Title

Emergence and host range expansion of an epidemic lineage of *Ralstonia solanacearum*

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

Members of the *Ralstonia solanacearum* species complex are soil-borne pathogens of important crop plants. Strain-level host specialisation is considered typical, yet outbreak monitoring in Martinique revealed the emergence of a new lineage with dramatically expanded host range. We show the outbreak was caused by the parallel expansion of lineages from two different *Ralstonia* species. One lineage of Asian origin retains a signature of host specialisation, while the emergent broad host range lineage originated in the Americas. The presence of similar integrative and conjugative elements in both lineages suggests they are exchanging genetic material. The emergent lineage acquired genes encoding novel catalytic and metabolic functions, but shows little variation among virulence genes considered host specificity determinants. The host range expansion event was not caused by the evolution of a virulence strategy targeting host immunity, but is more likely due to the acquisition of novel substrate utilisation pathways enhancing pathogen fitness in soil or plant tissues.

Disease outbreaks on crop plants cause severe losses for subsistence and commercial growers alike. Clarifying the mechanisms underlying disease emergence is of fundamental importance for safeguarding human health and agricultural productivity, and deepens our understanding of how microbes adapt to novel hosts and environments. Studies of the molecular mechanisms by which bacterial plant pathogens colonise and disrupt host immune defences have revealed toxins, cell wall degrading enzymes and virulence proteins secreted by the type 3 secretion system (T3SS) allow pathogens to establish infections. The identification and characterisation of pathogen virulence and host resistance genes underlying host-specific interactions provides insight into molecular mechanisms underlying successful infections, but the genetic determinants of host range remain undiscovered for the majority of plant pathogens and the extent to which these mechanisms are linked with major outbreak events remains unclear. Limited genomic data for plant pathogen populations sampled during prior outbreaks shows lineages isolated from different species have variable sets of virulence genes, suggesting horizontal gene transfer and host-mediated selection are critical for pathogen specialisation. We have very little understanding of the evolutionary processes and genomic changes associated with outbreak events however, and it is not clear how generalizable these findings are across bacterial plant pathogens, which remain relatively understudied despite the threat posed to agricultural health and food security.

The identification of a new lineage of *R. solanacearum* in the Caribbean island of Martinique represents a rare opportunity to investigate the emergence of a bacterial plant pathogen of global economic importance. We sought to identify the origins of an outbreak and the particular genomic changes associated with a host range expansion event by sequencing the largest collection of *Ralstonia* spp. sampled during an outbreak event. Multiple lineages of *R. solanacearum* are known to infect banana, which is cultivated on an industrial scale in the Caribbean, Central and South America. *R. solanacearum* is subject to extensive monitoring and movement controls to limit its spread, since it causes lethal wilting in susceptible hosts, disseminates rapidly and can persist for years in freshwater and soil. An outbreak of disease on cultivated ornamental plants (*Anthurium andreanum*) in Martinique in 1999 caused alarm when it was found to be caused by a lineage that is closely related to Rso2BS4, a banana-infecting lineage [1]. Additional infections on the island were soon reported in many hosts which *R. solanacearum* had not been previously known to infect [2]. The threat posed to banana cultivation and remarkable 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 [3], [4]. This work generated a remarkable collection of strains isolated from infected crops, asymptomatic weeds, soil and freshwater environments. We sequenced this collection and found the outbreak was caused by the expansion of two co-occurring, distantly related lineages. Each lineage was disseminated from well-established populations on the 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. Mobile elements are circulating between both lineages, particularly in areas where they co-occur. These mobile elements serve as vectors for genes that may confer novel metabolic functions. We found few changes in secreted virulence proteins, and postulate the host range expansion event is not due to the acquisition of genes linked with host defence manipulation, but is rather due to pathogen adaptation to the agricultural environment.

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 single isolate genomes sampled from crops, weeds, soil and freshwater environments 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, and the outbreak of infections documented in Martinique was caused by the expansion of two lineages from each species: Rps1S18 and Rso2BS4NPB (Fig. 1, Fig. S1). The center of origin of Rso and Rps is in the Americas and Asia, respectively. Multiple inter-continental dissemination events of both Rso and Rps lineages are evident, including global outbreak of Rso2BS1 on potato (*S. tuberosum*) and multiple introductions of distinct Rps lineages from Asia into the Americas (Fig. 1, Fig. S1).

The introduction of Rps lineages into the Americas resulted in their establishment and local diversification, primarily in solanaceous hosts. These introductions may have occurred via multiple sources, including the movement of infected plant material. Rps1S18 is the most frequently isolated Rps lineage in the Americas (Fig. 1, Fig. S1). In contrast, Rso has been diversifying in the Americas for some time. The Rso2A clade encompasses the greatest intraspecific diversity of all *Ralstonia* species and lineages sampled in French Guiana and Martinique. There is no evidence of recent clonal expansion among Rso2A lineages however, which instead form a large clade of primarily tomato-infecting strains (68%) sampled across a broad geographic range (Fig. 1, Fig. S1). Considerable diversity is also evident among Rso2B isolated from potato in Peru, Colombia and Brazil. The origin of cultivated potato is in Peru and wild relatives are distributed across the Americas, indicating Rso2B may have a long coevolutionary history with these hosts in the Americas [5].

Multiple Rso2B lineages are responsible for both regional and global outbreak events, most notably on banana (*Musa* spp.) and potato, commodity and subsistence crops of global importance. Rso2BS4NPB has the greatest documented host range of any lineage in the species complex, despite being primarily restricted to Martinique. In contrast, Rso2BS1 (formerly race 3 biovar 2), is responsible for a global outbreak on a single host (potato). Rso2BS1 has also been isolated from other *Solanum* spp., *Pelargonium* spp. (formerly *Geranium* spp.), surface water and soil. Although the centre of origin of Rso2BS1 appears to be in South America, its rapid global expansion suggests international trade in seed potatoes is the likeliest mechanism of long-distance dissemination, highlighting the elevated risk of disease transmission during movement of infected plant material [6] [7].

Outbreaks of Rso on banana are well documented in the Americas, where it is referred to as Moko disease. Moko disease is clearly polyphyletic, as it includes four Rso2A lineages primarily isolated in Brazil; two basal lineages in the Rso2B clade from which the global potato outbreak lineage emerged (Rso2BS25 [8]; Rso2BS3 [9]), and basal members of Rso2BS4. The latter is closely related to the lineage that underwent host range expansion in Martinique (Rso2BS4NPB). These banana-infecting

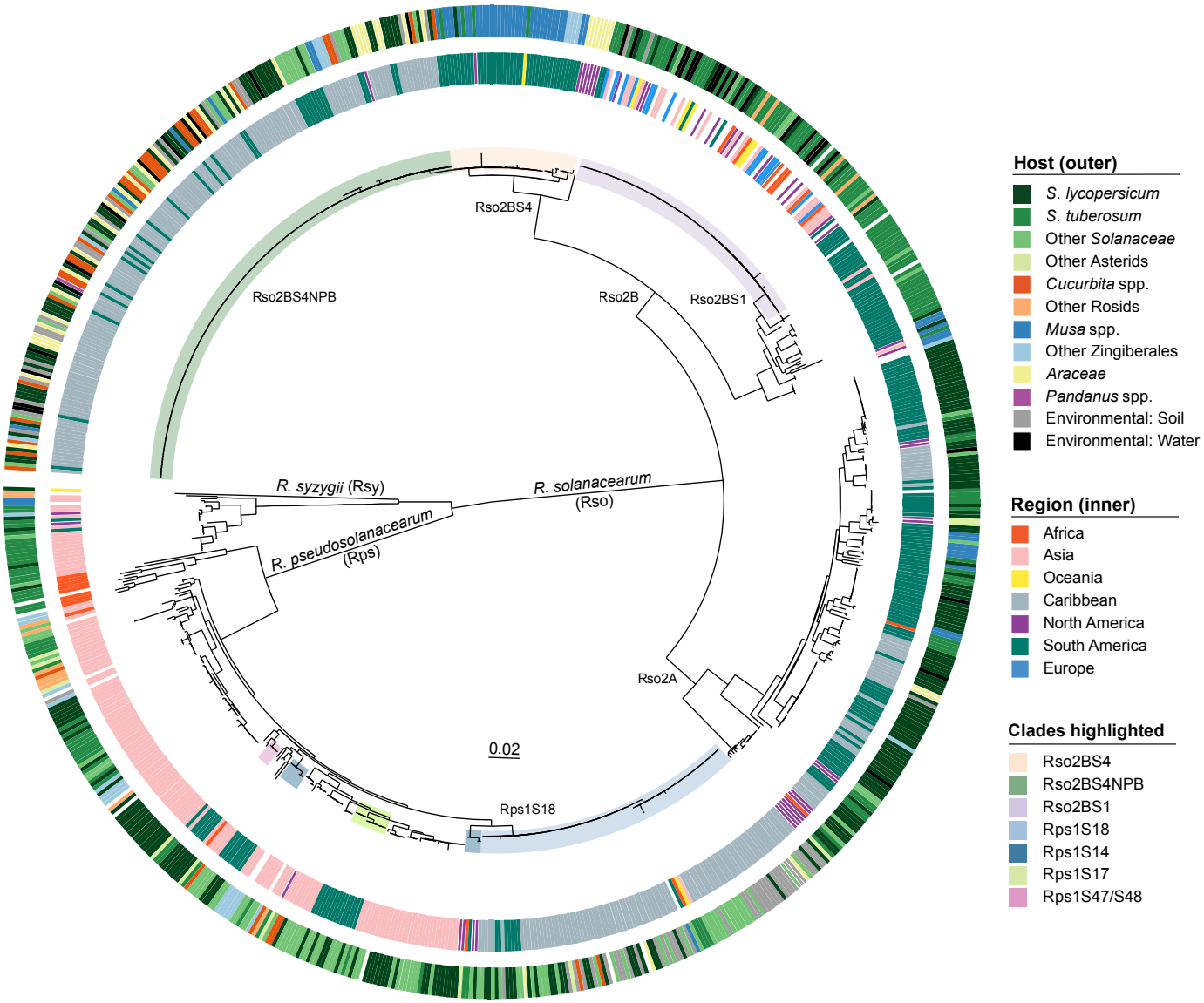
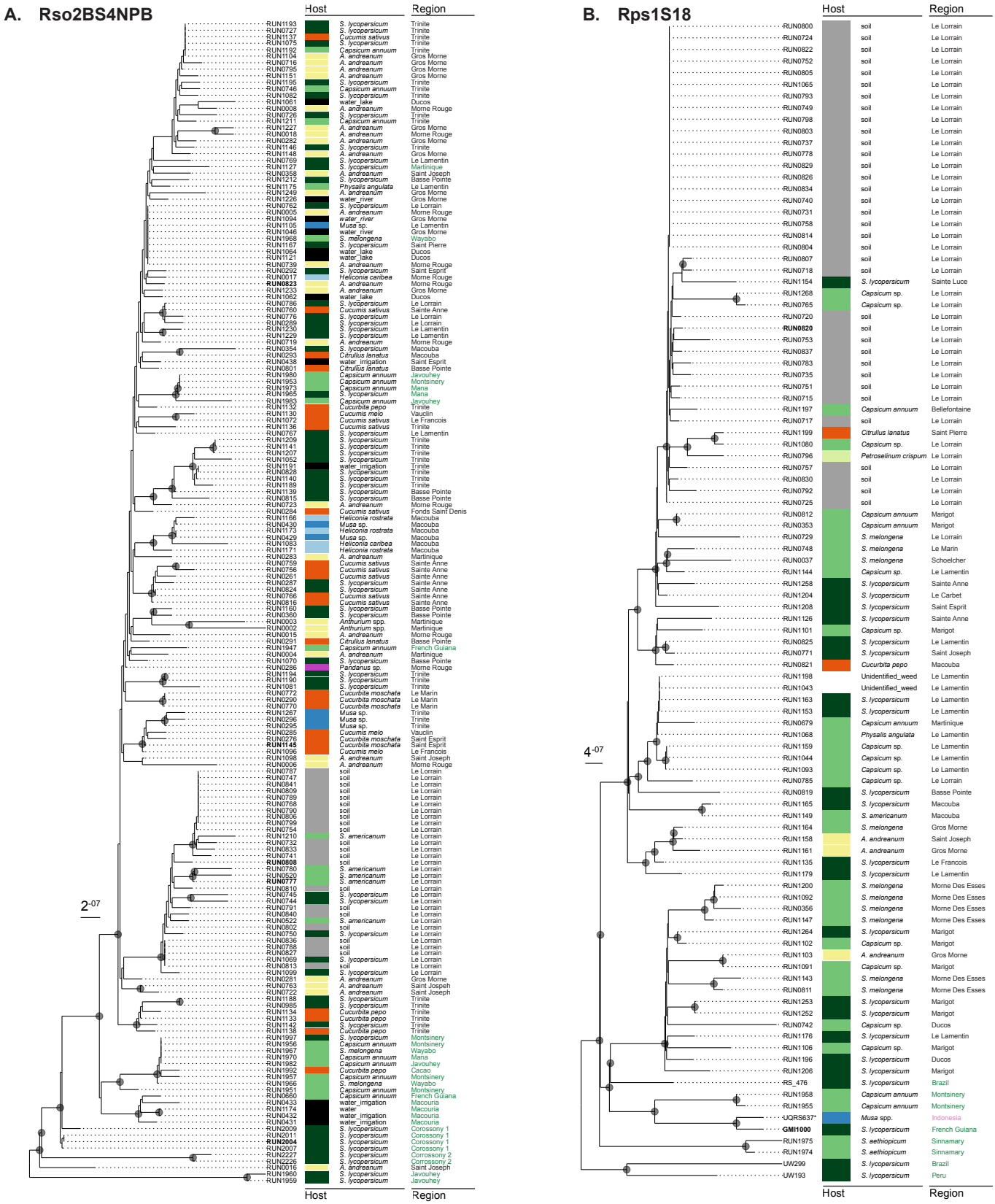


Fig. 1. Core gene tree of the *Ralstonia solanacearum* species complex. Maximum likelihood tree of 816 RSSC genomes inferred from a 137,423bp SNP alignment obtained after core gene identification, alignment generation and invariant site removal. Nodes with the bootstrap support values under 70 are collapsed. 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.

134 lineages are collectively responsible for outbreaks in Panama, Costa Rica, Colombia, Peru, Brazil and
135 French Guiana from at least the 1960s onwards (Fig. S2). Banana was introduced to the Americas as a
136 crop in the 16th century [10], so it is a matter of some interest that the most deeply branching basal
137 members of Rso2B are isolated from *Musa* spp. along with *Heliconia* spp. and *Anthurium* spp., both
138 endemic to the Americas (Fig. 1, Fig. S2) [11], [12]. The expansion of high-density single variety
139 banana monocultures is likely to offer multiple opportunities for Rso to emerge and spread in
140 agricultural areas, in turn increasing likelihood of spread to other crop varieties, both endemic and
141 introduced.

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

143 The species complex tree shows the surge in plant infections in Martinique is linked with the recent
144 clonal expansion of two lineages across the island: Rso2BS4NPB and Rps1S18 (Fig. 1). The latter lineage
145 is historically significant, as it is the source of the first sequenced *Ralstonia* genome and commonly used
146 model strain GMI1000, isolated in French Guiana in 1978. We investigated the origins of both lineages
147 using well resolved, non-recombinant core genome trees including closely related strains from mainland
148 populations (Fig. 2). South American mainland strains most closely related to the emergent Rso2BS4NPB
149 lineage in Martinique were all isolated from solanaceous hosts or freshwater sources on the mainland,
150 along with a single isolate from pumpkin (*Cucurbita pepo*). These basal isolates all infect hosts that
151 evolved in the Americas and represent the ancestral population from which Rso2BS4NPB emerged to
152 sweep across Martinique. Two strains (RUN2226 and RUN2227) were isolated from tomato plants grown
153 in areas planted the first season after rainforest clearing, suggesting the ancestral population is
154 established in uncultivated areas and has a long coevolutionary history with endemic host species. There
155 was evidently limited spread of Rso2BS4NPB on the mainland compared to the dramatic epidemic
156 expansion in Martinique. The derived lineage in Martinique exhibits limited genetic diversity, differing
157 by only 166 nonrecombinant SNPs in the core genome. A single Martinique isolate (RUN0016) isolated
158 from *A. andreanum* in Saint Joseph, Martinique groups with basal mainland isolates (Fig. 2, Fig. S3). The
159 detection of additional derived Rso2BS4NPB isolates from the same host species in Saint Joseph
160 suggests this location may be linked with the initial introduction of the emergent lineage in Martinique
161 (Fig. 2). Most Rso2BS4NPB isolates from Le Lorrain cluster together, regardless of whether they were
162 isolated from soil, cultivated or wild *Solanum* spp. Notably, Rso2BS4NPB was isolated from symptomatic
163 banana (*Musa* spp.) in Trinite, indicating the appellation 'Non-Pathogenic on Banana' (NPB) for this
164 lineage may not be wholly accurate.
165
166 The evolutionary history of Rps1S18 parallels that of Rso2BS4NPB: Martinique Rps1S18 is similarly
167 derived from a mainland population circulating on solanaceous hosts (Fig. 2, Fig. S4). Despite the large
168 number of sequenced Rps1S18 isolates, the lineage displays limited diversity in the core genome and
169 T3E repertoire, consistent with recent epidemic expansion. There is no evidence of a dramatic host
170 range expansion however: with few exceptions, the majority of Martinique Rps1S18 was isolated from
171 soil or solanaceous hosts. Unexpectedly, a *Musa* isolate UQRS637_UW745 from Indonesia was assigned
172 to the Rps1S18 lineage. This should be interpreted with caution however, as the assembly version of
173 UQRS637_UW745 used here was later suppressed on NCBI. There is evidence of structuring by sampling


A. Rso2BS4NPB
Host
Region
2-07
B. Rps1S18
Host
Region
4-07
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
Asia
Pangenome
Core genes
Soft core genes
Shell genes
Cloud genes
C. Pangenome fraction
Rso2BS4NPB
Rps1S18

175 **Fig. 2. Outbreak lineage core genome trees.** (A) Rso2BS4NPB maximum likelihood tree calculated from a
176 5,206,677 bp non-recombinant non-mobile core genome alignment with 195 SNPs and 73 parsimony informative
177 sites. Internal nodes with bootstrap value above 80 are indicated with black circles. Strain labels shown in bold are
178 are complete or near complete assemblies generated in this study and/or used as mapping references. (B) Rps1S18
179 maximum likelihood tree calculated from a 4,760,272 bp non-recombinant core genome alignment with 311 SNPs
180 and 116 parsimony informative sites. The asterisk for UQRS637 refers to a duplicated identical strain with a
181 different label: UW745. Strain labels shown in bold are closed genomes generated in this study and/or used as
182 mapping references. (C) Pangenome waffle chart representing the proportion of core (present in 99 - 100% of
183 genomes), soft-core (95 - 99% of genomes), shell (15 - 95% of genomes) and cloud genes (present in less than 15%
184 of genomes) identified by Panaroo, displayed separately for each lineage pangenome.
185 —
186
187 location for one site: just as in the Rso2BS4NPB tree, Rps1S18 isolates from Le Lorrain form a well-
188 supported clade. Most Le Lorrain isolates - whether Rps1S18 or Rso2BS4NPB - were recovered from soils
189 (Fig. 2). In contrast, freshwater isolations yielded solely Rso2BS4NPB.
190
191 We used well-resolved non-recombinant trees to determine the divergence time of Rps1S18 and
192 Rso2BS4NPB. The best estimated divergence times for the split between basal solanaceous and derived
193 host-range expanded Rso2BS4NPB sublineage is between the 1930s and 1980s (Fig. S5). Estimates were
194 also generated for Rps1S18 (1850s - 1930s), yet there was no temporal signal in the data (not shown).
195 The estimated divergence times for Rso2BS4NPB are also poorly supported, even after removing
196 potential outlier samples and collapsing weakly supported nodes in the phylogenies. The R-squared
197 value for root-to-tip regression is 0.05, and BactDating random date permutation analyses showed
198 estimated divergence times were better than random in only 50% of simulations, indicating there is no
199 temporal signal in the data. Pangenome analysis for both Rps1S18 and Rso2BS4NPB showed both
200 lineages have similar numbers of core and soft-core genes (4,565 for Rps1S18 and 4,792 for
201 Rso2BS4NPB), but remarkably different accessory genome sizes: 487 accessory genes in Rso2BS4NPB
202 and 2,636 in Rps1S18, amounting to 10% and 35% of each lineage pangenome, respectively (Fig. 2). The
203 larger Rps1S18 accessory genome suggests this lineage has higher rates of horizontal exchange, or that
204 it was introduced to Martinique prior to Rso2BS4NPB, and therefore had more time to acquire accessory
205 genes.

206 **Minimal changes in virulence gene repertoires linked with host range expansion**

207 Despite reduced genetic diversity indicative of very recent expansion, the Rso2BS4NPB isolates from
208 Martinique encompass the broadest known host range of any single lineage in the species complex (Fig.
209 2). We sought to identify the genes underlying host range expansion in Rso2BS4NPB by assessing
210 changes in the repertoire of Type 3 secreted effectors (T3Es). The delivery of virulence proteins by the
211 Type 3 secretion system (T3SS) is known to play an important role in successful host colonisation, and
212 the repertoire of T3Es carried by a strain is thought to play an important role in host specificity [13],
213 [14], [15]. We predicted the T3E repertoires in each genome and found no link between the emergence
214 of Rso2BS4NPB and any specific T3E gain or loss (Fig. 3, Fig. S1, Table S4). Rso2BS4NPB genomes display
215 nearly uniform T3E repertoires, consistent with recent population expansion (Fig. S2). A single T3E is

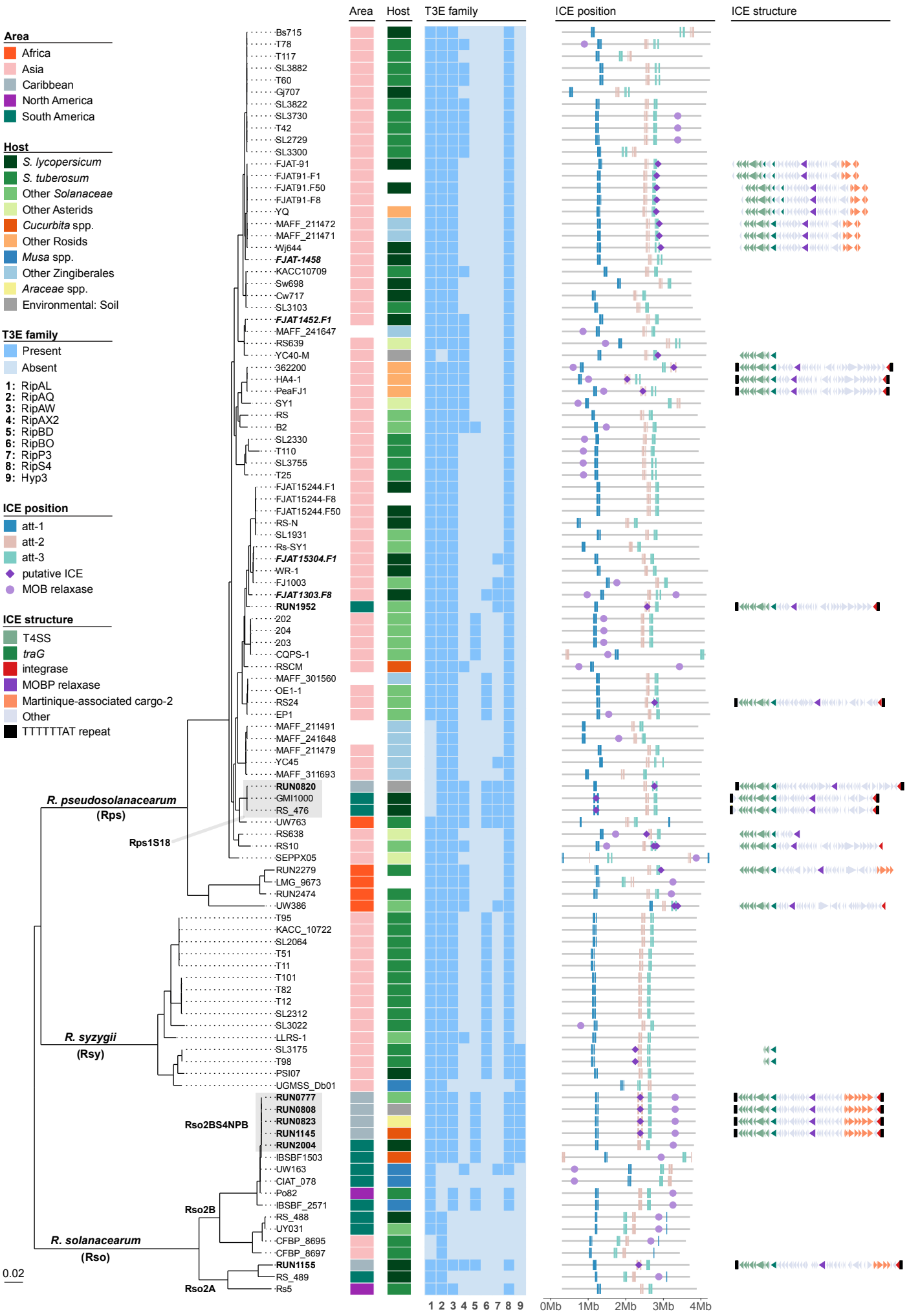

Area
Host
T3E family
ICE position
ICE structure
Area
Africa
Asia
Caribbean
North America
South America
Host
S. lycopersicum
S. tuberosum
Other Solanaceae
Other Asterids
Cucurbita spp.
Other Rosids
Musa spp.
Other Zingiberales
Araceae spp.
Environmental: Soil
T3E family
Present
Absent
1: RipAL
2: RipAQ
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
Martinique-associated cargo-2
Other
TTTTTTAT repeat
R. pseudosolanacearum
(Rps)
Rps1S18
R. syzygii
(Rsy)
Rso2BS4NPB
Rso2B
R. solanacearum
(Rso)
Rso2A
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

Fig. 3. Virulence gene and ICE distribution across the *R. solanacearum* species complex. Maximum likelihood tree of all completely assembled RSSC genomes available in this study. Strain labels shown in bold are complete or near complete assemblies generated in this study and/or used as mapping references. Strain labels shown in boldface italics represent clades collapsed due to low diversity. The geographical area and host of isolation are shown next to strain labels, followed by nine *Ralstonia* T3E families exhibiting presence/absence variation for Rso2BS4 and Rps1S18 clades. Dark blue in the T3E family column represents the presence of the T3E family in a strain and light blue represents its absence. Panel D shows the position of all attachment sites (here defined as regions syntenic to the 50kb up- and downstream flanking regions of the Rps1S18 GMI1000 (att-1, blue), Rso2BS4NPB RUN1145 (att-2, pink) and Rso2BS4NPB RUN1960 (att-3, green) attachment sites on chromosome 1 of each RSSC genome. Thin grey lines represent the length ranges of the primary chromosomes. MOB relaxases occurring in isolation (purple circles) are separately indicated from those adjacent to T4SS genes and forming putative ICEs (purple diamonds). The final panel shows the organization of putative ICEs identified in RSSC strains, with homologs of various functional categories highlighted in different colours. Black bars around ICEs represent the TTTTTTAT repeat motif delimiting putative ICE boundaries.

—

uniquely present in the emergent lineage and its sister lineage Rso2BS4 and absent in all other RSSC genomes: the hypothetical effector *T3E_Hyp13* (UW163_19700 in Rso2BS4 UW163, OFBKGH_20220 in Rso2BS4NPB RUN1145) (Fig. S1). Another T3E (*ripAX1*) is sufficiently divergent from other members of the effector family to be classified as a clade-specific gene by Panaroo pangenome analysis. BLASTn searches reveal that *ripAX1* is present in most strains across RSSC, but absent in *Heliconia* and *Musa* RsoS4 isolates from Colombia, other Rso2B *Musa* isolates and some *Rsy* strains (Fig. S2). In fact, while there is no evidence variation in T3E carriage is linked with the emergence and host range expansion of Rso2BS4NPB, T3E exchange may play a role in Rso2BS4 pathogenicity on banana. The Rso2BS4 isolates from cultivated endemic (*Heliconia* spp.) and introduced (banana, *Epipremnum* spp.) monocot hosts that are endemic and introduced (banana and) to the Americas exhibit more variation in effector repertoire than Rso2BS4NPB (Fig. 3, Fig. S2). Two effectors are present in Rso2BS4 and absent in Rso2BS4NPB: *ripAX2* and *ripP3*. The *ripAX2* gene is present in many other Rso2B and Rso2A clades, indicating loss in Rso2BS4NPB is likely, while *ripP3* is sparsely distributed in Rso2B. Four effector families exhibit evidence of likely loss in monocot-infecting Rso2BS4: *ripAQ*, *ripAW*, *ripBD* and the predicted *T3E_Hyp3* effector. The loss of *ripAQ* is most striking as the family is otherwise present in all other RSSC strains. The *ripAW* gene is missing from most Rso2BS4 banana isolates from Peru, most pandemic Rso2BS1 potato isolates and some *R. syzygii*, but is present in all other RSSC strains. Peruvian Rso2BS4 strains also lack *ripBD*, an effector otherwise common in Rso2A. *T3E_Hyp3* are absent from Rso2BS4 and patchily distributed in other Rso2B lineages. Since these effector families are present among related clades and/or across the species complex, the most parsimonious explanation for the pattern observed is they were lost in the ancestor of banana-infecting Rso2BS4 strains rather than being acquired by ancestral Rso2BS4NPB. Their link with the pathogenicity of Rso2BS4 on monocots *Musa* spp., *Heliconia* spp. and *Epipremnum* spp. is not known.

Ancestral acquisition of novel metabolic functions in the emergent lineage

Since the host range expansion of Rso2BS4NPB is likely to have occurred as a result of genomic changes in the ancestor of this lineage, we explored whether recombination events introducing novel genes

259 might be linked with the emergence event. Pangenome exploration revealed 84 genes were uniquely
260 acquired in the ancestor of Rso2BS4 and Rso2BS4NPB (Table S3). The acquired genes include enzymes
261 such as a TauD/TfdA family dioxygenase, which has been shown to degrade phenoxy herbicides in soil-
262 borne bacteria like *Burkholderia* and *Ralstonia eutropha* [16], [17], [18], [19], [20], [21]; a CxxC motif
263 protein (UW163_01710, OFBKGH_14170) associated with thiol-disulfide oxidoreductase activity and
264 oxidative damage responses [22], [23]; an electron acceptor flavodoxin (UW163_18870,
265 OFBKGH_23185) also involved in oxidative damage response [24]; and a carboxymuconolactone
266 decarboxylase (UW163_18885, OFBKGH_23170) in the protocatechuate catabolism pathway, linked
267 with conversion of lignin-derived aromatic compounds into tricarboxylic acid intermediates like
268 succinate and malate [25], [26]. The last two genes are located within 2.25kb on the secondary
269 chromosome of Rso2BS4 UW163. These metabolism-associated genes are involved in distinct reactions
270 and pathways, but are variously involved in responses to oxidative stress and aromatic compound
271 catabolism.

272 **Recombination between co-occurring pathogen lineages**

273 Since the co-occurrence of two pathogen lineages in the same host or niche provides opportunities for
274 between-lineage homologous recombination, we then explored the possibility Rps1S18 might serve as a
275 donor for novel genetic material for Rso2BS4NPB. We found no evidence of homologous recombination
276 between Rso2BS4NPB and Rps1S18 however (Fig. S6). RECOPHY identified some regions with reduced (<
277 10/kb) and elevated (120/kb) SNP density compared to the alignment-wide average of 53 SNPs/kb, yet
278 distance trees for those regions display the same topology as the core genome tree, indicating
279 homologous recombination between these lineages is unlikely to have occurred. Homologous
280 recombination is nevertheless occurring with unknown donors, as ClonalFrameML analysis for each
281 lineage indicates low but similar relative ratios of recombination to mutation (R/theta, 0.159 for
282 Rps1S18 and 0.152 for Rso2S4NPB).
283
284 Mobile elements constitute another means by which genetic material can be exchanged between
285 bacterial pathogens. Pangenome exploration and association testing revealed the presence of
286 integrative and conjugative elements (ICEs) circulating between the co-occurring lineages in Martinique
287 (Fig. 3). All but one of the 27 genes associated with isolation in specific regions in Martinique are located
288 on an ICE in Rso2BS4NPB; a kmer-based association test similarly showed 625 out of 2,367 Martinique-
289 associated kmers and all 83 soil-associated kmers map to the same ICE (Fig. S7, Table S11, Table S12). A
290 precise association between a cluster of eight ICE genes with isolation from Trinite and Morne Rouge
291 was found, while another two ICE gene clusters have overlapping associations with isolation from soil
292 and from Le Lorrain (Fig. S7). The gene clusters linked with specific regions of isolation correspond to
293 backbone regions and hotspots of cargo gene integration (cargo-1 and cargo-2) on the ICE (Fig. 4).
294
295 Since pangenome exploration and association tests showed Rso2BS4NPBB acquired an ICE, we placed
296 this in the context of *Ralstonia* ICE diversity by predicting all ICEs in the species complex using all 133
297 available complete assemblies (Fig. 3). We identified a total of 30 ICEs, of which 17 lack direct repeats
298 but satisfy other criteria for ICE identification (presence of T4SS gene clusters, a *mob* relaxase and

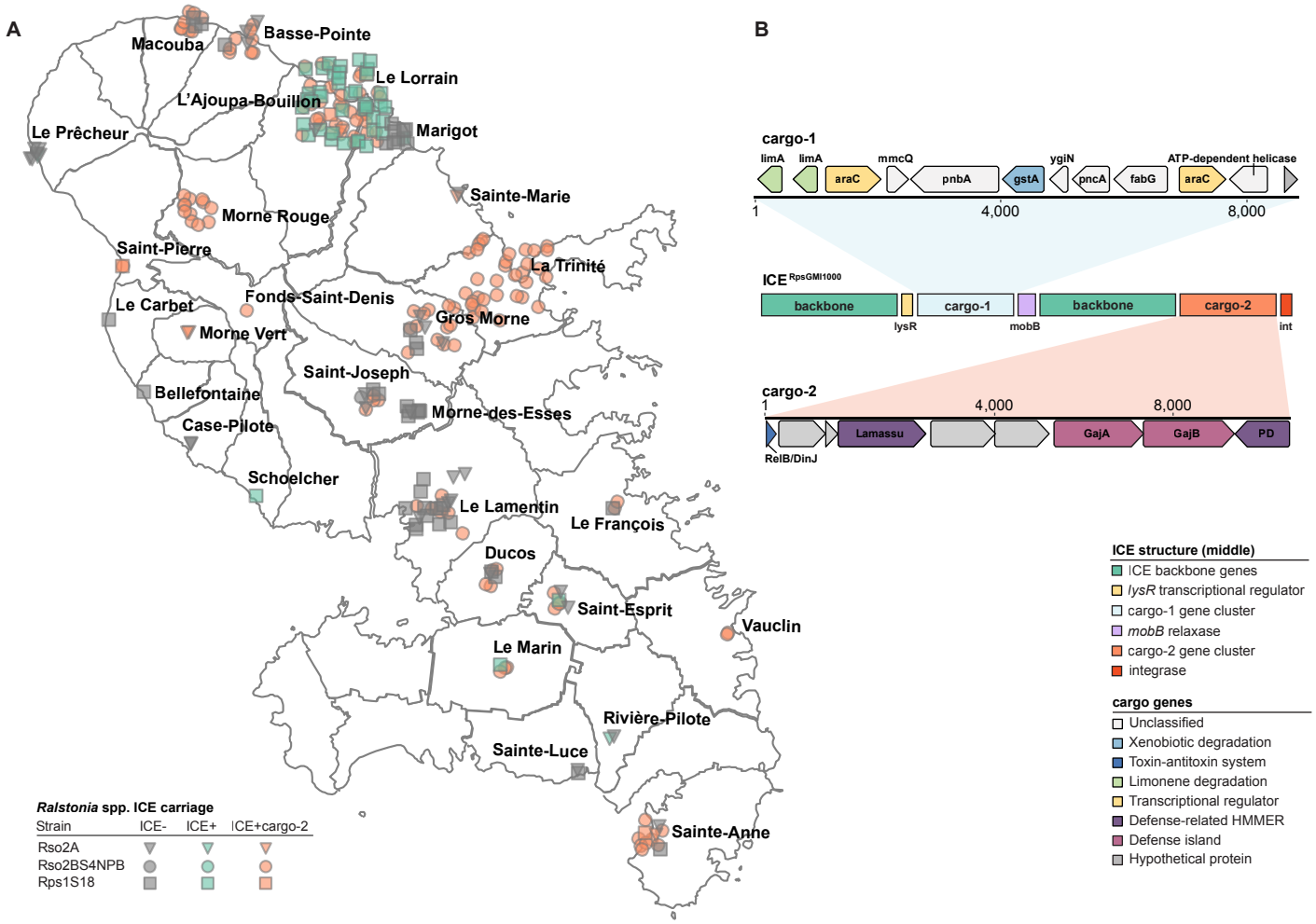


Fig. 4. *Ralstonia* ICE distribution and cargo gene carriage in Martinique. (A) Sampling location of all Martinique isolates sequenced in this study. The exact longitude and latitude of the sampling locations were jittered for the visualisation purpose. Rso2A (triangle), Rso2BS4NPB (circle), Rps1S18 (square) are represented with as grey (ICEless), green (ICE^{RpsGMI1000} with cargo-1) or orange (ICE^{RpsGMI1000} with cargo-1 and cargo-2) icons. (B). A diagram of ICE^{RpsGMI1000} structure is shown in the center, delineating the position of the backbone and cargo gene clusters. The coding sequences in each cargo region are shown above (cargo-1) and below (cargo-2) the diagram of ICE structure, coloured according to KEGG annotation. The first backbone region includes the T4SS and has a *lysR* family transcriptional regulator at the 3' end, followed by cargo-1. The second backbone region includes the *mob* relaxase, *traF*, *parA*, *repA*, *qseC* and *rdfS;* genes encoding a hypothetical protein and helix-turn-helix domain-containing protein, and DUF domain-containing proteins (DUF3800, DUF945, DUF2285 (*fseA*), DUF2840, DUF2958).

308 integrase sequence). Some genomes have *mob* relaxase genes flanked by direct repeats but lack
309 adjacent T4SS: these may be degenerate ICEs or integrative mobilizable elements (IMEs). Each genome
310 typically carries only a single ICE, integrated into one of three preferred chromosomal integration sites.
311 ICEs are more common among *R. pseudosolanacearum* (23/30) than *R. solanacearum* (5/30) and *R.*
312 *syzygii* (2/30). The five ICEs in *R. solanacearum* are exclusively found in Martinique isolates: four of these
313 are in Rso2BS4NPB and the other is in Rso2A RUN1155 (Fig. 3). The Rso2BS4NPB and Rso2A ICEs are
314 identical to each other and homologous to a 45.9kb ICE in Rps1S18 GMI1000. This particular ICE is also
315 present in seven other Rps genomes that collectively span the diversity of *R. pseudosolanacearum*. Since
316 this ICE is broadly distributed among Rps genomes and present in the reference Rps1S18 GMI1000, we
317 named it $ICE^{RpsGMI1000}$.
318
319 We then assessed the spatial structure of ICE carriage among the entire set of complete and draft
320 genomes representing the densely sampled Martinique population and found local hotspots of ICE
321 acquisition in Martinique (Fig. 4, Fig. S8). Identification of $ICE^{RpsGMI1000}$ presence and accessory cargo
322 carriage in every isolate shows that while there are multiple regions in which both lineages co-occur, Le
323 Lorrain Rps1S18 and Rso2BS4NPB isolates carry ICEs both with and without the cargo-2 gene cluster.
324 $ICE^{RsoGMI1000}$ acquisition by Rso2BS4NPB may have occurred in this region, followed by its dissemination
325 to other areas across the island, either in tandem or independent of host strain movement (Fig. 4, Fig.
326 S8).
327
328 The existence of recombination hotspots, where mobile elements appear more frequently exchanged
329 between bacterial genomes, is mirrored by the existence of recombination hotspots within the mobile
330 elements themselves. The $ICE^{GMI1000}$ has two cargo regions, named cargo-1 or cargo-2 according to their
331 integration site within the ICE. We found that while the cargo gene clusters vary, their overall function
332 does not: cargo-1 gene clusters always encode metabolism-associated genes and cargo-2 gene clusters
333 always encode bacterial defence systems (Fig. S9, Fig. S10). The cargo-1 region present in Rps1S18 and
334 Rso2BS4NPB (e.g. RUN0777 and RUN1960) includes two *limA* genes encoding limonene 1,2-epoxide
335 hydrolases [27] (OFBKGH_09550, OFBKGH_09555 in RUN1145); a phthalate ester hydrolase *pnbA*; an
336 antibiotic biosynthesis monooxygenase *ygiN* (also known as ABM, OFBKGH_09580); a glutathione S-
337 transferase *gstA* linked with the detoxification of organochlorine compounds (1-chloro-2,4-
338 dinitrobenzene) used as an herbicide and pesticides; a pyrazinamidase/nicotinamidase *pncA*, which
339 converts nicotinamide to nicotinic acid, and *fabG*, which encodes a reducing agent of ethyl 4-
340 chloroacetoacetate, an intermediate used in herbicide and pesticide production. The two distinct cargo-
341 2 clusters carried by Rps1S18 GMI1000 $ICE^{RpsGMI1000}$ and Rso2BS4NPB $ICE^{RpsGMI1000}$ each carry multiple
342 defence systems, including the anti-plasmid/phage Lamassu island and anti-viral Gabija islands [28],
343 [29]. The only other ICE present in a Rso genome (Rso2A RUN1155, also from Martinique) shares a
344 similar cargo-2 region with Rso2BS4NPB $ICE^{GMI1000}$ (Fig. 3).

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

346 The identification of an ICE shared between Rsp1S18 and Rso2BS4NPB along with multiple acquisition
347 events in the ancestor of Rso2BS4NPB from unknown donors suggests multiple recombination

348 mechanisms are introducing variation into the emergent lineage. Since prophage are known to
349 introduce accessory genes and modify bacterial host fitness, we queried the collection using the
350 PHASTEST phage prediction tool and found evidence of multiple insertions in the ancestor of the
351 emergent lineage. Interestingly, one of the three prophage identified is uniquely present in Rso2BS4NPB
352 and absent from all other sequenced RSSC genomes (Table S4, Table S7). 63 ortholog groups identified
353 as unique to Rso2BS4NPB in Panaroo pangenome analyses are located within the prophage located in
354 chromosome 1 (2,599,979 - 2,636,445 in Rso2BS4NPB RUN1145) as well as a 18kb region directly
355 downstream. The closest BLASTn hits include are soil isolates (*Ralstonia* metagenome-assembled
356 genome from estuary sediment in China, *Pandoraea pnomenusa* from sludge in Malaysia and *Massilia*
357 *putida* from soil in China), as well as *Xanthomonas albilineans*, a bacterial plant pathogen causing lethal
358 leaf scald in sugarcane from Brazil (Fig. S11). In addition to the core genes encoding the phage head, tail,
359 baseplate and restriction endonucleases, the Rso2BS4NPB-specific phage carries auxiliary genes whose
360 products are predicted to function as a circadian clock protein (*kaiA*, OFBKGH_12090 in RUN1145) and a
361 methyl-accepting chemotaxis protein (OFBKGH_12085). A 17.7Kb region downstream of the phage that
362 is also unique to Rso2B4NPB contains a flagellar motor switch encoding gene *fliM* (OFBKGH_12305). The
363 strong association between this prophage and genomic changes in the emergent lineage was
364 additionally confirmed by Pyseer kmer-based association test, as 1,007 out of 2,367 kmers associated
365 with isolation from Martinique in Rso2BS4NPB RUN1145 genome are mapped to the prophage and its
366 downstream auxiliary region (Fig. 5, Table S9). The other two intact prophages are also located on the
367 primary chromosome, and are present in Rso strains from Central and South America (Table S7). Further
368 prophage prediction in all 12 complete Rso2B genomes shows all Rso2BS4NPB (RUN0777, RUN0823,
369 RUN0808, RUN1145 and basal RUN2004) have three intact and one partial prophage insertion on the
370 primary chromosome, while three of four Rso2BS4 (UW163, Po82 and IBSBF1503) have at most a single
371 intact prophage on the primary chromosome and a single prophage on the secondary chromosome. The
372 exception is Rso2BS4 CIAT_078 (*Musa acuminata x balbisiana* AAB, Colombia, 2004), which has no intact
373 prophage predictions on the primary chromosome.
374
375 Alignment of complete assemblies for Rso2BS4 UW163 (from symptomatic *Musa acuminata* x *balbisiana*
376 AAB, Peru, 1967) and Rso2BS4NPB RUN1145 (from symptomatic *Cucurbita moschata*, Martinique, 2002)
377 confirms horizontal gene transfer mediated by the movement of prophage and other mobile genetic
378 elements has had a significant impact on Rso2BS4NPB genome evolution. Despite being isolated 45
379 years apart from different hosts and regions, these genomes are highly conserved across the first (Fig. 5)
380 and second chromosome (Fig. S12). The most striking difference between the two lineages are multiple
381 large insertions in Rso2BS4NPB, most of which are predicted to be prophage (Fig. 5). Additionally,
382 according to RECOPHY, the region with the highest SNP density in the Rso2BS4 core genome alignment
383 (1,060 out of 2,894 SNPs) is located in the region surrounding two prophages and an adjacent *fhaC-fhaB*
384 locus (Rso2BS4NPB RUN1145 chr1: 227,422-304,634). This region also contains a high density of
385 Martinique-associated kmers in RUN1145 and *Musa*-associated kmers in UW163. Cumulatively, the data

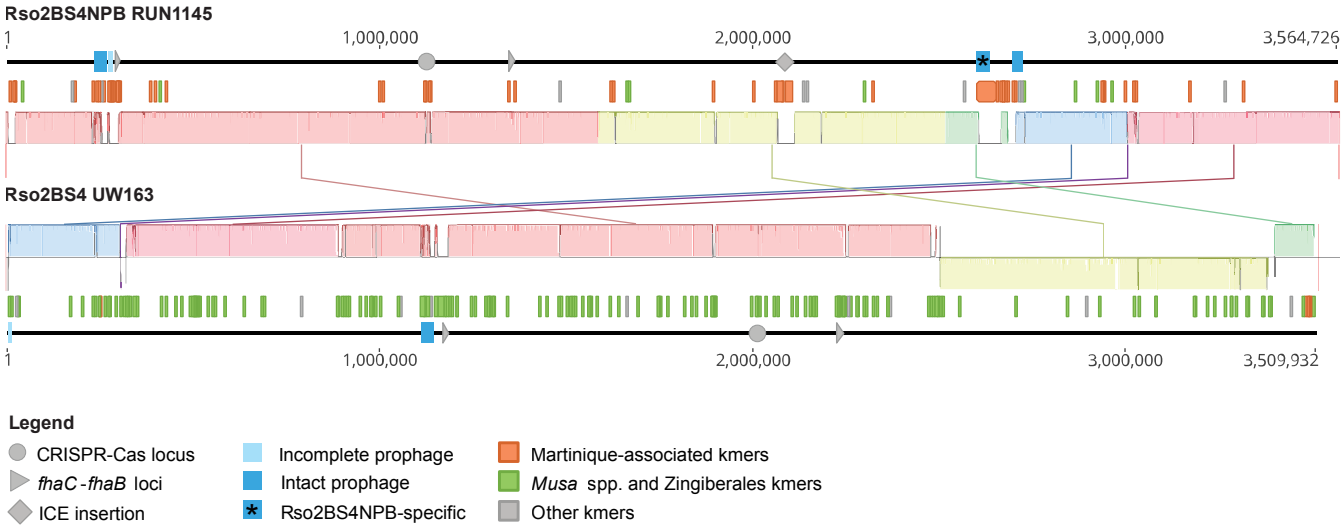


Fig. 5. Mobile elements and host-associated kmers in Rso2BS4 and Rso2BS4NPB. ProgressiveMauve alignment of the first chromosomes of Rso2BS4NPB RUN1145 and Rso2BS4 UW163 strains shown in grayscale together with mapping positions of kmers reported by GWAS and annotated chromosome length scales. The alignment starts from the Rso2BS4NPB RUN1145 origin of replication, while the Rso2BS4 UW163 origin of replication is located at approximately 882,000bp. The chromosome length scales are annotated with positions of prophages (blue rectangles), CRISPR-Cas loci (grey circles), *fhaC-fhaB* loci (grey arrows) and the ICE insertion in Rso2BS4 RUN1145 (grey diamond). The prophage uniquely found in Rso2B4NPB genomes is marked with an asterisk. The kmers that overlap or are located within 1kb of each other were merged. The kmer mapping positions (merged or not) are represented by coloured bars between chromosome scales and ProgressiveMauve locally colinear blocks. The kmers associated with isolation from Martinique are shown in orange, those associated with *Musa* spp. or Zingiberales isolation are shown in green. The kmers associated with any other trait are shown in grey.

396 suggests there were multiple phage infections in the ancestor of the emergent Rso2BS4NPB lineage,
397 either prior or soon after its dissemination to Martinique, introducing multiple novel genes that may be
398 linked with apparent host range expansion.

399 **CRISPR array reveals a history of viral encounter in the emergent Rso2BS4NPB lineage**

400 Evidence of repeated mobile element acquisition in the host-range expanded lineage Rso2BS4NPB
401 attests to the high frequency of encounter and exchange with members of the soil, freshwater and plant
402 microbiomes. Not all encounters are beneficial: members of the RSSC carry type I-E CRISPR-Cas systems
403 conferring immunity to bacteriophage via acquisition of spacer sequences from prior infections. We
404 found that Rso2BS4 UW163 and Rso2BS4NPB RUN1145 share conserved CRISPR-Cas operons (*cas3*,
405 *casA/cas8e*, *casB/cas11/cse2*, *cas7e*, *cas5e*, *cas6e*, *cas1e*, *cas2*), but vary in their CRISPR arrays. The
406 Rso2BS4NPB CRISPR 1 and CRISPR 2 arrays are nearly ten times longer than the Rso2BS4 arrays:
407 Rso2BS4 UW163 has 20 spacers compared to the 171 spacers in Rso2BS4NPB RUN1145 (of which 11 are
408 duplicates with 1-5bp differences). These genomes share one leader-proximal and 15 leader-distal
409 spacers. Our kmer-based association tests also show that the largest cluster of Martinique-associated
410 kmers that do not map to MGEs actually map to the leader-proximal position of CRISPR arrays in
411 Rso2BS4NPB, corresponding to more recent spacer acquisition than the leader-distal spacers shared
412 with Rso2BS4 (Fig. 5, Fig. S13) [30], [31]. This represents a rapid expansion of CRISPR-mediated
413 immunity in the emergent lineage, suggesting bacteria-virus interactions played an important role
414 during the expansion of Rso2BS4NPB across Martinique.
415
416 The identity of the viruses challenging Rso2BS4NPB was inferred by using spacer sequences as queries
417 for INPHARED phage database searches. A total of nine phage were identified using this approach (Table
418 S8). Three phage are targeted by spacers shared between Rso2BS4 and Rso2BS4NPB CRISPR systems,
419 and six are exclusively targeted by Rso2BS4NPB. Interestingly, five of the six phage targeted by
420 Rso2BS4NPB are known to infect Rps1 strains (RPZH3, RS603, RSMSuper, Eline, Raharianne), revealing
421 Rso2BS4NPB are encountering Rps phage, consistent with their frequent co-occurrence.

422 **Pathogen genes associated with host and location**

423 Finally, we took advantage of the extensive metadata collected by Wicker and colleagues during
424 sampling to run genome-wide tests for associations between genetic variation (either genes or kmers),
425 host and location of isolation. Many kmers (16,361) were linked with specific hosts, most frequently
426 *Musa* spp. (6,142 kmers), *Heliconia* spp. (4,560 kmers), *Capsicum annuum* (2,554 kmers) and members
427 of the Alismatales (*Anthurium* spp., *Epipremnum* spp., and *Pothos* spp., 2,369 cumulative kmers) (Table
428 S4). Mapping kmers against Rso2BS4NPB RUN1145 showed 25 *Musa*-associated kmers map to seven loci
429 including a region 5bp upstream of a gene involved in the degradation of aromatic compounds into
430 succinate and malate via the homoprotocatechuate pathway (5-carboxymethyl-2-hydroxymuconate
431 Delta-isomerase, OFBKGH_04945 in RUN1145) and a glycine hydroxymethyltransferase (Table S10. A
432 large fraction of *Musa*-associated kmers map to insertion sequences, two prophages, multiple intergenic
433 regions, hypothetical protein-encoding genes and one of three *cheW* genes encoding chemotaxis signal
434 receptor proteins. Interestingly, the histidine kinase protein CheA, which is known to directly interact

435 with CheW, is one of the most diverse proteins across both Rso2BS4NPB and Rps1S18 according to our
436 pangenome analysis and inspection of gene alignments. 52.6% (2,316/4,401) of kmers mapping to the
437 Rso2BS4 UW163 reference genome are associated with isolation from banana. In comparison, only two
438 banana-associated kmers map to the RpsS18 GMI1000 genome.
439
440 Gene-based panGWAS methods (SCOARY and Pyseer) were also applied to search for genes whose
441 presence or absence is associated with a particular host of isolation. Pyseer reported a total of 938
442 genes significantly associated with any region or host of isolation traits (Table S11). Among these,
443 multiple genes associated with isolation from *Musa* spp. (135), *Solanum melongena* (80), *Capsicum*
444 *annuum* (56) and *Heliconia* spp. (21) were identified. There was no overlap between the Pyseer and
445 SCOARY predictions for host of isolation, though there were for association with specific locations, at
446 varying spatial scales (Table S12). As mentioned, the majority of location-specific genes identified by
447 both methods are located on the $ICE^{RsoGMI1000}$. Outside mobile elements, the most common genes with
448 location-specific associations are *fhaC* and *fhaB* (Fig. 5). These encode type 5b two partner secretion
449 systems comprised of a transporter (FhaC) and extracellular hemagglutinin repeat-containing protein
450 (FhaB). Interestingly, kmers associated with isolation from two distinct regions can map to different
451 regions of the same locus. For example, four out of seven *fhaC-fhaB* loci in Rps1S18 GMI1000 have
452 China-associated kmers mapping to the 5' end of *fhaB* and Martinique-associated kmers mapping to the
453 3' end of the same gene. The 3' end of another *fhaB* gene is composed of alternating regions containing
454 either China or Martinique-associated kmers mapping. Fine-scale spatial associations with *fhaC-fhaB* are
455 also present, for example between an *fhaB* and the region of Le Lorrain in Martinique. Overlap with
456 predictions for host association occurs as well: kmers associated with isolation from both *Musa spp.* and
457 Peru map to four of seven *fhaC-fhaB* loci from Rso2BS4 UW163.

458 **Discussion**

459 This work reveals the evolutionary history of a globally distributed bacterial pathogen and highlights the
460 importance of recombination for disease emergence and host range expansion. We found the disease
461 outbreak affecting a broad range of plant hosts was caused by the dissemination and clonal expansion of
462 two distantly related lineages: *R. pseudosolanacearum* Rps1S18 and *R. solanacearum* Rso2BS4NPB. The
463 unique collection of *Ralstonia* spp. sampled across Martinique and French Guiana provides insight into
464 global and local patterns of pathogen dissemination, the role of recombination in pathogen emergence
465 and the genomic changes associated with host range expansion in Rso2BS4NPB. The data shows
466 Rso2BS4NPB did not emerge in Martinique, but is derived from an ancestral population still present on
467 the mainland, from which Rso2BS4NPB and banana-infecting Rso2BS4 lineages separately diverged.
468 Since basal Rso2BS4NPB isolates were predominantly isolated from *Solanaceae* in French Guiana and
469 there are no reports of a similar epidemic affecting a broad range of plant hosts on the mainland, the
470 genomic changes associated with host range expansion in Rso2BS4NPB may have occurred immediately
471 prior or immediately after dissemination to Martinique. We cannot exclude the possibility favourable
472 climatic conditions and other biotic or abiotic variables may have enhanced opportunities for infection
473 to occur, but the parallel expansion of Rps1S18 across Martinique provides additional evidence of a rare

474 host range expansion event, since Rps1S18 retains a clear signature of host specialisation while
475 Rso2BS4NPB displays the broadest host range yet detected in the species complex.
476
477 Despite the link between T3E carriage and host specific interactions in *Ralstonia* spp. and other bacterial
478 pathogens, we found strikingly little evidence that diversification in T3E repertoire is driving host range
479 expansion in Rso2BS4NPB. There is very little variation in effector carriage between the Rso2BS4NPB
480 and its sister lineage Rso2BS4, which infects banana. Only one hypothetical effector (*T3E_hyp3*) was
481 acquired and two (*ripAX2* and *ripP3*) were lost. We do not exclude the possibility that changes in
482 expression and allelic variation among T3Es may be responsible for altered host range. For instance,
483 *ripS3* is present in both Rso2BS4 and Rso2BS4NPB, but displays considerable sequence polymorphism
484 compared to all other effectors present across both lineages. These differences may be associated with
485 the altered pathogenicity of Rso2BS4NPB, but appear insufficient to account for the dramatic expansion
486 in pathogen host range, particularly since single effector deletions typically have limited consequences
487 for infection in the absence of recognition. We argue host range expansion in Rso2B4NPB is not due to
488 changes in effector-mediated disruption of recognition and immunity affecting its interactions with a
489 panoply of distantly related hosts, but is more likely due to the acquisition of novel genetic modules
490 altering bacterial metabolism and heightening competitive fitness.
491
492 Bacterial metabolism and manipulation of nutrient availability in the host is increasingly recognised as
493 an important component of pathogenicity, in mammalian and plant pathogens alike [32], [33], [34], [35].
494 For instance, genes whose products facilitate the use of xylem sap as a carbon source [36] and
495 horizontally acquired components of the denitrification pathway [37] have been shown to be important
496 for xylem colonisation and virulence in *Ralstonia*. Transcriptome sequencing of experimentally evolved
497 Rps1S18 GMI1000 showed fatty acid metabolism is upregulated in evolved clones of Rps1S18 GMI1000,
498 highlighting the importance of metabolism during adaptation to a novel host [38]. Metabolism is likely
499 carefully regulated according to infection stage, as was shown for Rso2BS1 UW551, which
500 downregulates carbohydrate metabolism and tricarboxylic acid-related pathways during late stages of
501 xylem infection [39]. We postulate the acquisition of novel pathways for substrate catabolism that
502 enhance bacterial growth either in the soil or within plant tissues, resulting in a sudden and explosive
503 increase in symptom production in the field.
504
505 The acquisition of an ICE with novel metabolic enzymes may have been pivotal to the host range
506 expansion of Rso2BS4NPB. ICEs facilitate rapid horizontal gene transfer of entire functional genetic
507 modules, allowing rapid adaptation to novel hosts and environmental niches [40]. All Rso2BS4NPB
508 strains in Martinique carry an ICE with a cargo-1 gene cluster encoding multiple enzymes and regulators.
509 Many cargo-1 genes are linked with the degradation and detoxification of environmental pollutants and
510 reactive xenobiotics. The functional significance of these acquisitions for Rso2BS4NPB virulence is
511 unclear, although there are striking links with the degradation of organic compounds present in the soil.
512 The phthalate ester hydrolase (*pnbA*) is linked with the degradation of phthalate esters, common
513 synthetic organic compounds that are known to be environmental pollutants [41]. Homologs of *gstA* and
514 *fabG* are known to degrade organochlorine compounds used as primary active ingredients (1-chloro-2,4-
515 dinitrobenzene) or intermediates (ethyl 4-chloroacetoacetate) in the production of herbicides and

516 pesticides [42], [43]. Strikingly, multiple cargo-1 genes have products linked with the manipulation of
517 host-derived metabolites. The limonene 1,2-epoxide hydrolases *limA* may act on host plant epoxides,
518 which are induced during both pathogen attack and methyl jasmonate application [44]. A homolog of
519 *ygiN* (ABM) is known to hydroxylate jasmonic acid in rice, attenuating plant defences during infection
520 [45], and pyrazinamidase/nicotinamidase (*pncA*) is known to convert nicotinamide to nicotinic acid. The
521 accumulation of nicotinic acid during grape xylem colonisation by *Xylella fastidiosa* suggests *pncA* may
522 be linked with vascular pathogenesis [46]. *Ralstonia* spp. degradation of plant metabolites has been
523 observed in Rso2BS1 UW551, which degrades phenolic compounds in xylem sap [47], and many
524 *Ralstonia* spp. genomes contain salicylic acid degradation pathways [48].
525
526 Recombination events introducing metabolic enzymes occurred beyond the ICE as well, in the ancestor
527 of Rso2BS4 and Rso2BS4NPB. These are similarly linked with the degradation of either plant-derived
528 compounds or protection against environmental toxicity or oxidative stress. The acquisition of a
529 TauD/TfdA family dioxygenase linked with synthetic auxin degradation may confer a fitness advantage in
530 niches frequently exposed to phenoxy herbicides. The ability to metabolise toxic or antimicrobial
531 compounds is not unusual in *Ralstonia*: *R. pickettii* degrades DDT pesticides [49] and *R. eutropha*
532 degrades chloroaromatic compounds [50]. The ancestral acquisition of a carboxymuconolactone
533 decarboxylase is also of considerable interest, since the protocatechuate catabolism pathway is
534 responsible for the conversion of lignin-derived aromatic compounds into amino acids and organic acids,
535 potentially fueling *Ralstonia* growth in plant tissues. These acquisitions, along with additional genomic
536 changes associated with Rso2BS4NPB adaptation to multi-host environments, are strong candidates for
537 broad host range determinants in this lineage.
538
539 Pathogen co-occurrence provides opportunities for mobile element exchange, as we have observed for
540 the ICE^{RpsGMI1000} movement between Rso2BS4NPB and Rps1S18 in Martinique. The Rso2BS4NPB
541 ICE^{RsoGMI1000} shares a similar backbone and identical cargo-1 cluster with the Rps1S18 ICE^{RsoGMI1000},
542 differing only in the carriage of distinct cargo-2 clusters [51]. Examination of the spatial patterns of
543 pathogen diversity and mobile element distribution show the ICE is present in only three Rps1S18 and all
544 Rso2B4NPB in Martinique (Fig. 4), as well as some basal members of the Rso2B4NPB lineage in French
545 Guiana. With the exception of a single Martinique Rso2A isolate also carrying the ICE, the other *R.*
546 *solanacearum* sequenced to date are otherwise curiously ICE-free, despite chromosomal homology at
547 preferred ICE integration sites and an abundance of direct repeats targeted by the ICE^{RsoGMI1000} integrase.
548 In contrast, we identified multiple additional ICEs circulating among *R. pseudosolanacearum*. The
549 apparent species-specific variation in ICE carriage is striking, we speculate ICE acquisition by Rso2B4NPB
550 may have occurred prior to or soon after Rso2BS4NPB dissemination to Martinique, perhaps via
551 coinfection with Rps1S18 or another *R. pseudosolanacearum*. It is not possible to reliably infer the
552 source and directionality of transfer events for mobile elements like ICEs, but their distribution among
553 *Ralstonia* spp. sampled across Martinique nevertheless suggests ICE^{RsoGMI1000} acquisition is linked with
554 Rso2BS4NPB emergence and host range evolution and provides insight into the dynamic evolution of
555 the mobile elements themselves. We are not aware of any prior observation that ICE cargo integration
556 sites can be so clearly partitioned by function, between either metabolism-associated or defence system
557 gene clusters.

558
559 Rso2BS4NPB is both a parasite of plants and a host in turn, defending itself against parasitism by
560 bacteriophage. The emergent lineage exhibits evidence of frequent encounter with potential parasites.
561 The expansion of CRISPR repeats and acquisition of multiple fragmented and intact prophages in
562 Rso2BS4NPB genomes is clear evidence of frequent exposure to phage; acquired immunity may provide
563 an additional advantage to Rso2B4NPB in the phage-rich soil environment [52]. Since some of the
564 Rso2BS4NPB CRISPR spacers with identifiable origins correspond to phage known to successfully infect
565 *R. pseudosolanacearum* strains like Rps1S18, coinfections are likely sites of exposure to lytic phage,
566 along with the acquisition of ICEs and prophage carrying accessory genes altering microbial lifestyle and
567 pathogenicity [53].
568
569 The emergence of a pandemic lineage infecting crop plants is frequently linked with the introduction
570 and expanded cultivation of susceptible host varieties [6], [54], [55], [56]. The center of origin of *R.*
571 *solanacearum* is in the Americas, where banana was introduced as a crop five centuries ago and became
572 intensively cultivated only in the last two centuries [57]. Infectious disease outbreaks occurred soon
573 after, most notoriously those caused by *Fusarium oxysporum* TR1 and *F. odoratissimum* TR4, which
574 previously eliminated the Gros Michel variety and now threaten the Cavendish variety [58], [59]. Banana
575 pathogenicity evolved independently multiple times in *R. solanacearum*. The repeated, independent
576 evolution of banana pathogenicity across the *Ralstonia* species complex aligns with the suggestion
577 pathogenicity on this host is a derived trait in Rso2BS4 [2], [3], [60]. Convergent trait evolution and a
578 high frequency of recombination breaking down linkage disequilibrium can contribute to the
579 identification of genes and loci linked with particular phenotypes. The banana-associated genes and
580 kmers identified here represent interesting candidates for study of the genetic basis of specialisation on
581 this specific host. Although further work is required to determine whether the candidates identified in
582 this work are responsible for banana pathogenicity in Rso and the increased fitness of Rso2BS4NPB in
583 soil and host environments, this work shows bacterial metabolism is likely to play an important role in
584 host adaptation, and reveals how ecological interactions with other members of the microbial
585 community shape pathogen evolution and contribute to the emergence of infectious disease.

586 ## Methods

587 ### Genome sequencing

588 *Ralstonia* spp. were collected in Martinique and French Guiana as described in Wicker *et al*. 2009 [3] and
589 Deberdt *et al*. 2014 [4]. Isolates collected after the initial sampling were assigned to phylotype and
590 sequevar by sequencing endoglucanase (*egl*) and mutation mismatch repair (*mutS*) genes. For this study,
591 we extracted genomic DNA from 487 strains using Promega Wizard, then used a modified Nextera
592 protocol for library preparation, as described in [61]. Paired-end sequencing with Illumina NextSeq 2000
593 was then performed. Since the first round of sequencing produced under 25-fold average coverage and
594 under 40-fold maximum coverage, a second round of library preparation and sequencing was
595 performed. Eight isolates (RUN0777, RUN0808, RUN0820, RUN0823, RUN1145, RUN1155, RUN1952
596 and RUN2004) were additionally sequenced using Oxford Nanopore Technology with a FLO-MIN106 flow
597 cell and SQK-RBK004 barcoding kit. Basecalling and demultiplexing was performed with
598 guppy_basecaller (v. 5.0.7) and guppy_barcoder v. 4.5.4 [62].

599 ### Genome assembly and annotation

600 Paired-end 150bp reads of both Illumina sequencing runs were filtered and trimmed with fastq_screen
601 [63], bbduk [64] and fastp [65], [66] (--dedup --detect_adapter_for_pe --cut_front 15 --length_required
602 50). Assembly was performed with SPAdes [67], [68] using both separate and pooled (i.e. both
603 sequencing runs) read sets. QUAST was used for assembly quality assessment [69]. Pooling both
604 sequencing runs improved average assembly N50 values from 73Kb to 195Kb and decreased average
605 contigs per assembly from 370 to 217, so assemblies generated from pooled sequencing runs were
606 used. After filtering out assemblies with under 50-fold coverage, over 250 contigs, total assembly length
607 exceeding 6Mb and N50 exceeding 75Kb, a total of 409 assemblies were retained for downstream
608 analyses.
609
610 Oxford Nanopore long read data was filtered with Filtlong (v. 0.2.1) [70] to remove reads under 1Kb and
611 the 5% of reads with the lowest quality values, then assembled with Trycycler (v. 0.5.4) [71], using Flye
612 [72], Miniasm [73] with Minipolish [74] and Raven [75] as underlying assemblers. Contigs that did not
613 pass the reconciling step of Trycycler pipeline were manually removed. Final Trycycler consensus
614 sequences were polished three times using long reads with Medaka [76], followed by two rounds of
615 short read polishing using BWA [77], [78] and Polypolish [79], using separate batches of the Illumina
616 reads produced by each sequencing run.
617
618 All assemblies were annotated by Bakta [80], [81] using Prodigal [82] gene models and a reference
619 protein database built using all complete *Ralstonia* spp. genome assemblies available from the RefSeq
620 database in October 2022.

621 ### NCBI datasets and reference genomes

622 All available *R. pseudosolanacearum, R. syzygii* and *R. solanacearum* genome assemblies were
623 downloaded from NCBI in February 2023 to provide additional phylogenetic context for analyses.

624 Assemblies with fewer than 200 contigs, total assembly length between 5.0 - 6.2Mb, N50 over 75Kb and
625 unflagged by the NCBI Genome datahub [83] for evidence of potential contamination were retained. A
626 total of 410 Ralstonia spp. genomes passed these filters: 370 *R. solanacearum* (Rso), 35 *R.*
627 *pseudosolanacearum* (Rps) and 5 *R. syzygii* (Rsy) genomes (Table S1). Three isolates were present in
628 duplicate in the NCBI filtered dataset (MAFF 211479, UW163 and UW551), the lowest quality duplicate
629 assembly was removed, producing a final NCBI dataset of 407 assemblies. These were then reannotated
630 with Bakta and merged with the 409 assemblies produced in this work for downstream pangenome
631 analyses. Metadata (host, geographical region and year of isolation) for genomes in the NCBI dataset
632 was obtained by inspecting Biosample entries. If the information was missing in Biosample entries, we
633 used RSSCdb records [9], [84] or publications describing the initial isolation of the strain of interest [85].

634 **Type 3 secreted effector identification**

635 Amino acid sequences of 118 *R. solanacearum* type 3 secreted effectors (T3Es) from the manually
636 curated RalstoT3E database [86] were downloaded and used as query sequences for DIAMOND BLASTp
637 [87] searches against proteomes of the Bakta-annotated assemblies, using the --ultrasensitive mode.
638 Hits were filtered using a S30L25 strategy: retaining sequences with alignment identity of at least 30%
639 and query sequence coverage of 25% or greater. Since the T3E query set includes multiple alleles of the
640 same effector families, hits were then grouped into T3E families by sorting first by identity and query
641 length coverage. T3E family assignment for each protein sequence was performed using the most
642 significant T3E hit. Effector presence and absence was visualised by a heatmap in R, omitting duplication
643 events. This approach was tested by comparing the predicted T3E repertoires in Rps1S18 GMI1000 and
644 Rsy RPSI07 with the ones provided in the RalstoT3E database. Average pairwise similarity of T3E
645 repertoires within RSSC clades was estimated by Scikit-learn function cosine_similarity [88] using the
646 information about presence-absence of effector families as an input binary matrix.

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

648 Panaroo [89] was used to identify core and flexible genomes for the RSSC dataset, i.e. the merged
649 dataset of 816 *Ralstonia* genomes sequenced in this study (409 strains) and downloaded from NCBI (407
650 strains). A concatenation of core genes was used as an input alignment for phylogenetic analysis of the
651 species complex. Alignment columns with gaps or ambiguities and invariant sites were removed with
652 SNPsites [90], producing a final alignment of 137,423bp. Tree inference was then performed using
653 IQtree [91] with a GTR+ASC model of sequence evolution, model for alignments that do not contain
654 constant sites, and 1000 bootstrap replicates. The ggtree R package [92] was used for all tree and
655 metadata visualisation.
656
657 Panaroo gene presence and absence tables for the RSSC dataset were queried using a custom Python
658 script to identify genes uniquely present in either Rps1S18, RsoS4 or Rso2BS4NPB. Genes are referred to
659 as clade-specific when they are present in at least 95% of all genomes in a given focal clade, and present
660 in no more than 10 non-focal clade assemblies.
661

662 The species complex tree from the set of core genes identified by Panaroo was used for clade
663 delineation and lineages assignment. Most genomes clustered according to the phylotypes initially
664 assigned by *egl* and *mutS* sequencing in Wicker *et al.* 2007 [3] and Deberdt *et al.* 2014 [4], although
665 some discrepancies are present. We simplify strain species, phylotype and sequevar references here
666 using the following format: [species][phylotype][sequevar]-[strain ID]. For instance, *R. solanacearum*
667 phylotype IIB sequevar 4 strains are referred to as Rso2BS4, where the 'S' precedes prior sequevar
668 assignment. Those strains that were not classified by sequevar have a shorter designation:
669 [species][phylotype]-[strain ID]. The clade designation in this work may differ from historically assigned
670 sequevar or even phylotype labels, since these were first assigned using *egl* and *mutS* sequences rather
671 than whole genome sequencing and core gene tree reconstruction. The criteria for membership in a
672 named clade is exceeding a minimum shared identity of 99% in the RSSC core gene alignment.
673
674 Separate pangenome analyses and core gene tree reconstructions were performed on three specific
675 clades: *R. pseudosolanacearum* phylotype 1 sequevar 18 (Rps1S18, 102 strains), *R. solanacearum*
676 phylotype 2 sequevar 4 (Rso2BS4, 226 strains), and the emergent lineage *R. solanacearum* phylotype 2
677 sequevar 4 Non-Pathogenic on Banana (Rso2BS4NPB, 180 strains). The strains forming each clade were
678 identified with reference to the RSSC core gene tree (Figure 1). Assemblies for IBSBF_2571
679 (GCA_003590605.1) and O12BS_UW407 (GCA_023076615.1) were excluded due to abnormally long
680 terminal branch lengths and P673 was excluded due to contradictory phylogenetic position of two
681 assembly versions: the most recent P673 assembly GCA_015698365.1 was assigned to Rso2BS4NPB
682 clade, while the earlier version GCA_000525615.1 grouped with other Rso2BS4 *Epipremnum* isolates
683 from the USA.
684
685 An additional iteration of clade-specific pangenome analysis was performed for both Rso2BS4NPB and
686 Rps1S18. Although strain IBSBF1503 (GCA_001587155.1) groups with other Rso2BS4NPB strains in the
687 RSSC core gene tree, the clade-specific tree revealed this strain is a more distant relative and more
688 properly referred to as an outgroup strain, therefore this was excluded from a second iteration.
689 Similarly, Rps1S18 strains CIP266_UW505 (GCA_023075875.1) and UW393 (GCA_023076675.1) form a
690 distant outgroup to other Rps1S18 strains, therefore these were excluded from the second iteration of
691 the Rps1S18 pangenome analysis.

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

693 All potential mobile elements, including genes annotated as integrases, transposases, prophage-related
694 genes and IS elements, were identified and masked for the downstream phylogenetic analysis in the
695 Bakta-annotated genome assemblies for Rps1S18 strain GMI1000 (GCA_000009125.1), RsoS4 strain
696 UW163 (GCA_001587135.1) and Rso2BS4NPB strain RUN1145. Prophage and IS elements were
697 predicted using the PHASTER, PHASTEST [93], [94], [95] and ISfinder [96] online tools. All gene
698 annotations containing keywords "phage" OR "transposase" OR "tail" OR "head" OR "integrase" OR
699 "terminase" OR "conjugal" OR "integrase" OR "integrative conjugative element" OR "conjugative
700 transfer" OR "conjugative coupling" were manually searched for and flagged in Geneious Prime [97]. The

701 results were converted into BED files, then sorted, merged and used for masking with Bedtools [98].
702 Masked references were used for reference-based contig mapping.

703 **Phylogenetic analyses**

704 Reference-based contig mapping was performed for the Rps1S18, RsoS4 and Rso2BS4NPB clades, as
705 defined by using RSSC and clade-specific core gene trees. Rps1S18 included 100 genomes, with GMI1000
706 as reference. The RsoS4 clade contained 226 genomes, with UW163 as reference. Rso2BS4NPB included
707 179 genomes, with RUN1145 as reference. The Brazilian *Cucumis sativus* isolate IBSBF1503 is referred to
708 as Rso2B4NPB in the literature and initially grouped with Rso2BS4NPB in the species complex tree, and,
709 so we initially used this strain as a reference for Rso2BS4NPB. While more closely related to Rso2B4NPB
710 than any other available complete *Rso* genome, we found IBSBF1503 is clearly an outgroup relative to
711 other Rso2BS4NPB isolates (Fig. S3). The average pairwise nucleotide diversity is 0.002% between
712 Rso2BS4NPB genomes, and an order of magnitude greater (0.02%) relative to IBSBF1503. According to
713 progressiveMauve whole genome alignments, Rso2BS4NPB isolate RUN1145 differs from RUN1960 and
714 IBSBF1503 by 134 and 1,215 SNPs, respectively. Since IBSBF1503 is too divergent to be considered a
715 member of the same lineage, we reclassified IBSBF1503 as Rso2BS4. We resequenced five Rso2BS4NPB
716 strains from Martinique using Nanopore long-read sequencing to generate closed hybrid assemblies and
717 used the complete assembly of Rso2BS4NPB RUN1145 as a reference for more reliable core genome
718 alignment-based phylogenetic inference of relationships between closely related strains. The Rps1S18
719 contig mapping using Rps1S18 GMI1000 as reference resulted in 4,760,272 bp consensus alignment with
720 311 SNPs and 116 parsimony informative sites.
721
722 Reads and contigs mappings were performed separately for each assembly with bwa-mem2 [77], [78] or
723 Minimap2 [99] using asm5 mode respectively. All resulting BAM alignments were sorted and indexed
724 with Samtools [100]. BCFtools mpileup was used to filter and retain bases with mapping quality above
725 60 [101]. Variants were called with bcftools-call and filtered by bcftools-filter using the expression
726 TYPE="snp" && DP>=X && QUAL>30 && AF1>=0.95 with X=10 for reads and X=1 for contig mapping.
727 Any alignment column with a gap site was removed. Consensus sequences were obtained using bcftools-
728 consensus, masking the sites with coverage depth below one as identified by Mosdepth [102]. The
729 consensus sequences for all strains in a given clade were concatenated into a FASTA alignment file. Sites
730 containing N characters were trimmed with goalign clean-sites [103]. Recombinant regions were
731 predicted and removed by ClonalFrameML [104] using an initial tree calculated by IQtree [91] under
732 GTR+F+I+R2 substitution model. The final Maximum likelihood trees were calculated from the non-
733 recombinant alignments for each clade by IQtree using the MFP (ModelFinder Plus) option for the best
734 model search [105]. The contig mapping approach resulted in better resolved and bootstrap supported
735 trees, therefore the trees generated by reads mapping were excluded from further analysis. Tree
736 visualisations were created with ggtree [92].

737 **Tree dating**

738 Root-to-tip regression and dating of the Rso2BS4NPB and Rps1S18 non-recombinant core genome trees
739 were performed with BactDating using 100,000 Markov Chain Monte Carlo iterations [106]. For the

740 Rso2BS4NPB tree, RsoS4 strain IBSBF1503 was added as an outgroup for tree calculation and dating, as
741 the initRoot function was otherwise unable to root the tree. TempEst [107] was used to confirm root-to-
742 tip regression results obtained with BactDating, setting the 'Best-fitting root' option of TempEst to 'R
743 squared'. Tree tips with missing sampling date information (Rps1S18 UQRS637_UW745 and UW193,
744 Rso2BS4NPB RUN0283 and RUN0739) and outliers (Rps1S18 UW299, RUN0679, RUN0037 and GMI1000)
745 identified by TempEst were dropped from the Rps1S18 and Rso2BS4NPB trees using the R function
746 ape::drop.tip. Branches with edge lengths of zero were manually collapsed into a single node by
747 Dendroscope [108]. A Dates Permutation Test was performed with BactDating by randomly generating
748 dates between 2000 and 2012 and randomly assigning these to tree tips. The simulated trees were
749 compared to the actual dating model using the modelcompare function of BactDating.

750 **CRISPR-Cas system identification**

751 CRISPR-Cas systems were identified with Bakta annotation of Rso2BS4 and Rso2BS4NPB genomes.
752 CRISPR repeats and spacers were then annotated in two closed references Rso2BS4-UW163 and
753 Rso2BS4NPB-RUN1145 using the CRISPRloci workflow [109]. The spacer sequences identified with
754 CRISPRloci were then used for BLASTn searches against the RefSeq Phage database downloaded from
755 PhageScope [110], NCBI Nucleotide collection (nr_nt) and Environmental samples (env_nt) databases
756 accessed in June 2024. The extracted spacers were also used for INPHARED database [111] searches
757 using the SpacePHARER [112] toolkit in August 2024.

758 **Inter and intraspecific recombination prediction**

759 Potential homologous recombination events within and between Rso2BS4 and Rps1S18 lineages were
760 investigated using the RECOPHY toolkit [113]. Two datasets were analysed: Rso2BS4 alone and in
761 combination with Rps1S18, using Rso2BS4NPB-RUN1145 as a reference in each case. Prior to RECOPHY
762 execution, all genome pairs with a Mash distance [114] below zero were identified and one member of
763 each pair was then randomly removed from the dataset. The final datasets included 49 genomes for the
764 Rso2BS4 analysis alone, and with an additional 64 Rps1S18 genomes for the merged analysis. The
765 regions of core alignment with increased and reduced density of SNPs per 1 kb block were extracted and
766 investigated separately. In order to find discrepancies with the core phylogeny, a simple Neighbor-
767 Joining tree was calculated for each of these regions using Geneious Prime software [97].

768 **Pangenome-wide association testing**

769 Scoary [115] and Pyseer [116] pan-GWAS analyses were used to investigate associations between gene
770 presence and absence patterns and the host or location of strain isolation. The broad (country) and fine-
771 scale (village) locations and all taxonomic levels of the host of isolation or environmental substrate were
772 converted from the metadata table into a binary traits table using the pandas.get_dummies function
773 [117]. For example, a strain isolated from a tomato plant in China would receive the value "1" in the
774 columns China, *Eudicots, Asterids, Solanales, Solanaceae, Solanum* and *S. lycopersicum*, and receive the
775 value "0" in all other columns of the traits table. This traits table was used as input for both Scoary and
776 Pyseer pan-GWAS analyses, along with the gene presence and absence table from Panaroo.
777

The significance of candidates identified with Scoary was assessed after multiple test corrections: individual (naive) p-value, Bonferroni adjusted p-value, Benjamini-Hochberg adjusted p-value, entire range of pairwise comparison p-values, and empirical p-value from 100 permutations, using thresholds of p=0.05 for all corrections. Significant genes identified by Scoary after all the corrections were further filtered using 95% sensitivity and 70% specificity thresholds. A lower specificity threshold was chosen due to the nature of the trait dataset: the traits table simply reflects the known host of isolation and the location of isolation. The actual host range of a particular strain may include additional hosts, closely or distantly related, just as closely related strains may be present in vastly different geographic regions. Since the Scoary and Pyseer pan-GWAS analyses aim to identify genes linked with survival in a particular host, environment or substrate, the sensitivity should be high. Those same genes may also be present in strains isolated from other plants however, therefore specificity is reduced. For example: a gene whose product is strongly associated with tomato (*S. lycopersicum*) infection should be overrepresented among tomato isolates in the dataset. Therefore we seek genes present in 95% of tomato isolates. We reduced the specificity threshold to 70% to allow for instances where another strain carries that same gene and is also capable of infecting tomato, but happens to have been isolated from a different host.

For the Pyseer analysis, the distance matrix from the RSSC core gene tree and the number of patterns for the likelihood ratio test (LRT) were calculated using phylogeny_distance.py and count_patterns.py scripts downloaded from the Pyseer GitHub page [118], accessed on January 2024. The pattern count resulted in 6,075 patterns for the RSSC, leading to the likelihood ratio test (LRT) p-value threshold of 3.8e-08 = 0.05/(patterns_number*traits_number). Hits with effect size (beta value) under 0.2 and phenotypic variance values (variant_h2) above 0.5 were filtered out.

Pan-GWAS of gene presence/absence data is an effective tool for the identification of genes associated with traits of interest, but is limited to the accessory genome and does not test for trait associations within the core genome and intergenic regions. This gap is filled with a kmer-based association test. Analogously to the method used for gene presence/absence data, Pyseer was applied to identify the kmers significantly associated with any host or location of isolation across the RSSC, using the same binary metadata table as described for the pan-GWAS analysis. The initial set of 31bp long kmers was generated for all the RSSC genomes using unitig-caller [119]. The distance matrix from the species complex core genes tree and the number of patterns for the likelihood ratio test (LRT) were calculated using phylogeny_distance.py and count_patterns.py scripts from the Pyseer GitHub page [118]. The pattern count reported 1,484,018 patterns for the RSSC kmers, producing a LRT p-value threshold of 5.0e-10 = 0.05/(patterns_number*traits_number). Hits with effect size (beta value) under 0.2 and phenotypic variance values (variant_h2) above 0.5 were filtered out. The presence and positions of significant kmers in Rso2BS4NPB reference strain RUN1145, RsoS4 reference strain UW163 and Rps1S18 reference strain GMI1000 was identified and stored in BED files using a custom Python script. For visualisation purposes, all kmers mapped on a genome within a 1kb distance were merged using the bedtools-merge function [98].

The panGWAS and kmer-based tests identify many genes associated with isolation from specific hosts. These tests vary in how they control for population structure, but some genes were predicted to have

820 associations with particular traits using multiple methods. Separate association tests were performed
821 using either spatial location or host/substrate of isolation metadata. We noticed that many genes have
822 overlapping predictions using both metadata types, particularly where it concerns isolates recovered
823 from soil or water. For example, gene presence associated with *S. lycopersicum* isolation often overlaps
824 with gene presence associated with *Solanaceae* isolation. All overlaps were investigated using a custom
825 Python script and visualised in an Upset plot using the UpSetR R package [120]. For the genes assigned
826 to multiple nested traits, the lowest level of taxonomic identification (e.g. species rather than genus) or
827 the finest level of geographical scale (e.g. village rather than country) for which significance was
828 identified were chosen. Overlaps within the Scoary, Pyseer pan-GWAS and kmer-based outputs were
829 identified using the same strategy.

830 **Identification of integrative and conjugative elements**

831 Complete integrative and conjugative elements were identified in 133 complete RSSC genome
832 assemblies using the following criteria: the presence of a type 4 secretion system (T4SS) required for
833 conjugal transfer between strains, a *mob* family relaxase and integrase, and the presence of 8bp direct
834 repeats (TTTTTTAT) encompassing the predicted ICE [121]. This 8bp repeat is described as an ICE
835 insertion site for Rps1S18 GMI1000 in the ICEberg 3.0 database [51]. We note this repeat occurs
836 between 50-70 times elsewhere in each RSSC genome. This does not preclude them from being actual
837 ICE integration sites: despite the apparent existence of preferred ICE integration sites in the RSSC, some
838 ICE integrases are known to be promiscuous and may target different repeats as attachment sites across
839 the genome. For example, ICE Tn916 targets AT-rich loci and integrates almost at random across
840 genomes [122]. MOBscan was used for identification of relaxases [123], followed by BLASTn sequence
841 homology searches for a T4SS and adjacent integrases [124] using Rso2BS4NPB RUN1960 T4SS and
842 integrase sequences as BLASTn queries, followed by T4SS locus confirmation with the TXSScan module
843 of MacSyFinder toolkit [125]. The flanking repeat sequence TTTTTTAT was identified using a custom
844 python script. In genomes where the repeat was not identified within 10kb of the T4SS, relaxase and
845 integrase, putative insertion sites were manually identified.
846
847 The genomic context of putative ICEs in the RSSC assemblies was inferred by aligning 50 kb upstream
848 and downstream flanking regions surrounding putative ICEs in Rps1S18 GMI1000, Rso2BS4NPB
849 RUN1145 and Rso2BS4NPB RUN1960 to the genomes. The presence of genes significantly associated
850 with isolation from Martinique by panGWAS analysis was also confirmed by BLASTn, retaining hits with
851 e-values under 1.0^{-5}, query length coverage over 10% and overall alignment length exceeding 1 kb.
852
853 The involvement of ICE cargo genes in specific metabolic pathways was predicted using KEGG Orthology
854 assignment with protein sequences from the cargo regions of interest, performing searches in the
855 database for the genus Ralstonia (taxa ID: 48736). The locus structure of putative ICEs was visualised
856 with the ggtree::geom_motif R function [92]. Complete ICE sequences identified in *Rps* 362200, *Rps*
857 RS24, *Rps* RUN1952, Rps1S18 RUN820, Rps1S18 GMI1000, Rso2BS4NPB RUN1960, RUN0777, Rso2A
858 RUN1155 and *Ralstonia pickettii* strain 12J were aligned to each other using the progressive Mauve
859 [126] algorithm implemented in Geneious [97].

860 **Geographical metadata display**

861 Information about the geographical region of isolation for each strain was obtained from the country of
862 isolation metadata collected for the RSSC strains set as described above. Counties of isolation were
863 converted into continents using pycountry Python package [127]. The “Caribbean” region of isolation
864 was manually assigned for all Martinique isolates to highlight the region of interest.
865
866 Longitude and latitude of the sampling site was recorded for 83/500 strains sampled in Martinique and
867 French Guiana. For the remaining 417/500 strains, the commune (administrative district) or the nearest
868 village was listed in the strain metadata. For these cases, approximate longitude and latitude values
869 were retrieved using the geolocator function of the geopy.geocoders Python package [128] using the
870 commune or nearest village names. Commune maps of Martinique and French Guiana in geojson format
871 were downloaded from the France Geojson collection [129]. The actual and estimated longitude and
872 latitude values for each isolate were visualised with the geopandas package [130].

873 **Data availability**

874 Sequence data generated for this work is available at NCBI SRA Accession XXX.
875

References

[1] P. Prior and M. Fegan, "Diversity and molecular detection of Ralstonia solanacearum race 2 strains by multiplex PCR," Jan. 2005.

[2] E. Wicker *et al.*, "Ralstonia solanacearum Strains from Martinique (French West Indies) Exhibiting a New Pathogenic Potential," *Appl. Environ. Microbiol.*, vol. 73, no. 21, pp. 6790–6801, Nov. 2007, doi: 10.1128/AEM.00841-07.

[3] E. Wicker, L. Grassart, R. Coranson-Beaudu, D. Mian, and P. Prior, "Epidemiological evidence for the emergence of a new pathogenic variant of Ralstonia solanacearum in Martinique (French West Indies)," *Plant Pathol.*, vol. 58, no. 5, pp. 853–861, 2009, doi: https://doi.org/10.1111/j.1365-3059.2009.02098.x.

[4] P. Deberdt *et al.*, "Diversity of Ralstonia solanacearum in French Guiana Expands Knowledge of the 'Emerging Ecotype,'" *Phytopathology®*, vol. 104, no. 6, pp. 586–596, Nov. 2013, doi: 10.1094/PHYTO-09-13-0264-R.

[5] D. Ugent, "The Potato," *Science*, vol. 170, no. 3963, pp. 1161–1166, Dec. 1970, doi: 10.1126/science.170.3963.1161.

[6] A. Abdurahman *et al.*, "Molecular characterization of Ralstonia solanacearum strains from Ethiopia and tracing potential source of bacterial wilt disease outbreak in seed potatoes," *Plant Pathol.*, vol. 66, no. 5, pp. 826–834, 2017, doi: 10.1111/ppa.12661.

[7] L. Tessema, W. Negash, R. Kakuhenzire, G. Belay, E. Seid, and M. Enyew, "Seed health trade-offs in adopting quality declared seed in potato farming systems," *Crop Sci.*, vol. 64, no. 3, pp. 1340–1348, 2024, doi: 10.1002/csc2.21025.

[8] G. M. R. Albuquerque *et al.*, "Moko Disease-Causing Strains of Ralstonia solanacearum from Brazil Extend Known Diversity in Paraphyletic Phylotype II," *Phytopathology®*, vol. 104, no. 11, pp. 1175–1182, Nov. 2014, doi: 10.1094/PHYTO-12-13-0334-R.

[9] *lowepowerlab/Ralstonia_Global_Diversity*. (May 31, 2022). Lowe-Power Lab (UC Davis). Accessed: Jun. 12, 2024. [Online]. Available: https://github.com/lowepowerlab/Ralstonia_Global_Diversity

[10] N. S. Price, "The origin and development of banana and plantain cultivation," in *Bananas and Plantains*, S. Gowen, Ed., Dordrecht: Springer Netherlands, 1995, pp. 1–13. doi: 10.1007/978-94-011-0737-2_1.

[11] W. J. Kress, "Systematics of Central American Heliconia (heliconiaceae) with Pendent Inflorescences," *J. Arnold Arbor.*, vol. 65, no. 4, pp. 429–532, 1984.

[12] T. B. Croat and R. Box, "Ecology and Life Forms of Araceae," vol. 11, 1988.

[13] C. J. Hueck, "Type III Protein Secretion Systems in Bacterial Pathogens of Animals and Plants," *Microbiol. Mol. Biol. Rev.*, vol. 62, no. 2, pp. 379–433, Jun. 1998, doi: 10.1128/mmbr.62.2.379-433.1998.

[14] M. Poueymiro *et al.*, "Two Type III Secretion System Effectors from Ralstonia solanacearum GMI1000 Determine Host-Range Specificity on Tobacco," *Mol. Plant-Microbe Interactions®*, vol. 22, no. 5, pp. 538–550, May 2009, doi: 10.1094/MPMI-22-5-0538.

[15] D. Landry, M. González-Fuente, L. Deslandes, and N. Peeters, "The large, diverse, and robust arsenal of Ralstonia solanacearum type III effectors and their in planta functions," *Mol. Plant Pathol.*, vol. 21, no. 10, pp. 1377–1388, 2020, doi: 10.1111/mpp.12977.

[16] Y. Suwa *et al.*, "Characterization of a Chromosomally Encoded 2,4-Dichlorophenoxyacetic Acid/␣-Ketoglutarate Dioxygenase from Burkholderia sp. Strain RASC," *APPL Env. MICROBIOL*, vol. 62, 1996.

[17] L. Han, D. Zhao, and C. Li, "Isolation and 2,4-D-degrading characteristics of Cupriavidus campinensis BJ71," *Braz. J. Microbiol. Publ. Braz. Soc. Microbiol.*, vol. 46, no. 2, pp. 433–441, Jun. 2015, doi: 10.1590/S1517-838246220140211.

[18] A. Zaprasis, Y.-J. Liu, S.-J. Liu, H. L. Drake, and M. A. Horn, "Abundance of novel and diverse tfdA-like

923 genes, encoding putative phenoxyalkanoic acid herbicide-degrading dioxygenases, in soil," *Appl.*
924 *Environ. Microbiol.*, vol. 76, no. 1, pp. 119–128, Jan. 2010, doi: 10.1128/AEM.01727-09.
925 [19] J. Baelum, T. Henriksen, H. C. B. Hansen, and C. S. Jacobsen, "Degradation of 4-chloro-2-
926 methylphenoxyacetic acid in top- and subsoil is quantitatively linked to the class III tfdA gene," *Appl.*
927 *Environ. Microbiol.*, vol. 72, no. 2, pp. 1476–1486, Feb. 2006, doi: 10.1128/AEM.72.2.1476-
928 1486.2006.
929 [20] D. T. Newby, T. J. Gentry, and I. L. Pepper, "Comparison of 2,4-Dichlorophenoxyacetic Acid
930 Degradation and Plasmid Transfer in Soil Resulting from Bioaugmentation with Two Different pJP4
931 Donors," *Appl. Environ. Microbiol.*, vol. 66, no. 8, pp. 3399–3407, Aug. 2000, doi:
932 10.1128/AEM.66.8.3399-3407.2000.
933 [21] H. P. Füchslin, I. Rüegg, J. R. Van Der Meer, and T. Egli, "Effect of integration of a GFP reporter gene
934 on fitness of Ralstonia eutropha during growth with 2,4-dichlorophenoxyacetic acid," *Environ.*
935 *Microbiol.*, vol. 5, no. 10, pp. 878–887, 2003, doi: 10.1046/j.1462-2920.2003.00479.x.
936 [22] P. T. Chivers, K. E. Prehoda, and R. T. Raines, "The CXXC Motif: A Rheostat in the Active Site,"
937 *Biochemistry*, vol. 36, no. 14, pp. 4061–4066, Apr. 1997, doi: 10.1021/bi9628580.
938 [23] S. Quan, I. Schneider, J. Pan, A. V. Hacht, and J. C. A. Bardwell, "The CXXC Motif Is More than a
939 Redox Rheostat *," *J. Biol. Chem.*, vol. 282, no. 39, pp. 28823–28833, Sep. 2007, doi:
940 10.1074/jbc.M705291200.
941 [24] J. Sancho, "Flavodoxins: sequence, folding, binding, function and beyond," *Cell. Mol. Life Sci.*, vol.
942 63, no. 7–8, pp. 855–864, Apr. 2006, doi: 10.1007/s00018-005-5514-4.
943 [25] D. Eulberg, S. Lakner, L. A. Golovleva, and M. Schlömann, "Characterization of a Protocatechuate
944 Catabolic Gene Cluster from Rhodococcus opacus 1CP: Evidence for a Merged Enzyme with 4-
945 Carboxymuconolactone-Decarboxylating and 3-Oxoadipate Enol-Lactone-Hydrolyzing Activity," *J.*
946 *Bacteriol.*, vol. 180, no. 5, pp. 1072–1081, Mar. 1998.
947 [26] D. B. Martim *et al.*, "Resolving the metabolism of monolignols and other lignin-related aromatic
948 compounds in Xanthomonas citri," *Nat. Commun.*, vol. 15, no. 1, p. 7994, Sep. 2024, doi:
949 10.1038/s41467-024-52367-6.
950 [27] F. Barbirato, J. C. Verdoes, J. A. M. de Bont, and M. J. van der Werf, "The Rhodococcus erythropolis
951 DCL14 limonene-1,2-epoxide hydrolase gene encodes an enzyme belonging to a novel class of
952 epoxide hydrolases," *FEBS Lett.*, vol. 438, no. 3, pp. 293–296, 1998, doi: 10.1016/S0014-
953 5793(98)01322-2.
954 [28] S. Doron *et al.*, "Systematic discovery of antiphage defense systems in the microbial pangenome,"
955 *Science*, vol. 359, no. 6379, p. eaar4120, Mar. 2018, doi: 10.1126/science.aar4120.
956 [29] R. Cheng *et al.*, "A nucleotide-sensing endonuclease from the Gabija bacterial defense system,"
957 *Nucleic Acids Res.*, vol. 49, no. 9, pp. 5216–5229, May 2021, doi: 10.1093/nar/gkab277.
958 [30] A. da Silva Xavier *et al.*, "Characterization of CRISPR-Cas systems in the Ralstonia solanacearum
959 species complex," *Mol. Plant Pathol.*, vol. 20, no. 2, pp. 223–239, Nov. 2018, doi:
960 10.1111/mpp.12750.
961 [31] J. He and M. W. Deem, "Heterogeneous Diversity of Spacers within CRISPR (Clustered Regularly
962 Interspaced Short Palindromic Repeats)," *Phys. Rev. Lett.*, vol. 105, no. 12, p. 128102, Sep. 2010, doi:
963 10.1103/PhysRevLett.105.128102.
964 [32] C. A. Rowley and M. M. Kendall, "To B12 or not to B12: Five questions on the role of cobalamin in
965 host-microbial interactions," *PLOS Pathog.*, vol. 15, no. 1, p. e1007479, Jan. 2019, doi:
966 10.1371/journal.ppat.1007479.
967 [33] Z. Freyberg and E. T. Harvill, "Pathogen manipulation of host metabolism: A common strategy for
968 immune evasion," *PLOS Pathog.*, vol. 13, no. 12, p. e1006669, Dec. 2017, doi:
969 10.1371/journal.ppat.1006669.
970 [34] J. Cai, Y. Jiang, E. S. Ritchie, A. P. Macho, F. Yu, and D. Wu, "Manipulation of plant metabolism by

971 pathogen effectors: more than just food," *FEMS Microbiol. Rev.*, vol. 47, no. 2, p. fuad007, Mar.
972 2023, doi: 10.1093/femsre/fuad007.
973 [35] J.-S. Eom *et al.*, "SWEETs, transporters for intracellular and intercellular sugar translocation," *Curr.*
974 *Opin. Plant Biol.*, vol. 25, pp. 53–62, Jun. 2015, doi: 10.1016/j.pbi.2015.04.005.
975 [36] C. D. Hamilton, O. R. Steidl, A. M. MacIntyre, C. G. Hendrich, and C. Allen, "Ralstonia solanacearum
976 Depends on Catabolism of Myo-Inositol, Sucrose, and Trehalose for Virulence in an Infection Stage–
977 Dependent Manner," *Mol. Plant-Microbe Interactions®*, vol. 34, no. 6, pp. 669–679, Jun. 2021, doi:
978 10.1094/MPMI-10-20-0298-R.
979 [37] A. N. Truchon *et al.*, "Plant-Pathogenic Ralstonia Phylotypes Evolved Divergent Respiratory
980 Strategies and Behaviors To Thrive in Xylem," *mBio*, vol. 14, no. 1, pp. e03188-22, Feb. 2023, doi:
981 10.1128/mbio.03188-22.
982 [38] R. Gopalan-Nair *et al.*, "Convergent Rewiring of the Virulence Regulatory Network Promotes
983 Adaptation of Ralstonia solanacearum on Resistant Tomato," *Mol. Biol. Evol.*, vol. 38, no. 5, pp.
984 1792–1808, May 2021, doi: 10.1093/molbev/msaa320.
985 [39] R. de Pedro-Jové *et al.*, "Dynamic expression of Ralstonia solanacearum virulence factors and
986 metabolism-controlling genes during plant infection," *BMC Genomics*, vol. 22, no. 1, p. 170, Mar.
987 2021, doi: 10.1186/s12864-021-07457-w.
988 [40] O. S. Gonçalves, M. V. de Queiroz, and M. F. Santana, "Potential evolutionary impact of integrative
989 and conjugative elements (ICEs) and genomic islands in the Ralstonia solanacearum species
990 complex," *Sci. Rep.*, vol. 10, no. 1, Art. no. 1, Jul. 2020, doi: 10.1038/s41598-020-69490-1.
991 [41] J. Ding *et al.*, "Properties of a Newly Identified Esterase from Bacillus sp. K91 and Its Novel Function
992 in Diisobutyl Phthalate Degradation," *PLoS ONE*, vol. 10, no. 3, p. e0119216, Mar. 2015, doi:
993 10.1371/journal.pone.0119216.
994 [42] N. Allocati, L. Federici, M. Masulli, and C. Di Ilio, "Glutathione transferases in bacteria," *FEBS J.*, vol.
995 276, no. 1, pp. 58–75, 2009, doi: 10.1111/j.1742-4658.2008.06743.x.
996 [43] J. E. L. Santos, M. A. Gómez, D. C. de Moura, M. Cerro-López, M. A. Quiroz, and C. A. Martínez-
997 Huitle, "Removal of herbicide 1-chloro-2,4-dinitrobenzene (DNCB) from aqueous solutions by
998 electrochemical oxidation using boron-doped diamond (BDD) and PbO2 electrodes," *J. Hazard.*
999 *Mater.*, vol. 402, p. 123850, Jan. 2021, doi: 10.1016/j.jhazmat.2020.123850.
1000 [44] V. Sauveplane *et al.*, "Arabidopsis thaliana CYP77A4 is the first cytochrome P450 able to catalyze the
1001 epoxidation of free fatty acids in plants," *FEBS J.*, vol. 276, no. 3, pp. 719–735, 2009, doi:
1002 10.1111/j.1742-4658.2008.06819.x.
1003 [45] R. N. Patkar *et al.*, "A fungal monooxygenase-derived jasmonate attenuates host innate immunity,"
1004 *Nat. Chem. Biol.*, vol. 11, no. 9, pp. 733–740, Sep. 2015, doi: 10.1038/nchembio.1885.
1005 [46] O. R. Feitosa-Junior *et al.*, "The Exometabolome of Xylella fastidiosa in Contact with
1006 Paraburkholderia phytofirmans Supernatant Reveals Changes in Nicotinamide, Amino Acids, Biotin,
1007 and Plant Hormones," *Metabolites*, vol. 14, no. 2, Art. no. 2, Feb. 2024, doi:
1008 10.3390/metabo14020082.
1009 [47] C. D. Hamilton, B. Zaricor, C. J. Dye, E. Dresserl, R. Michaels, and C. Allen, "Ralstonia solanacearum
1010 pandemic lineage strain UW551 overcomes inhibitory xylem chemistry to break tomato bacterial
1011 wilt resistance," *Mol. Plant Pathol.*, vol. 25, no. 1, p. e13395, 2024, doi: 10.1111/mpp.13395.
1012 [48] C. Baroukh, L. Cottret, E. Pires, R. Peyraud, A. Guidot, and S. Genin, "Insights into the metabolic
1013 specificities of pathogenic strains from the Ralstonia solanacearum species complex," *mSystems*,
1014 vol. 8, no. 4, pp. e00083-23, Jun. 2023, doi: 10.1128/msystems.00083-23.
1015 [49] A. S. Purnomo, A. Sariwati, and I. Kamei, "Synergistic interaction of a consortium of the brown-rot
1016 fungus Fomitopsis pinicola and the bacterium Ralstonia pickettii for DDT biodegradation," *Heliyon*,
1017 vol. 6, no. 6, p. e04027, Jun. 2020, doi: 10.1016/j.heliyon.2020.e04027.
1018 [50] N. Trefault *et al.*, "Genetic organization of the catabolic plasmid pJP4 from Ralstonia eutropha

1019 JMP134 (pJP4) reveals mechanisms of adaptation to chloroaromatic pollutants and evolution of
1020 specialized chloroaromatic degradation pathways," *Environ. Microbiol.*, vol. 6, no. 7, pp. 655–668,
1021 2004, doi: 10.1111/j.1462-2920.2004.00596.x.
1022 [51] "ICEberg 3.0." Accessed: May 28, 2024. [Online]. Available: https://tool2-
1023 mml.sjtu.edu.cn/ICEberg3/Browse_detail.php?iceid=335
1024 [52] J. B. Emerson *et al.*, "Host-linked soil viral ecology along a permafrost thaw gradient," *Nat.*
1025 *Microbiol.*, vol. 3, no. 8, pp. 870–880, Aug. 2018, doi: 10.1038/s41564-018-0190-y.
1026 [53] P. E. Dougherty *et al.*, "Widespread and largely unknown prophage activity, diversity, and function
1027 in two genera of wheat phyllosphere bacteria," *ISME J.*, vol. 17, no. 12, pp. 2415–2425, Dec. 2023,
1028 doi: 10.1038/s41396-023-01547-1.
1029 [54] H. C. McCann *et al.*, "Genomic Analysis of the Kiwifruit Pathogen Pseudomonas syringae pv.
1030 actinidiae Provides Insight into the Origins of an Emergent Plant Disease," *PLOS Pathog.*, vol. 9, no.
1031 7, p. e1003503, Jul. 2013, doi: 10.1371/journal.ppat.1003503.
1032 [55] H. C. McCann *et al.*, "Origin and Evolution of the Kiwifruit Canker Pandemic," *Genome Biol. Evol.*,
1033 vol. 9, no. 4, pp. 932–944, Apr. 2017, doi: 10.1093/gbe/evx055.
1034 [56] K. G. Pegg, L. M. Coates, W. T. O'Neill, and D. W. Turner, "The Epidemiology of Fusarium Wilt of
1035 Banana," *Front. Plant Sci.*, vol. 10, Dec. 2019, doi: 10.3389/fpls.2019.01395.
1036 [57] A. Drenth and G. Kema, "The Vulnerability of Bananas to Globally Emerging Disease Threats,"
1037 *Phytopathology®*, vol. 111, no. 12, pp. 2146–2161, Dec. 2021, doi: 10.1094/PHYTO-07-20-0311-
1038 RVW.
1039 [58] R. C. Ploetz, "Management of Fusarium wilt of banana: A review with special reference to tropical
1040 race 4," *Crop Prot.*, vol. 73, pp. 7–15, Jul. 2015, doi: 10.1016/j.cropro.2015.01.007.
1041 [59] F. A. García-Bastidas *et al.*, "First Report of Fusarium Wilt Tropical Race 4 in Cavendish Bananas
1042 Caused by Fusarium odoratissimum in Colombia," *Plant Dis.*, vol. 104, no. 3, pp. 994–994, Mar.
1043 2020, doi: 10.1094/PDIS-09-19-1922-PDN.
1044 [60] J. Beutler, S. Holden, S. Georgoulis, D. Williams, D. J. Norman, and T. M. Lowe-Power, "Whole
1045 Genome Sequencing Suggests that 'Nonpathogenicity on Banana (NPB)' Is the Ancestral State of the
1046 Ralstonia solanacearum IIB-4 Lineage," *PhytoFrontiers™*, vol. 3, no. 2, pp. 262–267, Sep. 2023, doi:
1047 10.1094/PHYTOFR-06-22-0068-SC.
1048 [61] J. I. Meier *et al.*, "Haplotype tagging reveals parallel formation of hybrid races in two butterfly
1049 species," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 118, no. 25, p. e2015005118, Jun. 2021, doi:
1050 10.1073/pnas.2015005118.
1051 [62] "Guppy protocol - Guppy software overview," Oxford Nanopore Technologies. Accessed: Jun. 12,
1052 2024. [Online]. Available: https://community.nanoporetech.com/protocols/Guppy-
1053 protocol/v/gpb_2003_v1_revax_14dec2018
1054 [63] S. W. Wingett and S. Andrews, "FastQ Screen: A tool for multi-genome mapping and quality
1055 control," Sep. 17, 2018, *F1000Research*: 7:1338. doi: 10.12688/f1000research.15931.2.
1056 [64] "BBMap," SourceForge. Accessed: Jun. 12, 2024. [Online]. Available:
1057 https://sourceforge.net/projects/bbmap/
1058 [65] S. Chen, Y. Zhou, Y. Chen, and J. Gu, "fastp: an ultra-fast all-in-one FASTQ preprocessor,"
1059 *Bioinformatics*, vol. 34, no. 17, pp. i884–i890, Sep. 2018, doi: 10.1093/bioinformatics/bty560.
1060 [66] S. Chen, "Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using
1061 fastp," *iMeta*, vol. 2, no. 2, p. e107, 2023, doi: 10.1002/imt2.107.
1062 [67] A. Bankevich *et al.*, "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell
1063 Sequencing," *J. Comput. Biol.*, vol. 19, no. 5, p. 455, May 2012, doi: 10.1089/cmb.2012.0021.
1064 [68] A. Prjibelski, D. Antipov, D. Meleshko, A. Lapidus, and A. Korobeynikov, "Using SPAdes De Novo
1065 Assembler," *Curr. Protoc. Bioinforma.*, vol. 70, no. 1, p. e102, 2020, doi:
1066 https://doi.org/10.1002/cpbi.102.

[69] A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler, “QUAST: quality assessment tool for genome assemblies,” *Bioinforma. Oxf. Engl.*, vol. 29, no. 8, pp. 1072–1075, Apr. 2013, doi: 10.1093/bioinformatics/btt086.
[70] R. Wick, *rrwick/Filtlong*. (Jun. 06, 2024). C++. Accessed: Jun. 12, 2024. [Online]. Available: https://github.com/rrwick/Filtlong
[71] R. R. Wick *et al.*, “Trycycler: consensus long-read assemblies for bacterial genomes,” *Genome Biol.*, vol. 22, no. 1, Art. no. 1, Dec. 2021, doi: 10.1186/s13059-021-02483-z.
[72] M. Kolmogorov, J. Yuan, Y. Lin, and P. A. Pevzner, “Assembly of long, error-prone reads using repeat graphs,” *Nat. Biotechnol.*, vol. 37, no. 5, pp. 540–546, May 2019, doi: 10.1038/s41587-019-0072-8.
[73] H. Li, “Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences,” *Bioinformatics*, vol. 32, no. 14, pp. 2103–2110, Jul. 2016, doi: 10.1093/bioinformatics/btw152.
[74] R. R. Wick and K. E. Holt, “Benchmarking of long-read assemblers for prokaryote whole genome sequencing,” Feb. 01, 2021, *F1000Research*: 8:2138. doi: 10.12688/f1000research.21782.4.
[75] R. Vaser and M. Šikić, “Time- and memory-efficient genome assembly with Raven,” *Nat. Comput. Sci.*, vol. 1, no. 5, pp. 332–336, May 2021, doi: 10.1038/s43588-021-00073-4.
[76] *nanoporetech/medaka*. (Jun. 07, 2024). Python. Oxford Nanopore Technologies. Accessed: Jun. 12, 2024. [Online]. Available: https://github.com/nanoporetech/medaka
[77] H. Li, “Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM,” May 26, 2013, *arXiv*: arXiv:1303.3997. doi: 10.48550/arXiv.1303.3997.
[78] Md. Vasimuddin, S. Misra, H. Li, and S. Aluru, “Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems,” in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, May 2019, pp. 314–324. doi: 10.1109/IPDPS.2019.00041.
[79] R. R. Wick and K. E. Holt, “Polypolish: Short-read polishing of long-read bacterial genome assemblies,” *PLOS Comput. Biol.*, vol. 18, no. 1, p. e1009802, Jan. 2022, doi: 10.1371/journal.pcbi.1009802.
[80] O. Schwengers, L. Jelonek, M. A. Dieckmann, S. Beyvers, J. Blom, and A. Goesmann, “Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification,” *Microb. Genomics*, vol. 7, no. 11, p. 000685, 2021, doi: 10.1099/mgen.0.000685.
[81] O. Schwengers, “Bakta database.” Zenodo, Feb. 20, 2023. doi: 10.5281/zenodo.7669534.
[82] D. Hyatt, G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, and L. J. Hauser, “Prodigal: prokaryotic gene recognition and translation initiation site identification,” *BMC Bioinformatics*, vol. 11, no. 1, p. 119, Mar. 2010, doi: 10.1186/1471-2105-11-119.
[83] “Genome,” NCBI. Accessed: Jun. 12, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=1310165,28097,305
[84] T. Lowe-Power and K. Chipman, “A Meta-analysis of the known Global Distribution and Host Range of the Ralstonia Species Complex,” Jul. 13, 2020, *bioRxiv*. doi: 10.1101/2020.07.13.189936.
[85] C. A. Boucher, F. Van Gijsegem, P. A. Barberis, M. Arlat, and C. Zischek, “Pseudomonas solanacearum genes controlling both pathogenicity on tomato and hypersensitivity on tobacco are clustered,” *J. Bacteriol.*, vol. 169, no. 12, pp. 5626–5632, Dec. 1987, doi: 10.1128/jb.169.12.5626-5632.1987.
[86] “Ralsto T3E.” Accessed: May 31, 2024. [Online]. Available: https://iant.toulouse.inra.fr/bacteria/annotation/site/prj/T3Ev3/
[87] B. Buchfink, C. Xie, and D. H. Huson, “Fast and sensitive protein alignment using DIAMOND,” *Nat. Methods*, vol. 12, no. 1, Art. no. 1, Jan. 2015, doi: 10.1038/nmeth.3176.
[88] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.
[89] G. Tonkin-Hill *et al.*, “Producing polished prokaryotic pangenomes with the Panaroo pipeline,” *Genome Biol.*, vol. 21, no. 1, p. 180, Jul. 2020, doi: 10.1186/s13059-020-02090-4.

1115 [90] A. J. Page *et al.*, “SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments,” *Microb.*
1116 *Genomics*, vol. 2, no. 4, p. e000056, 2016, doi: 10.1099/mgen.0.000056.
1117 [91] L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, and B. Q. Minh, “IQ-TREE: A Fast and Effective
1118 Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies,” *Mol. Biol. Evol.*, vol. 32, no.
1119 1, pp. 268–274, Jan. 2015, doi: 10.1093/molbev/msu300.
1120 [92] G. Yu, D. K. Smith, H. Zhu, Y. Guan, and T. T.-Y. Lam, “ggtree: an r package for visualization and
1121 annotation of phylogenetic trees with their covariates and other associated data,” *Methods Ecol.*
1122 *Evol.*, vol. 8, no. 1, pp. 28–36, 2017, doi: 10.1111/2041-210X.12628.
1123 [93] Y. Zhou, Y. Liang, K. H. Lynch, J. J. Dennis, and D. S. Wishart, “PHAST: a fast phage search tool,”
1124 *Nucleic Acids Res.*, vol. 39, no. Web Server issue, pp. W347-352, Jul. 2011, doi: 10.1093/nar/gkr485.
1125 [94] D. Arndt *et al.*, “PHASTER: a better, faster version of the PHAST phage search tool,” *Nucleic Acids*
1126 *Res.*, vol. 44, no. W1, pp. W16-21, Jul. 2016, doi: 10.1093/nar/gkw387.
1127 [95] D. S. Wishart *et al.*, “PHASTEST: faster than PHASTER, better than PHAST,” *Nucleic Acids Res.*, vol.
1128 51, no. W1, pp. W443–W450, May 2023, doi: 10.1093/nar/gkad382.
1129 [96] P. Siguier, J. Perochon, L. Lestrade, J. Mahillon, and M. Chandler, “ISfinder: the reference centre for
1130 bacterial insertion sequences,” *Nucleic Acids Res.*, vol. 34, no. Database issue, pp. D32–D36, Jan.
1131 2006, doi: 10.1093/nar/gkj014.
1132 [97] “Geneious | Bioinformatics Software for Sequence Data Analysis,” Geneious. Accessed: Jun. 12,
1133 2024. [Online]. Available: https://www.geneious.com/
1134 [98] A. R. Quinlan and I. M. Hall, “BEDTools: a flexible suite of utilities for comparing genomic features,”
1135 *Bioinforma. Oxf. Engl.*, vol. 26, no. 6, pp. 841–842, Mar. 2010, doi: 10.1093/bioinformatics/btq033.
1136 [99] H. Li, “Minimap2: pairwise alignment for nucleotide sequences,” *Bioinformatics*, vol. 34, no. 18, pp.
1137 3094–3100, Sep. 2018, doi: 10.1093/bioinformatics/bty191.
1138 [100] P. Danecek *et al.*, “Twelve years of SAMtools and BCFtools,” *GigaScience*, vol. 10, no. 2, p.
1139 giab008, Feb. 2021, doi: 10.1093/gigascience/giab008.
1140 [101] H. Li, “A statistical framework for SNP calling, mutation discovery, association mapping and
1141 population genetical parameter estimation from sequencing data,” *Bioinformatics*, vol. 27, no. 21,
1142 pp. 2987–2993, Nov. 2011, doi: 10.1093/bioinformatics/btr509.
1143 [102] B. S. Pedersen and A. R. Quinlan, “Mosdepth: quick coverage calculation for genomes and
1144 exomes,” *Bioinformatics*, vol. 34, no. 5, pp. 867–868, Mar. 2018, doi:
1145 10.1093/bioinformatics/btx699.
1146 [103] F. Lemoine and O. Gascuel, “Gotree/Goalign: toolkit and Go API to facilitate the development of
1147 phylogenetic workflows,” *NAR Genomics Bioinforma.*, vol. 3, no. 3, p. lqab075, Sep. 2021, doi:
1148 10.1093/nargab/lqab075.
1149 [104] X. Didelot and D. J. Wilson, “ClonalFrameML: Efficient Inference of Recombination in Whole
1150 Bacterial Genomes,” *PLOS Comput. Biol.*, vol. 11, no. 2, p. e1004041, Feb. 2015, doi:
1151 10.1371/journal.pcbi.1004041.
1152 [105] S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, and L. S. Jermiin, “ModelFinder:
1153 fast model selection for accurate phylogenetic estimates,” *Nat. Methods*, vol. 14, no. 6, pp. 587–
1154 589, Jun. 2017, doi: 10.1038/nmeth.4285.
1155 [106] X. Didelot, N. J. Croucher, S. D. Bentley, S. R. Harris, and D. J. Wilson, “Bayesian inference of
1156 ancestral dates on bacterial phylogenetic trees,” *Nucleic Acids Res.*, vol. 46, no. 22, p. e134, Dec.
1157 2018, doi: 10.1093/nar/gky783.
1158 [107] A. Rambaut, T. T. Lam, L. Max Carvalho, and O. G. Pybus, “Exploring the temporal structure of
1159 heterochronous sequences using TempEst (formerly Path-O-Gen),” *Virus Evol.*, vol. 2, no. 1, p.
1160 vew007, Jan. 2016, doi: 10.1093/ve/vew007.
1161 [108] D. H. Huson, D. C. Richter, C. Rausch, T. Dezulian, M. Franz, and R. Rupp, “Dendroscope: An
1162 interactive viewer for large phylogenetic trees,” *BMC Bioinformatics*, vol. 8, no. 1, p. 460, Nov. 2007,

1163 doi: 10.1186/1471-2105-8-460.
1164 [109] O. S. Alkhnbashi *et al.*, "CRISPRloci: comprehensive and accurate annotation of CRISPR–Cas
1165 systems," *Nucleic Acids Res.*, vol. 49, no. W1, pp. W125–W130, Jun. 2021, doi:
1166 10.1093/nar/gkab456.
1167 [110] R. H. Wang *et al.*, "PhageScope: a well-annotated bacteriophage database with automatic
1168 analyses and visualizations," *Nucleic Acids Res.*, vol. 52, no. D1, pp. D756–D761, Jan. 2024, doi:
1169 10.1093/nar/gkad979.
1170 [111] R. Cook *et al.*, "INfrastructure for a PHAge REference Database: Identification of Large-Scale
1171 Biases in the Current Collection of Cultured Phage Genomes," *PHAGE*, vol. 2, no. 4, pp. 214–223,
1172 Dec. 2021, doi: 10.1089/phage.2021.0007.
1173 [112] R. Zhang, M. Mirdita, E. Levy Karin, C. Norroy, C. Galiez, and J. Söding, "SpacePHARER: sensitive
1174 identification of phages from CRISPR spacers in prokaryotic hosts," *Bioinformatics*, vol. 37, no. 19,
1175 pp. 3364–3366, Oct. 2021, doi: 10.1093/bioinformatics/btab222.
1176 [113] T. Sakoparnig, C. Field, and E. van Nimwegen, "Whole genome phylogenies reflect the
1177 distributions of recombination rates for many bacterial species," *eLife*, vol. 10, p. e65366, Jan. 2021,
1178 doi: 10.7554/eLife.65366.
1179 [114] B. D. Ondov *et al.*, "Mash: fast genome and metagenome distance estimation using MinHash,"
1180 *Genome Biol.*, vol. 17, no. 1, p. 132, Jun. 2016, doi: 10.1186/s13059-016-0997-x.
1181 [115] O. Brynildsrud, J. Bohlin, L. Scheffer, and V. Eldholm, "Rapid scoring of genes in microbial pan-
1182 genome-wide association studies with Scoary," *Genome Biol.*, vol. 17, no. 1, p. 238, Nov. 2016, doi:
1183 10.1186/s13059-016-1108-8.
1184 [116] J. A. Lees, M. Galardini, S. D. Bentley, J. N. Weiser, and J. Corander, "pyseer: a comprehensive
1185 tool for microbial pangenome-wide association studies," *Bioinformatics*, vol. 34, no. 24, pp. 4310–
1186 4312, Dec. 2018, doi: 10.1093/bioinformatics/bty539.
1187 [117] The pandas development team, *pandas-dev/pandas: Pandas*. (Apr. 10, 2024). Zenodo. doi:
1188 10.5281/zenodo.10957263.
1189 [118] M. Galardini, *mgalardini/pyseer*. (Sep. 09, 2024). Python. Accessed: Jan. 20, 2024. [Online].
1190 Available: https://github.com/mgalardini/pyseer
1191 [119] G. Holley and P. Melsted, "Bifrost – Highly parallel construction and indexing of colored and
1192 compacted de Bruijn graphs," Aug. 13, 2019, *bioRxiv*. doi: 10.1101/695338.
1193 [120] J. R. Conway, A. Lex, and N. Gehlenborg, "UpSetR: an R package for the visualization of
1194 intersecting sets and their properties," *Bioinformatics*, vol. 33, no. 18, pp. 2938–2940, Sep. 2017,
1195 doi: 10.1093/bioinformatics/btx364.
1196 [121] E. Colombi *et al.*, "Comparative analysis of integrative and conjugative mobile genetic elements
1197 in the genus Mesorhizobium," *Microb. Genomics*, vol. 7, no. 10, p. 000657, 2021, doi:
1198 10.1099/mgen.0.000657.
1199 [122] F. Lu and G. Churchward, "Tn916 target DNA sequences bind the C-terminal domain of integrase
1200 protein with different affinities that correlate with transposon insertion frequency," *J. Bacteriol.*,
1201 vol. 177, no. 8, pp. 1938–1946, Apr. 1995, doi: 10.1128/jb.177.8.1938-1946.1995.
1202 [123] M. P. Garcillán-Barcia, S. Redondo-Salvo, L. Vielva, and F. de la Cruz, "MOBscan: Automated
1203 Annotation of MOB Relaxases," *Methods Mol. Biol. Clifton NJ*, vol. 2075, pp. 295–308, 2020, doi:
1204 10.1007/978-1-4939-9877-7_21.
1205 [124] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search
1206 tool," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, Oct. 1990, doi: 10.1016/S0022-2836(05)80360-2.
1207 [125] S. S. Abby, J. Cury, J. Guglielmini, B. Néron, M. Touchon, and E. P. C. Rocha, "Identification of
1208 protein secretion systems in bacterial genomes," *Sci. Rep.*, vol. 6, no. 1, p. 23080, Mar. 2016, doi:
1209 10.1038/srep23080.
1210 [126] A. C. E. Darling, B. Mau, F. R. Blattner, and N. T. Perna, "Mauve: Multiple Alignment of

Conserved Genomic Sequence With Rearrangements," *Genome Res.*, vol. 14, no. 7, pp. 1394–1403, Jul. 2004, doi: 10.1101/gr.2289704.

[127] *pycountry/pycountry*. (Aug. 28, 2024). Python. pycountry. Accessed: Feb. 04, 2021. [Online]. Available: https://github.com/pycountry/pycountry

[128] "Welcome to GeoPy's documentation! — GeoPy 2.4.1 documentation." Accessed: Jun. 12, 2024. [Online]. Available: https://geopy.readthedocs.io/en/stable/

[129] "france-geojson/regions/martinique/communes-martinique.geojson at master · gregoiredavid/france-geojson," GitHub. Accessed: May 31, 2024. [Online]. Available: https://github.com/gregoiredavid/france-geojson/blob/master/regions/martinique/communes-martinique.geojson

[130] J. V. den Bossche *et al.*, *geopandas/geopandas: v0.14.3*. (Jan. 31, 2024). Zenodo. doi: 10.5281/zenodo.10601680.

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.