Adaptative and ancient co-evolution of integrons with *Xanthomonas* genomes

Elena Colombi[1,2 *], Timothy M. Ghaly[1,2], Vaheesan Rajabal[1,2], Liam D.H. Elbourne[1,2], Michael Gillings[1] & Sasha Tetu[1,2]

[1] School of Natural Sciences, Macquarie University, Sydney, NSW, 2109, Australia
[2] ARC Centre of Excellence in Synthetic Biology, Sydney, NSW, 2109, Australia
*Corresponding author: elena.colombi@mq.edu.au

Keywords: genome evolution, horizontal gene transfer, niche adaptation, plant-pathogen, mobile genetic elements

ABSTRACT

Integrons are genetic elements that facilitate gene acquisition. They have been extensively studied in clinical bacteria, but their evolutionary role in phytopathogens remains underexplored. Here, we analysed complete genomes of *Xanthomonas* species to investigate the origin, distribution, and functional dynamics of integrons in this genus. We found that 93% of genomes harboured integrons. The integron-integrase gene *intI* was predominantly located downstream of *ilvD*, indicating an ancestral acquisition of integrons, predating diversification within the genus. Phylogenetic analyses support vertical inheritance of *intI*, with the exception of rare horizontal gene transfer events, notably in *X. arboricola*. Despite their widespread presence, full-length *intI* genes and active integron platforms are only retained in some species, especially *X. campestris*, which shows high integron gene cassette variability and functional integron activity. In contrast, species such as *X. cissicola* and *X. phaseoli* exhibit widespread *intI* inactivation, likely occurring early in their divergence, leading to more stable cassette arrays and conserved integron-associated phenotypes. The number and diversity of genes within cassette arrays varied significantly by species and, to a lesser extent, by the ecological context of plant host cultivation. While most cassettes encoded proteins without a known function, those with annotated roles were associated with stress response mechanism, competitive exclusion, and plant-associated functions. Together, our findings demonstrate that integrons in *Xanthomonas* likely originated from a single ancient acquisition event, preceding genus-wide speciation, and have co-evolved with *Xanthomonas* pathovars as they adapted to distinct plant hosts.

IMPACT STATEMENT

This study provides the first comprehensive genus-wide analysis of integron evolution and dynamics in *Xanthomonas*, a globally distributed and versatile plant pathogen. Integrons are

genetic platforms that allow horizontal transfer of genes. They are well characterised in human pathogens where they mediate transfer of antibiotic resistance genes, but little is known about their role in plant associated bacteria. We showed that in *Xanthomonas* the integron platform was ancestrally acquired, yet integrons have undergone repeated lineage-specific inactivation events. Despite widespread erosion of integron activity, some species such as *X. campestris* maintain robust and functionally diverse integrons that continue to shape genome plasticity. Notably, high cassette diversity, combined with the presence of rare and often uncharacterized genes within these arrays (some potentially involved in environmental sensing or host interaction) suggest that integrons may serve as reservoirs of adaptive potential. Our findings reshape current views of integron function beyond antibiotic resistance and highlight their long-term role in microbial evolution, niche adaptation, and genome innovation in plant-associated bacteria.

DATA SUMMARY

The authors confirm all supporting data, code and protocols have been provided within the article or through supplementary data files. Accession numbers of the genomes analysed in this manuscript are listed in Table S1.

INTRODUCTION

Integrons are genetic elements that facilitate horizontal gene transfer in Bacteria (Stokes & Hall, 1989), and Archaea (Ghaly *et al.*, 2022). They are ancient structures that have been involved in the evolution of bacterial genomes for hundreds of millions of years (Rowe-Magnus & Mazel, 2001). Functional integrons are composed of an integron-integrase gene (*intI*), an integron recombination site (*attI*), a promoter that drives the expression of the integrase (P_{int}) and generally also have a second promoter (P_c), oriented for expression of gene cassette array proteins (Hall & Collis, 1995, Mazel *et al.*, 1998, Escudero *et al.*, 2018).

Gene cassettes are mobile, non-replicating elements, which generally consist of an open reading frame (ORF) and an *attC* site. They exist as circular molecules when excised from a cassette array. Although cassettes generally consist of a single ORF, cassettes that have two ORFs or ORF-less cassettes with promoter activity have been observed (Michael & Labbate, 2010, Tansirichaiya *et al.*, 2019, Blanco *et al.*, 2024). The tyrosine recombinase IntI can capture circular cassettes and integrate them into the genome by mediating the recombination between *attC* and *attI* sites, forming arrays of 1 to > 200 gene cassettes. As well as mediating the insertion of cassettes, IntI can also excise cassettes from an array. Cassette arrays lacking an integrase are called CALIN (clusters of *attC* sites lacking integron-integrases) (Cury *et al.*, 2016), and can be mobilized *in trans* by a functional IntI.

Technically, integrons themselves are not mobile, whereas their gene cassette cargo is. Integron activity can facilitate horizontal gene cassette transfer between prokaryotes,

contributing to the rapid evolution and genetic diversity of Bacteria and Archaea (Escudero *et al.*, 2018). Because DNA damage and nutrient starvation induce activity of the integron integrase via bacterial SOS and stringent responses (Guerin *et al.*, 2009, Cambray *et al.*, 2011, Strugeon *et al.*, 2016), acquisition of novel genetic functions is preferentially triggered during periods of environmental challenge. This provides a mechanism for adaptive innovation when survival pressures are highest.

Integrons have been extensively studied because of their role in the accumulation and dissemination of antibiotic resistance genes in clinical isolates (Gillings *et al.*, 2008, Hipólito *et al.*, 2023). However, integrons are also common in other environments such as marine environments (Pereira *et al.*, 2020), soil (Ghaly *et al.*, 2019), and the plant rhizoplane (Qi *et al.*, 2024, Rajabal *et al.*, 2024). In these environments, the majority of gene cassettes remain functionally uncharacterised, but predicted functions encode extensive functional diversity including traits involved in microbe-host interactions (Gillings, 2014, Escudero *et al.*, 2018, Ghaly *et al.*, 2020).

Xanthomonas is a genus of Gram-negative bacteria in the class Gammaproteobacteria that contains pathogenic strains able to cause disease in more than 400 different plants. In common with other genera of plant pathogens, *Xanthomonas* also contains strains associated with plants, that do not cause disease symptoms (Vauterin *et al.*, 1996, Essakhi *et al.*, 2015, Garita-Cambronero *et al.*, 2017). *Xanthomonas* is a large genus that encompasses more than 35 species, which are subdivided into pathovars (Timilsina *et al.*, 2020).

Integrons have been described in some *Xanthomonas* isolates (Gillings *et al.*, 2005, Barionovi & Scortichini, 2006, Barionovi & Scortichini, 2008). Using PCR with site-specific primers annealing to the proximal part of *intI* and to the *attI* region, Gillings and colleagues screened 32 *Xanthomonas* strains representing 12 pathovars, and found that all the strains analysed had an integron. In all cases, *intI* was integrated downstream of the acid dehydratase gene, *ilvD*, however, the majority of integrase genes were predicted to be inactivated by frameshifts, stop codons, or large deletions. Groups of strains with the same deletions or stop codons/frameshifts in *intI* usually contained identical arrays of genes. There was no evidence of integrons generating diversity within pathovars but strong evidence for diversity between pathovars. With some minor exceptions, individual pathovars had distinct proximal gene cassette arrays, and every cassette identified was found only in one pathovar.

Complete genome sequences for hundreds of *Xanthomonas* isolates are now available, allowing for a systematic and detailed investigation of integrons in these important plant-associated bacteria. Here, we have examined the presence of integrons in all publicly available complete genomes of *Xanthomonas*, significantly expanding our knowledge of these elements, and showing how integrons have been a dynamic component during *Xanthomonas* genome evolution.

114

115 **METHODS**

116 ***Xanthomonas* classification and phylogeny**

117 Complete genomes of *Xanthomonas* were downloaded from the National Center for
118 Biotechnology Information (NCBI) Database (NCBI txid: 338) (708 genomes, last accessed
119 September 2024) (Table S1). Taxonomic classifications of the genomes was based on the
120 Genome Taxonomy Database (GTDB) (Parks *et al.*, 2022) release 2.2.0 using GTDB-Tk v2.4.0
121 (Chaumeil *et al.*, 2022) (Table S1). The command classify_wf was used with default settings.
122 GTDB-Tk first aligns 120 single-copy phylogenetic marker genes, then classifies each genome
123 based on its placement into domain-specific reference trees (built from 47,899 prokaryote
124 genomes), its relative evolutionary divergence, and average nucleotide identity to reference
125 genomes in the GTDB. Two genomes (GCA_041228075.1 and GCA_002240395.1) did not
126 belong to the *Xanthomonas* genus (*Luteibacter* and *Luteimonas*_D, respectively) and were
127 removed from the dataset. The GTDB-Tk classification system does not strictly adhere to the
128 traditional nomenclature of *Xanthomonas* strains, sometimes reassigning species. We use
129 GTDB-Tk classification in order to use a standardised microbial taxonomy method based on
130 genome phylogeny, and not to attempt to re-classify or rename isolates.

131 Phylogeny of the *Xanthomonas* genus was inferred using on the alignment of the 120
132 marker genes produced with the GTDB-Tk command classify_wf using GTDB-Tk command
133 infer (default setting with --gamma) which uses FastTree 2.1 (Price *et al.*, 2010) to build an
134 approximately-maximum-likelihood phylogenetic tree. The tree was rooted with
135 GCA_041228075.1 (*Luteibacter*) then the tip was removed with the R package ape 5.8-1
136 (Paradis & Schliep, 2019). Single-species phylogeny was built using Realphy (Bertels *et al.*,
137 2014) which maps genomes to a series of reference genomes via bowtie2 2.0.0-beta7
138 (Langmead & Salzberg, 2012). From these, multiple sequence alignments are constructed
139 and used to infer phylogenetic trees via PhyML 3.1 (Langmead & Salzberg, 2012). Trees are
140 visualised using the ggtree 3.16 (Yu *et al.*, 2017) and ggtreeExtra 2.3 (Xu *et al.*, 2021) R
141 packages.

142 **Analysis of integrons**

143 Integrons were identified with IntegronFinder 2.0.5 (Néron *et al.*, 2022) (parameters: --local-
144 max --gbk --promoter-attI) (Table S1). IntegronFinder 2.0.5 was run with default --calin-
145 threshold (default: 2), putative CALINs were identified if they carried at least 2 *attC* sites. To
146 determine the genomic location of the integrons, identified integron sequences were
147 extracted together with 5 kb of flanking regions upstream and downstream (fasta files
148 deposited at https://github.com/EC-MQ/Xanthomonas_integrons) and annotated with
149 Bakta 1.11.3 (Schwengers *et al.*, 2021). Taxonomic origins of integron gene cassette
150 recombination sites (*attC*s) were predicted using the script attC-taxa.sh (Ghaly *et al.*, 2021,

151 Ghaly *et al.*, 2022). The *attC*-taxa pipeline (https://github.com/timghaly/attC-taxa) can
152 detect *attC* sites that are conserved among one of 11 bacterial taxa, comprising:
153 Alteromonadales, Methylococcales, Oceanospirillales, Pseudomonadales, Vibrionales,
154 Xanthomonadales, Acidobacteria, Cyanobacteria, Deltaproteobacteria, Planctomycetes and
155 Spirochaetes.

156 *IntI* sequences identified by Integron Finder were aligned using PRANK (Löytynoja, 2013) run
157 with default settings, the alignment was stripped of gaps with the goalign 'clean sites'
158 command (Lemoine & Gascuel, 2021) and the tree inferred with FastTree 2.1 (Price *et al.*,
159 2010) using the generalized time-reversible model (--gtr).

160 We used multiple approaches to annotate the cassette-encoded proteins. Proteins were
161 functionally annotated with eggNOG-mapper v2 (Huerta-Cepas *et al.*, 2019, Cantalapiedra *et*
162 *al.*, 2021), executed in DIAMOND (Buchfink *et al.*, 2015) mode. Additionally, Foldseek (Van
163 Kempen *et al.*, 2024) was used to create a protein structural database with the ProstT5
164 protein language model (Heinzinger *et al.*, 2024). The database was then used to perform
165 structural alignment against the SwissProt database using a Foldseek search. Foldseek
166 convertalis was used to convert the alignment database in a tab-delimited output, only
167 functional annotations against the Swiss-Prot database with a coverage of 80% and an e-
168 value < 0.001 were considered.

169 To identify genes involved in functions of ecological relevance (trace gas oxidation, carbon
170 cycling, nitrogen cycling, sulphur cycling, phosphorus cycling, iron cycling, plant-microbe
171 interactions, and osmotic stress tolerance) we used the *EcoFoldDB-annotate* pipeline (Ghaly
172 *et al.*, 2025).

173 Plant growth-promoting traits were identified using the Plant Growth-Promoting Traits
174 Prediction tool (PGPT-Pred) (Ashrafi *et al.*, 2022). SignalP V6.0 (Almagro Armenteros *et al.*,
175 2019) was used for signal peptide predictions as a marker for identifying transmembrane or
176 secreted gene products. DefenseFinder (Tesson *et al.*, 2022) was used to detect known anti-
177 phage systems.

178 Unique gene orthologues encoded by the cassettes and their distributions were identified
179 with Proteinortho 6.0.22 (Klemm *et al.*, 2023) run with the -singles option.

180 To detect genes encoding type III secretion effectors (T3SE), we used a database of 66
181 previously characterized 66 *Xanthomonas* T3SE retrieved from the EuroXanth platform
182 (Costa *et al.*, 2024) (Biessy *et al.*, 2025). T3SE in the cassettes were identified with
183 Proteinortho v6.0.22 (Klemm *et al.*, 2023) by querying the curated effectors database
184 against the amino acid sequences of the cassette-encoded genes.

185 Phylogenetic signal in the number of *attC* sites across genomes was first assessed with
186 Pagel's λ tests using the R package caper 1.0.3 (Orme *et al.*, 2013). To evaluate the influence

of cultivation type on *attC* site number while accounting for shared evolutionary history, phylogenetic generalized least squares regression was conducted using the R package nlme 3.1 (Pinheiro *et al.*, 2017).

Horizontal gene transfer events between species

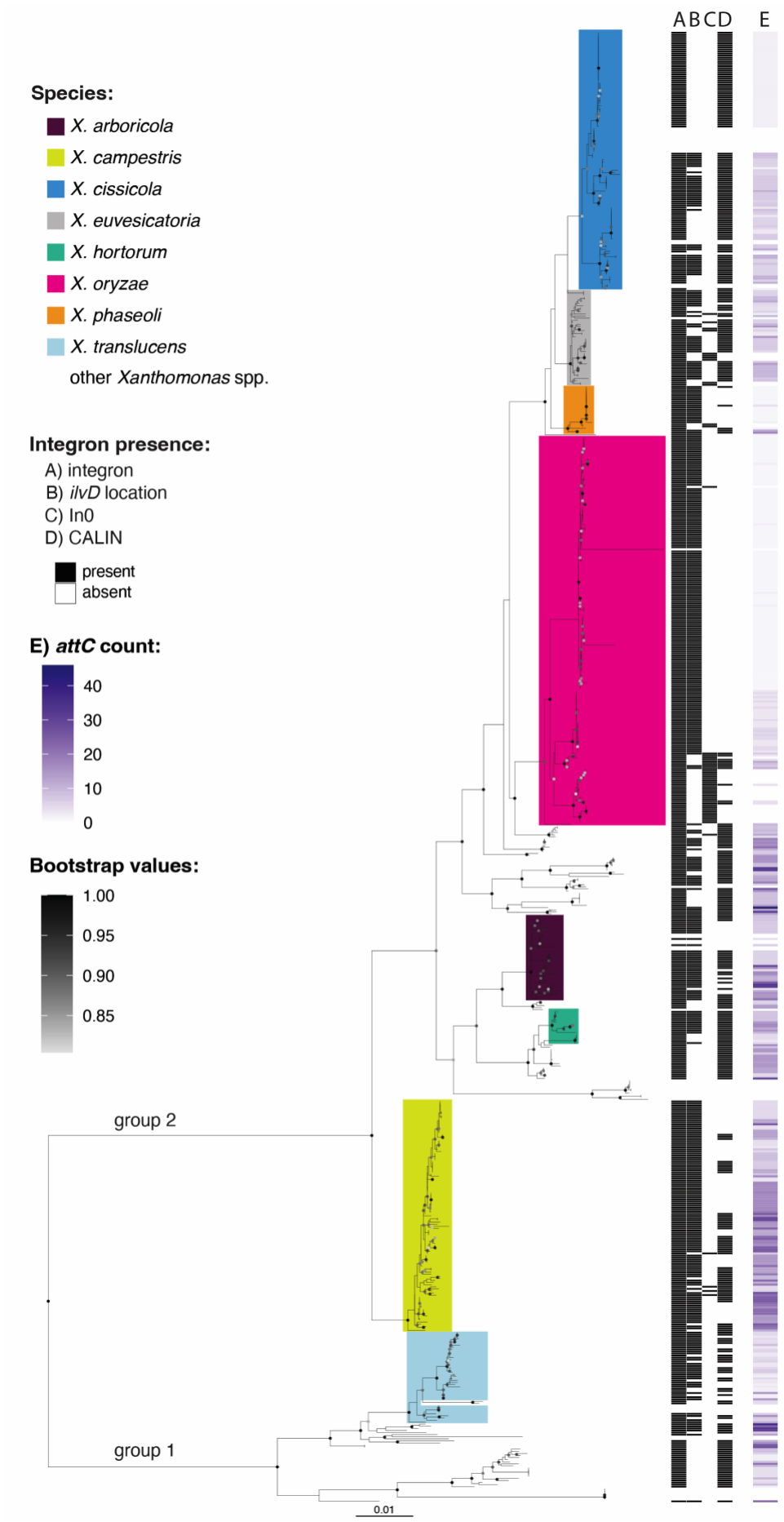
Alfy 1.0.5 (Domazet-Lošo & Haubold, 2011) was used to guide the detection of highly similar cassettes present in the integron arrays. Alfy was run with the -M option, and selecting only matches with a *P* value <0.05 within a sliding window of 100 bp. Only arrays sharing closest homology (recombination), with at least 10% of their array sequence, were considered. The output of this was further filtered, using average nucleotide identity comparisons to assess likelihood that shared cassettes represented probable horizontal gene transfer (HGT) events. FastANI 1.33 (Jain *et al.*, 2018) was run between the arrays and the whole genomes of the strains. We classified an event as HGT when the ANI identity between arrays exceeded 96%, the ANI between the genomes harbouring them was below 95%, and the array ANI was at least 2% higher than the genome ANI. Cytoscape (Shannon *et al.*, 2003) was then used to visualise the numbers of cassettes transferred between *Xanthomonas* species.

RESULTS AND DISCUSSION

The *Xanthomonas* integron platform is ancestral

Xanthomonas complete genomes classified based on the Genome Taxonomy Database (GTDB) as "*Xanthomonas*" (n=629) and "*Xanthomonas_A*" (n=78), known also as *Xanthomonas* group 2 and 1, respectively (Pena *et al.*, 2024) were analysed for integron features (Table S1). We retained the genomes of *Xanthomonas_A,* even though they are classified as a different genus by GTDB-Tk, because from a plant pathology perspective they are still considered *Xanthomonas*, and encompass *X. translucens,* an important causal agent of diseases in cereal crops and forage grasses (Sapkota *et al.*, 2020).

Using IntegronFinder 2.0.5 (Néron *et al.*, 2022), we identified integrons, In0 (integron-integrases that lack cassettes), or CALINs (clusters of *attC* sites lacking integron-integrases) in 93% of the genomes (657 genomes) (Figure 1, Table S1). *X. albilineans* (n=6, group 1) and *X. fragariae* (n=6, group 2) were the only species with more than 2 genome sequences available that did not harbour either an *intI* or *attC* sites. However, integrons have been previously detected in some isolates of *X. fragariae* via PCR amplification (Barionovi & Scortichini, 2006).


Species:
X. arboricola
X. campestris
X. cissicola
X. euvesicatoria
X. hortorum
X. oryzae
X. phaseoli
X. translucens
other Xanthomonas spp.
Integron presence:
A) integron
B) ilvD location
C) In0
D) CALIN
present
absent
E) attC count:
40
30
20
10
0
Bootstrap values:
1.00
0.95
0.90
0.85
group 2
group 1
0.01
A B C D
E

222 **Figure 1 – Distribution of integron platforms in *Xanthomonas* groups 1 and 2.** The maximum likelihood
223 phylogenetic tree is based on a concatenated alignment of 120 top-ranked marker proteins. The tree was
224 rooted at the midpoint. Tips are labelled with species names assigned with Genome Taxonomy Database
225 phylogeny. Coloured clades show species with more than 10 genomes available for analysis. A) Integrons
226 were classified as present if the isolates carried a functional integron, *intI* only or an *attC* array only, B)
227 Indicates the presence of a complete integron or *attC* sites in the *ilvD* locus, C) indicates the presence of
228 an *intI* lacking cassettes (In0), D) indicates the presence of a CALINs (cluster of *attC* sites) not in the *ilvD*
229 locus. E) indicates the number of *attC* sites (number of cassettes) in the genome. Only bootstrap values
230 0.80 are shown. Scale bar indicates substitutions per site.

231

232 Because integrons are widespread in *Xanthomonas*, our first question was whether the
233 acquisition of the integron module was ancestral. The integron-integrase encoding gene *intI*
234 was integrated downstream *ilvD* in all genomes, with the two exceptions being isolates of *X.*
235 *oryzae* and *X. cucurbitae*. Of the 187 *oryzae* isolates examined, *X. oryzae* pv. *oryzicola* (26
236 genomes) harbour a truncated *intI* (561 bp) that was not integrated downstream *ilvD.*
237 However, the integron locus (*intI* and cassette array) is surrounded by transposases, which
238 could indicate that the integron was moved via transposition, or could be a result of
239 assembly errors which are common in regions flanked by insertion sequences. In all seven *X.*
240 *cucurbitae* genomes, *intI* was not integrated downstream *ilvD.* However, downstream of *ilvD*
241 there are two genes without recognizable *attCs* which are encoded in the other integrons
242 (including a DUF1488 encoding gene which occurs in multiple species).

243 The different genomic locations of *intI* could be explained by either within-genome
244 transposition or recombination (*intI* was initially located downstream *ilvD* and then moved),
245 or by an independent acquisition of *intI* from another source which was integrated in a
246 different locus. To test this, we constructed the phylogeny of all the *Xanthomonas intIs*
247 (Figure S1, Table S2). If *intIs* located elsewhere from *ilvD* were acquired independently, they
248 would likely have a different evolutionary history, and form separate clades in the *intI*
249 phylogeny.

250 The phylogeny of *intI* clusters this gene within species, generally being independent of the
251 integration locus. *X. oryzae intIs* cluster together, and the *X. cucurbitae intIs* form a clade
252 with pv. *phaseoli* (Figure S1). The only notable exception is *X. arboricola*. *X. arboricola*
253 strains can carry one of two variant *intIs*. In two strains, *intI* clusters closest to the *X.*
254 *campestris intIs* and are truncated and likely not functional. Phylogenetically close *X.*
255 *arboricola* strains do not harbour an *intI* at all. The other *X. arboricola* strains carry *intI*
256 (either full-length or with an early stop codon) that are more distantly related to all other
257 *intIs*. These more distantly related *intI* could have been acquired from a source outside
258 *Xanthomonas*. We compared the *ilvD-intI* locus sequence in two representative *X. arboricola*
259 strains harbouring the two *intI* variants (GCA_018141705.1 and GCA_905367745.1). While
260 the *ilvD* genes share 97.1% average nucleotide identity (ANI), the *intIs* share 68.2% ANI, this

261 discrepancy in ANI values between two adjacent genes suggests a recombination event
262 (Ansari & Didelot, 2014), with abrupt shifts in ANI at gene boundaries a common indicator of
263 a recombination breakpoint.

264 We used two full length *intI* variants (one belonging to *X. arboricola*, one to *X. campestris*) as
265 a query in a BLASTn search on complete genomes in NCBI excluding *Xanthomonas*
266 (taxid:338). The two closest *intI*s belonged to *Lysobacter* sp. CECT 30171 (locus tag:
267 LYB30171_00805, 93% coverage and 74% identity against *X. campestris intI*) and
268 *[Pseudomonas] boreopolis* strain GO2 (locus tag: M3M27_18655, 92% coverage and 91%
269 identity against *X. arboricola intI*) (both genera in the Xanthomonadaceae family). Curiously,
270 in isolate GO2, *intI* was also located downstream of *ilvD*. We constructed a phylogenetic tree
271 using *Xanthomonas intI* sequences longer than 900 bp and included the *intIs* of CECT 30171
272 and GO2, using *Vibrio* sp. SCSIO 43136 *intI* as an outgroup (Figure 2). The *intI* variants in *X.*
273 *arboricola* were more closely related to the *intI* in GO2 than the other *Xanthomonas intIs*.
274 *ilvD* phylogeny instead clusters together all *X. arboricola* isolates within the *Xanthomonas*
275 group 2, and GO2 as basal both to *Xanthomonas* group 1 and 2 isolates (Figure S2). This
276 suggests that, in those *X*. *arboricola* strains, *intI* was acquired via horizontal gene transfer
277 (HGT) from another member of the Xanthomonadaceae family.

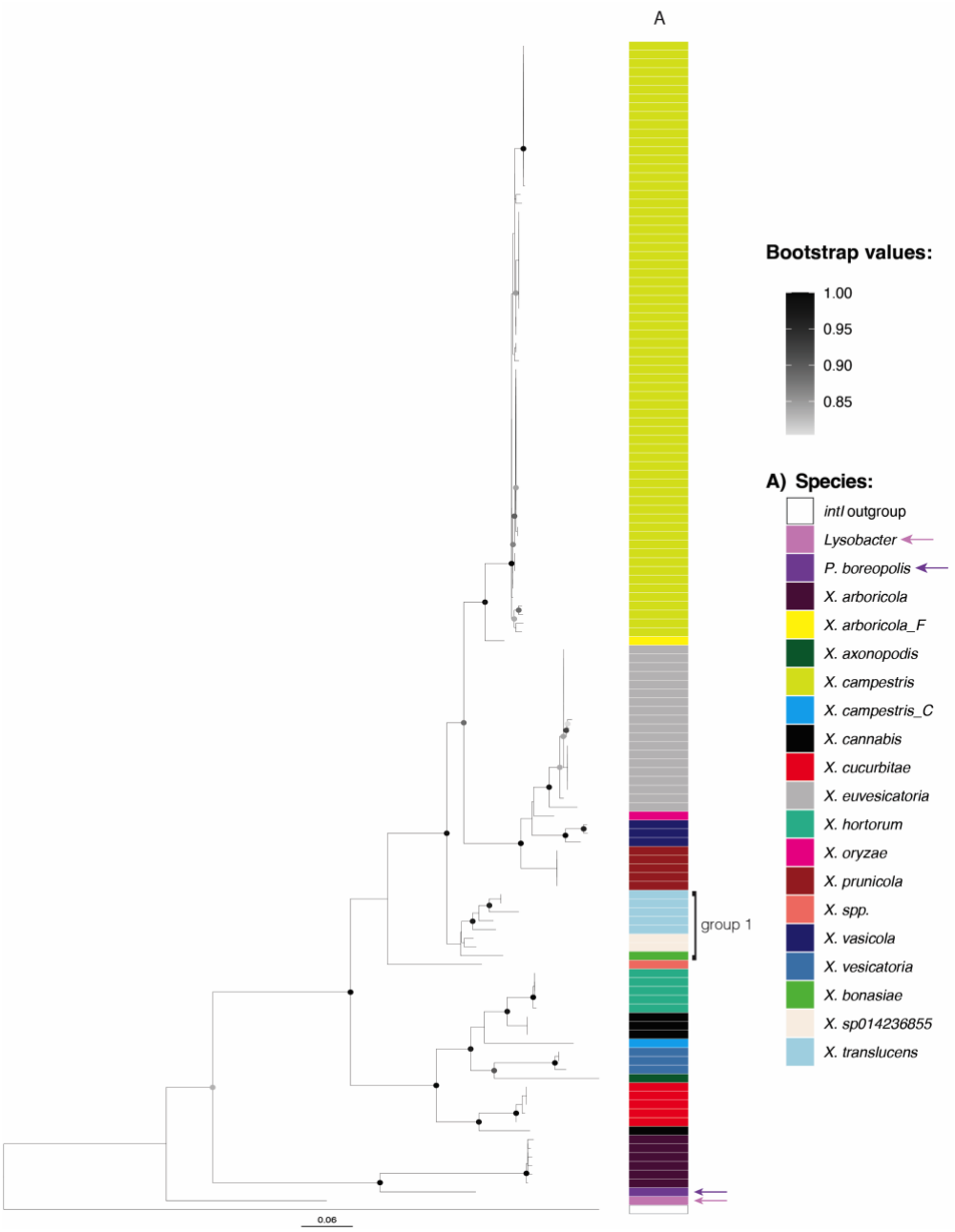


Figure 2 – Phylogeny of *intI*. Sequences of *intI* longer than 900 bp were aligned with PRANK, stripped of gaps and used to infer the tree with FastTree. *intI* from *Vibrio* sp. SCSIO 43136 (locus tag: J4N39_08275) was used as an outgroup. *intI* from group 1 *Xanthomonas* are indicated on the side of the plot, arrows indicate *intI*s of *Lysobacter* sp. CECT 30171 and [*Pseudomonas*] *boreopolis* strain GO2. The scale bar indicates substitutions per site. Only bootstrap values 0.80 are shown.

The phylogeny of *ilvD* and *intI* is incongruent, with the *ilvDs* of *Xanthomonas* group 1 forming a separate group in comparison to group 2 (Figure S2), while the group 1 *intIs* are nested within the most common *Xanthomonas intIs* (Figure 2). Genetic exchange between a restricted number of strains of group 2 and the entire group 1 clade has been reported (Pena *et al.*, 2024), therefore it is plausible that horizontal transfer of an *intI* gene from a

290 group 2 *Xanthomonas* species occurred during the early evolutionary history of group 1
291 *Xanthomonas*.

292 Together, this suggests that the acquisition of the integron platform was ancestral to the
293 *Xanthomonas* genus, and then *intI* diversified with the species. The exceptions are some *X.*
294 *arboricola* strains that could have recruited *intI* from another source, while group 1
295 *Xanthomonas* could have acquired *intI* from group 2 *Xanthomonas* in one single gene flow
296 event.

297 Additionally, approximately 50% of genomes harbor CALINs, here defined as clusters of at
298 least two *attC* sites, in loci other than the canonical *ilvD* region and in the absence of a
299 proximal *intI* gene. In genomes where *intI* is retained, this gene is consistently integrated
300 downstream of *ilvD,* while these CALINs show spatial separation, being located in other
301 regions of the genome. Among the species represented by more than 10 genomes, CALINs
302 not adjacent to *ilvD* were identified at seven distinct chromosomal loci. In *X. arboricola*, *X.*
303 *campestris*, *X. cissicola*, *X. euvesicatoria*, and *X. hortorum*, CALINs consistently occurred
304 adjacent to *lamG* and/or *secF*. In *X. cissicola*, additional CALINs were located near *recD* and a
305 gene encoding a NAD(P)/FAD-dependent oxidoreductase. In *X. oryzae*, two strains
306 possessed CALINs adjacent to an outer membrane protein gene, while in *X. translucens*,
307 CALINs were observed at two distinct genomic positions. One *X. campestris* strain
308 (GCA_028749605.1) harboured a CALIN on a conjugative plasmid. The relatively conserved
309 positioning of CALINs across species supports the hypothesis of ancestral integron activity.
310 This contrasts with experimental data from *Escherichia coli*, where cassette libraries
311 integrated via *attG* sites (consensus recombination sequences that are not *attC* or *attI*)
312 exhibited a broad distribution across numerous genomic loci, indicating that *de novo*
313 integration is typically less site-specific (Loot *et al.*, 2024).

314

315 **Acquisition of the integron platform is ancestral but its activity is progressively lost**

316 To investigate integron integrase activity and the complement of integron gene cassettes in
317 *Xanthomonas* genomes, we focused on species which had more than 10 genomes in our
318 dataset: *X. campestris* (n=111) (Figure 3), *X. cissicola* (n= 124) (Figure 4), *X. arboricola* (n=41)
319 (Figure S3), *X. translucens* (n=46) (Figure S4), *X. euvesicatoria* (n=46) (Figure S5), *X. oryzae*
320 (n=161) (Figure S6), *X. phaseoli* (n=23) (Figure S7), *X. hortorum* (n=12) (Figure S8). A detailed
321 description of the integrons in these eight species is reported in the Supplementary Results
322 Section.

323 Although integron associated sequences are common in *Xanthomonas* groups 1 and 2, most
324 strains appear to have lost integron integrase activity (*intI* is either truncated or absent), and
325 the distribution of functional integrons appears to be species specific. However, among the
326 species analysed, *X. campestris* stands out for maintaining a largely functional integron

platform (Figure 3). A high proportion (85.6%) of its genomes carry a full-length *intI* gene, and integrons are consistently flanked by extensive cassette arrays (ranging from 3 to 29 *attCs*, averaging 14.3 per genome). Cassette composition is highly variable even among strains of the same pathovar, suggesting that, in *X. campestris*, integron activity drives rapid diversification within genetically related groups. This robust activity is also evident in the gene content: 216 orthologous cassette-encoded genes were identified, 41.1% of which are singletons (observed in only one instance in this dataset). Despite this extensive variability, a gene encoding a DUF1488 domain-containing protein is present in the cassettes of 81.5% of genomes, while *symE* (a toxin-encoding gene) is found in 56.8% of the cassette arrays. This level of integron retention suggests that active integrons continue to play a key role in the adaptive evolution of *X. campestris*.

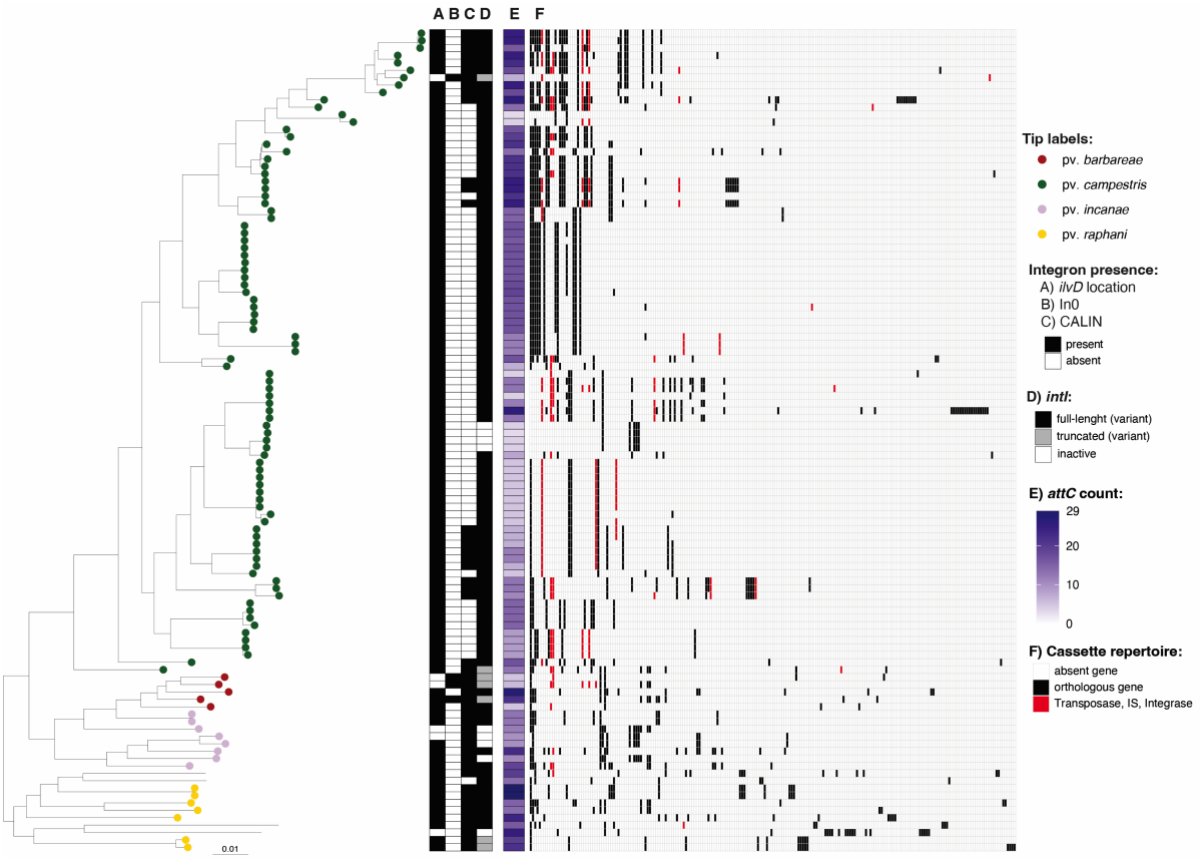


Figure 3 – Integrons in *X. campestris*. The phylogeny of *X. campestris* was built using Realphy, using GCA_000007145.1 as a reference genome. The tree was rooted with GCA_000972745 (*X. arboricola*) then the tip was removed from the tree. Tip labels show pathovars, which were assigned from the literature search (Table S3). The scale bar indicates substitutions per site. A) Indicates the presence of a complete integron or *attC* sites in the *ilvD* locus, B) indicates the presence of an *intI* lacking cassettes (In0), C) indicates the presence of a CALINs (cluster of *attC* sites) not in the *ilvD* locus, D) indicates whether IntI is predicted to be functional (full length) (Table S3). E) indicates the number of *attC* sites (number of cassettes) in the genome. Panel F) represents the distribution of orthologous genes among all cassettes carried by the corresponding isolate inferred with Proteinortho. Orthologous genes are listed in decreasing order based on the number of strains within the species that carry them.

349 *X. arboricola* exhibited a modest preservation of the *intI* gene with 7% of strains retaining a
350 full-length *intI.* In the remaining genomes, *intI* is either deleted or contains an early stop
351 codon. The *X. arboricola* phylogeny divides strains into two clades, clade A which includes
352 pathovars that cause diseases in *Prunus*, *Juglans*, and *Corylus* spp. (pvs. *pruni*, *juglandis*, and
353 *corylina*, respectively), and clade B which comprises many isolates lacking metadata (Figure
354 S3). Among clade A isolates, none possess a full-length *intI.* In contrast, clade B strains that
355 carry *intI*, either full-length or truncated (accounting for 58.8% of clade B isolates), harbor a
356 variant likely acquired via HGT, suggesting the acquisition event dated prior to the
357 diversification of this clade. Strains of clade B also carried more cassettes and a broader
358 array of genes compared to clade A.

359 *X. translucens* (Figure S4) and *X. euvesicatoria* (Figure S5) emphasize the dynamic nature of
360 integrons in *Xanthomonas.* Both species exhibit high frequencies of full-length *intI* (22.7%
361 and 45.6% respectively) and variability in cassette content is high (44.6% and 50%
362 respectively).

363 *X. oryzae* includes pathogens of rice (pv. *oryzae* and pv. *oryzicola*), and pathogens of a
364 pervasive weed species that grows along rivers and canals surrounding rice paddies (pv.
365 *leersiae*). However, only X11-5A, a weakly pathogenic strain (Triplett *et al.*, 2011) distantly
366 related to the above-mentioned pathovars carries a full-length *intI*. Divergent cassette
367 compositions were observed among the pathovars (Figure S6). This supports the hypothesis
368 that integron activity in *X. oryzae* was ancestrally active, followed by progressive
369 inactivation concomitant with pathovar diversification.

370 None of the *X. cissicola* (Figure 4) and X. *phaseoli* (Figure S7) isolates retain a full-length,
371 functional *intI*. However, 28.7% and 38% of the genes carried by the cassettes in these
372 species, respectively, appear in one isolate only (singletons). *X. cissicola* comprises isolates
373 commonly known as *X. citri*, an important pathogen able to infect many plants including
374 *Citrus* (pv. *citri*). *X. citri* pv. *citri* includes three recognized pathotypes: A, A*, and AwA. All
375 pv. *citri* genomes analysed here, which include both A and AwA pathotypes, harbor a CALIN
376 element adjacent to *secD* and lack the integrase gene *intI*. All the CALINs in the pv. *citri*
377 contain the same two gene cassettes, which are not present in any other genome analysed
378 in this study, suggesting that the integron platform was inactivated prior to pathotype
379 diversification (Figure 4). A recent genomic study of 95 pv. *citri* strains estimated that the
380 diversification of these pathotypes occurred 1,730 to 5,663 years ago (Patané *et al.*, 2019).
381 Given that citrus domestication is thought to have occurred at least 2,000 years ago (Rao *et*
382 *al.*, 2021), the inactivation event likely predates the domestication of the host genus. One of
383 the two cassettes encodes a protein homologous to members of the late embryogenesis
384 abundant (LEA) protein family, which is known to confer protection against water deficit in
385 bacteria (Raga-Carbajal *et al.*, 2022). This indicates that the cassette's function could be
386 oriented toward general environmental stress rather than mediating specific interactions
387 with the plant host.

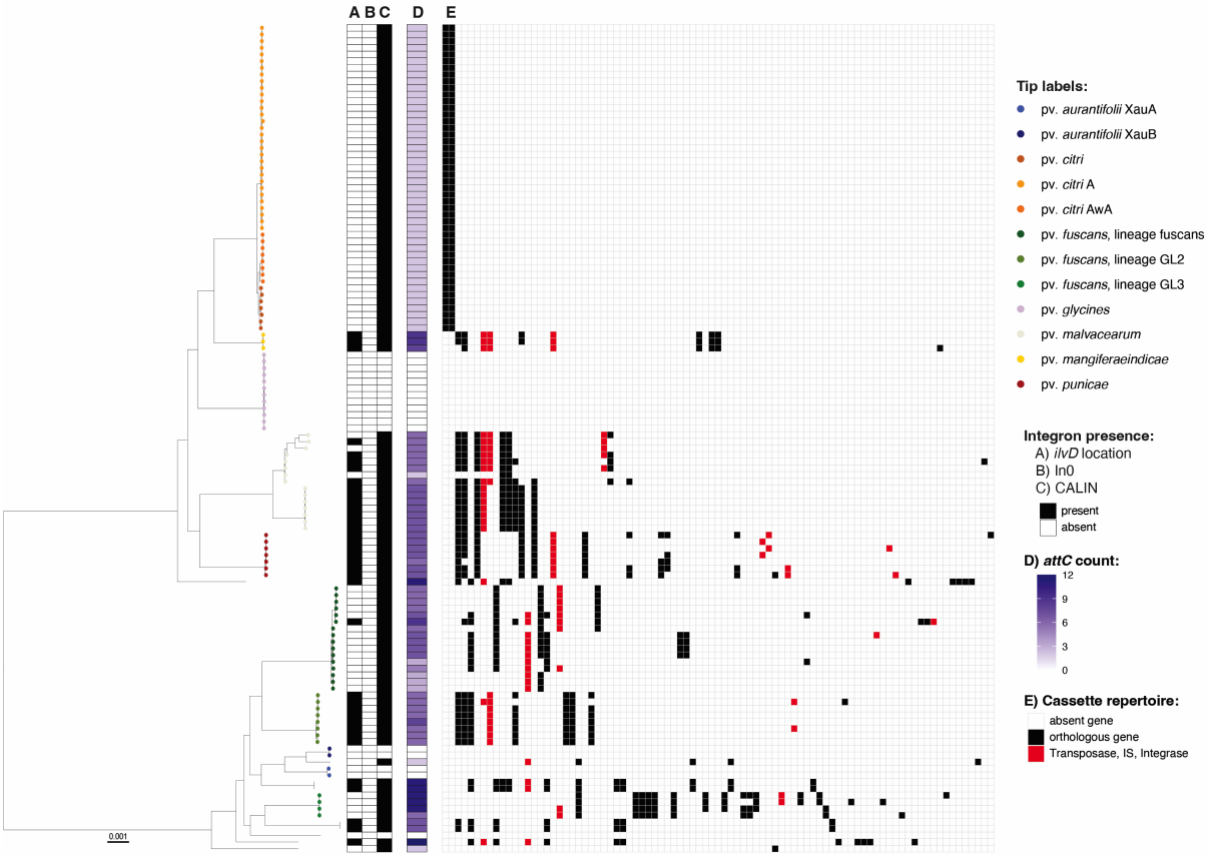


Figure 4 – Integrons in *X. cissicola*. The phylogeny of *X. cissicola* was built using Realphy, using GCA_000007165.1 as a reference genome. The tree was rooted with GCA_000009165.1 (*X. euvesicatoria*) then the tip was removed from the tree. Tip labels show pathovars, which were assigned from the literature search (Table S3). The scale bar indicates substitutions per site. A) Indicates the presence of a complete integron or *attC* sites in the *ilvD* locus, B) indicates the presence of an *intI* lacking cassettes (In0), C) indicates the presence of a CALINs (cluster of *attC* sites) not in the *ilvD* locus, D) indicates the number of *attC* sites (number of cassettes) in the genome (Table S3). Panel E) represents the distribution of orthologous genes among the all the cassettes carried by the corresponding isolate inferred with Proteinortho. Orthologous genes are given in decreasing order based on the number of strains within the species that carry them.

Gene cassette abundance is species specific.

We investigated whether variation in integron cassette array size, quantified by the number of *attC* sites per genome (Table S1), is primarily shaped by phylogeny or by the ecological context of isolation, categorized by cultivation type (crop, non-crop, crop tree, or ornamental host plants) (Table S4). We used only the species that had more than 10 genomes in our dataset, additionally, we excluded genomes lacking information on the host of isolation, and one strain isolated from mud. We first assessed whether variation in *attC* site counts exhibited a phylogenetic signal. Pagel's λ test indicated a very strong and highly

409 significant phylogenetic signal (λ=0.99, p < 2.22e-16), indicating that *attC* abundance is
410 strongly structured by evolutionary history. Given this strong phylogenetic structure, we
411 applied a phylogenetic generalized least squares regression to evaluate the effect of
412 cultivation type while accounting for shared ancestry. Isolates from ornamental plants had
413 significantly more *attC* sites (p < 0.001; mean = 12.1), whereas those from tree-crop
414 environments had fewer (p ≈ 0.017; mean = 3.05). No significant differences were observed
415 for crop (mean = 6.01) and non-crop (mean = 9.77) isolates (Figure S9). Overall, these results
416 suggest that while integron array size is strongly shaped by phylogenetic inertia, reflecting
417 higher similarity within rather than between species, ecological factors, such as
418 characteristics of the plant host, may also contribute. The prevalence of a functional
419 integron integrase appears to influence cassette abundance: species with predominantly
420 inactive integrases tend to carry fewer *attC* sites, likely due to gradual cassette loss.
421 However, the evolutionary forces determining why some lineages retain active *intI* while
422 others don't are unclear.

423 **Cassette arrays in *Xanthomonas* harbor diverse but poorly characterized gene functions**

424 Prediction of the taxonomic origins of integron *attC*s was performed with a classification
425 tool developed for *attC*s, which uses both sequence and structural homology information
426 (Ghaly *et al.*, 2021), to identify matches to known *attC* 'types'. The method classified all *attC*
427 sites in the *Xanthomonas* strains analysed as originating within the *Xanthomonadales*.
428 Therefore, there is no evidence of long-distance acquisition of cassettes, and the cassettes
429 appear to circulate at least within the *Xanthomonadales* family.

430 Arrays greatly varied in length across genomes, with a maximum of 46 cassettes in a single
431 genome (*X. arboricola_F* GCA_040182365.1) and 33 cassettes in a single array (*X. arboricola*
432 GCA_041475745.1 and GCA_041475785.1) (Table S1). Despite the majority of the isolates
433 carrying inactive integron modules, with *intI* deleted or truncated, the cassette arrays
434 appear to have been extremely dynamic before the inactivation of the integron integrase.

435 Overall, the 706 *Xanthomonas* genomes carried 1004 cassettes arrays, with a total of 4,773
436 cassettes, within these, Proteinortho (Klemm *et al.*, 2023) identified 1,087 different
437 orthologous genes. Of the 1,087 orthologous genes identified by Proteinortho, 476 (43.8%)
438 were present as singletons, unique to one isolate.

439 Of the 1,087 orthologous genes, only 11.9% could be classified into a known Clusters of
440 Orthologous Groups of proteins (COG) category (Galperin *et al.*, 2025) (Table S5). This
441 underrepresentation of known COGs among cassette proteins is well known and has been
442 reported for integrons from a range of different hosts and environments (Rowe-Magnus *et*
443 *al.*, 2003, Mazel, 2006, Boucher *et al.*, 2007, Ghaly *et al.*, 2024, Rajabal *et al.*, 2024). This
444 may be due to sampling bias in databases, with captured cassettes potentially being from

445 uncharacterised environmental organisms (Rodríguez Del Río *et al.*, 2024), or may be a
446 result of cassette genes being subject to high mutation rates after capture.

447 Of the most represented functions, excluding unknown and recombination and repair,
448 1.47% of the total cassette-encoded proteins were predicted to play a role in transcription,
449 0.9% in amino acid transport and metabolism, and 0.55% in defence mechanisms. To gain
450 further insight into possible cassette functions, we performed both sequence-based
451 homology searching against eggNOG 5.0 (Huerta-Cepas *et al.*, 2019), and protein structural
452 homology searching against the Swiss-Prot database. From this combined approach,
453 putative functions could only be assigned to 6.8% of the proteins. Of these annotated
454 cassette-encoded proteins, 47.7% were classified either as transposase, integrases or
455 insertion sequences (4.8% of the total). The presence of insertion sequences and
456 transposases within gene cassettes (Ghaly *et al.*, 2024), and targeting *attC* sites (Tetu &
457 Holmes, 2008, Post & Hall, 2009), has been frequently observed in past integron studies.

458 Signal peptides were identified in 11.6% of cassette-encoded proteins, using signalP6.0
459 (Teufel *et al.*, 2022). Again, transmembrane and secreted proteins are commonly encoded
460 by gene cassettes in Bacteria (Rowe-Magnus *et al.*, 2003, Ghaly *et al.*, 2023) and are
461 hypothesized to help facilitate interactions with their broader environment. Additionally,
462 PGPT-Pred (Ashrafi *et al.*, 2022) predicted 6.7% of cassette-encoded proteins to have plant
463 growth promoting traits such as biofertilization, plant signal production and stress control.
464 However, the most frequent category, representing 4.9% of total proteins, was associated
465 with competitive exclusion functions. These included antimicrobial resistance and
466 detoxification functions, and toxin-antitoxin systems, which in a plant pathogenic context
467 could contribute to niche adaptation by enhancing survival in a competitive microbial
468 environment. One gene was classified by the EcoFoldDB (Ghaly *et al.*, 2025) pipeline as
469 involved in spermidine production, which can play an important role in plant growth and
470 stress (Dunn & Becerra-Rivera, 2023).

471 DefenseFinder (Tesson *et al.*, 2022) identified 20 orthologous genes (1.8%) as genes
472 involved in defences against phages. Twelve different defence systems were detected, these
473 included restriction-modification and abortive infection systems, but also the newly
474 described Kiwa (Zhang *et al.*, 2023) and Shedu defence systems (Loeff *et al.*, 2025).

475 Using a database of previously characterized type three secretion system effectors (T3SEs)
476 retrieved from the EuroXanth platform (Costa *et al.*, 2024, Biessy *et al.*, 2025), we identified
477 two T3SEs: AvrBs3 (a Transcription Activator-Like effector, TALe) in African-like pv. *oryzae*
478 (Schepler-Luu *et al.*, 2023), and XopAF2 in three *X. vasicola* isolates. T3SEs are protein
479 secreted via the type III secretion system (T3SS) that suppress or induce plant defences.
480 TALes are a particular type of T3SE, once inside the host cell, they translocate to the nucleus
481 where their unique domain of tandemly arranged 34-amino acid repeats mediates binding
482 to specific promoter elements. Several TALes are known to induce host SWEET sucrose

483 uniporter genes, thereby facilitating sucrose efflux from xylem parenchyma into the
484 apoplasm at the infection sites. TALes may enhance disease by targeting susceptibility
485 genes, or may trigger a resistance response, and are therefore important for pathogen host
486 range and virulence (Schepler-Luu *et al.*, 2023).

487 Given that the majority of cassette-encoded proteins are of unknown function and often
488 occur as singletons (observed only in one strain), the phenotypic impact of these cassettes
489 remains unclear. It is uncertain whether they confer a selective advantage or are maintained
490 in the arrays through neutral processes. A cassette that is conserved across multiple species
491 is more likely to encode a beneficial function. The most widespread cassette, identified in 13
492 species, encodes *symE* and the non-coding small RNA *sRNA-Xcc1*. *sRNA-Xcc1* is known to be
493 activated by regulatory elements of the T3SS (Chen *et al.*, 2011), suggesting a potential role
494 in plant host colonization. In *E. coli*, *symE* forms part of the SymE-SymR toxin-antitoxin
495 system, where SymE induces nucleoid condensation, disrupts DNA replication and
496 transcription, and causes double-strand DNA breaks (Thompson *et al.*, 2022). Regulation in
497 *E. coli* occurs via the cis-encoded small RNA *symR* (Kawano *et al.*, 2007), whereas in the
498 integron cassettes, *sRNA-Xcc1* is located upstream of *symE*, suggesting a divergent
499 regulatory mechanism. Notably, unlike most integron-associated genes oriented to be
500 transcribed from the *Pc* promoter (typically located within *intI*), both *sRNA-Xcc1* and *symE*
501 are oriented in the opposite direction and not under *Pc* control.

502 Excluding transposases, the second most recurrent cassette, shared among 8 species,
503 encoded a protein of unknown function. The third most recurrent, present in 7 species,
504 included two hypothetical proteins and a gene conferring resistance to bleomycin.
505 Bleomycin resistance genes have been previously identified as enriched in gene cassettes
506 relative to total metagenomes in environmental samples, as shown by both cassette-
507 targeted amplicon sequencing and shotgun metagenomics (Ghaly *et al.*, 2023). One of the
508 recurring cassettes is an "empty" cassette which can occur multiple times in the same array.
509 This cassette could have promoter activity as it was previously suggested in other species
510 (Michael & Labbate, 2010, Tansirichaiya *et al.*, 2019, Blanco *et al.*, 2024).

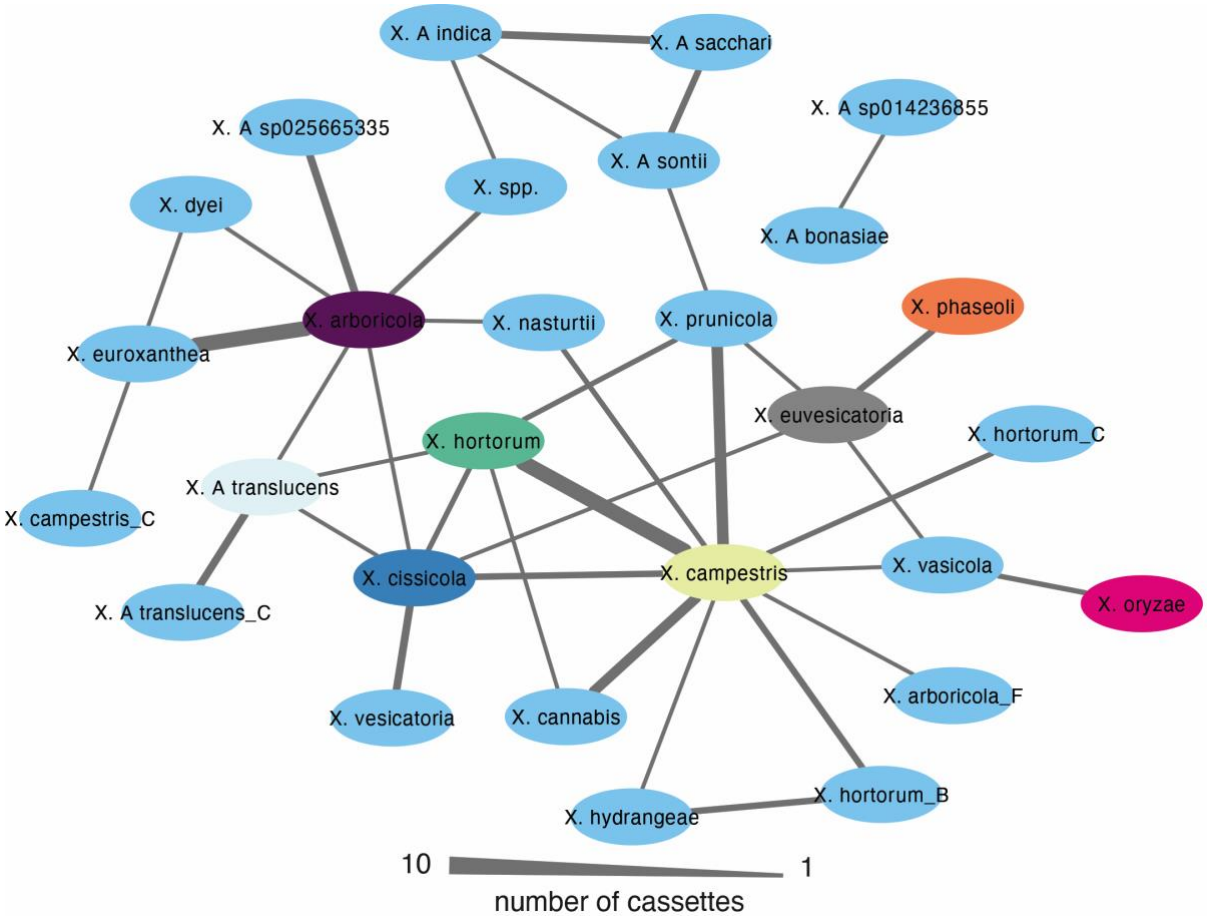
511

512 **Detection of inter-species horizontal transfer of cassette arrays**

513 To assess interspecies cassette gene sharing, we examined pairs of species that most
514 frequently share orthologous genes. As expected, species harbouring a greater number of
515 cassettes tend to share more orthologous genes with others. For example, *X. arboricola* and
516 *X. campestris* shared the highest number of orthologous genes (21), with *X. campestris*
517 sharing the most overall (57 orthologous genes across multiple species).

518 We then looked for more recent evidence of HGT of cassettes present in multiple species.
519 We used Alfy v1.0.5 (Domazet-Lošo & Haubold, 2011) to identify putative regions subject to

520 HGT. We further selected the regions identified by Alfy 1.0.5 selecting those based on the
521 difference in ANI between the cassette and the genomes harbouring them (Figure S10).
522 With this approach, we detected the movement of 95 cassettes between 38 pairs of species,
523 even between group 1 and 2 *Xanthomonas* (Figure 5). Figure 5 shows the number of
524 cassettes exchanged between species, without normalization for genome or cassette
525 abundance. While this approach does not account for differences in genome diversity within
526 and between species, or for uneven sampling depth (i.e. pandemic lineages are often
527 oversampled and appear less diverse in comparison to commensal isolates), it nevertheless
528 provides a direct view of observed horizontal transfer events. The highest flux of cassettes
529 (10) was between *X. arboricola* and *X. euroxanthea,* and between *X. campestris* and *X.*
530 *hortorum.* However, *X. campestris* is the species with the highest flux of cassettes compared
531 to any other species (36). One of the cassettes that showed evidence of horizontal transfer
532 was that carrying *symE* and *sRNA-Xcc1.*



533

534

535 **Figure 5 – Horizontal transfer of cassettes among *Xanthomonas* species.** Cytoscape network illustrating
536 transfer of cassettes between species. Nodes represent *Xanthomonas* species, species containing more
537 than 10 genomes in the dataset are coloured as in Figure 1 (tip labels). Line width between species is
538 proportional to the number of cassettes moved between species.

539

540 **CONCLUSIONS**

541 Our comprehensive analysis of 706 complete genomes reveals that the acquisition of the
542 integron platform is likely an ancestral event in the evolution of the *Xanthomonas* genus.
543 The widespread conservation of the *intI* gene downstream of *ilvD* across species, along with
544 *intI* phylogeny, strongly supports the early acquisition of the integron platform. Only in *X.*
545 *arboricola* clade B, the "original" *intI* appears to have been substituted via HGT with an *intI*
546 from another member of the Xanthomonadaceae family.

547 Despite the ancestral acquisition, our results show that integron activity, defined by the
548 presence of full-length, potentially functional *intI* genes, and large and diverse cassette
549 arrays, has been progressively lost in many *Xanthomonas* lineages. This inactivation appears
550 to be species-specific and often predates major evolutionary or ecological transitions, such
551 as pathovar diversification, such as in *X. cissicola* and *X. oryzae*. Nevertheless, in species
552 such as *X. campestris*, integrons remain robustly functional and continue to contribute to
553 genomic diversification, as evidenced by the presence of full-length *intIs*, high cassette
554 variability, and the presence of strain-specific gene cassettes in cassette arrays. Selective
555 pressures and ecological niches may favor the retention of an active integron system in this
556 species and not in the others, for reasons not clear at this time. Loss of integron integrase
557 activity in plant-associated pathovars may be favoured when it stabilizes cassette-associated
558 genetic traits that confer a fitness advantage. Once a beneficial cassette configuration is
559 established, inactivation of the integrase can prevent rearrangements or excisions that
560 would disrupt key functions, enabling these lineages to undergo clonal expansion. Integron
561 integrase activity may be also lost in the absence of positive selection to maintain it.
562 Integrons provide the greatest adaptive advantage in dynamic environments with a diverse
563 pool of accessible gene cassettes. Such diversity is potentially diminished in intensive
564 agricultural systems, where pathogens encounter repetitive host-pathogen cycles and
565 microbial communities are less diverse than in wild plant environments (Karasov *et al.*,
566 2018). Under such conditions, the selective pressure to maintain integrase activity might be
567 reduced. In agricultural settings, other types of mobile genetic elements carrying genes
568 already optimized by selection in donor strains may play a more prominent role in driving
569 plant-bacteria evolution (Cesbron *et al.*, 2015, Colombi *et al.*, 2023, Colombi *et al.*, 2024,
570 Jibrin *et al.*, 2024, Sparks *et al.*, 2025).

571 The phenotypic impact of the large pool of gene cassettes residing in *Xanthomonas*
572 genomes remains largely unresolved. Nearly half of orthologous genes carried by gene
573 cassettes occurred only once, contributing to strain level genetic variability. These rare
574 genes may represent a reservoir of adaptive potential, but could also be selectively neutral
575 or even transient. Cassettes in *Vibrio* chromosomal integron platforms have been suggested
576 to have been selected to be as neutral as possible (Blanco *et al.*, 2024). The frequent

577 occurrence of hypothetical proteins, including highly conserved cassettes ORFs, suggests
578 important functional roles that remain undiscovered. For instance, a cassette carrying the
579 DUF1488 encoding gene is present, a few times in multiple copies, in 16.7% of the genomes
580 across 6 species, sometimes with multiple copies within the same integron array. Among
581 cassette-encoded genes with predicted putative functions, many have predicted functions
582 linked to environmental interaction and thus have the potential to contribute to niche
583 adaptation. For example, the most widespread cassette in our dataset encodes *symE* and
584 the small coding sRNA-Xcc1, known to be activated by T3SS regulatory proteins (Chen *et al.*,
585 2011), which may influence bacterial growth during plant colonization. In *Xanthomonas*, we
586 found a fairly small fraction of cassette genes to be classified as involved defence
587 mechanisms, despite the emerging role of integron platforms in encoding anti-phage genes
588 (Darracq *et al.*, 2025, Getz *et al.*, 2025, Kieffer *et al.*, 2025). However, this may be an
589 underestimate, as work on large sedentary chromosomal integrons in *Vibrio cholerae*
590 showed that 20% of their cassettes with no predicted function encode for anti-phage
591 defence systems, despite these genes not being detected as such by bioinformatic tools
592 (Darracq *et al.*, 2025).

593 Importantly, our analysis revealed evidence of interspecies horizontal transfer of cassettes.
594 For both *X. campestris* and *X. arboricola* there is evidence of recent horizontal transfer of
595 integron-associated cassettes with other *Xanthomonas* species. Interestingly, however,
596 despite their high levels of cassette acquisition, there is little evidence of recent and direct
597 cassette exchange between these two species. Instead, *X. campestris* shows the highest
598 degree of gene sharing with *X. hortorum*, a species more closely related to *X. arboricola*
599 than to *X. campestris*. Conversely, the species with which *X. arboricola* shares the greatest
600 number of cassettes is *X. euroxanthea*, its closest phylogenetic neighbour among the
601 genomes analysed. These patterns suggest that both phylogenetic proximity and ecological
602 overlap, such as co-occurrence in the phyllosphere or endosphere during mixed infections,
603 have contributed to the dynamics of cassette exchange. The shared habitat provides
604 repeated opportunities for interspecies contact and genetic exchange, likely promoting the
605 horizontal dissemination of integron cassettes across the genus.

606 In summary, our analysis indicates that integron platforms were likely acquired early in the
607 evolution of the *Xanthomonas* genus and have since followed divergent evolutionary
608 trajectories. While many lineages show signs of integron inactivation, others, such as *X.*
609 *campestris*, retain active systems contributing to genomic diversification. The gene cassettes
610 carried by these integrons, including many with unknown or potentially adaptive functions,
611 add to strain-level variability and may influence niche adaptation. Patterns of horizontal
612 gene transfer suggest that both phylogenetic relatedness and ecological overlap shape
613 cassette exchange across species, highlighting integrons as dynamic elements in the
614 evolutionary landscape of *Xanthomonas*.

615

Funding information: This work is supported by funding from the ARC Centre of Excellence in Synthetic Biology (CE200100029).
Conflicts of interest: The authors declare that there is no conflict of interest.

REFERENCES

Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G & Nielsen H (2019) SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature biotechnology* **37**: 420-423.

Ansari MA & Didelot X (2014) Inference of the properties of the recombination process from whole bacterial genomes. *Genetics* **196**: 253-265.

Ashrafi S, Kuzmanović N, Patz S, Lohwasser U, Bunk B, Spröer C, Lorenz M, Elhady A, Frühling A & Neumann-Schaal M (2022) Two new Rhizobiales species isolated from root nodules of common sainfoin (Onobrychis viciifolia) show different plant colonization strategies. *Microbiology Spectrum* **10**: e01099-01022.

Barionovi D & Scortichini M (2006) Assessment of integron gene cassette arrays in strains of *Xanthomonas fragariae* and *X. arboricola* pvs. *fragariae* and *pruni*. *Journal of Plant Pathology* 279-284.

Barionovi D & Scortichini M (2008) Integron variability in *Xanthomonas arboricola* pv. *juglandis* and *Xanthomonas arboricola* pv. *pruni* strains. *FEMS microbiology letters* **288**: 19-24.

Bertels F, Silander OK, Pachkov M, Rainey PB & Van Nimwegen E (2014) Automated reconstruction of whole-genome phylogenies from short-sequence reads. *Molecular biology and evolution* **31**: 1077-1088.

Biessy A, Cadieux M, Ciotola M, McDuff F, Soufiane B, Laforest M & Filion M (2025) Characterization of a plant-pathogenic T3SS-lacking *Xanthomonas* strain isolated from common ragweed. *Plant Pathology* **74**: 308-319.

Blanco P, Hipólito A, García-Pastor L, Trigo da Roza F, Toribio-Celestino L, Ortega AC, Vergara E, San Millán Á & Escudero JA (2024) Identification of promoter activity in gene-less cassettes from *Vibrionaceae* superintegrons. *Nucleic acids research* **52**: 2961-2976.

Boucher Y, Labbate M, Koenig JE & Stokes H (2007) Integrons: mobilizable platforms that promote genetic diversity in bacteria. *Trends in microbiology* **15**: 301-309.

650 Buchfink B, Xie C & Huson DH (2015) Fast and sensitive protein alignment using DIAMOND.
651 *Nature methods* **12**: 59-60.

652 Cambray G, Sanchez-Alberola N, Campoy S, Guerin É, Da Re S, González-Zorn B, Ploy M-C,
653 Barbé J, Mazel D & Erill I (2011) Prevalence of SOS-mediated control of integron integrase
654 expression as an adaptive trait of chromosomal and mobile integrons. *Mobile DNA* **2**: 6.

655 Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P & Huerta-Cepas J (2021) eggNOG-
656 mapper v2: functional annotation, orthology assignments, and domain prediction at the
657 metagenomic scale. *Molecular biology and evolution* **38**: 5825-5829.

658 Cesbron S, Briand M, Essakhi S, Gironde S, Boureau T, Manceau C, Fischer-Le Saux M &
659 Jacques M-A (2015) Comparative genomics of pathogenic and nonpathogenic strains of
660 *Xanthomonas arboricola* unveil molecular and evolutionary events linked to
661 pathoadaptation. *Frontiers in Plant Science* **6**: 1126.

662 Chaumeil P-A, Mussig AJ, Hugenholtz P & Parks DH (2022) GTDB-Tk v2: memory friendly
663 classification with the genome taxonomy database. *Bioinformatics* **38**: 5315-5316.

664 Chen X-L, Tang D-J, Jiang R-P, He Y-Q, Jiang B-L, Lu G-T & Tang J-L (2011) *sRNA-Xcc1*, an
665 integron-encoded transposon-and plasmid-transferred trans-acting sRNA, is under the
666 positive control of the key virulence regulators HrpG and HrpX of *Xanthomonas campestris*
667 pathovar *campestris*. *RNA biology* **8**: 947-953.

668 Colombi E, Hill Y, Lines R, Sullivan JT, Kohlmeier MG, Christophersen CT, Ronson CW,
669 Terpolilli JJ & Ramsay JP (2023) Population genomics of Australian indigenous
670 *Mesorhizobium* reveals diverse nonsymbiotic genospecies capable of nitrogen-fixing
671 symbioses following horizontal gene transfer. *Microbial genomics* **9**: 000918.

672 Colombi E, Bertels F, Doulcier G, McConnell E, Pichugina T, Sohn KH, Straub C, McCann HC &
673 Rainey PB (2024) Rapid dissemination of host metabolism–manipulating genes via
674 integrative and conjugative elements. *Proceedings of the National Academy of Sciences* **121**:
675 e2309263121.

676 Costa J, Pothier JF, Bosis E, Boch J, Kölliker R & Koebnik R (2024) A community-curated
677 DokuWiki resource on diagnostics, diversity, pathogenicity, and genetic control of
678 xanthomonads. *Molecular Plant-Microbe Interactions* **37**: 347-353.

679 Cury J, Jové T, Touchon M, Néron B & Rocha EP (2016) Identification and analysis of
680 integrons and cassette arrays in bacterial genomes. *Nucleic acids research* **44**: 4539-4550.

681 Darracq B, Littner E, Brunie M, Bos J, Kaminski PA, Depardieu F, Slesak W, Debatisse K,
682 Touchon M & Bernheim A (2025) Sedentary chromosomal integrons as biobanks of bacterial
683 antiphage defense systems. *Science* **388**: eads0768.

684 Domazet-Lošo M & Haubold B (2011) Alignment-free detection of local similarity among
685 viral and bacterial genomes. *Bioinformatics* **27**: 1466-1472.

686 Dunn MF & Becerra-Rivera VA (2023) The biosynthesis and functions of polyamines in the
687 interaction of plant growth-promoting rhizobacteria with plants. *Plants* **12**: 2671.

688 Escudero JA, Loot C & Mazel D (2018) Integrons as adaptive devices. *Molecular mechanisms*
689 *of microbial evolution* 199-239.

690 Essakhi S, Cesbron S, Fischer-Le Saux M, Bonneau S, Jacques M-A & Manceau C (2015)
691 Phylogenetic and variable-number tandem-repeat analyses identify nonpathogenic
692 Xanthomonas arboricola lineages lacking the canonical type III secretion system. *Applied*
693 *and Environmental Microbiology* **81**: 5395-5410.

694 Galperin MY, Vera Alvarez R, Karamycheva S, Makarova KS, Wolf YI, Landsman D & Koonin
695 EV (2025) COG database update 2024. *Nucleic Acids Research* **53**: D356-D363.

696 Garita-Cambronero J, Palacio-Bielsa A, López MM & Cubero J (2017) Pan-genomic analysis
697 permits differentiation of virulent and non-virulent strains of Xanthomonas arboricola that
698 cohabit Prunus spp. and elucidate bacterial virulence factors. *Frontiers in microbiology* **8**:
699 573.

700 Getz LJ, Fairburn SR, Vivian Liu Y, Qian AL & Maxwell KL (2025) Integrons are anti-phage
701 defence libraries in *Vibrio parahaemolyticus*. *Nature Microbiology* 1-10.

702 Ghaly TM, Tetu SG & Gillings MR (2021) Predicting the taxonomic and environmental
703 sources of integron gene cassettes using structural and sequence homology of attC sites.
704 *Communications biology* **4**: 946.

705 Ghaly TM, Geoghegan JL, Alroy J & Gillings MR (2019) High diversity and rapid spatial
706 turnover of integron gene cassettes in soil. *Environmental Microbiology* **21**: 1567-1574.

707 Ghaly TM, Geoghegan JL, Tetu SG & Gillings MR (2020) The peril and promise of integrons:
708 beyond antibiotic resistance. *Trends in microbiology* **28**: 455-464.

709 Ghaly TM, Gillings MR, Rajabal V, Paulsen IT & Tetu SG (2024) Horizontal gene transfer in
710 plant microbiomes: integrons as hotspots for cross-species gene exchange. *Frontiers in*
711 *Microbiology* **15**: 1338026.

712 Ghaly TM, Rajabal V, Russel D, Colombi E & Tetu SG (2025) EcoFoldDB: Protein structure-
713 guided functional profiling of ecologically relevant microbial traits at the metagenome scale.
714 *bioRxiv* 2025.2004. 2002.646905.

715 Ghaly TM, Tetu SG, Penesyan A, Qi Q, Rajabal V & Gillings MR (2022) Discovery of integrons
716 in Archaea: Platforms for cross-domain gene transfer. *Science Advances* **8**: eabq6376.

717 Ghaly TM, Penesyan A, Pritchard A, Qi Q, Rajabal V, Tetu SG & Gillings MR (2022) Methods
718 for the targeted sequencing and analysis of integrons and their gene cassettes from complex
719 microbial communities. *Microbial genomics* **8**: 000788.

720 Ghaly TM, Rajabal V, Penesyan A, Coleman NV, Paulsen IT, Gillings MR & Tetu SG (2023)
721 Functional enrichment of integrons: Facilitators of antimicrobial resistance and niche
722 adaptation. *Iscience* **26**.

723 Gillings M, Boucher Y, Labbate M, Holmes A, Krishnan S, Holley M & Stokes H (2008) The
724 evolution of class 1 integrons and the rise of antibiotic resistance. *Journal of bacteriology*
725 **190**: 5095-5100.

726 Gillings MR (2014) Integrons: past, present, and future. *Microbiology and molecular biology*
727 *reviews* **78**: 257-277.

728 Gillings MR, Holley MP, Stokes H & Holmes AJ (2005) Integrons in *Xanthomonas*: a source of
729 species genome diversity. *Proceedings of the National Academy of Sciences* **102**: 4419-4424.

730 Guerin É, Cambray G, Sanchez-Alberola N, Campoy S, Erill I, Da Re S, Gonzalez-Zorn B, Barbé
731 J, Ploy M-C & Mazel D (2009) The SOS response controls integron recombination. *Science*
732 **324**: 1034-1034.

733 Hall RM & Collis CM (1995) Mobile gene cassettes and integrons: capture and spread of
734 genes by site-specific recombination. *Molecular microbiology* **15**: 593-600.

735 Heinzinger M, Weissenow K, Sanchez JG, Henkel A, Mirdita M, Steinegger M & Rost B (2024)
736 Bilingual language model for protein sequence and structure. *NAR Genomics and*
737 *Bioinformatics* **6**: lqae150.

738 Hipólito A, García-Pastor L, Vergara E, Jové T & Escudero JA (2023) Profile and resistance
739 levels of 136 integron resistance genes. *npj Antimicrobials and Resistance* **1**: 13.

740 Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR,
741 Letunic I, Rattei T & Jensen LJ (2019) eggNOG 5.0: a hierarchical, functionally and
742 phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses.
743 *Nucleic acids research* **47**: D309-D314.

744 Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT & Aluru S (2018) High throughput
745 ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature*
746 *communications* **9**: 5114.

747 Jibrin MO, Sharma A, Mavian CN, Timilsina S, Kaur A, Iruegas-Bocardo F, Potnis N,
748 Minsavage GV, Coutinho TA & Creswell TC (2024) Phylodynamic insights into global
749 emergence and diversification of the tomato pathogen *Xanthomonas hortorum* pv. *gardneri*.
750 *Molecular Plant-Microbe Interactions* **37**: 712-720.

751 Karasov TL, Almario J, Friedemann C*, et al.* (2018) *Arabidopsis thaliana* and *Pseudomonas*
752 pathogens exhibit stable associations over evolutionary timescales. *Cell host & microbe* **24**:
753 168-179. e164.

754 Kawano M, Aravind á & Storz G (2007) An antisense RNA controls synthesis of an SOS-
755 induced toxin evolved from an antitoxin. *Molecular microbiology* **64**: 738-754.

Kieffer N, Hipólito A, Ortiz-Miravalles L, Blanco P, Delobelle T, Vizuete P, Ojeda FM, Jové T, Jurenas D & García-Quintanilla M (2025) Mobile integrons encode phage defense systems. *Science* **388**: eads0915.

Klemm P, Stadler PF & Lechner M (2023) Proteinortho6: pseudo-reciprocal best alignment heuristic for graph-based detection of (co-) orthologs. *Frontiers in bioinformatics* **3**: 1322477.

Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**: 357-359.

Lemoine F & Gascuel O (2021) Gotree/Goalign: toolkit and Go API to facilitate the development of phylogenetic workflows. *NAR Genomics and Bioinformatics* **3**: lqab075.

Loeff L, Walter A, Rosalen GT & Jinek M (2025) DNA end sensing and cleavage by the Shedu anti-phage defense system. *Cell* **188**: 721-733. e717.

Loot C, Millot GA, Richard E, Littner E, Vit C, Lemoine F, Néron B, Cury J, Darracq B & Niault T (2024) Integron cassettes integrate into bacterial genomes via widespread non-classical attG sites. *Nature Microbiology* **9**: 228-240.

Löytynoja A (2013) Phylogeny-aware alignment with PRANK. *Multiple sequence alignment methods,*p.^pp. 155-170. Springer.

Mazel D (2006) Integrons: agents of bacterial evolution. *Nature Reviews Microbiology* **4**: 608-620.

Mazel D, Dychinco B, Webb VA & Davies J (1998) A distinctive class of integron in the *Vibrio cholerae* genome. *Science* **280**: 605-608.

Michael CA & Labbate M (2010) Gene cassette transcription in a large integron-associated array. *BMC genetics* **11**: 1-13.

Néron B, Littner E, Haudiquet M, Perrin A, Cury J & Rocha EP (2022) IntegronFinder 2.0: identification and analysis of integrons across bacteria, with a focus on antibiotic resistance in Klebsiella. *Microorganisms* **10**: 700.

Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, Isaac N & Pearse W (2013) The caper package: comparative analysis of phylogenetics and evolution in R. *R package version* **5**: 1-36.

Paradis E & Schliep K (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**: 526-528.

Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil P-A & Hugenholtz P (2022) GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic acids research* **50**: D785-D794.

Patané JS, Martins J, Rangel LT, Belasque J, Digiampietri LA, Facincani AP, Ferreira RM, Jaciani FJ, Zhang Y & Varani AM (2019) Origin and diversification of *Xanthomonas citri* subsp. *citri* pathotypes revealed by inclusive phylogenomic, dating, and biogeographic analyses. *BMC genomics* **20**: 1-23.

Pena MM, Bhandari R, Bowers RM, Weis K, Newberry E, Wagner N, Pupko T, Jones JB, Woyke T & Vinatzer BA (2024) Genetic and functional diversity help explain pathogenic, weakly pathogenic, and commensal lifestyles in the genus Xanthomonas. *Genome Biology and Evolution* **16**: evae074.

Pereira BM, Österlund T, Eriksson KM, Backhaus T, Axelson-Fisk M & Kristiansson E (2020) A comprehensive survey of integron-associated genes present in metagenomes. *BMC genomics* **21**: 1-14.

Pinheiro J, Bates D, DebRoy S, Sarkar D, Heisterkamp S, Van Willigen B & Maintainer R (2017) Package 'nlme'. *Linear and nonlinear mixed effects models, version* **3**: 274.

Post V & Hall RM (2009) Insertion sequences in the IS 1111 family that target the attC recombination sites of integron-associated gene cassettes. *FEMS microbiology letters* **290**: 182-187.

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

Qi Q, Ghaly TM, Rajabal V, Russell DH, Gillings MR & Tetu SG (2024) Vegetable phylloplane microbiomes harbour class 1 integrons in novel bacterial hosts and drive the spread of chlorite resistance. *Science of the Total Environment* **954**: 176348.

Raga-Carbajal E, Espin G, Ayala M, Rodríguez-Salazar J & Pardo-López L (2022) Evaluation of a bacterial group 1 LEA protein as an enzyme protectant from stress-induced inactivation. *Applied Microbiology and Biotechnology* **106**: 5551-5562.

Rajabal V, Ghaly TM, Egidi E, Ke M, Penesyan A, Qi Q, Gillings MR & Tetu SG (2024) Exploring the role of mobile genetic elements in shaping plant–bacterial interactions for sustainable agriculture and ecosystem health. *Plants, People, Planet* **6**: 408-420.

Rao MJ, Zuo H & Xu Q (2021) Genomic insights into citrus domestication and its important agronomic traits. *Plant Communications* **2**.

Rodríguez Del Río Á, Giner-Lamia J, Cantalapiedra CP, Botas J, Deng Z, Hernández-Plaza A, Munar-Palmer M, Santamaría-Hernando S, Rodríguez-Herva JJ & Ruscheweyh H-J (2024) Functional and evolutionary significance of unknown genes from uncultivated taxa. *Nature* **626**: 377-384.

Rowe-Magnus DA & Mazel D (2001) Integrons: natural tools for bacterial genome evolution. *Current opinion in microbiology* **4**: 565-569.

826 Rowe-Magnus DA, Guerout A-M, Biskri L, Bouige P & Mazel D (2003) Comparative analysis of
827 superintegrons: engineering extensive genetic diversity in the Vibrionaceae. *Genome*
828 *research* **13**: 428-442.

829 Sapkota S, Mergoum M & Liu Z (2020) The translucens group of Xanthomonas translucens:
830 Complicated and important pathogens causing bacterial leaf streak on cereals. *Molecular*
831 *plant pathology* **21**: 291-302.

832 Schepler-Luu V, Sciallano C, Stiebner M, Ji C, Boulard G, Diallo A, Auguy F, Char SN, Arra Y &
833 Schenstnyi K (2023) Genome editing of an African elite rice variety confers resistance against
834 endemic and emerging Xanthomonas oryzae pv. oryzae strains. *Elife* **12**: e84864.

835 Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J & Goesmann A (2021) Bakta:
836 rapid and standardized annotation of bacterial genomes via alignment-free sequence
837 identification. *Microbial genomics* **7**: 000685.

838 Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B &
839 Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular
840 interaction networks. *Genome research* **13**: 2498-2504.

841 Sparks AH, Adorada DL, Colombi E, Kelly LA, Young A, Knight NL & Vaghefi N (2025) Clonal
842 expansion from standing genetic variation underpins the evolution of *Curtobacterium*
843 *flaccumfaciens* pv. *flaccumfaciens* in Australia. *Phytopathology*.

844 Stokes Ht & Hall R (1989) A novel family of potentially mobile DNA elements encoding site-
845 specific gene-integration functions: integrons. *Molecular microbiology* **3**: 1669-1683.

846 Strugeon E, Tilloy V, Ploy M-C & Da Re S (2016) The stringent response promotes antibiotic
847 resistance dissemination by regulating integron integrase expression in biofilms. *MBio* **7**:
848 10.1128/mbio. 00868-00816.

849 Tansirichaiya S, Mullany P & Roberts AP (2019) Promoter activity of ORF-less gene cassettes
850 isolated from the oral metagenome. *Scientific Reports* **9**: 8388.

851 Tesson F, Hervé A, Mordret E, Touchon M, d'Humières C, Cury J & Bernheim A (2022)
852 Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nature*
853 *communications* **13**: 2561.

854 Tetu SG & Holmes AJ (2008) A family of insertion sequences that impacts integrons by
855 specific targeting of gene cassette recombination sites, the IS 1111-attC group. *Journal of*
856 *bacteriology* **190**: 4959-4970.

857 Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD, Winther O,
858 Brunak S, von Heijne G & Nielsen H (2022) SignalP 6.0 predicts all five types of signal
859 peptides using protein language models. *Nature biotechnology* **40**: 1023-1025.

860 Thompson MK, Nocedal I, Culviner PH, Zhang T, Gozzi KR & Laub MT (2022) *Escherichia coli*
861 SymE is a DNA-binding protein that can condense the nucleoid. *Molecular microbiology* **117**:
862 851-870.

863 Timilsina S, Potnis N, Newberry EA, Liyanapathiranage P, Iruegas-Bocardo F, White FF, Goss
864 EM & Jones JB (2020) Xanthomonas diversity, virulence and plant–pathogen interactions.
865 *Nature Reviews Microbiology* **18**: 415-427.

866 Triplett LR, Hamilton J, Buell C, Tisserat N, Verdier V, Zink F & Leach J (2011) Genomic
867 analysis of Xanthomonas oryzae isolates from rice grown in the United States reveals
868 substantial divergence from known X. oryzae pathovars. *Applied and Environmental*
869 *Microbiology* **77**: 3930-3937.

870 Van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CL, Söding J & Steinegger
871 M (2024) Fast and accurate protein structure search with Foldseek. *Nature biotechnology*
872 **42**: 243-246.

873 Vauterin L, Yang P, Alvarez A, Takikawa Y, Roth DA, Vidaver AK, Stall RE, Kersters K & Swings
874 J (1996) Identification of non-pathogenic Xanthomonas strains associated with plants.
875 *Systematic and applied microbiology* **19**: 96-105.

876 Xu S, Dai Z, Guo P, Fu X, Liu S, Zhou L, Tang W, Feng T, Chen M & Zhan L (2021) ggtreeExtra:
877 compact visualization of richly annotated phylogenetic data. *Molecular biology and*
878 *evolution* **38**: 4039-4042.

879 Yu G, Smith DK, Zhu H, Guan Y & Lam TTY (2017) ggtree: an R package for visualization and
880 annotation of phylogenetic trees with their covariates and other associated data. *Methods*
881 *in Ecology and Evolution* **8**: 28-36.

882 Zhang Z, Todeschini TC, Wu Y, Kogay R, Naji A, Rodriguez JC, Mondi R, Kaganovich D, Taylor
883 DW & Bravo JP (2023) Kiwa is a bacterial membrane-embedded defence supercomplex
884 activated by phage-induced membrane changes. *bioRxiv* 2023.2002. 2026.530102.

885

886 **FIGURE AND TABLE LEGENDS**

887 **Figure 1 – Distribution of integron platforms in *Xanthomonas* groups 1 and 2.** The
888 maximum likelihood phylogenetic tree is based on a concatenated alignment of 120 top-
889 ranked marker proteins. The tree was rooted at the midpoint. Tips are labelled with species
890 names assigned with Genome Taxonomy Database phylogeny. Coloured clades show
891 species with more than genomes. A) Integrons were classified as present if the isolates
892 carried a functional integron, *intI* only or an *attC* array only, B) Indicates the presence of a
893 complete integron or *attC* sites in the *ilvD* locus, C) indicates the presence of an *intI* lacking
894 cassettes (In0), D) indicates the presence of a CALINs (cluster of *attC* sites) not in the *ilvD*
895 locus. E) indicates the number of *attC* sites (number of cassettes) in the genome. Only
896 bootstrap values 0.80 are shown. Scale bar indicates substitutions per site.

897 **Figure 2 – Phylogeny of *intI*.** Sequences of *intI* longer than 900 bp were aligned with PRANK,
898 stripped of gaps and used to infer the tree with FastTree. *intI* from *Vibrio* sp. SCSIO 43136
899 (locus tag: J4N39_08275) was used as an outgroup. *intI* from group 1 *Xanthomonas* are

900 indicated on the side of the plot, arrows indicate *intI*s of *Lysobacter* sp. CECT 30171 and
901 [*Pseudomonas*] *boreopolis* strain GO2. The scale bar indicates substitutions per site. Only
902 bootstrap values 0.80 are shown.

903 **Figure 3 – Integrons in *X. campestris*.** The phylogeny of *X. campestris* was built using
904 Realphy, using GCA_000007145.1 as a reference genome. The tree was rooted with
905 GCA_000972745 (*X. arboricola*) then the tip was removed from the tree. Tip labels show
906 pathovars, which were assigned from the literature search (Table S3). The scale bar
907 indicates substitutions per site. A) Indicates the presence of a complete integron or *attC*
908 sites in the *ilvD* locus, B) indicates the presence of an *intI* lacking cassettes (In0), C) indicates
909 the presence of a CALINs (cluster of *attC* sites) not in the *ilvD* locus, D) indicates whether IntI
910 is predicted to be functional (full length) (Table S3). E) indicates the number of *attC* sites
911 (number of cassettes) in the genome. Panel F) represents the distribution of orthologous
912 genes among all cassettes carried by the corresponding isolate inferred with Proteinortho.
913 Orthologous genes are listed in decreasing order based on the number of strains within the
914 species that carry them.

915 **Figure 4 – Integrons in *X. cissicola*.** The phylogeny of *X. cissicola* was built using Realphy,
916 using GCA_000007165.1 as a reference genome. The tree was rooted with
917 GCA_000009165.1 (*X. euvesicatoria*) then the tip was removed from the tree. Tip labels
918 show pathovars, which were assigned from the literature search (Table S3). The scale bar
919 indicates substitutions per site. A) Indicates the presence of a complete integron or *attC*
920 sites in the *ilvD* locus, B) indicates the presence of an *intI* lacking cassettes (In0), C) indicates
921 the presence of a CALINs (cluster of *attC* sites) not in the *ilvD* locus, D) indicates the number
922 of *attC* sites (number of cassettes) in the genome (Table S3). Panel E) represents the
923 distribution of orthologous genes among the all the cassettes carried by the corresponding
924 isolate inferred with Proteinortho. Orthologous genes are given in decreasing order based
925 on the number of strains within the species that carry them.

926 **Figure 5 – Horizontal transfer of cassettes among *Xanthomonas* species.** Cytoscape
927 network illustrating transfer of cassettes between species. Nodes represent *Xanthomonas*
928 species, species containing more than 10 genomes in the dataset are coloured as in Figure 1
929 (tip labels). Line width between species is proportional to the number of cassettes moved
930 between species.

931 **Figure S1 – Phylogeny of *Xanthomonas intI*.** A maximum likelihood phylogenetic tree of all
932 *intI* in *Xanthomonas*. *IntI* from *Vibrio* sp. SCSIO 43136 (locus tag: J4N39_08275) was used as
933 an outgroup. The tree was inferred with FastTree and rooted with *Vibrio* sp. SCSIO 43136,
934 the scale bar indicates substitutions per site. * indicates the *intI*s are not integrated adjacent
935 to *ilvD*.

936 **Figure S2 – Phylogeny of *ilvD*.** The maximum-likelihood phylogenetic tree of all *ilvD* in
937 *Xanthomonas*. *ilvD* from *Vibrio* sp. SCSIO 43136 was used as an outgroup. The tree was
938 inferred with FastTree and rooted at midpoint, the scale bar indicates substitutions per site.

939 **Figure S3 – Integrons in *X. arboricola*.** The phylogeny of *X. arboricola* was built using
940 Realphy, using GCA_000972745.1 as a reference genome. The tree was rooted with
941 GCA_001908755.1 (*X. hortorum*) then the tip was removed from the tree. Tip labels show
942 pathovars, which were assigned via literature search (Table S3). The red triangle indicates
943 the probable acquisition of the *intI* variant. The scale bar indicates substitutions per site. A)
944 Indicates the presence of a complete integron or *attC* sites in the *ilvD* locus, B) indicates the
945 presence of an *intI* lacking cassettes (In0), C) indicates the presence of a CALINs (cluster of
946 *attC* sites) not in the ilvD locus, D) indicates whether IntI is predicted to be functional (full
947 length) (Table S3). E) indicates the number of *attC* sites (number of cassettes) in the
948 genome. Panel F) represents the distribution of orthologous genes among the all cassettes
949 carried by the corresponding isolate inferred with Proteinortho. Orthologous genes are
950 given in decreasing order based on the number of strains within the species that carry them.

951 **Figure S4 – Integrons in *X. translucens*.** The phylogeny of *X. translucens* was built using
952 Realphy, using GCA_001021935.1 as reference genome. The tree was rooted with
953 GCA_000007145.1 (*X. campestris*) then the tip removed from the tree. Tip labels show
954 pathovars, which were assigned via literature search (Table S3). The scale bar indicates
955 substitutions per site. A) Indicates the presence of a complete integron or *attC* sites in the
956 *ilvD* locus, B) inidicates the presence of an *intI* lacking cassettes (In0), C) indicate the
957 presence of a CALINs (cluster of *attC* sites) not in the *ilvD* locus, D) indicates whether IntI is
958 predicted to be functional (full length) (Table S3). E) indicates the number of *attC* sites
959 (number of cassettes) in the genome. Panel F) represents the distribution of orthologous
960 genes among the all the cassettes carried by the corresponding isolate interfered with
961 Proteinortho. Orthologous genes are ordered in decreasing order based on the number of
962 strains within the species that carry them.

963 **Figure S5 – Integrons in *X. euvesicatoria*.** The phylogeny of *X. euvesicatoria* was built using
964 Realphy, using GCA_000009165.1 as reference genome. The tree was rooted with
965 GCA_002759095.2 (*X. phaseoli*) then the tip removed from the tree. The scale bar indicates
966 substitutions per site. A) Indicates the presence of a complete integron or *attC* sites in the
967 *ilvD* locus, B) inidicates the presence of an *intI* lacking cassettes (In0), C) indicate the
968 presence of a CALINs (cluster of *attC* sites) not in the *ilvD* locus, D) indicates whether IntI is
969 predicted to be functional (full length) (Table S3). E) indicates the number of *attC* sites
970 (number of cassettes) in the genome. Panel F) represents the distribution of orthologous
971 genes among the all the cassettes carried by the corresponding isolate interfered with
972 Proteinortho. Orthologous genes are ordered in decreasing order based on the number of
973 strains within the species that carry them.

974 **Figure S6 – Integrons in *X. oryzae*.** The phylogeny of *X. oryzae* was built using Realphy, using
975 GCA_000007385.1 as reference genome. The tree was rooted with GCA_002759095.2 (*X.*
976 *phaseoli*) then the tip removed from the tree. Tip labels show pathovars, which were
977 assigned via literature search (Table S3). The scale bar indicates substitutions per site. A)
978 Indicates the presence of a complete integron or *attC* sites in the *ilvD* locus, B) inidicates the
979 presence of an *intI* lacking cassettes (In0), C) indicate the presence of a CALINs (cluster of
980 *attC* sites) not in the *ilvD* locus, D) indicates whether IntI is predicted to be functional (full
981 length). E) indicates the number of *attC* sites (number of cassettes) in the genome (Table
982 S3). Panel F) represents the distribution of orthologous genes among the all the cassettes
983 carried by the corresponding isolate interfered with Proteinortho. Orthologous genes are
984 ordered in decreasing order based on the number of strains within the species that carry
985 them.

986 **Figure S7 – Integrons in *X. phaseoli*.** The phylogeny of *X. phaseoli* was built using Realphy,
987 using GCA_008639345.1 as reference genome. The tree was rooted with GCA_000009165.1
988 (*X. euvesicatoria*) then the tip removed from the tree. Tip labels show pathovars, which
989 were assigned via literature search (Table S3). The scale bar indicates substitutions per site.
990 A) Indicates the presence of a complete integron or *attC* sites in the *ilvD* locus, B) inidicates
991 the presence of *intI* lacking cassettes (In0), C) indicate the presence of a CALINs (cluster of
992 *attC* sites) not in the *ilvD* locus, D) indicates whether IntI is predicted to be functional (full
993 length) (Table S3). E) indicates the number of *attC* sites (number of cassettes) in the
994 genome. Panel F) represents the distribution of orthologous genes among the all the
995 cassettes carried by the corresponding isolate interfered with Proteinortho. Orthologous
996 genes are ordered in decreasing order based on the number of strains within the species
997 that carry them.
998

999 **Figure S8 – Integrons in *X. hortorum*.** The phylogeny of *X. hortorum* was built using Realphy,
1000 using GCA_002285515.1 as reference genome. The tree was rooted with GCA_905367715.1
1001 (*X. arboricola*) then the tip removed from the tree. Tip labels show pathovars, which were
1002 assigned via literature search (Table S3). The scale bar indicates substitutions per site. A)
1003 Indicates the presence of a complete integron or *attC* sites in the *ilvD* locus, B) indicates the
1004 presence of an *intI* lacking cassettes (In0), C) indicate the presence of a CALINs (cluster of
1005 *attC* sites) not in the *ilvD* locus, D) indicates whether IntI is predicted to be functional (full
1006 length). E) indicates the number of *attC* sites (number of cassettes) in the genome (Table
1007 S3). Panel F) represents the distribution of orthologous genes among the all the cassettes
1008 carried by the corresponding isolate interfered with Proteinortho. Orthologous genes are
1009 ordered in decreasing order based on the number of strains within the species that carry
1010 them.

1011 **Figure S9 – Integron gene cassette counts in *Xanthomonas*.** Boxplot of the number of *attC*
1012 sites per genome grouped by plant of isolation and by species (only species with more than
1013 10 genomes are shown).

1014 **Figure S10 – Delta ANI values for selected events of interspecies transfer of cassettes.** A)
1015 Dot plot and B) density plot of the cassette-to-genome ANI differences (delta ANI). In A),
1016 black dots represent individual data points, while in B), the curve shows the density
1017 estimate of the distribution.
1018

1019
1020 **Table S1 – *Xanthomonas* genomes used in this study with integrons and number of**
1021 **cassettes identified.**

1022 **Table S2 – Coordinates of *intI* identified in *Xanthomonas.***

1023 **Table S3 – Selected *Xanthomonas* genomes with integrons and number of cassettes**
1024 **identified, pathovar information, and predicted *intI* functionality.**

1025 **Table S4 – Metadata associated with Selected *Xanthomonas* genomes.**

1026 **Table S5 – Identified COG categories of the 1,087 orthologous genes in non-redundant**
1027 **cassettes.**