Title

Interspecies exchange of a mobile genetic element during a plant disease outbreak

Authors

Daria Evseeva[1], Yann Pecrix[2], Marek Kucka[3], Catrin Weiler[1], Carolin Franzl[1], Markéta Vlková-Žlebková[1], Elena Colombi[4], Yingguang Frank Chan[3], Stéphane Poussier[2,5,6], Emmanuel Wicker[6], Honour McCann[1]*

Affiliations

[1] Max Planck Institute for Biology, Tübingen, Germany
[2] CIRAD UMR-PVBMT, Réunion, France
[3] Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany
[4] Macquarie University, Sydney, Australia
[5] Université de la Réunion, Réunion, France
[6] CIRAD UMR-PHIM, Montpellier, France
* Corresponding

Present address: M.K.: University of Cologne, Department of Translational Genomics, Cologne, Germany; Y.F.C.: Groningen Institute for Evolutionary Life Sciences (GELIFES), University of Groningen, Groningen, The Netherlands

Author ORCIDs

Evseeva https://orcid.org/0009-0001-6560-4572
Pecrix https://orcid.org/0000-0002-6537-3145
Kucka https://orcid.org/0000-0002-0928-4914
Franzl https://orcid.org/0000-0003-2467-4328
Vlková-Žlebková https://orcid.org/0000-0002-0272-5115
Colombi https://orcid.org/0000-0002-3753-4925
Chan https://orcid.org/0000-0001-6292-9681
Poussier https://orcid.org/0000-0003-0273-3518
Wicker https://orcid.org/0000-0003-0927-7404
McCann https://orcid.org/0000-0003-3061-337X

Keywords

Microbial evolution, disease emergence, bacterial pathogen, mobile elements, plant pathogen

Abstract

The evolutionary processes underlying disease outbreaks remain unknown for most bacterial plant pathogens. We sequenced an outbreak of lethal wilt disease in Martinique caused by two distantly related lineages in the *Ralstonia solanacearum* species complex. One lineage (*R. solanacearum* IIB-4NPB) exhibited the broadest host range ever documented for a single lineage in the field, while the other (*R. pseudosolanacearum* I-18) expanded in parallel but retained specialisation on solanaceaous hosts. Phylogenomic analysis of 407 outbreak isolates shows both lineages independently disseminated from mainland populations into Martinique. We resolved fine-scale geographic patterns of genomic diversity and identified spatial hotspots of interspecies mobile element exchange, resulting in the identification of a new family of *Ralstonia* integrative conjugative elements (ICEs) associated with the outbreak. ICE accessory gene integration sites display striking functional specialization and differentiation despite variable gene content: each site acquires only metabolism-associated or defence-related genes, respectively. This work provides insight into the origin and genomic changes associated with an outbreak of plant disease, and highlights the role of mobile elements in driving pathogen emergence.

Disease outbreaks on crop plants have severe consequences for subsistence and commercial growers alike, yet we have little understanding of the genetic mechanisms and evolutionary processes underlying pathogen emergence and adaptation to agricultural environments. Whole genome sequencing of emerging and established bacterial pathogens has led to significant new discoveries detailing the origins of outbreak lineages, dynamic selection pressure imposed by viral predation, and the role of mobile elements in the movement of antimicrobial resistance and virulence genes[1–4]. Despite recent notable exceptions, most studies focus on mammalian pathogens and it is unclear how generalizable their findings are to plant pathogens[5–10]. Diseases can spread quickly across vast areas used for food, forage and industrial crops, with severe consequences for agricultural productivity and economic stability. Clarifying the mechanisms underlying disease emergence in plant hosts is of both fundamental importance and practical value, and deepens our understanding of how microbes adapt to new hosts and environments.

The identification of a novel lineage of *Ralstonia solanacearum* in the Caribbean island of Martinique represents a rare opportunity to investigate the emergence of a bacterial plant pathogen of global economic importance[11]. Members of the *R. solanacearum* species complex (RSSC) are known to cause lethal infections in many different host species, including staple and commodity crops[12–14]. The species complex is comprised of *R. solanacearum* (Rso, also known as phylotype II and further subdivided into Rso IIA and IIB), *R. pseudosolanacearum* (Rps, phylotype I and III), and R. syzygii (Rsy, phylotype IV)[15,16]. Lineages identified within each species or phylotype were assigned to over 56 sequevars using sequence typing of the endoglucanase (*egl*) gene[17,18]. Since the pathogen disseminates rapidly and can persist for years in freshwater and soil, the pathogen is subject to extensive monitoring and movement controls to limit its spread. Clonally propagated staples like potato and banana are particularly threatened by emergence events, since pathogens can spread rapidly in areas with high host density and genetic homogeneity[19,20]. The threat posed to potato cultivation is so great that one Rso lineage is classified as a Select Agent in the United States[21]. Multiple lineages of *R. solanacearum* are known to infect banana, which is cultivated on an industrial scale in the Americas[8]. An outbreak of disease on a cultivated ornamental (*Anthurium andreanum*) in Martinique in 1999 caused alarm when it was found to be caused by a lineage closely related to the banana-infecting lineage Rso IIB sequevar 4 (IIB-4)[11]. Additional infections on the island were soon reported in many hosts which *R. solanacearum* had not been previously known to infect. The threat posed to banana cultivation and apparent host range expansion led to an extensive sampling survey of *Ralstonia* populations across Martinique, as well as coastal and inland areas of French Guiana on the South American mainland[23,24]. This work generated a remarkable collection of strains isolated from infected crops, asymptomatic weeds, soil and freshwater environments during a disease emergence event characterized by the rapid and unprecedented expansion of *Ralstonia* spp. infections across a broad range of hosts and geographic areas in Martinique[11]. Multi-locus sequence analysis (MLSA) showed both Rps and Rso isolates were recovered, but the origins, diversity and genomic changes accompanying the emergence of the host-range expanded lineage remain unclear. Since this collection provides a unique opportunity to explore the disease origins and investigate the spatial population structure of isolates collected from cultivated crops, wild hosts and environmental sources, we sequenced the largest collection of *Ralstonia* spp. ever sampled during an outbreak event.

Our study of 409 new genomes, paired with 407 publicly available genomes, provides unparalleled phylogenetic and spatial resolution of an extraordinarily well sampled outbreak of *Ralstonia* spp. in the Caribbean island of Martinique. Earlier MLSA studies showed two different lineages (Rso IIB-4NPB and Rps I-18) were circulating during the outbreak event; we show here each lineage was disseminated from well-established populations on the South American mainland, though one lineage is of more distant Asian origin and retains a signature of host specialisation. The lineage that underwent an apparent host range expansion arose from a population that has been diversifying in the Americas for much longer. Although each lineage is a member of a different species, we found mobile elements called integrative and conjugative elements (ICEs) are circulating between them, particularly in geographic areas where lineages co-occur. The ICEs have two cargo gene insertion hotspots that exhibit consistent functional differentiation: one hotspot is a target for metabolic gene acquisition, while the other is a hotspot for defense element acquisition. In contrast, there were very few changes in the repertoire of secreted virulence proteins linked with pathogenesis and host defense manipulation. These results challenge the view plant pathogen emergence is driven by virulence factor acquisition or innovation, highlighting instead the critical roles of horizontal transfer, microbial competition and metabolic flexibility in shaping apparent host range and ecological success. The expansion of Rso IIB-4NPB in particular, which exhibits the broadest documented host range of any Ralstonia lineage, was more strongly driven by the acquisition of mobile elements and expanded anti-phage defences than by virulence gene acquisition. As agricultural intensification continues to reshape plant-microbe interactions, pathogen lineages that can rapidly acquire and integrate novel genetic material may rapidly evolve to exploit novel niches and hosts. Surveillance strategies should therefore consider not just the pathogen, but the community and network of genetic exchange that can fuel its evolution.

Results

Global phylogeography of the *Ralstonia solanacearum* species complex

The global phylogeography of the *Ralstonia solanacearum* species complex (RSSC) reveals patterns of long-distance dissemination and epidemic expansion. Our analysis of 409 new whole genome sequences from *Ralstonia* isolated from crops, weeds, soil and freshwater environments during the outbreak paired with 407 genomes available on NCBI shows that diverse populations of both *R. solanacearum* (Rso) and *R. pseudosolanacearum* (Rps) are circulating in the Americas [11,23,24]. The outbreak documented in Martinique was caused by the expansion of two lineages from each species: Rps I-18 and Rso IIB-4NPB (Fig. 1, Fig. S1), consistent with the MLSA results in Wicker et al. 2007. The center of origin of Rso and Rps is in the Americas and Asia, respectively[12]. Multiple inter-continental dissemination events of both Rso and Rps lineages are evident in the tree, including a global outbreak of Rso IIB-1 on potato (*S. tuberosum*) and multiple intercontinental introductions of at least eight distinct Rps lineages from Asia into the Americas (Fig. 1, Fig. S1). Figure 1 shows the Rps lineages introduced to the Americas are divergent (with the exception of Rps I-14 and I-18) and provides phylogenetic context for their distinct origins, placing them in the context of the diversity of Asian Rps. The introduction of Rps lineages into the Americas resulted in their establishment and local diversification, primarily in solanaceous hosts.

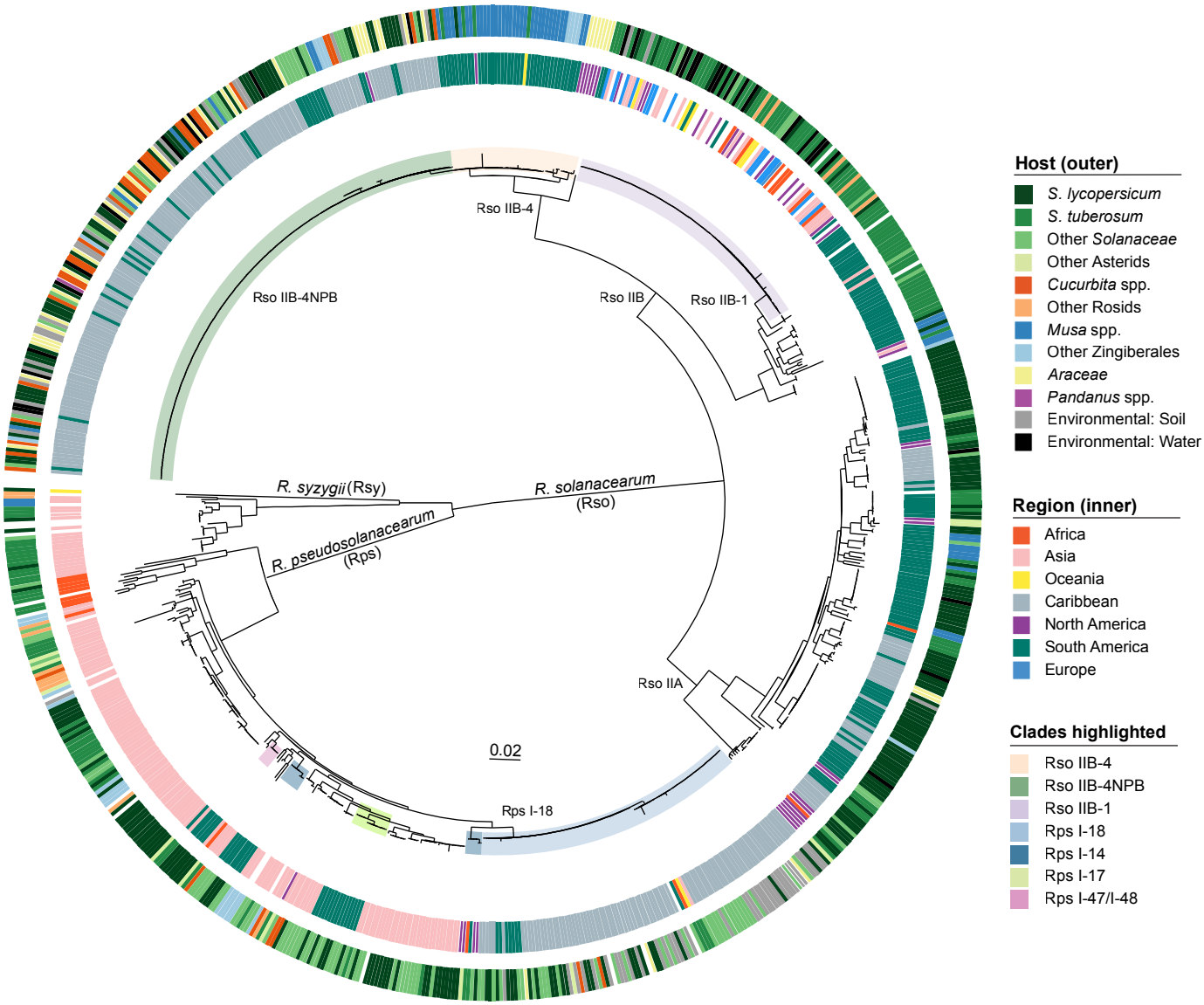
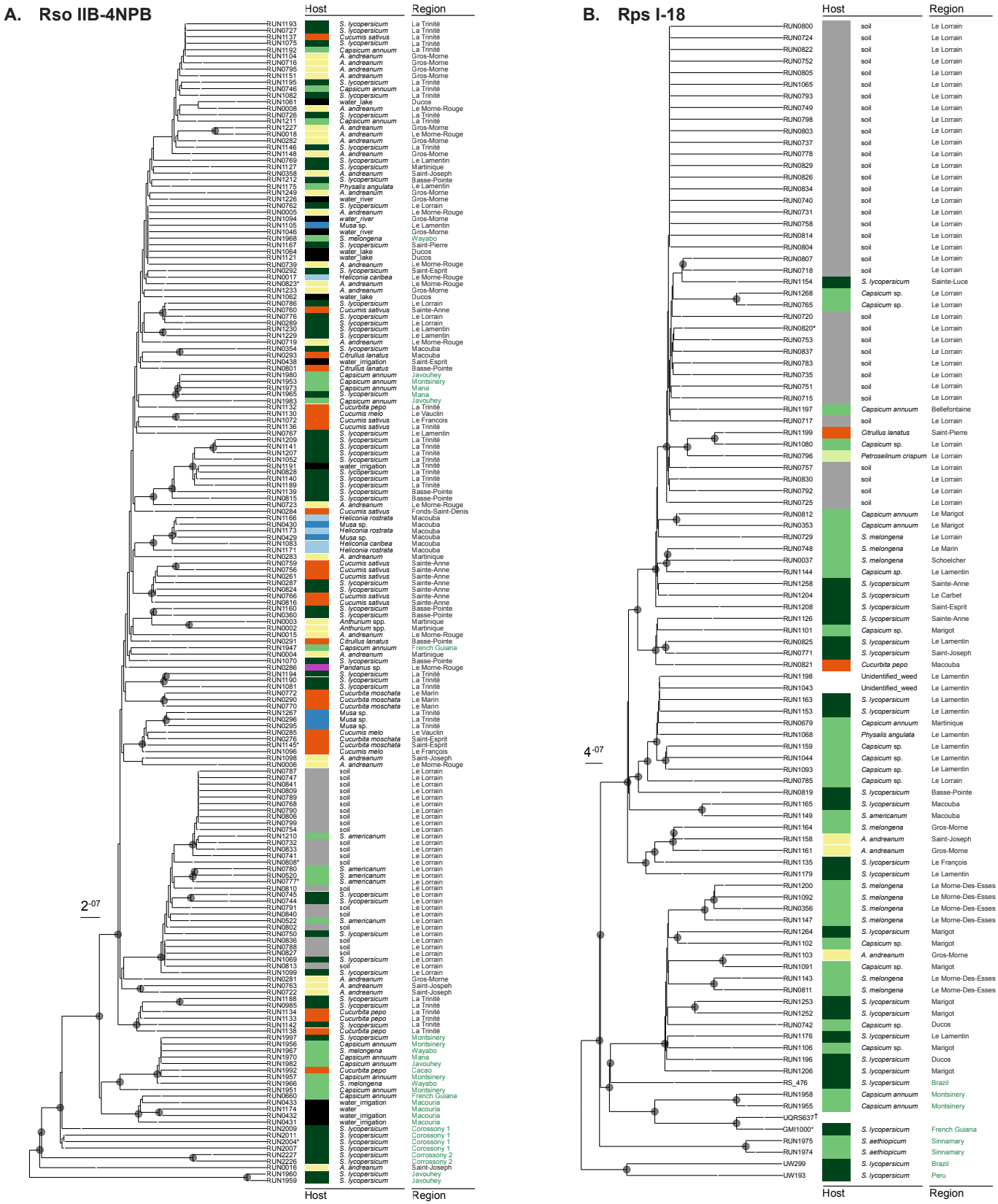


Fig. 1. Core gene tree of the *Ralstonia solanacearum* species complex. Maximum likelihood tree of 816 RSSC genomes inferred from a 137,423bp SNP alignment obtained after core gene identification, alignment generation and invariant site removal. Nodes with the bootstrap support values under 70 are collapsed. Clades representing Rso and Rps lineages of interest are highlighted in the tree. The geographic region and host of isolation are shown in the first and second rings, respectively (A and B). Martinique isolates are a subset of Caribbean isolates, and French Guiana isolates are a subset of South American isolates. With the exception of *S. lycopersicum* (tomato) and *S. tuberosum* (potato), the host of isolation for each isolate is shown collapsed to family level. Lower level taxonomic classifications within the same family are indicated with greater colour saturation. Missing data is shown in white.

128 Rps I-18 is widely distributed across Martinique, French Guiana, Brazil and Peru (Fig. 1, Fig. S1). This
129 lineage shares a recent common ancestor with Rps I-14, which is present in Guatemala, Costa Rica and
130 Colombia. At least three additional Rps lineages of Asian origin are present in the Americas: Rps I-13, Rps
131 I-14, Rps I-17 and Rps I-47/48. Although a single dissemination event could be responsible for the
132 introduction of a diverse pathogen population, it is more likely multiple independent introductions of
133 Rps occurred. The presence of Rps I in the Americas has been documented as early as 1967[25], yet the
134 genomic epidemiology work shown here provides the first clear evidence of the distinctly different
135 origins and fates of multiple intercontinental dissemination events from Asia into the Americas.
136
137 In contrast to Rps, Rso has been diversifying in the Americas for some time. Although two lineages (Rso
138 IIB-4NPB and Rps I-18) expanded in Martinique and French Guiana, it is the Rso IIA isolates that
139 encompass the greatest phylogenetic diversity. Rso IIA, primarily isolated from tomato (68%), is broadly
140 distributed across the Americas, though there is no evidence of recent epidemic expansion (Fig. 1, Fig.
141 S1). Rso IIB lineages are responsible for both regional and global outbreak events however, most notably
142 on banana and potato. Rso IIB-1 (formerly race 3 biovar 2 or R3bv2), responsible for a global outbreak
143 on potato, has also been isolated from other *Solanum* spp., *Pelargonium* spp. (formerly *Geranium* spp.),
144 surface water and soil. Although the centre of origin of Rso IIB-1 appears to be in South America, its
145 rapid global expansion suggests international trade in seed potatoes is the likeliest mechanism of long-
146 distance dissemination, highlighting the elevated risk of disease transmission during movement of
147 infected plant material[14,19,26,27]. Outbreaks of Rso on banana are well documented in the Americas,
148 where it is referred to as Moko disease. Moko disease is clearly polyphyletic, as outbreaks were caused
149 by four Rso IIA and three Rso IIB lineages, including basal members of Rso IIB-4 (Fig. 1)[28,29]. These
150 banana-infecting lineages are responsible for multiple regional outbreaks in Panama, Costa Rica,
151 Colombia, Peru, Brazil and French Guiana from at least the 1960s onwards (Fig. S2). Since banana was
152 only introduced to the Americas as a crop in the 16th century[30], it is a matter of some interest that the
153 most deeply branching basal members of Rso IIB are isolated from *Musa* spp. along with *Heliconia* spp.
154 and *Anthurium* spp., both endemic to the Americas (Fig. 1, Fig. S2)[31,32]. The expansion of high-density
155 single variety banana monocultures likely increased opportunities for the pathogen to emerge and
156 spread in agricultural areas, in turn increasing likelihood of spread to other crop varieties, both endemic
157 and introduced.

158 **Epidemic expansion of two lineages after their dissemination from South America**

159 The species complex tree confirms the surge in plant infections in Martinique is due to the recent clonal
160 expansion of two lineages across the island: Rso IIB-4NPB and Rps I-18 (Fig. 1). Rso IIB-4NPB is of
161 particular interest since it was isolated from the most diverse group of naturally infected hosts that has
162 ever been documented for *Ralstonia* to date. Rps I-18 expanded in tandem with Rso IIB-4NPB, however
163 members of this lineage originally only caused infections of solanaceous hosts in the field[11,23]. We
164 investigated the origins of both lineages using well resolved, non-recombinant core genome trees
165 including closely related strains from mainland populations (Fig. 2).



Host

- *S. lycopersicum*
- Other *Solanaceae*
- *Other Asterids*
- *Cucurbita* spp.
- *Musa* spp.
- Other Zingiberales
- *Araceae*
- *Pandanus* spp.
- Environmental: Soil
- Environmental: Water

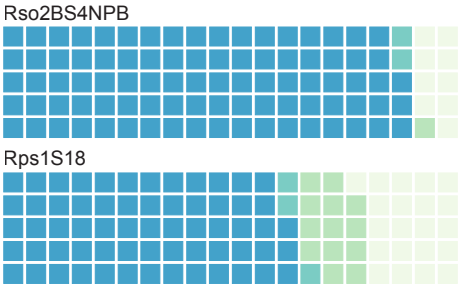
Region (text colour)

- Caribbean
- South America

Pangenome

- Core genes
- Soft core genes
- Shell genes
- Cloud genes

C. Pangenome fraction



—

Fig. 2. Outbreak lineage core genome trees. (A) Rso IIB-4NPB maximum likelihood tree calculated from a 5,206,677 bp non-recombinant non-mobile core genome alignment with 195 SNPs and 73 parsimony informative sites. Internal nodes with bootstrap value above 80 are indicated with black circles. Strain labels shown in bold are are complete or near complete assemblies generated in this study and/or used as mapping references. (B) Rps I-18 maximum likelihood tree calculated from a 4,760,272 bp non-recombinant core genome alignment with 311 SNPs and 116 parsimony informative sites. The dagger for UQRS637 indicates this assembly was later suppressed on NCBI by the uploader. Asterisks next to strain labels indicates closed genomes generated in this study and/or used as mapping references. (C) Pangenome waffle chart representing the proportion of core (present in 99 - 100% of genomes), soft-core (95 - 99% of genomes), shell (15 - 95% of genomes) and cloud genes (present in less than 15% of genomes) identified by Panaroo, displayed separately for each lineage pangenome.

—

Basal members of the Rso IIB-4NPB tree were mostly isolated from solanaceous hosts or freshwater sources on the South American mainland, in French Guiana (Fig. 2, Fig. S3). The basal group of isolates likely represents the ancestral mainland population from which Rso IIB-4NPB emerged to sweep across Martinique, since only one of these was isolated on Martinique (RUN0016). There was limited spread of the derived clade on the mainland: only six were isolated from solanaceous plants in French Guiana while the remaining 149 were isolated in Martinique during the epidemic. Although a mainland origin for Rso IIB-4NPB (and indeed Rps I-18) is most likely, it is possible there are more closely related ancestral populations that remain unsampled. The derived group of 155 isolates exhibits limited genetic diversity, differing by only 166 nonrecombinant SNPs in the core genome. Curiously, both basal and derived members of Rso IIB-4NPB were isolated from *Anthurium andreanum* in Saint Joseph, Martinique, suggesting this location and host may be linked with the initial introduction of the emergent lineage to the island (Fig. 2). Notably, Rso IIB-4NPB was also isolated from symptomatic banana in Trinité, indicating the appellation 'Non-Pathogenic on Banana' (NPB) for the lineage may not be wholly accurate.

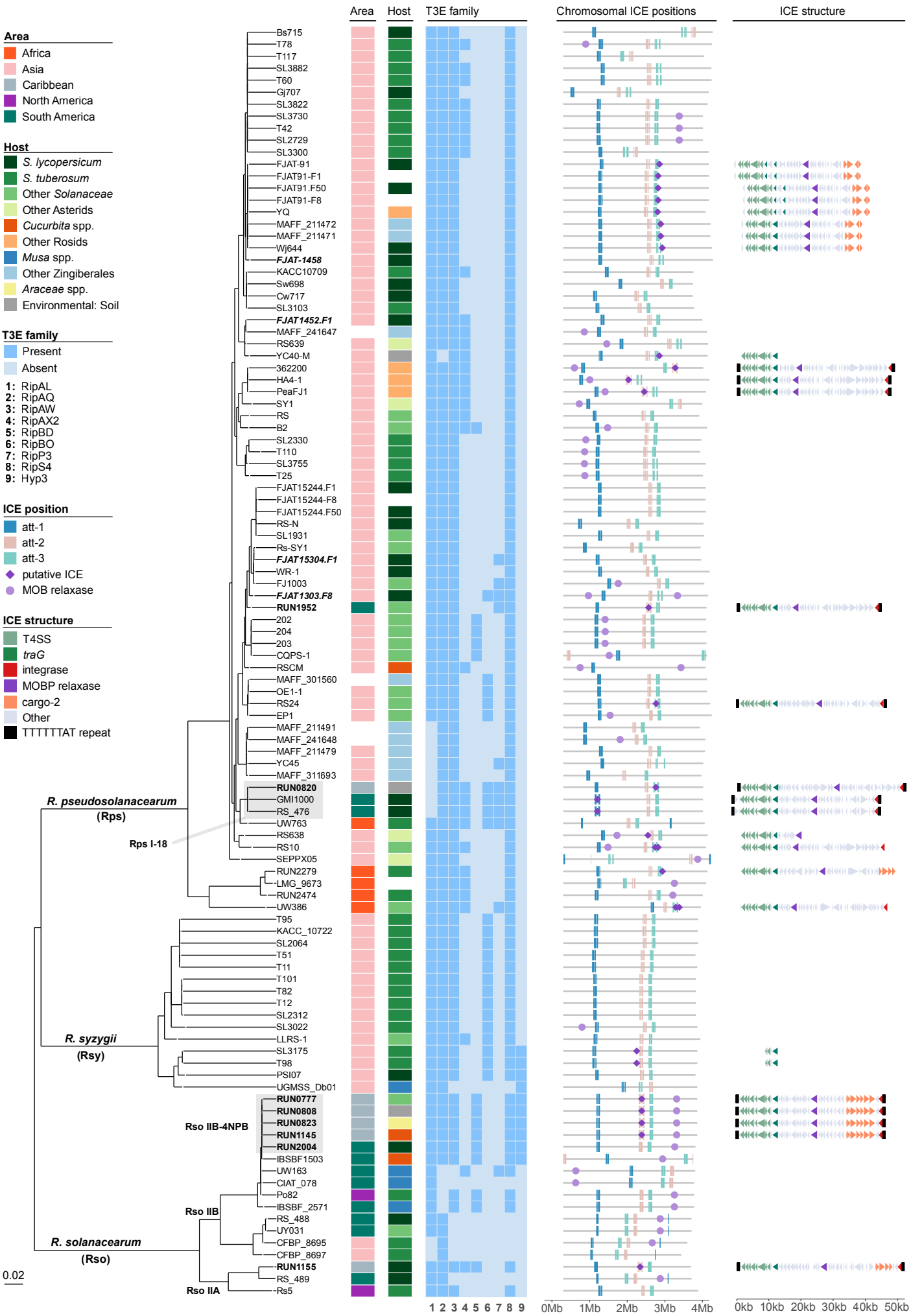
The evolutionary history of Rps I-18 parallels that of Rso IIB-4NPB: Martinique Rps I-18 is similarly derived from a broadly distributed population on the mainland (Fig. 2, Fig. S4). Despite the large number of sequenced Rps I-18 isolates, the lineage displays limited diversity in the core genome and T3E repertoire, consistent with recent epidemic expansion. In contrast to Rso IIB-4NPB, there is no evidence of expanded host range: with few exceptions, the majority of Martinique Rps I-18 was isolated from soil or solanaceous hosts (Fig. S5). There is evidence of spatial population structure for Rps I-18 and Rso IIB-4NPB alike: Le Lorrain Rso IIB-4NPB and Rps I-18 isolates form distinct and well supported clades in their respective trees, regardless of whether they were isolated from soil, cultivated or wild *Solanum* spp.

We used these well-resolved non-recombinant trees to determine the divergence time of Rps I-18 and Rso IIB-4NPB. BactDating showed there is some correlation between the sampling date and root-to-tip distance for Rso IIB-4NPB (R^2=0.16, p<0.0001) and permutation analysis showed the estimated divergence time outperformed random sampling in 77.6% of simulations. Though weakly supported, the estimated divergence time for the split between the basal mainland and derived Martinique population of Rso IIB-4NPB is between 1930 -1970 (Fig. S6). Rps I-18 exhibits little temporal signal (R^2=0.05, p=0.04),

210 yet date permutation analysis showed the estimated divergence time outperformed random sampling in
211 100% of simulations. The estimated divergence time for Rps I-18 is earlier, between 1870 - 1940 (Fig S7).
212 During pangenome analyses we found both Rps I-18 and Rso IIB-4NPB have similar numbers of core and
213 soft-core genes (4,565 for Rps I-18 and 4,792 for Rso IIB-4NPB), but different accessory genome sizes:
214 487 accessory genes in Rso IIB-4NPB and 2,323 in Rps I-18. While it is possible the difference in
215 accessory genome sizes (amounting to 10% and 35% of the Rso IIB-4NPB and Rps I-18 pangenome,
216 respectively) may be caused by differences in the rate or maintenance of horizontal acquisition events,
217 this may also reflect different introduction times to Martinique. The data cumulatively indicates Rps I-18
218 was introduced to Martinique prior to Rso IIB-4NPB.

219 **Ancestral changes in virulence and metabolic gene repertoires**

220 Despite the reduced genetic diversity and recent emergence of Rso IIB-4NPB, derived members of this
221 lineage have been isolated from the widest variety of hosts during natural infections and display the
222 broadest documented host range of any single lineage in the species complex (Fig. 2). Since the delivery
223 of virulence proteins by the Type 3 secretion system is known to play an important role in successful
224 host colonisation and the repertoire of Type 3 secreted effectors (T3Es) in turn is thought to play an
225 important role in host specificity[33–35], we identified the T3E repertoire across the entire dataset to
226 determine whether the gain or loss of specific T3Es is linked with the emergence of Rso IIB-4NPB (Fig. 3,
227 Fig. S1, Table S4). Rso IIB-4NPB genomes display nearly uniform T3E repertoires, consistent with the
228 recent expansion of this lineage, with no lineage-specific T3E changes compared to Rso IIB-4 (Fig. S2).
229 Multiple T3E changes occurred in both the common ancestor of Rso IIB-4NPB and Rso IIB-4, and in Rso
230 IIB-4 alone. For example, a single T3E is uniquely present in Rso IIB-4 and Rso IIB-4NPB, but absent in all
231 other RSSC genomes: the hypothetical effector *T3E_Hyp13* (UW163_19700 in Rso IIB-4 UW163,
232 OFBKGH_20220 in Rso IIB-4NPB RUN1145) (Fig. S1, Table S6). Another T3E (*ripAX1*) is sufficiently
233 divergent from other members of the effector family to be classified as a clade-specific gene by Panaroo
234 pangenome analysis. Two effectors are present in Rso IIB-4 and absent in Rso IIB-4NPB: *ripAX2* and *ripP3*
235 *(*(Fig. S1, Fig. S2, Table S6). The effector *ripAX2* has a broad distribution across Rso IIA and IIB genomes
236 and was therefore likely lost in Rso IIB-4NPB, while *ripP3* has a patchy distribution and was likely
237 acquired by Rso IIB-4. Four effectors appear to have been lost in basal Rso IIB-4: *ripAQ*, *ripAW*, *ripBD* and
238 the predicted *T3E_Hyp3* effector. The loss of *ripAQ* is most striking as the family is otherwise present in
239 all other RSSC strains. Since these effectors are present among related clades and/or across the species
240 complex, the most parsimonious explanation for the pattern observed is they were lost in the ancestor
241 of banana-infecting Rso IIB-4 strains rather than being acquired by ancestral Rso IIB-4NPB. In sum, the
242 Rso IIB-4 isolated from monocot hosts (*Heliconia* spp., banana and *Epipremnum* spp.) exhibit more
243 variation in effector repertoire than Rso IIB-4NPB (Fig. 3, Fig. S2) and contain more unique variants
244 statistically associated with the host of isolation (Supplementary Results p.16).
245
246 We investigated non-T3E acquisitions in the ancestor of Rso IIB-4NPB, and found 84 genes uniquely
247 acquired in the ancestor of both Rso IIB-4 and Rso IIB-4NPB (Table S3). These genes encode enzymes
248 such as a TauD/TfdA family dioxygenase, shown to degrade phenoxy herbicides in soil-borne bacteria
249 like *Burkholderia* spp. and *Cupriavidus necator* (formerly *Ralstonia eutropha*)[36–41]; a CxxC motif protein


Area
Host
T3E family
Chromosomal ICE positions
ICE structure
Area
Africa
Asia
Caribbean
North America
South America
Host
S. lycopersicum
S. tuberosum
Other Solanaceae
Other Asterids
Cucurbita spp.
Other Rosids
Musa spp.
Other Zingiberales
Araceae spp.
Environmental: Soil
T3E family
Present
Absent
1: RipAL
2: RipAQ
3: RipAW
4: RipAX2
5: RipBD
6: RipBO
7: RipP3
8: RipS4
9: Hyp3
ICE position
att-1
att-2
att-3
putative ICE
MOB relaxase
ICE structure
T4SS
traG
integrase
MOBP relaxase
cargo-2
Other
TTTTTTAT repeat
R. pseudosolanacearum
(Rps)
Rps I-18
R. syzygii
(Rsy)
R. solanacearum
(Rso)
Rso IIB-4NPB
Rso IIB
Rso IIA
0.02
Bs715
T78
T117
SL3882
T60
Gj707
SL3822
SL3730
T42
SL2729
SL3300
FJAT-91
FJAT91-F1
FJAT91.F50
FJAT91-F8
YQ
MAFF_211472
MAFF_211471
Wj644
FJAT-1458
KACC10709
Sw698
Cw717
SL3103
FJAT1452.F1
MAFF_241647
RS639
YC40-M
362200
HA4-1
PeaFJ1
SY1
RS
B2
SL2330
T110
SL3755
T25
FJAT15244.F1
FJAT15244-F8
FJAT15244.F50
RS-N
SL1931
Rs-SY1
FJAT15304.F1
WR-1
FJ1003
FJAT1303.F8
RUN1952
202
204
203
CQPS-1
RSCM
MAFF_301560
OE1-1
RS24
EP1
MAFF_211491
MAFF_241648
MAFF_211479
YC45
MAFF_311693
RUN0820
GMI1000
RS_476
UW763
RS638
RS10
SEPPX05
RUN2279
LMG_9673
RUN2474
UW386
T95
KACC_10722
SL2064
T51
T11
T101
T82
T12
SL2312
SL3022
LLRS-1
SL3175
T98
PSI07
UGMSS_Db01
RUN0777
RUN0808
RUN0823
RUN1145
RUN2004
IBSBF1503
UW163
CIAT_078
Po82
IBSBF_2571
RS_488
UY031
CFBP_8695
CFBP_8697
RUN1155
RS_489
Rs5
1 2 3 4 5 6 7 8 9
0Mb 1Mb 2Mb 3Mb 4Mb
0kb 10kb 20kb 30kb 40kb 50kb

251 —
252 **Fig. 3. Virulence gene and ICE distribution across the *R. solanacearum* species complex.** Maximum likelihood tree
253 of all completely assembled RSSC genomes available in this study (n=105). Strain labels shown in bold are complete
254 or near complete assemblies generated in this study and/or used as mapping references. Strain labels shown in
255 boldface italics represent clades collapsed due to low diversity. The geographical area and host of isolation are
256 shown next to strain labels, followed by nine *Ralstonia* T3E families exhibiting presence/absence variation for Rso
257 IIB-4 and Rps I-18 clades. Dark blue in the T3E family column represents the presence of the T3E family in a strain
258 and light blue represents its absence. Chromosomal ICE position displays the position of the ICE attachment sites
259 on the main chromosome of each RSSC genome – no ICEs were identified on the megaplasmid in any complete
260 genome. The attachment sites are here defined as regions syntenic to the 50kb up- and downstream flanking
261 regions of the Rps I-18 GMI1000 (att-1, blue), Rso IIB-4NPB RUN1145 (att-2, pink) and Rso IIB-4NPB RUN1960 (att-
262 3, green) attachment sites. Thin grey lines represent the length ranges of the chromosomes. MOB relaxases
263 occurring in isolation (purple circles) are separately indicated from those adjacent to T4SS genes and forming
264 putative ICEs (purple diamonds). The final panel shows the organization of putative ICEs identified in RSSC strains,
265 with homologs of various functional categories highlighted in different colours. RS10 and UW386 both have two
266 ICEs inserted in att-3. Black bars around ICEs represent the TTTTTTAT repeat motif delimiting putative ICE
267 boundaries.
268 —
269
270 (UW163_01710, OFBKGH_14170) associated with thiol-disulfide oxidoreductase activity and oxidative
271 damage responses[42,43]; an electron acceptor flavodoxin (UW163_18870, OFBKGH_23185) also involved
272 in oxidative damage response[44]; and a carboxymuconolactone decarboxylase (UW163_18885,
273 OFBKGH_23170) in the protocatechuate catabolism pathway, linked with the degradation of both lignin
274 precursors[45] and the conversion of lignin-derived aromatic compounds to tricarboxylic acid
275 intermediates like succinate and malate[46,47]. In sum, although the function of many horizontally acquired
276 genes remains unclear, some ancestral acquisitions appear to be linked with the catabolism of aromatic
277 compounds in soil or plant tissues.

278 **Integrative and conjugative elements circulating between co-occurring pathogen lineages**

279 Since the co-occurrence of two pathogen lineages in the same host or niche provides opportunities for
280 homologous recombination, we then explored whether Rps I-18 and Rso IIB-4NPB were exchanging
281 genetic material. We found no evidence of homologous recombination between Rso IIB-4NPB and Rps I-
282 18 (Fig. S8). RECOPHY identified some regions with reduced (< 10/kb) and elevated (120/kb) SNP density
283 compared to the alignment-wide average of 53 SNPs/kb, yet distance trees for those regions display the
284 same topology as the core genome tree, indicating homologous recombination between these lineages
285 is unlikely to have occurred[4748]. ClonalFrameML analysis indicated that although the ratio of
286 recombination to mutation in both lineages is low (R/theta = 0.159 for Rps I-18 and 0.152 for Rso IIB-
287 4NPB), homologous recombination is occurring with unknown donors.
288
289 Mobile elements constitute another means by which genetic material can be exchanged between
290 bacterial pathogens. Pangenome exploration and association testing alike revealed the presence of
291 integrative and conjugative elements (ICEs) present in co-occurring Rps I-18 and Rso IIB-4NPB, but
292 absent in Rso IIB-4 (Fig. 3). All but one of the 27 genes associated with isolation in Martinique by

293 SCOARY are located on an ICE, while a PySEER kmer-based association test similarly showed 625 out of
294 2,367 Martinique-associated kmers and all 83 soil-associated kmers map to the same ICE (Fig. S9, Table
295 S10, Table S11, Table S12). A clear association between isolation from Trinité and Morne Rouge and a
296 cluster of eight ICE genes was identified, and another two ICE gene clusters display overlapping
297 associations with isolation from soil and Le Lorrain (Fig. S9, Fig. S10, Fig. S11). The gene clusters linked
298 with specific locations of isolation in Martinique correspond to ICE backbone genes and cargo gene
299 modules .
300
301 Since the emergent Rso IIB-4NPB lineage acquired an ICE absent in Rso IIB-4, we predicted ICEs in all 133
302 complete *Ralstonia* genomes (Fig. 3). We identified a total of 30 ICEs, of which 17 lack direct repeats but
303 satisfy other criteria for ICE identification (presence of T4SS gene clusters, a *mob* relaxase and integrase
304 sequence). Some genomes have *mob* relaxase genes flanked by direct repeats but lack an adjacent T4SS:
305 these may be degenerate ICEs or integrative mobilizable elements (IMEs). Each genome typically carries
306 a single ICE integrated into one of three preferred chromosomal integration sites, though UW386 and
307 RS10 carry two ICEs. All ICEs identified in complete RSSC genomes are located on the main chromosome,
308 none were found on the megaplasmid. Most ICEs were identified in *R. pseudosolanacearum* (23/30
309 ICEs). Only two were found in in R. syzygii, and five in *R. solanacearum*. All five ICEs in *R. solanacearum*
310 are from Martinique isolates: four Rso IIB-4NPB and one Rso IIA isolate (RUN1155, Fig. 3). Strikingly, the
311 Rso IIB-4NPB and Rso IIA ICEs are identical to each other and homologous to a 45.9kb ICE present in
312 eight Rps genomes that collectively span the diversity of *R. pseudosolanacearum*. Since this ICE is
313 broadly distributed among Rps genomes and also present in the reference strain Rps I-18 GMI1000, we
314 named it $ICE^{RpsGMI1000}$. $ICE^{GMI1000}$ has two cargo regions of accessory gene integration we call 'cargo-1' and
315 'cargo-2', according to their integration site within the ICE (Fig. 4, Fig. S12).
316
317 We assessed the spatial structure of ICE carriage among the entire set of 7 complete and 275 draft
318 genomes representing the densely sampled Martinique population and found local hotspots of ICE
319 exchange in Martinique (Fig. 4). While there are multiple regions in which both lineages co-occur, it is
320 rare for these divergent lineages to share identical ICEs (Fig. S13). At a global scale, ICE carriage appears
321 more common among *R. pseudosolanacearum* than *R. solanacearum*, where we have only found them
322 in lineages circulating in Martinique: Rso IIB-4NPB and Rso IIA. ICEs can be gained and lost quite rapidly
323 however, and intensive sampling and sequencing of each lineage in Martinique shows that while ICE
324 carriage is nearly ubiquitous among Rso IIB-4NPB, Rps I-18 carriage is variable. There is a striking hotspot
325 of lineage and ICE co-occurrence in Le Lorrain: Rps I-18 and Rso IIB-4NPB isolated from the soil in this
326 region carry ICEs identical in their backbone and cargo-1 genes, but variable in the cargo-2 genes (Fig. 4,
327 Fig. S13, Fig. S14). There are additional instances of identical $ICE^{RpsGMI1000}$ carriage in Rps I-18 and Rso IIA
328 in Martinique, but none on the mainland. The presence of spatial hotspots of horizontal gene transfer in
329 the environment, where mobile element exchange appears more frequently between co-occurring
330 lineages, is mirrored by the presence of recombination hotspots within the mobile elements themselves.
331 While cargo gene clusters vary between different ICEs, their overall function does not: cargo-1 gene
332 clusters always encode metabolism-associated genes and cargo-2 gene clusters always encode bacterial
333 defence systems (Fig. S14).

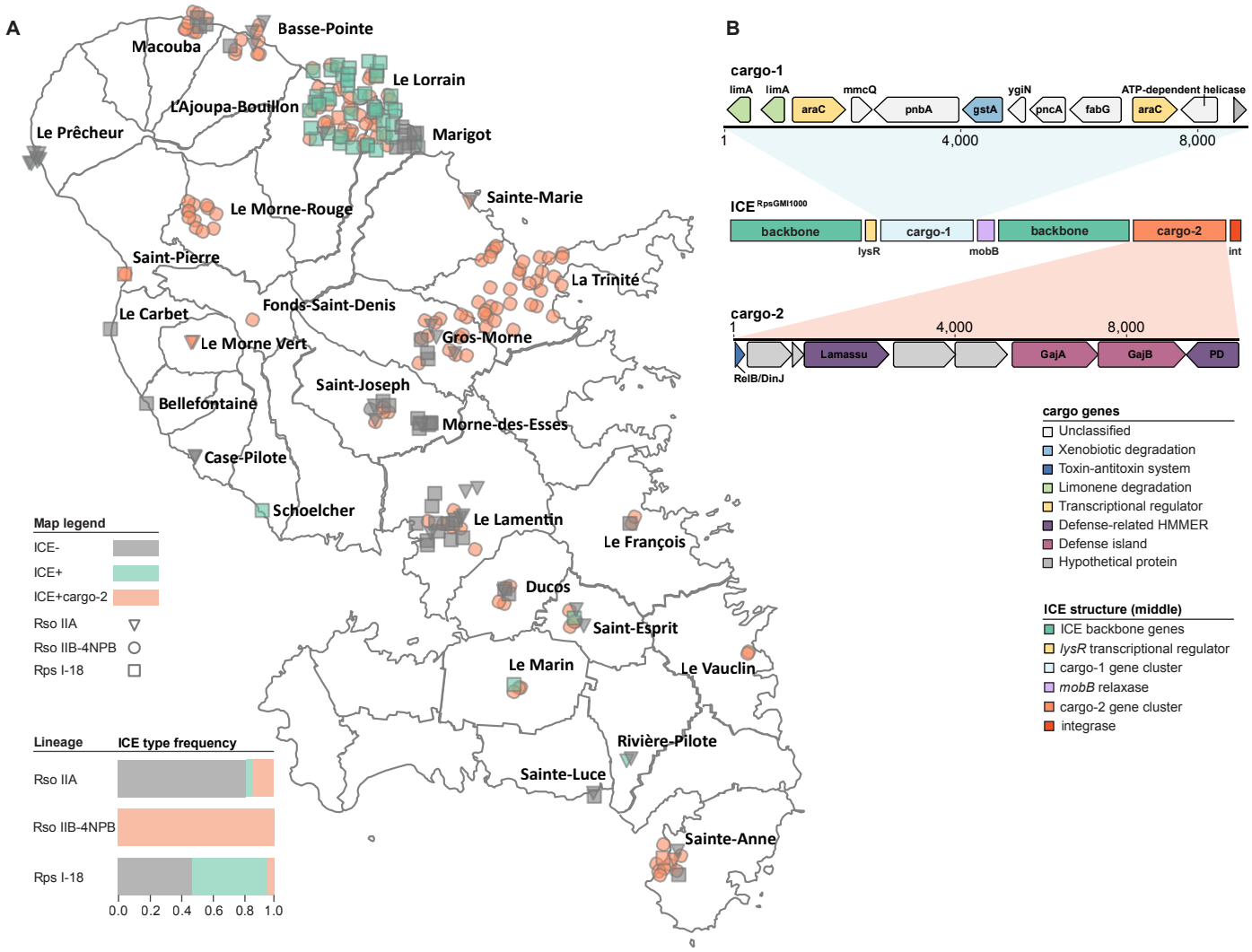


Fig. 4. *Ralstonia* ICE distribution and cargo gene carriage in Martinique. (A) Sampling location of all Martinique isolates sequenced in this study. The exact longitude and latitute of the sampling locations were jittered for visualisation purposes. Rso IIA (triangle), Rso IIB-4NPB (circle), Rps I-18 (square) are represented with grey (ICEless), green (ICE ^{GMI1000} with conserved cargo-1 and different cargo-2) or orange (ICE ^{GMI1000} with conserved cargo-1 and cargo-2) icons. The freuquency of strains carrying each ICE type is shown inset. (B) A diagram of ICE^{GMI1000} structure is shown in the top right, delineating the position of backbone and cargo gene clusters. The coding sequences are shown above (cargo-1) and below (cargo-2) for the conserved ICE, coloured according to KEGG annotation. The first backbone region includes the T4SS and has a *lysR* transcriptional regulator at the 3' end, followed by cargo-1. The second backbone region includes the mob relaxase, *traF*, *parA*, *repA*, *qseC* and *rdfS*, along with hypotheticals and helix-turn-helix and various DUF domain-containing proteins.

335
336 The metabolic cargo-1 content is identical in both Rps I-18 and Rso IIB-4NPB ICEs (e.g. RUN1145,
337 RUN0777 and RUN1960 includes two *limA* genes encoding limonene 1,2-epoxide hydrolases[48]
338 (OFBKGH_09550, OFBKGH_09555 in RUN1145); a phthalate ester hydrolase *pnbA*; an antibiotic
339 biosynthesis monooxygenase *ygiN* (also known as ABM, OFBKGH_09580); a glutathione S-transferase
340 *gstA* linked with the detoxification of organochlorine compounds (1-chloro-2,4-dinitrobenzene) used as
341 an herbicide and pesticides; a pyrazinamidase/nicotinamidase *pncA*, which converts nicotinamide to
342 nicotinic acid, and *fabG*, which encodes a reducing agent of ethyl 4-chloroacetoacetate, an intermediate
343 used in herbicide and pesticide production. Rps I-18 and Rso IIB-4NPB ICEs carry distinct cargo-2 clusters
344 encoding multiple defence systems, such as anti-plasmid/phage Lamassu and anti-viral Gabija
345 islands[49,50]. The only other ICE present in a Rso genome (Rso IIA RUN1155, also from Martinique) shares
346 a similar cargo-2 region with Rso IIB-4NPB ICE^{GMI1000} (Fig. 3, Fig. S14).

347 **Surge of prophage acquisition in the emergent lineage**

348 We found Rso IIB-4NPB acquired three *Caudoviricetes* prophages, one of which is uniquely present in
349 Rso IIB-4NPB and no other sequenced RSSC genome (Table S4, Table S7). The two other prophages are
350 present in other Rso circulating in Central and South America. All complete Rso IIB-4NPB genomes
351 (RUN0777, RUN0823, RUN0808, RUN1145 and basal RUN2004) have the three intact (and one partial)
352 prophage insertion on the chromosome, while the four complete Rso IIB-4 (UW163, Po82, IBSBF1503
353 and CIAT_078) have at most a single intact prophage on the chromosome and megaplasmid each.
354 Panaroo pangenome exploration and Pyseer kmer-based association tests alike uncover the Rso IIB-
355 4NPB-unique prophage and a downstream auxiliary region (Fig. S15, Table S9). Whole genome
356 alignment of Rso IIB-4 UW163 from *Musa acuminata* x *balbisiana* AAB (Peru, 1967) and Rso IIB-4NPB
357 RUN1145 from *Cucurbita moschata* (Martinique, 2002) confirms horizontal gene transfer in the form of
358 prophage and ICE movement has had a significant impact on Rso IIB-4NPB genome evolution. Despite
359 being isolated 45 years apart from different hosts and regions, these genomes are highly conserved
360 across the chromosome and megaplasmid (Fig. S15, Fig. S16). The most striking difference between the
361 two lineages are large insertions in Rso IIB-4NPB, most of which are predicted to be prophage. RECOPHY
362 indicates the region with the highest SNP density in the Rso IIB-4 core genome alignment (1,060 out of
363 2,894 SNPs) is adjacent to two prophages and a proximal *fhaC-fhaB* locus (Rso IIB-4NPB RUN1145 chr1:
364 227,422-304,634). This region also contains a high density of Martinique-associated kmers in RUN1145
365 and *Musa*-associated kmers in UW163. The closest BLASTn hits to the novel prophage are soil isolates
366 from Asia (*Ralstonia* metagenome-assembled genome from estuary sediment in China, *Pandoraea*
367 *pnomenusa* from sludge in Malaysia and *Massilia putida* from soil in China) and *Xanthomonas*
368 *albilineans*, a bacterial plant pathogen causing lethal leaf scald in sugarcane from Brazil (Fig. S17). In
369 addition to the core genes encoding the phage head, tail, baseplate and restriction endonucleases, the
370 Rso IIB-4NPB-specific prophage carries auxiliary genes whose products are predicted to function as a
371 circadian clock protein (*kaiA*, OFBKGH_12090 in RUN1145) and a methyl-accepting chemotaxis protein
372 (OFBKGH_12085). Cumulatively, the data shows ICE acquisition and prophage infections occurred
373 multiple times in the ancestor of the emergent Rso IIB-4NPB lineage, either prior or soon after its
374 dissemination to Martinique.

375 **CRISPR array reveals a history of viral encounter in the emergent Rso IIB-4NPB lineage**

376 Evidence of repeated mobile element acquisition in the host-range expanded lineage Rso IIB-4NPB
377 attests to the high frequency of encounter and exchange with members of the soil, freshwater and plant
378 microbiomes. The carriage of antiviral defense gene clusters on the ICE and presence of type I-E CRISPR-
379 Cas systems among RSSC genomes indicates the plant pathogen is frequently challenged by viral attack.
380 CRISPR-Cas systems confer immunity to phage via the acquisition of spacer sequences from prior
381 infections. We found that both Rso IIB-4 and Rso IIB-4NPB share conserved CRISPR-Cas operons (*cas3*,
382 *casA/cas8e*, *casB/cas11/cse2*, *cas7e*, *cas5e*, *cas6e*, *cas1e*, *cas2*), but vary in their CRISPR arrays. The Rso
383 IIB-4NPB CRISPR 1 and CRISPR 2 arrays are nearly ten times longer than the Rso IIB-4 arrays: Rso IIB-4
384 UW163 has 20 spacers compared to the 171 spacers in Rso IIB-4NPB RUN1145 (of which 11 are
385 duplicates with 1-5bp differences). These genomes share one leader-proximal and 15 leader-distal
386 spacers. The kmer-based association tests show that the largest cluster of Martinique-associated kmers
387 that do not map to ICEs or prophage actually map to the leader-proximal position of CRISPR arrays in
388 Rso IIB-4NPB, corresponding to more recent spacer acquisition than the leader-distal spacers shared
389 with Rso IIB-4 (Fig. S15, Fig. S18)[51,52]. This represents a rapid expansion of CRISPR-mediated immunity in
390 the emergent lineage, indicating the emerging lineage encountered phage during its expansion across
391 Martinique.
392
393 We inferred the identity of the viruses challenging Rso IIB-4NPB using spacer sequences as queries for
394 INPHARED phage database searches. A total of nine phage were identified using this approach (Table
395 S8). Three phage are targeted by spacers in Rso IIB-4 and Rso IIB-4NPB alike, and six are exclusively
396 targeted by Rso IIB-4NPB. Interestingly, five of the six phage targeted by Rso IIB-4NPB are known to
397 infect Rps phylotype I strains (RPZH3, RS603, RSMSuper, Eline, Raharianne)[53–55]. Rso IIB-4NPB immunity
398 to phage infecting Rps suggests they not only co-occur, but also encounter the same viral threats.

399 **Discussion**

400 Outbreaks of plant disease tend not to be well sampled outside the focal host, or even deeply sampled
401 within the focal host of interest, limiting our ability to infer outbreak origins or identify the genomic
402 changes associated with pathogen emergence. Our knowledge of pathogen host range and the selective
403 environments they are exposed to outside the acute infection stage is extremely limited. Infection
404 assays using a broad panel of potential hosts provide an indication of potential host range, yet these
405 experiments typically include high inoculum densities, mechanical damage and/or other methods of
406 bypassing the initial invasion stage. These results are therefore not likely to predict what is observed
407 during natural field infections. The densely sampled collection from Martinique and French Guiana is
408 therefore unique as it captures a disease outbreak that affected a wide range of hosts, including isolates
409 from asymptomatic and environmental reservoirs alike, across island and mainland populations in the
410 Caribbean and South America. This work provides unprecedented insight into the evolutionary history
411 and processes acting on a pathogen community during an outbreak event, revealing how mobile
412 elements exchange and interactions with viral predators have shaped their genomes.
413

414 The Martinique outbreak was caused by the parallel expansion of two distantly related pathogen
415 lineages: Rso IIB-4NPB and Rps I-18. While at present considered an unusual event[56], this is likely to
416 change with increased outbreak sampling and sequencing. Genetic exchange between endemic and
417 introduced epidemic pathogens has been documented[57,3]. Interestingly, the broad host range Rso IIB-
418 4NPB lineage emerged from a diverse endemic population that has been circulating in the Americas for
419 some time, while the lineage that ultimately originated in Asia retains a signature of specialization on
420 Solanaceae. Rso IIB-4NPB is currently unique in causing natural infections in the most diverse group of
421 hosts ever documented for a single lineage in the field. Our results show this emergence event was not
422 driven by T3E innovation and illustrates the risk of using limited sampling and whole genome sequencing
423 to inform candidate gene selection for studies of virulence gene function. For example, a strain
424 IBSBF1503 previously identified as Rso IIB-4NPB is in fact distantly related to the emergent lineage that
425 swept across Martinique[58], and there is very little variation in effector carriage between Rso IIB-4NPB
426 and Rso IIB-4: a single hypothetical effector (*T3E_hyp3*) was acquired and two (*ripAX2* and *ripP3*) were
427 lost. It is possible that regulatory changes or allelic variation in T3Es may alter apparent host range[59], but
428 the observed differences appear unlikely to account for the apparent expansion in host range compared
429 to Rso IIB-4, particularly since single effector deletions typically have limited consequences for
430 pathogenicity[60]. Although T3E changes associated with host range expansion in the emergent lineage
431 may have occurred prior to its dissemination to Martinique, the ancestral mainland population from
432 which Rso IIB4-NPB arose remains associated with *Solanaceae* infections. In sum, there is no evidence
433 the outbreak of Rso IIB-4NPB in Martinique is driven by T3E-mediated changes altering pathogen
434 interactions with a panoply of distantly related hosts. The genomic evidence instead shows the
435 acquisition of anti-viral defences and novel metabolic functions are more likely to have played an
436 important role in the successful establishment of the pathogen across a broad range of hosts during this
437 outbreak. This does not preclude the possibility outbreaks caused by related lineages are linked with T3E
438 acquisition, as may be the case for banana-pathogenic Rso IIB-4.
439
440 Resolving fine-scale geographic patterns of mobile element carriage revealed hotspots of ICE carriage
441 and interspecies exchange in Martinique, and provides insight into the dynamic evolution of the mobile
442 elements themselves. These ICEs exhibit conservation of function in both hotspots of accessory gene
443 acquisition: despite variable gene content the cargo-1 hotspot always consists of metabolism-associated
444 genes and cargo-2 always consists of defense-related genes. Investigation of mobile element carriage in
445 all complete genomes available for the species complex suggests that despite chromosomal homology
446 at preferred ICE integration sites and the presence of direct repeats targeted by the $ICE^{RpsGMI1000}$
447 integrase, ICE carriage is less common in Rso and Rsy compared to Rps. Expanded population level
448 sequencing cautions against inferences drawn from limited sequencing however, since we show MGE
449 carriage can vary from one sampling site to another even within a single lineage. Similarly, rapid
450 movement of ICEs prevent reliable inference of the source and directionality of transfer events, but their
451 absence from Rso IIB-4 and near ubiquity among Rso IIB-4NPB isolates suggests $ICE^{RpsGMI1000}$ acquisition
452 is linked with the outbreak in Martinique. Since ICEs mediate the horizontal transfer of entire functional
453 genetic modules, allowing rapid adaptation to novel hosts and environmental niches, the acquisition of
454 an ICE with novel metabolic enzymes may have been pivotal to the expansion of Rso IIB-4NPB.
455

456 All Rso IIB-4NPB strains in Martinique carry an ICE with cargo-1 genes whose products are linked with
457 the degradation of environmental pollutants, reactive xenobiotics and plant-derived metabolites. The
458 phthalate ester hydrolase (*pnbA*) is linked with the degradation of phthalate esters, common synthetic
459 organic compounds that are known to be environmental pollutants[61]. Homologs of *gstA* and *fabG* are
460 known to degrade organochlorine compounds used as primary active ingredients (1-chloro-2,4-
461 dinitrobenzene) or intermediates (ethyl 4-chloroacetoacetate) in the production of herbicides and
462 pesticides[62,63]. Multiple cargo-1 genes have products linked with the manipulation of host-derived
463 metabolites: the limonene 1,2-epoxide hydrolases *limA* can act on host plant epoxides, induced during
464 both pathogen attack and methyl jasmonate application[64]; a homolog of *ygiN* (ABM) is known to
465 hydroxylate jasmonic acid in rice, attenuating plant defences during infection[65], and
466 pyrazinamidase/nicotinamidase (*pncA*) is known to convert nicotinamide to nicotinic acid. The
467 accumulation of nicotinic acid during grape xylem colonisation by *Xylella fastidiosa* suggests *pncA* may
468 be linked with vascular pathogenesis[66]. *Ralstonia* spp. degradation of plant metabolites has elsewhere
469 been observed in Rso IIB-1 UY031, which degrades phenolic compounds in xylem sap[67], and many
470 *Ralstonia* spp. genomes contain salicylic acid degradation pathways[45,68,69]. Bacterial metabolism and
471 manipulation of nutrient availability in host plants is increasingly recognised as an important component
472 of pathogenicity, in mammalian and plant pathogens alike[70–73]. For instance, genes whose products
473 facilitate the use of xylem sap as a carbon source[74] and horizontally acquired components of the
474 denitrification pathway[75] have been shown to be important for xylem colonisation and virulence in
475 *Ralstonia*. Transcriptome sequencing of experimentally evolved Rps I-18 GMI1000 showed fatty acid
476 metabolism is upregulated in evolved clones of Rps I-18 GMI1000, highlighting the importance of
477 metabolism during adaptation to a novel host[76]. Metabolism is likely carefully regulated according to
478 infection stage, as was shown for Rso IIB-1 UW551, which downregulates carbohydrate metabolism and
479 tricarboxylic acid-related pathways during late stages of xylem infection[77]. The acquisition of novel
480 pathways for substrate catabolism may enhance bacterial growth in the soil and/or within plant tissues,
481 resulting in an explosive increase in symptom production across a wide range of hosts.
482
483 Recombination events introducing metabolic enzymes occurred beyond the ICE as well, in the ancestor
484 of Rso IIB-4 and Rso IIB-4NPB. These are similarly linked with the degradation of either plant-derived
485 compounds or protection against environmental toxicity or oxidative stress. The acquisition of a
486 TauD/TfdA family dioxygenase linked with synthetic auxin degradation may confer a fitness advantage in
487 niches frequently exposed to phenoxy herbicides. The ability to metabolise toxic or antimicrobial
488 compounds is not unusual in *Ralstonia* and close relatives: *R. pickettii* degrades DDT pesticides [1] and
489 *Cupriavidus necator* (formerly *R. eutropha*) degrades chloroaromatic compounds[78]. The ancestral
490 acquisition of a carboxymuconolactone decarboxylase is also of considerable interest, since the
491 protocatechuate catabolism pathway is responsible for the conversion of lignin-derived aromatic
492 compounds into amino acids and organic acids, potentially fueling *Ralstonia* growth in plant tissues.
493
494 Emerging pathogens may themselves serve as hosts for potential parasites. The antiphage defense in ICE
495 cargo-2, expansion of CRISPR repeats and acquisition of multiple fragmented prophage in Rso IIB-4NPB
496 attests to frequent viral encounter, and indicates immunity to predatory phage is likely to provide an

497 advantage to the pathogen[79]. Phage resistance is similarly implicated in the success of 7th pandemic
498 *Vibrio cholerae* lineages[80,4]. Since some of the Rso IIB-4NPB CRISPR spacers with identifiable origins
499 correspond to phage known to successfully infect *R. pseudosolanacearum* strains like Rps I-18,
500 coinfections are likely sites of exposure to lytic phage, along with the acquisition of ICEs and prophage
501 carrying accessory genes altering microbial lifestyle and pathogenicity[81]. Our findings emphasize the
502 value of integrating genomic, ecological and evolutionary perspectives to understand disease
503 emergence. As agricultural intensification continues to reshape plant-microbe interactions, pathogen
504 lineages that can rapidly acquire and integrate novel genetic material may be more likely to exploit
505 novel niches and hosts. This work underscores the need for field surveillance strategies that extend
506 beyond focal pathogen populations to the broader community and reservoir of genetic diversity that
507 fuels their evolution.

508 **Methods**

509 **Genome sequencing**

510 *Ralstonia* spp. were collected in Martinique and French Guiana as described in Wicker *et al*. 2009 and
511 Deberdt *et al*. 2014. Isolates collected after the initial sampling were assigned to phylotype and
512 sequevar by sequencing endoglucanase (*egl*) and mutation mismatch repair (*mutS*) genes. For this study,
513 we extracted genomic DNA from 487 strains using Promega Wizard, then used a modified Nextera
514 protocol for library preparation[82]. Paired-end sequencing with Illumina NextSeq 2000 was then
515 performed. Since the first round of sequencing produced under 25-fold average coverage and under 40-
516 fold maximum coverage, a second round of library preparation and sequencing was performed. Eight
517 isolates (RUN0777, RUN0808, RUN0820, RUN0823, RUN1145, RUN1155, RUN1952 and RUN2004) were
518 additionally sequenced using Oxford Nanopore Technology with a FLO-MIN106 flow cell and SQK-
519 RBK004 barcoding kit. Basecalling and demultiplexing was performed with guppy_basecaller (v. 5.0.7)
520 and guppy_barcoder v. 4.5.4[83].

521 **Genome assembly and annotation**

522 Paired-end 150bp reads of both Illumina sequencing runs were filtered and trimmed with
523 fastq_screen[84], bbduk[85] and fastp[86,87] (--dedup --detect_adapter_for_pe --cut_front 15 --length_required
524 50). Assembly was performed with SPAdes[88,89] using both separate and pooled (i.e. both sequencing
525 runs) read sets. QUAST was used for assembly quality assessment[90]. Pooling both sequencing runs
526 improved average assembly N50 values from 73Kb to 195Kb and decreased average contigs per
527 assembly from 370 to 217, so assemblies generated from pooled sequencing runs were used. After
528 filtering out assemblies with under 50-fold coverage, over 250 contigs, total assembly length exceeding
529 6Mb and N50 exceeding 75Kb, a total of 409 assemblies were retained for downstream analyses.
530
531 Oxford Nanopore long read data was filtered with Filtlong (v. 0.2.1)[91] to remove reads under 1Kb and
532 the 5% of reads with the lowest quality values, then assembled with Trycycler (v. 0.5.4)[92], using Flye[93],
533 Miniasm[94] with Minipolish[95] and Raven[96] as underlying assemblers. Contigs that did not pass the
534 reconciling step of Trycycler pipeline were manually removed. Final Trycycler consensus sequences were

535 polished three times using long reads with Medaka[97], followed by two rounds of short read polishing
536 using BWA[98,99] and Polypolish[95], using separate batches of the Illumina reads produced by each
537 sequencing run.
538
539 All assemblies were annotated by Bakta[100] using Prodigal[101] gene models and a reference protein
540 database built using all complete *Ralstonia* spp. genome assemblies available from the RefSeq database
541 in October 2022.

542 **NCBI datasets and reference genomes**

543 All available *R. pseudosolanacearum, R. syzygii* and *R. solanacearum* genome assemblies were
544 downloaded from NCBI in February 2023 to provide additional phylogenetic context for analyses.
545 Assemblies with fewer than 200 contigs, total assembly length between 5.0 - 6.2Mb, N50 over 75Kb and
546 unflagged by the NCBI Genome[102] for evidence of potential contamination were retained. A total of 410
547 *Ralstonia* spp. genomes passed filter (Table S1). Three isolates were present in duplicate in the NCBI
548 filtered dataset (MAFF 211479, UW163 and UW551), the lowest quality duplicate assembly was
549 removed, producing a final NCBI dataset of 407 assemblies. These were then reannotated with Bakta
550 and merged with the 409 assemblies produced in this work for downstream pangenome analyses.
551 Metadata (host, geographical region and year of isolation) for genomes in the NCBI dataset was
552 obtained by inspecting Biosample entries. If the information was missing in Biosample entries, we used
553 RSSCdb records[103,29] or publications describing the initial isolation of the strain of interest. Metadata for
554 the strains CIP128_UW455, UW258 and NCPPB282_UW491 was not used for the visualization due to
555 discrepancies.

556 ***Solanaceae* specialisation test**

557 Level of specialisation toward *Solanaceae* family hosts among three *Ralstonia* groups (Rso IIA, Rso IIB-
558 4NPB and Rps I-18) on Martinique was estimated by two Chi-squared Test for Independence using
559 function chi2_contingency from scipy package. Contingency table of infection frequencies was
560 composed from numbers of Rso IIA, Rso IIB-4NPB and Rps I-18 isolates from *Solanaceae* and non-
561 *Solanaceae* plants sampled on Martinique. Environmental isolates and the isolates with missing
562 metadata were not included. Tests were performed for two pairs Rso IIB-4NPB vs Rps I-18 and Rso IIB-
563 4NPB vs Rso IIA with the Null hypothesis "Probability of infecting *Solanaceae* family plan is independent
564 from the *Ralstonia* lineage". Observed and expected distribution were visualized by matplotlib package.

565 **Type 3 secreted effector identification**

566 Amino acid sequences of 118 *R. solanacearum* type 3 secreted effectors (T3Es) from the manually
567 curated RalstoT3E database[104] were downloaded and used as query sequences for DIAMOND BLASTp[105]
568 searches against proteomes of the Bakta-annotated assemblies, using the --ultrasensitive mode. Hits
569 were filtered using a S30L25 strategy: retaining sequences with alignment identity of at least 30% and
570 query sequence coverage of 25% or greater. Since the T3E query set includes multiple alleles of the
571 same effector families, hits were then grouped into T3E families by sorting first by identity and query
572 length coverage. T3E family assignment for each protein sequence was performed using the most

573 significant T3E hit. Effector presence and absence was visualised by a heatmap in R, omitting duplication
574 events. This approach was tested by comparing the predicted T3E repertoires in Rps I-18 GMI1000 and
575 Rsy RPSI07 with the ones provided in the RalstoT3E database. Average pairwise similarity of T3E
576 repertoires within RSSC clades was estimated by Scikit-learn function cosine_similarity[106] using the
577 information about presence-absence of effector families as an input binary matrix.

578 **Pangenome inference and core gene tree reconstruction**

579 Panaroo[107] was used to identify core and flexible genomes for the RSSC dataset, i.e. the merged dataset
580 of 816 *Ralstonia* genomes sequenced in this study (409 strains) and downloaded from NCBI (407
581 strains). A concatenation of core genes was used as an input alignment for phylogenetic analysis of the
582 species complex. Alignment columns with gaps or ambiguities and invariant sites were removed with
583 SNPsites[108], producing a final alignment of 137,423bp. Tree inference was then performed using
584 IQtree[109] with a GTR+ASC model of sequence evolution, model for alignments that do not contain
585 constant sites, and 1000 bootstrap replicates. The ggtree R package[110] was used for all tree and
586 metadata visualisation.
587
588 Panaroo gene presence and absence tables for the RSSC dataset were queried using a custom Python
589 script to identify genes uniquely present in either Rps I-18, Rso IIB-4 or Rso IIB-4NPB. Genes are referred
590 to as clade-specific when they are present in at least 95% of all genomes in a given focal clade, and
591 present in no more than 10 non-focal clade assemblies.
592
593 The species complex tree from the set of core genes identified by Panaroo was used for clade
594 delineation and lineages assignment. Most genomes clustered according to the phylotypes initially
595 assigned by *egl* and *mutS* sequencing in Wicker *et al.* 2007 and Deberdt *et al.* 2014, although some
596 discrepancies are present. We found previous sequevar designations can provide an inaccurate view of
597 strain relatedness, e.g. RUN2182 is classified as a Rps I-48, but groups with Rps I-47 in the core gene
598 tree. Some strains were also reclassified into different phylotypes based on the core gene tree and
599 RUN1259 was initially classified as Rso IIB but is in fact Rso IIA (Fig S1., Table S1). The clade designation
600 in this work may therefore differ from historically assigned sequevar or even phylotype labels, since
601 these were first assigned using *egl* and *mutS* sequences rather than whole genome sequencing and core
602 gene tree reconstruction. The criteria for membership in a named clade is exceeding a minimum shared
603 identity of 99% in the RSSC core gene alignment.
604
605 Separate pangenome analyses and core gene tree reconstructions were performed on three specific
606 clades: *R. pseudosolanacearum* phylotype 1 sequevar 18 (Rps I-18, 102 strains), *R. solanacearum*
607 phylotype 2 sequevar 4 (Rso IIB-4, 226 strains), and the emergent lineage *R. solanacearum* phylotype 2
608 sequevar 4 Non-Pathogenic on Banana (Rso IIB-4NPB, 180 strains). The strains forming each clade were
609 identified with reference to the RSSC core gene tree (Figure 1). Assemblies for IBSBF_2571
610 (GCA_003590605.1) and O12BS_UW407 (GCA_023076615.1) were excluded due to abnormally long
611 terminal branch lengths and P673 was excluded due to contradictory phylogenetic position of two
612 assembly versions: the most recent P673 assembly GCA_015698365.1 was assigned to Rso IIB-4NPB

613 clade, while the earlier version GCA_000525615.1 grouped with other Rso IIB-4 *Epipremnum* isolates
614 from the USA.
615
616 An additional iteration of clade-specific pangenome analysis was performed for both Rso IIB-4NPB and
617 Rps I-18. Although strain IBSBF1503 (GCA_001587155.1) groups with other Rso IIB-4NPB strains in the
618 RSSC core gene tree, the clade-specific tree revealed this strain is divergent and more properly referred
619 to as an outgroup strain, therefore this was excluded from a second iteration. Similarly, Rps I-18 strains
620 CIP266_UW505 (GCA_023075875.1) and UW393 (GCA_023076675.1) form a distant outgroup to other
621 Rps I-18 strains, therefore these were excluded from the second iteration of the Rps I-18 pangenome
622 analysis.

623 **Mobile element prediction in reference genome assemblies**

624 All potential mobile elements, including genes annotated as integrases, transposases, prophage-related
625 genes and IS elements, were identified and masked for the downstream phylogenetic analysis in the
626 Bakta-annotated genome assemblies for Rps I-18 strain GMI1000 (GCA_000009125.1), Rso IIB-4 strain
627 UW163 (GCA_001587135.1) and Rso IIB-4NPB strain RUN1145. Prophage and IS elements were
628 predicted using the PHASTER, PHASTEST[111–113] and ISfinder[114] online tools. All gene annotations
629 containing keywords "phage" OR "transposase" OR "tail" OR "head" OR "integrase" OR "terminase" OR
630 "conjugal" OR "integrase" OR "integrative conjugative element" OR "conjugative transfer" OR
631 "conjugative coupling" were manually searched for and flagged in Geneious Prime[115]. The results were
632 converted into BED files, then sorted, merged and used for masking with Bedtools[116]. Masked references
633 were used for reference-based contig mapping.

634 **Phylogenetic analyses**

635 Reference-based contig mapping was performed for the Rps I-18, Rso IIB-4 and Rso IIB-4NPB clades, as
636 defined by using RSSC and clade-specific core gene trees. Rps I-18 included 100 genomes, with GMI1000
637 as reference. The Rso IIB-4 clade contained 226 genomes, with UW163 as reference. Rso IIB-4NPB
638 included 179 genomes, with RUN1145 as reference. The Brazilian *Cucumis sativus* isolate IBSBF1503 is
639 referred to as Rso IIB-4NPB in the literature and initially grouped with Rso IIB-4NPB in the species
640 complex tree, and, so we initially used this strain as a reference for Rso IIB-4NPB. While more closely
641 related to Rso IIB-4NPB than any other available complete *Rso* genome, we found IBSBF1503 is clearly
642 an outgroup relative to other Rso IIB-4NPB isolates (Fig. S3). The average pairwise nucleotide diversity is
643 0.002% between Rso IIB-4NPB genomes, and an order of magnitude greater (0.02%) relative to
644 IBSBF1503. According to progressiveMauve whole genome alignments, Rso IIB-4NPB isolate RUN1145
645 differs from RUN1960 and IBSBF1503 by 134 and 1,215 SNPs, respectively. Since IBSBF1503 is too
646 divergent to be considered a member of the same lineage, we reclassified IBSBF1503 as Rso IIB-4. We
647 resequenced five Rso IIB-4NPB strains from Martinique using Nanopore long-read sequencing to
648 generate closed hybrid assemblies and used the complete assembly of Rso IIB-4NPB RUN1145 as a
649 reference for more reliable core genome alignment-based phylogenetic inference of relationships
650 between closely related strains. The Rps I-18 contig mapping using Rps I-18 GMI1000 as reference
651 resulted in 4,760,272 bp consensus alignment with 311 SNPs and 116 parsimony informative sites.

652
653 Reads and contigs mappings were performed separately for each assembly with bwa-mem2[98,99] or
654 Minimap2[94] using asm5 mode respectively. All resulting BAM alignments were sorted and indexed with
655 Samtools[117]. BCFtools mpileup was used to filter and retain bases with mapping quality above 60[118].
656 Variants were called with bcftools-call and filtered by bcftools-filter using the expression TYPE="snp" &&
657 DP>=X && QUAL>30 && AF1>=0.95 with X=10 for reads and X=1 for contig mapping. Any alignment
658 column with a gap site was removed. Consensus sequences were obtained using bcftools-consensus,
659 masking the sites with coverage depth below one as identified by Mosdepth[119]. The consensus
660 sequences for all strains in a given clade were concatenated into a FASTA alignment file. Sites containing
661 N characters were trimmed with goalign clean-sites[120]. Recombinant regions were predicted and
662 removed by ClonalFrameML[121] using an initial tree calculated by IQtree[109] under GTR+F+I+R2
663 substitution model. The final Maximum likelihood trees were calculated from the non-recombinant
664 alignments for each clade by IQtree using the MFP (ModelFinder Plus) option for the best model
665 search[122]. The contig mapping approach resulted in better resolved and bootstrap supported trees,
666 therefore the trees generated by reads mapping were excluded from further analysis. Tree visualisations
667 were created with ggtree[110].

668 **Tree dating**

669 Root-to-tip regression and dating of the Rso IIB-4NPB and Rps I-18 non-recombinant core genome trees
670 were performed with BactDating using 10,000 Markov Chain Monte Carlo iterations[123]. TempEst was
671 used to confirm root-to-tip regression results obtained with BactDating, setting the 'Best-fitting root'
672 option of TempEst to 'R squared', and for outlier detection by investigating 'Residuals' plots[124]. Tree tips
673 with missing sampling date information (Rps I-18 UQRS637_UW745 and UW193, Rso IIB-4NPB RUN0283
674 and RUN0739) and outliers (Rps I-18 UW393, GMI1000, RUN1974, UW299, RUN1975, RUN1958,
675 RUN1955, RUN0037, RUN0679 and CIP266_UW505, Rso IIB-4NPB RUN0003, RUN0002, RUN2227,
676 RUN2226, RUN1959, RUN1960, RUN0016, RUN2007, RUN2009, RUN2004, RUN2011, RUN0762,
677 RUN1046, RUN1121, RUN0005, RUN1094, RUN1064, RUN1951, RUN1968, RUN1105, RUN1167 and
678 RUN0006) identified by TempEst were dropped from the Rps I-18 and Rso IIB-4NPB alignments and trees
679 were re-calculated. Branches with edge lengths of zero were manually collapsed into a single node by
680 Dendroscope[125]. A Dates Permutation Test was performed 500 times for each tree with BactDating by
681 randomly generating dates between 2000 and 2012 and randomly assigning these to tree tips while
682 keeping the tree root (option: updateRoot=F). The simulated trees were compared to the actual dating
683 model using the modelcompare function of BactDating.

684 **CRISPR-Cas system identification**

685 CRISPR-Cas systems were identified with Bakta annotation of Rso IIB-4 and Rso IIB-4NPB genomes.
686 CRISPR repeats and spacers were then annotated in two closed references Rso IIB-4-UW163 and Rso IIB-
687 4NPB-RUN1145 using the CRISPRloci workflow[126]. The spacer sequences identified with CRISPRloci were
688 then used for BLASTn searches against the RefSeq Phage database downloaded from PhageScope[127],
689 NCBI Nucleotide collection (nr_nt) and Environmental samples (env_nt) databases accessed in June

690 2024. The extracted spacers were also used for INPHARED database[128] searches using the
691 SpacePHARER[129] toolkit in August 2024.

692 **Inter and intraspecific recombination prediction**

693 Potential homologous recombination events within and between Rso IIB-4 and Rps I-18 lineages were
694 investigated using the RECOPHY toolkit[130]. Two datasets were analysed: Rso IIB-4 alone and in
695 combination with Rps I-18, using Rso IIB-4NPB-RUN1145 as a reference in each case. Prior to RECOPHY
696 execution, all genome pairs with a Mash distance[131] below zero were identified and one member of each
697 pair was then randomly removed from the dataset. The final datasets included 49 genomes for the Rso
698 IIB-4 analysis alone, and with an additional 64 Rps I-18 genomes for the merged analysis. The regions of
699 core alignment with increased and reduced density of SNPs per 1 kb block were extracted and
700 investigated separately. In order to find discrepancies with the core phylogeny, a simple Neighbor-
701 Joining tree was calculated for each of these regions using Geneious Prime software[115].

702 **Pangenome-wide association testing**

703 Scoary and Pyseer pan-GWAS analyses were used to investigate associations between gene presence
704 and absence patterns and the host or location of strain isolation[132,133]. The broad (country) and fine-
705 scale (village) locations and all taxonomic levels of the host of isolation or environmental substrate were
706 converted from the metadata table into a binary traits table using the pandas.get_dummies function.
707 For example, a strain isolated from a tomato plant in China would receive the value "1" in the columns
708 China, *Eudicots, Asterids, Solanales, Solanaceae, Solanum* and *S. lycopersicum*, and receive the value "0"
709 in all other columns of the traits table. This traits table was used as input for both Scoary and Pyseer
710 pan-GWAS analyses, along with the gene presence and absence table from Panaroo.
711
712 The significance of candidates identified with Scoary was assessed after multiple test corrections:
713 individual (naive) p-value, Bonferroni adjusted p-value, Benjamini-Hochberg adjusted p-value, entire
714 range of pairwise comparison p-values, and empirical p-value from 100 permutations, using thresholds
715 of p=0.05 for all corrections. Significant genes identified by Scoary after all the corrections were further
716 filtered using 95% sensitivity and 70% specificity thresholds. A lower specificity threshold was chosen
717 due to the nature of the trait dataset: the traits table simply reflects the known host of isolation and the
718 location of isolation. The actual host range of a particular strain may include additional hosts, closely or
719 distantly related, just as closely related strains may be present in vastly different geographic regions.
720 Since the Scoary and Pyseer pan-GWAS analyses aim to identify genes linked with survival in a particular
721 host, environment or substrate, the sensitivity should be high. Those same genes may also be present in
722 strains isolated from other plants however, therefore specificity is reduced. For example: a gene whose
723 product is strongly associated with tomato (*S. lycopersicum*) infection should be overrepresented among
724 tomato isolates in the dataset. Therefore we seek genes present in 95% of tomato isolates. We reduced
725 the specificity threshold to 70% to allow for instances where another strain carries that same gene and
726 is also capable of infecting tomato, but happens to have been isolated from a different host.
727

728 For the Pyseer analysis, the distance matrix from the RSSC core gene tree and the number of patterns
729 for the likelihood ratio test (LRT) were calculated using phylogeny_distance.py and count_patterns.py
730 scripts downloaded from the Pyseer GitHub page, accessed on January 2024[133]. The pattern count
731 resulted in 6,075 patterns for the RSSC, leading to the likelihood ratio test (LRT) p-value threshold of
732 3.8e-08 = 0.05/(patterns_number*traits_number). Hits with effect size (beta value) under 0.2 and
733 phenotypic variance values (variant_h2) above 0.5 were filtered out.
734
735 Pan-GWAS of gene presence/absence data is an effective tool for the identification of genes associated
736 with traits of interest, but is limited to the accessory genome and does not test for trait associations
737 within the core genome and intergenic regions. This gap is filled with a kmer-based association test.
738 Analogously to the method used for gene presence/absence data, Pyseer was applied to identify the
739 kmers significantly associated with any host or location of isolation across the RSSC, using the same
740 binary metadata table as described for the pan-GWAS analysis. The initial set of 31bp long kmers was
741 generated for all the RSSC genomes using unitig-caller [1]. The distance matrix from the species complex
742 core genes tree and the number of patterns for the likelihood ratio test (LRT) were calculated using
743 Pyseer phylogeny_distance.py and count_patterns.py scripts[133]. The pattern count reported 1,484,018
744 patterns for the RSSC kmers, producing a LRT p-value threshold of 5.0e-10 =
745 0.05/(patterns_number*traits_number). Hits with effect size (beta value) under 0.2 and phenotypic
746 variance values (variant_h2) above 0.5 were filtered out. The presence and positions of significant kmers
747 in Rso IIB-4NPB reference strain RUN1145, Rso IIB-4 reference strain UW163 and Rps I-18 reference
748 strain GMI1000 was identified and stored in BED files using a custom Python script. For visualisation
749 purposes, all kmers mapped on a genome within a 1kb distance were merged using the bedtools-merge
750 function[116].
751
752 The panGWAS and kmer-based tests identify many genes associated with isolation from specific hosts.
753 These tests vary in how they control for population structure, but some genes were predicted to have
754 associations with particular traits using multiple methods. Separate association tests were performed
755 using either spatial location or host/substrate of isolation metadata. We noticed that many genes have
756 overlapping predictions using both metadata types, particularly where it concerns isolates recovered
757 from soil or water. We also found overlapping host associations at multiple taxonomic levels, for
758 example gene presence associated with *S. lycopersicum* isolation often overlaps with gene presence
759 associated with *Solanaceae* isolation. All overlaps were investigated using a custom Python script and
760 visualised in an Upset plot using the UpSetR R package[134]. For the genes assigned to multiple nested
761 traits, the lowest level of taxonomic identification (e.g. species rather than genus) or the finest level of
762 geographical scale (e.g. village rather than country) for which significance was identified were chosen.
763 Overlaps within the Scoary, Pyseer pan-GWAS and kmer-based outputs were identified using the same
764 strategy.

765 **Identification of integrative and conjugative elements**

766 Complete integrative and conjugative elements were identified in 133 complete RSSC genome
767 assemblies using the following criteria: the presence of a type 4 secretion system (T4SS) required for

768 conjugal transfer between strains, a *mob* family relaxase and integrase, and the presence of 8bp direct
769 repeats (TTTTTTAT) encompassing the predicted ICE[135]. This 8bp repeat is described as an ICE insertion
770 site for Rps I-18 GMI1000 in the ICEberg 3.0 database[136]. We note this repeat occurs between 50-70
771 times elsewhere in each RSSC genome. This does not preclude them from being actual ICE integration
772 sites: despite the apparent existence of preferred ICE integration sites in the RSSC, some ICE integrases
773 are known to be promiscuous and may target different repeats as attachment sites across the genome.
774 For example, ICE Tn916 targets AT-rich loci and integrates almost at random across genomes[137].
775 MOBscan was used for identification of relaxases[138], followed by BLASTn sequence homology searches
776 for a T4SS and adjacent integrases[139] using Rso IIB-4NPB RUN1960 T4SS and integrase sequences as
777 BLASTn queries, in turn followed by T4SS locus confirmation with the TXSScan module of MacSyFinder
778 toolkit[140]. The flanking repeat sequence TTTTTTAT was identified using a custom python script. In
779 genomes where the repeat was not identified within 10kb of the T4SS, relaxase and integrase, putative
780 insertion sites were manually identified.
781
782 The genomic context of putative ICEs in the RSSC assemblies was inferred by aligning 50 kb upstream
783 and downstream flanking regions surrounding putative ICEs in Rps I-18 GMI1000, Rso IIB-4NPB RUN1145
784 and Rso IIB-4NPB RUN1960 to the genomes. The presence of genes significantly associated with
785 isolation from Martinique by panGWAS analysis was also confirmed by BLASTn, retaining hits with e-
786 values under 10^{-5}, query length coverage over 10% and overall alignment length exceeding 1 kb.
787
788 The involvement of ICE cargo genes in specific metabolic pathways was predicted using KEGG Orthology
789 assignment with protein sequences from the cargo regions of interest, performing searches in the
790 database for the genus *Ralstonia* (taxa ID: 48736). The locus structure of putative ICEs was visualised
791 with the ggtree::geom_motif R function[110]. Complete ICE sequences identified in *Rps* 362200, *Rps* RS24,
792 *Rps* RUN1952, Rps I-18 RUN820, Rps I-18 GMI1000, Rso IIB-4NPB RUN1960, RUN0777, Rso IIA RUN1155
793 and *Ralstonia pickettii* strain 12J were aligned to each other using the progressive Mauve algorithm
794 implemented in Geneious[141].

795 **Geographical metadata display**

796 Information about the geographical region of isolation for each strain was obtained from the country of
797 isolation metadata collected for the RSSC strains set as described above. Counties of isolation were
798 converted into continents using pycountry Python package[142]. The "Caribbean" region of isolation was
799 manually assigned for all Martinique isolates to highlight the region of interest.
800
801 Longitude and latitude of the sampling site was recorded for 83/500 strains sampled in Martinique and
802 French Guiana. For the remaining 417/500 strains, the commune (administrative district) or the nearest
803 village was listed in the strain metadata. For these cases, approximate longitude and latitude values
804 were retrieved using the geolocator function of the geopy.geocoders Python package using the
805 commune or nearest village names[143]. Commune maps of Martinique and French Guiana in geojson
806 format were downloaded from the France Geojson collection[144]. The actual and estimated longitude and
807 latitude values for each isolate were visualised with the geopandas package[145].

808 **Data availability**

809 Sequence data generated for this work is available via NCBI SRA Accession PRJNA1196686. Annotated
810 assemblies are available via Zenodo at https://zenodo.org/records/15619409.

References

1. Keller, M. *et al.* Ancient *Yersinia pestis* genomes from across Western Europe reveal early diversification during the First Pandemic (541–750). *Proc. Natl. Acad. Sci. U.S.A.* **116**, 12363–12372 (2019).
2. Azarian, T. *et al.* Global emergence and population dynamics of divergent serotype 3 CC180 pneumococci. *PLoS Pathog* **14**, e1007438 (2018).
3. Lassalle, F. *et al.* Genomic epidemiology reveals multidrug resistant plasmid spread between *Vibrio cholerae* lineages in Yemen. *Nat Microbiol* **8**, 1787–1798 (2023).
4. Adams, D. W. *et al.* West African–South American pandemic *Vibrio cholerae* encodes multiple distinct phage defence systems. *Nat Microbiol* **10**, 1352–1365 (2025).
5. McCann, H. C. *et al.* Origin and Evolution of the Kiwifruit Canker Pandemic. *Genome Biology and Evolution* **9**, 932–944 (2017).
6. Quibod, I. L. *et al.* The Green Revolution shaped the population structure of the rice pathogen *Xanthomonas oryzae* pv. *oryzae*. *The ISME Journal* **14**, 492–505 (2020).
7. Landa, B. B. *et al.* Emergence of a Plant Pathogen in Europe Associated with Multiple Intercontinental Introductions. *Appl Environ Microbiol* **86**, e01521-19 (2020).
8. Roman-Reyna, V. *et al.* Live tracking of a plant pathogen outbreak reveals rapid and successive, multidecade plasmid reduction. *mSystems* **9**, e00795-23 (2024).
9. Colombi, E. *et al.* Rapid dissemination of host metabolism–manipulating genes via integrative and conjugative elements. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2309263121 (2024).
10. Timilsina, S. *et al.* Diversification of an emerging bacterial plant pathogen; insights into the global spread of *Xanthomonas euvesicatoria* pv. *perforans*. *PLoS Pathog* **21**, e1013036 (2025).
11. Wicker, E. *et al. Ralstonia solanacearum* Strains from Martinique (French West Indies) Exhibiting a New Pathogenic Potential ▿ †. *Applied and Environmental Microbiology* **73**, 6790–6801 (2007).
12. Wicker, E. *et al.* Contrasting recombination patterns and demographic histories of the plant pathogen *Ralstonia solanacearum* inferred from MLSA. *The ISME Journal* **6**, 961–974 (2011).
13. Genin, S. & Denny, T. P. Pathogenomics of the *Ralstonia solanacearum* Species Complex. *Annual Review of Phytopathology* **50**, 67–89 (2012).
14. Clarke, C. R. *et al.* Genome-Enabled Phylogeographic Investigation of the Quarantine Pathogen *Ralstonia solanacearum* Race 3 Biovar 2 and Screening for Sources of Resistance Against Its Core Effectors. *Phytopathology* **105**, 597–607 (2015).
15. Prior, P. *et al.* Genomic and proteomic evidence supporting the division of the plant pathogen *Ralstonia solanacearum* into three species. *BMC genomics* **17**, 90 (2016).
16. Safni, I., Subandiyah, S. & Fegan, M. Ecology, Epidemiology and Disease Management of *Ralstonia syzygii* in Indonesia. *Frontiers in Microbiology* **9**, 419 (2018).
17. Wicker, E. *et al.* A reference database of *Ralstonia solanacearum* egl-mutS haplotypes for global epidemiological surveillance of bacterial wilts. (2016).
18. Cellier, G., Gauche, M. M., Cheron, J. J. & Pecrix, Y. How to Conduct Phylogenetic Endoglucanase (*egl*) Inference Using the Reference *Ralstonia solanacearum* Species Complex Curated Database. *Phytopathology®* **115**, 548–554 (2025).
19. Dewberry, R. J. *et al.* Genotypic and Phenotypic Analyses Show *Ralstonia solanacearum* Cool Virulence is a Quantitative Trait Not Restricted to “Race 3 biovar 2”.
20. Drenth, A. & Kema, G. The Vulnerability of Bananas to Globally Emerging Disease Threats. *Phytopathology®* **111**, 2146–2161 (2021).
21. United States HHS and USDA Select Agents and Toxins List. https://www.selectagents.gov/sat/list.htm.

857 22. Ramírez, M., Moncada, R. N., Villegas-Escobar, V., Jackson, R. W. & Ramírez, C. A. Phylogenetic
858 and pathogenic variability of strains of *Ralstonia solanacearum* causing moko disease in Colombia.
859 *Plant Pathology* **69**, 360–369 (2020).
860 23. Wicker, E., Grassart, L., Coranson-Beaudu, R., Mian, D. & Prior, P. Epidemiological evidence for the
861 emergence of a new pathogenic variant of *Ralstonia solanacearum* in Martinique (French West
862 Indies). *Plant Pathology* **58**, 853–861 (2009).
863 24. Deberdt, P. *et al.* Diversity of *Ralstonia solanacearum* in French Guiana expands knowledge of the
864 'emerging ecotype'. *Phytopathology* **104**, 586–96 (2014).
865 25. Prior, P. & Steva, H. Characteristics of Strains of *Pseudomonas solanacearum* from the French
866 West Indies. *Plant Disease* **74**, 13–17 (1990).
867 26. Abdurahman, A. *et al.* Molecular characterization of *Ralstonia solanacearum* strains from Ethiopia
868 and tracing potential source of bacterial wilt disease outbreak in seed potatoes. *Plant Pathology*
869 **66**, 826–834 (2017).
870 27. Tessema, L. *et al.* Seed health trade-offs in adopting quality declared seed in potato farming
871 systems. *Crop Science* **64**, 1340–1348 (2024).
872 28. Albuquerque, G. M. R. *et al.* Moko Disease-Causing Strains of *Ralstonia solanacearum* from Brazil
873 Extend Known Diversity in Paraphyletic Phylotype II. *Phytopathology®* **104**, 1175–1182 (2014).
874 29. Lowe-Power Lab. Global distribution of plant pathogenic *Ralstonia* strains.
875 https://github.com/lowepowerlab/Ralstonia_Global_Diversity (2022).
876 30. Price, N. S. The origin and development of banana and plantain cultivation. in *Bananas and*
877 *Plantains* (ed. Gowen, S.) 1–13 (Springer Netherlands, Dordrecht, 1995). doi:10.1007/978-94-011-
878 0737-2_1.
879 31. Kress, W. J. Systematics of Central American Heliconia (Heliconiaceae) with pendent
880 inflorescences. *Journal of the Arnold Arboretum* **65**, 429–532 (1984).
881 32. Croat, T. B. *Ecology and Life Forms of Araceae*. vol. 11 (1990).
882 33. Hueck, C. J. Type III protein secretion systems in bacterial pathogens of animals and plants.
883 *Microbiology and molecular biology reviews : MMBR* **62**, 379–433 (1998).
884 34. Poueymiro, M. *et al.* Two Type III Secretion System Effectors from *Ralstonia solanacearum*
885 GMI1000 Determine Host-Range Specificity on Tobacco. *Molecular Plant-Microbe Interactions®*
886 **22**, 538–550 (2009).
887 35. Landry, D., González-Fuente, M., Deslandes, L. & Peeters, N. The large, diverse, and robust arsenal
888 of *Ralstonia solanacearum* type III effectors and their in planta functions. *Molecular Plant*
889 *Pathology* **21**, 1377–1388 (2020).
890 36. Suwa, Y. *et al.* Characterization of a chromosomally encoded 2,4-dichlorophenoxyacetic
891 acid/alpha-ketoglutarate dioxygenase from *Burkholderia* sp. strain RASC. *Appl Environ Microbiol*
892 **62**, 2464–2469 (1996).
893 37. Zaprasis, A., Liu, Y.-J., Liu, S.-J., Drake, H. L. & Horn, M. A. Abundance of Novel and Diverse *tfdA* -
894 Like Genes, Encoding Putative Phenoxyalkanoic Acid Herbicide-Degrading Dioxygenases, in Soil.
895 *Appl Environ Microbiol* **76**, 119–128 (2010).
896 38. Newby, D. T., Gentry, T. J. & Pepper, I. L. Comparison of 2,4-Dichlorophenoxyacetic Acid
897 Degradation and Plasmid Transfer in Soil Resulting from Bioaugmentation with Two Different pJP4
898 Donors. *Appl Environ Microbiol* **66**, 3399–3407 (2000).
899 39. Füchslin, H. P., Rüegg, I., Van Der Meer, J. R. & Egli, T. Effect of integration of a GFP reporter gene
900 on fitness of *Ralstonia eutropha* during growth with 2,4-dichlorophenoxyacetic acid.
901 *Environmental Microbiology* **5**, 878–887 (2003).
902 40. Bælum, J., Henriksen, T., Hansen, H. C. B. & Jacobsen, C. S. Degradation of 4-Chloro-2-
903 Methylphenoxyacetic Acid in Top- and Subsoil Is Quantitatively Linked to the Class III *tfdA* Gene.
904 *Appl Environ Microbiol* **72**, 1476–1486 (2006).

905 41. Han, L., Zhao, D. & Li, C. Isolation and 2,4-D-degrading characteristics of *Cupriavidus campinensis*
906 BJ71. *Braz. J. Microbiol.* **46**, 433–441 (2015).
907 42. Chivers, P. T., Prehoda, K. E. & Raines, R. T. The CXXC Motif: A Rheostat in the Active Site.
908 *Biochemistry* **36**, 4061–4066 (1997).
909 43. Quan, S., Schneider, I., Pan, J., Von Hacht, A. & Bardwell, J. C. A. The CXXC Motif Is More than a
910 Redox Rheostat. *Journal of Biological Chemistry* **282**, 28823–28833 (2007).
911 44. Sancho, J. Flavodoxins: sequence, folding, binding, function and beyond. *Cell. Mol. Life Sci.* **63**,
912 855–864 (2006).
913 45. Lowe, T. M., Ailloud, F. & Allen, C. Hydroxycinnamic Acid Degradation, a Broadly Conserved Trait,
914 Protects *Ralstonia solanacearum* from Chemical Plant Defenses and Contributes to Root
915 Colonization and Virulence. *MPMI* **28**, 286–297 (2015).
916 46. Eulberg, D., Lakner, S., Golovleva, L. A. & Schlömann, M. Characterization of a Protocatechuate
917 Catabolic Gene Cluster from *Rhodococcus opacus* 1CP: Evidence for a Merged Enzyme with 4-
918 Carboxymuconolactone-Decarboxylating and 3-Oxoadipate Enol-Lactone-Hydrolyzing Activity. *J*
919 *Bacteriol* **180**, 1072–1081 (1998).
920 47. Martim, D. B. *et al.* Resolving the metabolism of monolignols and other lignin-related aromatic
921 compounds in *Xanthomonas citri*. *Nat Commun* **15**, 7994 (2024).
922 48. Barbirato, F., Verdoes, J. C., De Bont, J. A. M. & Van Der Werf, M. J. The *Rhodococcus erythropolis*
923 DCL14 limonene-1,2-epoxide hydrolase gene encodes an enzyme belonging to a novel class of
924 epoxide hydrolases. *FEBS Letters* **438**, 293–296 (1998).
925 49. Doron, S. *et al.* Systematic discovery of antiphage defense systems in the microbial pangenome.
926 *Science* **359**, eaar4120 (2018).
927 50. Cheng, R. *et al.* A nucleotide-sensing endonuclease from the Gabija bacterial defense system.
928 *Nucleic Acids Research* **49**, 5216–5229 (2021).
929 51. He, J. & Deem, M. W. Heterogeneous Diversity of Spacers within CRISPR (Clustered Regularly
930 Interspaced Short Palindromic Repeats). *Phys. Rev. Lett.* **105**, 128102 (2010).
931 52. Da Silva Xavier, A. *et al.* Characterization of CRISPR-Cas systems in the *Ralstonia solanacearum*
932 species complex. *Molecular Plant Pathology* **20**, 223–239 (2019).
933 53. Trotereau, A. *et al.* High genomic diversity of novel phages infecting the plant pathogen *Ralstonia*
934 *solanacearum*, isolated in Mauritius and Reunion islands. *Sci Rep* **11**, 5382 (2021).
935 54. Gonçalves, O. S., Souza, F. D. O., Bruckner, F. P., Santana, M. F. & Alfenas-Zerbini, P. Widespread
936 distribution of prophages signaling the potential for adaptability and pathogenicity evolution of
937 *Ralstonia solanacearum* species complex. *Genomics* **113**, 992–1000 (2021).
938 55. Lin, Z. *et al.* Characterization and complete genome sequence analysis of the novel phage RPZH3
939 infecting *Ralstonia solanacearum*. *Arch Virol* **168**, 105 (2023).
940 56. Abou Fayad, A. *et al.* An unusual two-strain cholera outbreak in Lebanon, 2022-2023: a genomic
941 epidemiology study. *Nat Commun* **15**, 6963 (2024).
942 57. Dorman, M. J. *et al.* Genomics of the Argentinian cholera epidemic elucidate the contrasting
943 dynamics of epidemic and endemic *Vibrio cholerae*. *Nat Commun* **11**, 4918 (2020).
944 58. Ailloud, F. *et al.* Comparative genomic analysis of *Ralstonia solanacearum* reveals candidate genes
945 for host specificity. *BMC Genomics* **16**, 270 (2015).
946 59. Ailloud, F. *et al. In planta* comparative transcriptomics of host-adapted strains of *Ralstonia*
947 *solanacearum*. *PeerJ* **4**, e1549 (2016).
948 60. Wei, H.-L. *et al. Pseudomonas syringae* pv. *tomato* DC3000 Type III Secretion Effector Polymutants
949 Reveal an Interplay between HopAD1 and AvrPtoB. *Cell Host & Microbe* **17**, 752–762 (2015).
950 61. Ding, J. *et al.* Properties of a Newly Identified Esterase from *Bacillus* sp. K91 and Its Novel Function
951 in Diisobutyl Phthalate Degradation. *PLoS ONE* **10**, e0119216 (2015).

952 62. Allocati, N., Federici, L., Masulli, M. & Di Ilio, C. Glutathione transferases in bacteria. *The FEBS*
953 *Journal* **276**, 58–75 (2009).
954 63. Santos, J. E. L. *et al.* Removal of herbicide 1-chloro-2,4-dinitrobenzene (DNCB) from aqueous
955 solutions by electrochemical oxidation using boron-doped diamond (BDD) and PbO2 electrodes.
956 *Journal of Hazardous Materials* **402**, 123850 (2021).
957 64. Sauveplane, V. *et al. Arabidopsis thaliana* CYP77A4 is the first cytochrome P450 able to catalyze
958 the epoxidation of free fatty acids in plants. *The FEBS Journal* **276**, 719–735 (2009).
959 65. Patkar, R. N. *et al.* A fungal monooxygenase-derived jasmonate attenuates host innate immunity.
960 *Nat Chem Biol* **11**, 733–740 (2015).
961 66. Feitosa-Junior, O. R. *et al.* The Exometabolome of *Xylella fastidiosa* in Contact with
962 *Paraburkholderia phytofirmans* Supernatant Reveals Changes in Nicotinamide, Amino Acids,
963 Biotin, and Plant Hormones. *Metabolites* **14**, 82 (2024).
964 67. Hamilton, C. D. *et al. Ralstonia solanacearum* pandemic lineage strain UW551 overcomes
965 inhibitory xylem chemistry to break tomato bacterial wilt resistance. *Molecular Plant Pathology*
966 **25**, e13395 (2024).
967 68. Baroukh, C. *et al.* Insights into the metabolic specificities of pathogenic strains from the *Ralstonia*
968 *solanacearum* species complex. *mSystems* e00083-23 (2023) doi:10.1128/msystems.00083-23.
969 69. Lowe-Power, T. M. *et al.* Degradation of the Plant Defense Signal Salicylic Acid Protects *Ralstonia*
970 *solanacearum* from Toxicity and Enhances Virulence on Tobacco. *mBio* **7**, e00656-16 (2016).
971 70. Eom, J.-S. *et al.* SWEETs, transporters for intracellular and intercellular sugar translocation.
972 *Current Opinion in Plant Biology* **25**, 53–62 (2015).
973 71. Freyberg, Z. & Harvill, E. T. Pathogen manipulation of host metabolism: A common strategy for
974 immune evasion. *PLoS Pathog* **13**, e1006669 (2017).
975 72. Rowley, C. A. & Kendall, M. M. To B12 or not to B12: Five questions on the role of cobalamin in
976 host-microbial interactions. *PLoS Pathog* **15**, e1007479 (2019).
977 73. Cai, J. *et al.* Manipulation of plant metabolism by pathogen effectors: more than just food. *FEMS*
978 *Microbiology Reviews* **47**, fuad007 (2023).
979 74. Hamilton, C. D., Steidl, O. R., MacIntyre, A. M., Hendrich, C. G. & Allen, C. *Ralstonia solanacearum*
980 Depends on Catabolism of Myo-Inositol, Sucrose, and Trehalose for Virulence in an Infection
981 Stage–Dependent Manner. *MPMI* **34**, 669–679 (2021).
982 75. Truchon, A. N. *et al.* Plant-Pathogenic *Ralstonia* Phylotypes Evolved Divergent Respiratory
983 Strategies and Behaviors To Thrive in Xylem. *mBio* **14**, e03188-22 (2023).
984 76. Gopalan-Nair, R. *et al.* Convergent Rewiring of the Virulence Regulatory Network Promotes
985 Adaptation of *Ralstonia solanacearum* on Resistant Tomato. *Molecular Biology and Evolution* **38**,
986 1792–1808 (2021).
987 77. De Pedro-Jové, R. *et al.* Dynamic expression of *Ralstonia solanacearum* virulence factors and
988 metabolism-controlling genes during plant infection. *BMC Genomics* **22**, 170 (2021).
989 78. Trefault, N. *et al.* Genetic organization of the catabolic plasmid pJP4 from *Ralstonia eutropha*
990 JMP134 (pJP4) reveals mechanisms of adaptation to chloroaromatic pollutants and evolution of
991 specialized chloroaromatic degradation pathways. *Environmental Microbiology* **6**, 655–668 (2004).
992 79. Emerson, J. B. *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nature*
993 *microbiology* **3**, 870–880 (2018).
994 80. LeGault, K. N. *et al.* Temporal shifts in antibiotic resistance elements govern phage-pathogen
995 conflicts. *Science* **373**, eabg2166 (2021).
996 81. Dougherty, P. E. *et al.* Widespread and largely unknown prophage activity, diversity, and function
997 in two genera of wheat phyllosphere bacteria. *The ISME Journal* **17**, 2415–2425 (2023).
998 82. Meier, J. I. *et al.* Haplotype tagging reveals parallel formation of hybrid races in two butterfly
999 species. *Proceedings of the National Academy of Sciences* **118**, (2021).

83. Oxford Nanopore Technologies. Guppy Protocol - Guppy software overview. https://community.nanoporetech.com/protocols/Guppy-protocol/v/gpb_2003_v1_revax_14dec2018.
84. Wingett, S. W. & Andrews, S. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research* **7**, 1338 (2018).
85. BBMap. https://sourceforge.net/projects/bbmap/ (2024).
86. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
87. Chen, S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta* **2**, e107 (2023).
88. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* **19**, 455–477 (2012).
89. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes De Novo Assembler. *CP in Bioinformatics* **70**, e102 (2020).
90. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
91. Wick, R. Filtlong. https://github.com/rrwick/Filtlong.
92. Wick, R. R. *et al.* Trycycler: consensus long-read assemblies for bacterial genomes. *Genome Biol* **22**, 266 (2021).
93. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37**, 540–546 (2019).
94. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).
95. Wick, R. R. & Holt, K. E. Polypolish: Short-read polishing of long-read bacterial genome assemblies. *PLoS Comput Biol* **18**, e1009802 (2022).
96. Vaser, R. & Šikić, M. Time- and memory-efficient genome assembly with Raven. *Nat Comput Sci* **1**, 332–336 (2021).
97. Nanoporetech. Medaka. https://github.com/nanoporetech/medaka.
98. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://doi.org/10.48550/ARXIV.1303.3997 (2013).
99. Vasimuddin, Md., Misra, S., Li, H. & Aluru, S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* 314–324 (IEEE, Rio de Janeiro, Brazil, 2019). doi:10.1109/IPDPS.2019.00041.
100. Schwengers, O. *et al.* Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification: Find out more about Bakta, the motivation, challenges and applications, here. *Microbial Genomics* **7**, (2021).
101. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
102. NCBI Genome: *Ralstonia solanacearum*, *Ralstonia pseudosolanacearum*, *Ralstonia syzygii*. https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=1310165,28097,305.
103. Lowe-Power, T. M. *et al.* A Meta-analysis of the known Global Distribution and Host Range of the *Ralstonia* Species Complex. Preprint at https://doi.org/10.1101/2020.07.13.189936 (2020).
104. Ralsto T3E v3. https://iant.toulouse.inra.fr/bacteria/annotation/site/prj/T3Ev3/.
105. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59–60 (2015).
106. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

1047 107. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline.
1048 *Genome Biol* **21**, 180 (2020).
1049 108. Page, A. J. *et al.* SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments.
1050 *Microbial Genomics* **2**, (2016).
1051 109. Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective
1052 Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and*
1053 *Evolution* **32**, 268–274 (2015).
1054 110. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. GGTREE : an R package for visualization and
1055 annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol*
1056 *Evol* **8**, 28–36 (2017).
1057 111. Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J. & Wishart, D. S. PHAST: A Fast Phage Search Tool.
1058 *Nucleic Acids Research* **39**, W347–W352 (2011).
1059 112. Arndt, D. *et al.* PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic acids*
1060 *research* **44**, W16-21 (2016).
1061 113. Wishart, D. S. *et al.* PHASTEST: faster than PHASTER, better than PHAST. *Nucleic Acids Research*
1062 **51**, W443–W450 (2023).
1063 114. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. & Chandler, M. ISfinder: the reference centre for
1064 bacterial insertion sequences. *Nucleic Acids Research* **34**, D32–D36 (2006).
1065 115. Geneious: Bioinformatics Software for Sequence Data Analysis. https://www.geneious.com/.
1066 116. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features.
1067 *Bioinformatics* **26**, 841–842 (2010).
1068 117. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*
1069 **25**, 2078–9 (2009).
1070 118. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and
1071 population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993
1072 (2011).
1073 119. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes.
1074 *Bioinformatics* **34**, 867–868 (2018).
1075 120. Lemoine, F. & Gascuel, O. Gotree/Goalign: toolkit and Go API to facilitate the development of
1076 phylogenetic workflows. *NAR Genomics and Bioinformatics* **3**, lqab075 (2021).
1077 121. Didelot, X. & Wilson, D. J. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial
1078 Genomes. *PLoS Comput Biol* **11**, e1004041 (2015).
1079 122. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermiin, L. S. ModelFinder:
1080 fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**, 587–589 (2017).
1081 123. Didelot, X., Croucher, N. J., Bentley, S. D., Harris, S. R. & Wilson, D. J. Bayesian inference of
1082 ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Research* **46**, e134–e134 (2018).
1083 124. Rambaut, A., Lam, T. T., Carvalho, L. M. & Pybus, O. G. Exploring the temporal structure of
1084 heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution* **2**, vew007
1085 (2016).
1086 125. Huson, D. H. *et al.* Dendroscope: An interactive viewer for large phylogenetic trees. *BMC*
1087 *Bioinformatics* **8**, 460 (2007).
1088 126. Alkhnbashi, O. S. *et al.* CRISPRloci: comprehensive and accurate annotation of CRISPR–Cas
1089 systems. *Nucleic Acids Research* **49**, W125–W130 (2021).
1090 127. Wang, R. H. *et al.* PhageScope: a well-annotated bacteriophage database with automatic analyses
1091 and visualizations. *Nucleic Acids Research* **52**, D756–D761 (2024).
1092 128. Cook, R. *et al.* INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases
1093 in the Current Collection of Cultured Phage Genomes. *PHAGE* **2**, 214–223 (2021).

129. Zhang, R. *et al.* SpacePHARER: sensitive identification of phages from CRISPR spacers in prokaryotic hosts. *Bioinformatics* **37**, 3364–3366 (2021).
130. Sakoparnig, T., Field, C. & Nimwegen, E. van. Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species. *eLife* **10**, e65366 (2021).
131. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**, 132 (2016).
132. Brynildsrud, O., Bohlin, J., Scheffer, L. & Eldholm, V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol* **17**, 238 (2016).
133. Lees, J. A., Galardini, M., Bentley, S. D., Weiser, J. N. & Corander, J. pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics* **34**, 4310–4312 (2018).
134. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).
135. Colombi, E. *et al.* Comparative analysis of integrative and conjugative mobile genetic elements in the genus Mesorhizobium. *Microbial Genomics* **7**, (2021).
136. ICEberg 3.0: functional categorization and analysis of the integrative and conjugative elements in bacteria. https://tool2-mml.sjtu.edu.cn/ICEberg3/index.html.
137. Lu, F. & Churchward, G. Tn916 target DNA sequences bind the C-terminal domain of integrase protein with different affinities that correlate with transposon insertion frequency. *J Bacteriol* **177**, 1938–1946 (1995).
138. Garcillán-Barcia, M. P., Redondo-Salvo, S., Vielva, L. & De La Cruz, F. MOBscan: Automated Annotation of MOB Relaxases. in *Horizontal Gene Transfer* (ed. De La Cruz, F.) vol. 2075 295–308 (Springer US, New York, NY, 2020).
139. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402 (1997).
140. Abby, S. S. *et al.* Identification of protein secretion systems in bacterial genomes. *Sci Rep* **6**, 23080 (2016).
141. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Res.* **14**, 1394–1403 (2004).
142. pycountry: ISO country, subdivision, language, currency and script definitions. https://pypi.org/project/pycountry/.
143. geopy 2.4.1: geocoding python client. https://geopy.readthedocs.io/en/stable/.
144. France-geojson: Martinique communes. https://github.com/gregoiredavid/france-geojson/blob/master/regions/martinique/communes-martinique.geojson.
145. Joris Van den Bossche *et al.* geopandas/geopandas: v0.14.3. Zenodo https://doi.org/10.5281/ZENODO.10601680 (2024).

Acknowledgements

We acknowledge the support of the Max Planck Society in funding the Independent Research Group on Plant Pathogen Evolution (H.C.M.) and for providing an International Max Planck Research School Ph.D. scholarship (D.E.). Research was supported in part by grant NSF PHY-2309135, the Gordon and Betty Moore Foundation Grant No. 2919.02, and the Chan Zuckerberg Initiative DAF grant to the Kavli Institute for Theoretical Physics (KITP). This work was performed in accordance with the Nagoya Protocol on Access and Benefit Sharing (ABSCH-IRCC-FR-254778-1). We thank Gal Ofir and Sheila Roitman for helpful discussion, Ludovico Calabrese for assistance with Recophy and members of the McCann lab for discussion and comments on the manuscript.

Author contributions

D.E., Y.P., S.P., E.W. and H.C.M designed research, D.E., Y.P., M.K., C.W., C.F. and M.V.-Z. performed research; D.E., M.K., E.C., Y.F.C., E.W. and H.C.M. analysed data; D.E. and H.C.M wrote the paper; all authors commented upon and/or approved the manuscript.