Parallels between experimental and natural evolution of legume symbionts

Running title: Experimental versus natural evolution

Camille Clerissi[1,2,3], Marie Touchon[1,2], Delphine Capela[3], Mingxing Tang[3], Stéphane Cruveiller[4], Matthew A. Parker[5], Lionel Moulin[6], Catherine Masson-Boivin[3,*], Eduardo P.C. Rocha[1,2,*]

1, Microbial Evolutionary Genomics, Institut Pasteur, 28 rue Dr. Roux, 75015, Paris, France.

2, CNRS, UMR3525, 28 rue Dr. Roux, 75015, Paris, France.

3, LIPM, Université de Toulouse, INRA, CNRS, Castanet-Tolosan, France.

4, CEA/FAR, Institut de Génomique, 2 rue Gaston Crémieux, 91057, Evry, France.

5, Department of Biological Sciences, State University of New York, Binghamton, NY 13902, USA.

6, IRD, Cirad, Université de Montpellier, IPME, Montpellier, France

*** Materials & Correspondence**:

Eduardo P.C. Rocha, Microbial Evolutionary Genomics, Institut Pasteur, 25 Rue Dr. Roux, 75724 Paris, France. erocha@pasteur.fr

Catherine Masson-Boivin, LIPM, Université de Toulouse, INRA, CNRS, Castanet-Tolosan, France Catherine.masson@inra.fr

152 words for the abstract and 3626 words for the text (without Methods or legends).

Keywords: comparative genomics, experimental evolution, bacterial evolution, horizontal gene transfer, symbiosis, rhizobia.

Abstract

The emergence of symbiotic interactions has been studied using population genomics in nature and experimental evolution in the laboratory, but the parallels between these processes remain unknown. We compared the emergence of rhizobia after the horizontal transfer of a symbiotic plasmid in natural populations of *Cupriavidus taiwanensis*, over 10 MY ago, with the experimental evolution of symbiotic *Ralstonia solanacearum* for a few hundred generations. In spite of major differences in terms of time-span, environment, genetic background and phenotypic achievement, both processes resulted in rapid diversification dominated by purifying selection concomitant with acquisition of positively selected mutations. The latter were lacking in the plasmid carrying the genes responsible for the ecological transition. Instead, adaptation targeted the same set of genes leading to the co-option of the same quorum-sensing system. Our results provide evidence for similarities in experimental and natural evolutionary transitions and highlight the potential of comparisons between both processes to understand symbiogenesis.

Biological adaptations have traditionally been evaluated by inferring the evolutionary history of organisms from the genomic, morphological, and phenotypic comparison of natural isolates, including fossil records when they were available. Recently, these approaches have been increasingly complemented by experimental evolution studies. The latter can be done on controlled environments and provide nearly complete "fossil" records of past events because individuals from intermediate points in the experiment can be kept for later analysis [1,2]. Sequencing and phenotyping of evolved clones provides crucial information on the mechanisms driving adaptation in simplified environments. Yet, there is little data on the adaptation of lineages in the case of complex adaptations requiring numerous steps and even less on how they recapitulate natural processes (but see [3,4]), raising doubts on the applicability and relevance of experimental evolution studies to understand natural history [5].

Adaptations to new and complex environments, such as ecological transitions towards pathogenic or mutualistic symbiosis, are often initiated by the acquisition via horizontal transfer of genes that provide novel functionalities [6]. For example, the extreme virulence of *Shigella* spp., *Yersinia pestis*, or *Bacillus anthracis* results from the acquisition of plasmid-encoded virulence factors by otherwise poorly virulent clones. These novel genetic systems often require subsequent regulatory rewiring, a process that may take hundreds to millions of years *in natura* [7]. A striking case of horizontal gene transfer (HGT)-mediated transition towards mutualism concerns the rhizobium-legume symbiosis, a symbiosis of major ecological importance that contributes to ca. 25% of the global nitrogen cycling. Rhizobia induce the formation of new organs, the nodules, on the root of legumes, which they colonize intracellularly and in which they fix nitrogen to the benefit of the plant [8]. These symbiotic capacities emerged several times in the natural history of α- and β-Proteobacteria, from the horizontal transfer of the key symbiotic genes into soil free-living bacteria (*i.e.*, the *nod* genes for organ formation and the *nif/fix* genes for nitrogen fixation) [9–11]. This process resulted in hundreds of rhizobial species scattered in 14 known genera, including the genus *Cupriavidus* in β-proteobacteria [12].

Transition towards legume symbiosis has recently been tested at the laboratory time-scale using an experimental system [13]. A plant pathogen was evolved to become a legume symbiont by mimicking the natural evolution of rhizobia at an accelerated pace. First, the plasmid

pRalta^{LMG19424} - encoding the key genes allowing the symbiosis between *C. taiwanensis* LMG19424 [14] and *Mimosa* – was introduced into *Ralstonia solanacearum* GMI1000. The resulting chimera was further evolved under *Mimosa pudica* selective pressure. The chimeric ancestor, which was strictly extracellular and pathogenic on *Arabidopsis thaliana* - but not on *M. pudica* and unable to nodulate it - progressively adapted to become a legume symbiont during serial cycles of inoculation to the plant and subsequent re-isolation from nodules [13,15,16]. Several adaptive mutations driving acquisition and/or drastic improvement of nodulation and infection were previously identified [13,17,18]. Lab-evolution was accelerated by stress-responsive error-prone DNA polymerases encoded in the plasmid that increased the mutation load *ex planta* [19].

Here we compare the natural and experimental evolutions of *Mimosa* symbionts in the *Cupriavidus/Ralstonia* branch using population genomics and functional enrichment analyses. We traced the natural evolutionary history of *Cupriavidus taiwanensis* and provide evidence that, despite significant differences in terms of time frame, protagonists, and environmental context, there were very significant parallels in the two processes.

Results

Diversification of naturally and experimentally evolved *Mimosa* symbionts

We previously generated 18 independent symbiotic lineages of the *R. solanacearum* GMI1000-pRalta^{LMG19424} chimeras that nodulate *M. pudica* [15]. Each lineage was subject to 16 successive cycles of evolution in presence of the plant. We isolated one clone in each of the lineages after the final cycle to identify its genetic and phenotypic differences relative to the ancestor. The symbiotic performances of the evolved clones improved in the experiment with wide variations between lineages. Some clones were able to produce nodules massively and intracellularly infected (Fig. 1A). Yet none of them fixed nitrogen to the benefit of the plant at this stage. In addition to a total of ca. 1200 point mutations relative to the ancestral clones [15], we detected several large deletions in all clones (Fig. 1A). The positions of point mutations were different between lineages, but some genes and many functional categories were affected in parallel (Fig. 1B). In contrast, the deletions showed frequent parallelisms at the nucleotide level. They occurred in homologous regions of the symbiotic plasmid and were systematically flanked by transposable elements that probably mediated their loss by recombination (Table S1).

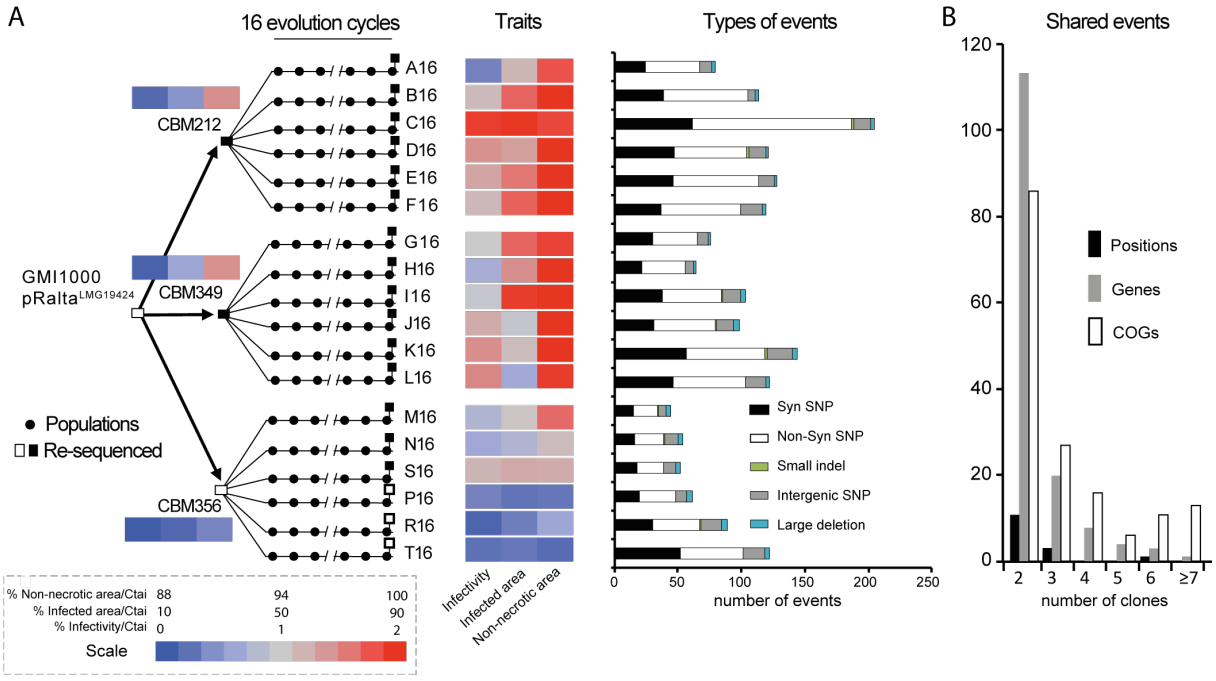


Figure 1. Experimental evolution of *Ralstonia* and associated symbiotic and genomic changes. A. An ancestor chimeric clone evolved to give origin to three clones able to nodulate *M. pudica*. Each of these clones was then evolved in 18 independent lineages using 16 serial

nodulation cycles. This process led to improved infectivity (number of viable bacteria recovered per nodule) and intracellularly-infected area per nodule section and a decrease of necrotic area per nodule section (heatmap on traits). Except clones CBM356, P16, R16 and T16 (white squares), all acquired the ability of intracellular infection (black squares.) The events identified at the end of the 16 evolution cycles for each lineage are indicated on the right (see list of deletions in Table S1 and other mutations in Table S12). **B**. Number of shared events between lineages, *i.e.* the number of positions, genes, and COG categories of genes that were mutated in two or more lineages.

We sequenced, or collected from public databanks, the genomes of 58 *Cupriavidus* strains to study the genetic changes associated with the natural emergence of *Mimosa* symbionts in the genus and to compare them with those observed in the experiment (see supplementary Text S1 and associated tables for data sources, coverage, and details of the results). The phylogeny of the genus core genome was well resolved, showing that 44 out of the 46 genomes with the *nod* and *nif* genes were in the monophyletic *C. taiwanensis* clade (Fig. 2). The two exceptions, strains UYPR2.512 and amp6, were placed afar from this clade in the phylogenetic tree and are clearly distinct species. *C. taiwanensis* strains are *bona fide* symbionts since they fixed nitrogen in symbiosis with *M. pudica* [20,21]. Unexpectedly, the average nucleotide identity (ANIb) values between *C. taiwanensis* strains were often lower than 94%, showing the existence of abundant polymorphism and suggesting that *C. taiwanensis* is not a single species, but a complex of several closely related ones (Fig. 2 and S1, Text S1, Table S2). Together, *C. taiwanensis* strains had a core genome of 3568 protein families and an open pan genome, 3.4 times larger than the average genome. Hence, this complex of species has very diverse gene repertoires.

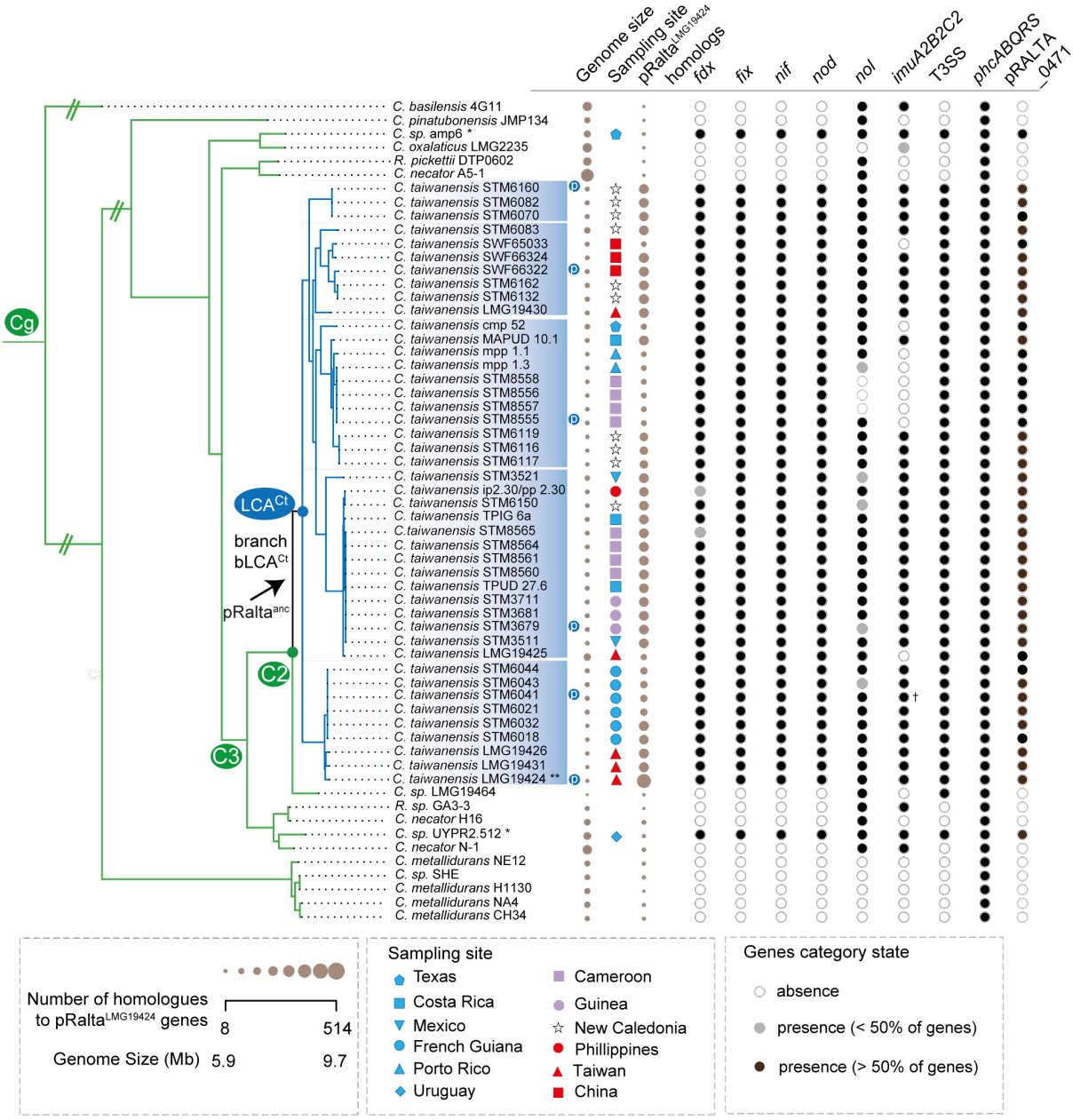


Figure 2. Distribution of symbiotic genes, the mutagenic cassette, T3SS, *imuA2B2C2* and *phcABQRS* within the 60 strains of *Cupriavidus*. See Fig. S1 for the complete tree of the genus *Cupriavidus* and *Ralstonia* without simplifications in branch length. The arrow indicates the most parsimonious scenarios for the acquisition of pRalta (inferred using the MPR function of the ape package in R). This is the branch before the LCA^{Ct}. The node LCA^{Ct} indicates the last common ancestor of *C. taiwanensis*. Circles indicate absence (white), presence of less than 50% of the genes (light grey) and presence of more than 50% of the genes (black). Note that most rhizobia possess the pRalta_0471 gene which is located downstream a *nod* box in LMG19424. The size of the circles for *Genome size* and *pRalta homologs* is proportional to the value of the variable. Sampling sites are coded according to geographic origins. Clusters were computed according to different thresholds of ANIb (as indicated in the text and in Figs. S1 and S8).

Symbols: Ct, C2, C3 and Cg: LCA of clades analyzed in this study. p (in a blue circle): plasmid re-sequenced by PacBio. *: two rhizobia are not part of *C. taiwanensis*. **: *C. taiwanensis* reference strain used as pivot to compute searches of orthologs. † In the PacBio version of this genome *imuABC* is very similar to that of the reference strain, but is encoded in another plasmid.

Parallel patterns of evolution upon the acquisition of the symbiotic plasmid

To compare the initial stages of adaptation in natural populations with those in experimental populations, we searched to identify when the rhizobial character (defined by the presence of the key symbiotic genes *nod* and *nif/fix*), was acquired in the genus *Cupriavidus* (Figs. 2 and S2). The most parsimonious reconstruction of the character in the phylogenetic tree revealed three independent transitions towards symbiosis: in the branch connecting the last common ancestor of *C. taiwanensis* and its immediate ancestor (branch before LCA^{Ct}, hereafter named $bLCA^{Ct}$), and in the terminal branches leading to strains UYPR2.512 and amp6. In agreement with these conclusions, we found very few homologs of the 514 $pRalta^{LMG19424}$ genes in the genomes of UYPR2.512 (8.3 %) or amp6 (6.4%) once the 32 symbiotic genes were excluded from the analysis. These few homologs in the plasmid also showed significantly lower values of sequence similarity than the core genes of the genus ($p<0.01$, Wilcoxon test). We then used birth-death models to identify the acquisitions of genes in the branch $bLCA^{Ct}$ (Fig. 2, Table S3). This analysis highlighted a set of 435 gene acquisitions that were present in $pRalta^{LMG19424}$, over-representing functions such as symbiosis, plasmid biology, and type 4 secretion system (Table S4). These results are consistent with a single initial acquisition of the plasmid in this clade. PacBio resequencing of five strains representative of the main lineages, putative novel species, of *C. taiwanensis* confirmed the ubiquitous presence of a variant of pRalta encoding the symbiotic genes (Table S5). Finally, while most individual *C. taiwanensis* core gene trees showed some level of incongruence with the concatenate core genome tree, an indication of recombination, this frequency was actually lower in the core genes of the plasmid ($p<0.04$, Fisher's exact test). Similarly, there were fewer signals of intragenic recombination in plasmid core genes (PHI, $p<0.001$, same test). This suggests that the plasmid inheritance was mostly vertical within *C. taiwanensis*. We thus concluded that the three rhizobial clades evolved independently and that the acquisition of the ancestral symbiotic plasmid of *C. taiwanensis* should be placed at the branch $bLCA^{Ct}$. The date of plasmid acquisition was estimated using a 16S rRNA clock in the range 12-16 MY ago. Although these dating procedures are only approximate, the values are consistent with the low

ANIb values within *C. taiwanensis* and are posterior to the radiation of its most typical host (*Mimosa* [22]).

Since the experiment only reproduced the initial stages of symbiogenesis, parallels between experimental and natural adaptation should be most striking at the branch $bLCA^{Ct}$, *i.e.*, during the onset of natural evolution towards symbiosis. The evolution experiment showed transient hypermutagenesis caused by the expression of the *imuA2B2C2* plasmid cassette *ex planta* [19]. The long timespan since the acquisition of the plasmid precluded the analysis of accelerated evolution in the branch $bLCA^{Ct}$ (relative to others). Yet, we were able to identify the *imuA2B2C2* cassette in most extant strains, suggesting that they could have played a role in the symbiotic evolution of *Cupriavidus*. We then searched for genes with an excess of recombination or nucleotide diversity in the branch $bLCA^{Ct}$, which revealed 90 recombining genes and 67 genes with an excess of genetic diversity in this branch relative to the *C. taiwanensis* sub-tree (Fig. S3 and Table S3). To identify the parallels between the experimental and natural processes, we identified the 2372 orthologs between the *R. solanacearum* and *C. taiwanensis* (Table S6), and added the 514 pRALTA genes in the chimera as orthologs. Clones of the evolution experiment accumulated significantly more mutations in genes whose orthologs had an excess of polymorphism at the onset of symbiosis in natural populations ($P<0.001$, Fisher's test; Tables S7 and S8), revealing a first parallel between the natural and experimental processes. A second parallel was identified in the overall regimes of natural selection. Both the substitutions in the core genes of *C. taiwanensis* (Fig. S4), and the mutations observed in the experiment [15] showed an excess of synonymous changes relative to the expected ones given the number of non-synonymous mutations. This shows a predominance of purifying selection in both processes, in spite of the observed adaptation towards symbiosis.

Adaptation in the genetic background, not in the symbiotic plasmid

The symbiotic plasmids carry many genes and induce a profound change in the lifestyle of the bacteria. We thus expected to identify changes in the plasmid reflecting its accommodation to the novel genetic background. The plasmid $pRalta^{LMG19424}$ accumulated an excess of synonymous substitutions and a vast majority of the genetic deletions observed in the experiment (Fig. 3 and Table S9). Natural populations also showed more deletions in the plasmid, since from the 413 genes present in $pRalta^{LMG19424}$ and inferred to be present in LCA^{Ct} only 12% were in the core genome, which is 6 times less than found among the

chromosomal genes present in *C. taiwanensis* LMG19424 and inferred to be present in LCA^{Ct} ($p<0.001$, Fisher's exact test, Fig. 3B). The few pRalta ^{LMG19424} core genes are related to the symbiosis or to typical plasmid functions (conjugation) (Fig. 3). The rate of recombination could not be measured on the genomes from the experiments because it is undetectable at this level of sequence similarity between clones (which presumably makes it less important as driver of diversification). The few plasmid core genes show lower recombination rates (PHI and SH analyses,both $p<0.01$) than the chromosomal ones.

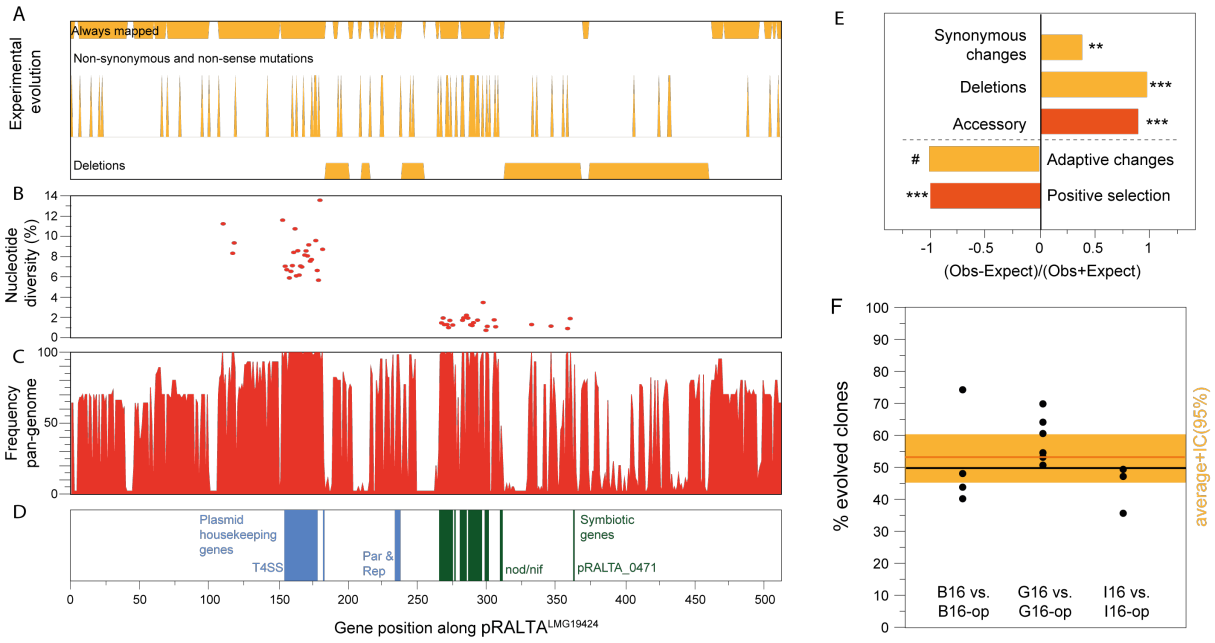


Figure 3. Analysis of the symbiotic plasmid of *Cupriavidus taiwanensis* LMG19424. A. Deletions, non-synonymous and non-sense mutations, and regions of the plasmid that could always be mapped to identify mutations in the experiment **B**. Nucleotide diversity of natural *C. taiwanensis* core genes: symbiotic genes accumulated much less diversity than the other genes. **C**. Frequency of each gene in the 44 *C. taiwanensis* (positional orthologs). **D**. Symbiotic and plasmid housekeeping genes. **E**. Observed over expected values for a number of traits in the plasmid natural (red) or experimental (orange) evolution (Tables S1, S3, and S8). **/*** significantly different from 1 ($P < 0.01/0.001$, Fisher's exact tests for all but the test for "Synonymous changes" which was made by permutations, see Methods and Table S9). # We could not find a single adaptive mutation in the plasmid in our previous works neither in the experiments in panel F. **F**. Impact of pRalta mutations on the *in planta* fitness of evolved clones. *M. pudica* plantlets were co-inoculated with pairs of strains at a 1:1 ratio and nodules were harvested at 21 dpi for bacteria counting. Each pair consisted of an evolved clone (B16, G16 or I16) and the same clone with the evolved pRalta replaced by the original one (B16-op, G16-op or I16-op). The orange horizontal bar represents the average and the large orange

rectangle the 95% interval of confidence of the average (that includes the value 50% indicating that the two types of clones are not significantly different in terms of fitness).

To evaluate whether the observed rapid plasmid diversification was driving the adaptation to symbiosis *in natura*, we compared the rates of positive selection on plasmid and chromosomal genes in *C. taiwanensis*. We identified 325 genes under positive selection in the clade, and 46 specifically in the branch leading to LCA^{Ct} (analysis of 1869 and 1676 core genes lacking evidence of recombination using PHI, respectively, Table S3). Surprisingly, all 325 genes under positive selection were chromosomal (none was found among the core genes of the plasmid, Fig. 3E). In parallel, all mutations previously identified as adaptive in the evolution experiment were chromosomal [13,17,18]. Since our previous analyses of mutations identified in the evolution experiment only focused on strongly adaptive genes, we evaluated the impact of $pRalta^{LMG19424}$ mutations on the symbiotic evolution of *R. solanacearum* by replacing the evolved plasmid with the original $pRalta^{LMG19424}$ in three evolved clones (B16, G16 and I16, thus generating strains B16-op, G16-op and I16-op, respectively). The relative *in planta* fitness of the new chimeras harboring the original plasmid were not significantly different from that of the experimentally evolved clones (Fig. 3F), showing that the adaptation of these strains did not involve mutations in the plasmid. Importantly, the original chimera had similar survival rates with and without the plasmid, suggesting that presence of the plasmid does not impact bacterial fitness in this respect (Tables S10 and S11). Although we cannot exclude that some events of positive selection in the plasmid may have passed undetected, nor that further symbiotic evolution of *R. solanacearum* will involve plasmid mutations, it appears that the genetic changes leading to improvement of the symbiotic traits mainly occurred in the chromosomes of *R. solanacearum* in the experiment, and of *C. taiwanensis* in nature, not on the plasmid carrying the symbiotic traits.

Parallel co-option of regulatory circuits

We identified 436 genes with non-synonymous or non-sense mutations in the experiment (Table S12). This set of genes over-represented virulence factors of *R. solanacearum*, including the T3SS effectors, EPS production, and a set of genes regulating (*phcBQS*) or directly regulated (*prhI*, *hrpG,* and *xpsR*) by the central regulator PhcA of the cell density system that controls virulence and pathogenicity in *R. solanacearum* [23] (Fig. 4A and Table S13). Among them, mutations in the structural T3SS component *hrcV*, or in the virulence

regulators *hrpG, prhI, vsrA*, and *efpR*, were demonstrated to be responsible for the acquisition or the drastic improvement of nodulation and/or infection [13,17,18].

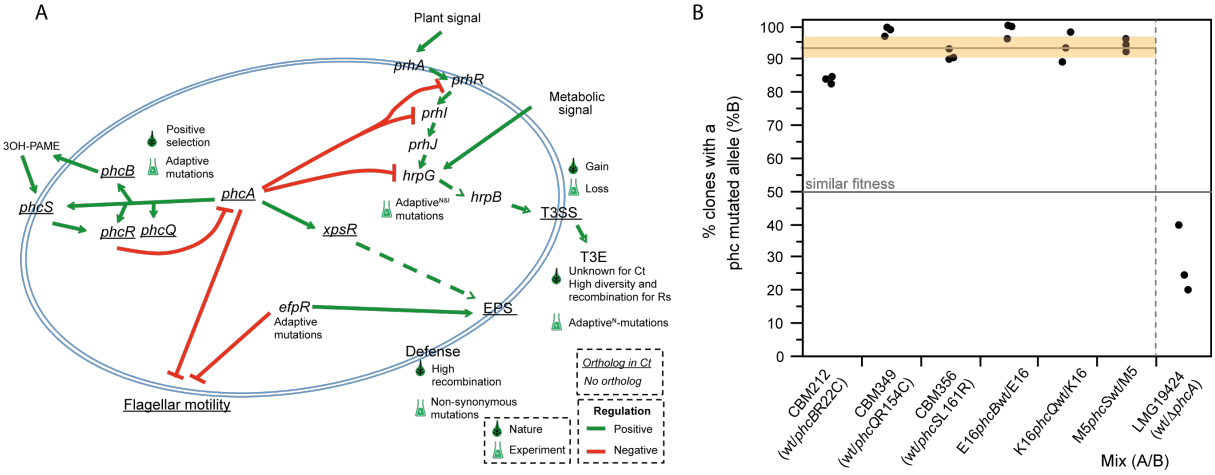


Figure 4. Virulence factors and regulatory pathways of *R. solanacearum* and their evolution in the evolution experiment. **A**. Schema of the major virulence factors and regulatory pathways mentioned in this study and their role in *R. solanacearum* (adapted from [23]). Adaptive^{N} and adaptive^{I}, represent the presence of adaptive mutations for nodulation and infection, respectively. Underline, genes or factors present in *C. taiwanensis*. The results of the enrichment analyses are in Tables S4, S13 and S21. **B**. Adaptive nature of the *phc* alleles evolved in the experiments and the recruitment of PhcA for symbiosis in the natural symbiont *C. taiwanensis* LMG19424. The horizontal grey line represents the average fitness of the evolved *phc* genes relative to the wild-type. The horizontal orange rectangle indicates the 95% interval of confidence for the mean. The results for *phc* are significantly different from the expected under the hypothesis that both variants are equally fit (horizontal line at 50%, $p < 0.005$, Wilcoxon test). The mean for the analysis of the mutant of PhcA (25%) is smaller than 50%, although the difference is at the edge of statistical significance (p=0.0597, two-side t-student test). The codes of the clones correspond to those indicated in Fig. 1.

We first turned our attention to the T3SS because its inactivation was required to activate symbiosis in the evolution experiment, presumably because some T3SS effectors block nodulation and early infection [13]. In contrast, the emergence of legume symbiosis *in natura* seems to be associated with the acquisition of T3SS since all rhizobial *Cupriavidus* strains of our sample encode a (chromosomal) T3SS, while most of the other *Cupriavidus* strains do not (Fig. 2). This apparent contradiction is solved by the fact that we could not find a single ortholog of the 77 T3SS effectors of *R. solanacearum* GMI1000 in *C. taiwanensis* LMG19424. Actually, it has been shown that a functional T3SS is not required for mutualistic

285 symbiosis of the latter with *M. pudica* [24], the only plant species used in the evolution
286 experiment.

287 We then focused on PhcA-associated genes since they accumulated an excess of mutations in
288 the experiment (Table S13). The *phc* system, which was only found intact in *Cupriavidus* and
289 *Ralstonia* (Table S14), regulates a reversible switch between two different physiological
290 states via the repression of the central regulator PhcA in *Ralstonia* [23] and *Cupriavidus* [25].
291 Interestingly, PhcA-associated genes were also enriched in substitutions *in natura*. Indeed, the
292 *phcBQRS* genes of the cell density-sensing system were among the 67 genes that exhibited an
293 excess of nucleotide diversity in the branch $bLCA^{Ct}$ relative to *C. taiwanensis* ("phcA-linked"
294 in Table S4). Strikingly, only seven genes showing an excess of diversity at $bLCA^{Ct}$ had
295 orthologs with mutations in the evolution experiment. Among these seven, only two also
296 showed signature of positive selection in *C. taiwanensis*: *phcB* and *phcS* (ongoing events,
297 Table S3).

298 Given the parallels between experimental and natural evolution regarding an over-
299 representation of changes in PhcA-associated genes, we enquired on the possibility that
300 mutations in the *phcB*, *phcQ* and *phcS* genes, detected in the evolved E16, K16 and M16
301 clones capable of nodule cell infection were adaptive for symbiosis with *M. pudica*. For this,
302 we introduced the mutated alleles of these genes in their respective nodulating ancestors,
303 CBM212, CBM349 and CBM356, and the wild-type allele in the evolved clones E16, K16
304 and M5 (M5 was used instead of M16, since genetic transformation failed in the latter clone
305 in spite of many trials). Competition experiments between the pairs of clones harboring the
306 wild type or the mutant alleles confirmed that these mutations were adaptive (Fig. 4B). The
307 evolved clones also showed better infectivity, since they contained more bacteria per nodule
308 (Fig. S5). On the other hand, we found that the Phc system plays a role in the natural *C.*
309 *taiwanensis*-*M. pudica* symbiosis: a *phcA* deletion mutant had lower nodulation
310 competitiveness than the wild-type *C. taiwanensis* (Fig. 4B), and lower infectiveness (Fig.
311 S6), when both strains were co-inoculated to *M. pudica*. Hence, the re-wiring of the *phc*
312 virulence regulatory pathway of *R. solanacearum* was involved in the evolution of symbiosis
313 in several lineages of the experimental evolution. In parallel, high genetic diversification
314 accompanied by positive selection of the homologous pathway was associated with the
315 transition to symbiosis in the natural evolution of *C. taiwanensis*.

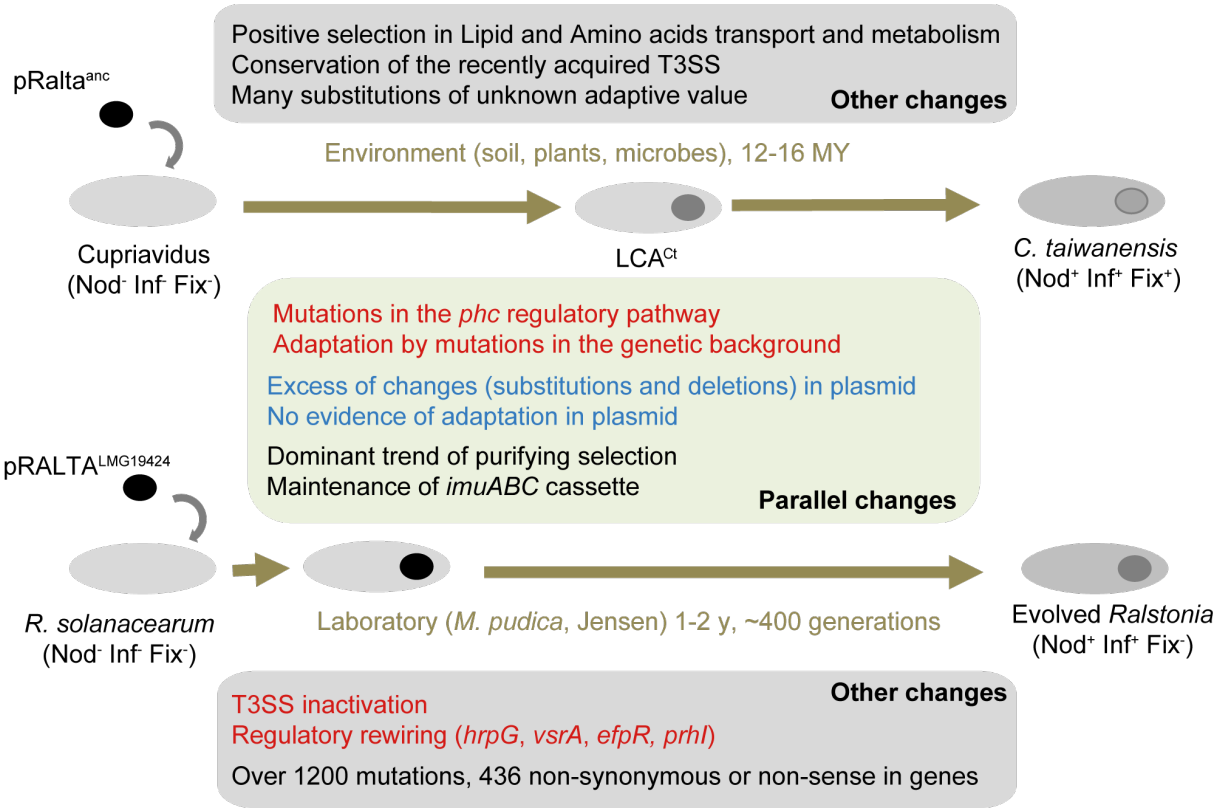
316

Discussion

Years of comparative genomics and loss of function approaches led to propose that most legume symbionts evolved in two-steps [8], *i. e.* acquisition of a set of essential symbiotic genes followed by subsequent adaptation of the resulting genome under plant selection pressure. Although, this evolutionary scenario has recently been validated in the laboratory [13,17], to which extent experimental evolution of symbionts parallels natural symbiogenesis was still unknown. Here, we highlighted several parallels between the experimental and *in natura* transitions towards legume symbiosis (Fig. 5). Such parallels were not necessarily expected, because the two processes differed in a number of fundamental points. The two species are from different genera and had different original lifestyles, saprophytic for *C. taiwanensis* and pathogenic for *R. solanacearum*. The conditions of the experimental evolution were extremely simplified and controlled, whereas natural environmental conditions were certainly very complex and changing. The time span of both processes was radically different, 12-16 MYA in nature, and ca. 400 bacterial generations per lineage in the experiment, providing very different magnitudes of genetic diversity. This precluded the identification of parallelisms at the scale of nucleotide positions due to excessive diversity in natural populations. Lastly, *C. taiwanensis* are well-adapted mutualistic symbionts of *Mimosa* spp., whereas the lab-evolution of *Ralstonia* is not yet achieved, none of the evolved clones being able to persist within nodule cells and fix nitrogen to the benefit of the plant.

The plasmid carrying the essential *nod* and *nif* genes drove the transition towards symbiosis in both processes. We expected that plasmid genes would show evidence of adaptation, either at the level of gene expression regulation or biochemical fine-tuning, to the novel genetic background and environmental conditions. Instead, the abundant substitutions observed in the plasmid seem to have a negligible role in the experiment and lack evidence of positive selection in nature. The cost of the plasmid has also not changed during the experiment. This suggests that the symbiotic genes acquired by *C. taiwanensis* in nature were already - like in the experiment - pre-adapted to establish a symbiotic association with *Mimosa* species. It is in agreement with proposals that pRalta was acquired from *Burkholderia* [26], which are ancient symbionts of *Mimosa* spp. [27]. This also suggests that adaptation following the acquisition of a large plasmid encoding traits driving ecological shifts does not require plasmid evolution. The fact that genetic adaptation to this novel complex trait only occurred in the background is a

348 testimony of the ability of mobile genetic elements to seamlessly plug novel functions in their
349 hosts.



350

351 **Figure 5. Overall similarities and differences between the experimental and natural**
352 **evolutionary processes described in this study.** Adaptive and non-adaptive changes are in
353 orange and blue, respectively.

354 Instead of affecting directly the novel genetic information, adaptive mutations seem to have
355 centered on the rewiring of regulatory modules to inactivate or co-opt native functions for the
356 novel trait. We previously showed that loss of the ability to express the T3SS was strictly
357 necessary for the early transition towards symbiosis in the experiment [13], and that subsequent
358 adaptation favored the re-use of regulatory modules leading to massive metabolic and
359 transcriptomic changes [17]. These phenotypic shifts occurred via mutations targeting regulatory
360 genes specific to *Ralstonia* (*e.g.*, *hrpG, prhI, efpR*, Rsc0965), which finely control the
361 expression of many virulence determinants [23,28,29]. Here, from the analysis of orthologs
362 between *R. solanacearum* and *C. taiwanensis,* we showed that several genes in the *phcBQRS*
363 operon both exhibited significant positive selection in *C. taiwanensis* populations and
364 accumulated adaptive mutations in the evolution experiment. In *R. solanacearum*, these genes
365 control the activity of the global virulence regulator PhcA via a cell density-dependent
366 mechanism [30]. Mutations in these genetic regulators are unlikely to cause adaptation by

367 attenuating the virulence of *Ralstonia*, since the chimeric ancestor is not pathogenic on *M.*
368 *pudica* (and these mutations induced the loss of pathogenicity on *Arabidopsis thaliana* [18]).
369 Since PhcA also plays a role in the natural *C. taiwanensis*-*M. pudica* symbiosis, we speculate
370 that adaptive mutations in the experiment and high diversification in nature on *phc* genes after
371 the acquisition of pRalta may reflect the re-wiring of a quorum-sensing system to sense the
372 environment for cues of when to express the novel mutualistic dialogue with eukaryotes.
373 Further work should determine if some of these mutations resulted in the integration of the
374 gene expression network of the plasmid in the broader network of the cell.

375 Very controlled experimental evolution studies show few similar parallel mutations between
376 replicates and require higher-order analyses at the level of genes, operons or pathways to
377 identify commonalities [31]. Here, the comparison of the natural evolution of *Mimosa* symbionts
378 in the *Cupriavidus* genus and the experimental symbiotic evolution of *Ralstonia* under *M.*
379 *pudica* selection pressure could not reveal parallel changes at the nucleotide level because of
380 the high diversity of natural populations. Yet, it showed that symbiotic adaptation occurred in
381 the recipient genome, with similar population genetic patterns, and involved changes in an
382 homologous central regulatory pathway in both processes. These parallels highlight the
383 potential of research projects integrating population genomics, molecular genetics, and
384 evolution experiment to provide insights on adaptation in nature and in the laboratory.
385 Therefore, experimental evolution appears not only useful to demonstrate the biological
386 plausibility of theoretical models in evolutionary biology, but also to enlighten the natural
387 history of complex adaptation processes.
388

Methods

Dataset for the experimental evolution. We used previously published data on the genomic changes observed in the experimental evolution of the chimera, including 21 bacterial clones (three ancestors and 18 evolved clones) [15]. We analyzed all the synonymous and non-synonymous mutations of each clone from these datasets (Table S12). Large deletions above 1 kb were first listed based on the absence of Illumina reads in these regions, and were then validated by PCR amplification using specific primers listed in Table S15. Primers were designed to amplify either one or several small fragments of the putative deleted regions or the junction of these deletions. All primer pairs were tested on all ancestors and final clones (Table S1).

Mutant construction. The pRalta in evolved *Ralstonia* clones B16, G16 and I16 or their derivatives, was replaced by the wild-type pRalta of *C. taiwanensis* LMG19424 strain as previously described [15], generating B16-op, G16-op and I16-op. Wild-type alleles of the *phcB*, *phcQ* and *phcS* genes and constitutively expressed reporter genes (GFP, mCherry) were introduced into *Ralstonia* evolved clones using the MuGent technique [32]. Briefly, this technique consisted in the co-transformation of two DNA fragments, one fragment carrying a kanamycin resistance cassette together with a gene coding a fluorophore and one unlabelled PCR fragment of *ca.* 6 kb carrying the point mutation to introduce, as previously described [17]. Co-transformants were first selected on kanamycin, then screened by PCR for the presence of the point mutation. M5, which possesses the *phcS* mutation, was used instead of M16 since M16 is no more transformable.

To construct the *phcA* deletion mutant of LMG19424, we used the pGPI-SceI/pDAI-SceI technique previously described [33]. Briefly the regions upstream and downstream *phcA* were amplified with the oCBM3413-3414 and oCBM3415-3416 primer pairs and the Phusion DNA polymerase (Thermo Fisher scientific). The two PCR products were digested with *Xba*I-*Bam*HI and *Bam*HI-*Eco*RI respectively and cloned into the pGPI-SceI plasmid digested by *Xba*I and *Eco*RI. The resulting plasmid was introduced into LMG19424 by triparental mating using the pRK2013 as helper plasmid. Deletion mutant were obtained after introduction of the pDAI-SceI plasmid encoding the I-SceI nuclease. LMG19424 *phcA* deletion mutants were verified by PCR using the oCBM3417-3418 and oCBM3419-3420 primer pairs corresponding to external and internal regions of *phcA*, respectively. Oligonucleotides used in these constructions are listed in Table S16.

Relative *in planta* fitness. *Mimosa pudica* seeds from Australia origin (B&T World Seed, Paguignan, France) were cultivated as previously described [15]. To measure the *in planta* relative fitness, a mix of two strains bearing different antibiotic resistance genes or fluorophores (5.10^5 bacteria of each strain per plant) were inoculated to 20 plants. Nodules were harvested 21 days after inoculation, pooled, surface sterilized and crushed. Dilutions of nodule crushes were spread on selective plates, incubated two days at 28°C, then colonies were counted using a fluorescent stereo zoom microscope V16 (Zeiss) when needed. Three independent experiments were performed for each competition.

Public genome dataset. We collected 13 genomes of *Cupriavidus* spp. (including three rhizobia) and 31 of *Ralstonia* from GenBank RefSeq and the MicroScope platform (http://www.genoscope.cns.fr/agc/microscope/home/) as available in September 2015. We removed the genomes that seemed incomplete or of poor quality, notably those smaller than 5 Mb and with L90>150 (defined as the smallest number of contigs whose cumulated length accounts for 90 % of the genome). All accession numbers are given in Table S17. Genomes of α- and β-Proteobacteria larger than 1 Mb and genomes of phages were downloaded from GenBank RefSeq as available in February 2013.

Sequencing, assembly, and annotation of Illumina data. The genomes of 43 *Mimosa* spp. isolates, a non rhizobial strain of *Cupriavidus* (strain LMG19464) as well as a *C. oxalaticus* strain (LMG2235) (Table S17), were sequenced at the GeT-PlaGe core facility, INRA Toulouse (get.genotoul.fr). DNA-seq libraries were prepared according to Biooscientific's protocol using the Biooscientific PCR free Library Prep Kit. Briefly, DNA was fragmented by sonication, size selection was performed using CLEANNA CleanPCR beads and adaptators were ligated to be sequenced. Library quality was assessed using an Advanced Analytical Fragment Analyser and libraries were quantified by qPCR using the Kapa Library Quantification Kit. DNA-seq experiments were performed on an Illumina HiSeq2000 sequencer using a paired-end read length of 2 x 100 bp with the HiSeq v3 reagent kit (LMG2235 and LMG19431) or on an Illumina MiSeq sequencer using a paired-end read length of 2 x 300 pb with the Illumina MiSeq v3 reagent kit (other strains). On average, genomes contained 99 contigs and an L90 of 29.

Genome assemblies were performed with the AMALGAM assembly pipeline (Automated MicrobiAL Genome AsseMbler; Cruveiller S. and Séjourné M., unpublished). The pipeline is a python script (v2.7.x and onward) that launches the various parts of the analysis and checks

that all tasks are completed without error. To date AMALGAM embeds SPAdes, ABySS [34], IDBA-UD [35], Canu [36], and Newbler [37]. After the assembly step, an attempt to fill scaffolds/contigs gaps is performed using the gapcloser software from the SOAPdenovo2 package [38]. Only one gap filling round was performed since launching gapcloser iteratively may lead to an over-correction of the final assembly. AMALGAM ends with the generation of a scaffolds/contigs file (fasta format) and a file describing the assembly in agp format (v2.0).

The genomes were subsequently processed by the MicroScope pipeline for complete structural and functional annotation [39]. Gene prediction was performed using the AMIGene software [40] and the microbial gene finding program Prodigal [41] known for its capability to locate the translation initiation site with great accuracy. The RNAmmer [42] and tRNAscan-SE [43] programs were used to predict rRNA and tRNA-encoding genes, respectively. Genome sequence and annotation was made publicly available (see accession numbers in Table S17).

PacBio sequencing. Library preparation and sequencing were performed according to the manufacturer's instructions "Shared protocol-20kb Template Preparation Using BluePippin Size Selection system (15kb-size cutoff)". At each step DNA was quantified using the Qubit dsDNA HS Assay Kit (Life Technologies). DNA purity was tested using the nanodrop (Thermofisher) and size distribution and degradation assessed using the Fragment analyzer (AATI) High Sensitivity DNA Fragment Analysis Kit. Purification steps were performed using 0.45X AMPure PB beads (Pacbio). 10µg of DNA was purified then sheared at 40kb using the meraruptor system (diagenode). A DNA and END damage repair step was performed on 5µg of sample. Then blunt hairpin adapters were ligated to the library. The library was treated with an exonuclease cocktail to digest unligated DNA fragments. A size selection step using a 13-15kb cutoff was performed on the BluePippin Size Selection system (Sage Science) with the 0.75% agarose cassettes, Marker S1 high Pass 15-20kb.

Conditioned Sequencing Primer V2 was annealed to the size-selected SMRTbell. The annealed library was then bound to the P6-C4 polymerase using a ratio of polymerase to SMRTbell at 10:1. Then after a magnetic bead-loading step (OCPW), SMRTbell libraries were sequenced on RSII instrument at 0.2nM with a 360 min movie. One SMRTcell was used for sequencing each library. Sequencing results were validated and provided by the Integrated next generation sequencing storage and processing environment NG6 accessible in the genomic core facility website [44].

485 **Core genomes.** Core genomes were computed using reciprocal best hits (hereafter named
486 RBH), using end-gap free Needleman-Wunsch global alignment, between the proteome of *C.*
487 *taiwanensis* LMG19424 or *R. solanacearum* GMI1000 (when the previous was not in the sub-
488 clade) as a pivot (indicated by ** on Fig. S1A) and each of the other 88 proteomes [45]. Hits
489 with less than 40 % similarity in amino acid sequence or more than a third of difference in
490 protein length were discarded. The lists of orthologs were filtered using positional
491 information. Positional orthologs were defined as RBH adjacent to at least two other pairs of
492 RBH within a neighbourhood of ten genes (five up- and five down-stream). We made several
493 sets of core genomes (see Fig. S1A): all the 89 strains (A1), 44 *C. taiwanensis* (Ct), Ct with
494 the closest outgroup (C2), Ct with the five closest outgroups (C3), the whole 60 genomes of
495 the genus *Cupriavidus* (Cg), and the 14 genomes of *R. solanacearum* (Rs). They were defined
496 as the intersection of the lists of positional orthologs between the relevant pairs of genomes
497 and the pivot (Table S18).

498 **Pan genomes.** Pan genomes describe the full complement of genes in a clade and were
499 computed by clustering homologous proteins in gene families. Putative homologs between
500 pairs of genomes were determined with blastp v2.2.18 (80 % coverage), and evalues (if
501 smaller than 10^{-4}) were used to infer protein families using SiLiX (v1.2.8, http://lbbe.univ-
502 lyon1.fr/SiLiX) [46]. To decrease the number of paralogs in pan genomes, we defined a minimal
503 identity threshold between homologs for each set. For this, we built the distribution of
504 identities for the positional orthologs of core genomes between the pivot and the most distant
505 genome in the set (Fig. S7), and defined an appropriate threshold in order to include nearly all
506 core genes but few paralogs (Table S19).

507 **Alignment and phylogenetic analyses.** Multiple alignments were performed on protein
508 sequences using Muscle v3.8.31 [47], and back-translated to DNA. We analyzed how the
509 concatenated alignment of core genes fitted different models of protein or DNA evolution
510 using IQ-TREE v1.3.8 [48]. The best model was determined using the Bayesian information
511 criterion (BIC). Maximum likelihood trees were then computed with IQ-TREE v1.3.8 using
512 the appropriate model, and validated via a ultrafast bootstrap procedure with 1000 replicates
513 [49] (Table S18). The maximum likelihood trees of each set of core genes were computed with
514 IQ-TREE v1.3.8 using the best model obtained for the concatenated multiple alignment.

515 In order to root the phylogeny based on core genes, we first built a tree using 16S rRNA
516 sequences of the genomes of *Ralstonia* and *Cupriavidus* genera analysed in this study and of

ten outgroup genomes of β-Proteobacteria. For this, we made a multiple alignment of the 16S sequences with INFERNAL v.1.1 (with default parameter) [50] using RF00177 Rfam model (v.12.1) [51], followed by manual correction with SEAVIEW to removed poorly aligned regions. The tree was computed by maximum likelihood with IQ-TREE using the best model (GTR+I+G4), and validated via an ultrafast bootstrap procedure with 1000 replicates.

To date the acquisition of the symbiotic plasmid in the branch $bLCA^{Ct}$, we computed the distances in the 16S rDNA tree between each strain and each of the nodes delimitating the branch $bLCA^{Ct}$ (respectively LCA^{Ct} and C2 in Fig. 2). The substitution rate of 16S in enterobacteria was estimated at ~1% per 50 MY of divergence [52], and we used this value as a reference.

Orthologs and pseudogenes of symbiotic genes, the mutagenic cassette, T3SS and PhcABQRS. We identified the positional orthologs of Cg for symbiotic genes, the mutagenic cassette, T3SS, and PhcABQRS using RBH and *C. taiwanensis* LMG19424 as a pivot (such as defined above). These analyses identify *bona fide* orthologs in most cases (especially within species), and provide a solid basis for phylogenetic analyses. However, they may miss genes that evolve fast, change location following genome rearrangements, or that are affected by sequence assembling (incomplete genes, small contigs without gene context, etc.). They also miss pseudogenes. Hence, we used a complementary approach to analyze in detail the genes of the symbiotic island in the plasmid, the mutagenic cassette, T3SS and PhcABQRS. Indeed, we searched for homologs of each gene in the reference genome in the other genomes using LAST v744 [53] and a score penalty of 15 for frameshifts. We discarded hits with evalues below 10^{-5}, with less than 40 % similarity in sequence, or aligning less than 50 % of the query. In order to remove most paralogs, we plotted values of similarity and patristic distances between the 59 *Cupriavidus* and the reference strain *C. taiwanensis* LMG19424 for each gene. We then manually refined the annotation using this analysis.

Evolution of gene families. We used Count (version downloaded in December 2015) [54] to study the past history of transfer, loss and duplication of the protein families of the pan genomes. The analysis was done using the core genomes reference phylogenies. We tested different models of gene content evolution using the tree of Cg (Table S18), and selected the best model using the Akaike information criterion (AIC) (Table S19). We computed the posterior probabilities for the state of the gene family repertoire at inner nodes with maximum likelihood and used a probability cutoff of 0.5 to infer the dynamics of gene families, notably

549 presence, gain, loss, reduction, and expansion for the branch leading to the last common
550 ancestor (LCA) of *C. taiwanensis* (LCA^{Ct}).

551 **Measures of similarity between genomes**. For each pair of genomes, we computed two
552 measures of similarity, one based on gene repertoires and another based on the sequence
553 similarity between two genomes. The gene repertoire relatedness (GRR) was computed as the
554 number of positional orthologs shared by two genomes divided by the number of genes in the
555 smallest one [55]. Pairwise average nucleotide identities (ANIb) were calculated using the pyani
556 Python3 module (https://github.com/widdowquinn/pyani), with default parameters [56]. We
557 used single-linkage clustering to group strains likely to belong to the same species. This was
558 done constructing a transitive closure of sequences with an ANIb higher than a particular
559 threshold (*i.e.*, >94%, 95% or 96%). We used BioLayout Express[3D] to visualize the graphs
560 representing the ANIb relationships and the resulting groups for each threshold (Fig. S8).

561 **Inference of recombination.** We identified recombination events using three different
562 approaches. We used the pairwise homoplasy index (PHI) test to look for incongruence
563 within each core gene multiple alignment (Ct and C3 datasets). We made 10,000 permutations
564 to assess the statistical significance of the results [57]. We used the SH-test, as implemented in
565 IQ-TREE v1.3.8 [48] (GTR+I+G4 model, 1000 RELL replicates), to identify incongruence
566 between the trees of each core gene and the concatenated multiple alignment of all core genes.
567 We used ClonalFrameML v10.7.5 [58] to infer recombination and mutational events in the
568 branch leading to the LCA^{Ct} using the phylogenetic tree of C3 (Table S18). The
569 transition/transversion ratios given as a parameter to ClonalFrameML were estimated with the
570 R package PopGenome v2.1.6 [59]. Lastly, ClonalFrameML was also used to compare the
571 relative frequency of recombination and mutation on the whole concatenated alignments of Ct
572 and Rs.

573 **Molecular diversity and adaptation.** Positive selection was identified using likelihood ratio
574 tests by comparing the M7 (beta) - M8 (beta&ω) models of codeml using PAML v4.8 [60]. We
575 used the independent phylogenetic tree of each gene family to avoid problems associated with
576 horizontal transfer (since many genes failed the SH-test for congruence with the core genome
577 phylogenetic tree). We removed from the analysis gene families that had incongruent
578 phylogenetic signals within the multiple alignment [61]. These correspond to the families for
579 which PHI identified evidence of recombination ($p < 0.05$).

We inferred the mutations arising in the branch leading to LCA^{Ct} using the phylogenetic tree build with the core genome of C3 (Ct and the five closest outgroups). First, we used ClonalFrameML to reconstruct the ancestral sequences of LCA^{Ct} and LCA^{C2} (accounting for recombination). Then, we estimated nucleotide diversity of each core gene for Ct, and between LCA^{Ct} and LCA^{C2} using the R package pegas. Finally, we used the branch-site model of codeml to identify positive selection on this branch for the core genes of C3 that lacked evidence of intragenic recombination (detected using PHI).

To infer the extent of purifying selection for Ct, we computed dN/dS values for each core genes between *C. taiwanensis* LMG19424 and the others strains of Ct using the yn00 model of PAML v4.8. We then plotted the average dN/dS of each strains with the patristic distances obtained from the tree of the concatenated multiple alignment of all core genes.

Functional annotations. We searched for the functions over-represented relative to a number of characteristics (recombination, nucleotide diversity, etc.). We analyzed COG categories, protein localizations, transporters, regulatory proteins and several pre-defined lists of genes of interest in relation to rhizobial symbiosis and to virulence.

We used COGnitor [62] as available on the MicroScope Platform (https//:www.genoscope.cns.fr/agc/microscope/home/) to class genes according to the COG categories (Tables S3 and S8). Protein subcellular localizations were predicted using PSORTb v3.0.2 (http://www.psort.org/psortb/) [63]. Transporters and regulatory proteins were inferred using TransportDB (http://www.membranetransport.org/) [64] and P2RP (http://www.p2rp.org/) [65], respectively. Protein secretion systems were identified using TXSScan (http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::txsscan) [66]. We manually checked and corrected the lists. Specific annotations were also defined for (i) *R. solanacearum* GMI1000: Type III effectors [67], PhcA-associated genes (*i.e.*, genes involved in the upstream regulatory cascade controlling the expression of phcA, and genes directly controlled by PhcA) [23,68,69], virulence [70], extracellular polysaccharides (EPS) [69,71], chemotaxis [72], twin-arginine translocation pathway (Tat) [73], Tat-secreted protein [74], and (ii) the pRalta of *C. taiwanensis* LMG19424: symbiotic genes [14], genes pertaining to plasmid biology (conjugation, replication, partition, based on the annotations [75]), and operons using ProOpDB (http://operons.ibt.unam.mx/OperonPredictor/) [76]. Lastly, we also annotated positional orthologs between *R. solanacearum* GMI1000 and *C. taiwanensis* LMG19424 according to specific annotations used for both strains (Tables S3 and S8).

Analysis of the mutations observed in the experimental evolution. To estimate differences between mutation rates on the three replicons of the chimera, we compared the observed number of synonymous mutations in each replicon to those obtained from simulations of genome evolution. First, we analyzed the distribution of synonymous mutations of the 18 final evolved clones in regions of the genome that were covered by sequencing data (some regions with repeats cannot be analyzed without ambiguity in the assignment of mutations). We built the mutation spectrum of the genome using these synonymous mutations, since they are expected to be the least affected by selection. Second, we performed 999 random experiments of genome evolution using the mutation spectrum and the total number of synonymous mutations obtained for the 18 final clones. With the results, we draw the distributions of the expected number of synonymous mutations in each replicon (under the null hypothesis that they occurred randomly). This data was then used to define intervals of confidence around the average values observed in the simulations.

Statistical analyses. In order to identify genes that evolved faster in the branch leading to LCA^{Ct}, we compared the nucleotide diversity of sequences for LCA^{Ct} and LCA^{C2} with those of the extant 44 *C. taiwanensis* using a regression analysis. Outliers above the regression line were identified using a one-sided prediction interval ($p < 0.001$) as implemented in JMP (JMP®, Version 10. SAS Institute Inc., Cary, NC, 1989-2007).

We computed functional enrichment analyses to identify categories over-represented in a focal set relative to a reference dataset. The categories that were used are listed above in the section *Functional annotations*. To account for the association of certain genes to multiple functional categories, enrichments were assessed by resampling without replacement the appropriated reference dataset (see Table S20) to draw out the expected null distribution for each category. More precisely, we made 999 random samples of the number of genes obtained for each analysis (positive selection, recombination, etc.) in the reference dataset. For each category, we then compared the observed value (in the focal set) to the expected distribution (in the reference dataset) to compute a p-value based on the number of random samples of the reference dataset that showed higher number of genes from the category.

We also compared the nucleotide diversity between sets of genes using the nonparametric Wilcoxon rank sum test ({stats}, wilcox.test).

Finally, we computed Fisher's exact tests (R package {stats}, fisher.test) to estimate the association between results of the natural and the experimental evolution, *i.e.*, to test whether

644 mutations found in the experimental evolution targeted genes that were found to be
645 significantly more diverse in the natural process.

646 P-values were corrected for multiple comparisons using Benjamini and Hochberg's method [77]
647 ({stats}, p.adjust).

648 Statistical analyses with R were done using version 3.1.3 (R: a language and environment for
649 statistical computing, 2008; R Development Core Team, R Foundation for Statistical
650 Computing, Vienna, Austria [http://www.R-project.org]).

651

652 **Acknowledgements**

653 This work was supported by funds from the French National research Agency (ANR-12-
654 ADAP-0014-01 and ANR-16-CE20-0011-01) the "Laboratoire d'Excellence (LABEX)"
655 TULIP (ANR-10-LABX-41) and France Génomique National infrastructure, funded as part of
656 "Investissement d'avenir" program managed by Agence Nationale pour la Recherche (contrat
657 ANR-10-INBS-09).

658 We thank Olaya Rendueles, Rémi Peyraud, Ludovic Cottret, Stéphane Genin, and Pedro
659 Couto Oliveira for helpful comments and suggestions. We thank Eddy Ngonkeu and Moussa
660 Diabate for help with the strain collection.

661

662 **Contributions**

663 CC, CM, and ER conceived the project, integrated the analyses, and wrote the draft of the
664 manuscript. CC, MTouchon, and ER made the computational analyses. DC and MTang
665 performed the experiments and analyzed the data. LM and MAP provided strains and data.
666 SC assembled and annotated the genomes. All authors contributed to the final text.

667

References

1. Kawecki, T. J. *et al.* Experimental evolution. *Trends Ecol. Evol.* **27,** 547–560 (2012).
2. Barrick, J. E. & Lenski, R. E. Genome dynamics during experimental evolution. *Nat. Rev. Genet.* **14,** 827–839 (2013).
3. Metzger, B. P. H., Yuan, D. C., Gruber, J. D., Duveau, F. & Wittkopp, P. J. Selection on noise constrains variation in a eukaryotic promoter. *Nature* **521,** 344–347 (2015).
4. Maddamsetti, R. *et al.* Synonymous genetic variation in natural isolates of *Escherichia coli* does not predict where synonymous substitutions occur in a long-term experiment. *Mol. Biol. Evol.* **32,** 2897–2904 (2015).
5. Bailey, S. F. & Bataillon, T. Can the experimental evolution programme help us elucidate the genetic basis of adaptation in nature? *Mol. Ecol.* **25,** 203–218 (2016).
6. Ochman, H. & Moran, N. A. Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science* **292,** 1096–1099 (2001).
7. Lercher, M. J. & Pal, C. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Mol. Biol. Evol.* **25,** 559–567 (2008).
8. Masson-Boivin, C., Giraud, E., Perret, X. & Batut, J. Establishing nitrogen-fixing symbiosis with legumes: how many rhizobium recipes? *Trends Microbiol.* **17,** 458–466 (2009).
9. Sullivan, J. T., Patrick, H. N., Lowther, W. L., Scott, D. B. & Ronson, C. W. Nodulating strains of *Rhizobium loti* arise through chromosomal symbiotic gene transfer in the environment. *Proc. Natl. Acad. Sci.* **92,** 8985–8989 (1995).
10. Moulin, L., Béna, G., Boivin-Masson, C. & Stępkowski, T. Phylogenetic analyses of symbiotic nodulation genes support vertical and lateral gene co-transfer within the *Bradyrhizobium* genus. *Mol. Phylogenet. Evol.* **30,** 720–732 (2004).

11. Nandasena, K. G., O'Hara, G. W., Tiwari, R. P. & Howieson, J. G. Rapid *in situ* evolution of nodulating strains for *Biserrula pelecinus* L. through lateral transfer of a symbiosis island from the original mesorhizobial inoculant. *Appl. Environ. Microbiol.* **72,** 7365–7367 (2006).

12. Remigi, P., Zhu, J., Young, J. P. W. & Masson-Boivin, C. Symbiosis within symbiosis: evolving nitrogen-fixing legume symbionts. *Trends Microbiol.* **24,** 63–75 (2016).

13. Marchetti, M. *et al.* Experimental evolution of a plant pathogen into a legume symbiont. *PLoS Biol.* **8,** e1000280 (2010).

14. Amadou, C. *et al.* Genome sequence of the -rhizobium *Cupriavidus taiwanensis* and comparative genomics of rhizobia. *Genome Res.* **18,** 1472–1483 (2008).

15. Marchetti, M. *et al.* Experimental evolution of rhizobia may lead to either extra- or intracellular symbiotic adaptation depending on the selection regime. *Mol. Ecol.* **26,** 1818–1831 (2017).

16. Marchetti, M. *et al.* Shaping bacterial symbiosis with legumes by experimental evolution. *Mol. Plant. Microbe Interact.* **27,** 956–964 (2014).

17. Capela, D. *et al.* Recruitment of a lineage-specific virulence regulatory pathway promotes intracellular infection by a plant pathogen experimentally evolved into a legume symbiont. *Mol. Biol. Evol.* **34,** 2503–2521 (2017).

18. Guan, S. H. *et al.* Experimental evolution of nodule intracellular infection in legume symbionts. *ISME J.* **7,** 1367–1377 (2013).

19. Remigi, P. *et al.* Transient hypermutagenesis accelerates the evolution of legume endosymbionts following horizontal gene transfer. *PLoS Biol* **12,** e1001942 (2014).

20. Klonowska, A. *et al.* Biodiversity of *Mimosa pudica* rhizobial symbionts (*Cupriavidus taiwanensis*, *Rhizobium mesoamericanum*) in New Caledonia and their adaptation to heavy metal-rich soils. *FEMS Microbiol. Ecol.* **81,** 618–635 (2012).

21. Chen, W.-M. *et al.* Legume symbiotic nitrogen fixation by beta-proteobacteria is widespread in nature. *J. Bacteriol.* **185,** 7266–7272 (2003).

22. Simon, M. F. *et al.* Recent assembly of the Cerrado, a neotropical plant diversity hotspot, by in situ evolution of adaptations to fire. *Proc. Natl. Acad. Sci.* **106,** 20359–20364 (2009).

23. Genin, S. & Denny, T. P. Pathogenomics of the *Ralstonia solanacearum* species complex. *Annu. Rev. Phytopathol.* **50,** 67–89 (2012).

24. Saad, M. M., Crevecoeur, M., Masson-Boivin, C. & Perret, X. The type 3 protein secretion system of *Cupriavidus taiwanensis* strain LMG19424 compromises symbiosis with *Leucaena leucocephala*. *Appl. Environ. Microbiol.* **78,** 7476–7479 (2012).

25. Garg, R. P. *et al.* Evidence that *Ralstonia eutropha* (*Alcaligenes eutrophus*) contains a functional homologue of the *Ralstonia solanacearum* Phc cell density sensing system. *Mol. Microbiol.* **38,** 359–367 (2000).

26. Mishra, R. P. N. *et al.* Genetic diversity of *Mimosa pudica* rhizobial symbionts in soils of French Guiana: investigating the origin and diversity of *Burkholderia phymatum* and other beta-rhizobia. *FEMS Microbiol. Ecol.* **79,** 487–503 (2012).

27. Bontemps, C. *et al. Burkholderia* species are ancient symbionts of legumes. *Mol. Ecol.* **19,** 44–52 (2010).

28. Peeters, N., Guidot, A., Vailleau, F. & Valls, M. *Ralstonia solanacearum*, a widespread bacterial plant pathogen in the post-genomic era. *Mol. Plant Pathol.* **14,** 651–662 (2013).

29. Perrier, A. *et al.* Enhanced in planta fitness through adaptive mutations in EfpR, a dual regulator of virulence and metabolic functions in the plant pathogen *Ralstonia solanacearum*. *PLOS Pathog.* **12,** e1006044 (2016).

30. Clough, S. J., Lee, K.-E., Schell, M. A. & Denny, T. P. A two-component system in *Ralstonia* (*Pseudomonas*) *solanacearum* modulates production of PhcA-regulated

742 virulence factors in response to 3-hydroxypalmitic acid methyl ester. *J. Bacteriol.* **179,**
743 3639–3648 (1997).
744 31. Tenaillon, O. *et al.* The molecular diversity of adaptive convergence. *Science* **335,** 457–
745 461 (2012).
746 32. Dalia, A. B., McDonough, E. & Camilli, A. Multiplex genome editing by natural
747 transformation. *Proc. Natl. Acad. Sci.* **111,** 8937–8942 (2014).
748 33. Daubech, B. *et al.* Spatio-temporal control of mutualism in legumes helps spread
749 symbiotic nitrogen fixation. *eLife* **6,** e28683 (2017).
750 34. Simpson, J. T. *et al.* ABySS: A parallel assembler for short read sequence data. *Genome*
751 *Res.* **19,** 1117–1123 (2009).
752 35. Peng, Y., Leung, H. C., Yiu, S.-M. & Chin, F. Y. IDBA-UD: a de novo assembler for
753 single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*
754 **28,** 1420–1428 (2012).
755 36. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer
756 weighting and repeat separation. *bioRxiv* 071282 (2017).
757 37. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre
758 reactors. *Nature* **437,** 376–380 (2005).
759 38. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de
760 novo assembler. *Gigascience* **1,** 18 (2012).
761 39. Vallenet, D. *et al.* MicroScope-an integrated microbial resource for the curation and
762 comparative analysis of genomic and metabolic data. *Nucleic Acids Res.* **41,** D636–D647
763 (2013).
764 40. Bocs, S. AMIGene: Annotation of MIcrobial Genes. *Nucleic Acids Res.* **31,** 3723–3726
765 (2003).

766 41. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site
767 identification. *BMC Bioinformatics* **11,** 119 (2010).
768 42. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes.
769 *Nucleic Acids Res.* **35,** 3100–3108 (2007).
770 43. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer
771 RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).
772 44. Mariette, J. *et al.* NG6: Integrated next generation sequencing storage and processing
773 environment. *BMC Genomics* **13,** 462 (2012).
774 45. Rocha, E. P. C. Inference and analysis of the relative stability of bacterial chromosomes.
775 *Mol. Biol. Evol.* **23,** 513–522 (2005).
776 46. Miele, V., Penel, S. & Duret, L. Ultra-fast sequence clustering from similarity networks
777 with SiLiX. *BMC Bioinformatics* **12,** 116 (2011).
778 47. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high
779 throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
780 48. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and
781 effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol.*
782 *Evol.* **32,** 268–274 (2015).
783 49. Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. Ultrafast approximation for
784 phylogenetic bootstrap. *Mol. Biol. Evol.* **30,** 1188–1195 (2013).
785 50. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches.
786 *Bioinformatics* **29,** 2933–2935 (2013).
787 51. Nawrocki, E. P. *et al.* Rfam 12.0: updates to the RNA families database. *Nucleic Acids*
788 *Res.* **43,** D130–D137 (2015).
789 52. Ochman, H. & Wilson, A. C. Evolution in Bacteria: evidence for a universal substitution
790 rate in cellular genomes. *J. Mol. Evol.* **26,** 74–86 (1987).

53. Sheetlin, S. L., Park, Y., Frith, M. C. & Spouge, J. L. Frameshift alignment: statistics and post-genomic applications. *Bioinformatics* **30,** 3575–3582 (2014).

54. Csuros, M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* **26,** 1910–1912 (2010).

55. Snel, B., Bork, P. & Huynen, M. A. Genome phylogeny based on gene content. *Nat. Genet.* **21,** 108–110 (1999).

56. Richter, M. & Rosselló-Móra, R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci.* **106,** 19126–19131 (2009).

57. Bruen, T. C. A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172,** 2665–2681 (2005).

58. Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLOS Comput. Biol.* **11,** e1004041 (2015).

59. Pfeifer, B., Wittelsburger, U., Ramos-Onsins, S. E. & Lercher, M. J. PopGenome: an efficient swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.* **31,** 1929–1936 (2014).

60. Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A.-M. K. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155,** 431–449 (2000).

61. Anisimova, M., Nielsen, R. & Yang, Z. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* **164,** 1229–1236 (2003).

62. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278,** 631–637 (1997).

63. Yu, N. Y. *et al.* PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* **26,** 1608–1615 (2010).

64. Ren, Q., Chen, K. & Paulsen, I. T. TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Res.* **35,** D274–D279 (2007).

65. Barakat, M., Ortet, P. & Whitworth, D. E. P2RP: a web-based framework for the identification and analysis of regulatory proteins in prokaryotic genomes. *BMC Genomics* **14,** 269 (2013).

66. Abby, S. S. *et al.* Identification of protein secretion systems in bacterial genomes. *Sci. Rep.* **6,** 23080 (2016).

67. Peeters, N. *et al.* Repertoire, unified nomenclature and evolution of the Type III effector gene set in the *Ralstonia solanacearum* species complex. *BMC Genomics* **14,** 1 (2013).

68. Yoshimochi, T., Hikichi, Y., Kiba, A. & Ohnishi, K. The global virulence regulator PhcA negatively controls the *Ralstonia solanacearum hrp* regulatory cascade by repressing expression of the PrhIR signaling proteins. *J. Bacteriol.* **191,** 3424–3428 (2009).

69. Huang, J., Yindeeyoungyeon, W., Garg, R. P., Denny, T. P. & Schell, M. A. Joint transcriptional control of xpsR, the unusual signal integrator of the *Ralstonia solanacearum* virulence gene regulatory network, by a response regulator and a LysR-type transcriptional activator. *J. Bacteriol.* **180,** 2736–2743 (1998).

70. Brumbley, S. M. & Denny, T. P. Cloning of wild-type *Pseudomonas solanacearum* phcA, a gene that when mutated alters expression of multiple traits that contribute to virulence. *J. Bacteriol.* **172,** 5677–5685 (1990).

71. Huang, J., Carney, B. F., Denny, T. P., Weissinger, A. K. & Schell, M. A. A complex network regulates expression of eps and other virulence genes of *Pseudomonas solanacearum*. *J. Bacteriol.* **177,** 1259–1267 (1995).

72. Yao, J. & Allen, C. Chemotaxis Is required for virulence and competitive fitness of the bacterial wilt pathogen *Ralstonia solanacearum*. *J. Bacteriol.* **188,** 3697–3708 (2006).

841 73. Poueymiro, M. & Genin, S. Secreted proteins from *Ralstonia solanacearum*: a hundred
842 tricks to kill a plant. *Curr. Opin. Microbiol.* **12,** 44–52 (2009).
843 74. Gonzalez, E. T., Brown, D. G., Swanson, J. K. & Allen, C. Using the *Ralstonia*
844 *solanacearum* Tat secretome to identify bacterial wilt virulence factors. *Appl. Environ.*
845 *Microbiol.* **73,** 3779–3786 (2007).
846 75. Cury, J., Touchon, M. & Rocha, E. P. C. Integrative and conjugative elements and their
847 hosts: composition, distribution and organization. *Nucleic Acids Res.* **45,** 8943–8956
848 (2017).
849 76. Taboada, B., Ciria, R., Martinez-Guerrero, C. E. & Merino, E. ProOpDB: Prokaryotic
850 Operon DataBase. *Nucleic Acids Res.* **40,** D627–D631 (2012).
851 77. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and
852 powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 289–300 (1995).
853