Modular evolution of secretion systems and virulence plasmids in a bacterial species complex

Running Title: **Modular evolution of *Agrobacterium* genomes**

Lin Chou[1#], Yu-Chen Lin[1#], Mindia Haryono[1], Mary Nia M. Santos[1,2,3], Shu-Ting Cho[1], Alexandra J. Weisberg[4], Chih-Feng Wu[1,4], Jeff H. Chang[4], Erh-Min Lai[1,2,5], Chih-Horng Kuo[1,2,5]*

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

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

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

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

[5] Biotechnology Center, National Chung Hsing University, Taichung, Taiwan

[#]Equal contribution; ordered alphabetically

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

Abstract

Background

Many bacterial taxa are species complexes and uncertainties regarding the organization of their genetic diversity challenge research efforts. We utilized *Agrobacterium tumefaciens*, a taxon known for its phytopathogenicity and applications in transformation, as a study system and devised strategies for investigating genome diversity and evolution of species complexes.

Results

We utilized 35 genome assemblies to achieve a comprehensive and balanced sampling of *A. tumefaciens*. Our confident inference of gene content and core-genome phylogeny supported a quantitative guideline for delineating 12 species and allowed for robust investigations of genes critical in fitness and ecology. For the type VI secretion system (T6SS) involved in interbacterial competition and thought to be conserved, we detected multiple losses and one horizontal gene transfer. For the tumor-inducing plasmids (pTi) and pTi-encoded type IV secretion system (T4SS) that are essential for agrobacterial phytopathogenicity, we uncovered novel diversity and hypothesized their involvement in shaping this species complex. Intriguingly, for both T6SS and T4SS, genes encoding structural components are highly conserved, whereas extensive diversity exists for genes encoding effectors and other proteins.

Conclusions

We demonstrated that the combination of a phylogeny-guided sampling scheme and an emphasis on high-quality assemblies provides a cost-effective approach for robust analysis in evolutionary genomics. Our strategies for multi-level investigations at scales that range from whole-genomes to intragenic domains and phylogenetic depths of between- and within-species are applicable to other bacteria. Finally, modularity observed in the molecular evolution of genes and domains is useful for inferring functional constraints and informing experimental works.

Keywords

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

Background

Investigation of species boundaries in bacteria is fundamentally important because species are the key identifiers for biological entities [1,2]. However, many bacterial groups are currently unresolved and are classified as species complexes. These uncertainties regarding species boundaries hamper research, communication, and policy-making such as in healthcare guidelines, pathogen quarantine regulations, and biological resource management. Based on barriers to homologous recombination, an analysis of > 20,000 bacterial genome sequences from 91 species belonging to 13 phyla found that 21 of the previously recognized species comprise multiple biological species [3]. These 21 groups include those that are important as pathogens (e.g., *Mycobacterium tuberculosis*, *Pseudomonas aeruginosa,* and *Vibrio cholerae*) or beneficial microbes (e.g., *Lactobacillus casei* and *Sinorhizobium meliloti*). This finding highlights the ubiquity of species complexes across bacterial lineages, even for those that are extensively studied.

For such complexes, comprehensive understanding of the genetic diversity organization is required for robust species delineation, which in turn is essential for providing a reliable framework to interpretate experimental findings and to gain insights into the biology. The use of genomic information has long been suggested as a powerful approach for defining species boundaries because the comprehensive genetic information can provide definitive and potentially quantitative guidelines [4]. However, several issues regarding genomic studies of bacterial species have remained unresolved. First, while genomospecies defined by overall genome divergence were suggested to represent distinct biological entities [5–9], the exact criteria for establishing

81 the species boundaries are disputed. Although 95% average nucleotide identity (ANI)
82 across the conserved parts of genomes was proposed as a universal boundary for
83 defining species in bacteria [10], this criterion was challenged [11]. Additionally, ANI
84 values alone do not provide information such as gene content or phylogenetic
85 relationships, which are critical in understanding biological entities [4]. Second,
86 phylogenetic relationships among closely related bacterial strains often cannot be
87 resolved with confidence, yet such information is fundamental for evolutionary analysis.
88 Third, comparative genomics studies are often limited by taxon sampling and/or
89 assembly quality of available genome sequences.

90 In this study, we utilized the *Agrobacterium tumefaciens* species complex, also
91 known as *Agrobacterium* biovar 1 [12], as the study system for developing strategies
92 that provide appropriate sampling and utilize multifaceted genomic analysis to resolve
93 species boundaries and to investigate molecular evolution of key traits. These bacteria
94 are known as the causative agents of crown gall disease that affects over 90 plant
95 families [13]. More importantly, the development of *Agrobacterium*-mediated
96 transformation has provided a critical tool for genetic manipulation in plant sciences and
97 agricultural biotechnology [14,15]. Due to their importance, this complex has been
98 studied for over a century and was found to harbor extensive phenotypic and genetic
99 diversity that continues to confound efforts to resolve their taxonomy [12–14]. Various
100 methods, such as DNA-DNA hybridization, biochemical characteristics, and molecular
101 markers have been used to define 10 genomospecies (i.e., G1-G9 and G13), which
102 have continually been associated to new nomenclature, a process that causes greater
103 confusion than resolution [5,6,16–18]. For example, the reference strain C58 used in

104 many *A. tumefaciens* studies [15,19,20] belongs to G8, which was renamed as
105 *Agrobacterium fabrum* in 2011 [21]. This has resulted in mixed usage of two names with
106 different meanings (i.e., *A. tumefaciens* for the entire complex and *A. fabrum* for G8) in
107 databases and literature. Compounding confusion, the name *Agrobacterium radiobacter*
108 refers to *A. tumefaciens* G4 [18,21], the entire *A. tumefaciens* complex [22], and the
109 more divergent *Agrobacterium* biovar 2 [23]. Hereafter, we use *A. tumefaciens* in
110 reference to the entire species complex and specific designations (i.e., G1-G9 and G13)
111 for genomospecies.

112 Previous characterizations found that *A. tumefaciens* strains have multipartite
113 genomes with one circular chromosome, one linear chromosome, and highly variable
114 plasmids [9,23–25]. Consistent with the high levels of genetic divergence inferred from
115 DNA-DNA hybridization [5], cross-genomospecies comparisons typically found that <
116 80% of the genes are conserved [7,26]. Strikingly, > 32,000 horizontal gene transfer
117 (HGT) events have been inferred to have shaped the evolutionary history of *A.*
118 *tumefaciens* [8]. Because the HGT patterns indicated co-transfers of genes that encode
119 coherent biochemical pathways, it was hypothesized that purifying selection on those
120 acquired gene clusters and overall gene content drove the ecological diversification
121 among genomospecies [8]. Moreover, the oncogenic plasmids that determine
122 *Agrobacterium* phytopathogenicity exhibit complex modularity and transmission
123 patterns, which further contributed to the diversification of these pathogens and their
124 global spread [9]. However, despite the progresses, those better-characterized
125 genomospecies (e.g., G1, G4, G7, and G8) and pathogenic strains were highly

126 overrepresented in previous genomics studies [7–9], and such biases may affect our
127 understanding of agrobacterial diversity and evolution.

128 To develop effective strategies for investigating bacterial species complexes
129 such as *A. tumefaciens*, we started by performing targeted genome sequencing for
130 strains in underrepresented lineages to achieve a comprehensive and balanced taxon
131 sampling of the study system. We also limited analyses to only high-quality assemblies,
132 which enabled detailed examinations of replicon-level synteny and confident inferences
133 of gene presence/absence. The global view of genomic diversity and resolved
134 phylogeny provided a robust framework for focused investigations of the genetic
135 elements involved in key aspects of agrobacterial fitness and ecology, namely the type
136 VI secretion system (T6SS) for interbacterial competition [7,27] and the virulence
137 plasmids for phytopathogenicity [13–15]. Taken together, the investigations, scaling
138 from whole-genome, whole-replicon, gene clusters, individual genes, and intragenic
139 protein domains, provided novel and detailed information on the evolution and genetic
140 diversity of bacteria important in plant pathology and biotechnology. Moreover, the
141 strategies developed in this work are applicable to the study of other bacterial species
142 complexes.

143

144 **Results**

145 Genome Sampling, Molecular Phylogeny, and Divergence

146 Based on existing knowledge of *A. tumefaciens* diversity [5,6,16–18] and
147 availability of genomics resources [7–9], we identified 12 strains that represent the six
148 poorly characterized genomospecies and two strains from the sister lineage

149 *Agrobacterium larrymoorei* to provide an ideal outgroup (Table 1). Whole-genome
150 sequencing with substantial efforts in iterative improvements of the assemblies based
151 on experimental and bioinformatic approaches were conducted for these 14 strains.
152 Additionally, we conducted multiple rounds of preliminary genome-scale phylogenetic
153 analysis to select 21 representatives from 98 *A. tumefaciens* genome assemblies
154 available from GenBank [28] (Table 1) to yield a dataset with maximal genetic diversity
155 without emphasis on including pathogenic strains. To ensure balanced sampling, we
156 selected between two and five strains for each of the 10 recognized *A. tumefaciens*
157 genomospecies. Importantly, 19 of these 35 assemblies, including nine produced in this
158 study, are complete and most others are nearly complete (i.e., average N50 = 1.3 Mb;
159 cf. the two chromosomes are ~2.9 and ~2.3 Mb, respectively).

160 Based on the comprehensive sampling of *A. tumefaciens* diversity, we identified
161 a core genome of 2,093 single-copy genes, which correspond to ~40% of the genes
162 annotated in each individual genome sequence. Compared to previous studies that
163 conducted genome-based phylogenetic analysis for *Agrobacterium* or higher taxonomic
164 ranks [8,9,29], the more focused sampling in this study yielded a higher core gene count
165 by one-to-two orders of magnitude. This increase in core gene count and the
166 improvement in taxon sampling allowed for the inference of a well-resolved maximum
167 likelihood phylogeny of the *A. tumefaciens* species complex (Figure 1A). Each of the 10
168 currently recognized genomospecies forms a distinct monophyletic clade with > 80%
169 bootstrap support. Additionally, we identified two novel genomospecies, G14 and G15,
170 each represented by a single strain. The pattern of overall genome similarities exhibits a
171 discrete multimodal distribution that supports use of a 95% ANI cutoff for delineating

172 bacterial species [10] and quantifies the divergence of the *A. tumefaciens* complex from
173 its most closely related sister lineage (Figure 1B).
174 The 12 *A. tumefaciens* genomospecies were classified into seven groups based
175 on 90% ANI, and further assigned to three supergroups according to the phylogeny
176 (Figure 1A). Based on a time-calibrated phylogeny [9], the most recent common
177 ancestor (MRCA) of *A. tumefaciens* emerged ~48 million years ago (Mya), the three
178 supergroups emerged ~40 Mya, the seven groups emerged ~30 Mya, and most of the
179 recognized genomospecies emerged ~2-10 Mya. The inferred rapid radiation in the
180 early history of *A. tumefaciens* likely prevented confident resolution of those deeper
181 relationships in previous studies [8,9,29]. Nevertheless, with improvements in the taxon
182 sampling of this work, we observed ~70% bootstrap support for those early nodes
183 (Figure 1A). This fully-resolved organismal tree provides a strong framework for our
184 downstream examination of gene and domain phylogenies.
185 The high levels of assembly completeness provided confident inference of gene
186 content comparisons. The principal coordinate analysis and hierarchical clustering
187 results indicated that all 12 *A. tumefaciens* genomospecies are similar to one another
188 while distinct from *A. larrymoorei* (Supplementary Figure S1A and S1C). Nonetheless,
189 with the exception of G4 and G7, these genomospecies are distinguishable based on
190 gene content (Supplementary Figure S1B and S1D). This finding suggests that despite
191 the extensive HGT inferred within this complex [8], the genomospecies defined by 95%
192 ANI likely represent distinct biological entities.
193

194 Diversity and Evolution of the T6SS Genes

195 The fully-resolved organismal phylogeny and confidence in gene content
196 inference, particularly for the cases of gene absence, provide a robust framework for
197 evolutionary analysis of key traits. We first focus on genes encoding the T6SS, a phage
198 tail-like contractile nanomachine commonly found in Proteobacteria and used to inject
199 effectors into eukaryotic or bacterial cells. The T6SS has major roles in pathogenesis,
200 symbiosis, and interbacterial competition [30–33]. For *A. tumefaciens*, the T6SS is a key
201 weapon for *in planta* competition between different genomospecies [7] and against
202 other bacteria [27]. Thus, investigating the diversity and evolution of T6SS genes may
203 shed light on a trait that influences the ecology and evolution of *A. tumefaciens*.

204 In a previous study that examined four *A. tumefaciens* genomospecies, T6SS-
205 mediated anti-bacterial activity was observed for all 11 strains sampled and thought to
206 be a conserved trait of this species complex [7]. To our surprise, among the 33 *A.*
207 *tumefaciens* strains examined in this work, a patchy distribution of the T6SS genes was
208 observed (Figure 1). Gene absences are in strains corresponding to previously under-
209 characterized genomospecies and were confirmed by examining syntenic regions and
210 using TBLASTN [34] to search entire genome sequences. For strains encoding a T6SS,
211 corresponding genes are consistently located on the linear chromosome and mostly
212 form a cluster of ~20 genes organized as two adjacent and oppositely oriented *imp* and
213 *hcp* operons [7,35] (Figure 2). Some strains harbor accessary loci containing *vgrG*
214 (involved in effector delivery) and other T6SS genes located elsewhere on the linear
215 chromosome [7,33,35] (Supplementary Figure S2). The T6SS gene phylogeny is largely
216 congruent with the species tree (Figure 2). One notable exception is that the MRCA of

217 G1 appears to have acquired the T6SS genes from a G8-related donor. Consistent with
218 this inference, the T6SS genes in G1 strains are located in a different chromosomal
219 location compared to strains of other genomospecies (Supplementary Figure S2).
220 Based on these observations, it is likely that a T6SS gene cluster was present in the
221 MRCA of G8-G6-G14-G2-G4-G7-G9-G3-G15 and at least two independent losses have
222 occurred in G6 and G14-G2. Regarding the ancestral state in the MRCA of the *A.*
223 *tumefaciens* complex, presence of the T6SS genes appears to be a more parsimonious
224 hypothesis based on the presence of these genes in the outgroup (Figure 2). However,
225 the lack of synteny conservation between the linear chromosomes of *A. tumefaciens*
226 and *A. larrymoorei* and the variable locations of *vgrG* homologs (Supplementary Figure
227 S2) suggest that multiple independent origins are also possible. For broader scales, the
228 T6SS genes have a patchy distribution among Rhizobiaceae [33,36], indicating that
229 these genes are not essential for these bacteria and have high rates of gains and
230 losses. Consistent with this, multiple pseudogenes confirmed by manual curation of
231 annotation were found (e.g., *tssH* in S56 and *tssA*/*vgrG* in ATCC31749) (Figure 2),
232 suggesting that for some strains these gene clusters are in the process of degradation
233 and will be eventually lost.
234 Examining synteny revealed that the *imp* operon, which encodes the majority of
235 T6SS structural components [37], is more conserved in gene composition and order
236 than the *hcp* operon, which often has different genes downstream of *vgrG* (Figure 2).
237 This genetic diversity may play a key role in interbacterial competition because genes
238 downstream of *vgrG* include those that encode effector and immunity (EI) protein pairs
239 [7]. The agrobacterial T6SS effectors often correspond to different toxins and the

240 cognate immunity proteins provide protection against self-intoxication [7,27,33,36]. The
241 rapid evolution of EI gene pairs is illustrated by three examples. First, despite the low
242 levels of sequence divergence among G1 strains (i.e., > 98% ANI in all pairwise
243 comparisons), different genes are found downstream of their *vgrG* homologs and this
244 variation is not consistent with either the species phylogeny or the T6SS core gene
245 phylogeny. Second, strain CFBP4996 of G7 has homologs of the same EI gene pair as
246 strains 12D1 and 183 of G4, rather than with other members of G7. Third, in both G3
247 strains, *vgrG* and the associated EI genes are located elsewhere on the linear
248 chromosome, rather than being a part of the *hcp* operon (Figure 2 and Supplementary
249 Figure S2). These results suggest that recombination involving gene modules has
250 contributed to the genetic diversity of these *A. tumefaciens* T6SS EI gene pairs.
251

252 Modularity of VgrG and Its Associated EI Pairs

253 The knowledge that *vgrG* homologs encode proteins with distinct C-terminal
254 domains responsible for binding specificities of different T6SS effectors for delivery
255 suggested that each *vgrG* homolog and its downstream EI gene pair may evolve as a
256 functional module [33,36]. Here, we sought to investigate the patterns of gene co-
257 occurrence and intra-module recombination to better understand the diversity and
258 evolution of these genes. For in-depth investigation of *vgrG* evolution, we began by
259 examining domain architecture of VgrG proteins and uncovered eight distinct domains
260 (Supplementary Figure S3). Based on differences in domain composition, the 44 *vgrG*
261 homologs, including 17 associated with the main T6SS gene cluster and 27 associated
262 with accessory loci, were classified into six major types and nine subtypes (Figure 3).

263 Only three of the domains are present in all VgrG variants. The N-terminal domain 1 is
264 the most conserved (Supplementary Figure S3) and the only one found in databases.
265 This domain corresponds to TIGR03361, which accounts for ~66-77% of the protein
266 length and the bulk of the structures that forms a trimeric complex analogous to a phage
267 tail spike [38,39] (Supplementary Figure S4). For domain 5 that was found in all *vgrG*
268 homologs belonging to subtypes A1-A3 and E1 (Figure 3), the presence of this domain
269 is perfectly correlated with the presence of a downstream DUF4123-domain-containing
270 gene (Figure 4). Because this DUF4123 domain acts as an adaptor/chaperone for
271 effector loading onto VgrG in *A. tumefaciens* [36] and *Vibrio cholerae* [40,41], this strong
272 co-occurrence suggests specific interactions between VgrG domain 5 and DUF4123.
273 Thus, combining domain analysis with gene co-occurrence provides a new strategy for
274 predicting the interacting domains of VgrG with other T6SS components.
275 Intriguingly, despite conservation of domain architecture within each subtype
276 (Figure 3), the phylogenies inferred from the three domains encoded by all *vgrG*
277 homologs do not have the same topology (Supplementary Figure S5). For domain 1,
278 sequences from the same subtype do not always form monophyletic groups. For
279 domains 2 and 3, the short sequence lengths limited the phylogenetic resolution;
280 nonetheless, low divergence within the same subtype and high divergence between
281 subtypes were observed. These patterns suggest that each domain-encoding region
282 evolved independently and can recombine between subtypes.
283 At the level of gene cluster organization, *vgrG* homologs within a subtype can
284 have distinct downstream genes (e.g., A1 and B1), regardless of whether they are
285 associated with the main T6SS gene cluster (Figure 4 and Supplementary Dataset

286 S1A). These findings suggest that in addition to domain shuffling among *vgrG*
287 homologs, recombination also facilitated novel *vgrG*-effector pairings in the evolution of
288 these T6SS genes.

289 When T6SS diversity was examined in a phylogenetic context, numbers and
290 types of *vgrG* homologs, as well as their linked EI genes, lack strong correlations with
291 species phylogeny (Figure 5). Based on our manual curation of *vgrG*-associated genes,
292 a total of 63 putative effector genes were identified (Supplementary Dataset S1A).
293 Among these, peptidoglycan-targeting toxins and nucleases are the two most commonly
294 found categories with 21 each. This finding is consistent with an investigation of > 1,000
295 T6SS effectors sampled from 466 species in the phylum Proteobacteria [32], which also
296 found these two as the dominant categories and suggested that they play
297 complementary roles in T6SS-mediated competition.

298

299 The Virulence Plasmids and Associated Genes

300 The tumor-inducing plasmids (pTi) are an important component of *A. tumefaciens*
301 genomes. These large accessory replicons harbor the virulence (*vir*) regulon genes that
302 encode the Vir proteins and type IV secretion system (T4SS) for processing and
303 delivering a transfer DNA (T-DNA) into plant cells, and are essential for agrobacterial
304 phytopathogenicity [13–15]. Among the 35 strains examined, we identified 15 pTi
305 sequences (Table 2). Two novel putative pTi (i.e., pTiCFBP4996 and pTiCFBP5473)
306 were found in the 14 newly sequenced strains. This efficiency of discovering novel pTi
307 types is surprising, given our previous study that defined pTi types I-VI was based on
308 comprehensive sampling of diverse historical collections containing 162 *Agrobacterium*

309 strains [9]. This finding demonstrates the importance and usefulness of a phylogeny-
310 guided approach for investigating genetic diversity. Our recent examination of > 4,000
311 Rhizobiaceae plasmids assigned these two novel pTi to types X and XI, and found that
312 both types are species-specific [42], These two novel pTi are distinctive in their large
313 sizes (Table 2), gene organization (Figure 6), gene content (Supplementary Figure S6),
314 and sequence divergence of core genes (Supplementary Figure S7). Moreover, their T-
315 DNA regions also differ from the typical sizes of ~18-26 kb observed in types I-III pTi
316 (Figure 6). For the tumorigenic strain CFBP5473 (Supplementary Figure S8), the
317 predicted T-DNA border sequences flank an extraordinarily large (~93 kb) region that
318 contains all of the *vir* regulon genes in addition to the typical T-DNA-associated genes
319 (e.g., synthesis of opine and plant hormone) (Figure 6). This reflects either a
320 translocation of a T-DNA border sequence or reliance on a non-canonical T-DNA border
321 sequence that we could not identify. For pTiCFBP4996, its 7-kb T-DNA is predicted to
322 contain only four genes (i.e., two correspond to opine synthesis and two encode
323 hypothetical proteins). Plant hormone synthesis genes, which are necessary to cause
324 visible disease symptoms, were not identified in this predicted T-DNA region or
325 elsewhere on this plasmid. Consistent with predictions, strain CFBP4996 did not induce
326 tumor formation when inoculated onto stems of tomato plants (Supplementary Figure
327 S8). Considering that pTiCFBP4996 encodes all essential *vir* genes required for T-DNA
328 processing and transfer, CFBP4996 may serve as a naturally disarmed strain capable
329 of T-DNA transfer without causing diseases.
330 For replicon-level comparisons, types II and III pTi are more similar to each other
331 than to type I pTi based on gene content (Supplementary Figure S6) and core gene

332 phylogeny (Supplementary Figure S7). Within type I, the two subtypes (i.e., I.a and I.b)
333 are distinguishable by gene content (Supplementary Figure S6) but do not form
334 mutually exclusive clades in the core-gene phylogeny (Supplementary Figure S7). All
335 putative pTi, including the two novel types and pTiS4 (i.e., type V from distantly-related
336 *Agrobacterium vitis*), contain genes for the T4SS that mediates T-DNA transfer into
337 plant cells (i.e., *virB1-B11* and *virD4*) and the corresponding two-component regulatory
338 system (i.e., *virA* and *virG*) (Figure 7). Strain 1D1609 is a notable case because its *virA*
339 and *virJ* are located on another plasmid, rather than the pTi [26]. For the other *vir*
340 regulon genes, several differences among pTi types were observed (Figure 7). For
341 example, while all pTi harbor a conserved *virE3a* that facilitates T-DNA protection and
342 entry into host [43], type I pTi harbor one or two additional copies of *virE3* that belong to
343 different sequence types (i.e., sharing the same annotation but classified as distinct
344 homologs due to sequence divergence). Similarly, *virF*, which encodes an F-box protein
345 that is a putative host-range determinant [44], can be classified into three sequence
346 types with distinct distributions. Those less well-characterized *vir* genes, such as *virD3*
347 [45,46], *virJ* [47], and *virP*, are also distributed differently among pTi types. Finally, in
348 addition to the presence/absence of individual genes, the overall organization of *vir*
349 regulons also differ among these pTi (Supplementary Figure S9). All type I pTi are
350 conserved in sharing a ~40 kb region that contains all *vir* genes. In comparison,
351 locations of *vir* genes are more variable among type II/III pTi; *virF/P* (and *virQ/H* if
352 present) are located ~5-50 kb away from the main *vir* gene cluster. Other than *vir*
353 genes, the gene content and organization of T-DNA are also different (Figure 8). All
354 type I pTi have one single T-DNA region, while types II and III have two and type V have

355 four, respectively [9] (Figure 6). Within the T-DNA regions, the plant hormone synthesis
356 genes (i.e., *tms1*/*iaaM*, *tms2*/*iaaH*, and *ipt*) are the most conserved ones, while others
357 are more variable (Figure 8). Taken together, this genetic variation may contribute to the
358 host range differences observed among strains harboring different types of pTi [48].
359

360 **Discussion**

361 Biological Entities at the Species Level and Above

362 Based on the divergence of core gene sequences (Figure 1A) and gene content
363 (Supplementary Figure S1), 95% ANI is a reliable approach for defining species within
364 the *A. tumefaciens* complex, as is the case for most other bacteria [10]. The discrete
365 multimodal distribution of genome similarities (Figure 1B) suggested that there are
366 genetic barriers between different genomospecies, which may be explained by neutral
367 processes and/or selection [8,10]. Regardless of the exact mechanisms, these patterns
368 supported application of the biological species concept, which is based on genetic
369 barriers, to these *A. tumefaciens* genomospecies and other bacteria [3]. With the
370 continuing drop in sequencing cost, it is foreseeable that ANI analysis can serve as a
371 practical or even a standard approach for accurate classification of additional strains,
372 which in turn could facilitate research and communication, and ideally leads to
373 improvements in bacterial taxonomy for basic works and applications.

374 At the above-species level, *A. tumefaciens* genomospecies exhibited some
375 intriguing patterns of genome divergence. In a previous study that compared ~90,000
376 prokaryotic genomes, it was extremely rare to find ANI values in the range of 82-96%
377 [10]. In other words, strains either belong to the same natural biological entity at the

378 species level and have > 95% ANI, or belong to different species and have < 82% ANI.
379 This observation could be due to biases in the sampling of available genomes, or the
380 all-against-all pairwise comparisons included mostly distantly-related species [11]. In
381 our study that provided a detailed examination of closely-related species, the ~85-93%
382 ANI among *A. tumefaciens* genomospecies (Figure 1B) indicated that the *A.*
383 *tumefaciens* complex is indeed a coherent entity with high divergence from its closest
384 sister lineage within the same genus (i.e., *A. larrymoorei*). The driving forces for
385 maintaining species complexes and the prevalence of such above-species level entities
386 are interesting questions that require further investigations. For the *A. tumefaciens*
387 complex, although the nomenclature originated from its phytopathogenicity, it is well-
388 established that this group contains both pathogenic and non-pathogenic strains that
389 differ in the possession of an oncogenic plasmid (i.e., pTi) or not [12]. The promiscuous
390 nature of their pTi [9,49–51] (Figure 6) suggested that lineages within this complex may
391 experience frequent transitions between pathogenic and non-pathogenic lifestyles, and
392 such shared ecological niches may be the force that maintains the coherence of this
393 species complex. Compared to sister lineages (e.g., *A. larrymoorei* and *A. rubi*), the
394 more diverse host range of *A. tumefaciens* [12,13] may have facilitated the divergence
395 of this complex into multiple genomospecies. To test this hypothesis, better sampling of
396 these sister lineages is required.
397 At the genus level and above, genome-based classification is more challenging.
398 The ANI approach is expected to have limited resolution when nucleotide sequence
399 identities are below ~80% [10]. Moreover, the fractions of genome sequences alignable
400 for ANI value calculation are highly variable for genus-level comparisons [52], which

401 raises concerns on the robustness of applying the ANI method to higher taxonomic
402 ranks. To resolve this challenge, analysis of protein sequence divergence among core
403 genes was proposed as a suitable approach [53]. However, while it may be desirable to
404 establish a standardized taxonomy with a defined range of genomic divergence for each
405 taxonomic rank, large variations in the divergence values at a given rank were observed
406 among different taxonomic groups in previous attempts [52–54]. These variations
407 created situations where some families contain higher divergence levels than some
408 orders or lower divergence levels than some genera, even after normalization and a full
409 revision of the current taxonomy [53]. Such situations demonstrated the challenges of
410 establishing a new standardized taxonomy even when the practical issues of
411 transitioning from the current taxonomy are not considered, and perhaps is to be
412 expected given the highly variable evolutionary rates across different lineages [55].
413 Given these considerations, other aspects of biology (e.g., physiology, ecology, etc)
414 may play more important roles in defining those higher taxonomic ranks.
415
416 Units and Modularity of Molecular Evolution
417 For evolutionary studies, the levels at which selection and other processes
418 operate on have been a topic that received much attention [56]. For prokaryotes, levels
419 from the entire genome to individual functional domains within genes are of particular
420 interest. Based on our results, all of these levels must be considered to comprehend the
421 complex patterns.
422 At the whole-genome level, the clear species boundaries based on overall
423 similarity (Figure 1 and Supplementary Figure S1) suggested that the entire genome

424 largely evolves as a single coherent unit. This result is consistent with previous findings
425 that at the global level HGT has very little impact on the reconstruction of organismal
426 phylogeny [57,58], despite the extensive HGT inferred in bacterial evolution [58–61] and
427 the importance of HGT in adaptation [62–64]. A possible explanation for these
428 seemingly conflicting observations is that most of the acquired genes are lost quickly
429 [63], presumably due to the strong mutational bias towards deletions observed in
430 bacteria [65,66]. Additionally, acquired genes are subjected to the selection that drives
431 species diversification [8], which is expected to act on all genes in a genome together.
432 At the level of individual replicons, chromosomes and plasmids certainly have
433 distinct evolutionary histories (Figure 6). Because novel chromosome/plasmid
434 combinations may lead to speciation [50], and the spread of plasmids have important
435 implications on the evolution of virulence [9,67] and antimicrobial resistance [68], further
436 investigations on the evolution of plasmids and their compatibilities with chromosomes
437 are important [9,69,70]. Additionally, for bacteria with multiple chromosomes, examining
438 the evolution of individual chromosomes may provide novel insights. In the case of *A.*
439 *tumefaciens*, the multipartite genome was hypothesized to originate from intragenomic
440 gene transfer from the ancestral circular chromosome to a plasmid, followed by
441 linearization of this plasmid to form the secondary chromosome [23]. The secondary
442 chromosome is known to exhibit higher levels of divergence in overall organization,
443 gene content, and sequences [23,26]. In this regard, it is curious to note that the
444 apparently rapid-evolving T6SS genes are all located on the secondary chromosome,
445 rather than the more conserved primary chromosome. For future studies, it may be

446 interesting to compare the molecular evolution of T6SS and other genes between
447 species with mono- and multi-partite genomes.
448 At the levels of gene clusters and below, several interesting observations were
449 made based on the loci of the two secretion systems investigated in this work. First,
450 although these two systems may provide some fitness advantages (e.g., T6SS for
451 interbacterial competitions and T4SS for host exploitation), complex patterns of gains
452 and losses were observed (Figures 2 and 6). These patterns suggested that there is not
453 a strong selective pressure to maintain these genes and non-adaptive stochastic
454 processes are important. Furthermore, there is a certain degree of modularity regarding
455 their evolution, such that the presence patterns are all-or-none and no partial cluster
456 was found for the chromosomal T6SS genes or the plasmid-encoded T4SS genes. This
457 is particularly evident for the T6SS genes, as when the main cluster was lost (e.g., G2
458 and G6), no accessory *vgrG* loci located elsewhere was found (Supplementary Figure
459 S3). Second, for both systems, genes for the structural components are conserved,
460 while those for the effectors and others are not (Figures 2, 5, and 7). This is consistent
461 with the expectation that opposite selective forces may act on these two categories of
462 genes, with purifying selection against changes to preserve the apparatus of a
463 functional secretion system and positive selection for more diverse effectors and other
464 accessory components. Third, modularity that may reflect functional constraints were
465 observed at finer scales. For example, each *vgrG* is linked to its cognate
466 effector/chaperone genes, and each effector is linked to its cognate immunity gene.
467 Such modularity is expected to be maintained by selection, similar to the observations
468 regarding co-transfers of genes involved in associated biochemical pathways [8].

469 However, the linkage could be broken down by recombination at within- or between-
470 species levels, as evident in the diversity of *vgrG* gene neighborhoods, even for those
471 homologs belonging to the same type (Figure 4). Fourth, at the level of individual genes,
472 the within-genome diversity of *vgrG* homologs (Figure 5) provides further support to the
473 hypothesis that HGT is more important than duplications in driving gene family
474 expansions in bacteria [71]. Finally, at the intra-gene level, within- or between-species
475 recombination may be important in generating novel combinations of domains, thus
476 promoting the diversification of homologs (Figure 3 and Supplementary Figure S6).
477 Taken together, these observations illustrated the complexity of biological
478 systems. While it is difficult to draw up generalized rules or to estimate the relative
479 importance of each evolutionary process at different levels, it is important to consider
480 and examine these complexities to better understand organisms of interest.
481
482 Conclusions
483 In summary, by using a group of important bacteria as the study system, this
484 work utilized a strategy of phylogeny-conscious genome sampling for systematic
485 investigations of a species complex. This approach is not only cost-effective but also
486 critical in obtaining an unbiased picture that does not over-emphasize certain
487 subgroups. Additionally, the emphasis on generating and utilizing high-quality
488 assemblies improves the confidence in gene content analysis. With the continuing
489 advancements in sequencing technologies and bioinformatic tools, this emphasis
490 becomes increasingly accessible. For this study system, our examination of biological
491 boundaries at the species level and above improves the understanding of how natural

492 biodiversity is organized. The targeted analysis of those secretion system genes and
493 oncogenic plasmids provides novel insights regarding the key genetic variations
494 involved in the fitness and ecology of these soil-borne phytopathogens that need to
495 compete in complex microbiota and invade plant hosts. Moreover, the multi-level
496 analysis of their genetic diversity from whole-genome to intra-genic domains highlights
497 the complexity of these biological systems. The strategy and findings of this work
498 provide useful guides for future studies of other bacteria.
499

500 **Materials and Methods**

501 Genome Sequencing

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

506 The procedure for whole-genome shotgun sequencing was based on that
507 described in our previous studies [7,72,73]. All bioinformatics tools were used with the
508 default settings unless stated otherwise. Briefly, total genomic DNA was prepared using
509 the Wizard Genomic DNA purification kit (Promega, USA). The Illumina paired-end
510 sequencing libraries were prepared using KAPA LTP Library Preparation Kits (Roche
511 Sequencing, USA) with a targeted insert size of ~550 bp. The Illumina MiSeq platform
512 was used to generate 300x2 reads with an average coverage of 306-fold per strain
513 (range: 141- to 443-fold). The raw reads were quality trimmed using a Q20 cutoff and
514 used for *de novo* assembly based on Velvet v1.2.10 [74] with the settings "-exp_cov

515 auto -min_contig_lgth 2000 -scaffolding no". The contigs were oriented by mapping to
516 those complete genome assemblies available (Table 1) using MAUVE v2015-02-13
517 [75]. Due to the difficulties of identifying appropriate reference genomes for several
518 evolutionary branches, four strains (i.e., CFBP5473, CFBP5875, CFBP5877, and
519 CFBP6623) were selected for PacBio long-read sequencing and PacBio HGAP v3
520 assembly. These PacBio-based assemblies were used as a guide for scaffolding, rather
521 than the finalized results.

522 To improve the Illumina-based draft assemblies, an iterative process was used to
523 examine the raw reads mapping results and to incorporate gap-filling results based on
524 PCR and Sanger sequencing. This process was repeated until the complete assembly
525 was obtained or the draft assembly could not be improved further. The finalized
526 assemblies were submitted to the National Center for Biotechnology Information (NCBI)
527 and annotated using the Prokaryotic Genome Annotation Pipeline (PGAP) [76].

528

529 Comparative and Evolutionary Analysis

530 The genomes analyzed are listed in Table 1. The procedures for genome
531 comparisons were based on those described in our previous studies [7,77–79]. Briefly,
532 pairwise genome similarities were calculated using FastANI v1.1 [10]. For comparisons
533 of plasmids, FastANI was executed with the custom settings that reduced fragment
534 length to 1,000 bp and minimum matched fragments to 25. For global alignments of
535 chromosomes and plasmids, the syntenic regions were identified by BLASTN v2.6.0
536 [34] and visualized using genoPlotR v0.8.9 [80]. For gene content comparison, BLASTP
537 v2.6.0 [34] with e-value cutoff set to $1e^{-15}$ and OrthoMCL v1.3 [81] were used to infer the

538 homologous gene clusters. The result was converted into a matrix of 35 genomes by
539 17,058 clusters, with the value in each cell corresponding to the copy number. This
540 matrix was further converted into a Jaccard distance matrix among genomes using the
541 VEGAN package v2.5-6 in R, then processed using the principal coordinates analysis
542 function in the APE package [82] and visualized using ggplot2 v3.3.2 [83]. The
543 hierarchical clustering analysis was performed using PVCLUST v3.4.4 [84].
544 For phylogenetic analysis, homologous sequences were aligned using MUSCLE
545 v3.8.31 [85] for maximum likelihood inference by PhyML v.3.3.20180214 [86]. The
546 proportion of invariable sites and the gamma distribution parameter were estimated
547 from the data set, the number of substitute rate categories was set to four. The
548 bootstrap supports were estimated based on 1,000 replicates.
549
550 Analysis of the Type VI Secretion System Genes
551 To identify the T6SS-associated genes, C58 [27,36,37] and other strains [7,33]
552 that have been characterized experimentally were used as the references. Based on the
553 known T6SS genes in these genomes, homologous genes in other genomes were
554 identified based on the OrthoMCL result. To screen for novel T6SS effector, chaperone,
555 and immunity genes, genes that are located near *vgrG* (i.e., three upstream and ten
556 downstream) were examined manually by using the NCBI conserved domain database
557 (CDD) [87] and the Phyre2 protein fold recognition server [88]; the e-value cutoff was
558 set to 0.01. A few genes with a hit to the DUF4123 domain (i.e., a known T6SS
559 chaperone) but have an e-value above the cutoff were manually added back to the list
560 of T6SS-associated genes (e.g., CFBP5477_RS20350). To confirm the absence of

561 specific T6SS genes, the genome sequences were used as the subjects and the protein
562 sequences of known genes were used as the queries to run TBLASTN searches.
563 For classification of the *vgrG* homologs, we developed a domain-based scheme.
564 The conserved N-terminal TIGR03361 domain was first identified by the NCBI CDD
565 searches. A global alignment of all homologs was used to determine the exact
566 boundaries of this domain. After this TIGR03361 domain was removed, the remaining
567 C-terminal sequences were processed using MEME v5.1.1 [89] to identify conserved
568 domains that meet these criteria: (1) present in at least two sequences, (2) with zero or
569 one occurrence per sequence, (3) with a size between 30 and 300 a.a., and (4) with an
570 e-value lower than 0.05. The results were manually curated to break down large
571 domains that are composed of smaller domains. Pairwise BLASTP searches were
572 conducted to verify that each domain is unique and no two domains have a BLASTP e-
573 value of lower than 1e-05. For each domain, the consensus sequence was generated
574 using Jalview v2.10.5 [90] and sequence conservation was visualized using WebLogo
575 server v3 [91]. For functional prediction, the consensus sequence of each domain was
576 used to query against NCBI CDD and Phyre2. Additionally, one representative from
577 each subtype of *vgrG* homologs was used for structure modeling using Phyre2 with the
578 “normal” mode. The chain D of PA0091 VgrG1 (PDB identifier: 4MTK) was selected as
579 the template. The predicted structures were visualized using PyMOL v1.2r3pre
580 (Schrödinger, USA).
581 For the EI gene pairs identified, EI1 through EI11 were named based on the
582 nomenclature proposed previously [7] and novel pairs were named starting from EI12.
583 When only the putative effector (E) or the putative immunity (I) genes were found, those

584 genes were classified in the format of “E??” or “I??”, respectively. For some of the EI
585 pairs that were identified previously based on adjacency to T6SS genes but lacked
586 high-confidence annotation (i.e., EI2, EI3, EI5, and EI8), we chose a more conservative
587 approach and annotated those genes as hypothetical proteins.
588
589 Analysis of the Tumor-inducing Plasmids and Type IV Secretion System Genes
590 The list of 20 putative pTi sequences analyzed is provided in Table 2. These
591 included all of the 15 complete pTi sequences determined in this study and five
592 representatives from GenBank that are important in *Agrobacterium* research [15]. The
593 pTi typing was performed based on k-mer profile clustering with a reference set of 143
594 oncogenic plasmids in Rhizobiaceae [9] and a second set that contains > 4,000
595 Rhizobiaceae plasmids [42]. For T-DNA identification, putative T-DNA borders were
596 identified based on the motif YGRCAGGATATATNNNNNKGTMAWN [92]. Genes
597 involved in opine metabolism [93] and T4SS [94] were identified based on the
598 annotation and homologous gene clustering results produced by OrthoMCL.
599 Additionally, putative T4SS effectors were identified using the T4SEpre tool [95] in
600 EffectiveDB [96] with the minimal score set to 0.8. All protein sequences of pTi-encoded
601 genes were used as the queries.
602
603 Tumorigenesis Assay
604 Tomato tumorigenesis assays [97] were performed to evaluate the virulence of
605 selected strains. The plants (cultivar Known-You 301) were maintained in growth
606 chambers with a 16-/8-h light/dark regime and a constant temperature of 22°C.

607 Inoculation was performed on three-week old seedlings. Bacterial strains were
608 transferred from stock to 5 mL 523 broth and cultured overnight at 28°C in a shaker
609 incubator (250 rpm), then sub-cultured for four hours prior to inoculation. Bacterial cells
610 were washed and resuspended in 0.9% NaCl solution with a concentration of OD_{600} 0.2.
611 The stem was punctured with a sterilized sewing needle and 5 μL of bacterial
612 suspension was added to the wounding site. The plants were collected three weeks
613 after inoculation and 1-cm stem segments centered at the wounding site were cut for
614 weighing.
615

616 **Data Availability**

617 The 14 new genome sequences are available in NCBI under BioProject
618 accessions PRJNA534385-PRJNA534397 and PRJNA534399.
619

620 **Competing Interests**

621 The authors declare that they have no competing interests.
622

623 **Author Contribution**

624 Conceptualization: CHK
625 Funding acquisition: JHC, EML, CHK
626 Investigation: LC, YCL, MH, MNS, AJW, CFW
627 Methodology: LC, YCL, MNS, STC, CHK
628 Project administration: CHK
629 Supervision: JHC, EML, CHK

630 Validation: LC, YCL, MH, MNS, AJW

631 Visualization: LC, YCL

632 Writing – original draft: LC, YCL, CHK

633 Writing – review & editing: LC, YCL, AJW, CFW, JHC, EML, CHK

634

635 **Funding**

636 Research in the Chang lab was supported by the National Institute of Food and
637 Agriculture, US Department of Agriculture awards 2014-51181-22384 and 2020-51181-
638 32154. Research in the Lai lab was supported by Academia Sinica and the Ministry of
639 Science and Technology of Taiwan (MOST 104-2311-B-001-025-MY3 and MOST 107-
640 2311-B-001-019-MY3). Research in the Kuo lab was supported by Academia Sinica and
641 the Ministry of Science and Technology of Taiwan (MOST 109-2628-B-001-012). The
642 funders had no role in study design, data collection and interpretation, or the decision to
643 submit the work for publication.

644

645 **Acknowledgement**

646 We thank Ai-Ping Chen, Hsin-Ying Chiang, Shu-Jen Chou, Mei-Jane Fang, Ya-Yi
647 Huang, and Wen-Sui Lo for technical assistance. Sophien Kamoun provided helpful
648 comments that improved the writing of this manuscript. The bacterial strains were
649 imported under the permits 103-B-003 and 104-B-002 issued by the Council of
650 Agriculture of Taiwan. The Sanger sequencing service and the Illumina sequencing
651 library preparation service were provided by the Genomic Technology Core (Institute of
652 Plant and Microbial Biology, Academia Sinica). The Illumina MiSeq sequencing service

653 was provided by the Genomics Core (Institute of Molecular Biology, Academia Sinica).
654 The PacBio sequencing and data processing service was provided by Genomics BioSci
655 & Tech. Co. Ltd. (New Taipei City, Taiwan). The Institute of Plant and Microbial Biology
656 (Academia Sinica) and the Department of Botany and Plant Pathology (Oregon State
657 University) provided computing resources.
658

659 **References**

660 1. Rosselló-Mora R, Amann R. The species concept for prokaryotes. FEMS Microbiol Rev.
661 2001;25:39–67.

662 2. Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. The bacterial species challenge: making
663 sense of genetic and ecological diversity. Science. 2009;323:741–6.

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

666 4. Konstantinidis Konstantinos T, Ramette Alban, Tiedje James M. The bacterial species
667 definition in the genomic era. Philos Trans R Soc B Biol Sci. 2006;361:1929–40.

668 5. Popoff MY, Kersters K, Kiredjian M, Miras I, Coynault C. Position taxonomique de souches de
669 *Agrobacterium* d'origine hospitalière. Ann Inst Pasteur Microbiol. 1984;135:427–42.

670 6. Costechareyre D, Bertolla F, Nesme X. Homologous recombination in *Agrobacterium*:
671 potential implications for the genomic species concept in bacteria. Mol Biol Evol. 2009;26:167–
672 76.

673 7. Wu C-F, Santos MNM, Cho S-T, Chang H-H, Tsai Y-M, Smith DA, et al. Plant-pathogenic
674 *Agrobacterium tumefaciens* strains have diverse type VI effector-immunity pairs and vary in in-
675 planta competitiveness. Mol Plant Microbe Interact. 2019;32:961–71.

676 8. Lassalle F, Planel R, Penel S, Chapulliot D, Barbe V, Dubost A, et al. Ancestral genome
677 estimation reveals the history of ecological diversification in *Agrobacterium*. Genome Biol Evol.
678 2017;9:3413–31.

679 9. Weisberg AJ, Davis EW, Tabima J, Belcher MS, Miller M, Kuo C-H, et al. Unexpected
680 conservation and global transmission of agrobacterial virulence plasmids. Science. American
681 Association for the Advancement of Science; 2020;368:eaba5256.

682 10. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI
683 analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun.
684 2018;9:5114.

685 11. Murray CS, Gao Y, Wu M. Re-evaluating the evidence for a universal genetic boundary
686 among microbial species. Nat Commun. 2021;12:4059.

687 12. Young JM. *Agrobacterium*—taxonomy of plant-pathogenic *Rhizobium* species. In: Tzfira T,
688 Citovsky V, editors. Agrobacterium Biol Biotechnol [Internet]. Springer New York; 2008 [cited
689 2016 Oct 21]. p. 183–220. Available from: http://link.springer.com/chapter/10.1007/978-0-387-
690 72290-0_5

691 13. Kado CI. Historical account on gaining insights on the mechanism of crown gall
692 tumorigenesis induced by *Agrobacterium tumefaciens*. Front Microbiol. 2014;5:340.

693 14. Nester EW. *Agrobacterium*: nature's genetic engineer. Front Plant Sci. 2015;5:730.

694 15. Hwang H-H, Yu M, Lai E-M. *Agrobacterium*-mediated plant transformation: biology and
695 applications. Arab Book. 2017;15:e0186.

696 16. Mougel C, Thioulouse J, Perrière G, Nesme X. A mathematical method for determining
697 genome divergence and species delineation using AFLP. Int J Syst Evol Microbiol.
698 2002;52:573–86.

699 17. Portier P, Saux MF-L, Mougel C, Lerondelle C, Chapulliot D, Thioulouse J, et al.
700 Identification of genomic species in *Agrobacterium* biovar 1 by AFLP genomic markers. Appl
701 Environ Microbiol. American Society for Microbiology; 2006;72:7123–31.

702 18. Costechareyre D, Rhouma A, Lavire C, Portier P, Chapulliot D, Bertolla F, et al. Rapid and
703 efficient identification of *Agrobacterium* species by *recA* allele analysis: Agrobacterium recA
704 diversity. Microb Ecol. 2010;60:862–72.

705 19. Hellens R, Mullineaux P, Klee H. A guide to *Agrobacterium* binary Ti vectors. Trends Plant
706 Sci. 2000;5:446–51.

707 20. Lee L-Y, Gelvin SB. T-DNA binary vectors and systems. Plant Physiol. 2008;146:325–32.

708 21. Lassalle F, Campillo T, Vial L, Baude J, Costechareyre D, Chapulliot D, et al. Genomic
709 species are ecological species as revealed by comparative genomics in *Agrobacterium*
710 *tumefaciens*. Genome Biol Evol. 2011;3:762–81.

711 22. Young JM, Pennycook SR, Watson DRW. Proposal that *Agrobacterium radiobacter* has
712 priority over *Agrobacterium tumefaciens*. Request for an Opinion. Int J Syst Evol Microbiol.
713 2006;56:491–3.

714 23. Slater SC, Goldman BS, Goodner B, Setubal JC, Farrand SK, Nester EW, et al. Genome
715 sequences of three *Agrobacterium* biovars help elucidate the evolution of multichromosome
716 genomes in bacteria. J Bacteriol. 2009;191:2501–11.

717 24. Goodner B, Hinkle G, Gattung S, Miller N, Blanchard M, Qurollo B, et al. Genome sequence
718 of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. Science.
719 2001;294:2323–8.

720 25. Wood DW, Setubal JC, Kaul R, Monks DE, Kitajima JP, Okura VK, et al. The genome of the
721 natural genetic engineer *Agrobacterium tumefaciens* C58. Science. 2001;294:2317–23.

722 26. Haryono M, Cho S-T, Fang M-J, Chen A-P, Chou S-J, Lai E-M, et al. Differentiations in gene
723 content and expression response to virulence induction between two *Agrobacterium* strains.
724 Front Microbiol. 2019;10:1554.

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

728 28. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Ostell J, Pruitt KD, et al. GenBank.
729 Nucleic Acids Res. 2018;46:D41–7.

730 29. Ormeño-Orrillo E, Servín-Garcidueñas LE, Rogel MA, González V, Peralta H, Mora J, et al.
731 Taxonomy of rhizobia and agrobacteria from the Rhizobiaceae family in light of genomics. Syst
732 Appl Microbiol. 2015;38:287–91.

733 30. Hernandez RE, Gallegos-Monterrosa R, Coulthurst SJ. Type VI secretion system effector
734 proteins: Effective weapons for bacterial competitiveness. Cell Microbiol. 2020;22:e13241.

735 31. Jurėnas D, Journet L. Activity, delivery, and diversity of Type VI secretion effectors. Mol
736 Microbiol. 2021;115:383–94.

737 32. Smith WPJ, Vettiger A, Winter J, Ryser T, Comstock LE, Basler M, et al. The evolution of
738 the type VI secretion system as a disintegration weapon. PLOS Biol. Public Library of Science;
739 2020;18:e3000720.

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

743 34. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+:
744 architecture and applications. BMC Bioinformatics. 2009;10:421.

745 35. Wu H-Y, Chung P-C, Shih H-W, Wen S-R, Lai E-M. Secretome analysis uncovers an Hcp-
746 family protein secreted via a type VI secretion system in *Agrobacterium tumefaciens*. J
747 Bacteriol. American Society for Microbiology Journals; 2008;190:2841–50.

748 36. Bondage DD, Lin J-S, Ma L-S, Kuo C-H, Lai E-M. VgrG C terminus confers the type VI
749 effector transport specificity and is required for binding with PAAR and adaptor–effector
750 complex. Proc Natl Acad Sci. 2016;113:E3931–40.

751 37. Lin J-S, Ma L-S, Lai E-M. Systematic dissection of the *Agrobacterium* type VI secretion
752 system reveals machinery and secreted components for subcomplex formation. PLOS ONE.
753 2013;8:e67647.

754 38. Pukatzki S, Ma AT, Revel AT, Sturtevant D, Mekalanos JJ. Type VI secretion system
755 translocates a phage tail spike-like protein into target cells where it cross-links actin. Proc Natl
756 Acad Sci. National Academy of Sciences; 2007;104:15508–13.

757 39. Leiman PG, Basler M, Ramagopal UA, Bonanno JB, Sauder JM, Pukatzki S, et al. Type VI
758 secretion apparatus and phage tail-associated protein complexes share a common evolutionary
759 origin. Proc Natl Acad Sci. National Academy of Sciences; 2009;106:4154–9.

760 40. Liang X, Moore R, Wilton M, Wong MJQ, Lam L, Dong TG. Identification of divergent type VI
761 secretion effectors using a conserved chaperone domain. Proc Natl Acad Sci. National
762 Academy of Sciences; 2015;112:9106–11.

763 41. Unterweger D, Kostiuk B, Ötjengerdes R, Wilton A, Diaz-Satizabal L, Pukatzki S. Chimeric
764 adaptor proteins translocate diverse type VI secretion system effectors in *Vibrio cholerae*.
765 EMBO J. John Wiley & Sons, Ltd; 2015;34:2198–210.

766 42. Weisberg AJ, Miller M, Ream W, Grünwald NJ, Chang JH. Rapid diversification of plasmids
767 in a genus of pathogenic and nitrogen fixing bacteria. Philos Trans R Soc Lond B Biol Sci. 2021;

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

770 44. Jarchow E, Grimsley NH, Hohn B. virF, the host-range-determining virulence gene of
771 *Agrobacterium tumefaciens*, affects T-DNA transfer to Zea mays. Proc Natl Acad Sci. National
772 Academy of Sciences; 1991;88:10426–30.

773 45. Vogel AM, Das A. The *Agrobacterium tumefaciens* virD3 gene is not essential for
774 tumorigenicity on plants. J Bacteriol. American Society for Microbiology Journals;
775 1992;174:5161–4.

776 46. Lin T-S, Kado CI. The virD4 gene is required for virulence while virD3 and orf5 are not
777 required for virulence of *Agrobacterium tumefaciens*. Mol Microbiol. 1993;9:803–12.

778 47. Pan SQ, Jin S, Boulton MI, Hawes M, Gordon MP, Nester EW. An *Agrobacterium* virulence
779 factor encoded by a Ti plasmid gene or a chromosomal gene is required for T-DNA transfer into
780 plants. Mol Microbiol. 1995;17:259–69.

781 48. Hwang H-H, Wu ET, Liu S-Y, Chang S-C, Tzeng K-C, Kado CI. Characterization and host
782 range of five tumorigenic *Agrobacterium tumefaciens* strains and possible application in plant
783 transient transformation assays. Plant Pathol. 2013;62:1384–97.

784 49. Hooykaas PJJ, Klapwijk PM, Nuti MP, Schilperoort RA, Rörsch A. Transfer of the
785 *Agrobacterium tumefaciens* Ti plasmid to avirulent agrobacteria and to *Rhizobium ex planta*. J
786 Gen Microbiol. 1977;98:477–84.

787 50. Haryono M, Tsai Y-M, Lin C-T, Huang F-C, Ye Y-C, Deng W-L, et al. Presence of an
788 *Agrobacterium*-type tumor-inducing plasmid in *Neorhizobium* sp. NCHU2750 and the link to
789 phytopathogenicity. Genome Biol Evol. 2018;10:3188–95.

790 51. Rathore DS, Mullins E. Alternative non-*Agrobacterium* based methods for plant
791 transformation. In: Roberts JA, editor. Annu Plant Rev Online [Internet]. Hoboken, New Jersey,
792 USA: John Wiley & Sons, Ltd.; 2018 [cited 2020 Dec 16]. p. 891–908. Available from:
793 https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119312994.apr0659

794 52. Barco RA, Garrity GM, Scott JJ, Amend JP, Nealson KH, Emerson D. A genus definition for
795 Bacteria and Archaea based on a standard genome relatedness index. mBio. American Society
796 for Microbiology; 2020;11:e02475-19.

797 53. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, et al. A
798 standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of
799 life. Nat Biotechnol. Nature Publishing Group; 2018;36:996–1004.

800 54. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. A complete
801 domain-to-species taxonomy for Bacteria and Archaea. Nat Biotechnol. Nature Publishing
802 Group; 2020;38:1079–86.

803 55. Kuo C-H, Ochman H. Inferring clocks when lacking rocks: the variable rates of molecular
804 evolution in bacteria. Biol Direct. 2009;4:35.

805 56. Okasha S. Evolution and the Levels of Selection [Internet]. Oxford University Press; 2006
806 [cited 2021 Jan 25]. Available from:
807 https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199267972.001.000
808 1/acprof-9780199267972

809 57. Daubin V, Moran NA, Ochman H. Phylogenetics and the cohesion of bacterial genomes.
810 Science. 2003;301:829–32.

811 58. Choi I-G, Kim S-H. Global extent of horizontal gene transfer. Proc Natl Acad Sci.
812 2007;104:4489–94.

813 59. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial
814 innovation. Nature. 2000;405:299–304.

815 60. Dagan T, Artzy-Randrup Y, Martin W. Modular networks and cumulative impact of lateral
816 transfer in prokaryote genome evolution. Proc Natl Acad Sci. 2008;105:10039–44.

817 61. Chan CX, Beiko RG, Darling AE, Ragan MA. Lateral transfer of genes and gene fragments
818 in prokaryotes. Genome Biol Evol. 2009;1:429–38.

819 62. Pál C, Papp B, Lercher MJ. Adaptive evolution of bacterial metabolic networks by horizontal
820 gene transfer. Nat Genet. 2005;37:1372–5.

821 63. Kuo C-H, Ochman H. The fate of new bacterial genes. FEMS Microbiol Rev. 2009;33:38–43.

822 64. Wiedenbeck J, Cohan FM. Origins of bacterial diversity through horizontal genetic transfer
823 and adaptation to new ecological niches. FEMS Microbiol Rev. 2011;35:957–76.

824 65. Mira A, Ochman H, Moran NA. Deletional bias and the evolution of bacterial genomes.
825 Trends Genet. 2001;17:589–96.

826 66. Kuo C-H, Ochman H. Deletional bias across the three domains of life. Genome Biol Evol.
827 2009;1:145–52.

828 67. Sundin GW. Genomic insights into the contribution of phytopathogenic bacterial plasmids to
829 the evolutionary history of their hosts. Annu Rev Phytopathol. 2007;45:129–51.

830 68. Bennett PM. Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic
831 resistance genes in bacteria. Br J Pharmacol. 2009;153:S347–57.

832 69. Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EPC, Cruz F de la. Mobility of plasmids.
833 Microbiol Mol Biol Rev. 2010;74:434–52.

834 70. Redondo-Salvo S, Fernández-López R, Ruiz R, Vielva L, de Toro M, Rocha EPC, et al.
835 Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids.
836 Nat Commun. 2020;11:3602.

837 71. Treangen TJ, Rocha EPC. Horizontal transfer, not duplication, drives the expansion of
838 protein families in prokaryotes. PLOS Genet. 2011;7:e1001284.

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

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

843 74. Zerbino DR, Birney E. Velvet: algorithms for *de novo* short read assembly using de Bruijn
844 graphs. Genome Res. 2008;18:821–9.

845 75. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved
846 genomic sequence with rearrangements. Genome Res. 2004;14:1394–403.

847 76. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, et al. NCBI
848 prokaryotic genome annotation pipeline. Nucleic Acids Res. 2016;44:6614–24.

849 77. Lo W-S, Chen L-L, Chung W-C, Gasparich GE, Kuo C-H. Comparative genome analysis of
850 *Spiroplasma melliferum* IPMB4A, a honeybee-associated bacterium. BMC Genomics.
851 2013;14:22.

852 78. Lo W-S, Gasparich GE, Kuo C-H. Convergent evolution among ruminant-pathogenic
853 *Mycoplasma* involved extensive gene content changes. Genome Biol Evol. 2018;10:2130–9.

854 79. Cho S-T, Kung H-J, Huang W, Hogenhout SA, Kuo C-H. Species boundaries and molecular
855 markers for the classification of 16SrI phytoplasmas inferred by genome analysis. Front
856 Microbiol. Frontiers; 2020;11:1531.

857 80. Guy L, Roat Kultima J, Andersson SGE. genoPlotR: comparative gene and genome
858 visualization in R. Bioinformatics. 2010;26:2334–5.

859 81. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic
860 genomes. Genome Res. 2003;13:2178–89.

861 82. Popescu A-A, Huber KT, Paradis E. ape 3.0: New tools for distance-based phylogenetics
862 and evolutionary analysis in R. Bioinformatics. 2012;28:1536–7.

863 83. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. New York, NY: Springer-
864 Verlag New York; 2016 [cited 2019 Aug 1]. Available from: https://ggplot2.tidyverse.org

865 84. Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical
866 clustering. Bioinformatics. 2006;22:1540–2.

867 85. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput.
868 Nucleic Acids Res. 2004;32:1792–7.

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

871 87. Marchler-Bauer A, Zheng C, Chitsaz F, Derbyshire MK, Geer LY, Geer RC, et al. CDD:
872 conserved domains and protein three-dimensional structure. Nucleic Acids Res. 2013;41:D348–
873 52.

874 88. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for
875 protein modeling, prediction and analysis. Nat Protoc. 2015;10:845–58.

876 89. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME Suite: tools for
877 motif discovery and searching. Nucleic Acids Res. 2009;37:W202–8.

878 90. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview version 2—a
879 multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009;25:1189–91.

880 91. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator.
881 Genome Res. 2004;14:1188–90.

882 92. Conner AJ, Barrell PJ, Baldwin SJ, Lokerse AS, Cooper PA, Erasmuson AK, et al.
883 Intragenic vectors for gene transfer without foreign DNA. Euphytica. 2007;154:341–53.

884 93. Vladimirov IA, Matveeva TV, Lutova LA. Opine biosynthesis and catabolism genes of
885 *Agrobacterium tumefaciens* and *Agrobacterium rhizogenes*. Russ J Genet. 2015;51:121–9.

886 94. Bhatty M, Laverde Gomez JA, Christie PJ. The expanding bacterial type IV secretion
887 lexicon. Res Microbiol. 2013;164:620–39.

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

890 96. Eichinger V, Nussbaumer T, Platzer A, Jehl M-A, Arnold R, Rattei T. EffectiveDB—updates
891 and novel features for a better annotation of bacterial secreted proteins and Type III, IV, VI
892 secretion systems. Nucleic Acids Res. Oxford Academic; 2016;44:D669–74.

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

896

897

898 **Tables**

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

903

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

909

910

911 **Figure Legends**

912 **Figure 1.** Relationships among representatives of the *Agrobacterium tumefaciens*
913 species complex. The sister species *Agrobacterium larrymoorei* (*A. l.*) is included as the
914 outgroup. (A) Maximum likelihood phylogeny based on a concatenated alignment of
915 2,093 single-copy genes shared by all 35 genomes (635,594 aligned amino acid sites).
916 Bootstrap support values in the range of 60-80% are labeled. Strains with complete
917 genome assemblies are highlighted with an asterisk ('*'). The genomospecies
918 assignments (i.e., G1-G9 and G13-G15) are labeled to the right of strain names. The
919 three *A. tumefaciens* supergroups are indicated by the colored background of the
920 genomospecies assignments. Information to the right of the genomospecies
921 assignments shows the grouping of genomes according to different cutoff values of
922 genome-wide average nucleotide identity (ANI), the presence/absence of type VI
923 secretion system (T6SS)-encoding gene cluster (green: present; white: absent), copy
924 number of *vgrG* (white background: absent), number of plasmids, and the tumor-
925 inducing plasmid (pTi) type based on k-mer profile (white background: absent). (B)
926 Pairwise genome similarities based on the percentages of genomic segments mapped
927 and the ANI values.

928

929 **Figure 2.** Phylogeny and organization of the T6SS gene cluster. The maximum
930 likelihood phylogeny was inferred based on a concatenated alignment (5,960 aligned
931 amino acid sites) of 14 core T6SS genes, including *tagE*, *tagF*, *tssM*, *tssL*, *tssK*, *tagH*,
932 *tssG*, *tssF*, *tssE*, *tagJ*, *tssC40*, *tssC41*, *tssB*, and *tssD*. Two other core genes, *tssA* and
933 *tssH*, are excluded because some homologs are pseudogenized. Genes downstream of

934 *tssD* (e.g., *tai*, *tae*, and *vgrG*) are excluded due to variable presence. The species
935 phylogeny on the right is based on **Figure 1**. Genes are color-coded according to
936 annotation, syntenic regions are indicated by gray blocks.
937
938 **Figure 3.** Domain organization and classification of *vgrG* homologs. Six major types
939 (i.e., A-F) and nine subtypes were identified and labeled on the right. For each homolog,
940 the genomospecies assignment is provided in square brackets, followed by the locus
941 tag. The gene names are provided in parenthesis for those functionally characterized
942 homologs (i.e., *vgrG1-2* for C58 homologs and *vgrGa-d* for 1D1609 homologs).
943
944 **Figure 4.** Gene neighborhoods of *vgrG* homologs. The grouping and labeling of *vgrG*
945 homologs are based on the convention used in **Figure 3**. For each *vgrG* homolog, three
946 upstream genes are plotted to illustrate if it is associated with the main T6SS gene
947 cluster or not, and 10 downstream genes are plotted to illustrate putative
948 effector/immunity genes. Two A1-type homologs from the strain S56 (i.e.,
949 AGR1B_RS27940 and AGR1B_RS27960) are in close association with each other and
950 plotted together.
951
952 **Figure 5.** Phylogenetic distribution of *vgrG* homologs and T6SS effector/immunity
953 genes. The species tree is based on **Figure 1**. Gene presence is illustrated with colored
954 cells in the heatmap, gene copy numbers are labeled when applicable.
955

956 **Figure 6.** Molecular phylogeny and global alignment of pTi. The maximum likelihood
957 phylogeny was inferred based on the concatenated alignment of 21 core genes and
958 8,534 aligned amino acid sites (**Supplementary Figure S6A**). The species phylogeny
959 on the right is based on **Figure 1**. The pTi sequences derived from this study are
960 highlighted in bold. For those with relevant information available, the genomospecies
961 assignments are indicated in square brackets and the opine types are indicated in
962 parentheses. For the alignment, all plasmids are visualized in linear form starting from
963 the replication genes. Other gene clusters are color-coded according to functions.
964 Syntenic regions are indicated by gray blocks.
965
966 **Figure 7.** Distribution of key *vir* genes among the putative pTi. The pTi sequences
967 derived from this study are highlighted in bold. For those with relevant information
968 available, the genomospecies assignments are indicated in square brackets and the
969 opine types are indicated in parentheses. Gene presence and absence are indicated by
970 filled and empty circles, respectively. The main *vir* genes include the components of the
971 type IV secretion system (T4SS; *virB1-B11* and *virD4*). For *virE3*, the three sequence
972 types are listed separately. For *virF* and *virD3*, the sequence types are labeled inside
973 the circles. For 1D1609, the genes *virA* and *virJ* are located on another plasmid and
974 plotted as absent in this figure. The locus tags are provided in **Supplementary Dataset**
975 **S1B**.
976
977 **Figure 8.** Organization of the transfer DNA (T-DNA) on tumor-inducing plasmids (pTi).
978 Two unusual putative pTi sequences (i.e., pTiCFBP4996 and pTiCFBP4573) are

979 excluded, other pTi sequences derived from this study are highlighted in bold. For those
980 with relevant information available, the genomospecies assignments are indicated in
981 square brackets and the opine types are indicated in parentheses. Genes are color-
982 coded according to annotation, syntenic regions are indicated by gray blocks.
983

984 **Supplementary Materials**

985 **Supplementary Figure S1.** Gene content dissimilarity among the *Agrobacterium*
986 genomes. (A) and (B): principal coordinate analysis with and without the outgroup *A.*
987 *larrymoorei*, respectively. The % variance explained by each axis is provided in
988 parentheses. (C) and (D): hierarchical clustering with and without the outgroup *A.*
989 *larrymoorei*, respectively.

990

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

993

994 **Supplementary Figure S3.** Logo plots of the putative protein domains identified among
995 *vgrG* homologs. For each domain, the length and the number of homologs with the
996 domain is labeled. Domain 1 is the only domain with a corresponding database entry
997 (TIGR03361).

998

999 **Supplementary Figure S4.** Predicted structures of VgrG homologs. Regions are
1000 colored according to the scheme used in the domain analysis (**Figure. 3**). The chain D
1001 of PA0091 VgrG1 (PDB identifier: 4MTK) from *Pseudomonas aeruginosa* was selected
1002 as the template. The C-terminal parts that could not be confidently inferred are omitted.
1003 In all cases, the coverage (i.e., percentage of the sequence included in the structure
1004 prediction) are at least 75%, the sequence identity to the template is at least 30% and
1005 the confidence score is 100%.

1006

1007 **Supplementary Figure S5.** Maximum likelihood phylogenies of *vgrG*-associated
1008 domains. (A) Domain 1 (TIGR03361; VI_Rhs_Vgr super family), (B) Domain 2
1009 (unknown function), and (C) Domain 3 (unknown function).
1010
1011 **Supplementary Figure S6.** Principal coordinate analysis of gene content among the
1012 putative pTi analyzed.
1013
1014 **Supplementary Figure S7.** Maximum likelihood phylogeny of pTi based on the
1015 concatenated alignment of shared single-copy genes. (A) All of the 20 pTi sequences
1016 analyzed; 21 core genes and 8,534 aligned amino acid sites. (B) Excluding the two
1017 novel pTi; 40 core genes and 15,473 aligned amino acid sites, all branches received >
1018 80% bootstrap support.
1019
1020 **Supplementary Figure S8.** Tomato tumor assay of strains 12D1, CFBP4996, and
1021 CFBP5473. Mock was inoculated with sterilized water as a negative control and strain
1022 C58 was included as a positive control. Strain 12D1 harbors a plasmid with opine
1023 transporter and catabolism genes but lacks *vir* regulon genes and identifiable T-DNA.
1024 CFBP4996 and CFBP5473 harbor novel types of putative tumor-inducing plasmids
1025 (pTi). (A) Tomato stems at three weeks after inoculation. Scale bar: 0.25 cm. (B) Weight
1026 distribution of five biological replicates (1-cm segments of the stem centered at the
1027 inoculation site). The letters indicate ANOVA results.
1028

1029 **Supplementary Figure S9.** Gene organization of the *vir* regulons on pTi. Syntenic
1030 regions are indicated by grey blocks. The virulence (*vir*) genes are highlighted in red,
1031 the conjugation (*tra*) genes are highlighted in yellow, and other genes are plotted in
1032 white.
1033
1034 **Supplementary Dataset S1.** Detailed information of the analyzed genes. (A) List of
1035 *vgrG*-associated genes. Information including genomic location, RefSeq annotation, and
1036 domain prediction are included. (B) Locus tags of the *vir* regulon genes on pTi. The
1037 *virA*/*J* of 1D1609 are located on another plasmid and are highlighted by "*". The *virB7* of
1038 pTiChry5 and pTiEU6 are unannotated in the GenBank RefSeq records so no locus tag
1039 is available but the gene presence was confirmed by BLASTN searches.
1040

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.

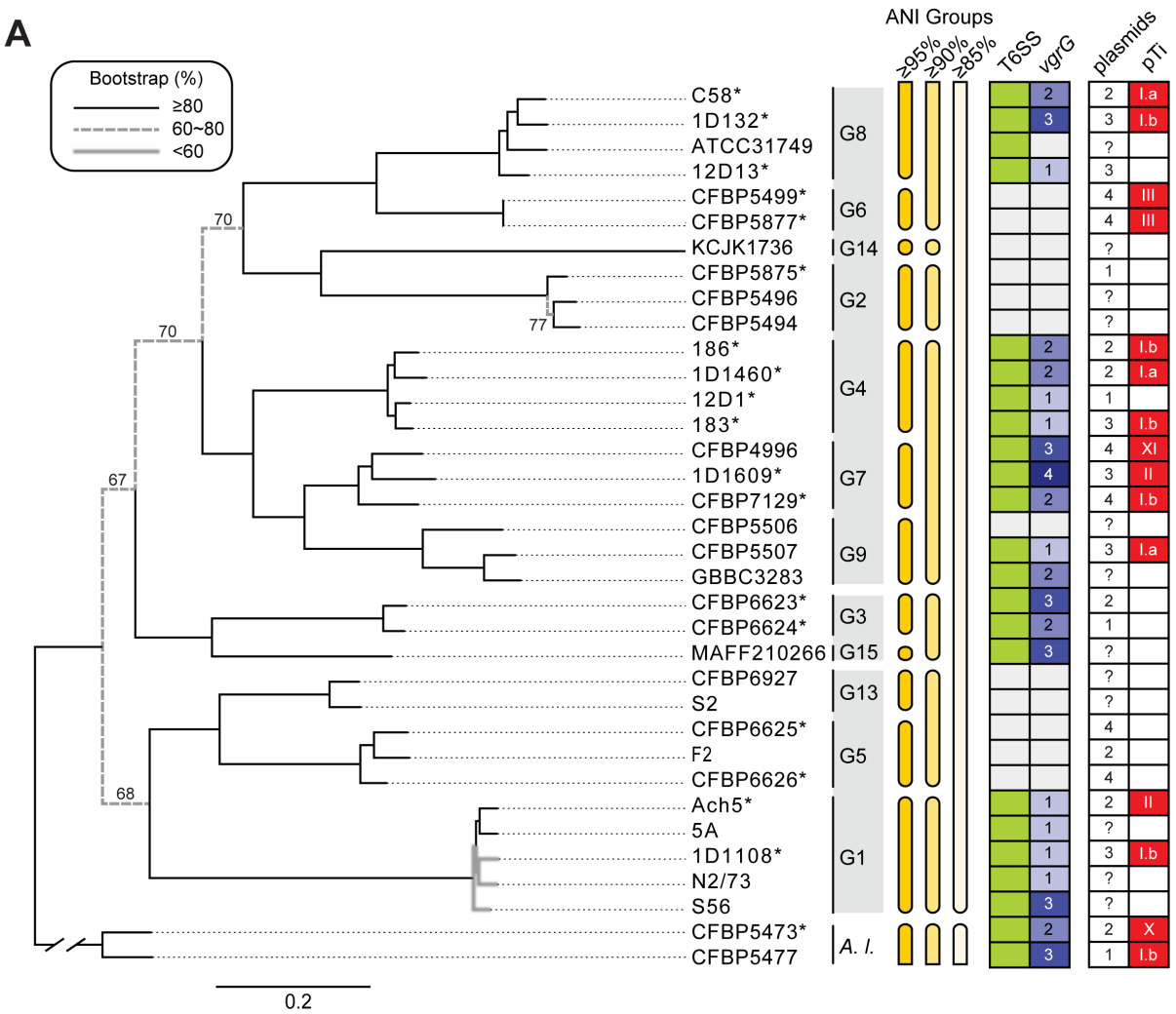
At G7	1D1609	GCF_002943835	Complete	5,539	254	CA, USA	*Medicago sativa*
At G7	CFBP4996	GCF_005144435	10 contigs	5,741	213	UK	*Flacourtia ramontchi*
At G7	CFBP7129	GCF_005221405	Complete	5,927	285	Tunisia	*Pyrus communis*
At G8	12D13	GCF_003667945	Complete	5,083	178	?	?
At G8	1D132	GCF_003667725	Complete	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 G14	KCJK1736	GCF_001641425	41 contigs	4,933	162	FL, USA	*Bos taurus* feces
At G15	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	XI	?	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	X	?	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.

Figures


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
G14
G2
G4
G7
G9
G3
G15
G13
G5
G1
A. I.
70
70
77
67
68
0.2

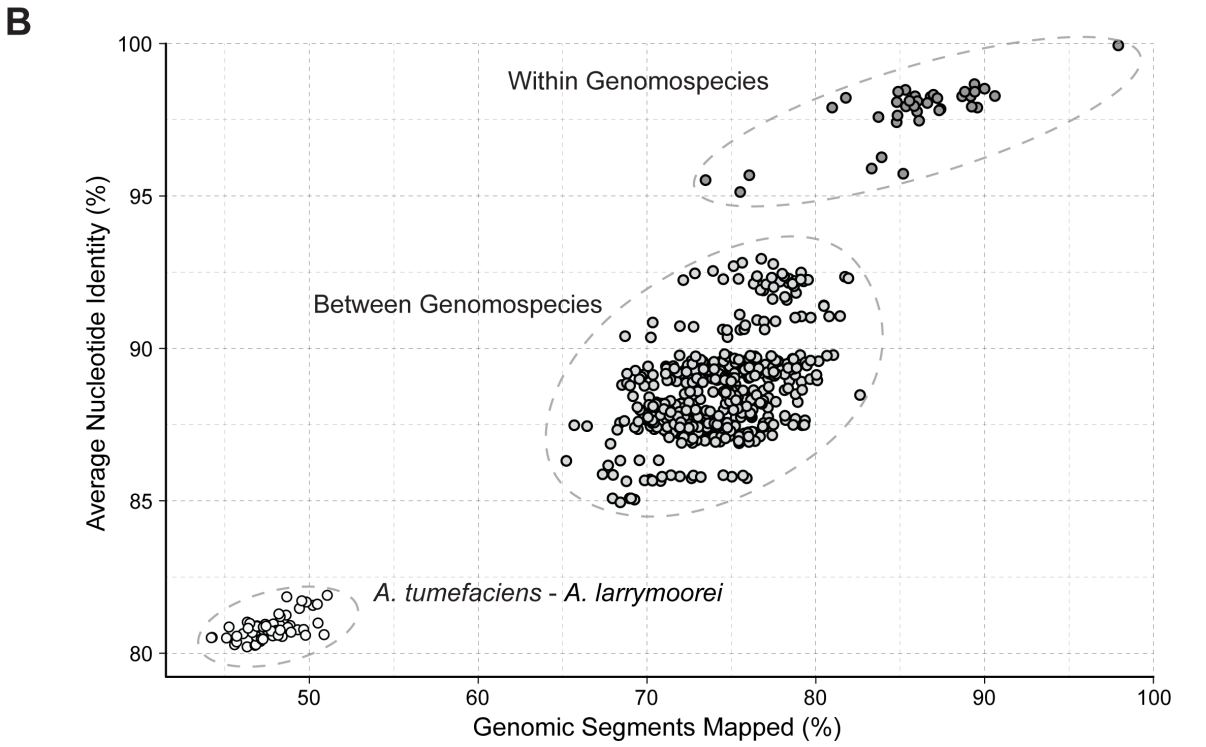

B
Within Genomospecies
Between Genomospecies
A. tumefaciens - A. larrymoorei
Average Nucleotide Identity (%)
Genomic Segments Mapped (%)
100
95
90
85
80
50
60
70
80
90
100

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-G15) 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 tumor-inducing plasmid (pTi) type based on k-mer profile (white background: absent). (B) Pairwise genome similarities based on the percentages of genomic segments mapped and the ANI values.

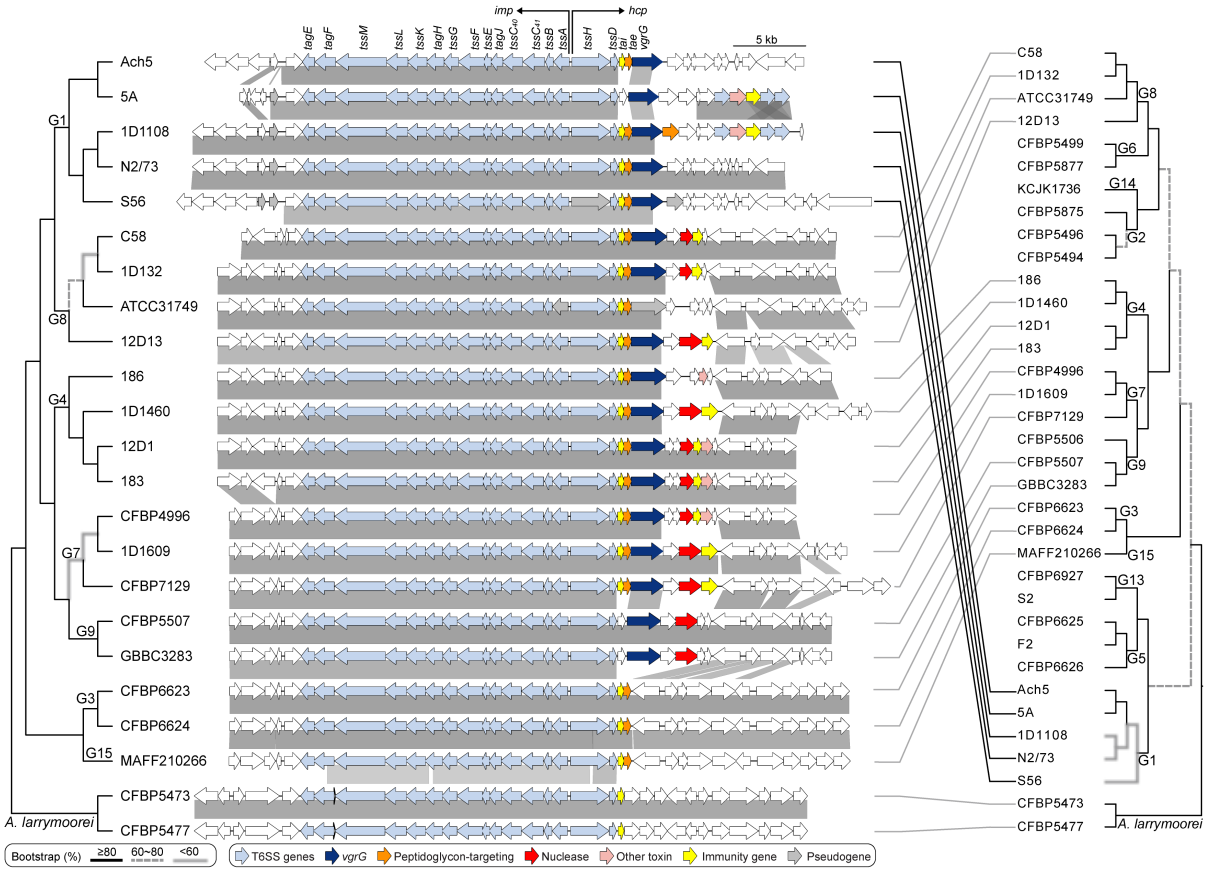


Figure 2. Phylogeny and organization of the T6SS gene cluster. The maximum likelihood phylogeny was inferred based on a concatenated alignment (5,960 aligned amino acid sites) of 14 core T6SS genes, including *tagE*, *tagF*, *tssM*, *tssL*, *tssK*, *tagH*, *tssG*, *tssF*, *tssE*, *tagJ*, *tssC40*, *tssC41*, *tssB*, and *tssD*. Two other core genes, *tssA* and *tssH*, are excluded because some homologs are 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.

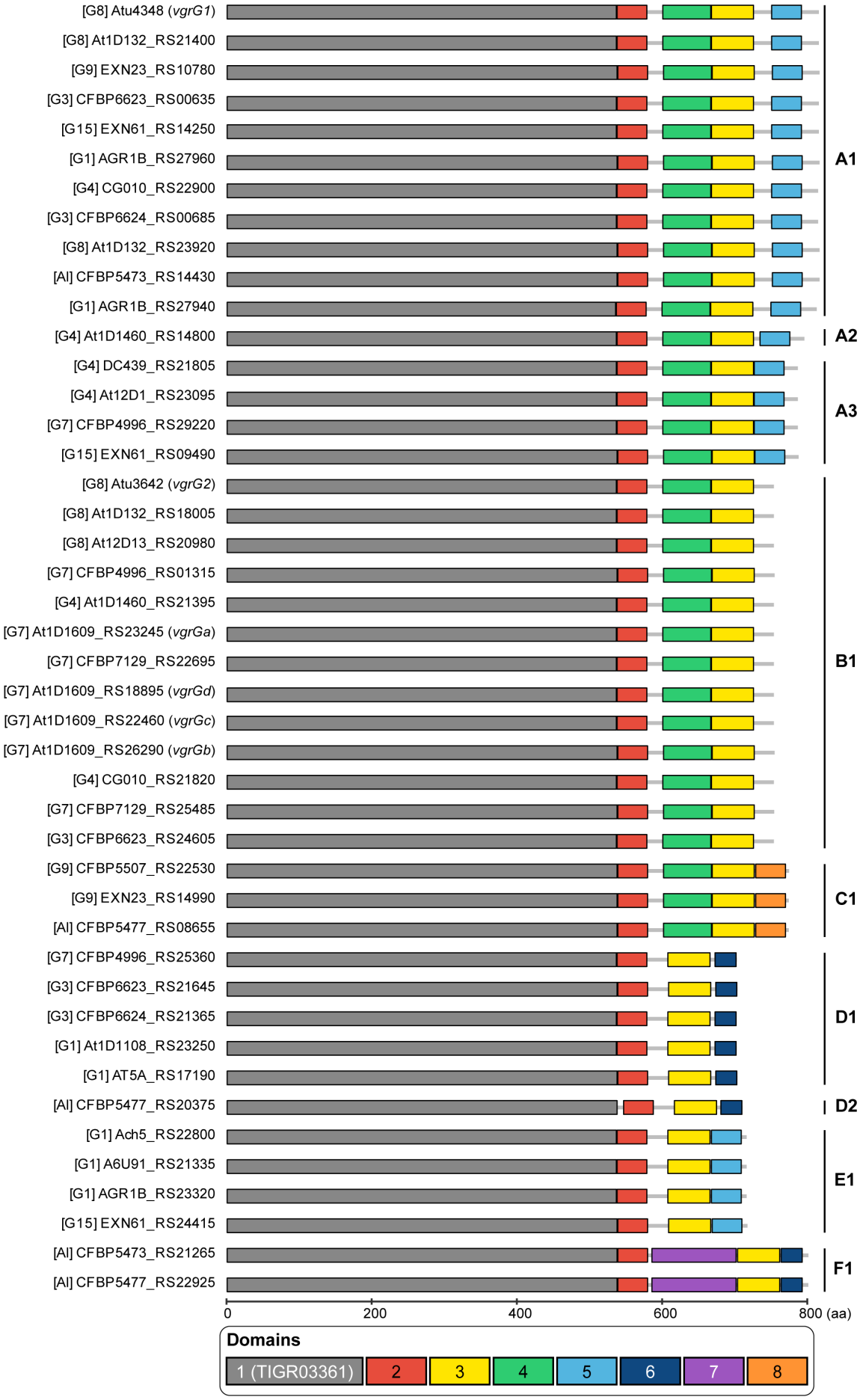

[G8] Atu4348 (vgrG1)
[G8] At1D132_RS21400
[G9] EXN23_RS10780
[G3] CFBP6623_RS00635
[G15] EXN61_RS14250
[G1] AGR1B_RS27960
[G4] CG010_RS22900
[G3] CFBP6624_RS00685
[G8] At1D132_RS23920
[AI] CFBP5473_RS14430
[G1] AGR1B_RS27940
[G4] At1D1460_RS14800
[G4] DC439_RS21805
[G4] At12D1_RS23095
[G7] CFBP4996_RS29220
[G15] 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
[G15] 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).

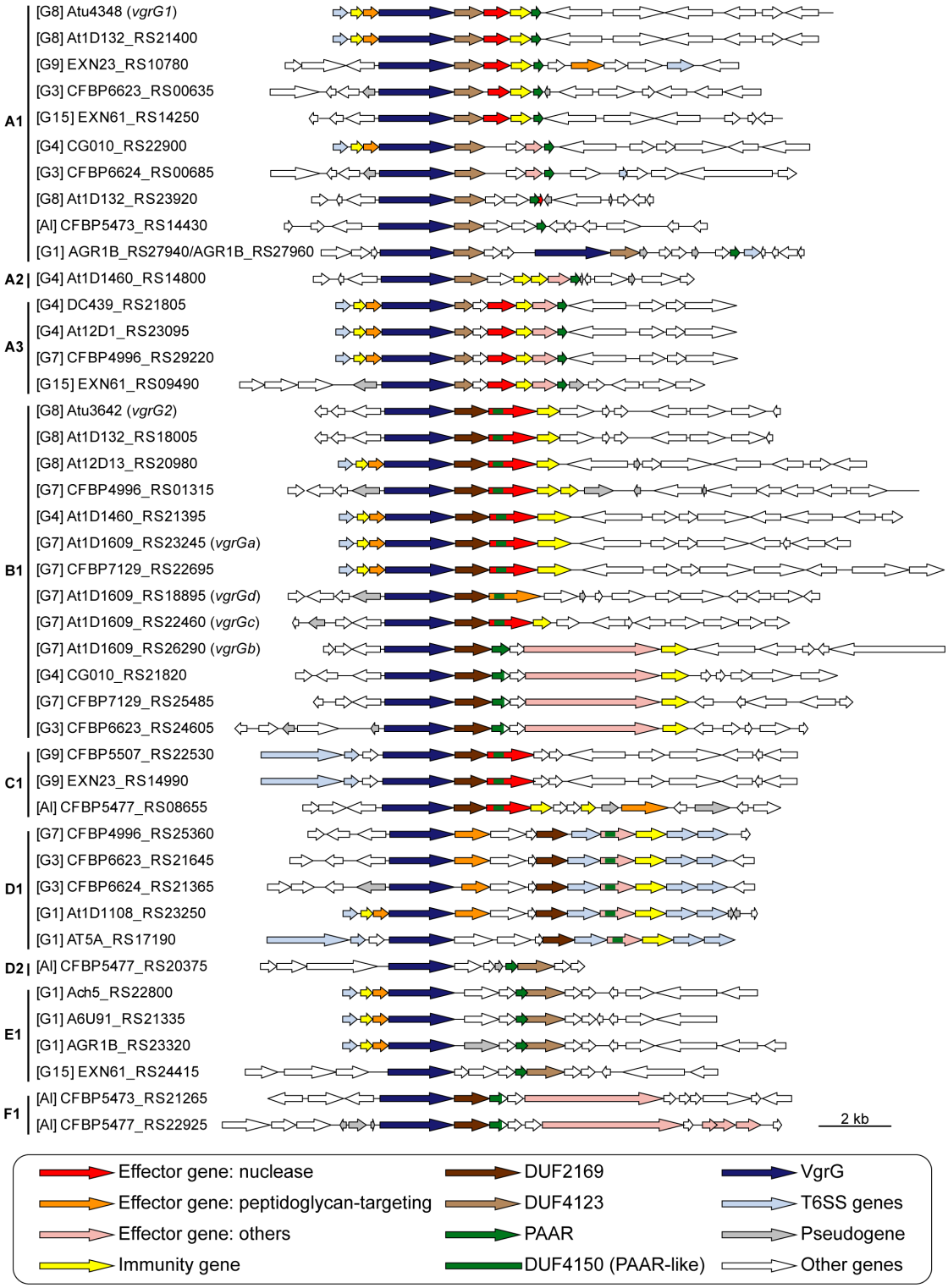


Figure 4. Gene neighborhoods of *vgrG* homologs. The grouping and labeling of *vgrG* homologs are based on the convention used in **Figure 3**. For each *vgrG* homolog, three upstream genes are plotted to illustrate if it is associated with the main T6SS gene cluster or not, and 10 downstream genes are 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.

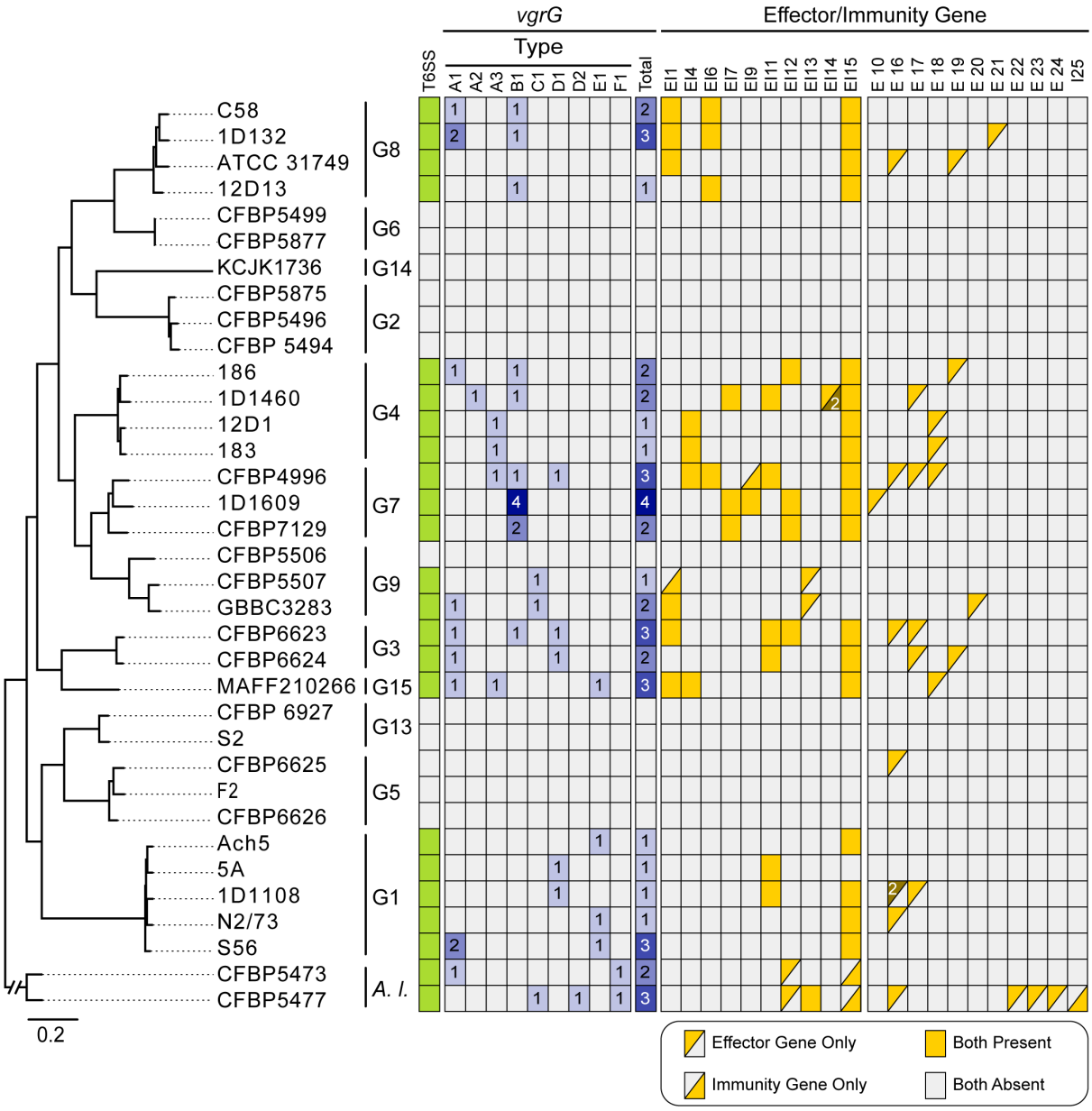


Figure 5. Phylogenetic distribution of *vgrG* homologs and T6SS effector/immunity genes. The species tree is based on **Figure 1**. Gene presence is illustrated with colored cells in the heatmap, gene copy numbers are labeled when applicable.

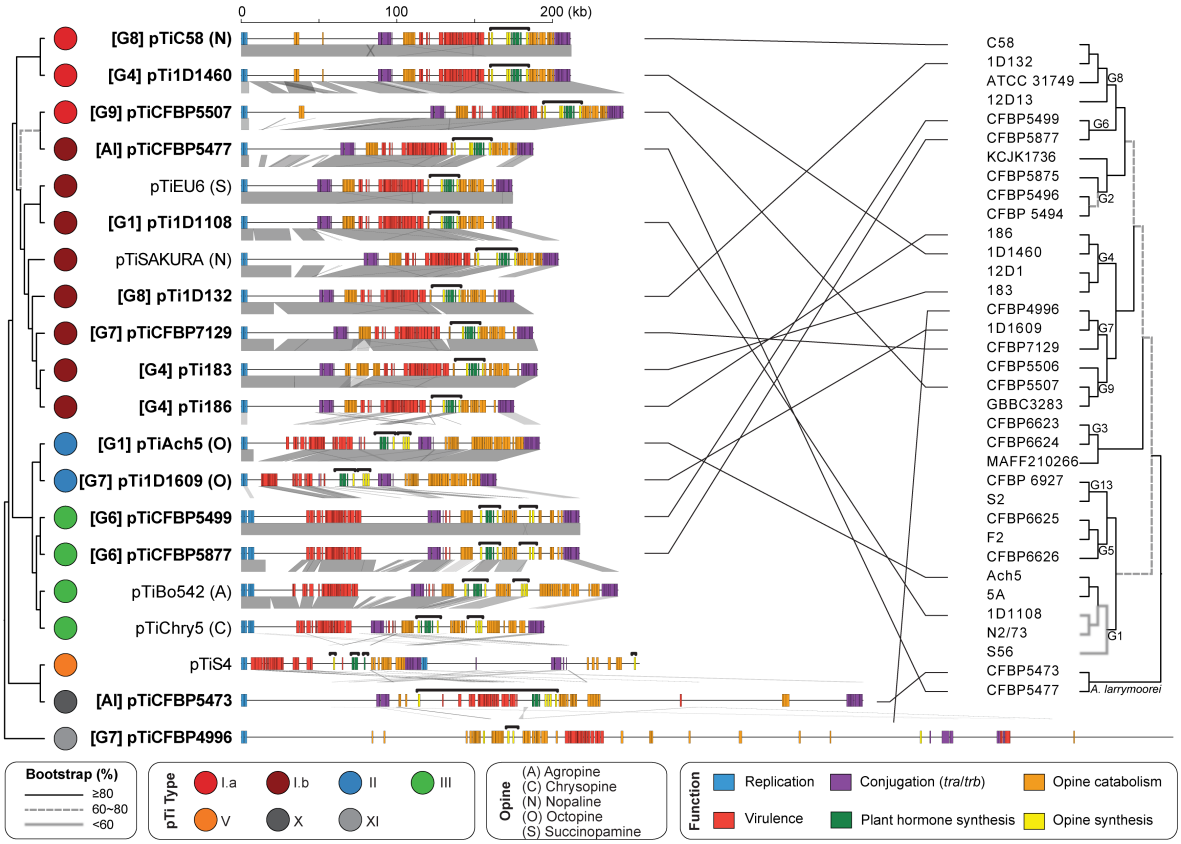


Figure 6. Molecular phylogeny and global alignment of pTi. The maximum likelihood phylogeny was inferred based on the concatenated alignment of 21 core genes and 8,534 aligned amino acid sites (**Supplementary Figure S6A**). The species phylogeny on the right is based on **Figure 1**. The pTi sequences 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.

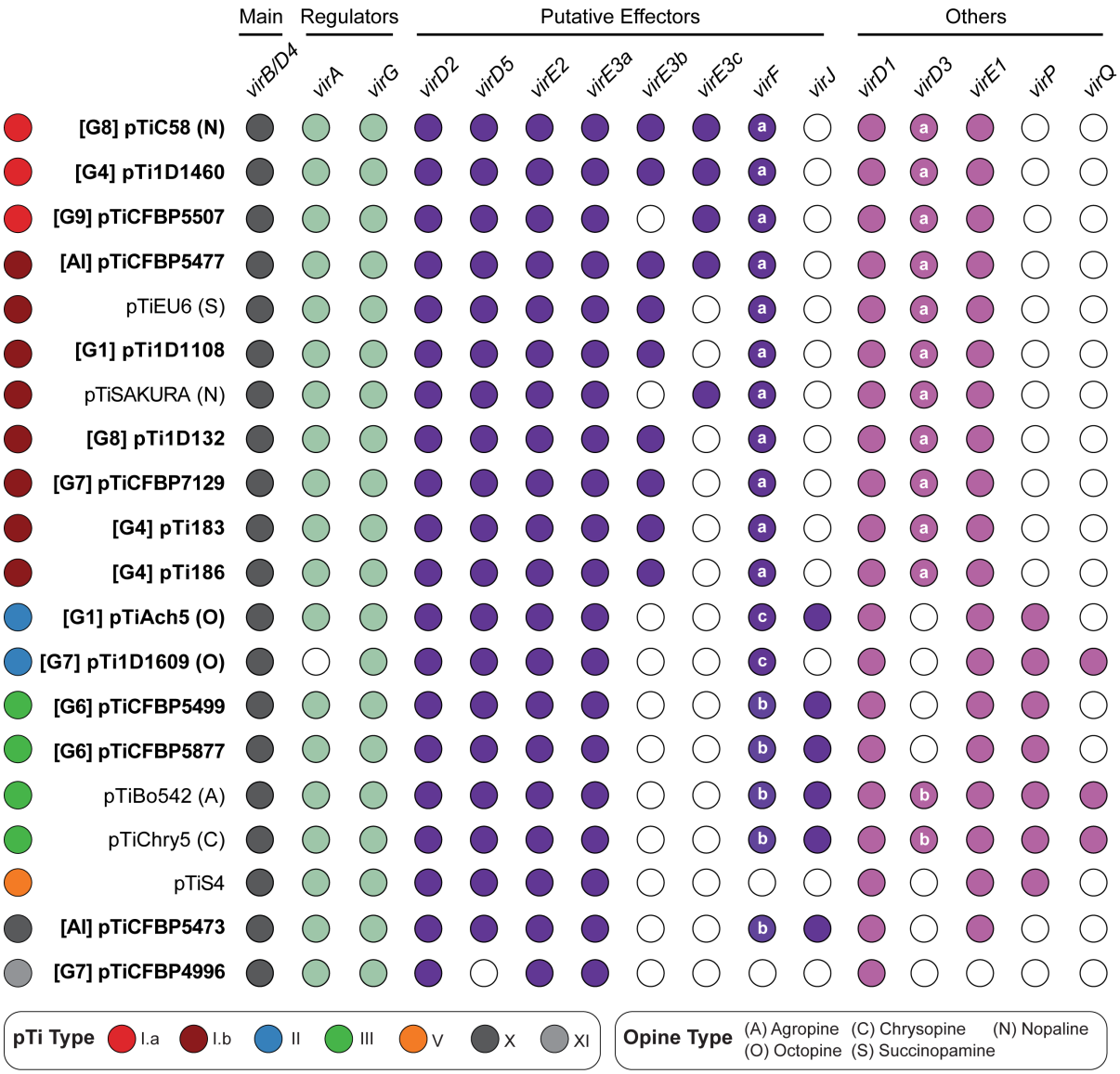


Figure 7. Distribution of key *vir* genes among the putative pTi. The pTi sequences derived from this study are highlighted in bold. For those with relevant information available, the genomospecies assignments are indicated in square brackets and the opine types are indicated in parentheses. Gene presence 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 Dataset S1B**.

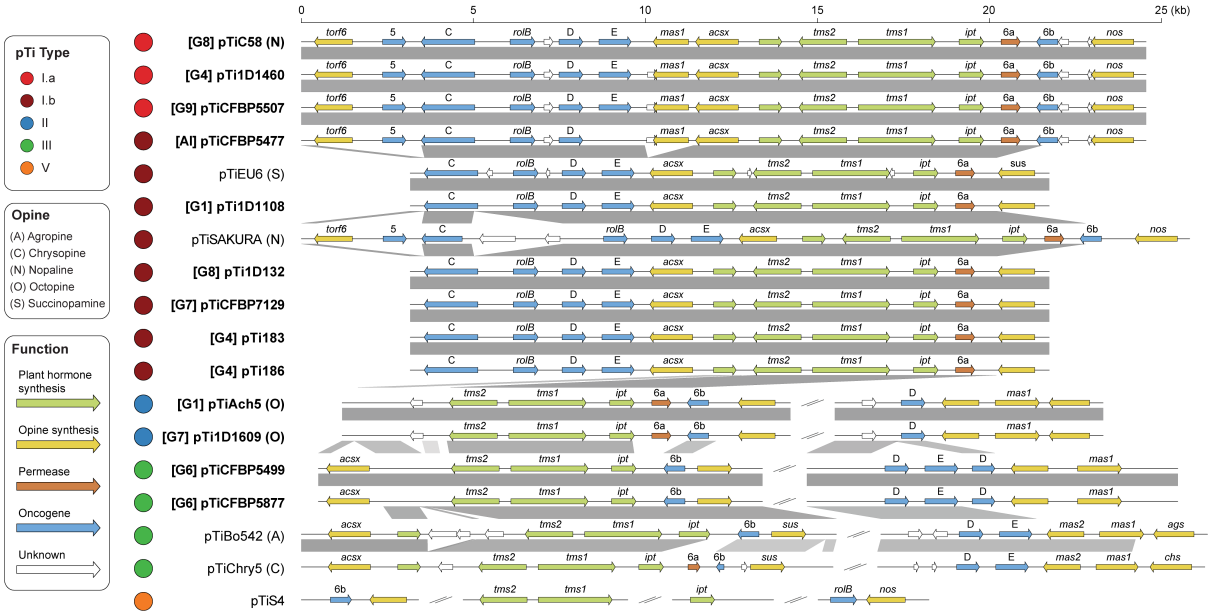
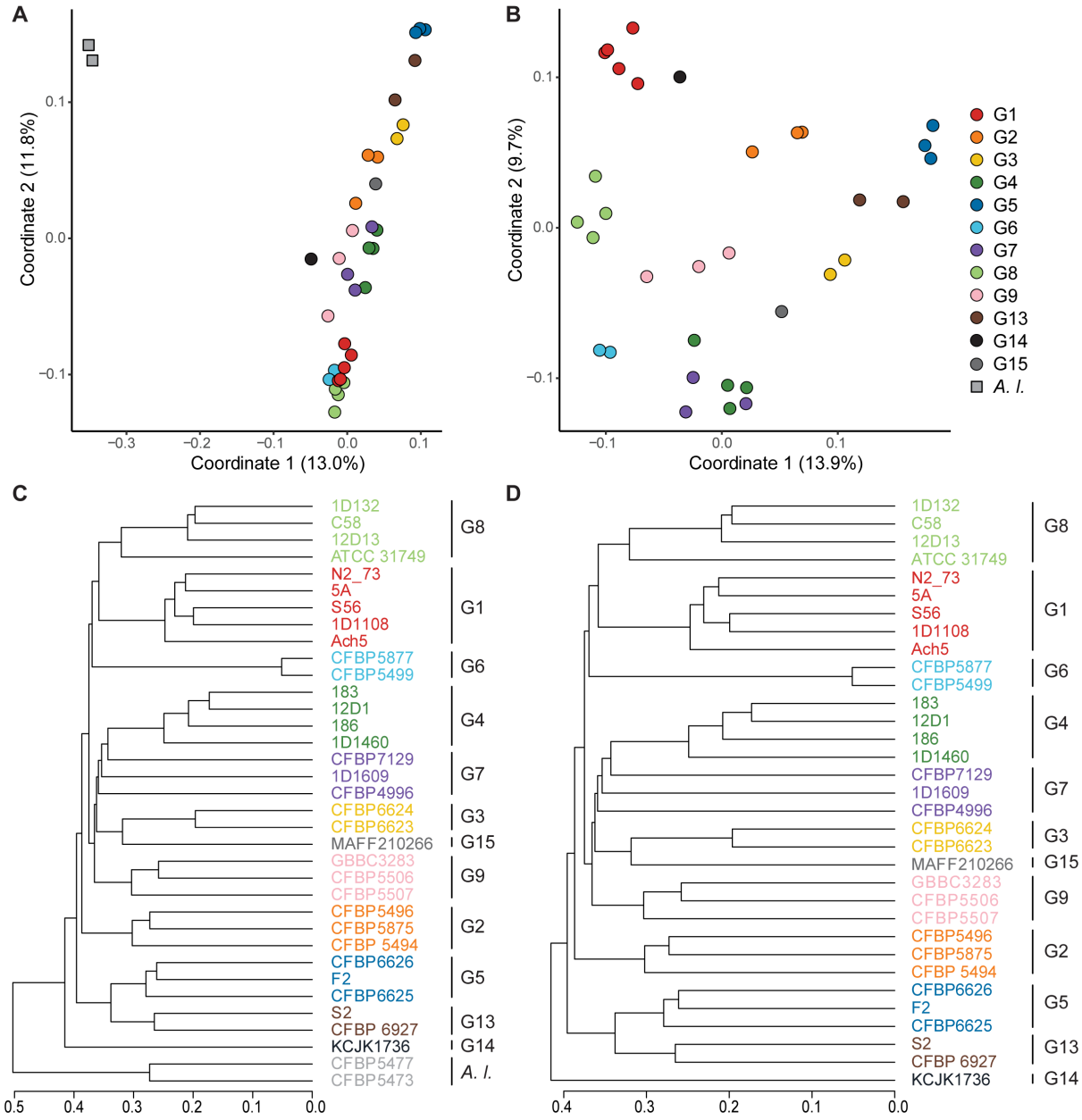
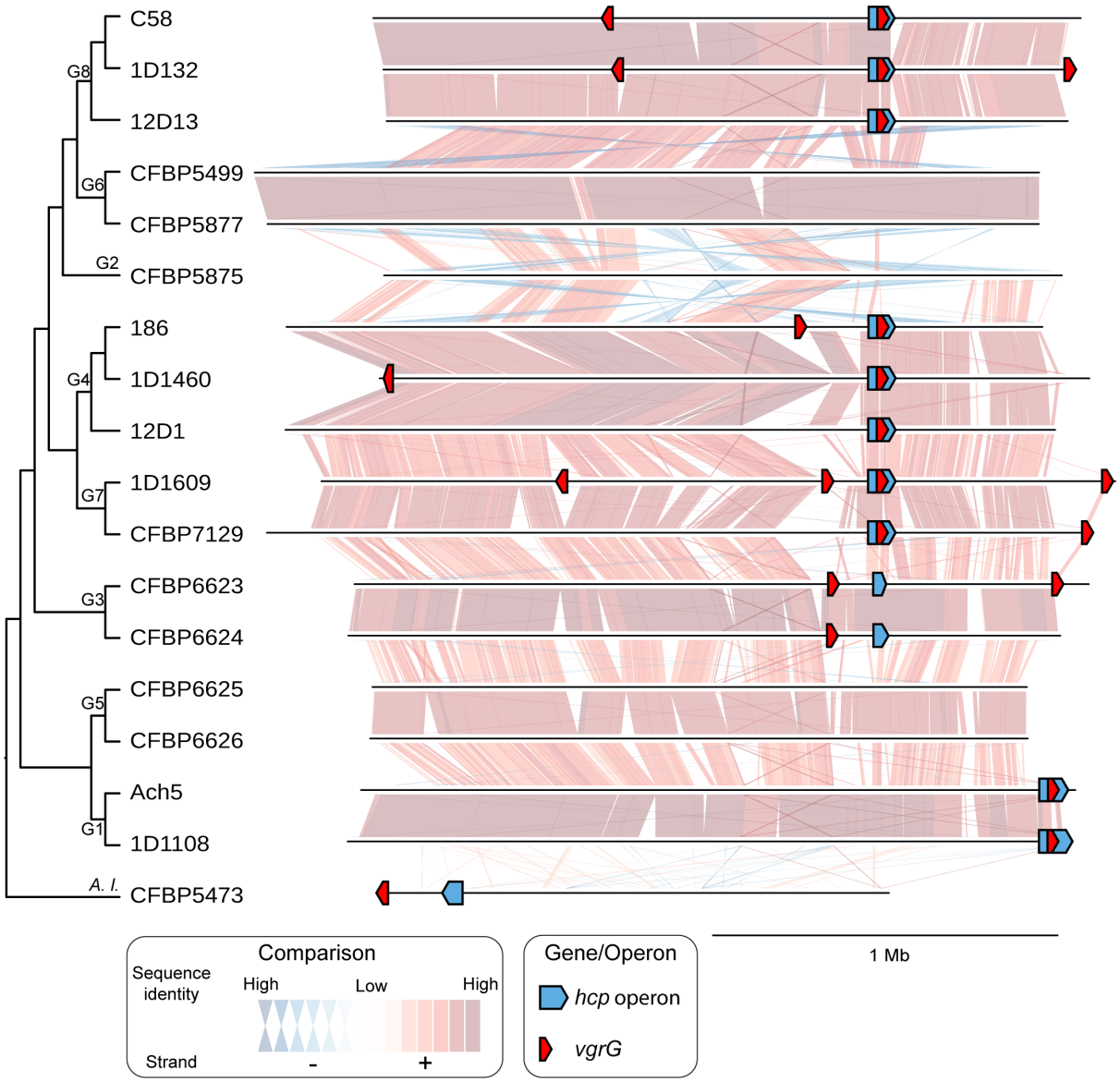


Figure 8. Organization of the transfer DNA (T-DNA) on tumor-inducing plasmids (pTi). Two unusual putative pTi sequences (i.e., pTiCFBP4996 and pTiCFBP4573) are excluded, other pTi sequences derived from this study are highlighted in bold. For those with relevant information available, the genomospecies 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

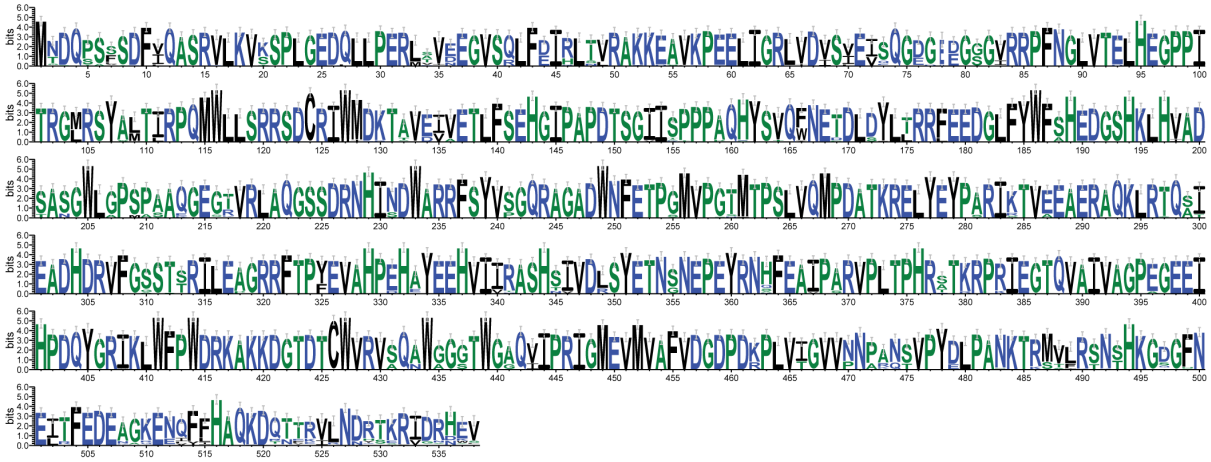


Supplementary Figure S1. Gene content dissimilarity among the *Agrobacterium* genomes. (A) and (B): principal coordinate analysis with and without the outgroup *A. larrymoorei*, respectively. The % variance explained by each axis is provided in parentheses. (C) and (D): hierarchical clustering with and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

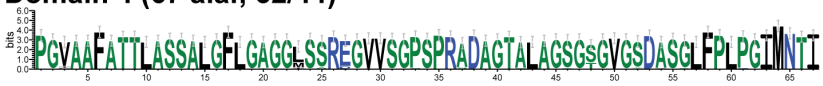
Domain 1 (538 a.a; 44/44; TIGR03361)



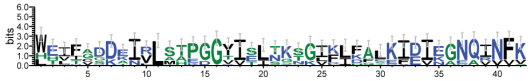
Domain 2 (42 a.a.; 44/44)

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

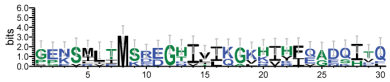
Domain 4 (67 a.a.; 32/44)



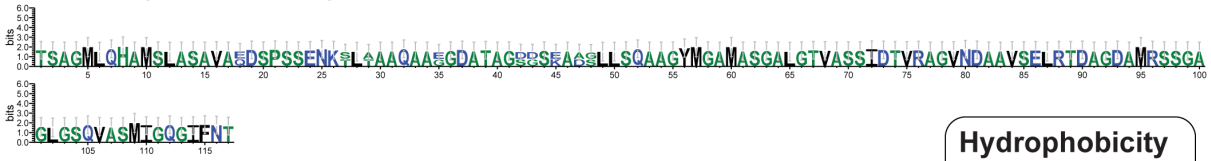
Domain 5 (42 a.a.; 20/44)



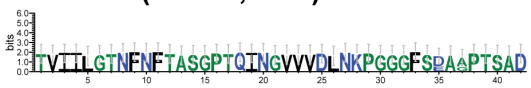
Domain 6 (30 a.a.; 8/44)



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

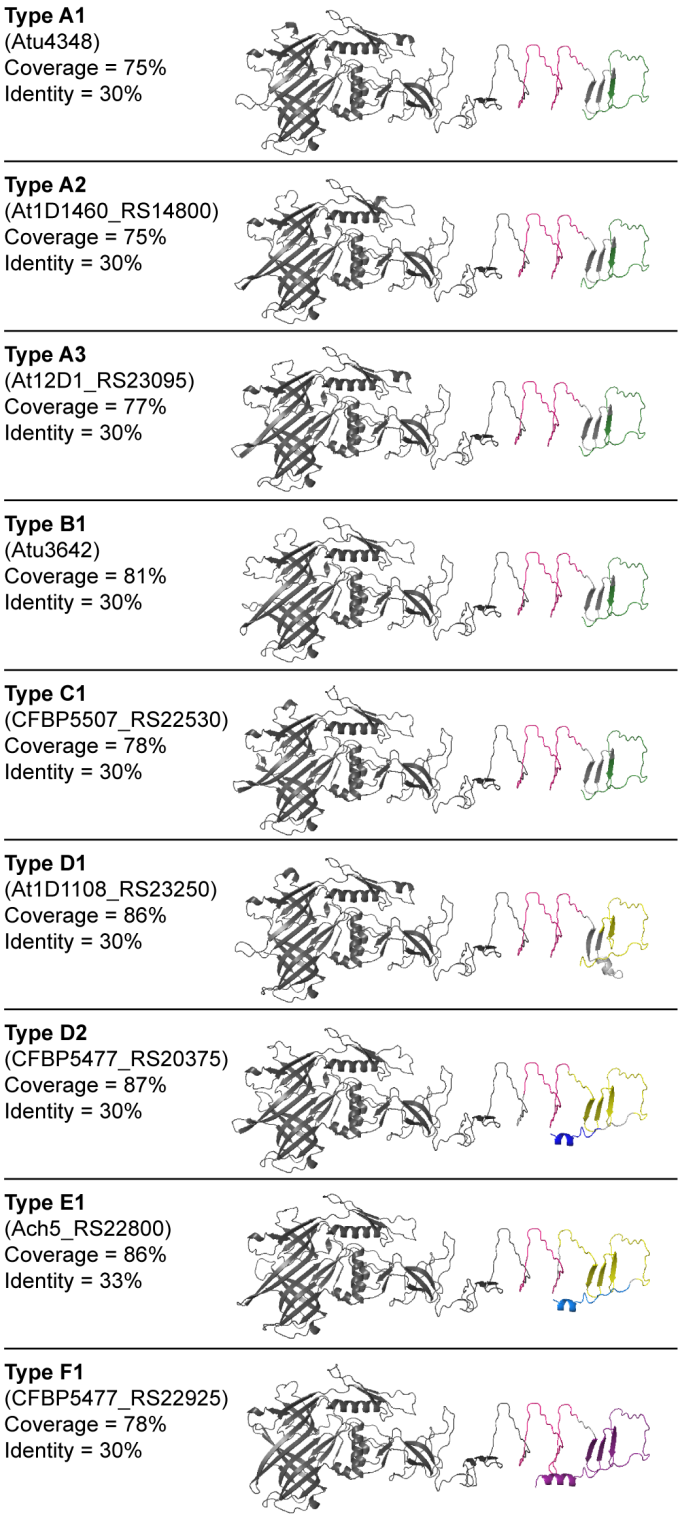


Domain 8 (42 a.a.; 2/44)

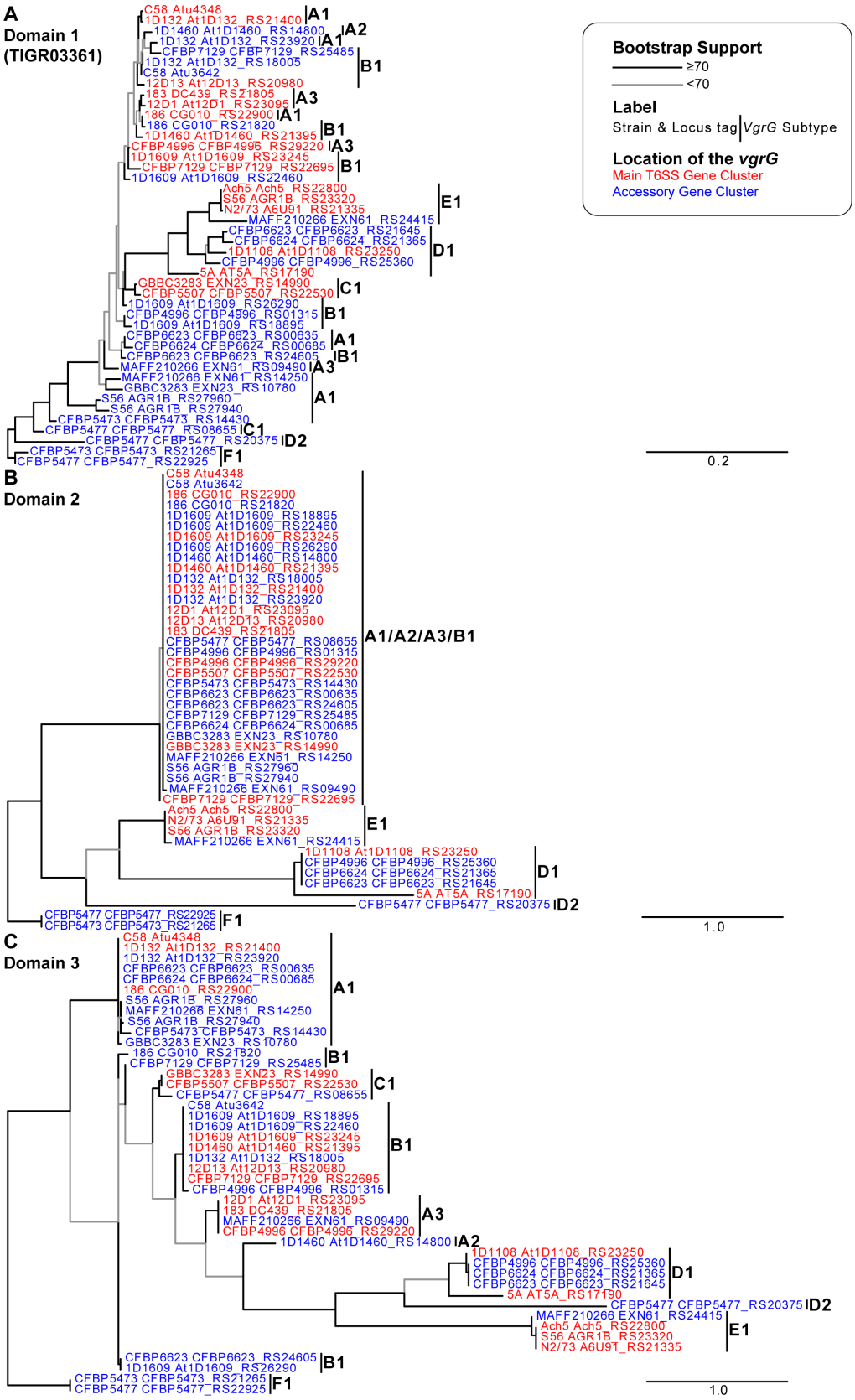


Hydrophobicity
- Hydrophilic
- Neutral
- Hydrophobic

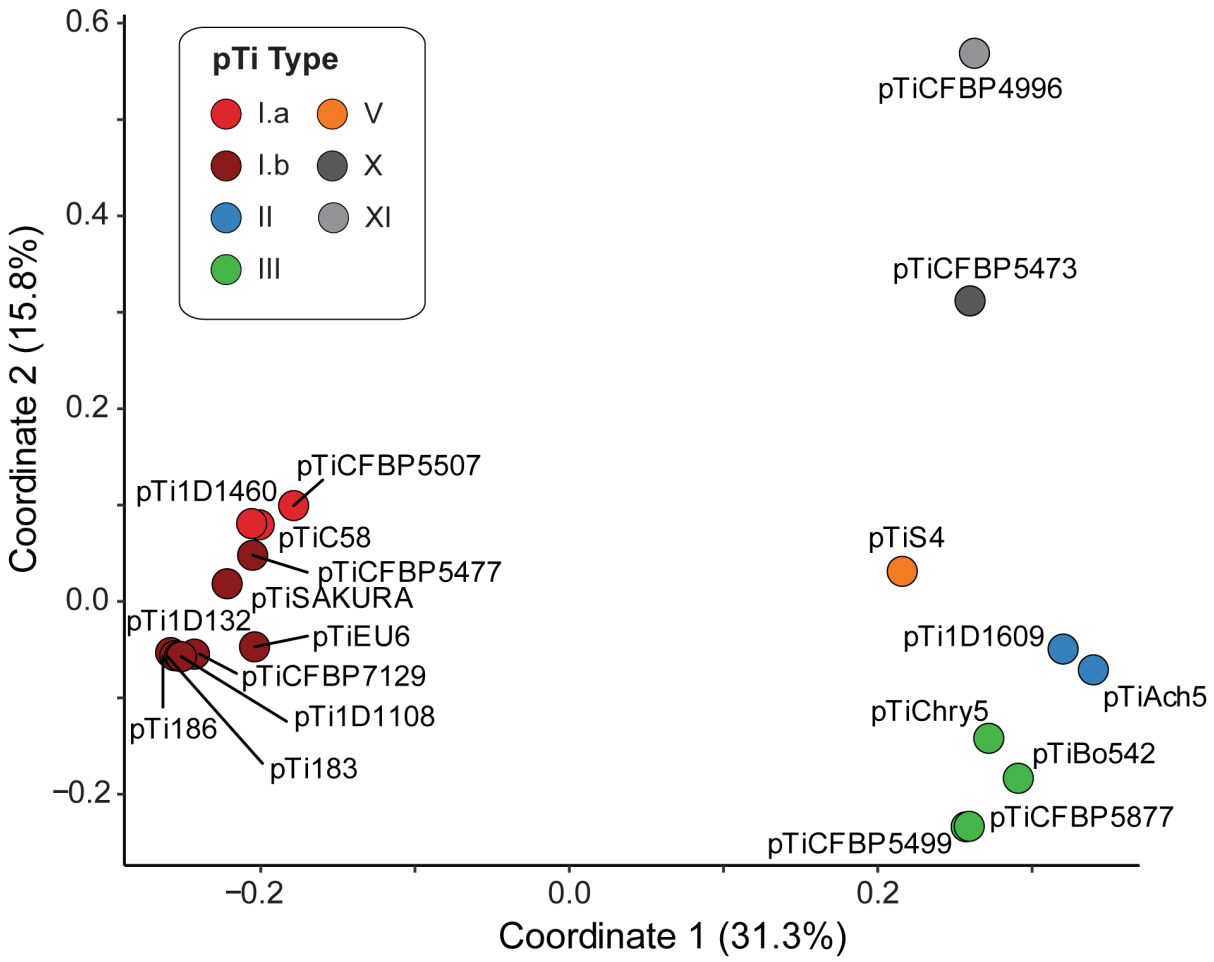
Supplementary Figure S3. Logo plots of the putative protein domains identified among *vgrG* homologs. For each domain, the length and the number of homologs with the domain is labeled. Domain 1 is the only domain with a corresponding database entry (TIGR03361).



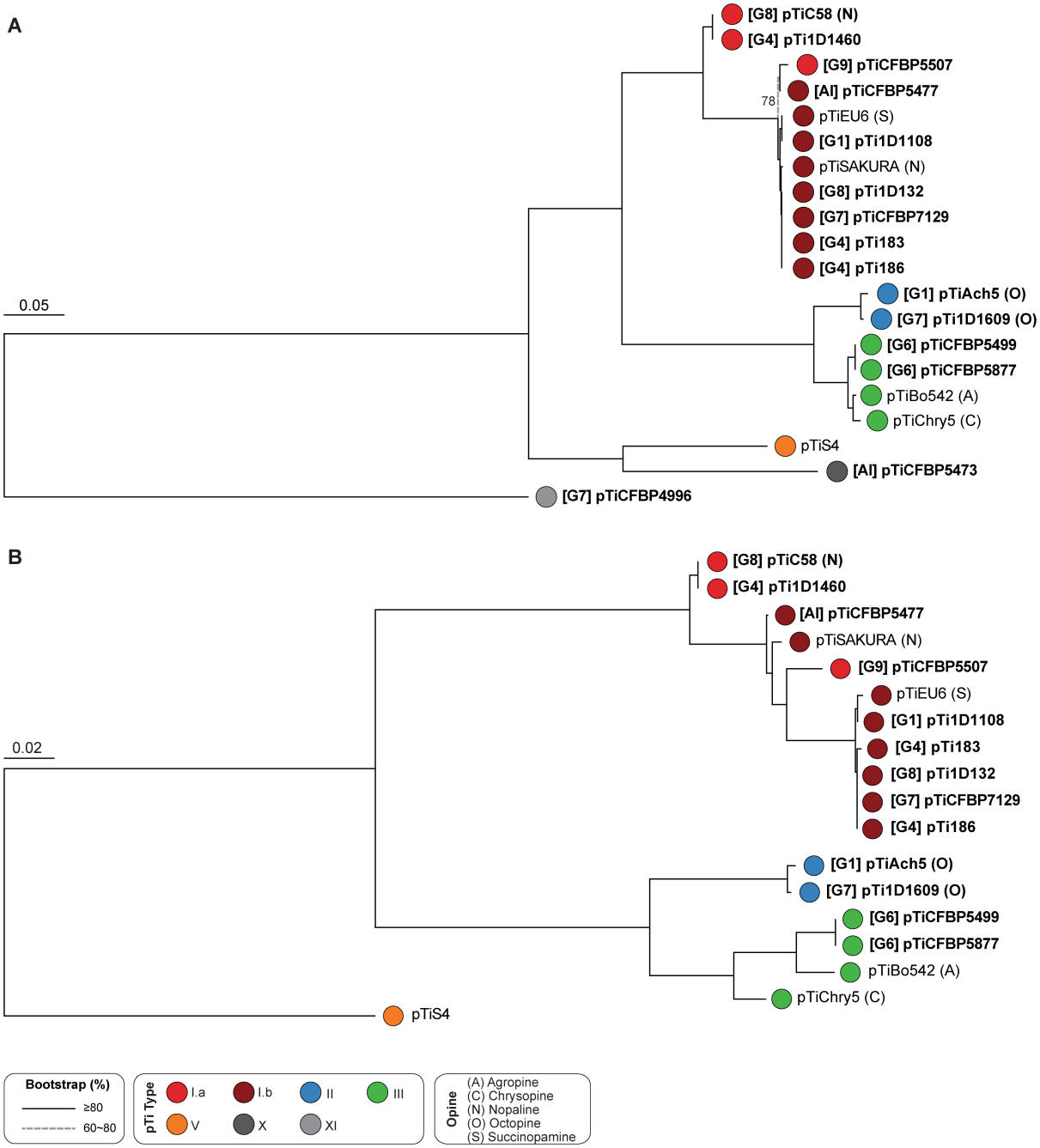
Supplementary Figure S4. Predicted structures of VgrG homologs. Regions are colored according to the scheme used in the domain analysis (**Figure. 3**). The chain D of PA0091 VgrG1 (PDB identifier: 4MTK) from *Pseudomonas aeruginosa* was selected as the template. The C-terminal parts that could not be 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%.



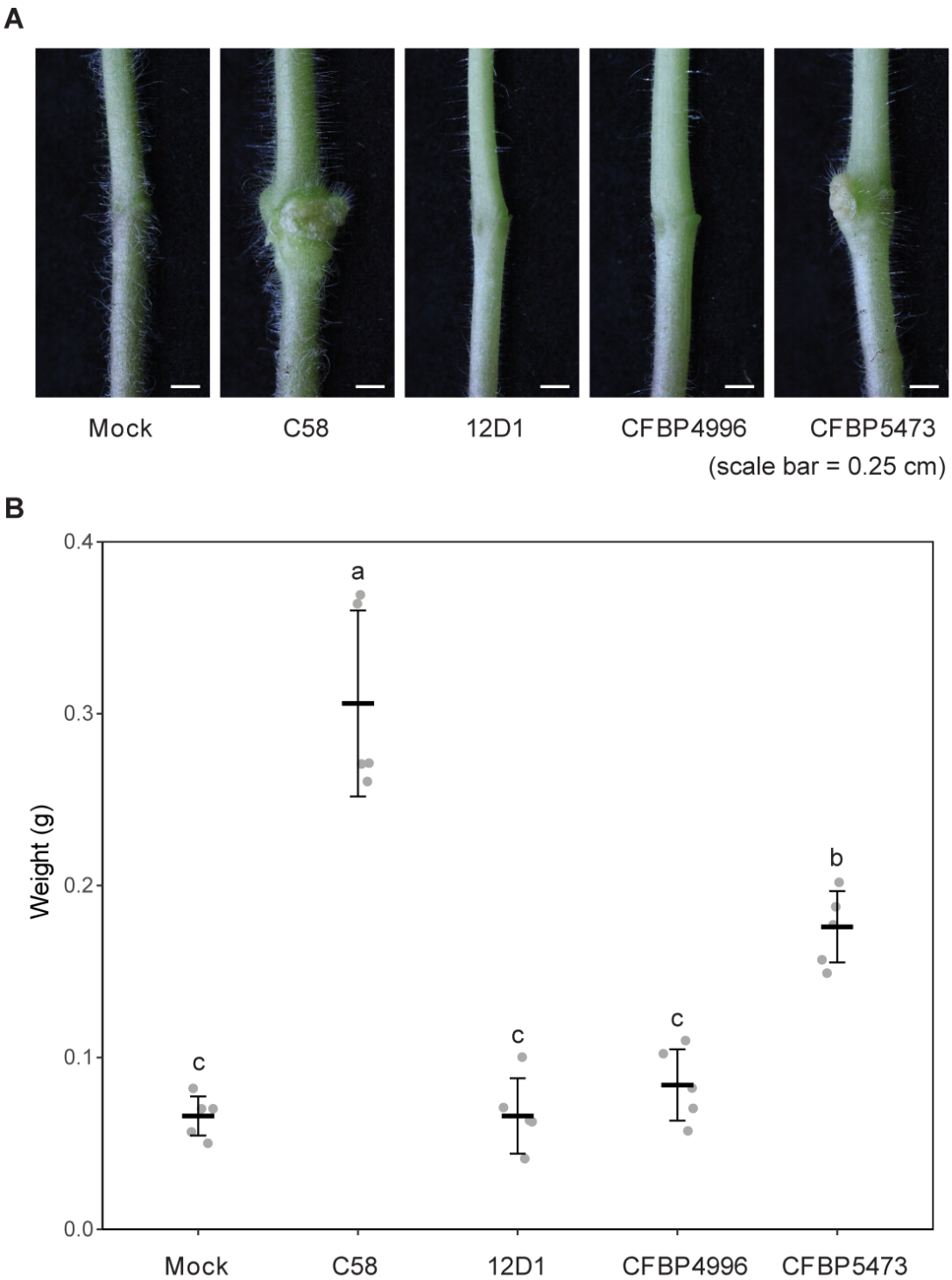
Supplementary Figure S5. Maximum likelihood phylogenies of *vgrG*-associated domains. (A) Domain 1 (TIGR03361; VI_Rhs_Vgr super family), (B) Domain 2 (unknown function), and (C) Domain 3 (unknown function).



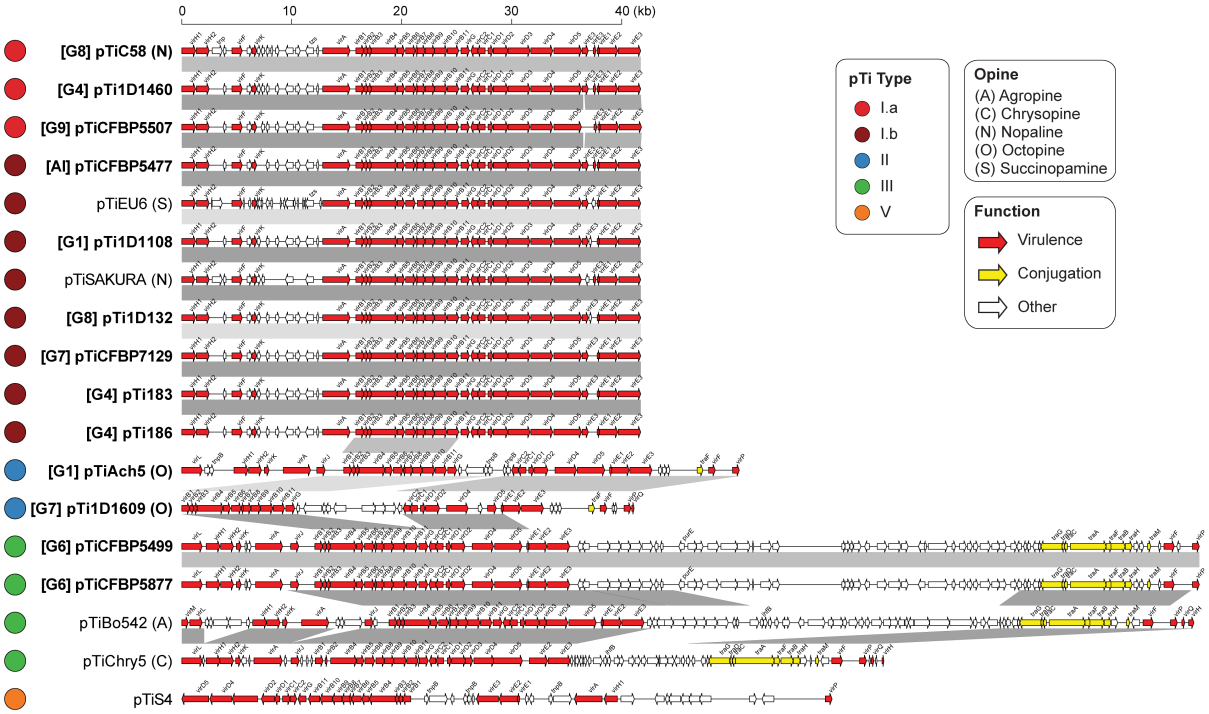
Supplementary Figure S6. Principal coordinate analysis of gene content among the putative pTi analyzed.



Supplementary Figure S7. 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 novel pTi; 40 core genes and 15,473 aligned amino acid sites, all branches received > 80% bootstrap support.



Supplementary Figure S8. Tomato tumor assay of strains 12D1, CFBP4996, and CFBP5473. Mock was inoculated with sterilized water as a negative control and strain C58 was included as a positive control. 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 Figure S9. 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.

Supplementary Dataset S1. Detailed information of the analyzed genes. (A) List of *vgrG*-associated genes. Information including genomic location, RefSeq annotation, and domain prediction are included. (B) Locus tags of the *vir* regulon genes on pTi. 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 was confirmed by BLASTN searches.