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 a new lineage with dramatically expanded host range had emerged. Our analysis shows the outbreak in Martinique was caused by the expansion of lineages from two different *Ralstonia* species. One lineage is of Asian origin and 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 (ICEs) in both lineages suggests they are exchanging genetic material. The emergent lineage acquired novel prophage along with genes encoding novel catalytic and metabolic functions, but shows little variation among the virulence proteins considered host specificity determinants. We postulate that host range expansion is not caused by the evolution of a virulence strategy targeting host immunity, but is rather due to the acquisition of novel pathways for substrate metabolism enhancing pathogen fitness in soil or plant tissues.

Introduction

Clarifying the mechanisms underlying pathogen emergence and adaptation to novel hosts is of fundamental importance in deepening our understanding of how microbes evolve and for safeguarding human health and agricultural productivity. Pathogen infections of crop plants impose severe losses on both small-scale and commercial growers. Studies of the molecular mechanisms by which 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 repertoire of virulence factors is shaped by selection in different hosts, and results in the emergence of specialised pathogen lineages. The identification and characterisation of pathogen virulence and host resistance genes underlying host-specific interactions provide insight into molecular mechanisms underlying disease establishment. Sequencing isolates shows recombination and horizontal gene transfer play important roles in shaping virulence gene repertoires, but the genetic determinants of pathogen host range remain undiscovered for the majority of plant pathogens. We have similarly very little understanding of the evolutionary processes and genomic changes associated with outbreak events.

We sought to identify the evolutionary origins of an outbreak lineage and the particular genomic changes associated with its emergence and shift in host range by sequencing the largest collection of *Ralstonia* spp. sampled during an outbreak event. The emergence of a new lineage of *R. solanacearum* in the Caribbean island of Martinique represents a rare opportunity to address fundamental questions in a bacterial plant pathogen of global economic importance. Multiple lineages of *R. solanacearum* are known to infect banana, which is cultivated on an industrial scale in the Caribbean, Central and South America. The pathogen is considered highly aggressive, causing lethal wilting symptoms in susceptible hosts. *R. solanacearum* is subject to extensive monitoring and movement controls to limit its spread, since it is also known for rapid dissemination and long-term persistence in freshwater and soil, not to mention lethal wilting disease in susceptible hosts. An outbreak of disease on cultivated ornamental plants (*Anthurium andreanum*) in Martinique in 1999 caused alarm when the strain recovered was found to be closely related to 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]. An extensive study was therefore initiated to sample *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 used this collection to explore the origins of this unusual host range expansion and found the outbreak was caused by two co-occurring, distantly related lineages. One lineage retains a signature of host specialisation and was initially introduced from Asia, while the lineage that underwent apparent host range expansion originated from a population diversifying in the Americas for some time. Mobile elements are circulating between both lineages, serving as vehicles for the movement of 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 significant patterns of dissemination and epidemic expansion in Central and South America. 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 are evident in the data. A global outbreak of Rso2BS1 on potato (*S. tuberosum*) has an ultimate source in the Americas, while at least four Rps lineages have been introduced to the Americas from Asia (Rps1S14, Rps1S17, Rps1S47, Rps1S48).

The introduction of several Rps lineages into the Americas resulted in their establishment and local diversification, primarily in solanaceous hosts. Rps1S18 is the most frequently isolated Rps lineage in the Americas, where a secondary dissemination event resulted in its introduction to Martinique. Despite the large number of sequenced Rps1S18 isolates and their distribution across Martinique, French Guiana, Brazil and Peru, the lineage displays limited diversity in the core genome and T3E repertoire, consistent with recent epidemic expansion (Fig. 1, Fig. S1). Higher resolution phylogenetic inference of the history and diversity of Rps1S18 shows the most basal members of the lineage are *S. tuberosum* isolates from South Africa (UW393) and Indonesia (UW505) (Fig. S2). Rps1S18 shares recent ancestry with Rps1S14, isolated from solanaceous hosts in Central America, South America and Senegal. The introduction of Rps lineages into Central and South America may have occurred from multiple sources prior to 1961, potentially via the movement of infected plant material. Rps1S18 then diversified across Central and South America prior to its introduction into Martinique, where it displays a broader host range of all Rps lineages in the Americas. Rps1S18 isolates in Martinique were found on non-solanaceous hosts like *Cucurbita* spp., *Anthurium* spp., fallow soil and unidentified weeds, while other Rps isolates from the Americas were sampled from solanaceous hosts.

There is far greater diversity of Rso in the Americas compared to Rps: the center of origin of Rso and its solanaceous hosts alike spans Central and South America. Among the two *Ralstonia* species and multiple lineages sampled in both French Guiana and Martinique, Rso2A encompasses the greatest intraspecific pathogen diversity. There is no evidence of recent clonal expansion among Rso2A lineages, which instead form a broad 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 [76].

Multiple Rso2B lineages are responsible for both regional and global outbreak events. Emerging in South America, the Rso2BS1 (formerly race 3 biovar 2) lineage spread globally, causing bacterial wilt infections on potato (*Solanum tuberosum*). Rso2B1 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, the recent global expansion suggests the likeliest mechanism of long-distance dissemination is via international trade in infected seed potatoes,

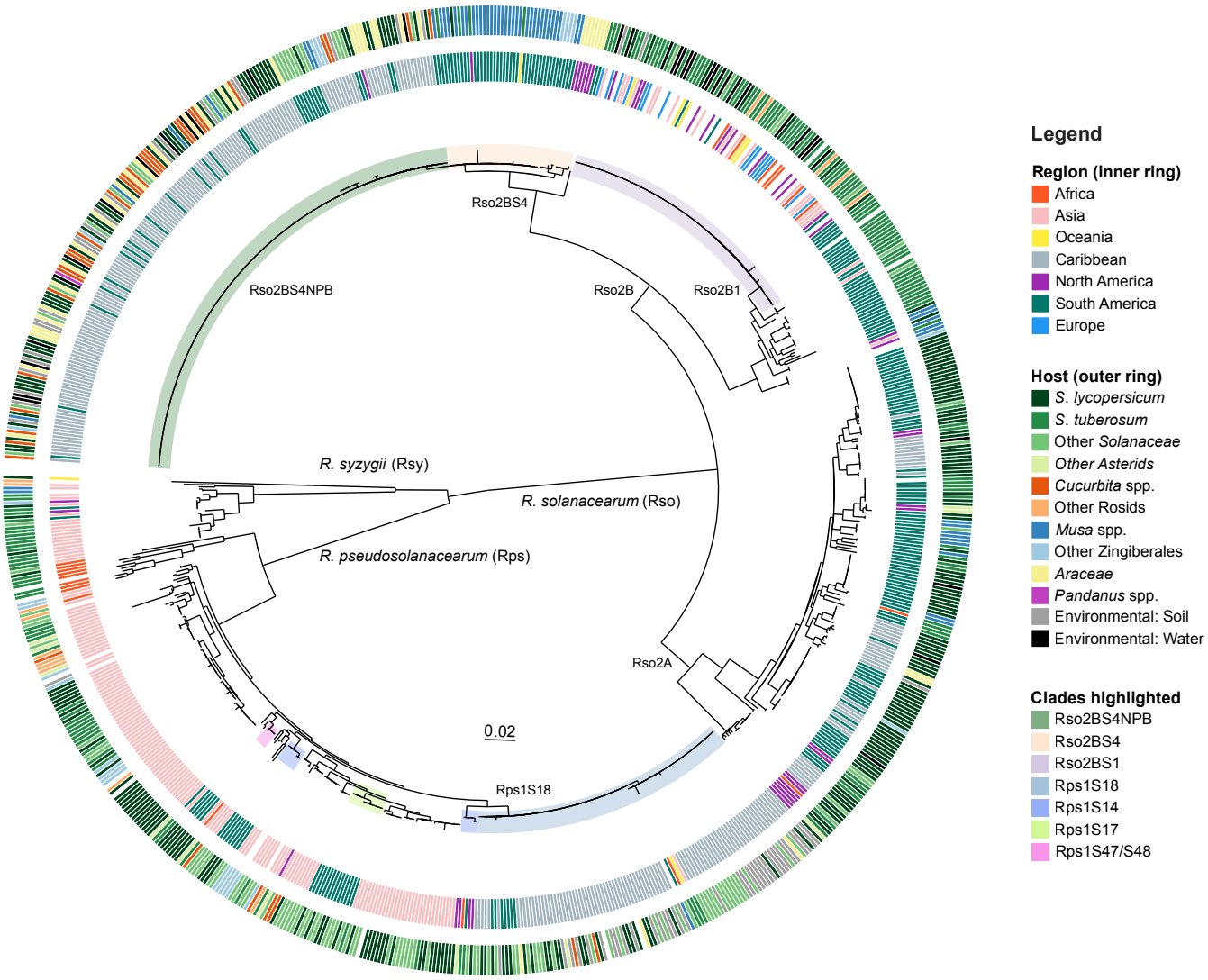


Fig. 1. Core gene tree of the Ralstonia solanacearum species complex. Maximum likelihood tree of 816 RSSC genomes inferred from a 137,423bp SNP alignment obtained after core gene identification, alignment generation and invariant site removal. Nodes with the bootstrap support values under 70 are collapsed. 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. Environmental isolates from soil and water are shown in grey and black respectively. A single *Pandanus* spp. isolate is shown in purple. Missing data is shown in white.

highlighting the elevated risk of disease transmission during movement of infected plant material [77] [78]. Rso2B lineages are also responsible for Moko disease outbreaks in South American banana plantations, while Rso2BS4NPB has the greatest documented host range of any lineage in the species complex, despite being primarily restricted to Martinique.

Banana (*Musa* spp.) is a globally significant commodity and subsistence crop, whose cultivation typically involves high density single variety monocultures. Multiple outbreaks of Rso on banana have been documented in the Americas, where it is referred to as Moko disease. The Rso lineages infecting banana are clearly polyphyletic, and include four Rso2A lineages primarily isolated in Brazil, two basal lineages in the Rso2B clade from which the global potato outbreak lineage emerged (Rso2BS25 [79] and Rso2BS3 [27]) and basal members of Rso2BS4. The latter are closely related to the lineage that underwent host range expansion in Martinique: Rso2BS4NPB. These banana-infecting lineages are collectively responsible for outbreaks of Moko disease in Panama, Costa Rica, Colombia, Peru, Brazil and French Guiana from the 1960s onwards (Fig. 1, Fig. S3). Banana was introduced to the Americas as a crop in the 16th century [80], so it is a matter of some interest that the most deeply branching basal members of Rso2B are isolated from *Musa* spp. along with *Heliconia* spp. and *Anthurium* spp., both endemic to the Americas [81], [82]. The expansion of plantation-style banana cultivation is likely to have afforded multiple opportunities for Rso to emerge and spread in agricultural areas, allowing it to spread to other crop varieties, both endemic and introduced.

Epidemic expansion of two lineages after their dissemination from South America

The species complex tree shows the surge in plant infections in Martinique is linked with the recent clonal expansion of two lineages across the island: Rso2BS4NPB and Rps1S18 (Fig. 1). The latter lineage is historically significant, as it is the source of the first sequenced *Ralstonia* genome and commonly used model strain GMI1000, isolated in French Guiana in 1978. We investigated the origins of both lineages using well resolved, non-recombinant core genome trees including closely related strains from mainland populations (Fig. 2).

South American mainland strains most closely related to the emergent Rso2BS4NPB lineage in Martinique were all isolated from solanaceous hosts or freshwater sources on the mainland, along with a single isolate from pumpkin (*Cucurbita pepo*). These basal isolates all infect hosts that evolved in the Americas and represent the ancestral population from which Rso2BS4NPB emerged to sweep across Martinique. Two strains (RUN2226 and RUN2227) were isolated from tomato plants grown in areas planted the first season after rainforest clearing, suggesting the ancestral population has a long coevolutionary history with hosts endemic to the area. There was evidently limited spread of Rso2BS4NPB on the mainland compared to the dramatic epidemic expansion of the derived Martinique isolates. The derived lineage has extremely low genetic diversity, differing by only 166 (among them 73 parsimony informative) nonrecombinant SNPs in the core genome. There is a single Martinique isolate (RUN0016) isolated from *A. andreanum* in Saint Joseph, Martinique that groups with basal mainland isolates. The detection of additional derived Rso2BS4NPB isolates from the same host species in Saint Joseph suggest there is an epidemiological link with the introduction and spread of the emergent lineage in Martinique (Fig. S2).

The apparent source and dissemination of Rps1S18 parallels the history of Rso2BS4NPB: Martinique Rps1S18 is similarly derived from a mainland population circulating on solanaceous hosts (Fig. 2b). There is no evidence of a dramatic host range expansion however: with few exceptions, the majority of Martinique Rps1S18 was isolated from soil or solanaceous hosts. The sole exceptions are two

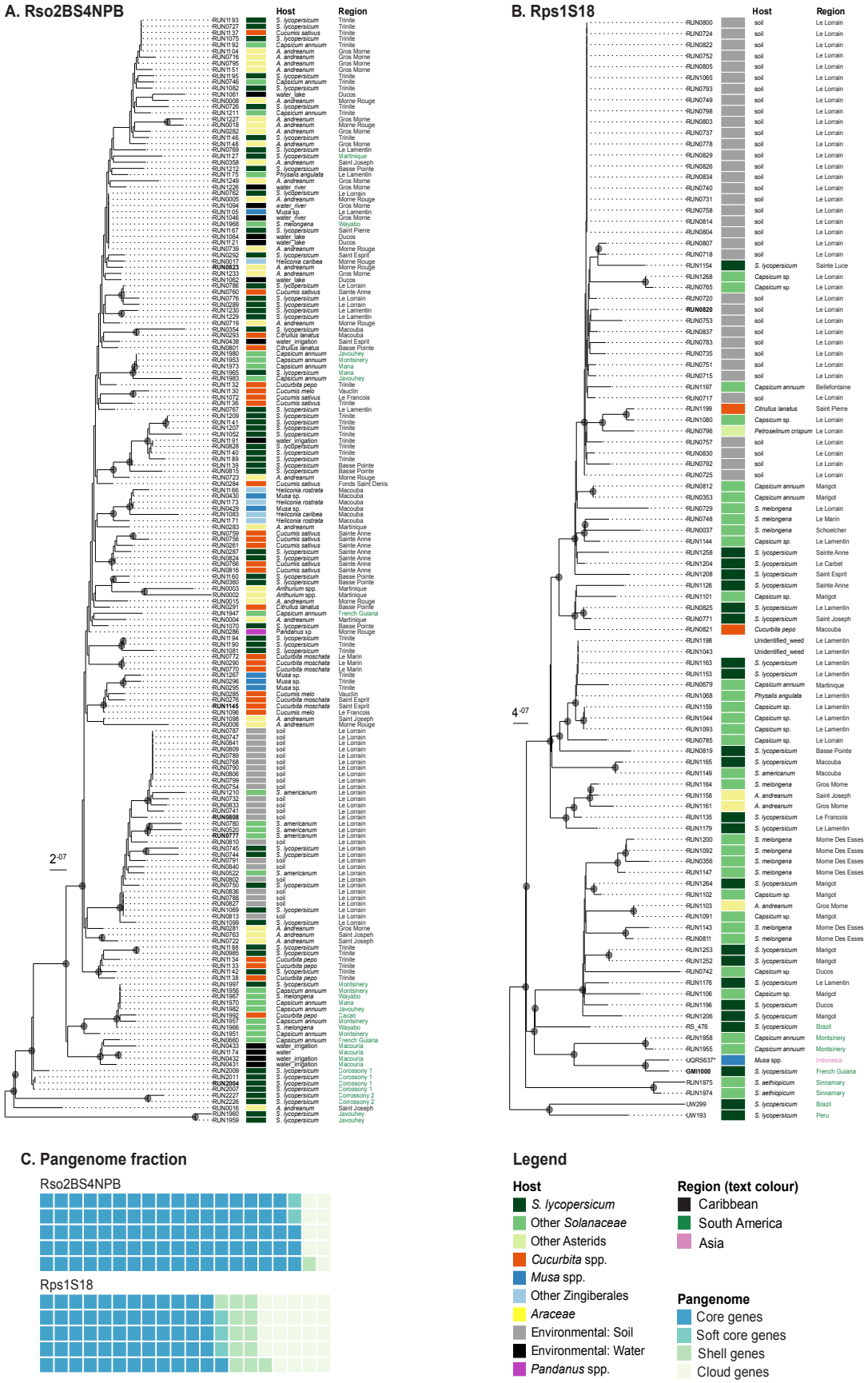


Fig. 2. Outbreak lineage core genome trees. (A) Rso2BS4NPB maximum likelihood tree calculated from a 5,206,677 bp non-recombinant non-mobile core genome alignment with 195 SNPs and 73 parsimony informative sites. Internal nodes with bootstrap value above 80 are indicated with black circles. Strain labels shown in bold are closed genomes generated in this study and/or used as mapping references. (B) Rps1S18 maximum likelihood tree calculated from a 4,760,272 bp non-recombinant core genome alignment with 311 SNPs and 116 parsimony informative sites. The asterisk for UQRS637 refers to identical strain under different label: UW745. Strain labels shown in bold are closed genomes generated in this study and/or used as mapping references. (C) Pangenome waffle chart representing the proportion of core (present in 99 - 100% of genomes), soft-core (95 - 99% of genomes), shell (15 - 95% of genomes) and cloud genes (present in less than 15% of genomes) identified by Panaroo, displayed separately for each lineage pangenome.

180 isolates from cucurbits and three from *A. andreanum*. Unexpectedly, one *Musa* isolate
181 UQRS637_UW745 from Indonesia was assigned to the Rps1S18 lineage. This should be interpreted
182 with caution however, as the assembly version of UQRS637_UW745 used here was later suppressed
183 on NCBI. There is evidence of structuring by sampling location for one site: just as in the
184 Rso2BS4NPB tree, Rps1S18 isolates from Le Lorrain form a well-supported clade. Most Le Lorrain
185 isolates - whether Rps1S18 or Rso2BS4NPB - were recovered from soils (Fig. 2).
186
187 We used well-resolved non-recombinant trees to determine the divergence time of Rps1S18 and
188 Rso2BS4NPB. The best estimates for the divergence time of the Rps1S18 lineage is between the
189 1850s - 1930s (Fig. S5), while estimated divergence times for the split between basal solanaceous and
190 derived host-range expanded Rso2BS4NPB sublineage is between the 1930s and 1980s (Fig. S6).
191 However, the estimated divergence dates are poorly supported as there is insufficient temporal signal
192 in the data even after removing potential outlier samples and collapsing weakly supported nodes in the
193 phylogenies. R-squared values for root-to-tip regressions were 0.05 and 0.09 for Rso2BS4NPB and
194 Rps1S18, respectively, and random date permutation analyses by BactDating showed estimated
195 divergence times for both lineages were better than random in only 50% of simulations, indicating
196 there is no temporal signal in the data. Pangenome analysis for each lineage showed both lineages
197 have similar numbers of core and soft-core genes (4,565 for Rps1S18 and 4,792 for Rso2BS4NPB),
198 but remarkably different accessory genome sizes: 487 accessory genes in Rso2BS4NPB and 2,636 in
199 Rps1S18, amounting to 10% and 35% of each lineage pangenome, respectively (Fig. 2). The larger
200 Rps1S18 accessory genome suggests this lineage has higher rates of horizontal exchange, or that it
201 was introduced to Martinique prior to Rso2BS4NPB, and therefore had more time to acquire
202 accessory genes.

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

204 We hypothesized the co-occurrence of two pathogen lineages in the same host or niche provide
205 opportunities for between-lineage homologous recombination. ClonalFrameML analysis for each
206 lineage indicated they exhibit similar relative ratios of recombination to mutation (R/theta, 0.159 for
207 Rps1S18 and 0.152 for Rso2S4NPB). RECOPHY found no evidence of between-lineage homologous
208 recombination however (Fig. S7). Distance trees for each region in the core genome alignment with
209 reduced SNP density (under 10 SNPs/kb) and one high SNP density region (120 SNPs/kb compared
210 to an alignment-wide average of 53 SNPs/kb) display the same topology as the core genome tree,
211 indicating homologous recombination is unlikely to have occurred in these regions.
212
213 Although we found no evidence of homologous recombination between Rso2BS4NPB and Rps1S18,
214 pangenome analysis and association testing revealed the presence of an integrative and conjugative
215 element (ICE) that may have been exchanged between these co-occurring lineages in Martinique (Fig.
216 3). We identified ICEs across the entire species complex using the set of 133 complete RSSC
217 genomes, and found 13 complete ICEs integrated in three chromosomal integration sites. Another 17
218 putative ICEs were identified that lack flanking direct repeats, but satisfy the other criteria for ICE
219 identification (presence of T4SS gene clusters, a *mob* relaxase and integrase sequence). Some
220 genomes have *mob* relaxase genes framed with direct repeats, but lack adjacent T4SS, these may be
221 relics of degenerate ICEs or integrative mobilizable elements (IMEs). Isolates typically carry a single
222 ICE, and all but seven of the 30 putative ICEs identified in the species complex are present in *R.*
223 *pseudosolanacearum* genomes alone: two partial ICEs are present in *R. syzygii* and five complete
224 ICEs are exclusively present in *R. solanacearum* from Martinique. The ICEs in Rso2BS4NPB
225 genomes from Martinique are identical to each other, and homologous to a 45.9kb ICE in Rps1S18

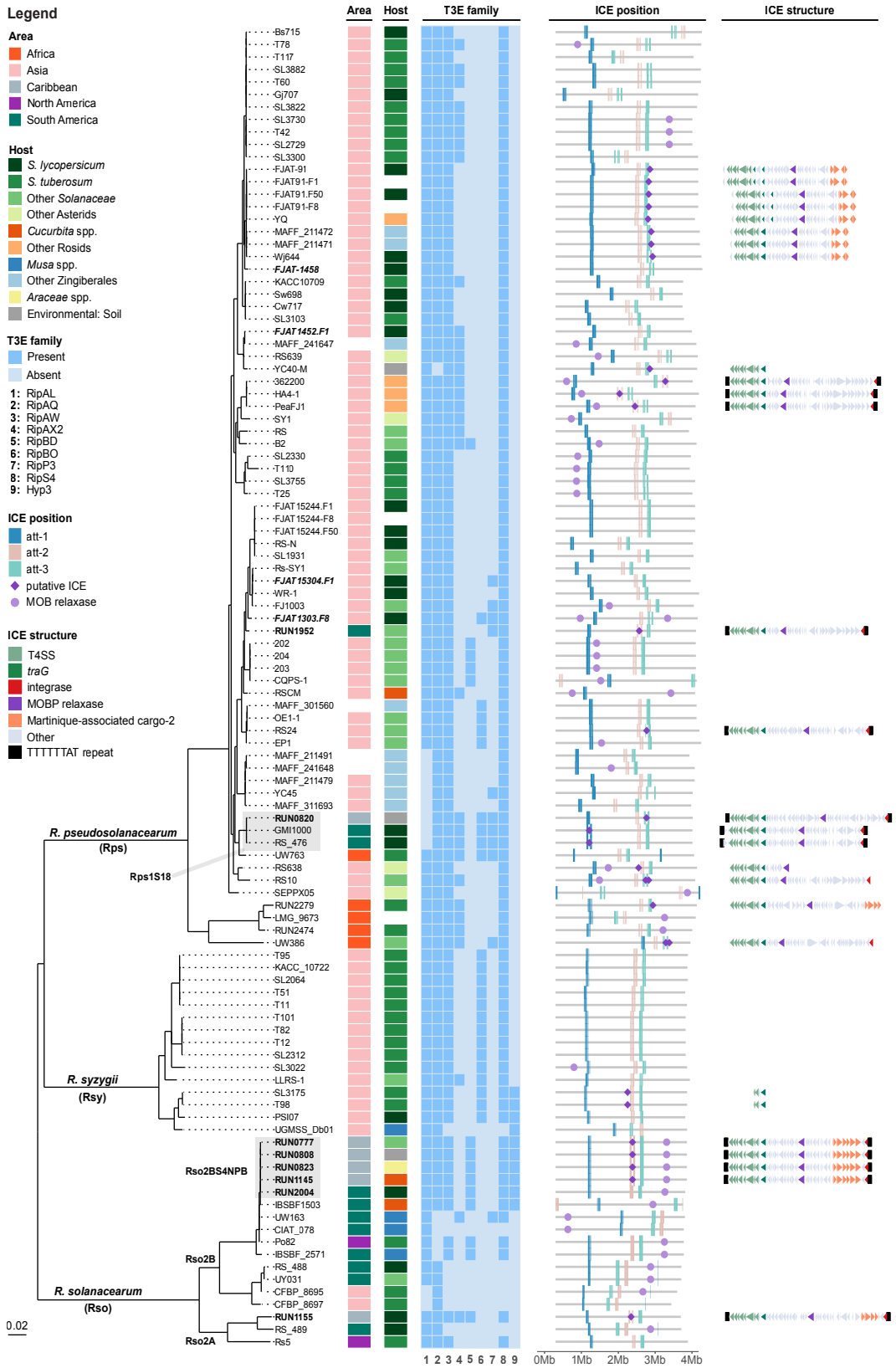


Fig. 3. Virulence gene and ICE distribution across the R. solanacearum species complex. Maximum likelihood tree of all completely assembled RSSC genomes available in this study. Labels of eight strains from Martinique and French Guiana sequenced by long-read technology during this study are shown in bold. The labels shown in bold italic represent clades collapsed due to low within-clade diversity (Fig. S13). Geographical region and host of isolation are shown for each strain. Nine T3E families exhibiting presence/absence variation for Rso2BS4 and Rps1S18 clades are shown in the third metadata column: dark blue indicates T3E family presence, light blue indicates absence. Fourth column shows position of all annotated relaxases and 100kb ICE attachment regions sharing synteny with ICEs found in Rps1S18 GMI1000 (att-1, blue), Rso2BS4NPB RUN1145 (att-2, pink) and Rso2BS4NPB RUN1960 (att-3, green) on chromosome 1 of each RSSC strain. Thin grey lines represent the length ranges of chromosome 1. Purple dots represent MOB relaxases, bold purple diamonds are relaxases located next to T4SS genes. Final column shows the composition of putative ICEs identified in RSSC strains. Green CDSs represent homologs to T4SS genes in Rso2BS4NPB and Rps1S18 genomes; orange CDSs represent CDS associated with Martinique by panGWAS. Purple and red CDS indicate MOBP relaxases and integrases respectively. Black bars around the loci represent the TTTTTTAT repeat motif delimiting putative ICE boundaries.

226 GMI1000. The ICE is also present in seven other Rps genomes that collectively span the diversity of
227 *R. pseudosolanacearum*. Within *R. solanacearum* it is solely present in Rso2BS4NPB and Rso2A
228 RUN1155, another isolate from Martinique (Fig. 3). Since this ICE is broadly distributed among Rps
229 and present in the reference Rps1S18 GMI1000, we named it $ICE^{RpsGMI1000}$.
230
231 Our exploration of the spatial structure of genetic variation revealed hotspots of ICE exchange among
232 the densely sampled Martinique populations. All but one of the 27 genes associated with isolation in
233 specific regions in Martinique are present on the $ICE^{RsoGMI1000}$ (Table S11, Table S12, Fig. S10). A
234 kmer-based approach also found 625 out of 2,367 Martinique-associated kmers and all 83 soil-
235 associated kmers map to the $ICE^{RsoGMI1000}$. Precise association between one cluster of eight genes on
236 the $ICE^{RsoGMI1000}$ and isolation from Trinite and Morne Rouge was found, while another two gene
237 clusters have overlapping associations with isolation from soil and from Le Lorrain (Fig. S10, Fig.
238 S14). These clusters correspond to two backbone regions and two hotspots of cargo gene integration
239 (cargo-1 and cargo-2, Fig. 5). Identification of $ICE^{RpsGMI1000}$ presence and cargo region carriage in
240 every Martinique isolate sequenced shows that while there are multiple regions in which both lineages
241 co-occur, Rps1S18 and Rso2BS4NPB isolates in Le Lorrain carry ICEs both with and without the
242 cargo-2 gene cluster. The strain and ICE-level diversity in Le Lorrain suggests it is a region of
243 heightened exchange. Rso2BS4NPB acquisition of $ICE^{RsoGMI1000}$ may have occurred in this region,
244 followed by dissemination to other areas across the island.
245
246 The $ICE^{GMI1000}$ cargo-1 gene cluster carries metabolism-associated genes, while the cargo-2 cluster
247 carries bacterial defense systems (Fig. S15). The cargo-1 region present in Rps1S18 and
248 Rso2BS4NPB (e.g. RUN0777 and RUN1960) includes two *limA* genes encoding limonene 1,2-
249 epoxide hydrolases [83] (OFBKGH_09550, OFBKGH_09555 in RUN1145); a phthalate ester
250 hydrolase *pnbA*; an antibiotic biosynthesis monooxygenase *ygiN* (also known as ABM,
251 OFBKGH_09580); a glutathione S-transferase *gstA* linked with the detoxification of organochlorine
252 compounds (1-chloro-2,4-dinitrobenzene) used as an herbicide and pesticides; a
253 pyrazinamidase/nicotinamidase *pncA*, which converts nicotinamide to nicotinic acid, and *fabG*, which
254 encodes a reducing agent of ethyl 4-chloroacetoacetate, an intermediate used in herbicide and
255 pesticide production.
256
257 The cargo gene clusters cargo-1 and cargo-2 are simply named according to the distinct integration
258 site within the ICE. While the cargo-1 clusters always encode genes linked with metabolism, the
259 cargo-2 gene clusters all encode defence systems. For example, the distinct cargo-2 clusters carried by
260 Rps1S18 GMI1000 $ICE^{RpsGMI1000}$ and Rso2BS4NPB $ICE^{RpsGMI1000}$ each carry multiple defence
261 systems, including the anti-plasmid/phage Lamassu island and anti-viral Gabija islands [84], [85]. The
262 only other ICE present in Rso outside of Rso2BS4NPB is in Rso2A RUN1155, also from Martinique,
263 shares a similar cargo-2 region with Rso2BS4NPB $ICE^{GMI1000}$ (Fig. 5, 6). ICE-mediated horizontal
264 transfer of gene modules encoding novel metabolic function may be linked with increased
265 pathogenicity and epidemic expansion of both Rso2BS4NPB and Rps1S18 in Martinique. The
266 surprising cargo gene partitioning between metabolic function and defense systems has never
267 previously been reported for ICEs.

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

269 Despite the reduced genetic diversity indicative of very recent expansion, the Rso2BS4NPB isolates
270 from Martinique encompass the broadest known host range of any single lineage in the species
271 complex (Fig. 2) The only host this lineage has not been isolated from is *S. tuberosum*. The

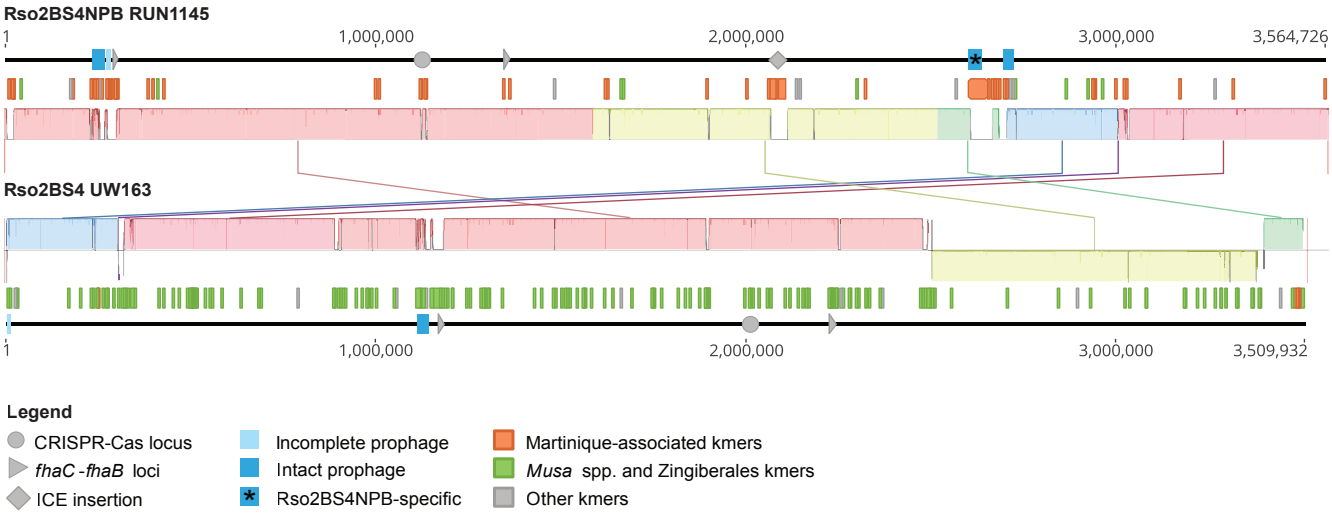


Fig. 4. Mobile elements and host-associated kmers in Rso2BS4 and Rso2BS4NPB. ProgressiveMauve alignment of the first chromosomes of Rso2BS4NPB RUN1145 and Rso2BS4 UW163 strains shown in grayscale together with mapping positions of kmers reported by GWAS and annotated chromosome length scales. The alignment starts from the RUN1145 origin of replication. UW163 origin of replication is located at the approximate position 882,000. The chromosome length scales are shown as black lines above RUN1145 and below UW163 and annotated with positions of prophages (blue rectangles), CRISPR-Cas loci (grey stars), fhaC-fhaB loci (grey arrows) and ICE insertion in RUN1145 (grey diamond). The prophage found uniquely in Rso2B4NPB strains is marked with an asterisk. Merged mapping positions of the kmers are represented by rectangles between the Mauve alignment and chromosome length scales. Merging was applied to overlapping kmers and kmers located closer than 1kb from each other. The kmers frequently found in the strains sampled in Martinique are shown in orange. The kmers frequently found in *Musa* spp. or Zingiberales isolates are shown in green. The kmers associated with any other trait are shown in grey.

Rso2BS4NPB tree shows strain clustering by host and location. Notably, most Rso2BS4NPB and Rps1S18 isolates from Le Lorrain cluster together, regardless of whether they were isolated from soil, cultivated or wild *Solanum* spp. Both basal and derived Rso2BS4NPB were isolated from *Solanum* and *Cucurbita* spp. in Trinite. Rso2BS4NPB was also isolated from symptomatic banana (*Musa* spp.) in that location (RUN0296, RUN0295), indicating the appellation 'Non-Pathogenic on Banana' (NPB) for this lineage may not be wholly accurate. Although Rps1S18 and Rso2BS4NPB alike were isolated from soil, freshwater isolations yielded Rso2BS4NPB alone.

We sought to identify the genes underlying host range expansion in Rso2BS4NPB by assessing changes in the repertoire of Type 3 secreted effectors (T3Es). The delivery of virulence proteins by the Type 3 secretion system (T3SS) is known to play an important role in successful host colonisation [86], [87]. The repertoire of T3Es carried by a strain is thought to play an important role in host specificity [88], [89]. We assessed the T3E repertoires across each genome and found no link between the emergence of Rso2BS4NPB and any specific T3E gain or loss (Table S4, Fig. 3 and Fig. S1). Rso2BS4NPB genomes display nearly uniform T3E repertoires, consistent with recent population expansion (Fig. S1). However one T3E is uniquely present among the emergent lineage and its sister lineage Rso2BS4: the hypothetical effector T3E_Hyp13 (UW163_19700 in Rso2BS4 UW163, OFBKGH_20220 in Rso2BS4NPB RUN1145) which is otherwise absent in all other RSSC genomes (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 from in *Heliconia* and *Musa* RsoS4 isolates from Colombia, other Rso2B *Musa* isolates and some *Rsy* strains (Fig. S3). In sum, there is no evidence of T3E repertoire variation that may be linked with the host range expansion or emergence of Rso2BS4NPB specifically.

The Rso2BS4 isolates from cultivated monocot hosts that are endemic (*Heliconia* spp.) and introduced (*Musa* spp. and *Epipremnum* spp.) to the Americas exhibit more variation in effector repertoire than Rso2BS4NPB. RipAX2 and RipP3 are present in Rso2BS4 and absent in Rso2BS4NPB (Fig. 3, Fig. S3). RipAX2 is present in many other Rso2B and Rso2A clades, indicating effector loss in Rso2BS4NPB is likely, while RipP3 is sparsely distributed in Rso2B. Four effector families exhibit evidence of likely loss in the monocot-infecting Rso2BS4: RipAQ, RipAW, RipBD and a predicted effector called T3E_Hyp3. The loss of RipAQ is most striking as the family is otherwise present in all other RSSC strains. RipAW is missing from most Rso2BS4 banana isolates from Peru, most of the pandemic Rso2BS1 potato isolates and some *R. s*yzygii, but is present in all other RSSC strains. RipBD is missing from Peruvian Rso2BS4 strains, but is otherwise common in Rso2A. Members of the hypothetical effector family T3E_Hyp3 are infrequently represented across the RSSC: present in Rso2BS4NPB, absent in the closely related sister clade of Rso2BS4, patchily distributed in other Rso2B and *R. syzygii*, and absent elsewhere. Since these effector families are present among related clades and/or across the species complex, the most parsimonious explanation of the observed pattern is they were lost in the ancestor of some 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

The host range expansion and epidemic spread of Rso2BS4NPB occurred as a result of genomic changes in the ancestor of this lineage. We therefore explored whether recombination events introducing novel genes might be linked with the emergence event. Pangenome exploration revealed

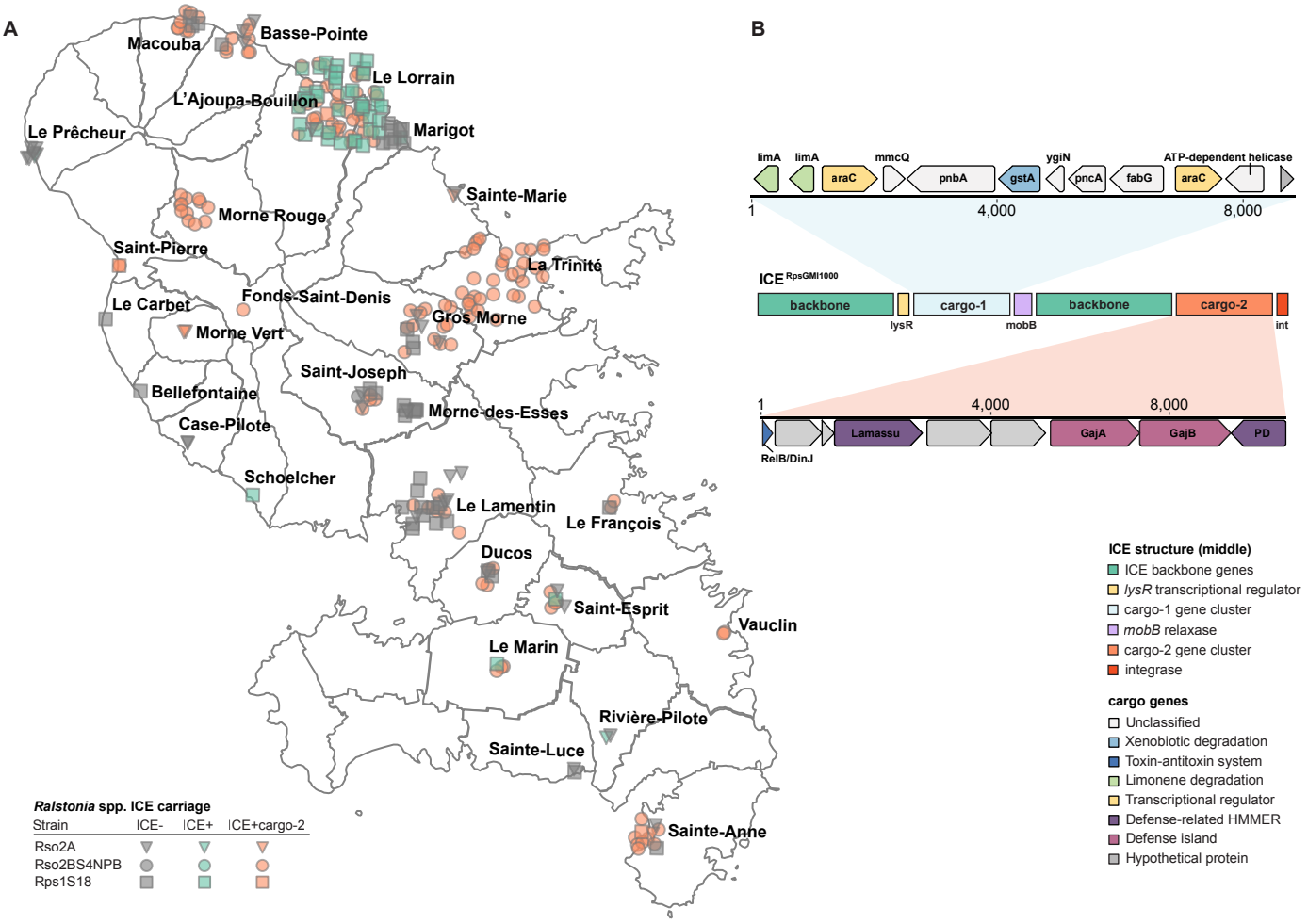


Fig. 5. Ralstonia ICE distribution and cargo gene carriage in Martinique. (A) Sampling location of all Martinique isolates sequenced in this study. The exact longitude and latitude of the sampling locations were jittered for the visualisation purpose. Rso2A (triangle), Rso2BS4NPB (circle), Rps1S18 (square) are represented with as grey (ICEless), green (ICERpsGMI1000 with cargo-1) or orange (ICERpsGMI1000 with cargo-1 and cargo-2) icons. (B). A diagram of ICERpsGMI1000 structure is shown in the center displaying 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 their KEGG annotation. Proteins assigned to a KEGG Ontology term not classified in a metabolic pathway are displayed in light blue. 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, rdfS* as well as a gene encoding a helix-turn-helix domain-containing protein, a hypothetical protein and various DUF domain-containing proteins (DUF3800, DUF945, DUF2285 (fseA) [139], DUF2840 and DUF2958). The second cargo-2 is flanked by an integrase at the 3' end of the ICE.

318 84 genes were uniquely acquired in the ancestor of Rso2BS4 and Rso2BS4NPB (Table S3). The
319 acquired genes include enzymes such as a TauD/TfdA family dioxygenase, which has been shown to
320 degrade phenoxy herbicides in soil-borne bacteria like *Burkholderia* and *Ralstonia eutropha* [90],
321 [91], [92], [93], [94], [95]; a CxxC motif protein (UW163_01710, OFBKGH_14170) associated with
322 thiol-disulfide oxidoreductase activity and oxidative damage responses [96], [97]; an electron
323 acceptor flavodoxin (UW163_18870, OFBKGH_23185) also involved in oxidative damage response
324 [98]; and a carboxymuconolactone decarboxylase (UW163_18885, OFBKGH_23170) in the
325 protocatechuate catabolism pathway, linked with conversion of lignin-derived aromatic compounds
326 into tricarboxylic acid intermediates like succinate and malate [99], [100]. The last two genes are
327 located within 2.25kb on the secondary chromosome of Rso2BS4 UW163. These metabolism-
328 associated genes are involved in distinct reactions and pathways, but are variously involved in
329 responses to oxidative stress and aromatic compound catabolism.

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

331 The identification of an ICE shared between Rsp1S18 and Rso2BS4NPB along with multiple
332 acquisition events from unknown donors suggests multiple mechanisms of recombination are
333 introducing selectable variation into the emergent lineage. Since prophage are known to introduce
334 accessory genes and modify bacterial host fitness, we queried the collection using the PHASTEST
335 phage prediction tool and found evidence of multiple insertions in the ancestor of the emergent
336 lineage. Strikingly, one of the three prophage identified is uniquely present in Rso2BS4NPB and
337 absent from all other sequenced RSSC genomes (Table S4, Table S7). 63 ortholog groups identified
338 as unique to Rso2BS4NPB in Panaroo pangenome analyses are located within the prophage located in
339 chromosome 1 (2,599,979 - 2,636,445 in Rso2BS4NPB RUN1145) as well as a 18kb region directly
340 downstream. The closest BLASTn hits include are soil isolates (*Ralstonia* metagenome-assembled
341 genome from estuary sediment in China, *Pandoraea pnomenusa* from sludge in Malaysia and
342 *Massilia putida* from soil in China), as well as *Xanthomonas albilineans*, a bacterial plant pathogen
343 causing lethal leaf scald in sugarcane from Brazil (Fig. S11). In addition to the core genes encoding
344 the phage head, tail, baseplate and restriction endonucleases, the Rso2BS4NPB-specific phage carries
345 auxiliary genes whose products are predicted to function as a circadian clock protein (*kaiA*,
346 OFBKGH_12090 in RUN1145) and a methyl-accepting chemotaxis protein (OFBKGH_12085). A
347 17.7Kb region downstream of the phage that is also unique to Rso2B4NPB contains a flagellar motor
348 switch encoding gene *fliM* (OFBKGH_12305). The strong association between this prophage and
349 genomic changes in the emergent lineage was additionally confirmed by Pyseer kmer-based
350 association test, as 1,007 out of 2,367 kmers associated with isolation from Martinique in
351 Rso2BS4NPB RUN1145 genome are mapped to the prophage and its downstream auxiliary region
352 (Fig. 4, Table S9).
353
354 The other two intact prophages are also located on the primary chromosome, one of these is only 60kb
355 downstream of the Rso2BS4NPB-specific prophage (Table S7). These additional prophage are not
356 unique to Rso2BS4NPB, BLASTn hits recover Rso strains from Central and South America. Further
357 prophage prediction in all 12 complete Rso2B genomes shows all Rso2BS4NPB (RUN0777,
358 RUN0823, RUN0808, RUN1145 and basal RUN2004) have three intact and one partial prophage
359 insertion on the primary chromosome, while three of four Rso2BS4 (UW163, Po82 and IBSBF1503)
360 have at most a single intact prophage on the primary chromosome and a single prophage on the
361 secondary chromosome. The exception is Rso2BS4 CIAT_078 (*Musa acuminata x balbisiana* AAB,
362 Colombia, 2004), which has no intact prophage predictions on the primary chromosome.
363

364 Alignment of complete assemblies for Rso2BS4 UW163 (symptomatic *Musa acuminata* x *balbisiana*
365 AAB, Peru, 1967) and Rso2BS4NPB RUN1145 (symptomatic *Cucurbita moschata*, Martinique,
366 2002) confirms horizontal gene transfer mediated by the movement of prophage and other mobile
367 genetic elements has had a significant impact on Rso2BS4NPB genome evolution. Despite being
368 isolated 45 years apart from different hosts and regions, these genomes are highly conserved across
369 the first (Fig. 4) and second chromosome (Fig. S9). The 4,989 Mb core genome alignment of
370 Rso2BS4NPB varies by only 2,894 SNPs (Fig. S8). The most striking difference between the two
371 lineages are multiple large insertions in Rso2BS4NPB, most of which are predicted to be prophage
372 (Fig. 4). Additionally, according to RECOPHY, the region with the highest SNP density in the
373 Rso2BS4 core genome alignment (1,060 out of 2,894 SNPs) is located in the region surrounding two
374 prophages and an adjacent *fhaC-fhaB* locus (Rso2BS4NPB RUN1145 chr1: 227,422-304,634). This
375 region also contains a high density of Martinique-associated kmers in RUN1145 and *Musa*-associated
376 kmers in UW163. Cumulatively, the data suggests there were multiple phage infections in the
377 ancestor of the emergent Rso2BS4NPB lineage, either prior or soon after its dissemination to
378 Martinique, introducing multiple novel genes that may be linked with apparent host range expansion.

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

380 We found repeated evidence of recent mobile element acquisition in the host-range expanded lineage
381 Rso2BS4NPB, attesting to a high frequency of encounter and exchange with members of the soil,
382 freshwater and plant microbiomes. Not all encounters are beneficial: members of the RSSC carry type
383 I-E CRISPR-Cas systems conferring immunity to bacteriophage via acquisition of spacer sequences
384 from prior infections. We found that Rso2BS4 UW163 and Rso2BS4NPB RUN1145 share conserved
385 CRISPR-Cas operons (*cas3*, *casA/cas8e*, *casB/cas11/cse2*, *cas7e*, *cas5e*, *cas6e*, *cas1e*, *cas2*), but vary
386 in their CRISPR arrays. The Rso2BS4NPB CRISPR 1 and CRISPR 2 arrays are nearly ten times
387 longer than the Rso2BS4 arrays: Rso2BS4 UW163 has 20 spacers compared to the 171 spacers in
388 Rso2BS4NPB RUN1145 (of which 11 are duplicates with 1-5bp differences). These genomes share
389 one leader-proximal and 15 leader-distal spacers. Our kmer-based association tests also show that the
390 largest cluster of Martinique-associated kmers that do not map to MGEs map to the leader-proximal
391 position of CRISPR arrays in Rso2BS4NPB, corresponding to more recent spacer acquisition than the
392 leader-distal spacers shared with Rso2BS4 (Fig. 4, Fig. S12) [101] [102]. This represents a rapid
393 expansion of CRISPR-mediated immunity in the emergent lineage, suggesting bacteria-virus
394 interactions played an important role during the expansion of Rso2BS4NPB across Martinique.
395
396 The identity of the viruses challenging Rso2BS4NPB was inferred by using spacer sequences as
397 queries for INPHARED phage database searches. A total of nine phage were identified using this
398 approach (Table S8). Three phage are targeted by spacers shared between Rso2BS4 and
399 Rso2BS4NPB CRISPR systems, and six are exclusively targeted by Rso2BS4NPB. Interestingly, five
400 of the six phage targeted by Rso2BS4NPB are known to infect Rps1 strains (RPZH3, RS603,
401 RSMSuper, Eline, Raharianne), revealing Rso2BS4NPB are encountering Rps phage, consistent with
402 their frequent co-occurrence.

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

404 We took advantage of the extensive metadata recorded by Wicker and colleagues during sampling to
405 run genome-wide tests for associations between genetic variation (either genes or kmers), host and
406 location of isolation. Many kmers (16,361) were linked with specific hosts, most frequently *Musa* spp.
407 (6,142 kmers), *Heliconia* spp. (4,560 kmers), *Capsicum annuum* (2,554 kmers) and members of the
408 Alismatales (*Anthurium* spp., *Epipremnum* spp., and *Pothos* spp., 2,369 cumulative kmers) (Table

409 S4). Mapping kmers against Rso2BS4NPB RUN1145 showed 25 *Musa*-associated kmers map to
410 seven loci including a region 5bp upstream of a gene involved in the degradation of aromatic
411 compounds into succinate and malate via the homoprotocatechuate pathway (5-carboxymethyl-2-
412 hydroxymuconate Delta-isomerase, OFBKGH_04945 in RUN1145) and a glycine
413 hydroxymethyltransferase (Table S10. A large fraction of *Musa*-associated kmers map to insertion
414 sequences, two prophages, multiple intergenic regions, hypothetical protein-encoding genes and one
415 of three *cheW* genes encoding chemotaxis signal receptor proteins. Interestingly, the histidine kinase
416 protein CheA, which is known to directly interact with CheW, is one of the most diverse proteins
417 across both Rso2BS4NPB and Rps1S18 according to our pangenome analysis and inspection of gene
418 alignments. 52.6% (2,316/4,401) of kmers mapping to the Rso2BS4 UW163 reference genome are
419 associated with isolation from banana. In comparison, only two banana-associated kmers map to the
420 RpsS18 GMI1000 genome.
421
422 Gene-based panGWAS methods (SCOARY and Pyseer) were also applied to search for genes whose
423 presence or absence is associated with a particular host of isolation. Pyseer reported a total of 938
424 genes significantly associated with any region or host of isolation traits (Table S11). Among these,
425 multiple genes associated with isolation from *Musa* spp. (135), *Solanum melongena* (80), *Capsicum*
426 *annuum* (56) and *Heliconia* spp. (21) were identified. There was no overlap between the Pyseer and
427 SCOARY predictions for host of isolation, though there were for association with specific locations,
428 at varying spatial scales (Table S12). As mentioned, the majority of location-specific genes identified
429 by both methods are located on the $ICE^{RsoGMI1000}$. Outside mobile elements, the most common genes
430 with location-specific associations are *fhaC* and *fhaB* (Fig. 4). These encode type 5b two partner
431 secretion systems comprised of a transporter (FhaC) and extracellular hemagglutinin repeat-
432 containing protein (FhaB). Interestingly, kmers associated with isolation from two distinct regions can
433 map to different regions of the same locus. For example, four out of seven *fhaC-fhaB* loci in Rps1S18
434 GMI1000 have China-associated kmers mapping to the 5' end of *fhaB* and Martinique-associated
435 kmers mapping to the 3' end of the same gene. The 3' end of another *fhaB* gene is composed of
436 alternating regions containing either China or Martinique-associated kmers mapping. Fine-scale
437 spatial associations with *fhaC-fhaB* are also present, for example between an *fhaB* and the region of
438 Le Lorrain in Martinique. Overlap with predictions for host association occurs as well: kmers
439 associated with isolation from both *Musa spp.* and Peru map to four of seven *fhaC-fhaB* loci from
440 Rso2BS4 UW163.

441 **Discussion**

442 This work reveals the evolutionary history of a globally distributed pathogen and highlights particular
443 genomic changes linked with a remarkable host range expansion event. We found that an outbreak of
444 disease affecting a broad range of plant species was caused by the dissemination and clonal expansion
445 of two distantly related lineages: *R. pseudosolanacearum* Rps1S18 and *R. solanacearum*
446 Rso2BS4NPB. The Rps1S18 lineage retains a signature of host specialisation, while Rso2BS4NPB
447 displays the broadest host range yet detected in the species complex. Strikingly, despite the link
448 between T3E carriage and host specific interactions in *Ralstonia* spp. and other bacterial pathogens,
449 we found little evidence that diversification in T3E repertoire is driving the apparent host range
450 expansion of Rso2BS4NPB. There is very little variation in effector carriage between the
451 Rso2BS4NPB and its sister lineage Rso2BS4, which infects banana. Only one hypothetical effector
452 (*T3E_hyp3*) was acquired and two (*ripAX2* and *ripP3*) were lost. We do not exclude the possibility
453 that changes in expression and allelic variation among T3Es may be responsible for altered host
454 range. For instance, *ripS3* is present in both Rso2BS4 and Rso2BS4NPB, but displays considerable

455 sequence polymorphism compared to all other effectors present across both lineages. These
456 differences may be associated with the altered pathogenicity of Rso2BS4NPB, but appear insufficient
457 to account for the apparently dramatic expansion in pathogen host range, particularly since single
458 effector deletions have limited consequences for infection in the absence of recognition. We argue the
459 causes of apparent host range expansion in Rso2B4NPB are not effector-mediated disruption of
460 recognition and immunity across a broad range of hosts, but more likely due to enhanced fitness due
461 to the acquisition of gene clusters conferring growth on novel carbon sources.
462
463 Bacterial metabolism and manipulation of nutrient availability in the host is increasingly recognised
464 as an important component of pathogenicity, in mammalian and plant pathogens alike [103], [104],
465 [105], [106]. For instance, genes whose products facilitate the use of xylem sap as a carbon source
466 [107] and horizontally acquired components of the denitrification pathway [108] have been shown to
467 be important for xylem colonisation and virulence in *Ralstonia*. Transcriptome sequencing of
468 experimentally evolved Rps1S18 GMI1000 showed fatty acid metabolism is upregulated in evolved
469 clones of Rps1S18 GMI1000, highlighting the importance of metabolism during adaptation to a novel
470 host [109]. Carbohydrate metabolism and tricarboxylic acid-related pathways were downregulated in
471 Rso2BS1 UW551 during the late stage of xylem infection [110]. We postulate the acquisition of novel
472 pathways for substrate catabolism resulted in enhanced bacterial growth either in the soil or within
473 plant tissues, resulting in a sudden and explosive increase in symptom production in the field.
474
475 The acquisition of an ICE with novel metabolic enzymes may have been pivotal to the host range
476 expansion of Rso2BS4NPB. ICEs facilitate rapid horizontal gene transfer of entire functional genetic
477 modules, allowing rapid adaptation to novel hosts and environmental niches [111]. All Rso2BS4NPB
478 strains in Martinique carry an ICE with a cargo-1 gene cluster encoding multiple enzymes and
479 regulators. Many of these are linked with the degradation and detoxification of environmental
480 pollutants and reactive xenobiotics. The functional significance of these acquisitions for
481 Rso2BS4NPB virulence is unclear, although there are striking links with degradation of both plant-
482 derived and organic compounds present in the soil. The limonene 1,2-epoxide hydrolases *limA* may
483 act on plant-derived epoxides, which are induced during pathogen attack and methyl jasmonate
484 application [112]. The phthalate ester hydrolase (*pnbA*) is linked with the degradation of phthalate
485 esters, common synthetic organic compounds that are known to be environmental pollutants [113].
486 Homologs of *gstA* and *fabG* are known to degrade organochlorine compounds used as primary active
487 ingredients (1-chloro-2,4-dinitrobenzene) or intermediates (ethyl 4-chloroacetoacetate) in the
488 production of herbicides and pesticides [114], [115]. Strikingly, two genes in cargo-1 are linked with
489 the manipulation of host-derived metabolites: a homolog of *ygiN* (ABM) hydroxylates jasmonic acid
490 in rice, attenuating plant defences during infection [116]. Finally, the pyrazinamidase/nicotinamidase
491 (*pncA*) is known to convert nicotinamide to nicotinic acid and is linked with xylem colonisation of
492 grape by *Xylella fastidiosa*, where nicotinic acid accumulation is observed during infection [117].
493 *Ralstonia* spp. degradation of plant metabolites has been observed in Rso2BS1 UW551, which
494 degrades phenolic compounds in xylem sap [118], and many *Ralstonia* spp. genomes contain salicylic
495 acid degradation pathways [119].
496
497 Recombination events introducing metabolic enzymes occurred beyond the ICE as well, in the
498 ancestor of Rso2BS4 and Rso2BS4NPB. These are similarly linked with protection against
499 environmental toxicity or oxidative stress, like the acquisition of a TauD/TfdA family dioxygenase
500 that may degrade phenoxy herbicides. The ability to metabolise toxic or antimicrobial compounds is
501 not unusual in *Ralstonia*: *R. pickettii* is known to degrade DDT pesticides [120] and *R. eutropha* is
502 able to degrade chloroaromatic compounds [121]. The ancestral acquisition of a

503 carboxymuconolactone decarboxylase is also of considerable interest, since the protocatechuate
504 catabolism pathway is responsible for the conversion of lignin-derived aromatic compounds into
505 amino acids and organic acids, fueling *Ralstonia*'s growth in plant tissues.
506
507 There are likely additional genomic changes associated with Rso2BS4NPB's adaptation to multi-host
508 environments. However, we found fewer genomic variants linked to a particular host of isolation than
509 to spatial sampling location. Rso2B4NPB moreover shows reduced pairwise diversity among core and
510 accessory genes, and a smaller set of accessory genes compared to Rps1S18 in Martinique, indicating
511 the host range expansion is not due to higher rates of exchange in the flexible genome. Although we
512 cannot exclude the possibility favourable climatic conditions and other biotic or abiotic variables may
513 have enhanced opportunities for infection to occur, the introduction and expansion of another lineage
514 in Martinique provides additional evidence Rso2BS4NPB underwent a rare host range expansion
515 event: Rps1S18 retains a clear signature of host specialisation and was mainly isolated from *Solanum*
516 spp.
517
518 The coexistence of both Rso2BS4NPB and Rps1S18 in Martinique and likelihood of coinfection is
519 reflected by the presence of the $ICE^{RpsGMI1000}$ in both lineages. The Rso2BS4NPB $ICE^{RsoGMI1000}$ shares
520 a similar backbone and identical cargo-1 cluster with the Rps1S18 $ICE^{RsoGMI1000}$, differing only in the
521 carriage of distinct cargo-2 clusters [66]. Examination of the spatial patterns of pathogen diversity and
522 mobile element distribution show the ICE is present in only three Rps1S18 and all Rso2B4NPB in
523 Martinique (Fig. 5), as well as some basal members of the Rso2B4NPB lineage in French Guiana.
524 With the exception of a single Martinique Rso2A isolate also carrying the ICE, the other *R.*
525 *solanacearum* sequenced to date are otherwise curiously ICE-free, despite chromosomal homology at
526 preferred ICE integration sites and an abundance of direct repeats targeted by the $ICE^{RsoGMI1000}$
527 integrase. In contrast, we identified multiple additional ICEs circulating among *R.*
528 *pseudosolanacearum*. While we cannot explain the pattern of species-specific variation in ICE
529 carriage, we speculate ICE acquisition by Rso2B4NPB likely occurred prior to or soon after
530 Rso2BS4NPB dissemination to Martinique, perhaps via coinfection with Rps1S18 or another *R.*
531 *pseudosolanacearum*. It is not possible to reliably infer the source and directionality of transfer events
532 for mobile elements like ICEs. Rps1S18 in Martinique may have acquired cargo-1 via recombination
533 with the Rso2B4NPB ICEs for instance, and Rso2BS4NPB may have acquired the $ICE^{RsoGMI1000}$ from
534 an unsampled lineage of *Ralstonia* sp., or a less related member of the microbial community
535 altogether. The distribution of ICEs among the *Ralstonia* spp. sampled across Martinique nevertheless
536 provide some insight into how intermicrobial interactions resulting in horizontal gene transfer shape
537 pathogen evolution, and may be linked with the emergence of epidemic lineages.
538
539 We found additional genomic signatures of host-parasite interactions in Rso2BS4NPB, which exhibits
540 evidence of frequent encounter with its own potential parasites. The expansion of CRISPR repeats and
541 acquisition of multiple fragmented and intact prophages in Rso2BS4NPB genomes provide clear
542 evidence of exposure to phage infection. Acquired immunity provides an additional advantage to
543 Rso2B4NPB in the phage-rich soil environment [122]. Some of the CRISPR spacers with identifiable
544 origins correspond to phage known to successfully infect *R. pseudosolanacearum* strains like
545 Rps1S18. Coinfections of the same plant host or co-occurrence in the same environmental niche may
546 therefore represent an avenue for Rso2BS4NPB exposure to Rps1S18 phage. Rso2BS4NPB also
547 acquired multiple prophage, including one which has not yet been identified in any other sequenced
548 member of the species complex. The acquisition of prophage is frequently accompanied by the
549 acquisition of accessory gene clusters, which may also encode novel functions altering microbial
550 lifestyle and pathogenicity [123].

Resolving the origins of disease outbreaks in crop plants remains challenging, as the establishment of sampling programs typically occurs years after pathogens emerge and spread in susceptible hosts. Sampling is then typically limited to symptomatic hosts of interest, excluding adjacent crops and uncultivated plants and substrates, and sequencing in turn may be sporadic, dependent on available resources and how it may contribute to plant breeding efforts. These constraints limit our ability to identify disease reservoirs and infer patterns of dissemination in agricultural crops. This unique collection of *Ralstonia* spp. sampled across Martinique and French Guiana provides insight into the genomic changes associated with the emergence and host range expansion of Rso2BS4NPB, the microbial interactions shaping bacterial evolution, and patterns of dissemination between continental and island pathogen populations. The data shows Rso2BS4NPB did not emerge in Martinique, but is derived from an ancestral population still present on the mainland, from which the Rso2BS4NPB and banana infecting Rso2BS4 lineages diverged. The Rso2BS4 isolates from Peru and Colombia form a sister lineage that is associated with banana infection. The isolation of Rso2BS4NPB from fields in French Guiana that were only cleared of forest for cultivation one season prior additionally suggests Rso2BS4NPB became established prior to the outbreak in Martinique. The basal French Guiana Rso2BS4NPB isolates were predominantly isolated from *Solanaceae* however. The genomic changes associated with host range expansion in Rso2BS4NPB may have occurred immediately prior or subsequent to dissemination to Martinique, since a similar epidemic across a broad range of hosts was not identified in French Guiana.

The emergence of a pandemic lineage infecting crop plants is frequently linked with the introduction and expanded cultivation of susceptible host varieties [124], [125], [126], [77]. Banana was introduced as a crop to the Americas five centuries ago, and intensive industrial cultivation only began in the last two centuries [127]. Infectious disease outbreaks occurred soon after, most notoriously those caused by *Fusarium oxysporum* TR1 and *F. odoratissimum* TR4, which previously eliminated and currently threaten the Gros Michel and Cavendish varieties, respectively [128] [129]. Banana pathogenicity evolved independently multiple times in *R. solanacearum*, which has a center of origin in the Americas. Rso2BS4NPB is closely related to Rso2B4 strains responsible for outbreaks in banana plantations in Peru and Colombia, but exhibits reduced pathogenicity on this host. The convergent evolution of banana pathogenicity across the species complex aligns with observations of Rso2BS4NPB's lack of pathogenicity on banana and suggestion pathogenicity on this host is a derived trait in Rso2BS4 [2], [3], [130]. Convergent trait evolution and high frequency of recombination across the species complex enabled the identification of multiple genes and loci linked with isolation from banana. These represent interesting candidates for study of the genetic basis of specialisation on this host, along with the acquisitions we hypothesize are responsible for increased fitness of Rso2BS4NPB in environmental and plant-associated niches, and the consequent outbreak of this lineage across a wide variety of plant hosts. Cumulatively, this work provides insight into the origins of disease emergence caused by *Ralstonia* spp., and reveals how ecological interactions with co-occurring bacterial species and bacteriophage alike can shape pathogen evolution and contribute to the emergence of infectious disease.

592 **Methods**

593 **Genome sequencing**

594 *Ralstonia* spp. were collected in Martinique and French Guiana as described in Wicker *et al*. 2009 [3]
595 and Deberdt *et al*. 2014 [4]. Isolates collected after the initial sampling were assigned to phylotype
596 and sequevar by sequencing endoglucanase (*egl*) and mutation mismatch repair (*mutS*) genes. For this
597 study, we extracted genomic DNA from 487 strains using Promega Wizard, then used a modified
598 Nextera protocol for library preparation, as described in [5]. Paired-end sequencing with Illumina
599 NextSeq 2000 was then performed. Since the first round of sequencing produced under 25-fold
600 average coverage and under 40-fold maximum coverage, a second round of library preparation and
601 sequencing was performed. Eight isolates (RUN0777, RUN0808, RUN0820, RUN0823, RUN1145,
602 RUN1155, RUN1952 and RUN2004) were additionally sequenced using Oxford Nanopore
603 Technology with a FLO-MIN106 flow cell and SQK-RBK004 barcoding kit. Basecalling and
604 demultiplexing was performed with guppy_basecaller (v. 5.0.7) and guppy_barcoder v. 4.5.4 [6].

605 **Genome assembly and annotation**

606 Paired-end 150bp reads of both Illumina sequencing runs were filtered and trimmed with fastq_screen
607 [7], bbduk [8] and fastp [9], [10] (--dedup --detect_adapter_for_pe --cut_front 15 --length_required
608 50). Assembly was performed with SPAdes [11], [12] using both separate and pooled (i.e. both
609 sequencing runs) read sets. QUAST was used for assembly quality assessment [13]. Pooling both
610 sequencing runs improved average assembly N50 values from 73Kb to 195Kb and decreased average
611 contigs per assembly from 370 to 217, so assemblies generated from pooled sequencing runs were
612 used. After filtering out assemblies with under 50-fold coverage, over 250 contigs, total assembly
613 length exceeding 6Mb and N50 exceeding 75Kb, a total of 409 assemblies were retained for
614 downstream analyses.
615
616 Oxford Nanopore long read data was filtered with Filtlong (v. 0.2.1) [14] to remove reads under 1Kb
617 and the 5% of reads with the lowest quality values, then assembled with Trycycler (v. 0.5.4) [15],
618 using Flye [16], Miniasm [17], Minipolish [18] and Raven [19] as underlying assemblers. Contigs that
619 did not pass the reconciling step of Trycycler pipeline were manually removed. Final Trycycler
620 consensus sequences were polished three times using long reads with Medaka [20], followed by two
621 rounds of short read polishing using BWA [21] and Polypolish [22], using separate batches of the
622 Illumina reads produced by each sequencing run.
623
624 All assemblies were annotated by Bakta [23], [24] using Prodigal [25] gene models and a reference
625 protein database built using all complete *Ralstonia* spp. genome assemblies available from the RefSeq
626 database in October 2022.

627 **NCBI datasets and reference genomes**

628 All available *R. pseudosolanacearum, R. syzygii* and *R. solanacearum* genome assemblies were
629 downloaded from NCBI in February 2023 to provide additional phylogenetic context for analyses.
630 Assemblies with fewer than 200 contigs, total assembly length between 5.0 - 6.2Mb, N50 over 75Kb
631 and unflagged by the NCBI Genome datahub [26] for evidence of potential contamination were
632 retained. A total of 410 Ralstonia spp. genomes passed these filters: 370 *R. solanacearum* (Rso), 35 *R.*
633 *pseudosolanacearum* (Rps) and 5 *R. syzygii* (Rsy) genomes (Table S1). Three isolates were present in
634 duplicate in the NCBI filtered dataset (MAFF 211479, UW163 and UW551), the lowest quality
635 duplicate assembly was removed, producing a final NCBI dataset of 407 assemblies. These were then

636 reannotated with Bakta and merged with the 409 assemblies produced in this work for downstream
637 pangenome analyses. Metadata (host, geographical region and year of isolation) for genomes in the
638 NCBI dataset was obtained by inspecting Biosample entries. If the information was missing in
639 Biosample entries, we used RSSCdb records [27], [28] or publications describing the initial isolation
640 of the strain of interest [29].

641 **Type 3 secreted effector identification**

642 Amino acid sequences of 118 *R. solanacearum* type 3 secreted effectors (T3Es) from the manually
643 curated RalstoT3E database [30] were downloaded and used as query sequences for DIAMOND
644 BLASTp [31] searches against proteomes of the Bakta-annotated assemblies, using the --
645 ultrasensitive mode. Hits were filtered using a S30L25 strategy: retaining sequences with alignment
646 identity of at least 30% and query sequence coverage of 25% or greater. Since the T3E query set
647 includes multiple alleles of the same effector families, hits were then grouped into T3E families by
648 sorting first by identity and query length coverage. T3E family assignment for each protein sequence
649 was performed using the most significant T3E hit. Effector presence and absence was visualised by a
650 heatmap in R, omitting duplication events. This approach was tested by comparing the predicted T3E
651 repertoires in Rps1S18-GMI1000 and Rsy-RPSI07 with the ones provided in the RalstoT3E database.
652 Average pairwise similarity of T3E repertoires within RSSC clades was estimated by Scikit-learn
653 function cosine_similarity [32] using the information about presence-absence of effector families as
654 an input binary matrix.

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

656 Panaroo [33] was used to identify core and flexible genomes for the RSSC dataset, i.e. the merged
657 dataset of 816 *Ralstonia* genomes sequenced in this study (409 strains) and downloaded from NCBI
658 (407 strains). A concatenation of core genes was used as an input alignment for phylogenetic analysis
659 of the species complex. Alignment columns with gaps or ambiguities and invariant sites were
660 removed with SNPsites [34], producing a final alignment of 137,423bp. Tree inference was then
661 performed using IQtree [35] with a GTR+ASC model of sequence evolution, model for alignments
662 that do not contain constant sites, and 1000 bootstrap replicates. The ggtree R package [36] was used
663 for all tree and metadata visualisation.
664
665 Panaroo gene presence and absence tables for the RSSC dataset were queried using a custom Python
666 script to identify genes uniquely present in either Rps1S18, RsoS4 or Rso2BS4NPB. Genes are
667 referred to as clade-specific when they are present in at least 95% of all genomes in a given focal
668 clade, and present in no more than 10 non-focal clade assemblies.
669
670 The species complex tree from the set of core genes identified by Panaroo was used for clade
671 delineation and lineages assignment. Most genomes clustered according to the phylotypes initially
672 assigned by *egl* and *mutS* sequencing in Wicker *et al.* 2007 [3] and Deberdt *et al.* 2014 [4], although
673 some discrepancies are present. We simplify strain species, phylotype and sequevar references here
674 using the following format: [species][phylotype][sequevar]-[strain ID]. For instance, *R. solanacearum*
675 phylotype IIB sequevar 4 strains are referred to as Rso2BS4, where the 'S' precedes prior sequevar
676 assignment. Those strains that were not classified by sequevar have a shorter designation:
677 [species][phylotype]-[strain ID]. The clade designation in this work may differ from historically
678 assigned sequevar or even phylotype labels, since these were first assigned using *egl* and *mutS*
679 sequences rather than whole genome sequencing and core gene tree reconstruction. The criteria for

membership in a named clade is exceeding a minimum shared identity of 99% in the RSSC core gene alignment.

Separate pangenome analyses and core gene tree reconstructions were performed on three specific clades: *R. pseudosolanacearum* phylotype 1 sequevar 18 (Rps1S18, 102 strains), *R. solanacearum* phylotype 2 sequevar 4 (Rso2BS4, 226 strains), and the emergent lineage *R. solanacearum* phylotype 2 sequevar 4 Non-Pathogenic on Banana (Rso2BS4NPB, 180 strains). The strains forming each clade were identified with reference to the RSSC core gene tree (Figure 1). Assemblies for IBSBF_2571 (GCA_003590605.1) and O12BS_UW407 (GCA_023076615.1) were excluded due to abnormally long terminal branch lengths and P673 was excluded due to contradictory phylogenetic position of two assembly versions: the most recent P673 assembly GCA_015698365.1 was assigned to Rso2BS4NPB clade, while the earlier version GCA_000525615.1 grouped with other Rso2BS4 *Epipremnum* isolates from the USA.

An additional iteration of clade-specific pangenome analysis was performed for both Rso2BS4NPB and Rps1S18. Although strain IBSBF1503 (GCA_001587155.1) groups with other Rso2BS4NPB strains in the RSSC core gene tree, the clade-specific tree revealed this strain is a more distant relative and more properly referred to as an outgroup strain, therefore this was excluded from a second iteration. Similarly, Rps1S18 strains CIP266_UW505 (GCA_023075875.1) and UW393 (GCA_023076675.1) form a distant outgroup to other Rps1S18 strains, therefore these were excluded from the second iteration of the Rps1S18 pangenome analysis.

Mobile element prediction in reference genome assemblies

All potential mobile elements, including genes annotated as integrases, transposases, prophage-related genes and IS elements, were identified and masked for the downstream phylogenetic analysis in the Bakta-annotated genome assemblies for Rps1S18 strain GMI1000 (GCA_000009125.1), RsoS4 strain UW163 (GCA_001587135.1) and Rso2BS4NPB strain RUN1145. Prophage and IS elements were predicted using the PHASTER, PHASTEST [37], [38], [39] and ISfinder [40] online tools. All gene annotations containing keywords "phage" OR "transposase" OR "tail" OR "head" OR "integrase" OR "terminase" OR "conjugal" OR "integrase" OR "integrative conjugative element" OR "conjugative transfer" OR "conjugative coupling" were manually searched for and flagged in Geneious Prime [41]. The results were converted into BED files, then sorted, merged and used for masking with Bedtools [42]. Masked references were used for reference-based contig mapping.

Phylogenetic analyses

Reference-based contig mapping was performed for the Rps1S18, RsoS4 and Rso2BS4NPB clades, as defined by using RSSC and clade-specific core gene trees. Rps1S18 included 100 genomes, with GMI1000 as reference. The RsoS4 clade contained 226 genomes, with UW163 as reference. Rso2BS4NPB included 179 genomes, with RUN1145 as reference. The Brazilian *Cucumis sativus* isolate IBSBF1503 is referred to as Rso2B4NPB in the literature and initially grouped with Rso2BS4NPB in the species complex tree, and, so we initially used this strain as a reference for Rso2BS4NPB. While more closely related to Rso2B4NPB than any other available complete *Rso* genome, we found IBSBF1503 is clearly an outgroup relative to other Rso2BS4NPB isolates (Fig. S4). The average pairwise nucleotide diversity is 0.002% between Rso2BS4NPB genomes, and an order of magnitude greater (0.02%) relative to IBSBF1503. According to progressiveMauve whole genome alignments, Rso2BS4NPB isolate RUN1145 differs from RUN1960 and IBSBF1503 by 134

725 and 1,215 SNPs, respectively. Since IBSBF1503 is too divergent to be considered a member of the
726 same lineage, we reclassified IBSBF1503 as Rso2BS4. We resequenced five Rso2BS4NPB strains
727 from Martinique using Nanopore long-read sequencing to generate closed hybrid assemblies and used
728 the complete assembly of Rso2BS4NPB RUN1145 as a reference for more reliable core genome
729 alignment-based phylogenetic inference of relationships between closely related strains. The Rps1S18
730 contig mapping using Rps1S18 GMI1000 as reference resulted in 4,760,272 bp consensus alignment
731 with 311 SNPs and 116 parsimony informative sites.
732
733 Reads and contigs mappings were performed separately for each assembly with bwa-mem2 [21], [43]
734 or Minimap2 [44] using asm5 mode respectively. All resulting BAM alignments were sorted and
735 indexed with Samtools [45]. BCFtools mpileup was used to filter and retain bases with mapping
736 quality above 60 [45]. Variants were called with bcftools-call and filtered by bcftools-filter using the
737 expression TYPE="snp" && DP>=X && QUAL>30 && AF1>=0.95 with X=10 for reads and X=1
738 for contig mapping. Any alignment column with a gap site was removed. Consensus sequences were
739 obtained using bcftools-consensus, masking the sites with coverage depth below one as identified by
740 Mosdepth [46]. The consensus sequences for all strains in a given clade were concatenated into a
741 FASTA alignment file. Sites containing N characters were trimmed with goalign clean-sites [47].
742 Recombinant regions were predicted and removed by ClonalFrameML [48] using an initial tree
743 calculated by IQtree [35] under GTR+F+I+R2 substitution model. The final Maximum likelihood
744 trees were calculated from the non-recombinant alignments for each clade by IQtree using the MFP
745 (ModelFinder Plus) option for the best model search [49]. The contig mapping approach resulted in
746 better resolved and bootstrap supported trees, therefore the trees generated by reads mapping were
747 excluded from further analysis. Tree visualisations were created with ggtree [36].

748 **Tree dating**

749 Root-to-tip regression and dating of the Rso2BS4NPB and Rps1S18 non-recombinant core genome
750 trees were performed with BactDating using 100,000 Markov Chain Monte Carlo iterations [50]. For
751 the Rso2BS4NPB tree, RsoS4 strain IBSBF1503 was added as an outgroup for tree calculation and
752 dating, as the initRoot function was otherwise unable to root the tree. TempEst [51] was used to
753 confirm root-to-tip regression results obtained with BactDating, setting the 'Best-fitting root' option
754 of TempEst to 'R squared'. Tree tips with missing sampling date information (Rps1S18
755 UQRS637_UW745 and UW193, Rso2BS4NPB RUN0283 and RUN0739) and outliers (Rps1S18
756 UW299, RUN0679, RUN0037 and GMI1000) identified by TempEst were dropped from the
757 Rps1S18 and Rso2BS4NPB trees using the R function ape::drop.tip. Branches with edge lengths of
758 zero were manually collapsed into a single node by Dendroscope [52]. A Dates Permutation Test was
759 performed with BactDating by randomly generating dates between 2000 and 2012 and randomly
760 assigning these to tree tips. The simulated trees were compared to the actual dating model using the
761 modelcompare function of BactDating.

762 **CRISPR-Cas system identification**

763 CRISPR-Cas systems were identified with Bakta annotation of Rso2BS4 and Rso2BS4NPB genomes.
764 CRISPR repeats and spacers were then annotated in two closed references Rso2BS4-UW163 and
765 Rso2BS4NPB-RUN1145 using the CRISPRloci workflow [53]. The spacer sequences identified with
766 CRISPRloci were then used for BLASTn searches against the RefSeq Phage database downloaded
767 from PhageScope [54], NCBI Nucleotide collection (nr_nt) and Environmental samples (env_nt)
768 databases accessed in June 2024. The extracted spacers were also used for INPHARED database [55]
769 searches using the SpacePHARER [56] toolkit in August 2024.

770 **Inter and intraspecific recombination prediction**

771 Potential homologous recombination events within and between Rso2BS4 and Rps1S18 lineages were
772 investigated using the RECOPHY toolkit [57]. Two datasets were analysed: Rso2BS4 alone and in
773 combination with Rps1S18, using Rso2BS4NPB-RUN1145 as a reference in each case. Prior to
774 RECOPHY execution, all genome pairs with a Mash distance [58] below zero were identified and one
775 member of each pair was then randomly removed from the dataset. The final datasets included 49
776 genomes for the Rso2BS4 analysis alone, and with an additional 64 Rps1S18 genomes for the merged
777 analysis. The regions of core alignment with increased and reduced density of SNPs per 1 kb block
778 were extracted and investigated separately. In order to find discrepancies with the core phylogeny, a
779 simple Neighbor-Joining tree was calculated for each of these regions using Geneious Prime software
780 [41].

781 **Pangenome-wide association testing**

782 Scoary [59] and Pyseer [60] pan-GWAS analyses were used to investigate associations between gene
783 presence and absence patterns and the host or location of strain isolation. The broad (country) and
784 fine-scale (village) locations and all taxonomic levels of the host of isolation or environmental
785 substrate were converted from the metadata table into a binary traits table using the
786 pandas.get_dummies function [61]. For example, a strain isolated from a tomato plant in China would
787 receive the value "1" in the columns China, *Eudicots, Asterids, Solanales, Solanaceae, Solanum* and
788 *S. lycopersicum*, and receive the value "0" in all other columns of the traits table. This traits table was
789 used as input for both Scoary and Pyseer pan-GWAS analyses, along with the gene presence and
790 absence table from Panaroo.
791
792 The significance of candidates identified with Scoary was assessed after multiple test corrections:
793 individual (naive) p-value, Bonferroni adjusted p-value, Benjamini-Hochberg adjusted p-value, entire
794 range of pairwise comparison p-values, and empirical p-value from 100 permutations, using
795 thresholds of p=0.05 for all corrections. Significant genes identified by Scoary after all the corrections
796 were further filtered using 95% sensitivity and 70% specificity thresholds. A lower specificity
797 threshold was chosen due to the nature of the trait dataset: the traits table simply reflects the known
798 host of isolation and the location of isolation. The actual host range of a particular strain may include
799 additional hosts, closely or distantly related, just as closely related strains may be present in vastly
800 different geographic regions. Since the Scoary and Pyseer pan-GWAS analyses aim to identify genes
801 linked with survival in a particular host, environment or substrate, the sensitivity should be high.
802 Those same genes may also be present in strains isolated from other plants however, therefore
803 specificity is reduced. For example: a gene whose product is strongly associated with tomato (*S.*
804 *lycopersicum*) infection should be overrepresented among tomato isolates in the dataset. Therefore we
805 seek genes present in 95% of tomato isolates. We reduced the specificity threshold to 70% to allow
806 for instances where another strain carries that same gene and is also capable of infecting tomato, but
807 happens to have been isolated from a different host.
808
809 For the Pyseer analysis, the distance matrix from the RSSC core gene tree and the number of patterns
810 for the likelihood ratio test (LRT) were calculated using phylogeny_distance.py and count_patterns.py
811 scripts downloaded from the Pyseer GitHub page [62], accessed on January 2024. The pattern count
812 resulted in 6,075 patterns for the RSSC, leading to the likelihood ratio test (LRT) p-value threshold of
813 3.8e-08 = 0.05/(patterns_number*traits_number). Hits with effect size (beta value) under 0.2 and
814 phenotypic variance values (variant_h2) above 0.5 were filtered out.
815

816 Pan-GWAS of gene presence/absence data is an effective tool for the identification of genes
817 associated with traits of interest, but is limited to the accessory genome and does not test for trait
818 associations within the core genome and intergenic regions. This gap is filled with a kmer-based
819 association test. Analogously to the method used for gene presence/absence data, Pyseer was applied
820 to identify the kmers significantly associated with any host or location of isolation across the RSSC,
821 using the same binary metadata table as described for the pan-GWAS analysis. The initial set of 31bp
822 long kmers was generated for all the RSSC genomes using unitig-caller [63]. The distance matrix
823 from the species complex core genes tree and the number of patterns for the likelihood ratio test
824 (LRT) were calculated using phylogeny_distance.py and count_patterns.py scripts from the Pyseer
825 GitHub page [62]. The pattern count reported 1,484,018 patterns for the RSSC kmers, producing a
826 LRT p-value threshold of 5.0e-10 = 0.05/(patterns_number*traits_number). Hits with effect size (beta
827 value) under 0.2 and phenotypic variance values (variant_h2) above 0.5 were filtered out. The
828 presence and positions of significant kmers in Rso2BS4NPB reference strain RUN1145, RsoS4
829 reference strain UW163 and Rps1S18 reference strain GMI1000 was identified and stored in BED
830 files using a custom Python script. For visualisation purposes, all kmers mapped on a genome within a
831 1kb distance were merged using the bedtools-merge function [42].
832
833 The panGWAS and kmer-based tests identify many genes associated with isolation from specific
834 hosts. These tests vary in how they control for population structure, but some genes were predicted to
835 have associations with particular traits using multiple methods. Separate association tests were
836 performed using either spatial location or host/substrate of isolation metadata. We noticed that many
837 genes have overlapping predictions using both metadata types, particularly where it concerns isolates
838 recovered from soil or water. For example, gene presence associated with *S. lycopersicum* isolation
839 often overlaps with gene presence associated with *Solanaceae* isolation. All overlaps were
840 investigated using a custom Python script and visualised in an Upset plot using the UpSetR R package
841 [64]. For the genes assigned to multiple nested traits, the lowest level of taxonomic identification (e.g.
842 species rather than genus) or the finest level of geographical scale (e.g. village rather than country) for
843 which significance was identified were chosen. Overlaps within the Scoary, Pyseer pan-GWAS and
844 kmer-based outputs were identified using the same strategy.

845 **Identification of integrative and conjugative elements**

846 Complete integrative and conjugative elements were identified in 133 complete RSSC genome
847 assemblies using the following criteria: the presence of a type 4 secretion system (T4SS) required for
848 conjugal transfer between strains, a *mob* family relaxase and integrase, and the presence of 8bp direct
849 repeats (TTTTTTAT) encompassing the predicted ICE [65]. This 8bp repeat is described as an ICE
850 insertion site for Rps1S18 GMI1000 in the ICEberg 3.0 database [66]. We note this repeat occurs
851 between 50-70 times elsewhere in each RSSC genome. This does not preclude them from being actual
852 ICE integration sites: despite the apparent existence of preferred ICE integration sites in the RSSC,
853 some ICE integrases are known to be promiscuous and may target different repeats as attachment sites
854 across the genome. For example, ICE Tn916 targets AT-rich loci and integrates almost at random
855 across genomes [67]. MOBscan was used for identification of relaxases [68], followed by BLASTn
856 sequence homology searches for a T4SS and adjacent integrases [69] using Rso2BS4NPB RUN1960
857 T4SS and integrase sequences as BLASTn queries, followed by T4SS locus confirmation with the
858 TXSScan module of MacSyFinder toolkit [70]. The flanking repeat sequence TTTTTTAT was
859 identified using a custom python script. In genomes where the repeat was not identified within 10kb
860 of the T4SS, relaxase and integrase, putative insertion sites were manually identified.
861

862 The genomic context of putative ICEs in the RSSC assemblies was inferred by aligning 50 kb
863 upstream and downstream flanking regions surrounding putative ICEs in Rps1S18 GMI1000,
864 Rso2BS4NPB RUN1145 and Rso2BS4NPB RUN1960 to the genomes. The presence of genes
865 significantly associated with isolation from Martinique by panGWAS analysis was also confirmed by
866 BLASTn, retaining hits with e-values under 1.0^{-5}, query length coverage over 10% and overall
867 alignment length exceeding 1 kb.
868
869 The involvement of ICE cargo genes in specific metabolic pathways was predicted using KEGG
870 Orthology assignment with protein sequences from the cargo regions of interest, performing searches
871 in the database for the genus Ralstonia (taxa ID: 48736). The locus structure of putative ICEs was
872 visualised with the ggtree::geom_motif R function [36]. Complete ICE sequences identified in *Rps*
873 362200, *Rps* RS24, *Rps* RUN1952, Rps1S18 RUN820, Rps1S18 GMI1000, Rso2BS4NPB RUN1960,
874 RUN0777, Rso2A RUN1155 and *Ralstonia pickettii* strain 12J were aligned to each other using the
875 progressive Mauve [71] algorithm implemented in Geneious [41].

876 **Geographical metadata display**

877 Information about the geographical region of isolation for each strain was obtained from the country
878 of isolation metadata collected for the RSSC strains set as described above. Counties of isolation were
879 converted into continents using pycountry Python package [72]. The "Caribbean" region of isolation
880 was manually assigned for all Martinique isolates to highlight the region of interest.
881
882 Longitude and latitude of the sampling site was recorded for 83/500 strains sampled in Martinique and
883 French Guiana. For the remaining 417/500 strains, the commune (administrative district) or the
884 nearest village was listed in the strain metadata. For these cases, approximate longitude and latitude
885 values were retrieved using the geolocator function of the geopy.geocoders Python package [73] using
886 the commune or nearest village names. Commune maps of Martinique and French Guiana in geojson
887 format were downloaded from the France Geojson collection [74]. The actual and estimated longitude
888 and latitude values for each isolate were visualised with the geopandas package [75].

889 ## Supplemental Results

890 **Phylogeography of Rso2BS4NPB and Rps1S18**

891 The Rso2BS4NPB tree shows the Gros Morne sampling site is adjacent to Trinite and also harbours
892 strains from multiple clades of Rso2BS4NPB, though all but two isolates (which were obtained from
893 lake water) were from the same host species endemic to the Americas: *A. andreanum*. Similarly, all
894 but one isolate from Sainte Anne group together, despite the fact they were isolated from dissimilar
895 hosts: tomato and cucumber (*Cucumis sativus*). There is a clade of Macouba village isolates from
896 *Heliconia* spp. and from *Musa* spp., though the latter are from plants with atypical symptoms.

897 **Identification of ICEs in the *Ralstonia solanacearum* species complex**

898 All eight $ICE^{RpsGMI1000}$ identified in Rps genomes have integrated into one of three attachment (att)
899 sites on the primary chromosome, numbered incrementally in order of distance from the origin of
900 replication in Rps1S18 GMI1000 (i.e. att-1, att-2 and att-3). Two Rps genomes have $ICE^{RpsGMI1000}$ in
901 att-1, four Rps genomes have $ICE^{RpsGMI1000}$ in att-2, and three genomes have $ICE^{RpsGMI1000}$ in att-3.
902 ICE attachment sites frequently overlap with tRNA genes, but this is not the case here, and
903 comparison of the flanking regions shows structural rearrangements have occurred [131]. The att-1
904 site appears least preferred for ICE integration, and is located between genes encoding a hypothetical

905 protein and a purine metabolism enzyme phosphoribosylformylglycinamidine cyclo-ligase. The att-2
906 is located between XRE family transcriptional regulator and inositol monophosphatase genes, while
907 att-3 is located between genes encoding a universal stress protein and isocitrate lyase. All five ICEs
908 identified in complete Rso assemblies are $ICE^{RpsGMI1000}$ integrated into att-2 sites in Martinique
909 isolates, although two Rso2BS4NPB draft assemblies (RUN1960 and RUN1959) contain an ICE
910 integrated into the att-3 site. The att-3 is the preferred integration site in Rps genomes for most
911 complete or partial ICEs, $ICE^{RpsGMI1000}$ or otherwise.
912
913 Rps ICEs retain core backbone genes (*trb* conjugative system, *traG* coupling protein, *mob* relaxase
914 and integrase) but vary considerably in cargo, i.e. gene clusters that do not encode functions critical to
915 the ICE lifecycle, but are highly variable between ICEs and may confer novel fitness-enhancing
916 functions to the bacterial host. Conserved and cargo regions in the broader family of ICEs were
917 delineated using ProgressiveMauve alignment of complete ICEs from five Rps genomes representing
918 five clades in the phylogenetic tree, three Rso (Rso2A and two Rso2BS4NPB with ICEs in either att-2
919 or att-3) and one outgroup *R. picketii* genome (Fig. 5). The cargo-1 in Rps1S18 RUN0820 and Rps
920 RUN1155 ICEs encodes genes associated with carbohydrate metabolism and folate biosynthesis
921 [132], and Rps RUN0820 ICE cargo-1 encodes various reductases, oxidases and dehydrogenases
922 involved in the metabolism of fucose, a sugar found on plant cell surface [133] (gene ID
923 PEIJFM_11570).

924 **Virulence gene repertoires in the *Ralstonia solanacearum* species complex**

925 Patterns of T3E family presence and absence correlate with genetic distance, i.e. strains within the
926 same species or the same clade share similar T3E repertoires. The average pairwise similarity within
927 T3E families is higher within Rso (92%) and Rps (94%) than it is across the entire species complex
928 (86%). *R. syzygii* (Rsy) exhibits the highest diversity among T3E families (88%), despite being less
929 well sampled than both *Rso* and *Rps*. The core effector arsenal of the species complex is comprised of
930 31 T3E families present in more than 99% of all RSSC genomes: *ripA3, ripA4, ripA5, ripB, ripC1,*
931 *ripD, ripE1, ripG4, ripG7, ripH2, ripM, ripR, ripS2, ripS5, ripU, ripW, ripY, ripZ, ripAB, ripAC,*
932 *ripAD, ripAE, ripAI, ripAJ, ripAM, ripAN, ripAO, ripAP, ripAR, ripAY, ripTPS*. Some T3Es were
933 found in only a single RSSC genome: RipO2 is present solely within *Rsy* strain NCPPB_3219 from
934 clove (*Syzygium aromaticum*), RipBN is limited to Rps3S29 strain CMR15 from tomato (*S.*
935 *lycopersicum*) and the hypothetical effector Hyp17 is present solely in Rps strain Tb04 from tobacco
936 (*Nicotiana tabacum*) (Fig. S1). Five T3E families were not identified in any of the RSSC genomes
937 despite being listed in the RalstoT3 database (RipG8, RipAF2, RipBE, RipBP and RipBQ). This may
938 be due to filtering and assignment to a T3E family based on similarity of the best hit, or the exclusion
939 of the relevant genomes from our dataset. A single Australian isolate grouping with the Rso2BS4
940 strains carrying RipP3 was also isolated from *S. tuberosum*.
941
942 Rps1S18 displays a uniform effector repertoire, with one exception: RipS4 is variable across
943 Rps1S18. The gain and loss of RipS4 does not appear to be linked with host, substrate or location of
944 isolation however. The same effector is often missing in other *Rps* clades circulating in French
945 Guiana, but present in the majority of *Rps* strains from Asia. Rps1S18 also lacks RipAL, an effector
946 otherwise present in almost every other RSSC genome. Conversely, Rps1S18 acquired RipBO, an
947 effector almost entirely absent from every sequenced Rps and Rso (Fig. 3). The effector RipAX2 has
948 been demonstrated to trigger specific resistance in eggplant [134].

949 **CRISPR-Cas loci in Rso2BS4 and Rso2BS4NPB**

950 The CRISPR-Cas operon (*cas3*, *casA/cas8e*, *casB/cas11/cse2*, *cas7e*, *cas5e*, *cas6e*, *cas1e*, *cas2*) is
951 conserved between Rso2BS4-UW163 and Rso2BS4NPB RUN1145 (a single nucleotide frameshift
952 mutation resulting in a premature stop in Rso2BS4 UW163 *cas3* is likely a sequencing error as it is
953 located in a region of 6 consecutive cytosines). The kmer-based association tests also show that the
954 largest cluster of Martinique-associated kmers that do not map to MGEs map to the leader-proximal
955 position of CRISPR arrays in Rso2BS4NPB. The 31bp kmers partially overlap with the near identical
956 29bp CRISPR 1 and CRISPR 2 repeat sequences, as well as up to 7bp of spacer sequence, with
957 CRISPR repeats sometimes differing in the final 1-2bp.
958
959 The INPHARED phage database was queried using Rso2BS4 and Rso2BS4NPB spacers, yielding hits
960 to nine different Ralstonia phages: RPZH3 [135], RS603, RSMSuper [136], Eline, Raharianne [137],
961 YO010_2, YO134_2, YO026_1 and YO158_1 [138] (Table S8). Spacers shared between CRISPR
962 arrays of both Rso2BS4NPB RUN1145 and Rso2BS4 UW163 map to YO134_2, YO026_1 and
963 YO158_1 phage. The other spacers with hits in the INPHARED database are among the set unique to
964 and recently acquired by Rso2BS4NPB RUN1145. None of the Rso2BS4NPB RUN1145 spacers
965 aligned to the RefSeq Phage or env_nt databases. NCBI-nr BLASTn reported two hits for
966 Rso2BS4NPB unique spacers: Rps3S23 (LN899820.1) and *Burkholderia* sp. S-53 (CP090483 and
967 CP090484) genomes.

968 **Spatial distribution of genetic variation**

969 There are many genes found solely in association with specific location, at varying spatial scales
970 (Table S12). The overlapping set of genes identified by both SCOARY and Pyseer panGWAS
971 includes 27 genes associated with isolation from specific districts (communes) in Martinique (Fig.
972 S10, Table S11, Table S12). Many kmers (48,477) were identified as being associated with a specific
973 district (commune) or country (Table S9). Very few had overlapping geographic associations (for
974 example between country and district), but 32% (15,747/48,477) of these had overlapping
975 associations between location and host of isolation. For example, some kmers show evidence of
976 association with isolation from both banana and Peru. Mapping kmers against reference genomes
977 shows there is ample genetic variation associated with Martinique among both Rso2BS4NPB and
978 Rps1S18. This is not unique to Martinique: there are also China and Peru-specific kmer associations
979 (Table S10). 93.5% of the kmers mapping to Rso2BS4NPB RUN1145 are associated with isolation in
980 Martinique.

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, Jun. 2014, doi: 10.1094/PHYTO-09-13-0264-R.

[5] J. I. Meier *et al.*, "Haplotype tagging reveals parallel formation of hybrid races in two butterfly species," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 118, no. 25, p. e2015005118, Jun. 2021, doi: 10.1073/pnas.2015005118.

[6] "Guppy protocol - Guppy software overview," Oxford Nanopore Technologies. Accessed: Jun. 12, 2024. [Online]. Available: https://community.nanoporetech.com/protocols/Guppy-protocol/v/gpb_2003_v1_revax_14dec2018

[7] S. W. Wingett and S. Andrews, "FastQ Screen: A tool for multi-genome mapping and quality control," Sep. 17, 2018, *F1000Research*: 7:1338. doi: 10.12688/f1000research.15931.2.

[8] "BBMap," SourceForge. Accessed: Jun. 12, 2024. [Online]. Available: https://sourceforge.net/projects/bbmap/

[9] S. Chen, Y. Zhou, Y. Chen, and J. Gu, "fastp: an ultra-fast all-in-one FASTQ preprocessor," *Bioinformatics*, vol. 34, no. 17, pp. i884–i890, Sep. 2018, doi: 10.1093/bioinformatics/bty560.

[10] S. Chen, "Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp," *iMeta*, vol. 2, no. 2, p. e107, 2023, doi: 10.1002/imt2.107.

[11] A. Bankevich *et al.*, "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing," *J. Comput. Biol.*, vol. 19, no. 5, p. 455, May 2012, doi: 10.1089/cmb.2012.0021.

[12] A. Prjibelski, D. Antipov, D. Meleshko, A. Lapidus, and A. Korobeynikov, "Using SPAdes De Novo Assembler," *Curr. Protoc. Bioinforma.*, vol. 70, no. 1, p. e102, 2020, doi: https://doi.org/10.1002/cpbi.102.

[13] 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.

[14] R. Wick, *rrwick/Filtlong*. (Jun. 06, 2024). C++. Accessed: Jun. 12, 2024. [Online]. Available: https://github.com/rrwick/Filtlong

[15] R. R. Wick *et al.*, "Trycycler: consensus long-read assemblies for bacterial genomes," *Genome Biol.*, vol. 22, no. 1, p. 266, Sep. 2021, doi: 10.1186/s13059-021-02483-z.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] *nanoporetech/medaka*. (Jun. 07, 2024). Python. Oxford Nanopore Technologies. Accessed: Jun. 12, 2024. [Online]. Available: https://github.com/nanoporetech/medaka

[21] 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.

[22] 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.

[23] 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.

[24] O. Schwengers, "Bakta database." Zenodo, Feb. 20, 2023. doi: 10.5281/zenodo.7669534.

[25] 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.

[26] "Genome," NCBI. Accessed: Jun. 12, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=1310165,28097,305

[27] *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

[28] 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.

[29] 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.

[30] "Ralsto T3E." Accessed: May 31, 2024. [Online]. Available: https://iant.toulouse.inra.fr/bacteria/annotation/site/prj/T3Ev3/

[31] 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.

[32] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.

[33] 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.

[34] A. J. Page *et al.*, "SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments," *Microb. Genomics*, vol. 2, no. 4, p. e000056, 2016, doi: 10.1099/mgen.0.000056.

[35] L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, and B. Q. Minh, "IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies," *Mol. Biol. Evol.*, vol. 32, no. 1, pp. 268–274, Jan. 2015, doi: 10.1093/molbev/msu300.

[36] G. Yu, D. K. Smith, H. Zhu, Y. Guan, and T. T.-Y. Lam, "ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data," *Methods Ecol. Evol.*, vol. 8, no. 1, pp. 28–36, 2017, doi: 10.1111/2041-210X.12628.

[37] Y. Zhou, Y. Liang, K. H. Lynch, J. J. Dennis, and D. S. Wishart, "PHAST: a fast phage search tool," *Nucleic Acids Res.*, vol. 39, no. Web Server issue, pp. W347-352, Jul. 2011, doi: 10.1093/nar/gkr485.

[38] D. Arndt *et al.*, "PHASTER: a better, faster version of the PHAST phage search tool," *Nucleic Acids Res.*, vol. 44, no. W1, pp. W16-21, Jul. 2016, doi: 10.1093/nar/gkw387.

[39] D. S. Wishart *et al.*, "PHASTEST: faster than PHASTER, better than PHAST," *Nucleic Acids Res.*, vol. 51, no. W1, pp. W443–W450, May 2023, doi: 10.1093/nar/gkad382.

[40] P. Siguier, J. Perochon, L. Lestrade, J. Mahillon, and M. Chandler, "ISfinder: the reference centre for bacterial insertion sequences," *Nucleic Acids Res.*, vol. 34, no. Database issue, pp. D32–D36, Jan. 2006, doi: 10.1093/nar/gkj014.

[41] "Geneious | Bioinformatics Software for Sequence Data Analysis," Geneious. Accessed: Jun. 12, 2024. [Online]. Available: https://www.geneious.com/

[42] A. R. Quinlan and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *Bioinforma. Oxf. Engl.*, vol. 26, no. 6, pp. 841–842, Mar. 2010, doi: 10.1093/bioinformatics/btq033.

[43] 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.

[44] H. Li, “Minimap2: pairwise alignment for nucleotide sequences,” *Bioinformatics*, vol. 34, no. 18, pp. 3094–3100, Sep. 2018, doi: 10.1093/bioinformatics/bty191.
[45] H. Li, “A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data,” *Bioinformatics*, vol. 27, no. 21, pp. 2987–2993, Nov. 2011, doi: 10.1093/bioinformatics/btr509.
[46] B. S. Pedersen and A. R. Quinlan, “Mosdepth: quick coverage calculation for genomes and exomes,” *Bioinformatics*, vol. 34, no. 5, pp. 867–868, Mar. 2018, doi: 10.1093/bioinformatics/btx699.
[47] F. Lemoine and O. Gascuel, “Gotree/Goalign: toolkit and Go API to facilitate the development of phylogenetic workflows,” *NAR Genomics Bioinforma.*, vol. 3, no. 3, p. lqab075, Sep. 2021, doi: 10.1093/nargab/lqab075.
[48] X. Didelot and D. J. Wilson, “ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes,” *PLOS Comput. Biol.*, vol. 11, no. 2, p. e1004041, Feb. 2015, doi: 10.1371/journal.pcbi.1004041.
[49] S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, and L. S. Jermiin, “ModelFinder: fast model selection for accurate phylogenetic estimates,” *Nat. Methods*, vol. 14, no. 6, pp. 587–589, Jun. 2017, doi: 10.1038/nmeth.4285.
[50] X. Didelot, N. J. Croucher, S. D. Bentley, S. R. Harris, and D. J. Wilson, “Bayesian inference of ancestral dates on bacterial phylogenetic trees,” *Nucleic Acids Res.*, vol. 46, no. 22, p. e134, Dec. 2018, doi: 10.1093/nar/gky783.
[51] A. Rambaut, T. T. Lam, L. Max Carvalho, and O. G. Pybus, “Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen),” *Virus Evol.*, vol. 2, no. 1, p. vew007, Jan. 2016, doi: 10.1093/ve/vew007.
[52] D. H. Huson, D. C. Richter, C. Rausch, T. Dezulian, M. Franz, and R. Rupp, “Dendroscope: An interactive viewer for large phylogenetic trees,” *BMC Bioinformatics*, vol. 8, no. 1, p. 460, Nov. 2007, doi: 10.1186/1471-2105-8-460.
[53] O. S. Alkhnbashi *et al.*, “CRISPRloci: comprehensive and accurate annotation of CRISPR–Cas systems,” *Nucleic Acids Res.*, vol. 49, no. W1, pp. W125–W130, Jun. 2021, doi: 10.1093/nar/gkab456.
[54] R. H. Wang *et al.*, “PhageScope: a well-annotated bacteriophage database with automatic analyses and visualizations,” *Nucleic Acids Res.*, vol. 52, no. D1, pp. D756–D761, Jan. 2024, doi: 10.1093/nar/gkad979.
[55] R. Cook *et al.*, “INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes,” *PHAGE*, vol. 2, no. 4, pp. 214–223, Dec. 2021, doi: 10.1089/phage.2021.0007.
[56] R. Zhang, M. Mirdita, E. Levy Karin, C. Norroy, C. Galiez, and J. Söding, “SpacePHARER: sensitive identification of phages from CRISPR spacers in prokaryotic hosts,” *Bioinformatics*, vol. 37, no. 19, pp. 3364–3366, Oct. 2021, doi: 10.1093/bioinformatics/btab222.
[57] T. Sakoparnig, C. Field, and E. van Nimwegen, “Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species,” *eLife*, vol. 10, p. e65366, Jan. 2021, doi: 10.7554/eLife.65366.
[58] B. D. Ondov *et al.*, “Mash: fast genome and metagenome distance estimation using MinHash,” *Genome Biol.*, vol. 17, no. 1, p. 132, Jun. 2016, doi: 10.1186/s13059-016-0997-x.
[59] O. Brynildsrud, J. Bohlin, L. Scheffer, and V. Eldholm, “Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary,” *Genome Biol.*, vol. 17, no. 1, p. 238, Nov. 2016, doi: 10.1186/s13059-016-1108-8.
[60] J. A. Lees, M. Galardini, S. D. Bentley, J. N. Weiser, and J. Corander, “pyseer: a comprehensive tool for microbial pangenome-wide association studies,” *Bioinformatics*, vol. 34, no. 24, pp. 4310–4312, Dec. 2018, doi: 10.1093/bioinformatics/bty539.
[61] T. pandas development team, *pandas-dev/pandas: Pandas*. (Apr. 10, 2024). Zenodo. doi: 10.5281/zenodo.10957263.
[62] M. Galardini, *mgalardini/pyseer*. (Sep. 09, 2024). Python. Accessed: Jan. 20, 2024. [Online]. Available: https://github.com/mgalardini/pyseer
[63] G. Holley and P. Melsted, “Bifrost – Highly parallel construction and indexing of colored and compacted de Bruijn graphs,” Aug. 13, 2019, *bioRxiv*. doi: 10.1101/695338.

[64] J. R. Conway, A. Lex, and N. Gehlenborg, "UpSetR: an R package for the visualization of intersecting sets and their properties," *Bioinformatics*, vol. 33, no. 18, pp. 2938–2940, Sep. 2017, doi: 10.1093/bioinformatics/btx364.
[65] E. Colombi *et al.*, "Comparative analysis of integrative and conjugative mobile genetic elements in the genus Mesorhizobium," *Microb. Genomics*, vol. 7, no. 10, p. 000657, 2021, doi: 10.1099/mgen.0.000657.
[66] "ICEberg 3.0." Accessed: May 28, 2024. [Online]. Available: https://tool2-mml.sjtu.edu.cn/ICEberg3/Browse_detail.php?iceid=335
[67] F. Lu and G. Churchward, "Tn916 target DNA sequences bind the C-terminal domain of integrase protein with different affinities that correlate with transposon insertion frequency," *J. Bacteriol.*, vol. 177, no. 8, pp. 1938–1946, Apr. 1995, doi: 10.1128/jb.177.8.1938-1946.1995.
[68] M. P. Garcillán-Barcia, S. Redondo-Salvo, L. Vielva, and F. de la Cruz, "MOBscan: Automated Annotation of MOB Relaxases," *Methods Mol. Biol. Clifton NJ*, vol. 2075, pp. 295–308, 2020, doi: 10.1007/978-1-4939-9877-7_21.
[69] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, Oct. 1990, doi: 10.1016/S0022-2836(05)80360-2.
[70] S. S. Abby, J. Cury, J. Guglielmini, B. Néron, M. Touchon, and E. P. C. Rocha, "Identification of protein secretion systems in bacterial genomes," *Sci. Rep.*, vol. 6, no. 1, p. 23080, Mar. 2016, doi: 10.1038/srep23080.
[71] 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.
[72] *pycountry/pycountry*. (Aug. 28, 2024). Python. pycountry. Accessed: Feb. 04, 2021. [Online]. Available: https://github.com/pycountry/pycountry
[73] "Welcome to GeoPy's documentation! — GeoPy 2.4.1 documentation." Accessed: Jun. 12, 2024. [Online]. Available: https://geopy.readthedocs.io/en/stable/
[74] "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
[75] J. V. den Bossche *et al.*, *geopandas/geopandas: v0.14.3*. (Jan. 31, 2024). Zenodo. doi: 10.5281/zenodo.10601680.
[76] D. Ugent, "The Potato," *Science*, vol. 170, no. 3963, pp. 1161–1166, Dec. 1970, doi: 10.1126/science.170.3963.1161.
[77] 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.
[78] 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.
[79] 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.
[80] 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.
[81] W. J. Kress, "Systematics of Central American Heliconia (heliconiaceae) with Pendent Inflorescences," *J. Arnold Arbor.*, vol. 65, no. 4, pp. 429–532, 1984.
[82] T. B. Croat and R. Box, "Ecology and Life Forms of Araceae," vol. 11, 1988.
[83] F. Barbirato, J. C. Verdoes, J. A. M. de Bont, and M. J. van der Werf, "The Rhodococcus erythropolis DCL14 limonene-1,2-epoxide hydrolase gene encodes an enzyme belonging to a novel class of epoxide hydrolases," *FEBS Lett.*, vol. 438, no. 3, pp. 293–296, 1998, doi: 10.1016/S0014-5793(98)01322-2.
[84] S. Doron *et al.*, "Systematic discovery of antiphage defense systems in the microbial

pangenome," *Science*, vol. 359, no. 6379, p. eaar4120, Mar. 2018, doi: 10.1126/science.aar4120.

[85] R. Cheng *et al.*, "A nucleotide-sensing endonuclease from the Gabija bacterial defense system," *Nucleic Acids Res.*, vol. 49, no. 9, pp. 5216–5229, May 2021, doi: 10.1093/nar/gkab277.

[86] 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.

[87] J. E. Galán and A. Collmer, "Type III Secretion Machines: Bacterial Devices for Protein Delivery into Host Cells," *Science*, vol. 284, no. 5418, pp. 1322–1328, May 1999, doi: 10.1126/science.284.5418.1322.

[88] 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.

[89] 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.

[90] 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.

[91] 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.

[92] A. Zaprasis, Y.-J. Liu, S.-J. Liu, H. L. Drake, and M. A. Horn, "Abundance of novel and diverse tfdA-like genes, encoding putative phenoxyalkanoic acid herbicide-degrading dioxygenases, in soil," *Appl. Environ. Microbiol.*, vol. 76, no. 1, pp. 119–128, Jan. 2010, doi: 10.1128/AEM.01727-09.

[93] J. Baelum, T. Henriksen, H. C. B. Hansen, and C. S. Jacobsen, "Degradation of 4-chloro-2-methylphenoxyacetic acid in top- and subsoil is quantitatively linked to the class III tfdA gene," *Appl. Environ. Microbiol.*, vol. 72, no. 2, pp. 1476–1486, Feb. 2006, doi: 10.1128/AEM.72.2.1476-1486.2006.

[94] D. T. Newby, T. J. Gentry, and I. L. Pepper, "Comparison of 2,4-Dichlorophenoxyacetic Acid Degradation and Plasmid Transfer in Soil Resulting from Bioaugmentation with Two Different pJP4 Donors," *Appl. Environ. Microbiol.*, vol. 66, no. 8, pp. 3399–3407, Aug. 2000, doi: 10.1128/AEM.66.8.3399-3407.2000.

[95] H. P. Füchslin, I. Rüegg, J. R. Van Der Meer, and T. Egli, "Effect of integration of a GFP reporter gene on fitness of Ralstonia eutropha during growth with 2,4-dichlorophenoxyacetic acid," *Environ. Microbiol.*, vol. 5, no. 10, pp. 878–887, 2003, doi: 10.1046/j.1462-2920.2003.00479.x.

[96] P. T. Chivers, K. E. Prehoda, and R. T. Raines, "The CXXC Motif: A Rheostat in the Active Site," *Biochemistry*, vol. 36, no. 14, pp. 4061–4066, Apr. 1997, doi: 10.1021/bi9628580.

[97] S. Quan, I. Schneider, J. Pan, A. V. Hacht, and J. C. A. Bardwell, "The CXXC Motif Is More than a Redox Rheostat *," *J. Biol. Chem.*, vol. 282, no. 39, pp. 28823–28833, Sep. 2007, doi: 10.1074/jbc.M705291200.

[98] J. Sancho, "Flavodoxins: sequence, folding, binding, function and beyond," *Cell. Mol. Life Sci.*, vol. 63, no. 7–8, pp. 855–864, Apr. 2006, doi: 10.1007/s00018-005-5514-4.

[99] D. Eulberg, S. Lakner, L. A. Golovleva, and M. Schlömann, "Characterization of a Protocatechuate Catabolic Gene Cluster from Rhodococcus opacus 1CP: Evidence for a Merged Enzyme with 4-Carboxymuconolactone-Decarboxylating and 3-Oxoadipate Enol-Lactone-Hydrolyzing Activity," *J. Bacteriol.*, vol. 180, no. 5, pp. 1072–1081, Mar. 1998.

[100] D. B. Martim *et al.*, "Resolving the metabolism of monolignols and other lignin-related aromatic compounds in Xanthomonas citri," *Nat. Commun.*, vol. 15, no. 1, p. 7994, Sep. 2024, doi: 10.1038/s41467-024-52367-6.

[101] A. da Silva Xavier *et al.*, "Characterization of CRISPR-Cas systems in the Ralstonia solanacearum species complex," *Mol. Plant Pathol.*, vol. 20, no. 2, pp. 223–239, Nov. 2018, doi: 10.1111/mpp.12750.

[102] J. He and M. W. Deem, "Heterogeneous Diversity of Spacers within CRISPR (Clustered

Regularly Interspaced Short Palindromic Repeats),” *Phys. Rev. Lett.*, vol. 105, no. 12, p. 128102, Sep. 2010, doi: 10.1103/PhysRevLett.105.128102.

[103] C. A. Rowley and M. M. Kendall, “To B12 or not to B12: Five questions on the role of cobalamin in host-microbial interactions,” *PLOS Pathog.*, vol. 15, no. 1, p. e1007479, Jan. 2019, doi: 10.1371/journal.ppat.1007479.

[104] Z. Freyberg and E. T. Harvill, “Pathogen manipulation of host metabolism: A common strategy for immune evasion,” *PLOS Pathog.*, vol. 13, no. 12, p. e1006669, Dec. 2017, doi: 10.1371/journal.ppat.1006669.

[105] J. Cai, Y. Jiang, E. S. Ritchie, A. P. Macho, F. Yu, and D. Wu, “Manipulation of plant metabolism by pathogen effectors: more than just food,” *FEMS Microbiol. Rev.*, vol. 47, no. 2, p. fuad007, Mar. 2023, doi: 10.1093/femsre/fuad007.

[106] J.-S. Eom *et al.*, “SWEETs, transporters for intracellular and intercellular sugar translocation,” *Curr. Opin. Plant Biol.*, vol. 25, pp. 53–62, Jun. 2015, doi: 10.1016/j.pbi.2015.04.005.

[107] C. D. Hamilton, O. R. Steidl, A. M. MacIntyre, C. G. Hendrich, and C. Allen, “Ralstonia solanacearum Depends on Catabolism of Myo-Inositol, Sucrose, and Trehalose for Virulence in an Infection Stage–Dependent Manner,” *Mol. Plant-Microbe Interactions®*, vol. 34, no. 6, pp. 669–679, Jun. 2021, doi: 10.1094/MPMI-10-20-0298-R.

[108] A. N. Truchon *et al.*, “Plant-Pathogenic Ralstonia Phylotypes Evolved Divergent Respiratory Strategies and Behaviors To Thrive in Xylem,” *mBio*, vol. 14, no. 1, pp. e03188-22, Feb. 2023, doi: 10.1128/mbio.03188-22.

[109] R. Gopalan-Nair *et al.*, “Convergent Rewiring of the Virulence Regulatory Network Promotes Adaptation of Ralstonia solanacearum on Resistant Tomato,” *Mol. Biol. Evol.*, vol. 38, no. 5, pp. 1792–1808, May 2021, doi: 10.1093/molbev/msaa320.

[110] R. de Pedro-Jové *et al.*, “Dynamic expression of Ralstonia solanacearum virulence factors and metabolism-controlling genes during plant infection,” *BMC Genomics*, vol. 22, no. 1, p. 170, Mar. 2021, doi: 10.1186/s12864-021-07457-w.

[111] C. M. Johnson and A. D. Grossman, “Integrative and Conjugative Elements (ICEs): What They Do and How They Work,” *Annu. Rev. Genet.*, vol. 49, pp. 577–601, 2015, doi: 10.1146/annurev-genet-112414-055018.

[112] V. Sauveplane *et al.*, “Arabidopsis thaliana CYP77A4 is the first cytochrome P450 able to catalyze the epoxidation of free fatty acids in plants,” *FEBS J.*, vol. 276, no. 3, pp. 719–735, 2009, doi: 10.1111/j.1742-4658.2008.06819.x.

[113] J. Ding *et al.*, “Properties of a Newly Identified Esterase from Bacillus sp. K91 and Its Novel Function in Diisobutyl Phthalate Degradation,” *PLoS ONE*, vol. 10, no. 3, p. e0119216, Mar. 2015, doi: 10.1371/journal.pone.0119216.

[114] N. Allocati, L. Federici, M. Masulli, and C. Di Ilio, “Glutathione transferases in bacteria,” *FEBS J.*, vol. 276, no. 1, pp. 58–75, 2009, doi: 10.1111/j.1742-4658.2008.06743.x.

[115] J. E. L. Santos, M. A. Gómez, D. C. de Moura, M. Cerro-López, M. A. Quiroz, and C. A. Martínez-Huitle, “Removal of herbicide 1-chloro-2,4-dinitrobenzene (DNCB) from aqueous solutions by electrochemical oxidation using boron-doped diamond (BDD) and PbO2 electrodes,” *J. Hazard. Mater.*, vol. 402, p. 123850, Jan. 2021, doi: 10.1016/j.jhazmat.2020.123850.

[116] R. N. Patkar *et al.*, “A fungal monooxygenase-derived jasmonate attenuates host innate immunity,” *Nat. Chem. Biol.*, vol. 11, no. 9, pp. 733–740, Sep. 2015, doi: 10.1038/nchembio.1885.

[117] O. R. Feitosa-Junior *et al.*, “The Exometabolome of Xylella fastidiosa in Contact with Paraburkholderia phytofirmans Supernatant Reveals Changes in Nicotinamide, Amino Acids, Biotin, and Plant Hormones,” *Metabolites*, vol. 14, no. 2, Art. no. 2, Feb. 2024, doi: 10.3390/metabo14020082.

[118] C. D. Hamilton, B. Zaricor, C. J. Dye, E. Dresserl, R. Michaels, and C. Allen, “Ralstonia solanacearum pandemic lineage strain UW551 overcomes inhibitory xylem chemistry to break tomato bacterial wilt resistance,” *Mol. Plant Pathol.*, vol. 25, no. 1, p. e13395, 2024, doi: 10.1111/mpp.13395.

[119] C. Baroukh, L. Cottret, E. Pires, R. Peyraud, A. Guidot, and S. Genin, “Insights into the metabolic specificities of pathogenic strains from the Ralstonia solanacearum species complex,”

mSystems, vol. 8, no. 4, pp. e00083-23, Jun. 2023, doi: 10.1128/msystems.00083-23.

[120] A. S. Purnomo, A. Sariwati, and I. Kamei, "Synergistic interaction of a consortium of the brown-rot fungus Fomitopsis pinicola and the bacterium Ralstonia pickettii for DDT biodegradation," *Heliyon*, vol. 6, no. 6, p. e04027, Jun. 2020, doi: 10.1016/j.heliyon.2020.e04027.

[121] N. Trefault *et al.*, "Genetic organization of the catabolic plasmid pJP4 from Ralstonia eutropha JMP134 (pJP4) reveals mechanisms of adaptation to chloroaromatic pollutants and evolution of specialized chloroaromatic degradation pathways," *Environ. Microbiol.*, vol. 6, no. 7, pp. 655–668, 2004, doi: 10.1111/j.1462-2920.2004.00596.x.

[122] J. B. Emerson *et al.*, "Host-linked soil viral ecology along a permafrost thaw gradient," *Nat. Microbiol.*, vol. 3, no. 8, pp. 870–880, Aug. 2018, doi: 10.1038/s41564-018-0190-y.

[123] P. E. Dougherty *et al.*, "Widespread and largely unknown prophage activity, diversity, and function in two genera of wheat phyllosphere bacteria," *ISME J.*, vol. 17, no. 12, pp. 2415–2425, Dec. 2023, doi: 10.1038/s41396-023-01547-1.

[124] H. C. McCann *et al.*, "Genomic Analysis of the Kiwifruit Pathogen Pseudomonas syringae pv. actinidiae Provides Insight into the Origins of an Emergent Plant Disease," *PLOS Pathog.*, vol. 9, no. 7, p. e1003503, Jul. 2013, doi: 10.1371/journal.ppat.1003503.

[125] H. C. McCann *et al.*, "Origin and Evolution of the Kiwifruit Canker Pandemic," *Genome Biol. Evol.*, vol. 9, no. 4, pp. 932–944, Apr. 2017, doi: 10.1093/gbe/evx055.

[126] K. G. Pegg, L. M. Coates, W. T. O'Neill, and D. W. Turner, "The Epidemiology of Fusarium Wilt of Banana," *Front. Plant Sci.*, vol. 10, Dec. 2019, doi: 10.3389/fpls.2019.01395.

[127] A. Drenth and G. Kema, "The Vulnerability of Bananas to Globally Emerging Disease Threats," *Phytopathology®*, vol. 111, no. 12, pp. 2146–2161, Dec. 2021, doi: 10.1094/PHYTO-07-20-0311-RVW.

[128] R. C. Ploetz, "Management of Fusarium wilt of banana: A review with special reference to tropical race 4," *Crop Prot.*, vol. 73, pp. 7–15, Jul. 2015, doi: 10.1016/j.cropro.2015.01.007.

[129] F. A. García-Bastidas *et al.*, "First Report of Fusarium Wilt Tropical Race 4 in Cavendish Bananas Caused by Fusarium odoratissimum in Colombia," *Plant Dis.*, vol. 104, no. 3, pp. 994–994, Mar. 2020, doi: 10.1094/PDIS-09-19-1922-PDN.

[130] J. Beutler, S. Holden, S. Georgoulis, D. Williams, D. J. Norman, and T. M. Lowe-Power, "Whole Genome Sequencing Suggests that 'Nonpathogenicity on Banana (NPB)' Is the Ancestral State of the Ralstonia solanacearum IIB-4 Lineage," *PhytoFrontiers™*, vol. 3, no. 2, pp. 262–267, Sep. 2023, doi: 10.1094/PHYTOFR-06-22-0068-SC.

[131] K. L. Menard and A. D. Grossman, "Selective Pressures to Maintain Attachment Site Specificity of Integrative and Conjugative Elements," *PLoS Genet.*, vol. 9, no. 7, p. e1003623, Jul. 2013, doi: 10.1371/journal.pgen.1003623.

[132] M. Neumann, G. Mittelstädt, F. Seduk, C. Iobbi-Nivol, and S. Leimkühler, "MocA Is a Specific Cytidylyltransferase Involved in Molybdopterin Cytosine Dinucleotide Biosynthesis in *Escherichia coli**," *J. Biol. Chem.*, vol. 284, no. 33, pp. 21891–21898, Aug. 2009, doi: 10.1074/jbc.M109.008565.

[133] T. Conway and L. O. Ingram, "Similarity of Escherichia coli propanediol oxidoreductase (fucO product) and an unusual alcohol dehydrogenase from Zymomonas mobilis and Saccharomyces cerevisiae.," *J. Bacteriol.*, vol. 171, no. 7, pp. 3754–3759, Jul. 1989.

[134] A. Morel *et al.*, "The eggplant AG91-25 recognizes the Type III-secreted effector RipAX2 to trigger resistance to bacterial wilt (Ralstonia solanacearum species complex)," *Mol. Plant Pathol.*, vol. 19, no. 11, pp. 2459–2472, 2018, doi: 10.1111/mpp.12724.

[135] Z. Lin, G. Gu, C. Chen, T. Zhou, F. Hu, and X. Cai, "Characterization and complete genome sequence analysis of the novel phage RPZH3 infecting Ralstonia solanacearum," *Arch. Virol.*, vol. 168, no. 4, p. 105, Mar. 2023, doi: 10.1007/s00705-023-05737-2.

[136] O. S. Gonçalves, F. de O. Souza, F. P. Bruckner, M. F. Santana, and P. Alfenas-Zerbini, "Widespread distribution of prophages signaling the potential for adaptability and pathogenicity evolution of *Ralstonia solanacearum* species complex," *Genomics*, vol. 113, no. 3, pp. 992–1000, May 2021, doi: 10.1016/j.ygeno.2021.02.011.

[137] A. Trotereau, C. Boyer, I. Bornard, M. J. B. Pécheur, C. Schouler, and C. Torres-Barceló, "High genomic diversity of novel phages infecting the plant pathogen Ralstonia solanacearum, isolated in Mauritius and Reunion islands," *Sci. Rep.*, vol. 11, no. 1, p. 5382, Mar. 2021, doi:

10.1038/s41598-021-84305-7.
[138] S. T. E. Greenrod, M. Stoycheva, J. Elphinstone, and V.-P. Friman, "Global diversity and distribution of prophages are lineage-specific within the Ralstonia solanacearum species complex," *BMC Genomics*, vol. 23, no. 1, p. 689, Oct. 2022, doi: 10.1186/s12864-022-08909-7.
[139] W. J. Jowsey *et al.*, "DUF2285 is a novel helix-turn-helix domain variant that orchestrates both activation and antiactivation of conjugative element transfer in proteobacteria," *Nucleic Acids Res.*, vol. 51, no. 13, pp. 6841–6856, Jul. 2023, doi: 10.1093/nar/gkad457.

999 **List of Supplemental Tables**

1000 Table S1. RSSC genomes and associated metadata

1001 Table S2. Panaroo pangenome gene presence and absence across RSSC

1002 Table S3. Rso2BS4-specific genes

1003 Table S4. Rso2BS4NPB-specific genes

1004 Table S5. Rps1S18-specific genes

1005 Table S6. T3E family distribution across the RSSC

1006 Table S7. Phages identified in 17 complete RSSC genomes

1007 Table S8. Rso2BS4NPB CRISPR spacer BLAST hits

1008 Table S9. Pyseer kmer-based trait associations

1009 Table S10. Pyseer kmer mapping positions in Rso2BS4, Rso2BS4NPB and Rps1S18-GMI1000
1010 reference genomes

1011 Table S11. Pyseer panGWAS results

1012 Table S12. SCOARY panGWAS results

1013 Table S13. MOBP relaxase positions identified by MOBscan

1014 Table S14. Assembly data for all genomes acquired from NCBI

981 **List of Supplemental Figures**

982 Fig. S1. Core gene tree of the *Ralstonia solanacearum* species complex

983 Fig. S2. Rps1S18 core genome tree with outgroup.

984 Fig. S3. Rso2BS4 core genome tree

985 Fig. S4. Rso2BS4NPB core genome tree with outgroup

986 Fig. S5. Divergence dating of Rps1S18

987 Fig. S6. Divergence dating of Rso2BS4NPB

988 Fig. S7. RECOPHY analysis for reduced set of Rso2BS4 and RpsS18 genomes

989 Fig. S8. RECOPHY analysis for reduced set of Rso2BS4 strains

990 Fig. S9. Alignment of Rso2BS4NPB and Rso2BS4 secondary chromosomes

991 Fig. S10. Complete panGWAS results displayed against RSSC core gene tree

992 Fig. S11. Rso2BS4NPB-specific prophage BLAST results

993 Fig. S12. CRISPR-Cas loci in Rso2BS4 and Rso2BS4NPB

994 Fig. S13. Distribution of ICEs across the RSSC

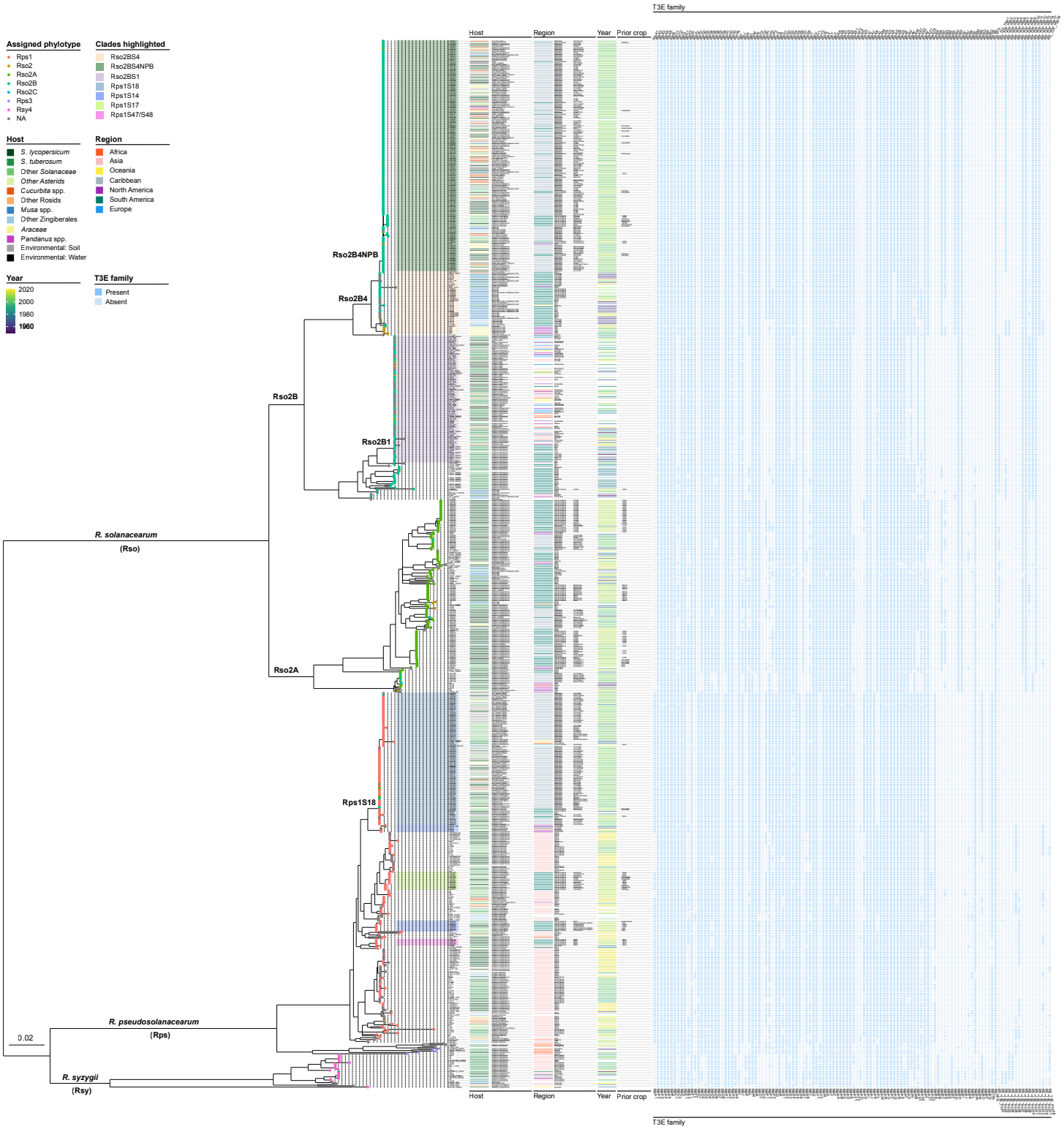
995 Fig. S14. Geographic distribution of *Ralstonia* spp. ICEs sampled in Martinique

996 Fig. S15. ICE cargo regions across multiple Rso and Rps strains

997 Fig. S16. ICE synteny across multiple Rso and Rps strains

998

Figure S1. Core gene tree of the *Ralstonia solanacearum* species complex



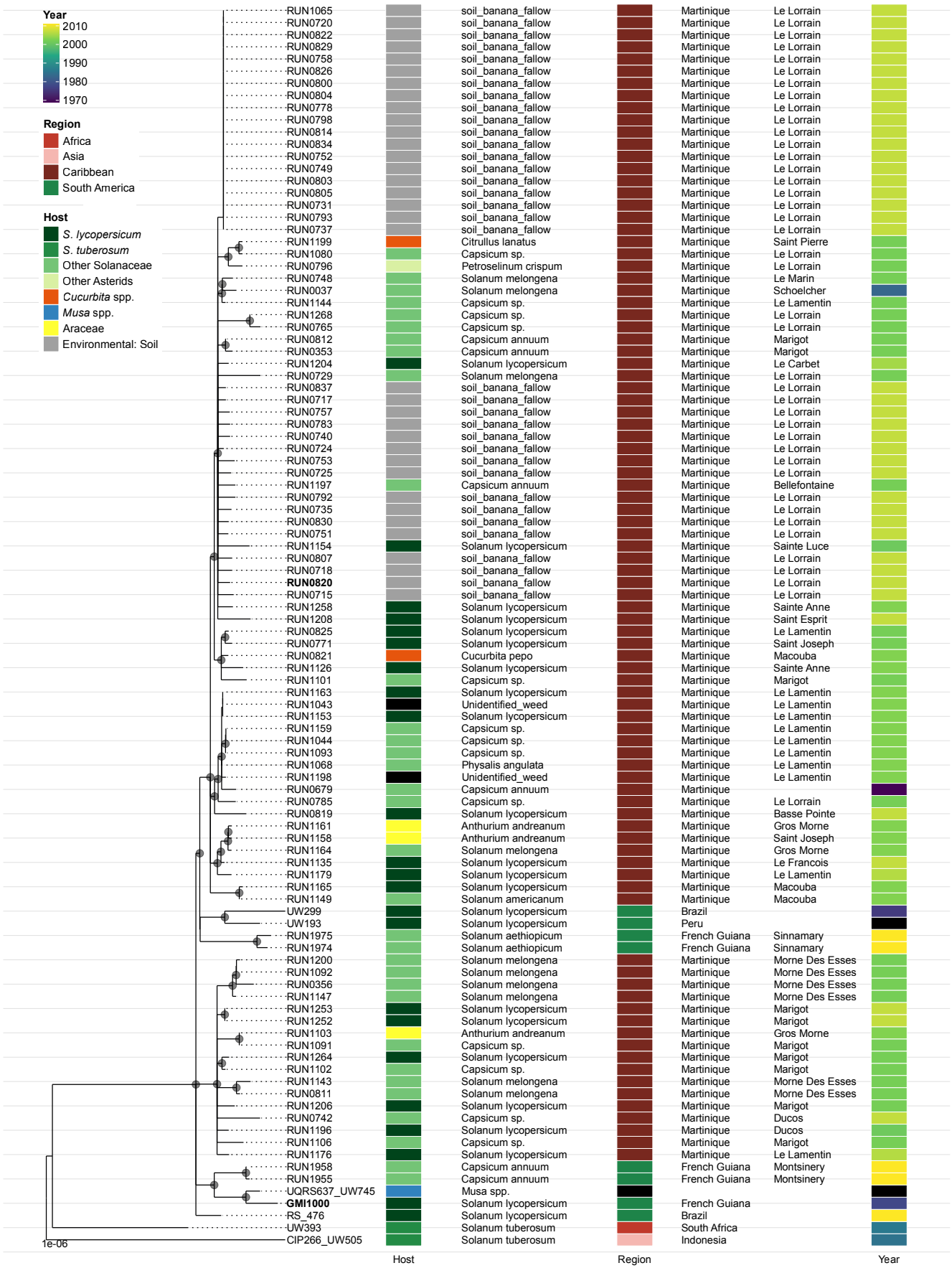


Fig. S2. Rps1S18 core genome tree with outgroup. Maximum likelihood tree of all Rps1S18 genomes together with UW393 and CIP266_UW505 outgroup strains calculated from contigs alignment against GMI1000 reference. Nodes with the bootstrap support below 70 are collapsed. Host, geographical region and year of isolation are shown as in Figure S1. Clades representing Rso and Rps lineages of interest are highlighted in the tree. Labels of the reference strains and the eight strains from Martinique and French Guiana sequenced by long-read technology during this study are shown in bold.

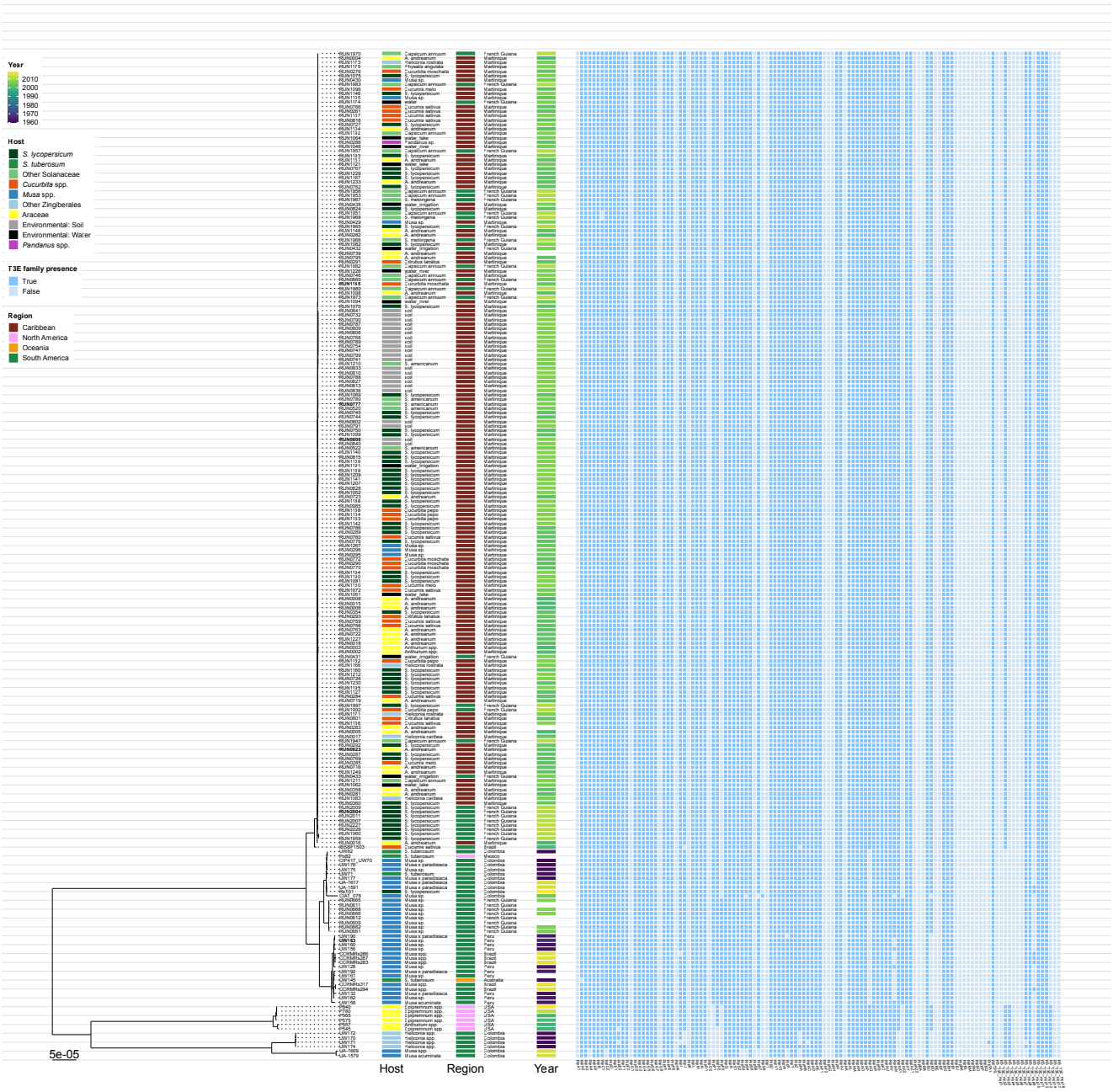


Fig. S3. Rso2BS4 core genome tree. Maximum likelihood tree of all RsoS4 genomes calculated from contigs alignment against UW163 reference. Nodes with the bootstrap support below 70 are collapsed. Host, geographical region and year of isolation are shown as in Figure S1. Clades representing Rso and Rps lineages of interest are highlighted in the tree. Labels of the reference strains and the eight strains from Martinique and French Guiana sequenced by long-read technology during this study are shown in bold. Presence and absence patterns of differentially presented T3E families are shown for each strain on the right panel.

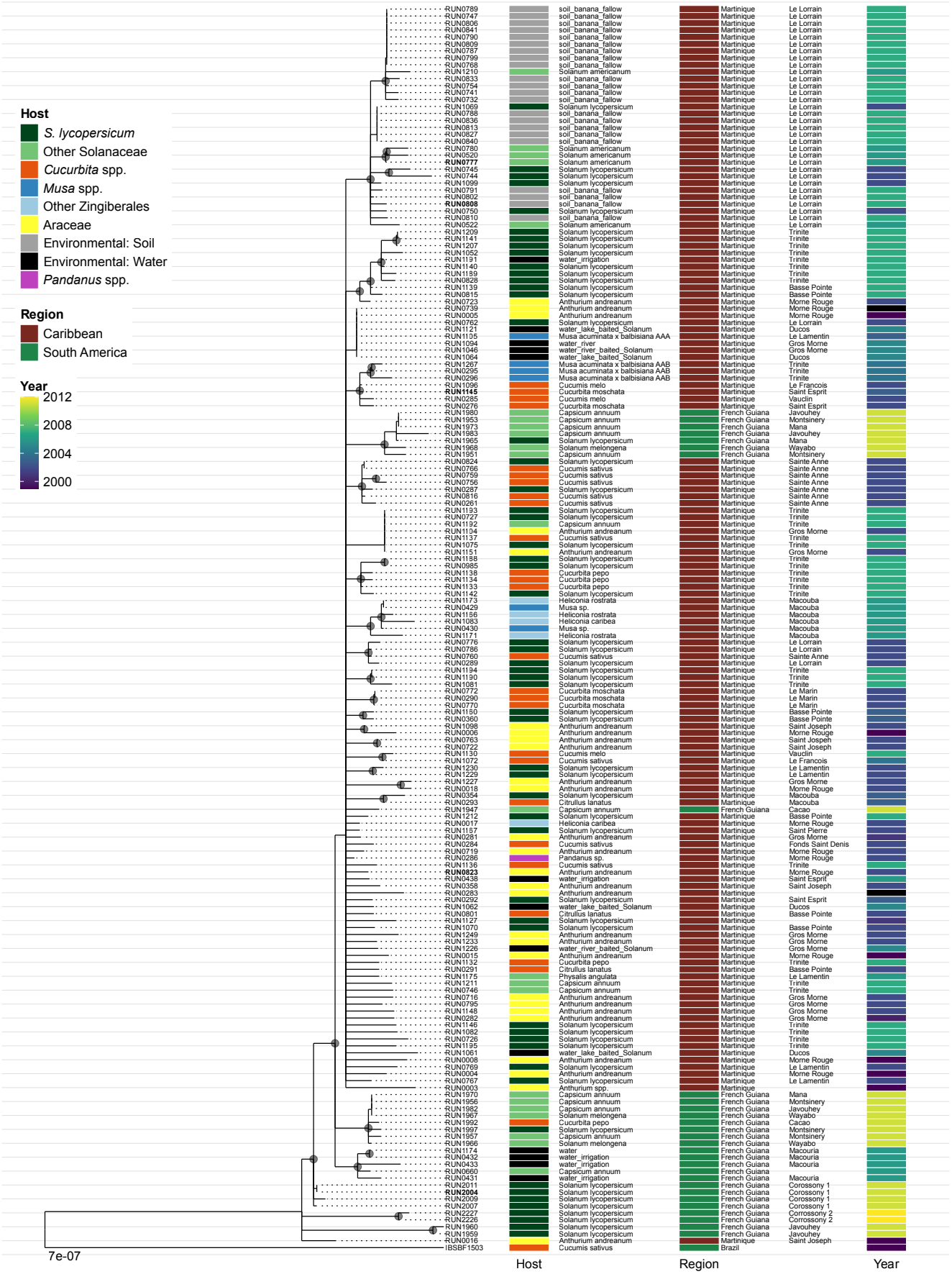


Fig. S4. Rso2BS4NPB core genome tree with outgroup. Maximum likelihood tree of all Rso2BS4NPB genomes together with an outgroup strain IBSBF1503 calculated from contigs alignment against IBSBF1503 reference. Nodes with the bootstrap support below 70 are collapsed. Host, geographical region and year of isolation are shown as in Figure S1. Clades representing Rso and Rps lineages of interest are highlighted in the tree. Labels of the reference strains and the eight strains from Martinique and French Guiana sequenced by long-read technology during this study are shown in bold.

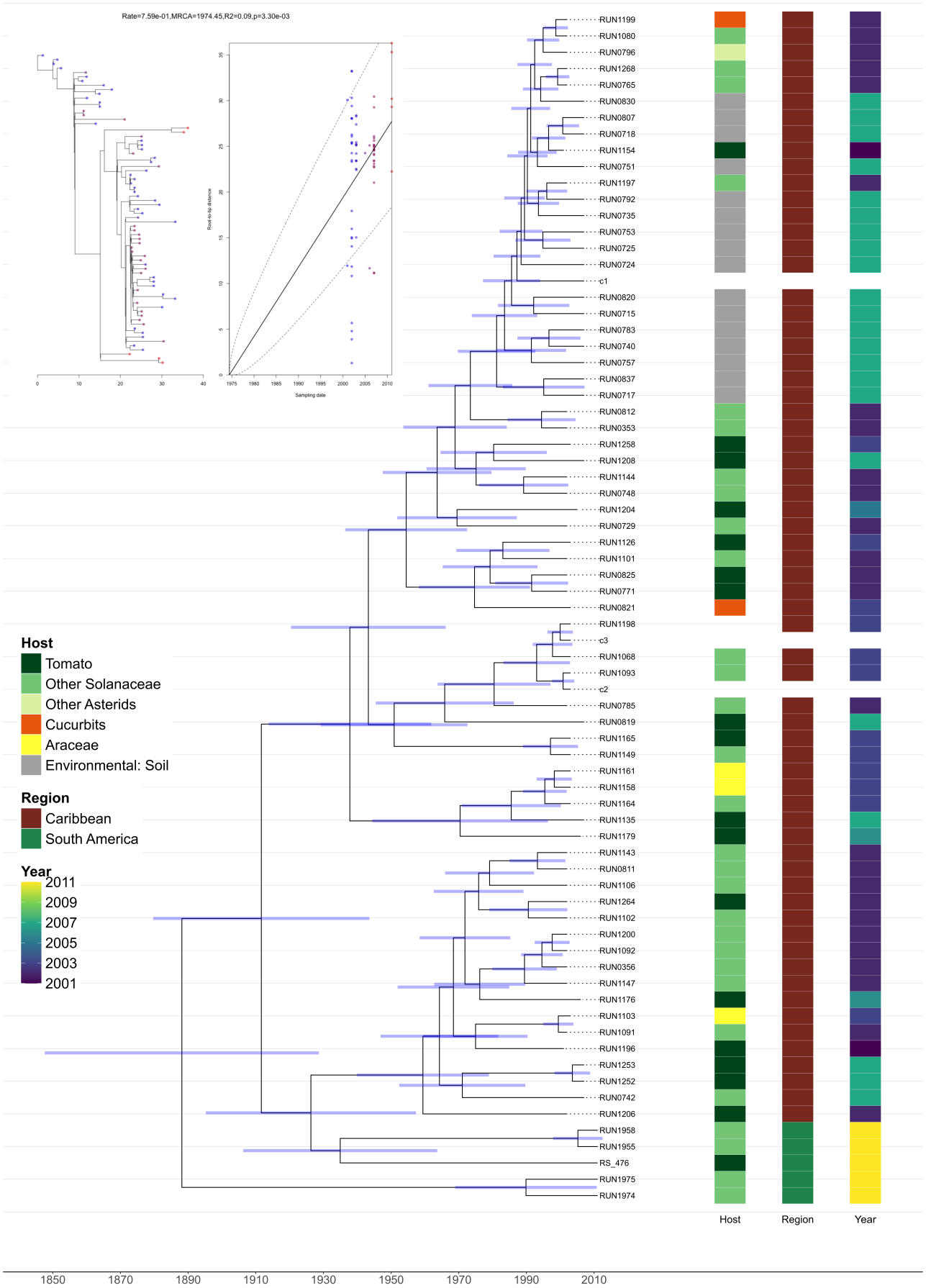


Fig. S5. Divergence dating of Rps1S18. Maximum likelihood tree of all Rps1S18 with the temporal dating confidence intervals calculated by BactDating shown as blue horizontal lines. The clades collapsed due to zero-branch length are labelled c1 to c3. The c1 node includes the strains RUN0720, RUN1065, RUN0822, RUN0829, RUN0758, RUN0826, RUN0800, RUN0804, RUN0778, RUN0798, RUN0814, RUN0834, RUN0752, RUN0749, RUN0803, RUN0805, RUN0731, RUN0793, RUN0737. The c2 node includes the strains RUN1044 and RUN1159. The c3 node includes the strains RUN1043, RUN1163 and RUN1153. Displayed together with root-to-tip regression plots generated by BactDating for the Rps1S18 strains (upper-left panel).

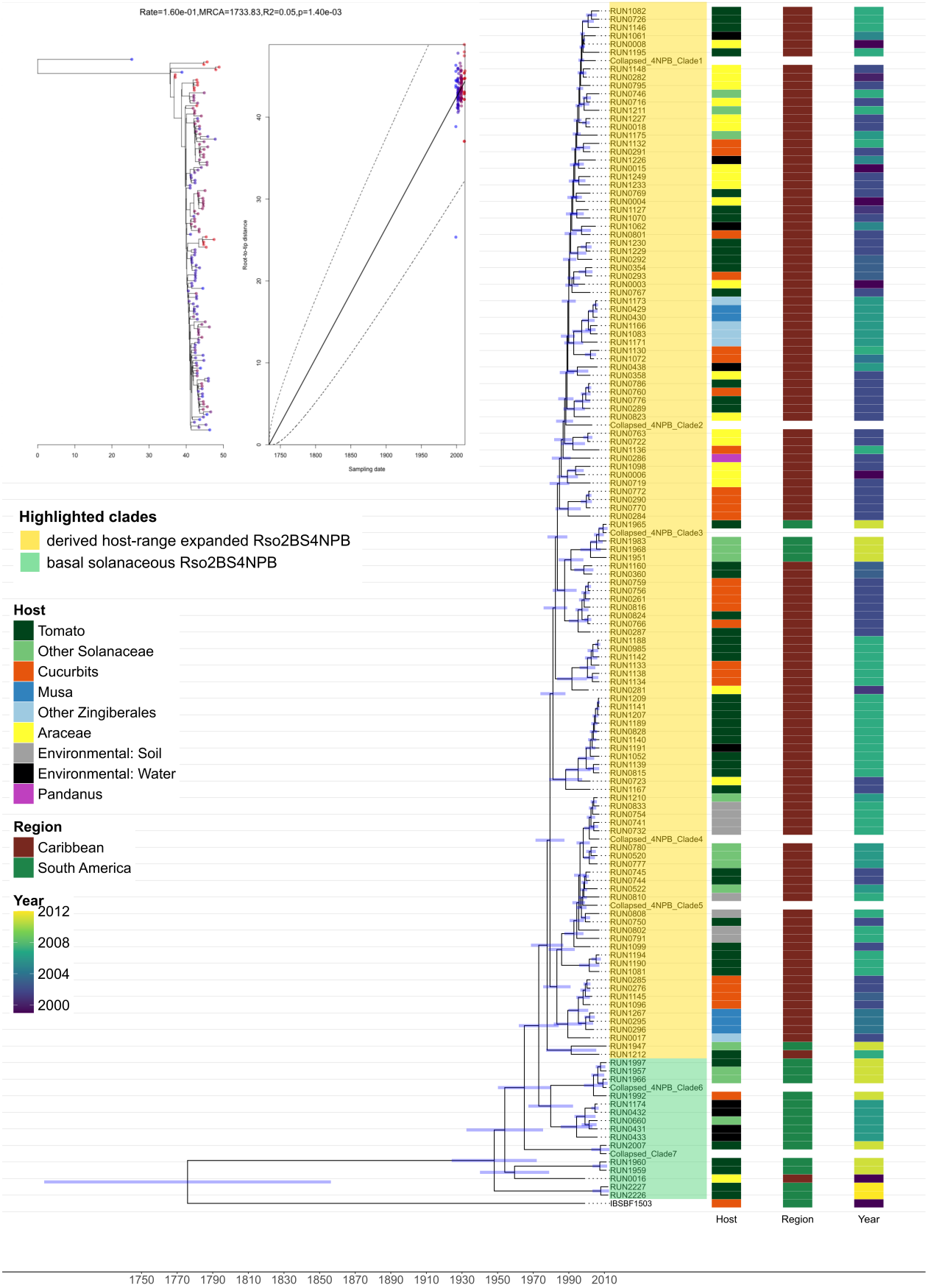
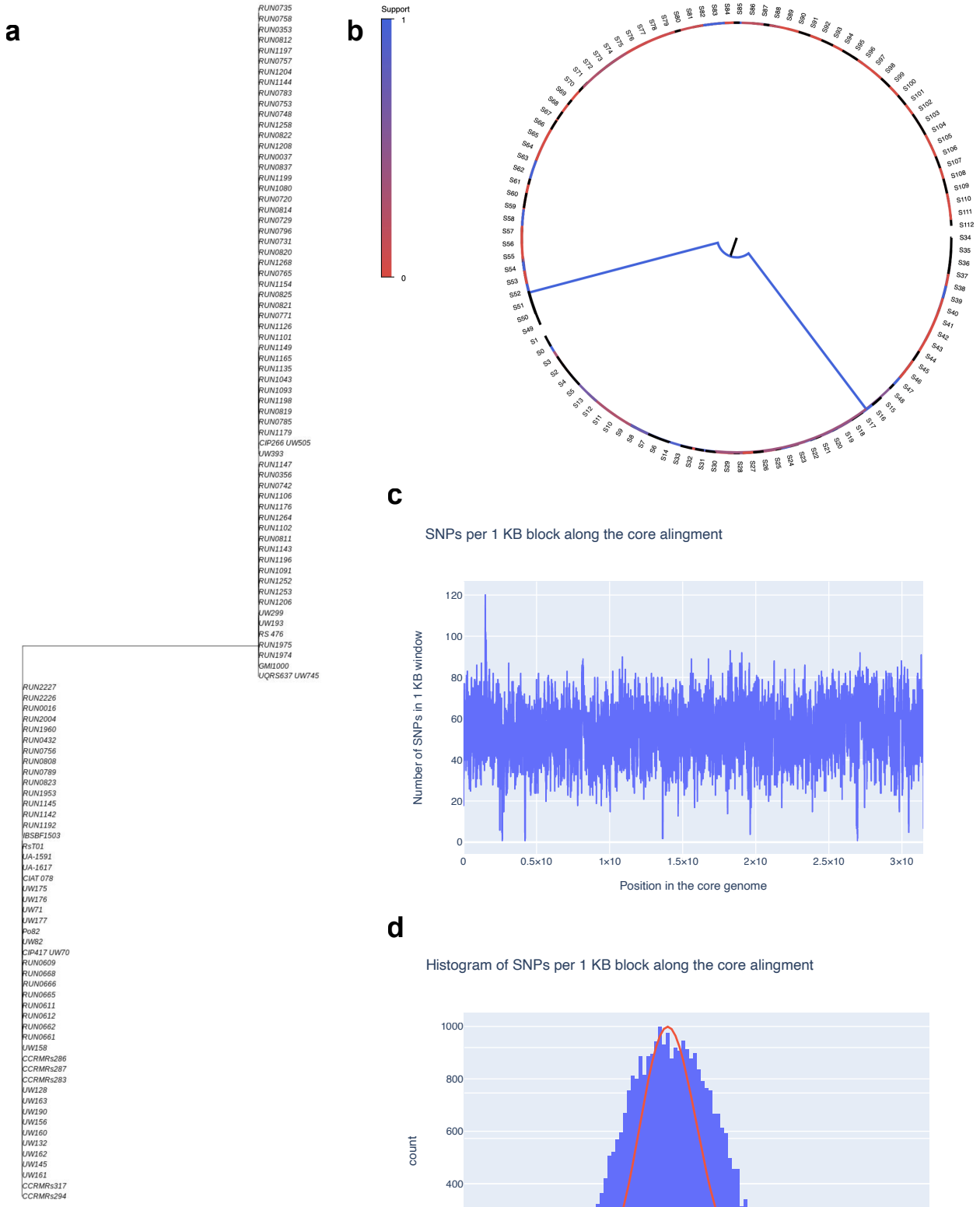


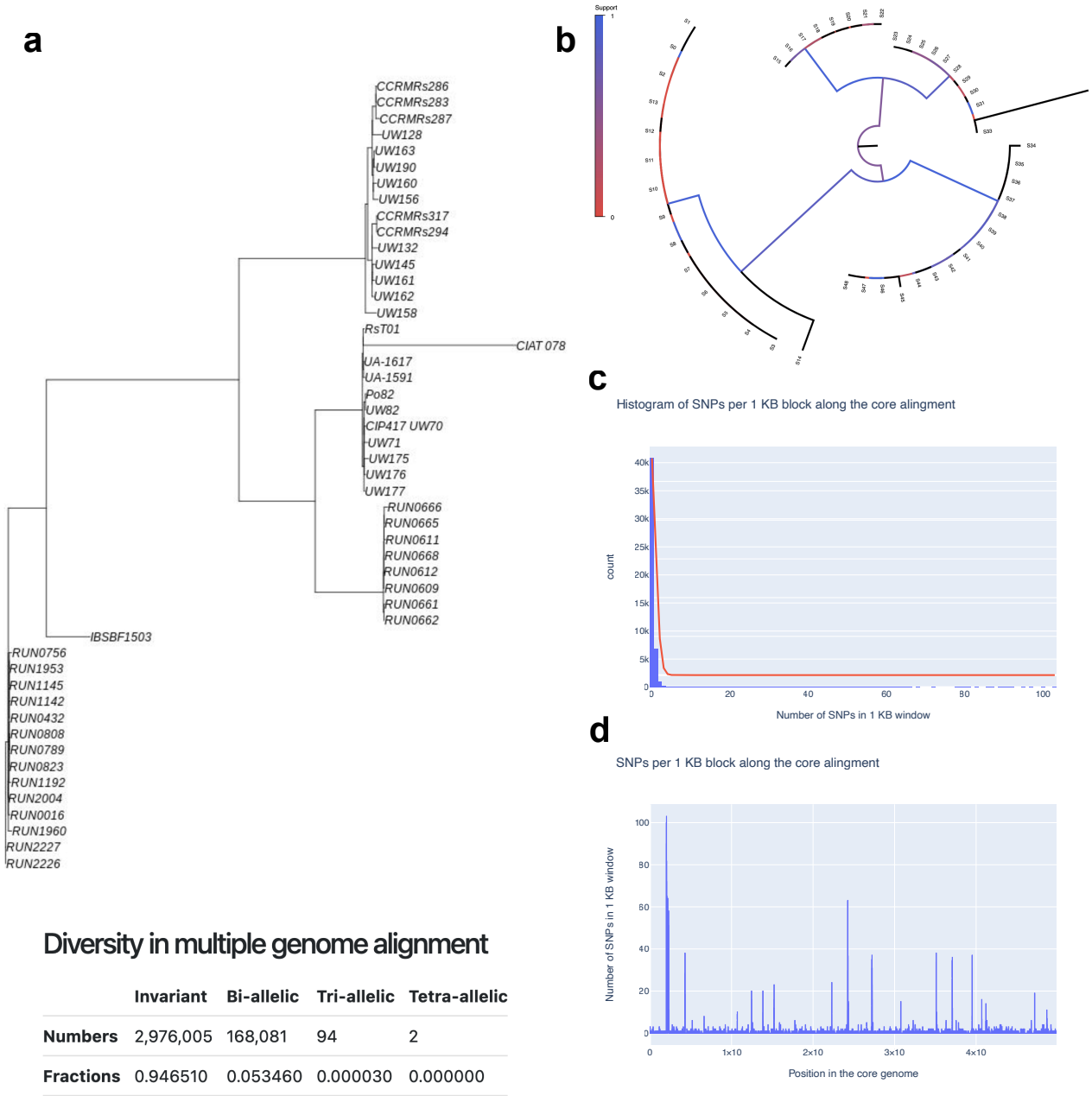
Fig. S6. Divergence dating of Rso2BS4NPB. Maximum likelihood tree of all Rso2BS4NPB with the temporal dating confidence intervals calculated by BactDating shown as blue horizontal lines. The clades collapsed due to zero-branch length are labelled Collapsed_4NPB_Clade1 to Collapsed_4NPB_Clade7. Collapsed_4NPB_Clade1 contains strains RUN0727, RUN1193, RUN1192, RUN1104, RUN1137, RUN1075, RUN1151. Collapsed_4NPB_Clade2 contains strains RUN0005, RUN0739, RUN0762, RUN1121, RUN1105, RUN1094, RUN1046, RUN1064. Collapsed_4NPB_Clade3 contains strains RUN1953, RUN1980, RUN1973. Collapsed_4NPB_Clade4 contains strains RUN0747, RUN0789, RUN0806, RUN0841, RUN0790, RUN0809, RUN0787, RUN0799, RUN0768. Collapsed_4NPB_Clade5 contains strains RUN0788, RUN1069, RUN0836, RUN0813, RUN0827, RUN0840. Collapsed_4NPB_Clade6 contains strains RUN1956, RUN1970, RUN1982, RUN1967. Collapsed_4NPB_Clade7 contains strains RUN2004, RUN2011. Displayed together with root-to-tip regression plots generated by BactDating for the Rso2BS4NPB strains (upper-left panel).



Diversity in multiple genome alignment

	Invariant	Bi-allelic	Tri-allelic	Tetra-allelic
Numbers	4,986,078	2,873	19	2
Fractions	0.999420	0.000580	0.000000	0.000000

Fig. S7. RECOPHY analysis for reduced set of Rso2BS4 and RpsS18 genomes. (a) Core genome phylogenetic tree generated by PhyML. (b) The same tree in circular form with the branches coloured according to the branch support values. (c) Plot representing SNPs counts per 1 kb of core alignment. (d) Histogram of SNPs counts per 1 kb.



Diversity in multiple genome alignment

	Invariant	Bi-allelic	Tri-allelic	Tetra-allelic
Numbers	2,976,005	168,081	94	2
Fractions	0.946510	0.053460	0.000030	0.000000

Fig. S8. RECOPHY analysis for reduced set of Rso2BS4 strains. (a) Core genome phylogenetic tree generated by PhyML. (b) The same tree in circular form with the branches coloured according to the branch support values. (c) Plot representing SNPs counts per 1 kb of core alignment. (d) Histogram of SNPs counts per 1 kb.

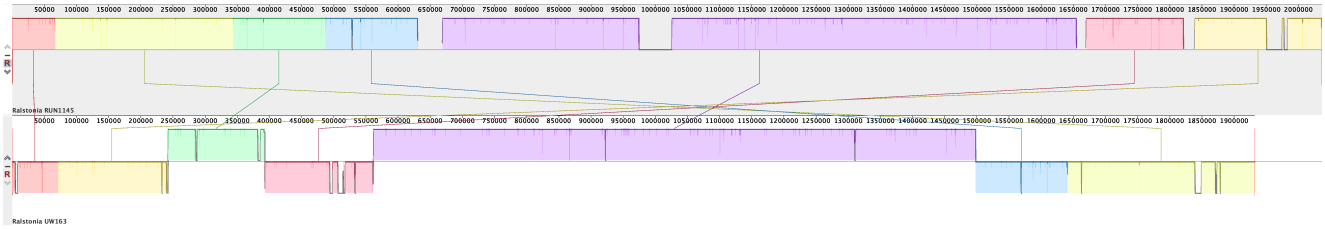


Fig. S9. Alignment of Rso2BS4NPB and Rso2BS4 secondary chromosomes. ProgressiveMauve alignment of the second chromosomes of Rso2BS4NPB RUN1145 and Rso2BS4-UW163 strains.

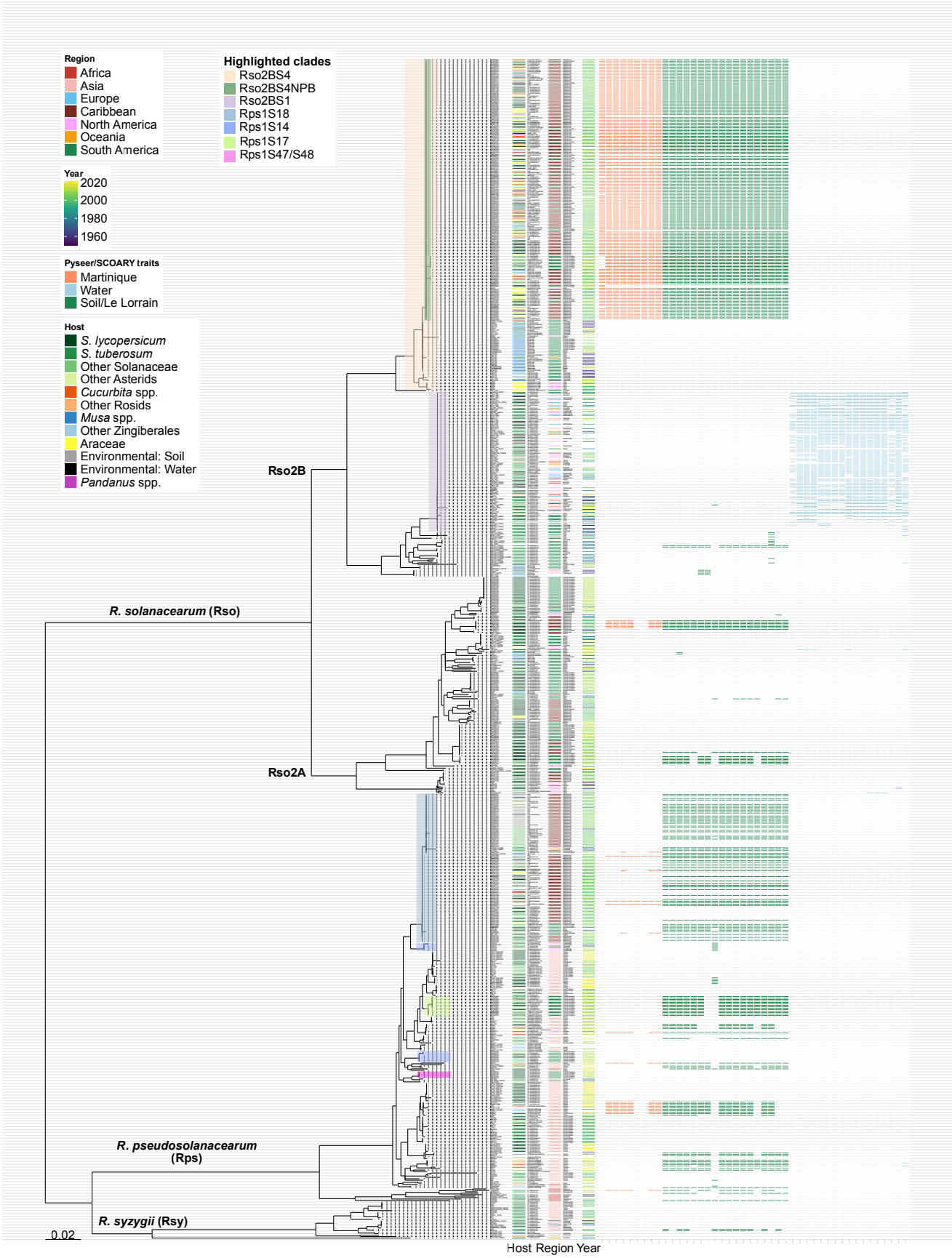


Fig. S10. Complete panGWAS results displayed against RSSC core gene tree. Maximum likelihood tree of all RSSC genomes with the metadata as described in Figure S1. Clades representing Rso and Rps lineages of interest are highlighted in the tree. Labels of the reference strains and the eight strains from Martinique and French Guiana sequenced by long-read technology during this study are shown in bold. Presence and absence patterns of all genes reported by both Scoary and Pyseer panGWAS analysis is shown on the right. The genes found in association with sampling in Martinique and Soil/Le Lorrain traits are shown in orange and green, respectively. The genes found in association with water samples are shown in blue.

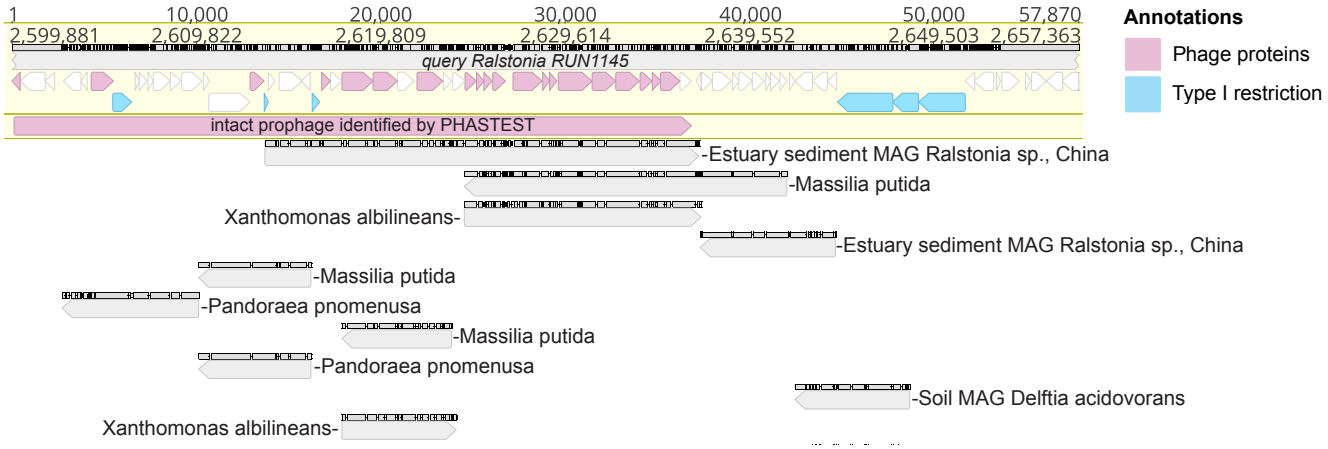


Fig. S11. Rso2BS4NPB-specific prophage BLAST results. Prophage insertion at the 2,6-2,66Mb position within the Rso2BS4NPB RUN1145 genome containing the phage unique for Rso2BS4NPB. Pink arrows represent the CDS annotated as phage proteins: tail, head, phage terminase, and baseplate. CDS annotated as parts of the Type I restriction-modification system are shown in blue. Large pink arrow below CDS designates the position of an intact phage identified by PHASTER. Grey arrows further below indicate the hits reported by BLAST search of the insertion region against the NCBI-nr database.

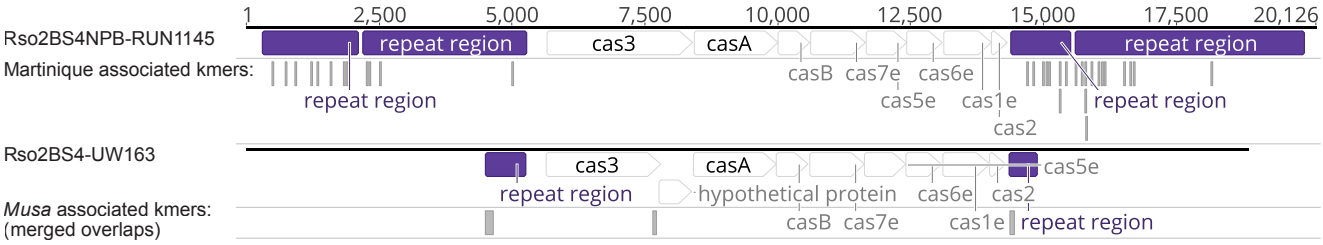


Fig. S12. CRISPR-Cas loci in Rso2BS4 and Rso2BS4NPB. GWAS kmers mapped to CRISPR-Cas loci located on the primary chromosome at the positions 1,119,872-1,139,675 and 2,010,416-2,020,801 in Rso2BS4NPB RUN1145 and Rso2BS4-UW163 strains respectively. CRISPR-repeat regions are highlighted in purple. All kmers aligned to RUN1145 are associated with geographical origin from Martinique. All kmers aligned to UW163 are associated with isolation from *Musa* spp. Because many kmers mapped to UW163 overlap, the merged kmers regions are displayed. RUN1145 kmers are displayed without modifications.

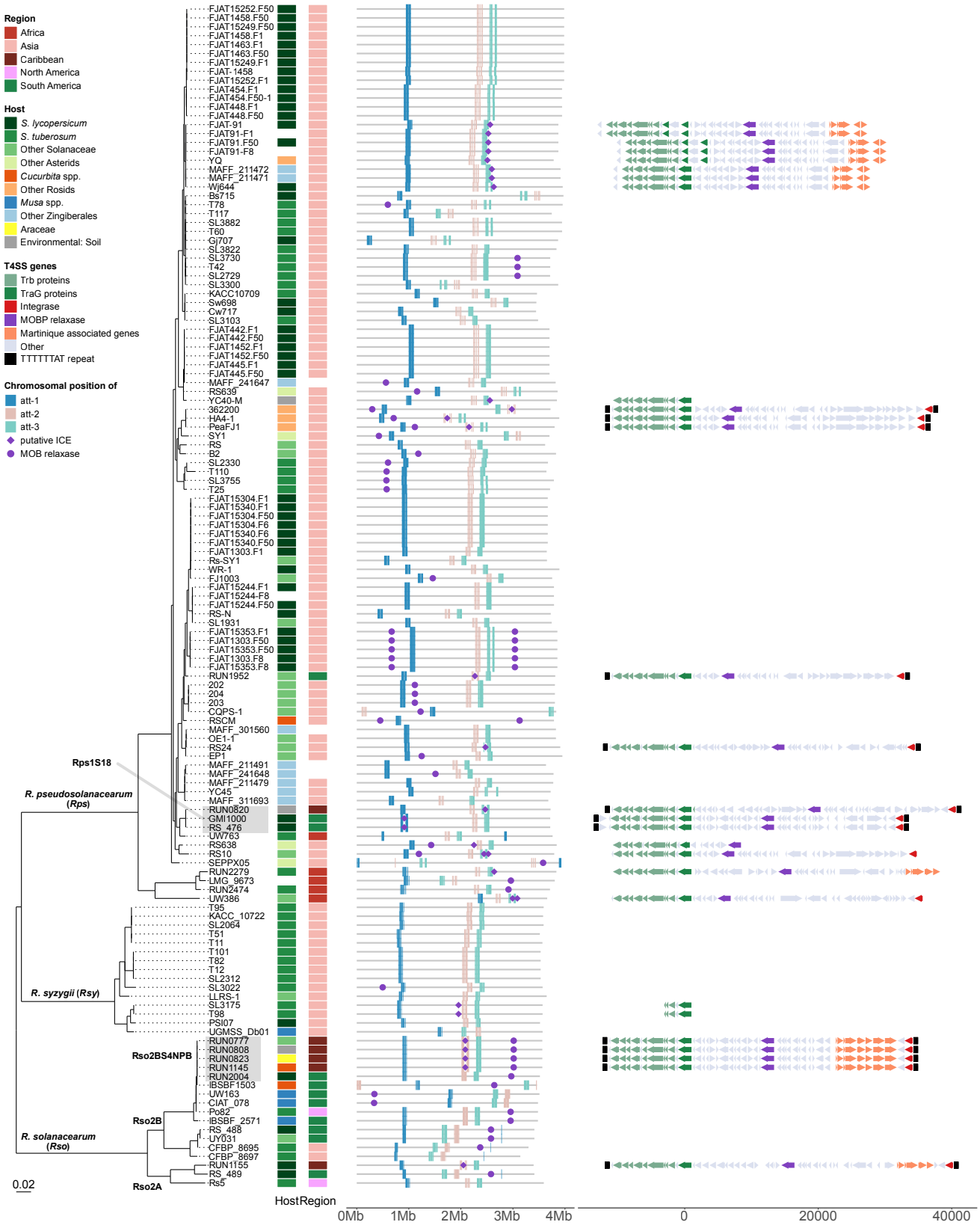


Fig. S13. Distribution of ICEs across the RSSC. Maximum likelihood tree of all completely assembled RSSC genomes available in this study together with the localization of all annotated relaxases and regions syntenic to the ICEs flanking regions, and composition of putative ICEs. The RSSC tree is the same as shown in Fig. 3 though with all nodes displayed, regardless of support.

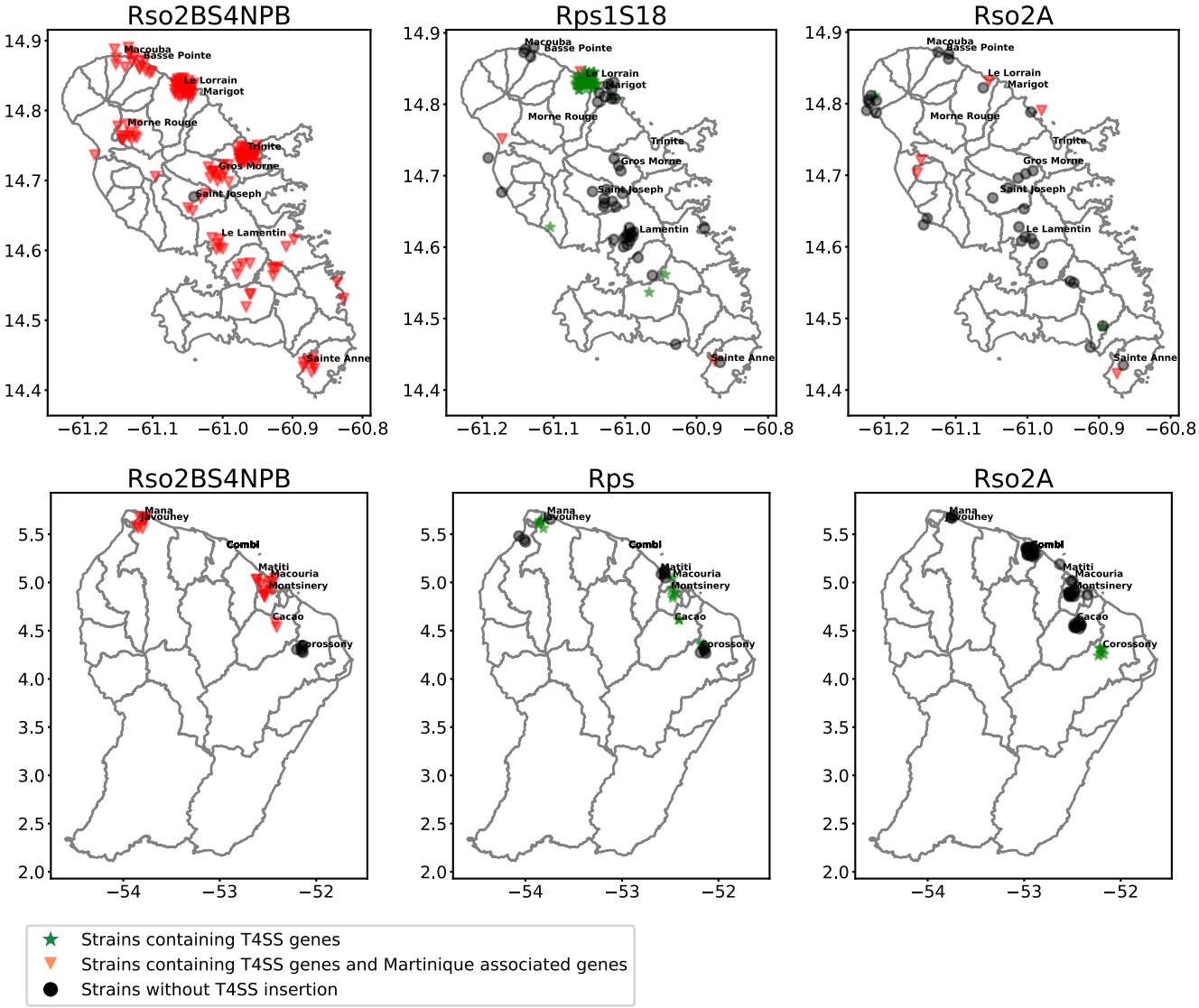


Fig. S14. Geographic distribution of Ralstonia spp. ICEs sampled in Martinique. Sampling sites of all strains collected in Martinique and French Guiana and assembled in this study. The exact longitude and latitude of the sampling locations were jittered for the visualisation purpose. The first and second row of the six panels show Martinique and French Guiana maps respectively with the locations where each of Rso2BS4NPB, Rps and Rso2A strain was sampled. All Rps strains found on Martinique are Rps1S18, while the Rps found in French Guiana belong to other lineages. Strains shown as green star-shaped dots contain T4SS genes homologous to T4SS found in Rso2BS4NPB genomes. Red diamond-shaped dots represent strains that, together with T4SS genes, contain the group of genes found in association with Martinique by panGWAS. Black circles represent the strains without both insertions. Only the geographical locations where more than five strains were sampled are labelled with their names.

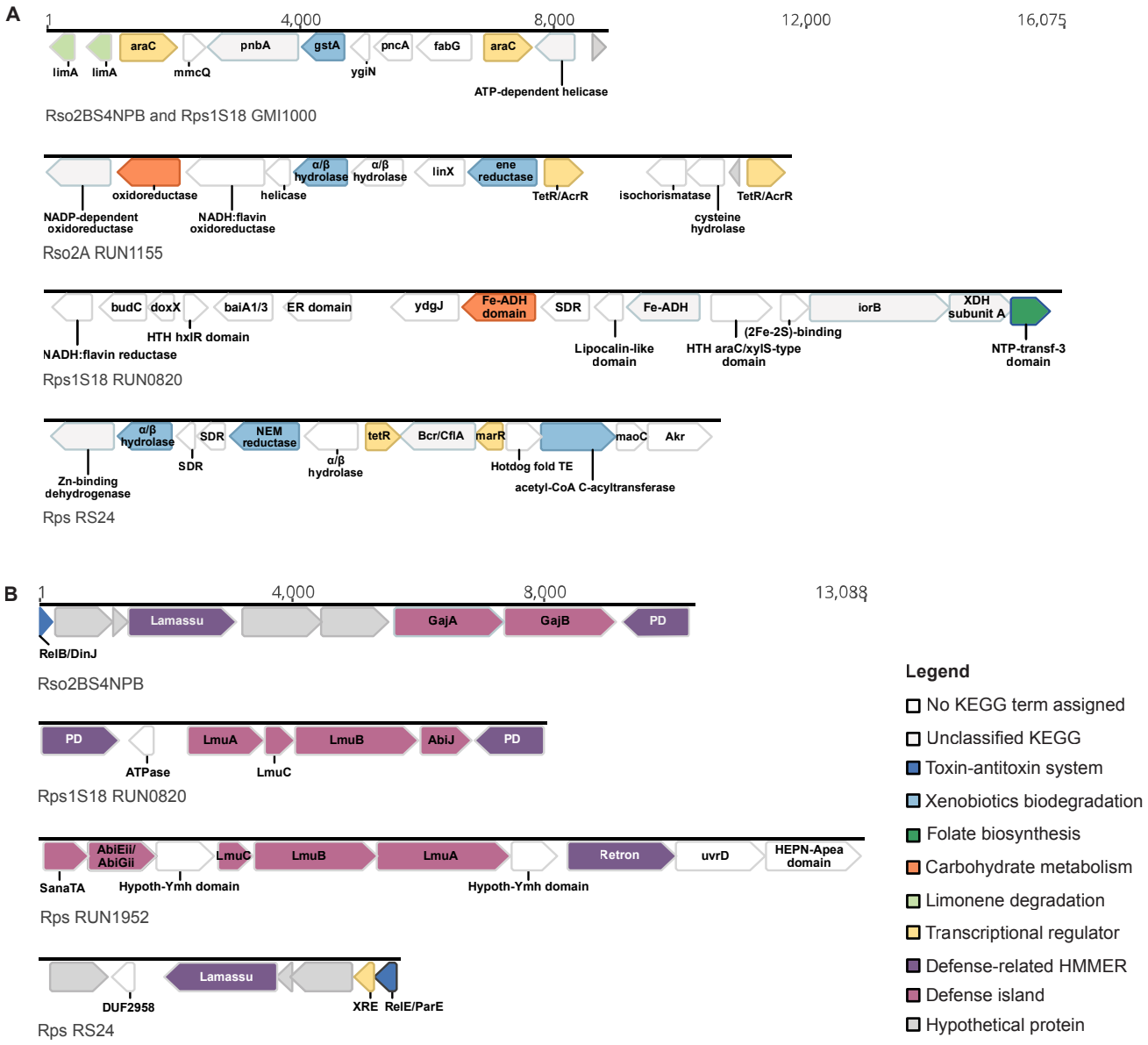


Fig. S15. ICE cargo regions across multiple Rso and Rps strains. (A) Functional annotations within cargo-1 regions of the ICEs found in all Rso2BS4NPB strains, Rps1S18 GMI1000, Rso2A RUN1155, Rps1S18 RUN0820 and Rps-Rs24. Annotations of cargo-1 regions from Rso2BS4NPB and Rps1S18 GMI1000 are identical and shown together. The CDS are coloured according to their KEGG annotation. Proteins assigned to a KO-term not classified as a part of a metabolic pathway are displayed in faint blue. Proteins identified as a part of xenobiotics biodegradation, folate biosynthesis, carbohydrate metabolism or limonene degradation pathways are shown in blue, dark blue, orange and red respectively. Transcriptional regulators and hypothetical proteins are shown in yellow and black respectively. (B) Functional annotations within cargo-2 regions found in all Rso2BS4NPB strains, Rps1S18 RUN0820, Rps RUN1952 and Rps Rs24. CDS are coloured as in (a). Additionally, members of three toxin-antitoxin systems RelE/ParE, AbiEii/AbiGii and RelB/DinJ are shown in pink.

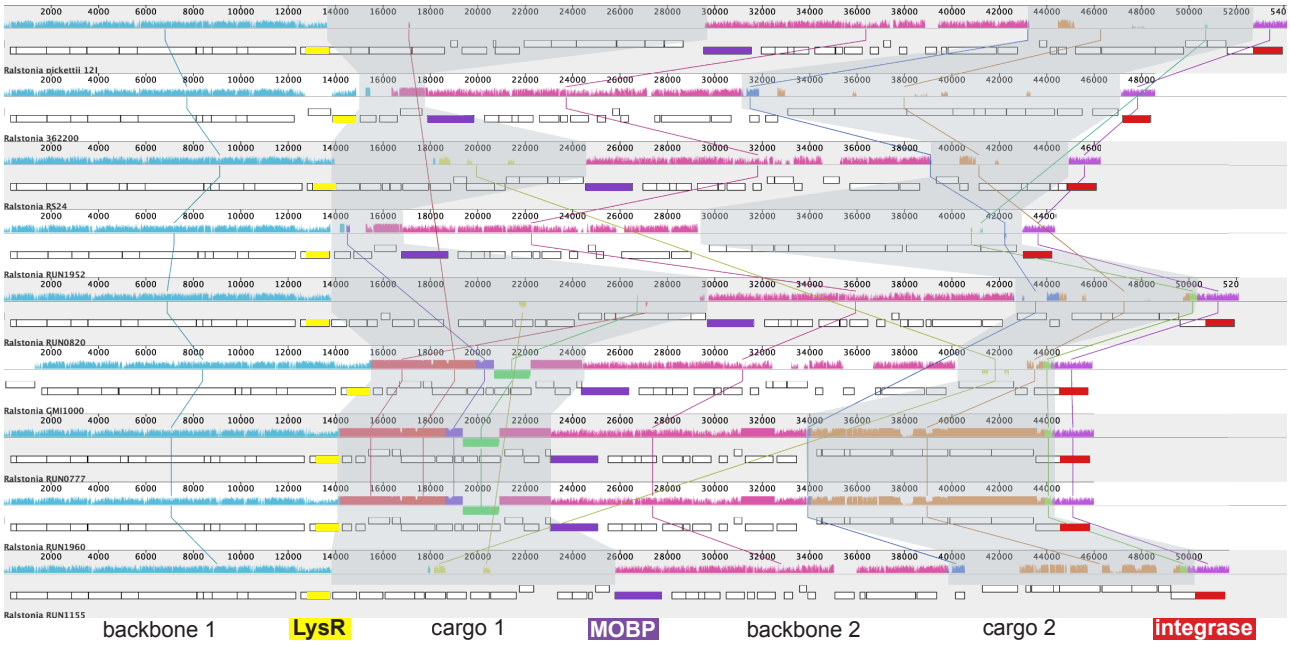


Fig. S16. ICE synteny across multiple Rso and Rps strains. ProgressiveMauve alignment of ICEs identified in Ralstonia pickettii, two Asian Rps strains (RS24 and 362200), three Rso and three Rps strains sampled in Martinique and French Guiana. The first backbone region includes the T4SS and has a LysR family transcriptional regulator at the 3' end, followed by a variable region called cargo-1. The second conserved region (backbone 2) spans the central 8kbp of the ICE and includes the mob relaxase, traF, parA, repA, qseC, rdfS as well as a gene encoding a helix-turn-helix domain-containing protein, a hypothetical protein and various DUF domain-containing proteins (DUF3800, DUF945, DUF2285 (fseA) [139], DUF2840 and DUF2958). A second variable region (cargo-2) is flanked by the integrase at the 3' end of the ICE. Blue and pink coloured LCBs indicate two conserved ICE backbones. Two variable cargo regions are highlighted by grey shades. LysR transcriptional regulator, MOBP family relaxase and integrase are displayed as yellow, purple and red CDSs respectively.