Parallels between experimental and natural evolution of legume symbionts

Camille Clerissi1,2,3, Marie Touchon1,2, Delphine Capela3, Mingxing Tang3, Stéphane Cruveiller4, Clémence Genthon5, Céline Lopez-Roques5, Matthew A. Parker6, Lionel Moulin7, Catherine Masson-Boivin3 & Eduardo P.C. Rocha1,2

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. Here we compare the emergence of rhizobia after the horizontal transfer of a symbiotic plasmid in natural populations of \textit{Cupriavidus taiwanensis}, over 10 MY ago, with the experimental evolution of symbiotic \textit{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 genetic diversification dominated by purifying selection. We observe no adaptation in the plasmid carrying the genes responsible for the ecological transition. Instead, adaptation was associated with positive selection in a set of genes that led to the co-option of the same quorum-sensing system in both processes. 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 can be studied using genomic or phenotypic comparisons of natural isolates, including fossil records when they are available, as well as experimental and population analyses of fitness variation. 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. Sequencing and phenotyping of evolved clones provides crucial information on the mechanisms driving adaptation in simplified environments. Yet, there are little data on the adaptation of lineages when the process is complex (requires numerous steps). There is even less data on how these experiences recapitulate natural processes (but see refs. 3,4), raising doubts on the applicability and relevance of experimental evolution studies to understand natural history.

Many descriptions of adaptations involving ecological transitions towards pathogenic or mutualistic symbiosis include an initial acquisition via horizontal transfer of genes that provide novel functionalities. For example, the extreme virulence of Shigella spp., Yersinia spp., or Bacillus anthracis results from the acquisition of plasmid-encoded virulence factors by otherwise poorly virulent clones. Adaptation is often coupled with the genetic rewiring of the recipient genome, a process that may take hundreds to millions of years in nature, and may require specific genetic backgrounds. A striking case of transition mediated by horizontal gene transfer towards mutualism concerns the rhizobium-legume symbiosis, a symbiosis of major ecological importance that contributes to 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.

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), and were further shaped under plant selection pressure. Indeed, legumes have developed control mechanisms that allow the selection of most compatible and beneficial symbionts.

There are now hundreds of known rhizobial species scattered in 14 known genera, including the genus Cupriavidus in β-proteobacteria.

Transition towards legume symbiosis has recently been tested at the laboratory time-scale using an experimental system. A plant pathogen was evolved to become a legume symbiont by mimicking the natural evolution of rhizobia at an accelerated pace. First, the plasmid pRaltaLMG19424—encoding the key genes allowing the symbiosis between Cupriavidus taiwanensis LMG19424 and Mimosa—was introduced into Ralstonia solanacearum GMI11000. 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. Several adaptive mutations driving acquisition and/or drastic improvement of nodulation and infection were previously identified. Lab-evolution was accelerated by stress-responsive error-prone DNA polymerases encoded in the plasmid which increased the mutation load ex planta.

Here, we trace the natural evolutionary history of C. taiwanensis, a Mimosa rhizobium, and compare it to the experimental evolution of Ralstonia into M. pudica symbionts, using population genomics and functional enrichment analyses. We specifically focused on patterns of evolution that were previously highlighted by experimental evolution: accumulation of genetic diversity, general patterns of natural selection, chromosomal vs. plasmid adaptation, and evolution of orthologous genes implicated in symbiotic adaptation (type III secretion system, global regulators, mutagenic cassettes). We provide evidence that, despite fundamental differences in terms of time frame, protagonists, environmental context, and symbiosis achievement, there were significant parallels in the two processes.

**Results**

**Diversification of experimentally evolved Mimosa symbionts.** We previously generated 18 independent symbiotic lineages of the R. solanacearum GMI11000-pRaltaLMG19424 chimeras that nodulate M. pudica. 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 (Supplementary Data 1). The symbiotic performances of the evolved clones improved in the experiment with wide variations between lineages. Fifteen out of the 18 final clones were able to induce the formation of intracellularly infected nodules (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, we detected several large deletions in all clones (Fig. 1a).

Convergent evolution has been observed in previous evolution experiments. Thus, we first identified the parallels between the evolved clones for SNP, indels and large deletions. Almost all genetic deletions occurred in homologous regions of the symbiotic plasmid and were systematically flanked by transposable elements that probably mediated their loss by recombination (Supplementary Table 1). These regions had almost only genes of unknown function. Point mutations showed fewer parallelisms. Out of 1147 positions identified as mutated in the final clones, only 12 were found at the same nucleotide position in more than one clone (Fig. 1b). Even if these positions were rare, they were observed (O) more frequently than expected (E) by chance ((O − E)/(O + E) = 0.98, P = 0.01, test based on simulated mutations, see Methods). We then aggregated intragenic mutations per gene and found that the number of genes with mutations in more than one clone was slightly larger than expected by chance ((O − E)/(O + E) = 0.10, P = 0.02, same test). Distribution of point mutations present in more than one lineage was also not random in terms of COG functions ((O − E)/(O + E) = 0.23, P = 0.01, same test). Similarly to previous studies, this analysis highlights that parallel mutations, even if rare, were more frequent than expected.

**Genetic diversification of naturally evolved Mimosa symbionts.** 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 Note 1 and associated tables for data sources, coverage, and details of the results). We identified 1844 orthologs present in all genomes of the genus (genus core genome). The phylogeny of the genus, using this core genome, was very well resolved since only a few nodes within C. taiwanensis show values of bootstrap lower than 90% (Supplementary Fig. 1 and Supplementary Data 2). The tree shows 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 (Fig. 2 and Supplementary Data 3). Several of them were shown to be bona fide symbionts since they fix nitrogen in symbiosis with their host.

---

**ARTICLE**

**NATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-04778-5**

---

**NATURE COMMUNICATIONS**

**DOI: 10.1038/s41467-018-04778-5**

---

**www.nature.com/naturecommunications**
To prepare the comparisons between experimental and natural evolution we characterized the levels of genetic diversity within C. taiwanensis strains. 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, Supplementary Figs. 2 and 3, Supplementary Note 1, Supplementary Data 4). We then identified the genes in the core genome (those with orthologs in all strains), and in the pan genome (those present in at least one strain) of C. taiwanensis. Together, C. taiwanensis strains had a core genome of 3568 protein families and a large pan genome, 3.4 times larger than the average genome. Hence, this complex of species has very diverse gene repertoires and core genes that accumulated more genetic diversity than would be expected for a single bacterial species.

It was proposed that C. taiwanensis evolved as a symbiont recently, following the acquisition of the symbiotic genes. To test this hypothesis, we first evaluated how many times the rhizobial character (defined by the presence of the core genes of the symbiotic locus nod-nif-fix) was independently acquired in the genus Cupriavidus (Fig. 2 and Supplementary Fig. 1). This analysis involves phylogenetic reconstruction of the character in the phylogenetic tree involves the genus Cupriavidus. The events identified at the end of the 16 evolution cycles for each lineage are indicated on the right (see list of deletions in Supplementary Table 1 and other mutations in Supplementary Data 1). To reconstruct ancestral states (presence of rhizobial genes). The analysis involves phylogenetic reconstruction of the genus and loss of genes in a tree using maximum likelihood (see Methods). This analysis showed that the most likely reconstruction of the character in the phylogenetic tree involves 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 (Supplementary Fig. 4). In agreement with this proposition, we found very few homologs of the 514 pRalta genes in the genomes of UYPR2.512 (8.3 %) or amp6 (6.4%) once the 32 symbiotic genes were excluded from the analysis. Furthermore, the sequence similarity between these few homologs was significantly smaller than those of core genes (P < 0.01, Wilcoxon test).

We then used the birth–death models to identify all acquisitions of genes in the branch bLCA Ct (Fig. 2 and Supplementary Data 5). This analysis highlighted a set of 435 gene acquisitions that were present in pRalta, over-representing functions such as symbiosis, plasmid biology, and components of type IV secretion systems (Supplementary Data 6). These results are consistent with a single initial acquisition of the plasmid in this clade at the branch bLCA Ct, evidenced by the presence of a variant of pRalta encoding the symbiotic genes (Supplementary Table 2). Finally, while most individual C. taiwanensis core gene trees showed some level of incongruence with the concatenate core genome tree (2897 out of 3568, SH test), an indication of recombination, this frequency was actually lower in the core genes of the plasmid (SH, (O − E)/(O + E) = −0.41, significant difference: P < 0.04, Fisher’s exact test). Similarly, there were signals of
intrageneric recombination in many genes of the core genome (1699 out of 3568, pairwise homoplasy index (PHI) test), but these were significantly less frequent in plasmid core genes (PHI, \( \frac{O - E}{O + E} = -0.92 \), significant difference: \( P < 0.001 \), Fisher’s exact test). This suggests that the plasmid inheritance was mostly vertical within \( C. taiwanensis \). Particularly, the observation that plasmid core genes are more congruent with the core genome tree than the other core genes further reinforces the scenario of a single ancient integration of a \( pRalta \) ancestor at the \( \text{LCA}^{C1} \). We thus concluded that the \( pRalta \) ancestor at the \( \text{LCAC}^{C1} \) indicates the last common ancestor of \( C. taiwanensis \). 

### Figure 2

**Distribution of symbiotic genes, the mutagenic \( \text{imuA}^{2B2C2} \) cassette, TSSS, and \( \text{phcABQRS} \) within the 60 strains of \( \text{Cupriavidus} \).** See Supplementary Fig. 2 for the complete tree of the genus \( \text{Cupriavidus} \) and \( \text{Ralstonia} \) without simplifications in branch length. The arrow indicates the most parsimonious scenario for the acquisition of the \( pRalta \) ancestor (inferred using birth-death models of symbiotic gene acquisitions and the MPR function of the ape package in R). This is the branch before the \( \text{LCAC}^{C1} \). The node \( \text{LCA}^{C1} \) indicates the last common ancestor of \( C. taiwanensis \). Circles indicate absence (white), presence of less than 50% of the genes (light gray) 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 \( \text{LMG}19424 \). 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 Results and in Supplementary Figs. 2 and 3). Symbols: C1, C2, C3, and Cg: LCA of clades analyzed in this study. f (in a blue circle): plasmid re-sequenced by PacBio. *: two rhizobia are not part of \( C. taiwanensis \). **: \( C. taiwanensis \) type strain used as pivot to compute searches of orthologs. 1 In the PacBio version of this genome \( \text{imuA}^{2B2C2} \) is very similar to that of the reference strain, but is encoded in another plasmid.

### Table

| Genes category state          | Genes          | Sampling site |
|------------------------------|----------------|---------------|
| Absence                      | phcABQRS       | Texas         |
| Presence (> 50% of genes)    | \( pRalta \)    | Cameroon      |
| Presence (< 50% of genes)    | phcABQRS       | Guinea        |
| Presence (> 50% of genes)    | \( pRalta \)    | New Caledonia |
| Presence (> 50% of genes)    | \( pRalta \)    | Philippines   |
| Presence (> 50% of genes)    | \( pRalta \)    | Taiwan        |
| Presence (> 50% of genes)    | \( pRalta \)    | China         |

### Diagram

![Diagram showing distribution of symbiotic genes, the mutagenic \( \text{imuA}^{2B2C2} \) cassette, TSSS, and \( \text{phcABQRS} \) within the 60 strains of \( \text{Cupriavidus} \).]
Parallel evolution after acquisition of the symbiotic plasmid.

Since the experiment only reproduced the initial stages of symbiogenesis, we put forward the hypothesis that the parallels between experimental and natural adaptation should be most striking at the branch bLCA\textsuperscript{Ct} (i.e., during the onset of natural evolution towards symbiosis. In the experiment, a mechanism of transient hypermutagenesis was shown to accelerate the symbiotic evolution of bacterial populations under plant selective pressure\textsuperscript{24}. The symbiotic plasmid transferred to R. solanacearum carries aimuA2B2C2 cassette encoding stress-responsive error-prone DNA polymerase that increased the mutation rate of the recipient genome in the rhizosphere. The long time span since the acquisition of the plasmid in nature and the lack of internal time-calibration points precluded the analysis of accelerated evolution in the branch bLCA\textsuperscript{Ct} (relative to others). However, we could identify the imuA2B2C2 cassette in most extant strains. To evaluate the possibility that symbiotic plasmids with and without the mutagenesis cassette were independently acquired, we analyzed the patterns of sequence similarity between genes in the imuA2B2C2 cassette and in nif-nod locus and compared these values with the distribution of similarity in all core genes. The results show distribution of sequence similarity in these genes as expected if they were acquired only once (they are within the range of variation of core genes, Supplementary Figs. 5 and 6). Analysis of the genomes re-sequenced using PacBio technology confirmed the presence of the imuA2B2C2 locus on the symbiotic plasmid, except in strain STM6041, where this cassette was present on another plasmid (and showed lower sequence similarity, Supplementary Fig. 5). The conservation of the plasmid cassette suggests it has been under selection. Whether it has played a role in the symbiotic evolution of Cupriavidus cannot be assessed in this study, because we lack calibration points in the tree to infer an acceleration in the rate of molecular evolution in the branch bLCA\textsuperscript{Ct}.

We put forward the hypothesis that genes with an excess of polymorphism in the branch bLCA\textsuperscript{Ct} (relative to the other branches) were more likely to have endured adaptive changes. To identify such genes, we took all the genes in the core genome that were present in the branch bLCA\textsuperscript{Ct} (core genome C3 in Supplementary Data 2) and computed their genetic diversity in C. taiwanensis. We also computed the ancestral sequences of these genes, accounting for recombination using ClonalFrameML, at the node LCA\textsuperscript{Ct} and at the node immediately ancestral (C2 in Fig. 2). We then computed for each gene the number of changes between these two nodes (these are the changes that accumulated in the branch bLCA\textsuperscript{Ct}), and compared this value to the diversity of genes in C. taiwanensis. This analysis revealed 67 genes with a clear excess of changes in bLCA\textsuperscript{Ct} relative to the expected values given the diversity in the species (Supplementary Fig. 7 and Supplementary Data 5). They corresponded to genes with an excess of polymorphism in bLCA\textsuperscript{Ct}. To study the parallels between the experimental and natural processes, we identified the 2372 orthologs between R. solanacearum and C. taiwanensis (Supplementary Data 7), and added the 514 pRalta genes in the chimera as orthologs. These are the genes that have orthologs in the two systems and can thus be queried to identify parallels. We found that final clones of the evolution experiment had 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; Supplementary Tables 3 and 4). Hence, there is a significant overlap in the genes that mutated in the experiment and diverged quickly in nature upon plasmid acquisition. This revealed a first parallel between the natural and experimental processes.

The genomic rates of non-synonymous substitutions in natural populations are systematically smaller than those of synonymous substitutions\textsuperscript{31,32}. Accordingly, the substitutions in the core genes of C. taiwanensis showed an excess of synonymous changes (Supplementary Fig. 8). In contrast, experimental evolution studies often show that adaptation occurs by the fixation of an excess of non-synonymous changes\textsuperscript{25,33–36}, including in R. solanacearum\textsuperscript{37}. Yet, we identified an excess of synonymous mutations over non-synonymous mutations in the evolution experiment\textsuperscript{20}. Both processes are thus characterized by a predominance of purifying selection.

Adaptation occurred 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\textsuperscript{LMG19424} accumulated an excess of synonymous substitutions and the majority of the genetic deletions observed in the experiment (Fig. 3 and Supplementary Table 5). Interestingly, whereas core genes in the nod-nif locus were very conserved in natural isolates (Supplementary Fig. 6), this locus had many mutations in the experiment (Supplementary Data 1). Natural populations also showed more deletions in the plasmid, since from the 413 genes present in pRalta\textsuperscript{LMG19424} and inferred to be present in LCA\textsuperscript{Ct}, only 12% were in the core genome. This is six times less than found among the chromosomal genes present in C. taiwanensis LMG19424 and inferred to be present in LCA\textsuperscript{Ct} (P < 0.001, Fisher’s exact test, Fig. 3b). Notably, the few pRalta\textsuperscript{LMG19424} core genes are related to the symbiosis or to typical plasmid functions (conjugation) (Fig. 3).

In order to test if the observed rapid plasmid genetic diversification was driving the adaptation to symbiosis in nature, 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 bLCA\textsuperscript{Ct} (analysis of 1869 and 1676 core genes lacking evidence of recombination using PHI, respectively, Supplementary Data 5). Surprisingly, all 325 genes under positive selection were chromosomal (none was found among the core genes of the plasmid, P = 0.001, \( \chi^2 \) test, Fig. 3e). This suggests that adaptation took place mostly in the chromosomes.

To test this hypothesis in the experimental study, we reanalyzed all mutations previously identified as adaptive in the evolution experiment. They were all chromosomal\textsuperscript{18,22,23}. Since our previous analyses of mutations identified in the evolution experiment only focused on strongly adaptive genes, we evaluated the impact of pRalta\textsuperscript{LMG19424} mutations on the symbiotic evolution of R. solanacearum by replacing the evolved plasmid with the original pRalta\textsuperscript{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 ex planta survival rates with and without the plasmid (Supplementary Fig. 9), showing that plasmid carriage does not have a fitness cost in this respect (Supplementary Tables 6 and 7). Although we cannot exclude that some events of positive selection in the plasmid may have passed undetected, nor that further symbiotic evolutionary events of positive selection in the plasmid carrying the symbiotic traits. Importantly, the original chimera had similar ex planta survival rates with and without the plasmid (Supplementary Fig. 9), showing that plasmid carriage does not have a fitness cost in this respect (Supplementary Tables 6 and 7).

Parallel co-option of regulatory circuits. To search for parallel adaptive mutations, we first analyzed the 436 genes with non-synonymous or non-sense mutations in the experiment (Supplementary Data 1). This set of genes over-represented virulence factors of *R. solanacearum* (47 genes), including the T3SS effectors, EPS production, and a set of genes regulating (phcBQS) or directly regulated (phl, hrpG, and xpsR) by the central regulator PhcA of the cell density system that controls virulence and pathogenicity in *R. solanacearum* (Fig. 4a and Supplementary Data 8). Among this set of genes, mutations in the structural T3SS component hrcV, or affecting the virulence regulators hrpG, phl, vsrA, and efpR, were demonstrated to be responsible for the acquisition or the drastic improvement of nodulation and/or infection [18,22,23]. Only 10 of the 47 mutated virulence-associated genes have an ortholog in *C. taiwanensis*: *vsrD, xpsR, Rsp0736, issD, three phc genes (phcQ, phcB, and phcS)* and three structural genes of the T3SS (*hrcV, hrcR, and hrpQ*). We focused on the phc system and the T3SS to evaluate their respective roles in the experimental and natural processes. 

Silencing of the T3SS and its effectors was required to activate symbiosis in the evolution experiment, presumably because some effectors block nodulation and early infection [18]. We searched for the T3SS in the *Cupriavidus* genomes to test if the onset of symbiosis was associated with the acquisition or the loss of a T3SS. In contrast to the evolution experiment, the emergence of legume symbiosis in nature 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). To understand this difference between the two processes, we searched for orthologs of the 77 T3SS effectors of *R. solanacearum* GMI1000 in *C. taiwanensis* LMG19424, but we found no single ortholog for these genes (Supplementary Data 5). In complement, it has been shown that a functional T3SS is not required for mutualistic symbiosis of the latter with *M. pudica* [39], the only plant species used in the evolution experiment. Hence, the differences between the two processes seem to be caused by selection for silencing some *R. solanacearum* effectors that are lacking in *C. taiwanensis*.

We then focused on PhcA-associated genes since they accumulated an excess of mutations in the experiment (Supplementary Data 8). The phc system, which was only found intact in
Cupriavidus and Ralstonia (Supplementary Data 10), regulates a reversible switch between two different physiological states via the repression of the central regulator PhcA in Ralstonia\textsuperscript{40} and Cupriavidus.\textsuperscript{46} Interestingly, PhcA-associated genes were also enriched in substitutions in natura. Indeed, the phcBQRS genes of the cell density-sensing system were among the 67 genes that exhibited an excess of nucleotide diversity in the branch bLCA\textsuperscript{Ct} relative to C. taiwanensis ("phcA-linked" in Supplementary Data 6). Strikingly, only seven genes showing an excess of diversity at bLCA\textsuperscript{Ct} had orthologs with mutations in the evolution experiment. Among these seven, two also showed a signature of positive selection in C. taiwanensis: phcB and phcS (ongoing experiments, Supplementary Data 5).

Given the parallels between experimental and natural evolution regarding an over-representation of changes in PhcA-associated genes, we enquired on the possibility that mutations in the phcB, phcQ and phcS genes, detected in the evolved E16, K16 and M16 clones capable of nodule cell infection were adaptive for symbiosis with M. pudica. For this, we introduced the mutated alleles of these genes in their respective nodulating ancestors, CBM212, CBM349, and CBM356, and the wild-type allele in the evolved clones E16, K16, and M5 (M5 was used instead of M16, since genetic transformation failed in the latter clone in spite of many trials). Competition experiments between the pairs of clones harboring the wild-type or the mutant alleles confirmed that these mutations were adaptive (Fig. 4b). The evolved clones also showed better infectivity as measured by the number of bacteria per nodule (Supplementary Fig. 10). On the other hand, we found that the Phc system plays a role in the natural C. taiwanensis-M. pudica symbiosis, since a phcA deletion mutant had lower in planta fitness than the wild-type C. taiwanensis (Fig. 4b), and lower infectiveness (Supplementary Fig. 11), when both strains were co-inoculated to M. pudica. Hence, the rewiring of the phc virulence regulatory pathway of R. solanacearum was involved in the evolution of symbiosis in several lineages of the experimental evolution. In parallel, high genetic diversification accompanied by positive selection of the homologous pathway was associated with the transition to symbiosis in the natural evolution of C. taiwanensis.

Discussion

Years of comparative genomics and reverse genetics approaches led to propose that most legume symbionts evolved in two-steps\textsuperscript{12}, i.e., acquisition of a set of essential symbiotic genes by horizontal transfer followed by subsequent adaptation of the resulting genome under plant selective pressure. Recent experimental work was able to confirm this scenario up to the point where plants nodulate and bacteria produce intracellular infection\textsuperscript{18,22}. Yet, it was unclear if there were parallels between the experimental and the natural evolution. Such parallels were not necessarily expected, because the two processes differed in a number of fundamental points. The two species are from different genera, share only 2140 orthologues (excluding pRALTA), 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. 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. In spite of these differences, we highlighted several parallels between the experimental and in natura transitions towards legume symbiosis (Fig. 5). We also highlighted some clear differences—concerning the T3SS and its effectors—and were unable to test some parallels because of the
suggests that the symbiotic genes acquired by the recipient did not seem to have a cost in ex planta culture conditions. This is consistent with evidence of positive selection in nature. Moreover, the plasmid genes would show evidence of adaptation to the novel environmental conditions. Instead, the plasmid genes would show evidence of adaptation to the novel environmental conditions. The fact that genetic adaptation to this large plasmid encoding traits driving ecological shifts does not require plasmid evolution. The observation of these parallels suggests that adaptation following the acquisition of a plasmid may inactivate or co-opt native functions for novel functions in their hosts.

Adaptive mutations were found on regulatory modules, the rewiring of which may inactivate or co-opt native functions for the novel trait. We previously showed that loss of the ability to express the T3SS was strictly necessary for the early transition towards symbiosis in the experiment, and that subsequent adaptation favored the re-use of regulatory modules leading to massive metabolic and transcriptomic changes. These phenotypic shifts occurred via mutations targeting regulatory genes specific to Ralstonia (e.g., hrpG, prhI, efpR, Rsc0965), which finely control the expression of many virulence determinants.

Here, from the analysis of orthologs between Ralstonia chimeras and C. taiwanensis, we showed that several genes in the phcBQRS operon both exhibited significant positive selection in C. taiwanensis populations and accumulated adaptive mutations in the evolution experiment. In R. solanacearum, these genes control the activity of the global virulence regulator PhcA via a cell density-dependent mechanism. Mutations in this pathway are unlikely to cause adaptation by attenuating the virulence of Ralstonia, because the chimeric ancestor is not pathogenic on Mimosa. Since PhcA also plays a role in the natural C. taiwanensis–M. pudica symbiosis, we speculate that adaptive mutations in the experiment and high genetic diversification in nature on phc genes after the acquisition of pRalta may reflect the rewiring of a quorum-sensing system to sense the environment for cues of when to express the novel mutualistic dialogue with eukaryotes. Further work should determine if some of these mutations resulted in the integration of the gene expression network of the plasmid in the broader network of the cell.

Fig. 5 Overall similarities and differences between the experimental and natural evolutionary processes described in this study. Known adaptive and non-adaptive changes are in red and blue, respectively.
highlights the potential of research projects integrating population genomics, molecular genetics, and evolution experiment to provide insights on adaptation in nature and in the laboratory. Therefore, experimental evolution appears not only useful to demonstrate the biological plausibility of theoretical models in evolutionary biology, but also to enlighten the natural history of complex adaptation processes.

Methods

Dataset for the experimental evolution. We used previously published data on the genomic changes observed in the experimental evolution of the clade, including 21 bacterial clones (three ancestors and 18 evolved clones)48. We analyzed all the synonymous and non-synonymous mutations of each clone from these datasets (Supplementary Data 1). Large deletions above 1 kb were first listed based on the absence of Illumina reads in these regions, and were then verified by PCR amplification using specific primers listed in Supplementary Table 8. 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 (Supplementary Table 1).

Mutant construction. Strains used in this study are indicated in Supplementary Table 6. The pRalta in evolved Ralstonia clones B16, G16, and I16 was replaced by the wild-type pRalta of R. solanacearum LMG19424 strain as previously described49, generating B16-op, G16-op, and I16-op. Wild-type alleles of the phcA, phcQ, and gfp (except pRalta from strain GMI1000, which was replaced by pRalta from strain LMG2235) were reintroduced into Ralstonia evolved clones using the MuCgent technique50. 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 described51. 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-Sclel/pDAI-Sclel technique55. Briefly, the regions upstream and downstream of pDAI-Sclel were amplified with the oCBM3143−3144 and oCBM3415−3416 primer pairs and the Phusion DNA polymerase (Thermo Fisher Scientific). The two PCR products were digested with Xbal-BamHI and BamHI-EcoRI respectively and cloned into the pGPI-Sclel plasmid digested by Xbal and EcoRI. The resulting plasmid was introduced into LMG19424 by triparental mating using the pRK2013 as helper plasmid. Deletion mutants were obtained after introduction of the pDAI-Sclel plasmid encoding the I-Sclel 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 Supplementary Table 9.

Relative in planta fitness. Mimosa pudica seeds from Australia origin (B&T World Seed, Pauquian, France) were surface sterilized for 15 min in a concentrated sulfuric acid, rinsed with sterile water and incubated for further 10 min in 2.4% sodium hypochlorite solution. Seeds were soaked for 24 h in sterile water at 28 °C under agitation then germinated on soft agar plates in darkness at 28 °C for 24 h. Seedlings were grown in N-free conditions in Gibson tubes filled with 20 ml of Fabaeas solid medium56 and 50 ml of 1/4th strength of Jensen liquid medium57. To measure the in planta relative fitness, a mix of two strains bearing different antibiotic resistance genes or fluorophores (5×10^6 bacteria of each strain per plant) were inoculated to 20 plants. All nodules were harvested 21 days after inoculation, pooled, surface sterilized, and crushed. Dilutions of nodule crushed were spread on selective plates, incubated 2 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.

Survival measurement. Single colonies of GM10000 (RCM1068) or GM10000-pRalta (RCM1069) were grown overnight in rich BG medium58 and 10^5 bacteria were inoculated to Gibson tubes filled with quarter-strength Jensen medium59. Colonies containing 1 Mb or more of deletions were scored using a fluorescence stereo zoom microscope V16 (Zeiss) when needed. Three independent experiments were performed.

Production and analysis of Illumina sequences. The genomes of 43 Mimosa spp. isolates, a non-rhizobial strain of Cupriavidus (strain LMG19464) as well as a C. caviae strain (strain LMG2235) (Supplementary Data 3), were sequenced using the GeT-PlaGe core facility, INRA Toulouse (get.genotoul.fr). DNA-seq libraries were prepared according to Bioanant’s protocol using the Bioanant’s PCR-free Library Prep Kit. Briefly, DNA was fragmented by sonication, size selection was performed using CLEANNA CleanPCR beads and adapters were ligated to be sequenced. Library quality was assessed using an Advanced Analytical Fragment Analyzer 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 × 100 bp with the HiSeq v3 chemistry (LMG2235) or on an Illumina MiSeq sequencer using a paired-end read length of 2 × 300 bp with the Illumina MiSeq v3 reagent kit (Table 1). On average, genomes contained 99 contigs and an L90 of 29.

Other genomes were sequenced using the MicroScope pipeline for complete structural and functional annotation60. Gene prediction was performed using GeneMarkSS and biological gene prediction was performed using Glimmer61 known for its capability to locate the translation initiation site with great accuracy. The RNAmmer62 and rNAScan-SE63 programs were used to predict tRNA and rRNA-encoding genes, respectively.

PacBio sequencing. Library preparation and sequencing were performed according to the manufacturer’s instructions “Shared protocol-20 kb Template Preparation Using BluePippin Size Selection system (15kb-size cutoff)”. At each step, DNA was quantified using the Qubit 4.0 DNA HS Assay Kit (Life Technologies), DNA purity was tested using the nanodrop (Thermo Fisher) and size distribution and degradation assessed using the Fragment analyzer (AATI) High Sensitivity DNA Fragment Analysis Kit. Purification steps were performed using 0.45× AMPure PB beads (Pacbio). Ten micrograms of DNA was purified then sheared at 40 kb using the microbead system (diagnostics). A DNA and END damage repair step was performed on 5 µg of sample. Then blunt hairpin adapters were ligated to the library. The library was treated with an exconuclease cocktail to digest unligated DNA fragments. A size selection step using a 13–15 kb cutoff was performed on the BluePippin Size Selection system (Sage Science) with the 0.75× agarose cassettes, Marker SI high pass 15–24 kb.

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 1:10. Then after a magnetic bead-lodging step (OCFW), SMRTbell libraries were sequenced on RSII instrument at 0.2 mreads with a 360× SMRTbell movie. One SMRTCell movie was used for sequencing. The resulting results were validated and provided by the Integrated Next generation sequencing storage and processing environment NG6 accessible in the genome core facility website64.

Core genomes. Core genomes were computed using reciprocal best hits (hereafter named RBH), using end-gap free Needleman–Wunsch global alignment, between the proteomes of C. taiwanensis LMG19424 or R. solanacearum GMI1000 (when the previous was not in the sub-clade) as a pivot (indicated by ** on Supplementary Fig. 2A) and each of the other 88 proteomes65. Hits with less than 40% similarity in amino acid sequence or more than a third of difference in protein length were discarded. The lists of orthologs were filtered using positional information. Positional orthologs were defined as RBH adjacent to at least two other pairs of RBH within the same neighborhood of ten genes (five up- and five downstream). A DNA and RNA analysis of the proteomes of all species/gene families and a sub-sequence of highly similar orthologs was performed using the AMIGene software57 and the microbial gene identification program (MIGene)66. A paralog in the core genome was defined as the intersection of the lists of orthologs between two species/gene families. A paralog in the core genome was defined as the intersection of the lists of orthologs between two species/gene families.
genome in the set (Supplementary Fig. 12), and defined an appropriate threshold in order to include nearly all core genes but few paralogs (Supplementary Table 10).

**Alignment and phylogenetic analyses.** Multiple alignments were performed on protein sequences using Muscle v3.8.3164, and back-translated to DNA. We analyzed how the concatenated alignment of core genes fitted different models of protein or DNA evolution using IQ-TREE v1.3.865. The best model was determined using the Bayesian information criterion. Maximum likelihood trees were then computed using IQ-TREE v1.3.8 using the appropriate model validated via a ultrafast bootstrap procedure with 1000 replicates (Supplementary Data 2). The maximum likelihood trees of each set of core genes were computed with IQ-TREE v1.3.8 using the best model obtained for the concatenated multiple alignment.

In order to root the phylogeny based on core genes, we first built a tree using 16S rRNA genes of the genomes of *Ralstonia* and *Capriovivida* genera analyzed in this study and of ten outgroup genomes of *β*-Proteobacteria. For this, we made a multiple alignment of the 16S rRNA genes with INFERNAL v.1.1 (with default parameter)67 using RF00177 Rfam model (v.12.1)68, followed by manual correction with SEAVIEW to remove poorly aligned regions. The tree was computed with 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 blCA3, we computed the distances in the 16S rRNA genes tree between each strain and each of the scaffolds delimiting the reticulately lCA3 and lCA2 in Fig. 2. The substitution rate of 16S rRNA genes in enterobacteria was estimated at ~1% per 50 MY of divergence69, and we used this value as a reference.

**Orthologs and pseudogenes of specific families of genes.** We identified the positional orthologs of Cg for symbiotic genes, the mutagentic cassette, TSSS, and PhaABQRS using RBH and *C. taiwanensis* LM914924 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 mutagentic cassette, TSSS and PhaABQRS. We searched for homologs of each gene in the reference genome in the other genomes using LAST v744 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 obtained from the tree of the concatenated multiple alignment of all core genes.

**Evolution of gene families using birth–death models.** We used Count (version downloaded in December 2015)31 to study the past history of transfer, loss and protein sequences using Muscle v3.8.3164, and back-translated to DNA. We performed the analysis gene families that had incongruent phylogenetic signals within the multiple alignment78. These correspond to the families for which PHI identified evidence of recombination (P < 0.05).

We inferred the mutations arising in the branch leading to lCA2 using the phylogentic tree build with the core genome of C3 (Ct and the five closest Clones). First, we used ClonalFrameML to reconstruct the ancestral sequences of lCA2 and lCA2 (accounting for recombination). Then, we estimated nucleotide diversity of each core gene for Ct, and between lCA2 and lCA2 using the R package peaks. 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 intrageneric 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* LM914924 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. We used the dN/dS values to determine if each core genes between *C. taiwanensis* LM914924 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.

**Orthologs and pseudogenes of specific families of genes.** We identified the positional orthologs of Cg for symbiotic genes, the mutagentic cassette, TSSS, and PhaABQRS using RBH and *C. taiwanensis* LM914924 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 mutagentic cassette, TSSS and PhaABQRS. We searched for homologs of each gene in the reference genome in the other genomes using LAST v744 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 obtained from the tree of the concatenated multiple alignment of all core genes.

**Evolution of gene families using birth–death models.** We used Count (version downloaded in December 2015)31 to study the past history of transfer, loss and duplication of the protein families of the pan genomes. The analysis was done using the count program of the ClonalFrameML. We computed the posterior probability of the birth, death and duplication of each protein families of genes for Ct. We used the corresponding trees of each protein families of genes to estimate the number of SNP of each replicon for the 18 final cloned in the branch leading to LCACt. We compared the 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 cloned 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 can be analyzed without ambiguity in the assignment of mutations. We built the mutation spectrum of the genome using these synonymous mutations, since they can be analyzed without ambiguity in the assignment of mutations. We built the mutation spectrum of the genome using these synonymous mutations, since they can be analyzed without ambiguity in the assignment of mutations. We built the mutation spectrum of the genome using these synonymous mutations, since they can be analyzed without ambiguity in the assignment of mutations.
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 overrepresented 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 Supplementary Table 1) 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 mutations found in the experimental evolution targeted genes that were found to be significantly more diverse in the natural process.

P values were corrected for multiple comparisons using Benjamini and Hochberg’s method (stats, P.adjust). Statistical analyses with R were done using version 3.1.3 (R: a language and foundation for statistical computing, 2008; R Development Core Team, R Foundation for Statistical Computing, Vienna, Austria (http://www.R-project.org)).

Data availability. Genome sequence and annotation were made publicly available (GenBank BioProject: PRJEB23670, see accession numbers in Supplementary Data 3).

Received: 18 January 2018 Accepted: 11 May 2018
Published online: 11 June 2018

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, 827–833 (2013).
3. Metzger, B. P. H., Yuan, D. C., Gruber, J. D., Duveau, F. & Wittkop, 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 Ralstonia solanacearum: evidence for targeted adaptation. Proc. Natl. Acad. Sci. USA 113, 4485–4489 (2016).
5. Barrick, J. E. & Batalion, 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. Ben-Gurion, R. & Shafferman, A. Essential virulence determinants of different Yersinia species are carried on a common plasmid. Plasmid 5, 183–187 (1981).
8. Bernardini, M. L., Mounier, J., d’Hauteville, H., Coquis-Rondon, M. & Sansonneti, J. P. Identification of iczA, a plasmid locus of Shigella flexneri that governs bacterial intra- and intercellular spread through interaction with F-actin. Proc. Natl. Acad. Sci. USA 86, 3867–3871 (1989).
9. Okinaka, R. T. et al. Sequence and organization of pXO1, the large plasmid of Ralstonia solanacearum. Plasmid 17, 458–466 (1997).
10. Sullivan, J. T., Pollard, 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. USA 92, 8985–8989 (1995).
11. Moulin, L., Bénia, G., Boivin-Masson, C. & Stepkowski, 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).
12. Nandakena, K. G., O’Hara, G. W., Tiwari, R. P. & Howison, J. G. Rapid in situ evolution of nodulating strains for B. p. mediterraneus through lateral transfer of a symbiosis island from the original mesorhizobial inoculant. Appl. Environ. Microbiol. 72, 7365–7367 (2006).
13. Masson-Boivin, C. & Sachs, J. L. Symbiotic nitrogen fixation by rhizobia - the roots of a success story. Curr. Opin. Plant Biol. 44, 7–15 (2018).
14. 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).
15. Marchetti, M. et al. Experimental evolution of a plant pathogen into a legume symbiont. PLoS Biol. 8, e1000280 (2010).
16. Amadou, C. et al. Genome sequence of the –rhizobium C. taiwanensis and comparative genomics of rhizobia. Genome Res. 18, 1472–1483 (2008).
17. 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).
18. Marchetti, M. et al. Shaping bacterial symbiosis with legumes by experimental evolution. Mol. Plant Microbe Interact. 27, 956–964 (2014).
19. 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. Ecol. Evol. 34, 2503–2521 (2017).
20. Guan, S. H. et al. Experimental evolution of nodule intracellular infection in legume symbionts. Isme J. 7, 1367–1377 (2013).
21. Remigi, P. et al. Transient hypermutagenesis accelerates the evolution of legume endosymbionts following horizontal gene transfer. PLoS Biol. 10, e1001942 (2012).
22. Tenaillon, O. et al. The molecular diversity of adaptive convergence. Science 335, 457–461 (2012).
23. Marvig, R. L., Sommer, L. M., Molin, S. & Johansen, H. K. Convergent evolution and adaptation of Pseudomonas aeruginosa within patients with cystic fibrosis. Nat. Genet. 47, 57–64 (2014).
24. Klouowksa, 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).
25. Chen, W.-M. et al. Legume symbiotic nitrogen fixation by β-proteobacteria is widespread in nature. J. Bacteriol. 185, 7266–7272 (2003).
26. Andam, C. P., Mondo, S. J. & Parker, M. A. Monophyly of nodA and nifH genes across Texan and Costa Rican populations of C. taiwanensis and C. nodicola. Appl. Environ. Microbiol. 73, 4686–4690 (2007).
27. 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. USA 106, 20359–20364 (2009).
28. Kuo, E. C. P. et al. Comparisons of dN/dS are time dependent for closely related bacterial genomes. J. Theor. Biol. 239, 226–235 (2006).
29. Perret, X. et al. The type 3 protein secretion system of C. taiwanensis and C. nodicola. Appl. Environ. Microbiol. 78, 6746–6757 (2012).
30. Paterson, S. et al. Antagonistic coevolution accelerates molecular evolution. Nature 464, 275–278 (2010).
31. Beaumont, H. J., Gallee, J., Kost, C., Ferguson, G. C. & Rainey, P. B. Experimental evolution of bet hedging. Nature 462, 90–93 (2009).
32. Barrick, J. E. et al. Genome evolution and adaptation in a long-term experiment with Escherichia coli. Nature 461, 1243–1247 (2009).
33. Perfeito, L., Fernandes, L., Mota, C. & Gordo, I. Adaptive mutations in bacteria: high rate and small effects. Science 317, 813–815 (2007).
34. Guidot, A. et al. Multithost experimental evolution of the pathogen Ralstonia solanacearum unveils genes involved in adaptation to plants. Mol. Biol. Evol. 31, 2913–2928 (2014).
35. Guten, S. et al. Pathogen evolutionary consequences of the pathogen Ralstonia solanacearum complex on the type 3 protein secretion system of C. taiwanensis strain LMG19424 compromises symbiosis with Leucaena leucocephala. Appl. Environ. Microbiol. 78, 4776–4779 (2012).
36. Saad, M. M., Crevecoeur, M., Masson-Boivin, C. & Perret, X. The type 3 protein secretion system of C. taiwanensis strain LMG19424 compromises symbiosis with Leucaena leucocephala. Appl. Environ. Microbiol. 58, 359–367 (2000).
37. Mishra, R. P. et al. Genetic diversity of Mimosa pudica rhizobial symbionts in soils of Indian Guiana: investigating the origin and diversity of Bradyrhizobium phaseoli and other beta-rhizobia. FEMS Microbiol. Ecol. 79, 487–503 (2012).
42. Bontemps, C. et al. *Burkholderia* species are ancient symbionts of legumes. *Mol. Ecol.* 19, 41–52 (2010).

43. Poukhov, N., Guégan, F., Vaillée, F. & Valls, M. *Ralstonia solanacearum*, a widespread bacterial plant pathogen in the post-genomic era. *Mol. Plant Pathol.* 14, 651–662 (2013).

44. 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, e1e006444 (2016).

45. Clough, S. J., Lee, K.-E., Schell, M. A. & Denny, T. P. A two-component system in *Ralstonia* (Pseudomonas) *solanacearum* modulates production of PhoA-regulated virulence factors in response to 3-hydroxypalmitic acid methyl ester. *J. Bacteriol.* 179, 3639–3648 (1997).

46. Dalia, A. B., McDonough, E. & Camilli, A. Multiplex genome editing by transposon delivery. *Nat. Protoc.* 11, 8937–8942 (2014).

47. Daubech, B. et al. Spatio-temporal control of mutualism in legumes helps spread symbiotic nitrogen fixation. *eLife* e62883 (2017).

48. Fahraeus, G. The invention of clover root hairs by nodule bacteria studied by a simple glass slide technique. *J. Gen. Microbiol.* 16, 374–381 (1957).

49. Jensen, H. L. Nitrogen fixation in leguminous plants. I. General characters of root nodule bacteria isolated from species of *Medicago* and *Trifolium* in Australia. *Proc. Int. Soc.* 67, 98–110 (1942).

50. Rouillon, M. & Cruveiller, S. AMALGAM: an automatic assembler tool for microbial genomes. https://doi.org/10.5281/zenodo.1239599 (2018).

51. Simpson, J. T. et al. ABySS: a parallel assembler for short read sequence data. *Genome Res.* 19, 1117–1123 (2009).

52. Peng, Y., Leung, H. C., Yiu, S.-M. & Chin, F. Y. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28, 1420–1428 (2012).

53. Koren, S. et al. Canu: scalable and accurate long-read assembly via iterative k-mer weighting and repeat separation. *Genome Res.* 27, 722–736 (2017).

54. Luo, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1, 18 (2012).

55. Mergulhas, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 370–378 (2005).

56. Miege, V., Penel, S. & Duret, L. Ultra-fast sequence clustering from similarity networks with SILIX. *BMCI Bioinformatics* 12, 116 (2011).

57. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797 (2004).

58. Moucheron, N. et al. Clan HOMSTRAD: a new database of homologous protein families. *Proteins: Struct., Funct., Bioinform.* 68, 631–642 (2007).

59. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274 (2015).

60. Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* 30, 1188–1195 (2013).

61. Nicholas, B. J. & Miyamoto, M. S. PRODAP: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964 (1997).

62. Kare, S. et al. Cloning of wild-type *Pseudomonas solanacearum* phcA, a gene that when mutated alters expression of multiple traits that contribute to virulence. *J. Bacteriol.* 170, 5577–5585 (1997).

63. 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.* 170, 5577–5585 (1997).

64. Gonzalez, E. T., Brown, D. G., Swanson, J. K. & Allen, C. Using the TULIP (ANR-10-LABX-41) and France Génomique National infrastructure, funded as part of “Investissement d’avenir” program managed by Agence Nationale pour la Recherche (contract ANR-10-INBS-09). We thank Olaya Rendueles, Rémi Peyraud, Ludovic Cotrett, Stéphane Genin, and Pedro Couto Oliveira for helpful comments and suggestions. We thank Eddy Ngonkeu and Moussa Diabate for help with the strain collection. We thank Céline Vandecasteele for the analysis of PacBio data. Sequencing was performed at the GEV core facility, Toulouse, France, supported in part by France Génomique National infrastructure (contract ANR-10-INBS-09).

**Acknowledgements**

This work was supported by funds from the French National Research Agency (ANR-12-ADAP-0014-01 and ANR-16-CE20-0011-01) and the “Labelexcellence (LABEX)” project “PROKARYOT” (ANR-11-LABX-41). The authors are grateful to the French National Research Agency (ANR) for funding this work as part of “Investissement d’avenir” program managed by Agence Nationale pour la Recherche (contract ANR-10-INBS-09). We thank Olaya Rendueles, Rémi Peyraud, Ludovic Cotrett, Stéphane Genin, and Pedro Couto Oliveira for helpful comments and suggestions. We thank Eddy Ngonkeu and Moussa Diabate for help with the strain collection. We thank Céline Vandecasteele for the analysis of PacBio data. Sequencing was performed at the GEV core facility, Toulouse, France, supported in part by France Génomique National infrastructure (contract ANR-10-INBS-09).

**Author contributions**

C.C., C.M.-B., and E.P.C.R. conceived the project, integrated the analyses, and wrote the draft of the manuscript. C.C., M.Ta., and E.P.C.R. made the computational analyses. D.C. and M.Ta. performed the experiments and analyzed the data. L.M. and M.A.P. provided the

---

**References**

64. Gonzalez, E. T., Brown, D. G., Swanson, J. K. & Allen, C. Using the TULIP (ANR-10-LABX-41) and France Génomique National infrastructure, funded as part of “Investissement d’avenir” program managed by Agence Nationale pour la Recherche (contract ANR-10-INBS-09). We thank Olaya Rendueles, Rémi Peyraud, Ludovic Cotrett, Stéphane Genin, and Pedro Couto Oliveira for helpful comments and suggestions. We thank Eddy Ngonkeu and Moussa Diabate for help with the strain collection. We thank Céline Vandecasteele for the analysis of PacBio data. Sequencing was performed at the GEV core facility, Toulouse, France, supported in part by France Génomique National infrastructure (contract ANR-10-INBS-09).

**Author contributions**

C.C., C.M.-B., and E.P.C.R. conceived the project, integrated the analyses, and wrote the draft of the manuscript. C.C., M.Ta., and E.P.C.R. made the computational analyses. D.C. and M.Ta. performed the experiments and analyzed the data. L.M. and M.A.P. provided
strains and data. C.G. and C.L.R. performed sequencing. S.C. assembled and annotated the genomes. All authors contributed to the final text.

Additional information
Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-04778-5.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.