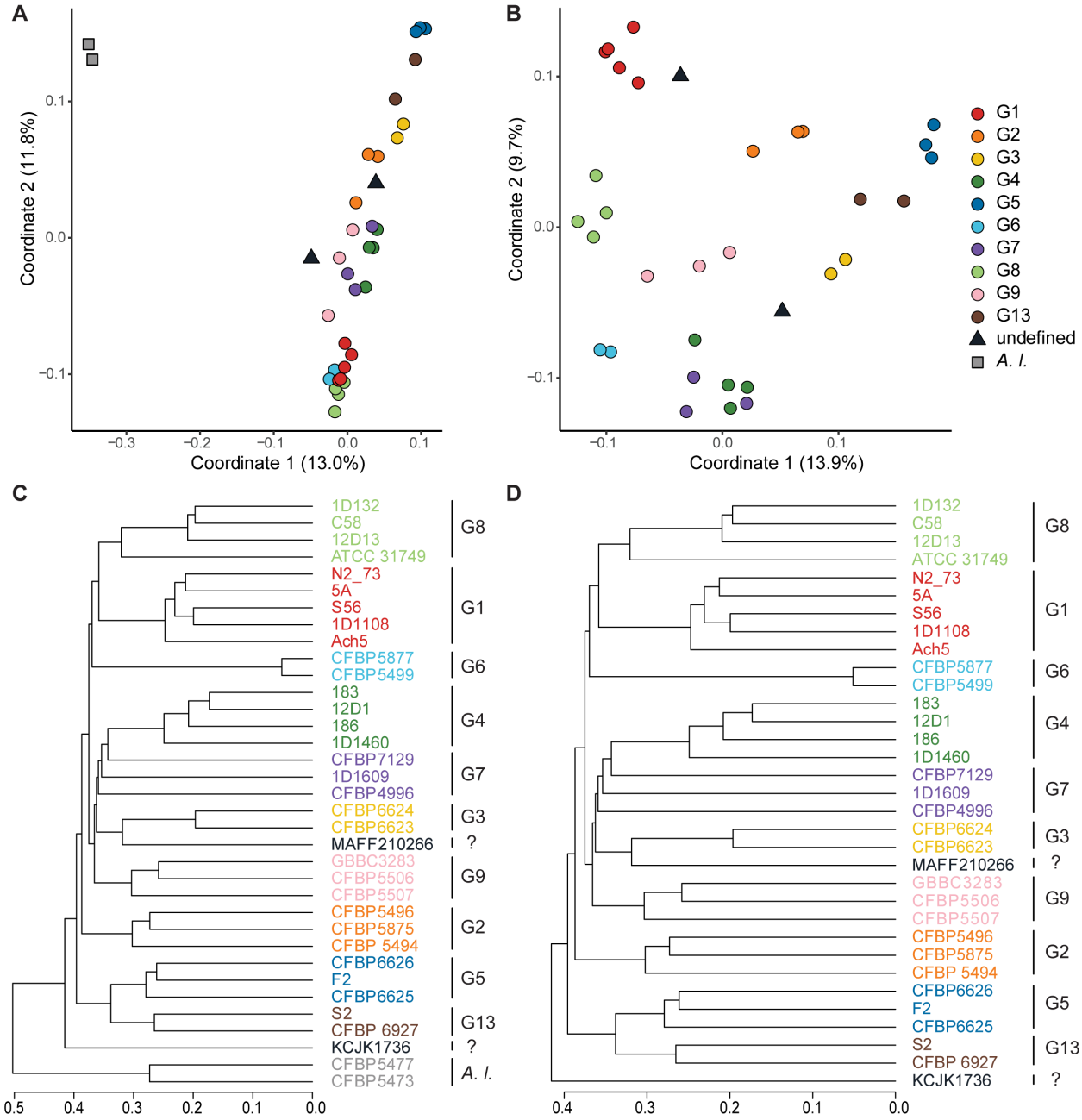


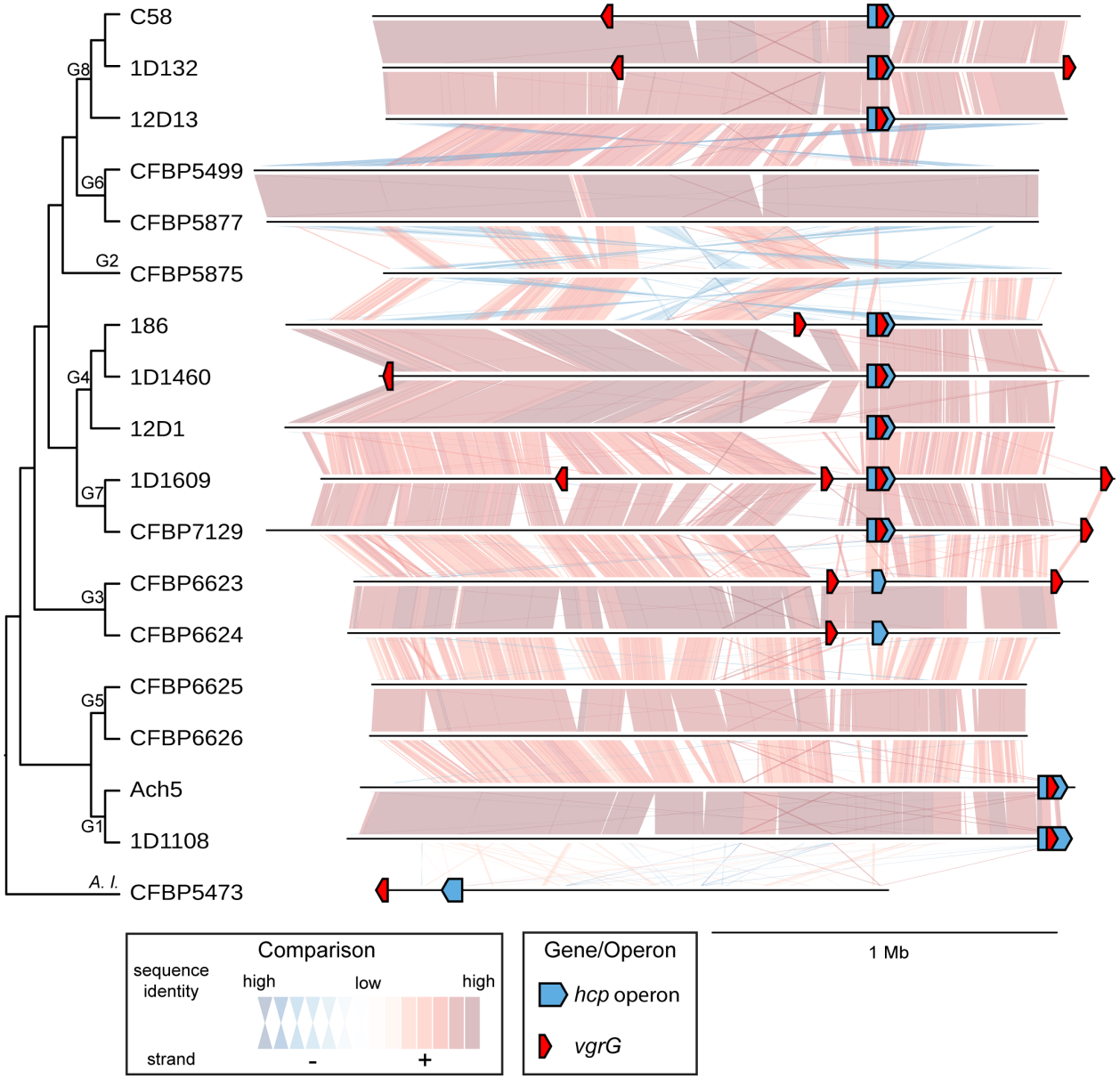
Figure 8. Organization of the transfer DNA (T-DNA) on tumor-inducing plasmids (pTi). Two unusual putative pTi sequences (i.e., pTiCFBP4996 and pTiCFBP4573) are excluded, other pTi sequences derived from this study are highlighted in bold. For those with relevant information available, the genomospecies assignments are indicated in square brackets and the opine types are indicated in parentheses. Genes are color-coded according to annotation, syntenic regions are indicated by gray blocks.

Supplementary Materials



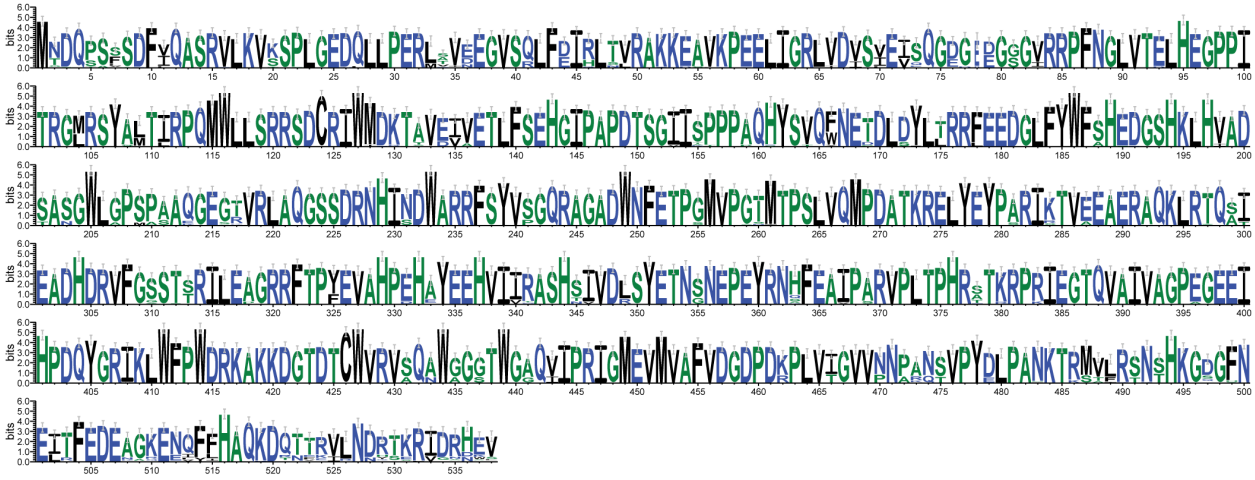
Supplementary Figure S1. Gene content dissimilarity among the *Agrobacterium* genomes. (A) and (B): principal coordinate analysis with and without the outgroup *A. larrymoorei*, respectively. The % variance explained by each axis is provided in

parentheses. (C) and (D): hierarchical clustering with and without the outgroup *A. larrymoorei*, respectively.



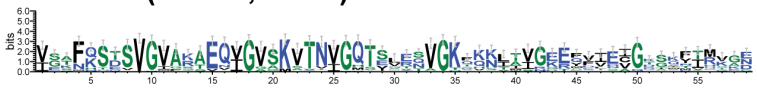
Supplementary Figure S2. Global alignment of the linear chromosomes. Locations of T6SS-*hcp* operons and *vgrG* homologs are labeled.

Domain 1 (538 a.a; 44/44; TIGR03361)

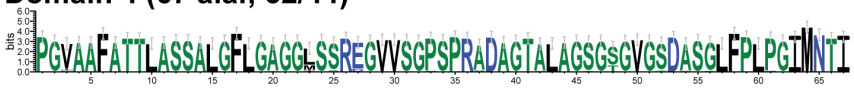


Domain 2 (42 a.a.; 44/44)

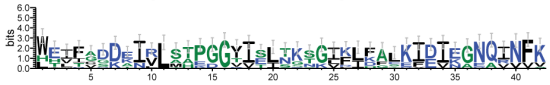
Domain 3 (59 a.a.; 44/44)



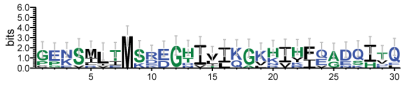
Domain 4 (67 a.a.; 32/44)



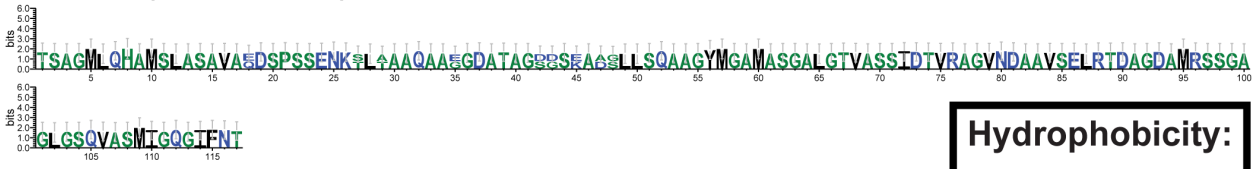
Domain 5 (42 a.a.; 20/44)



Domain 6 (30 a.a.; 8/44)



Domain 7 (117 a.a.; 3/44)



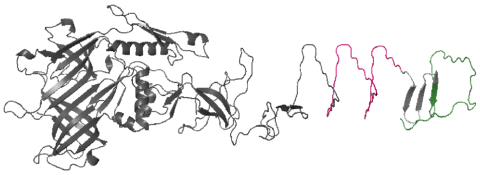
Domain 8 (42 a.a.; 2/44)

Hydrophobicity:
- Hydrophilic
- Neutral
- Hydrophobic

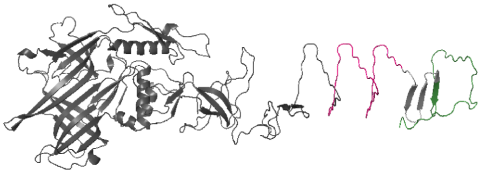
Supplementary Figure S3. Logo plots of the putative protein domains identified among *vgrG* homologs. For each domain, the length and the number of homologs with the

domain is labeled. Domain 1 is the only domain with corresponding database entry (TIGR03361).

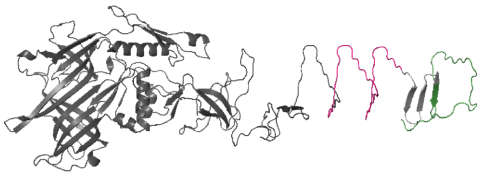
Type A1
(Atu4348)
Coverage = 75%
Identity = 30%



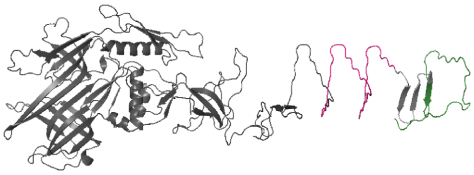
Type A2
(At1D1460_RS14800)
Coverage = 75%
Identity = 30%



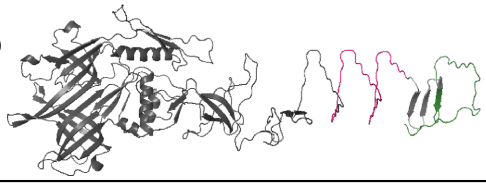
Type A3
(At12D1_RS23095)
Coverage = 77%
Identity = 30%



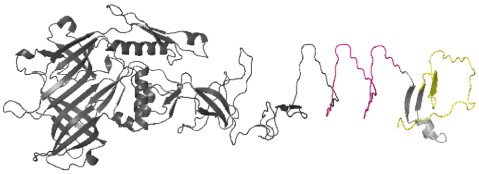
Type B1
(Atu3642)
Coverage = 81%
Identity = 30%



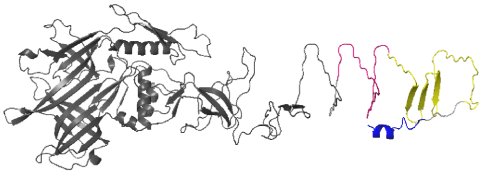
Type C1
(CFBP5507_RS22530)
Coverage = 78%
Identity = 30%



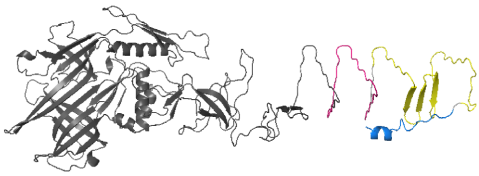
Type D1
(At1D1108_RS23250)
Coverage = 86%
Identity = 30%



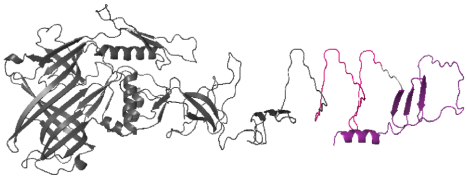
Type D2
(CFBP5477_RS20375)
Coverage = 87%
Identity = 30%



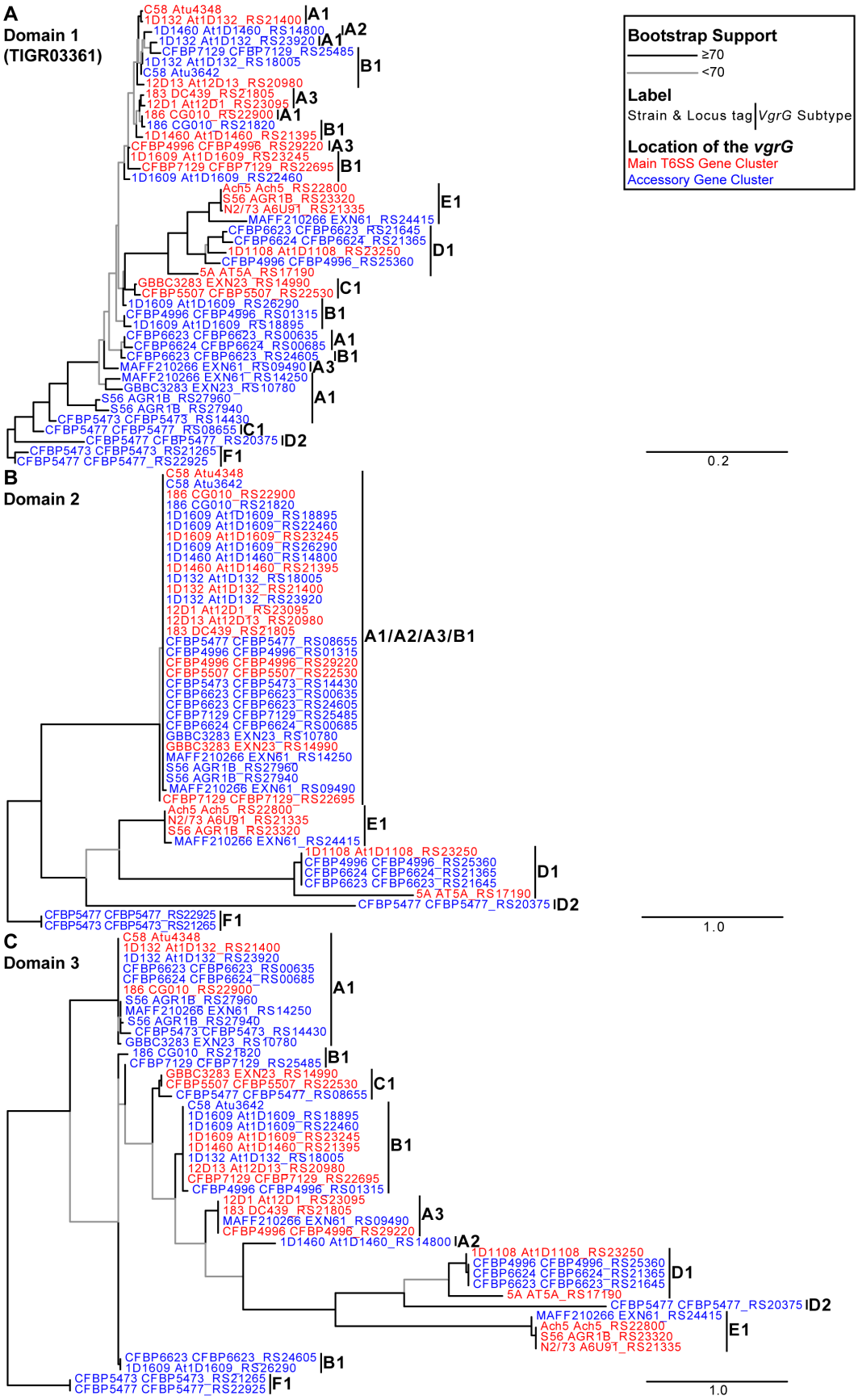
Type E1
(Ach5_RS22800)
Coverage = 86%
Identity = 33%



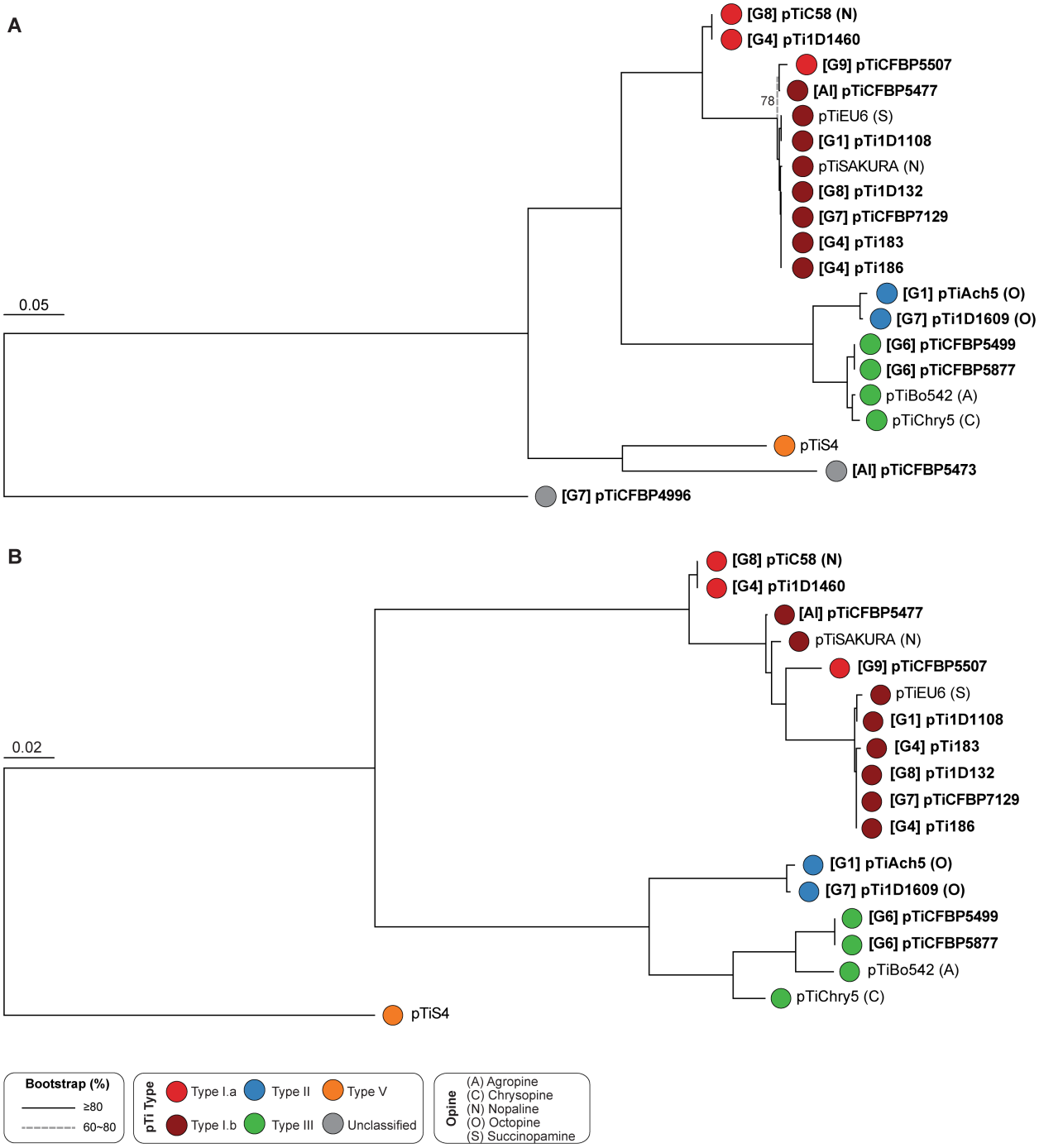
Type F1
(CFBP5477_RS22925)
Coverage = 78%
Identity = 30%



Supplementary Figure S4. Predicted structures of VgrG homologs. Regions are colored according to the scheme used in the domain analysis (**Figure. 3**). The chain D of PA0091 VgrG1 (PDB identifier: 4MTK) from *Pseudomonas aeruginosa* was selected as the template. The C-terminal parts that could not be confidently inferred are omitted. In all cases, the coverage (i.e., percentage of the sequence included in the structure prediction) are at least 75%, the sequence identity to the template is at least 30% and the confidence score is 100%.

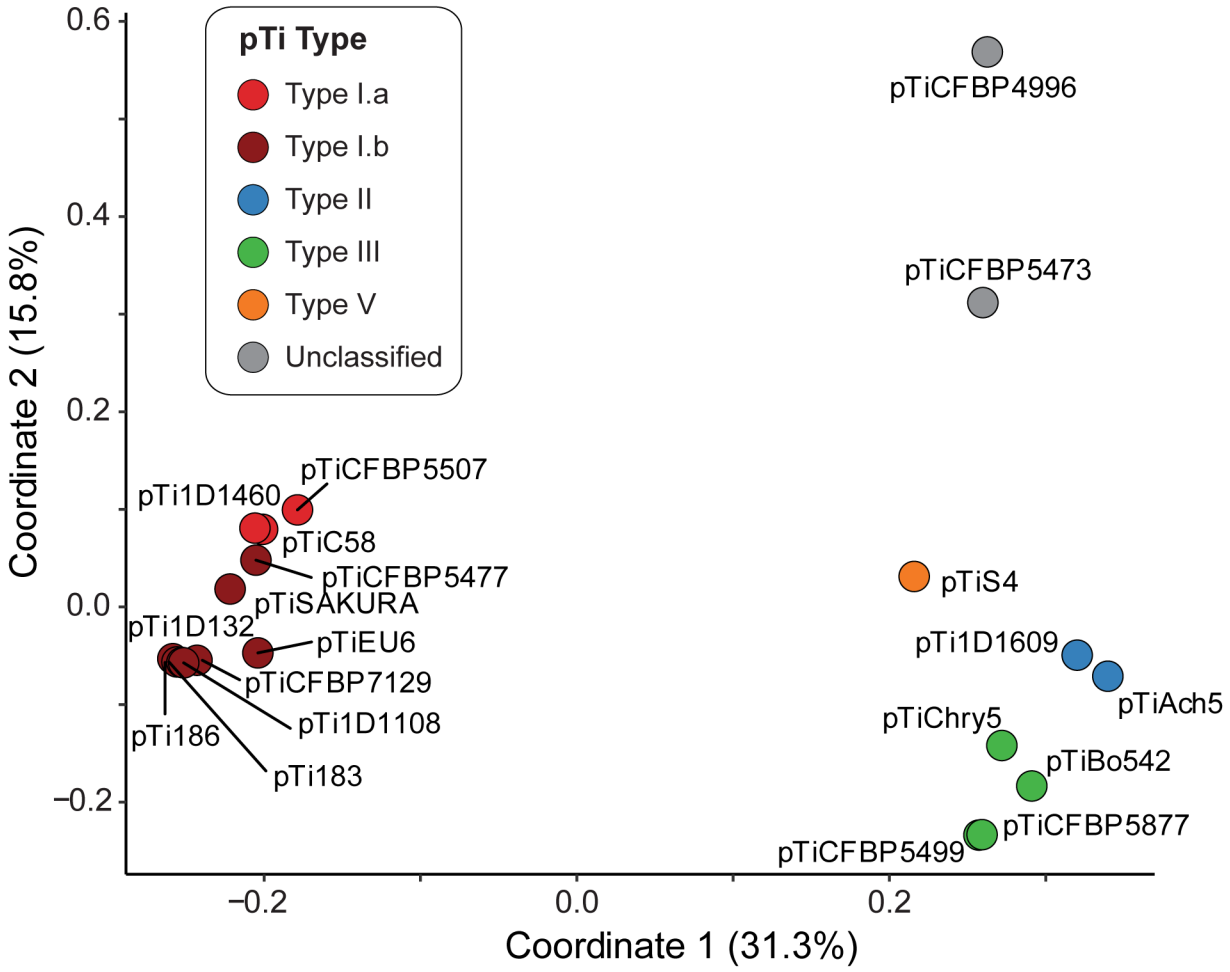

A
Domain 1
(TIGR03361)
C58 Atu4348
1D132 At1D132_RS21400
A1
1D1460 At1D1460_RS14800
A2
1D132 At1D132_RS23920
A1
CFBP7129 CFBP7129_RS25485
1D132 At1D132_RS18005
B1
C58 Atu3642
12D13 At12D13_RS20980
183 DC439_RS21805
A3
12D1 At12D1_RS23095
186 CG010_RS22900
A1
186 CG010_RS21820
B1
1D1460 At1D1460_RS21395
CFBP4996 CFBP4996_RS29220
A3
1D1609 At1D1609_RS23245
CFBP7129 CFBP7129_RS22695
B1
1D1609 At1D1609_RS22460
Ach5 Ach5_RS22800
S56 AGR1B_RS23320
N2/73 A6U91_RS21335
E1
MAFF210266 EXN61_RS24415
CFBP6623 CFBP6623_RS21645
CFBP6624 CFBP6624_RS21365
1D1108 At1D1108_RS23250
CFBP4996 CFBP4996_RS25360
D1
5A AT5A_RS17190
GBBC3283 EXN23_RS14990
CFBP5507 CFBP5507_RS22530
C1
1D1609 At1D1609_RS26290
CFBP4996 CFBP4996_RS01315
B1
1D1609 At1D1609_RS18895
CFBP6623 CFBP6623_RS00635
A1
CFBP6624 CFBP6624_RS00685
CFBP6623 CFBP6623_RS24605
B1
MAFF210266 EXN61_RS09490
A3
MAFF210266 EXN61_RS14250
GBBC3283 EXN23_RS10780
A1
S56 AGR1B_RS27960
S56 AGR1B_RS27940
CFBP5473 CFBP5473_RS14430
CFBP5477 CFBP5477_RS08655
C1
CFBP5477 CFBP5477_RS20375
D2
CFBP5473 CFBP5473_RS21265
CFBP5477 CFBP5477_RS22925
F1
Bootstrap Support
≥70
<70
Label
Strain & Locus tag | VgrG Subtype
Location of the vgrG
Main T6SS Gene Cluster
Accessory Gene Cluster
0.2
B
Domain 2
C58 Atu4348
C58 Atu3642
186 CG010_RS22900
186 CG010_RS21820
1D1609 At1D1609_RS18895
1D1609 At1D1609_RS22460
1D1609 At1D1609_RS23245
1D1609 At1D1609_RS26290
1D1460 At1D1460_RS14800
1D1460 At1D1460_RS21395
1D132 At1D132_RS18005
1D132 At1D132_RS21400
1D132 At1D132_RS23920
12D1 At12D1_RS23095
12D13 At12D13_RS20980
183 DC439_RS21805
CFBP5477 CFBP5477_RS08655
CFBP4996 CFBP4996_RS01315
CFBP4996 CFBP4996_RS29220
CFBP5507 CFBP5507_RS22530
CFBP5473 CFBP5473_RS14430
CFBP6623 CFBP6623_RS00635
CFBP6623 CFBP6623_RS24605
CFBP7129 CFBP7129_RS25485
CFBP6624 CFBP6624_RS00685
GBBC3283 EXN23_RS10780
GBBC3283 EXN23_RS14990
MAFF210266 EXN61_RS14250
S56 AGR1B_RS27960
S56 AGR1B_RS27940
MAFF210266 EXN61_RS09490
CFBP7129 CFBP7129_RS22695
A1/A2/A3/B1
Ach5 Ach5_RS22800
N2/73 A6U91_RS21335
S56 AGR1B_RS23320
MAFF210266 EXN61_RS24415
E1
1D1108 At1D1108_RS23250
CFBP4996 CFBP4996_RS25360
CFBP6624 CFBP6624_RS21365
CFBP6623 CFBP6623_RS21645
5A AT5A_RS17190
D1
CFBP5477 CFBP5477_RS20375
D2
CFBP5477 CFBP5477_RS22925
CFBP5473 CFBP5473_RS21265
F1
1.0
C
Domain 3
C58 Atu4348
1D132 At1D132_RS21400
1D132 At1D132_RS23920
CFBP6623 CFBP6623_RS00635
CFBP6624 CFBP6624_RS00685
186 CG010_RS22900
S56 AGR1B_RS27960
MAFF210266 EXN61_RS14250
S56 AGR1B_RS27940
CFBP5473 CFBP5473_RS14430
GBBC3283 EXN23_RS10780
A1
186 CG010_RS21820
CFBP7129 CFBP7129_RS25485
B1
GBBC3283 EXN23_RS14990
CFBP5507 CFBP5507_RS22530
CFBP5477 CFBP5477_RS08655
C1
C58 Atu3642
1D1609 At1D1609_RS18895
1D1609 At1D1609_RS22460
1D1609 At1D1609_RS23245
1D1460 At1D1460_RS21395
1D132 At1D132_RS18005
12D13 At12D13_RS20980
CFBP7129 CFBP7129_RS22695
CFBP4996 CFBP4996_RS01315
B1
12D1 At12D1_RS23095
183 DC439_RS21805
MAFF210266 EXN61_RS09490
CFBP4996 CFBP4996_RS29220
A3
1D1460 At1D1460_RS14800
A2
1D1108 At1D1108_RS23250
CFBP4996 CFBP4996_RS25360
CFBP6624 CFBP6624_RS21365
CFBP6623 CFBP6623_RS21645
5A AT5A_RS17190
D1
CFBP5477 CFBP5477_RS20375
D2
MAFF210266 EXN61_RS24415
Ach5 Ach5_RS22800
S56 AGR1B_RS23320
N2/73 A6U91_RS21335
E1
CFBP6623 CFBP6623_RS24605
1D1609 At1D1609_RS26290
B1
CFBP5473 CFBP5473_RS21265
CFBP5477 CFBP5477_RS22925
F1
1.0

Supplementary Figure S5. Maximum likelihood phylogenies of *vgrG*-associated domains. (A) Domain 1 (TIGR03361; VI_Rhs_Vgr super family), (B) Domain 2 (unknown function), and (C) Domain 3 (unknown function).

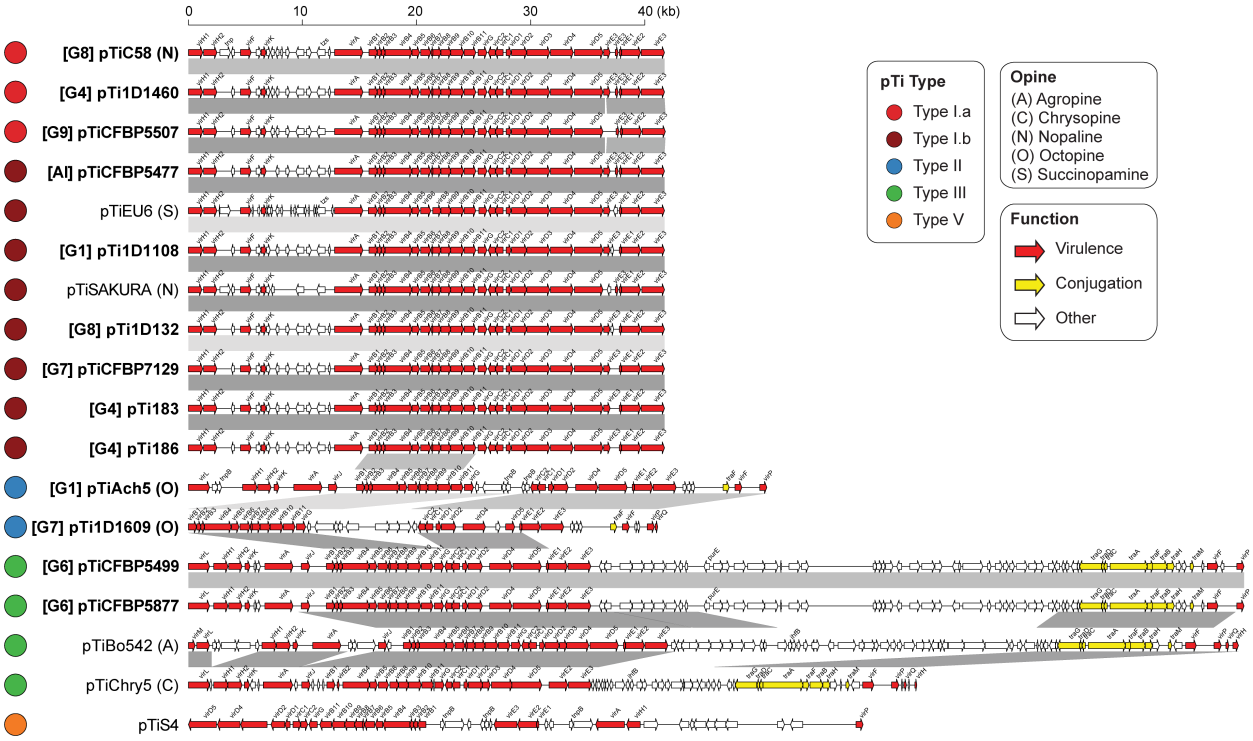


Supplementary Figure S6. Maximum likelihood phylogeny of pTi based on the concatenated alignment of shared single-copy genes. (A) All of the 20 pTi sequences analyzed; 21 core genes and 8,534 aligned amino acid sites. (B) Excluding the two

unclassified pTi; 40 core genes and 15,473 aligned amino acid sites, all branches received > 80% bootstrap support.

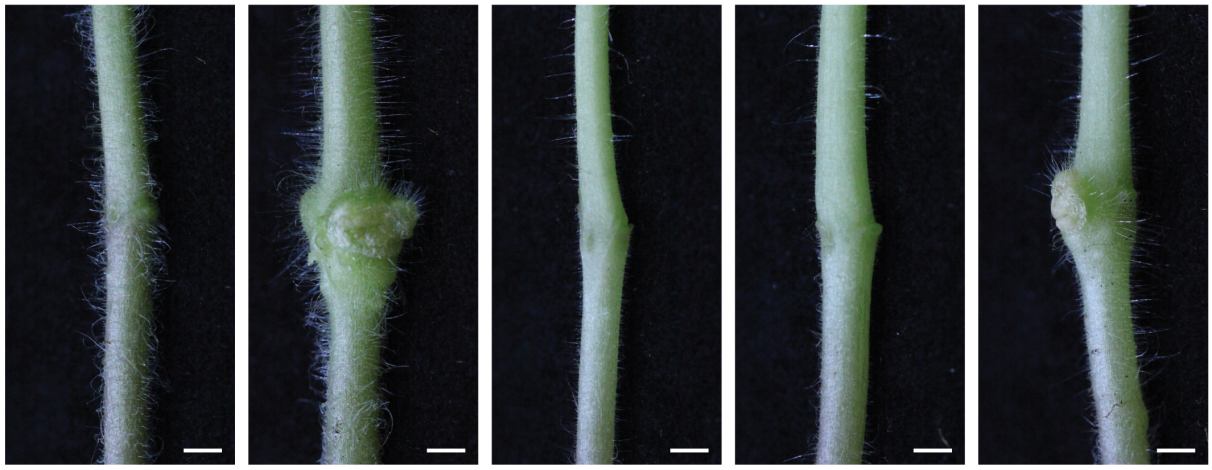


Supplementary Figure S7. Principal coordinate analysis of gene content among the pTi analyzed.



Supplementary Figure S8. Gene organization of the *vir* regulons on pTi. Syntenic regions are indicated by grey blocks. The virulence (*vir*) genes are highlighted in red, the conjugation (*tra*) genes are highlighted in yellow, and other genes are plotted in white.

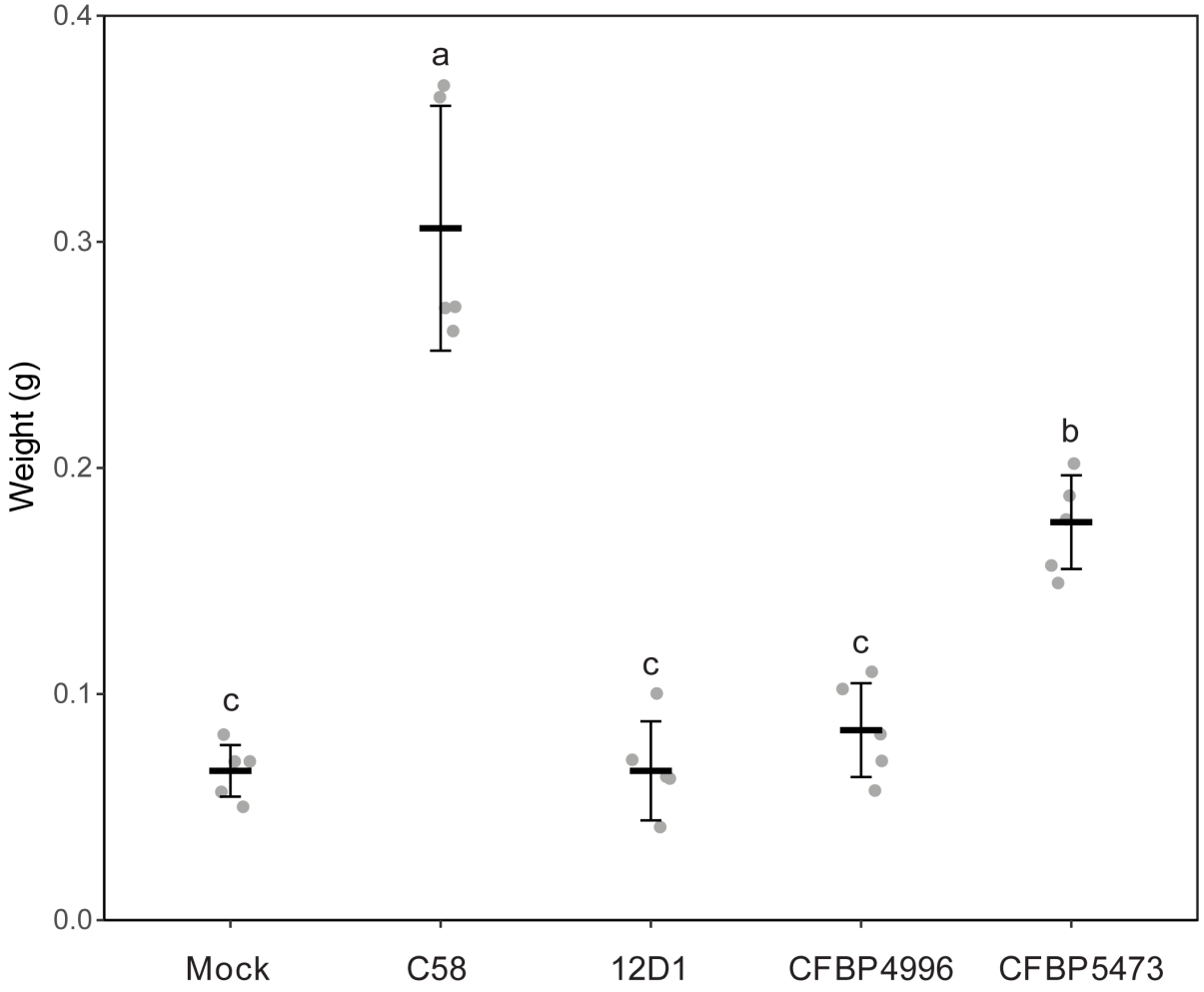
A



Mock C58 12D1 CFBP4996 CFBP5473

(scale bar = 0.25 cm)

B



Supplementary Figure S9. Tomato tumor assay of strains 12D1, CFBP4996, and CFBP5473. Mock was inoculated with sterilized water as a negative control and C58 was included as a positive control. The strain 12D1 harbors a plasmid with opine transporter and catabolism genes but lacks *vir* regulon genes and identifiable T-DNA. CFBP4996 and CFBP5473 harbor novel types of putative tumor-inducing plasmids (pTi). (A) Tomato stems at three weeks after inoculation. Scale bar: 0.25 cm. (B) Weight distribution of five biological replicates (1-cm segments of the stem centered at the inoculation site). The letters indicate ANOVA results.

Supplementary Data Set S1. List of *vgrG*-associated genes. Information including genomic location, RefSeq annotation, and domain prediction are included.

Supplementary Data Set S2. Locus tags of the *vir* regulon genes. The *virA*/*J* of 1D1609 are located on another plasmid and are highlighted by "*". The *virB7* of pTiChry5 and pTiEU6 are unannotated in the GenBank RefSeq records so no locus tag is available, but the gene presence could be confirmed by BLASTN searches.