

ARTICLE

DOI: 10.1038/s41467-018-04778-5

OPEN

# Parallels between experimental and natural evolution of legume symbionts

Camille Clerissi<sup>1,2,3</sup>, Marie Touchon<sup>1,2</sup>, Delphine Capela <sup>3</sup>, Mingxing Tang<sup>3</sup>, Stéphane Cruveiller<sup>4</sup>, Clémence Genthon<sup>5</sup>, Céline Lopez-Roques<sup>5</sup>, Matthew A. Parker<sup>6</sup>, Lionel Moulin<sup>7</sup>, Catherine Masson-Boivin<sup>3</sup> & Eduardo P.C. Rocha <sup>1,2</sup>

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 *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 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.

<sup>1</sup> Microbial Evolutionary Genomics, Institut Pasteur, 28 rue Dr. Roux, 75015 Paris, France. <sup>2</sup> UMR3525, CNRS, 28 rue Dr. Roux, 75015 Paris, France. <sup>3</sup> LIPM, Université de Toulouse, INRA, CNRS, 31326 Castanet-Tolosan, France. <sup>4</sup> LABGeM, Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Université d'Evry, Université Paris-Saclay, 2 rue Gaston Crémieux, 91057 Evry, France. <sup>5</sup> INRA, US 1426, GeT-PlaGe, Genotoul, 31326, Castanet-Tolosan, France. <sup>6</sup> Department of Biological Sciences, State University of New York, 4400 Vestal Parkway East, PO Box 6000, Binghamton, NY 13902, USA. <sup>7</sup> IRD, Cirad, Université de Montpellier, IPME, 911, avenue Agropolis—BP64501, 34394 Montpellier Cedex 5, France. Correspondence and requests for materials should be addressed to C.M.-B. (email: [Catherine.masson@inra.fr](mailto:Catherine.masson@inra.fr)) or to E.P.C.R. (email: [erocha@pasteur.fr](mailto:erocha@pasteur.fr))

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<sup>1,2</sup>. 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<sup>5</sup>.

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<sup>6</sup>. 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<sup>7–9</sup>. Adaptation is often coupled with the genetic rewiring of the recipient genome, a process that may take hundreds to millions of years in nature<sup>10</sup>, and may require specific genetic backgrounds<sup>11</sup>. 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 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<sup>12</sup>. These symbiotic capacities emerged several times in the natural history of  $\alpha$ - and  $\beta$ -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<sup>13–15</sup>. Indeed, legumes have developed control mechanisms that allow the selection of most compatible and beneficial symbionts<sup>16</sup>. There are now hundreds of known rhizobial species scattered in 14 known genera, including the genus *Cupriavidus* in  $\beta$ -proteobacteria<sup>17</sup>.

Transition towards legume symbiosis has recently been tested at the laboratory time-scale using an experimental system<sup>18</sup>. 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<sup>LMG19424</sup>—encoding the key genes allowing the symbiosis between *Cupriavidus taiwanensis* LMG19424<sup>19</sup> 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<sup>18,20,21</sup>. Several adaptive mutations driving acquisition and/or drastic improvement of nodulation and infection were previously identified<sup>18,22,23</sup>. Lab-evolution was accelerated by stress-responsive error-prone DNA polymerases encoded in the plasmid which increased the mutation load ex planta<sup>24</sup>.

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 cassette). 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

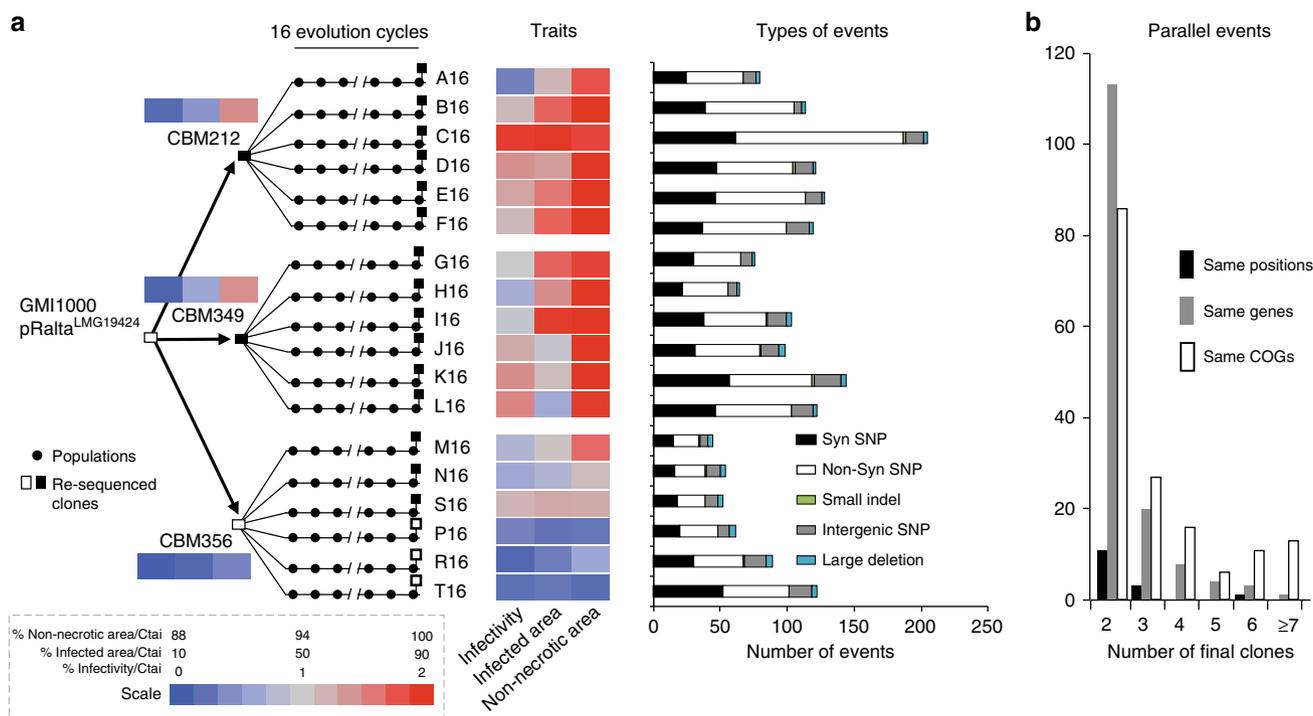
### Diversification of experimentally evolved *Mimosa* symbionts.

We previously generated 18 independent symbiotic lineages of the *R. solanacearum* GMI1000-pRalta<sup>LMG19424</sup> chimeras that nodulate *M. pudica*<sup>20</sup>. 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<sup>20</sup>, we detected several large deletions in all clones (Fig. 1a).

Convergent evolution has been observed in previous evolution experiments<sup>25,26</sup>. 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<sup>25,26</sup>, 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<sup>27,28</sup>.



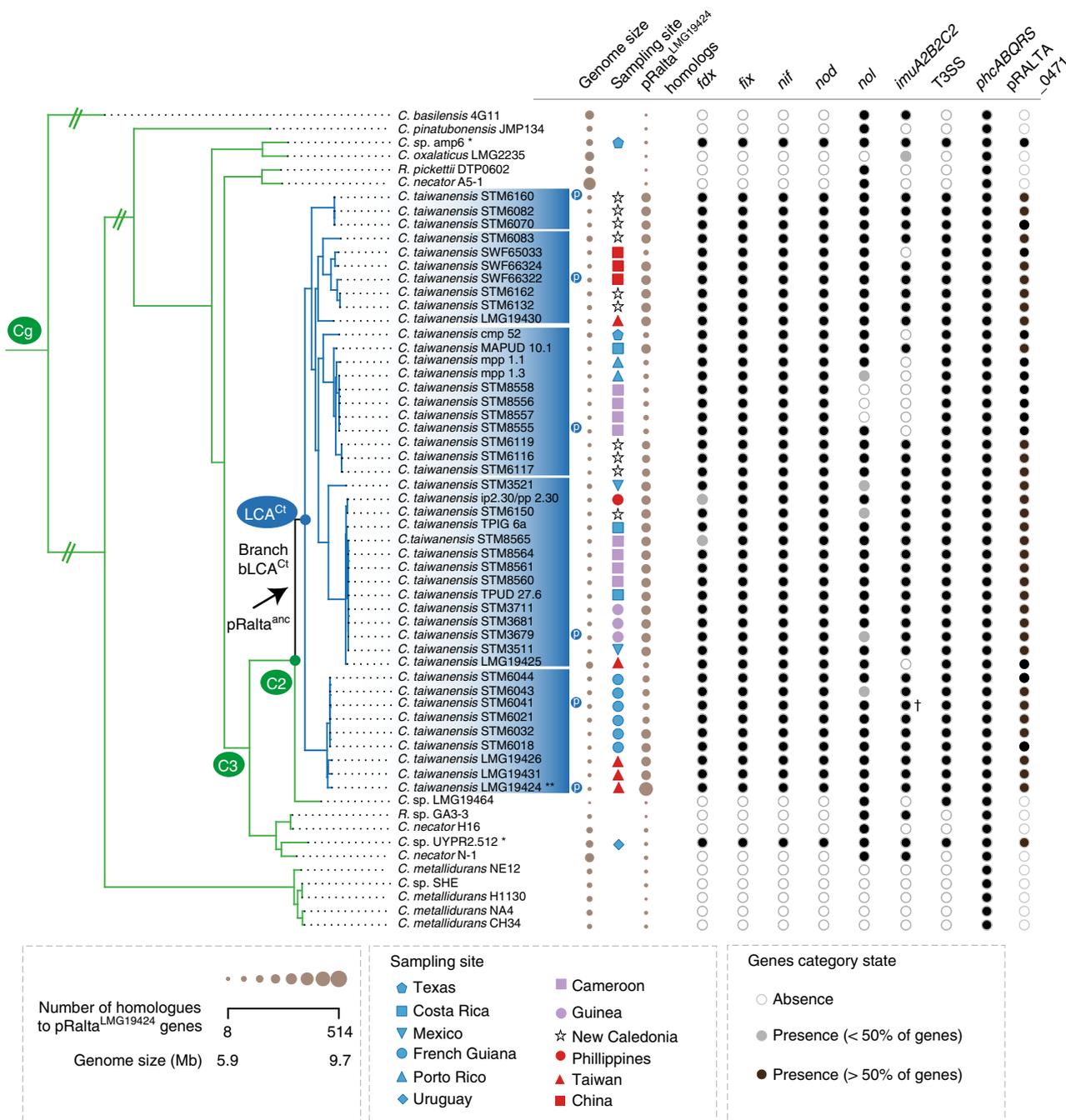
**Fig. 1** Experimental evolution of *Ralstonia* and associated symbiotic and genomic changes. **a** An ancestral chimeric clone evolved to give origin to three clones able to nodulate *M. pudica*. These clones were then evolved in 18 independent lineages using 16 serial nodulation cycles. This process 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 Supplementary Table 1 and other mutations in Supplementary Data 1). **b** Black bars indicate the number of nucleotides mutated at the same position in two or more clones. Gray bars indicate the number of genes mutated in two or more clones. White bars indicate the number of COG functions with mutations in two or more clones

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<sup>19,29</sup>. 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 genus and inference of ancestral states (presence of rhizobial genes). The phylogenetic inference is very robust; the three clades are separated in the tree by nodes of high statistical confidence (100% bootstrap, Supplementary Fig. 2). To reconstruct ancestral states, we used birth–death models that describe the rate of gain 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  $pRalta^{LMG19424}$  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  $pRalta^{LMG19424}$ , 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  $bLCA^{Ct}$ . PacBio resequencing of five strains representative of the main lineages, one for each putative novel species of *C. taiwanensis*, confirmed the ubiquitous 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



**Fig. 2** Distribution of symbiotic genes, the mutagenic *imuA2B2C2* cassette, T3SS, and *phcABQRS* within the 60 strains of *Cupriavidus*. See Supplementary Fig. 2 for the complete tree of the genus *Cupriavidus* and *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 LCA<sup>Ct</sup>. The node LCA<sup>Ct</sup> 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<sub>0471</sub> 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 Results and in Supplementary Figs. 2 and 3). 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* type strain used as pivot to compute searches of orthologs. † In the PacBio version of this genome *imuA2B2C2* is very similar to that of the reference strain, but is encoded in another plasmid

intragenic 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,  $(O - E)/(O + E) = -0.92$ , 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 LCA<sup>Ct</sup>. 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<sup>Ct</sup>. The date of plasmid acquisition was estimated using a 16S rRNA genes clock in the range 12–16 MY ago. Although these dating procedures are only approximate, the values are consistent with the low ANI<sub>b</sub> values within *C. taiwanensis*. They place the acquisition of the plasmid after the radiation of its most typical host (*Mimosa*<sup>30</sup>).

### 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<sup>Ct</sup>, 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<sup>24</sup>. The symbiotic plasmid transferred to *R. solanacearum* carries a *imuA2B2C2* cassette encoding stress-responsive error-prone DNA polymerases 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<sup>Ct</sup> (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<sup>Ct</sup>.

We put forward the hypothesis that genes with an excess of polymorphism in the branch bLCA<sup>Ct</sup> (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<sup>Ct</sup> (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<sup>Ct</sup> 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<sup>Ct</sup>), 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<sup>Ct</sup> 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<sup>Ct</sup>. 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.

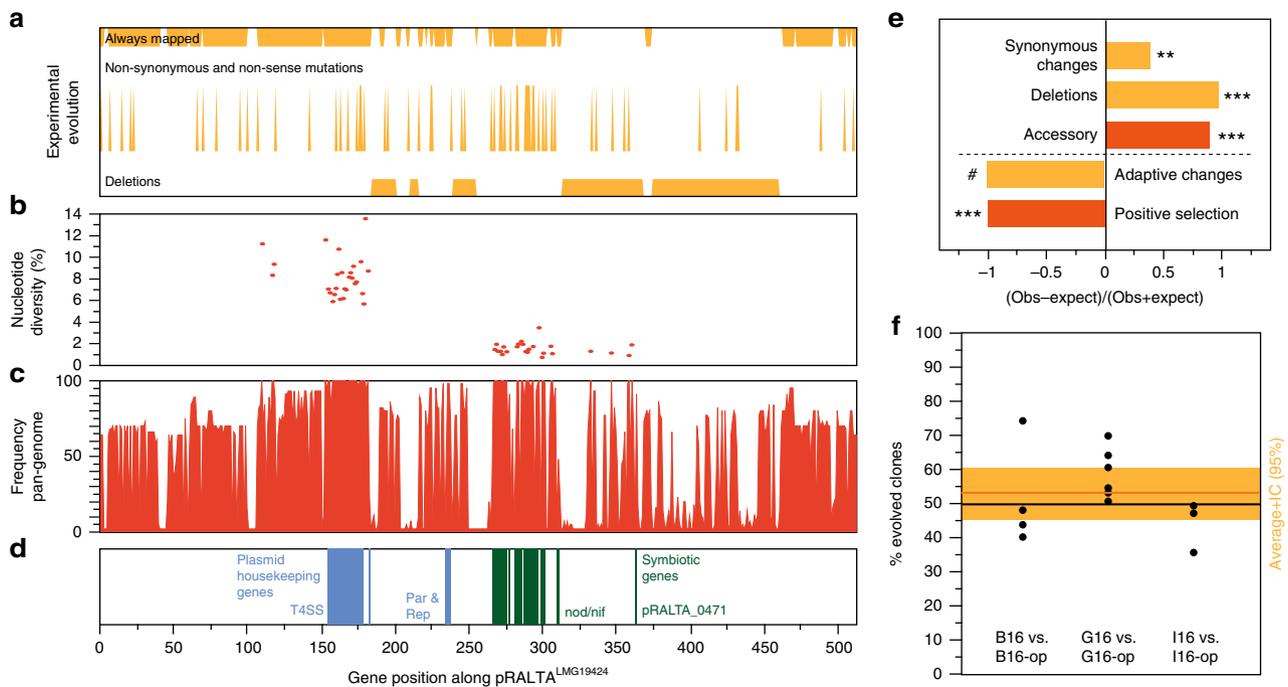
The genomic rates of non-synonymous substitutions in natural populations are systematically smaller than those of synonymous substitutions<sup>31,32</sup>. Accordingly, the substitutions in the core genes of *C. taiwanensis* showed an excess of synonymous changes (Supplementary Fig. 8). In contrast, experimental evolution studies often show that adaptation occurs by the fixation of an excess of non-synonymous changes<sup>25,33–36</sup>, including in *R. solanacearum*<sup>37</sup>. Yet, we identified an excess of synonymous mutations over non-synonymous mutations in the evolution experiment<sup>20</sup>. 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<sup>LMG19424</sup> 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 *nif-nod* 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<sup>LMG19424</sup> and inferred to be present in LCA<sup>Ct</sup>, 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<sup>Ct</sup> ( $P < 0.001$ , Fisher's exact test, Fig. 3b). Notably, the few pRalta<sup>LMG19424</sup> 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 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 bLCA<sup>Ct</sup> (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 re-analyzed all mutations previously identified as adaptive in the evolution experiment. They were all chromosomal<sup>18,22,23</sup>. Since our previous analyses of mutations identified in the evolution experiment only focused on strongly adaptive genes, we evaluated the impact of pRalta<sup>LMG19424</sup> mutations on the symbiotic evolution of *R. solanacearum* by replacing the evolved plasmid with the original pRalta<sup>LMG19424</sup> 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



**Fig. 3** Analysis of the symbiotic plasmid of *C. 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 (Supplementary Tables 1, 4, and Data 5). \*\*/\*\* 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 Supplementary Table 5). # 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 or I16) and the same clone with the evolved pRalta replaced by the original one (B16-op, G16-op or I16-op). Each dot represents an independent experiment. The orange horizontal bar represents the average and the large orange rectangle the 95% interval of confidence of the average (t Student). This interval includes the value 50% indicating that the two types of clones are not significantly different in terms of fitness

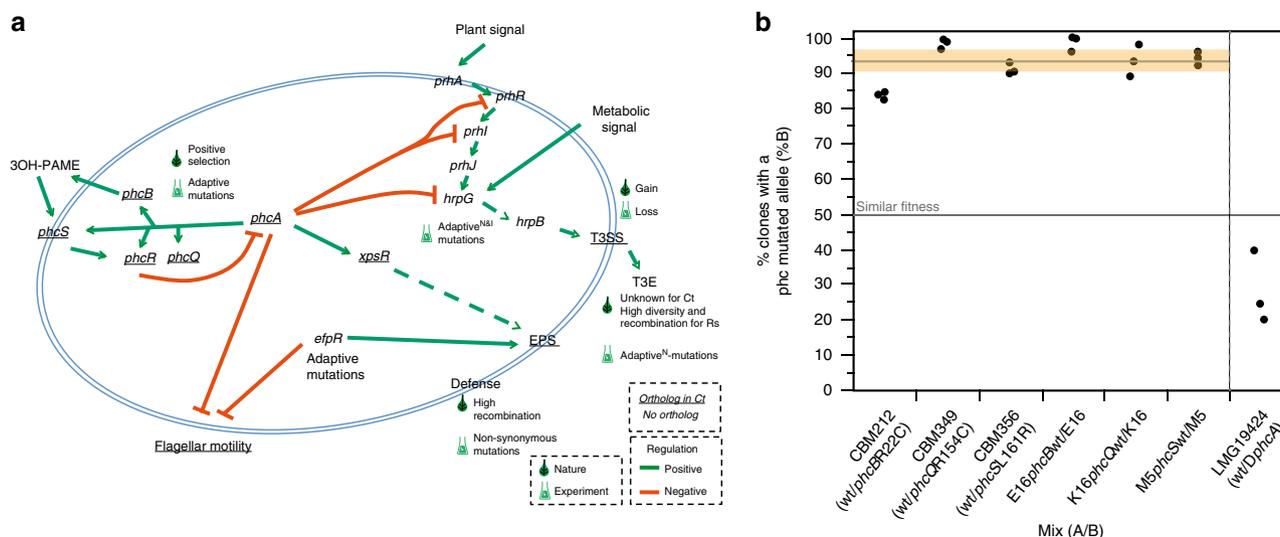
the plasmid. Importantly, the original chimera had similar ex planta survival rates with and without the plasmid (Supplementary Fig. 9), showing that plasmid carriage does not have a fitness cost in this respect (Supplementary Tables 6 and 7). Although we cannot exclude that some events of positive selection in the plasmid may have passed undetected, nor that further symbiotic evolution of *Ralstonia* chimera will involve plasmid mutations, it appears that the genetic changes leading to improvement of the symbiotic traits occurred mainly in the chromosomes of the chimera in the experiment, and of *C. taiwanensis* in nature, not on the plasmid carrying the symbiotic traits.

**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 (*prhI*, *hrpG*, and *xpsR*) by the central regulator PhcA of the cell density system that controls virulence and pathogenicity in *R. solanacearum*<sup>38</sup> (Fig. 4a and Supplementary Data 8). Among this set of genes, mutations in the structural T3SS component *hrcV*, or affecting 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<sup>18,22,23</sup>. Only 10 of the 47 mutated

virulence-associated genes have an ortholog in *C. taiwanensis*: *vsrD*, *xpsR*, *Rsp0736*, *tssD*, 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<sup>18</sup>. 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 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). 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*<sup>39</sup>, 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



**Fig. 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 ref. <sup>38</sup>). Adaptive<sup>N</sup> and adaptive<sup>I</sup> represent the presence of adaptive mutations for nodulation (N) and infection (I), respectively. Underline, genes or factors present in *C. taiwanensis*. The results of the enrichment analyses are in Supplementary Data 6, 8, and 9. **b** Adaptive nature of the *Ralstonia phc* alleles evolved in the experiment and the recruitment of PhcA for symbiosis in the natural symbiont *C. taiwanensis* LMG19424. Each dot represents an independent experiment. The horizontal gray 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's test). The codes of the clones correspond to those indicated in Fig. 1

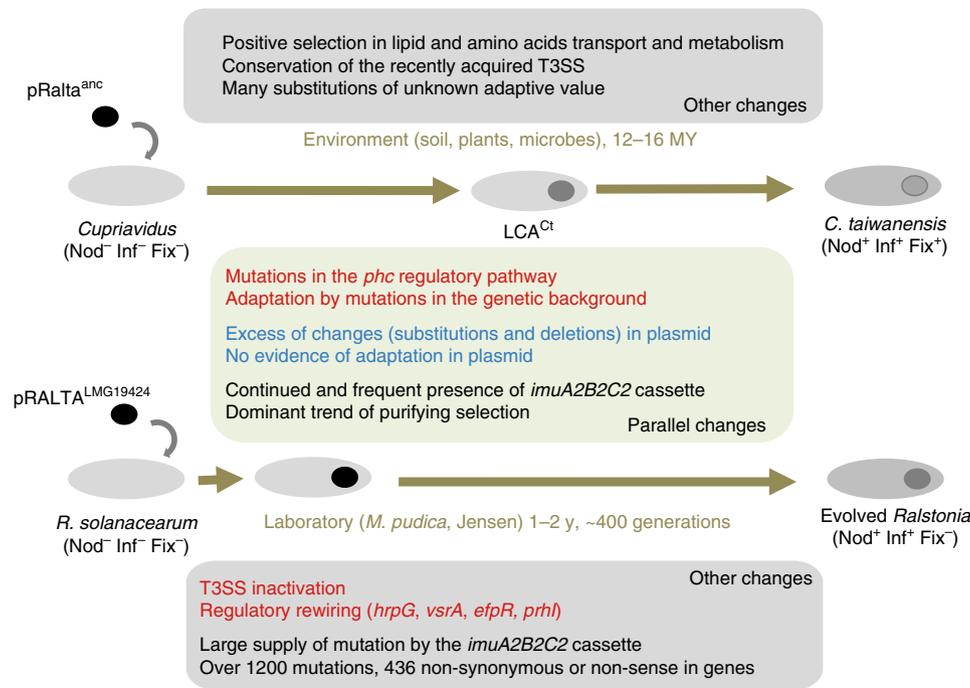
*Cupriavidus* and *Ralstonia* (Supplementary Data 10), regulates a reversible switch between two different physiological states via the repression of the central regulator PhcA in *Ralstonia*<sup>38</sup> and *Cupriavidus*<sup>40</sup>. Interestingly, PhcA-associated genes were also enriched in substitutions 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<sup>Ct</sup> relative to *C. taiwanensis* (“*phcA*-linked” in Supplementary Data 6). Strikingly, only seven genes showing an excess of diversity at bLCA<sup>Ct</sup> had orthologs with mutations in the evolution experiment. Among these seven, only two also showed a signature of positive selection in *C. taiwanensis*: *phcB* and *phcS* (ongoing events, 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<sup>12</sup>, 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<sup>18,22</sup>. 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



**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

high genetic diversity identified in natural populations, e.g., high mutation rates at bLCA<sup>Ct</sup> possibly caused by the *imuA2B2C2* cassette encoding stress-responsive error-prone DNA polymerases. We focus our discussion on the parallels between natural and experimental processes because there is little published data regarding them.

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 to the novel genetic background and environmental conditions. Instead, the abundant substitutions observed in the plasmid seem to have a negligible role in the experimental symbiotic adaptation and lack evidence of positive selection in nature. Moreover, the plasmid did not seem to have a cost in ex planta culture conditions. 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. This is in agreement with proposals that pRalta was acquired from *Burkholderia*<sup>41</sup>, which are ancient symbionts of *Mimosa* spp.<sup>42</sup>. 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 testimony of the ability of mobile genetic elements to seamlessly plug 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<sup>18</sup>, and that subsequent adaptation favored the re-use of regulatory modules leading to massive metabolic and transcriptomic changes<sup>22</sup>. 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<sup>38,43,44</sup>. Here, from the analysis of orthologs between *Ralstonia* chimera 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<sup>45</sup>. Mutations in this pathway are unlikely to cause adaptation by attenuating the virulence of *Ralstonia*, because the chimeric ancestor is not pathogenic on *M. pudica*<sup>18</sup>. 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.

Very controlled experimental evolution studies show few identical mutations between replicates and require the aggregation of mutations in genes, operons or pathways to identify commonalities<sup>25</sup>. This was the case of the initial evolution of three clones nodulating *Mimosa*, which occurred by three mutations in two different genes of the virulence pathway of *R. solanacearum*<sup>18</sup>. Here, the comparison of the natural evolution of *Mimosa* symbionts in the *Cupriavidus* genus and the experimental symbiotic evolution of *Ralstonia* under *M. pudica* selection pressure could not reveal parallel changes at the nucleotide level because of the high diversity of natural populations. Yet, it showed that symbiotic adaptation occurred in the recipient genome, in a selection regime where most non-synonymous mutations were purged (purifying selection) but some were adaptive and allowed the continuous improvement of the symbiotic traits in the experiment and, presumably given its efficiency, in nature. Adaptive changes affected homologous genes implicated in central regulatory pathways in both processes highlighting how natural selection can lead to the co-option of homologous structures in different lineages submitted to similar evolutionary challenges. The observation of these parallels

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 chimera, including 21 bacterial clones (three ancestors and 18 evolved clones)<sup>20</sup>. 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 validated 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 *C. taiwanensis* LMG19424 strain as previously described<sup>20</sup>, 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<sup>46</sup>. 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 unlabeled PCR fragment of ca. 6 kb carrying the point mutation to introduce, as previously described<sup>22</sup>. 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<sup>47</sup>. 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 mutants 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 Supplementary Table 9.

**Relative in planta fitness.** *Mimosa pudica* seeds from Australia origin (B&T World Seed, Pagnignan, France) were surface sterilized for 15 min in 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 Fahraeus solid medium<sup>48</sup> and 50 ml of 1/4th strength of Jensen liquid medium<sup>49</sup>. To measure the in planta relative fitness, a mix of two strains bearing different antibiotic resistance genes or fluorophores ( $5 \times 10^5$  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 crushes 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 GMI1000 (RCM1068) or GMI1000-pRalta (RCM1069) were grown overnight in rich BG medium<sup>10</sup> and  $10^7$  bacteria were inoculated to Gibson tubes filled with quarter-strength Jensen medium<sup>24</sup> alone or containing in addition two *M. pudica* plantlets. Bacteria were counted by plating. Twelve independent experiments were performed.

**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/agg/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 Supplementary Data 3. Genomes of  $\alpha$ - and  $\beta$ -Proteobacteria larger than 1 Mb and genomes of phages were downloaded from GenBank RefSeq as available in February 2013.

**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. oxalaticus* strain (LMG2235) (Supplementary Data 3), were sequenced at the GeT-PlaGe core facility, INRA Toulouse (get.genotoul.fr). DNA-seq libraries were prepared according to Bioscientific's protocol using the Bioscientific PCR-free Library Prep Kit. Briefly, DNA was fragmented by sonication, size selection was performed using CLEANNA CleanPCR beads and adaptors 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  $2 \times 100$  bp with the HiSeq v3 reagent kit (LMG2235 and LMG19431) or on an Illumina MiSeq sequencer using a paired-end read length of  $2 \times 300$  bp 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 (<https://zenodo.org/record/1239599#.WumbmtaxWqq>))<sup>50</sup>. 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<sup>51</sup>, IDBA-UD<sup>52</sup>, Canu<sup>53</sup>, and Newbler<sup>54</sup>. After the assembly step, an attempt to fill scaffolds/contigs gaps is performed using the gapcloser software from the SOAPdenovo2 package<sup>55</sup>. 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 app format (v2.0).

The genomes were subsequently processed by the MicroScope pipeline for complete structural and functional annotation<sup>56</sup>. Gene prediction was performed using the AMIGene software<sup>57</sup> and the microbial gene finding program Prodigal<sup>58</sup> known for its capability to locate the translation initiation site with great accuracy. The RNAMmer<sup>59</sup> and tRNAscan-SE<sup>60</sup> programs were used to predict rRNA and tRNA-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 dsDNA 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 $\times$  AMPure PB beads (Pacbio). Ten micrograms of DNA was purified then sheared at 40 kb using the meraruptor system (diagenode). A DNA and END damage repair step was performed on 5  $\mu$ g 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 13–15 kb cutoff was performed on the BluePippin Size Selection system (Sage Science) with the 0.75% agarose cassettes, Marker S1 high Pass 15–20 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 10:1. Then after a magnetic bead-loading step (OCPW), SMRTbell libraries were sequenced on RSII instrument at 0.2 nM 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<sup>61</sup>.

**Core genomes.** Core genomes were computed using reciprocal best hits (hereafter named RBH), using end-gap free Needleman–Wunsch global alignment, between the proteome 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 proteomes<sup>62</sup>. 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 a neighborhood of ten genes (five up- and five downstream). We made several sets of core genomes (see Supplementary Fig. 2A): all the 89 strains (A1), 44 *C. taiwanensis* (Ct), Ct with the closest outgroup (C2), Ct with the five closest outgroups (C3), the whole 60 genomes of the genus *Cupriavidus* (Cg), and the 14 genomes of *R. solanacearum* (Rs). They were defined as the intersection of the lists of positional orthologs between the relevant pairs of genomes and the pivot (Supplementary Data 2).

**Pan genomes.** Pan genomes describe the full complement of genes in a clade and were computed by clustering homologous proteins in gene families. Putative homologs between pairs of genomes were determined with blastp v2.2.18 (80% coverage), and evalues (if smaller than  $10^{-4}$ ) were used to infer protein families using SiLiX v1.2.8 (<http://lbe.univ-lyon1.fr/SiLiX>)<sup>63</sup>. To decrease the number of paralogs in pan genomes, we defined a minimal identity threshold between homologs for each set. For this, we built the distribution of identities for the positional orthologs of core genomes between the pivot and the most distant

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 v3.8.31<sup>64</sup>, 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 v1.3.8<sup>65</sup>. The best model was determined using the Bayesian information criterion. Maximum likelihood trees were then computed with IQ-TREE v1.3.8 using the appropriate model, and validated via a ultrafast bootstrap procedure with 1000 replicates<sup>66</sup> (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 *Cupriavidus* genera analyzed in this study and of ten outgroup genomes of  $\beta$ -Proteobacteria. For this, we made a multiple alignment of the 16S rRNA genes with INFERNAL v.1.1 (with default parameter)<sup>67</sup> using RF00177 Rfam model (v.12.1)<sup>68</sup>, followed by manual correction with SEAVIEW to remove 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<sup>Ct</sup>, we computed the distances in the 16S rRNA genes tree between each strain and each of the nodes delimitating the branch bLCA<sup>Ct</sup> (respectively LCA<sup>Ct</sup> and C2 in Fig. 2). The substitution rate of 16S rRNA genes in enterobacteria was estimated at ~1% per 50 MY of divergence<sup>69</sup>, 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 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. We searched for homologs of each gene in the reference genome in the other genomes using LAST v744<sup>70</sup> and a score penalty of 15 for frameshifts. We discarded hits with e-values 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 using birth–death models.** We used Count (version downloaded in December 2015)<sup>71</sup> 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 (Supplementary Data 2), and selected the best model using the Akaike information criterion (Supplementary Table 10). 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 presence, gain, loss, reduction, and expansion for the branch leading to the last common ancestor (LCA) of *C. taiwanensis* (LCA<sup>Ct</sup>).

**Measures of similarity between genomes.** For each pair of genomes, we computed two measures of similarity, one based on gene repertoires and another based on the sequence similarity between two genomes. The gene repertoire relatedness was computed as the number of positional orthologs shared by two genomes divided by the number of genes in the smallest one<sup>72</sup>. Pairwise average nucleotide identities (ANIb) were calculated using the pyani Python3 module (<https://github.com/widdowquinn/pyani>), with default parameters<sup>73</sup>. We used single-linkage clustering to group strains likely to belong to the same species. This was done constructing a transitive closure of sequences with an ANIb higher than a particular threshold (i.e., >94, 95, or 96%). We used BioLayout Express<sup>3D</sup> to visualize the graphs representing the ANIb relationships and the resulting groups for each threshold (Supplementary Fig. 3).

**Inference of recombination.** We identified recombination events using three different approaches. We used the PHI test to look for incongruence within each core gene multiple alignment (Ct and C3 datasets). We made 10,000 permutations to assess the statistical significance of the results<sup>74</sup>. We used the SH-test, as implemented in IQ-TREE v1.3.8<sup>65</sup> (GTR + I + G4 model, 1000 REL replicates), to identify incongruence between the trees of each core gene and that of the concatenated multiple alignment of all core genes. We used ClonalFrameML v10.7.5<sup>75</sup> to infer recombination and mutational events in the branch leading to the LCA<sup>Ct</sup> using the phylogenetic tree of C3 (Supplementary Data 2). The transition/transversion ratios given as a parameter to ClonalFrameML were estimated

with the R package PopGenome v2.1.6<sup>76</sup>. Lastly, ClonalFrameML was also used to compare the relative frequency of recombination and mutation on the whole concatenated alignments of Ct and Rs.

**Molecular diversity and adaptation.** Positive selection was identified using likelihood ratio tests by comparing the M7 (beta)–M8 (beta&w) models of codeml using PAML v4.8<sup>77</sup>. We used the independent phylogenetic tree of each gene family to avoid problems associated with horizontal transfer (since many genes failed the SH-test for congruence with the core genome phylogenetic tree). We removed from the analysis gene families that had incongruent phylogenetic signals within the multiple alignment<sup>78</sup>. 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 LCA<sup>Ct</sup> 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<sup>Ct</sup> and LCA<sup>C2</sup> (accounting for recombination). Then, we estimated nucleotide diversity of each core gene for Ct, and between LCA<sup>Ct</sup> and LCA<sup>C2</sup> 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<sup>79</sup> as available on the MicroScope Platform (<https://www.genoscope.cns.fr/agg/microscope/home/>) to class genes according to the COG categories (Supplementary Table 4 and Data 5). Protein subcellular localizations were predicted using PSORTb v3.0.2 (<http://www.psorth.org/psorth/>)<sup>80</sup>. Transporters and regulatory proteins were inferred using TransportDB (<http://www.membranetransport.org/>)<sup>81</sup> and P2RP (<http://www.p2rp.org/>)<sup>82</sup>, respectively. Protein secretion systems were identified using TXSScan (<http://mobyle.pasteur.fr/cgi-bin/portal.py#forms:txsscan>)<sup>83</sup>. We manually checked and corrected the lists. Specific annotations were also defined for (i) *R. solanacearum* GMI1000: Type III effectors<sup>84</sup>, PhcA-associated genes (i.e., genes involved in the upstream regulatory cascade controlling the expression of phcA, and genes directly controlled by PhcA)<sup>38,85,86</sup>, virulence<sup>87</sup>, extracellular polysaccharides (EPS)<sup>86,88</sup>, chemotaxis<sup>89</sup>, twin-arginine translocation pathway (Tat)<sup>90</sup>, Tat-secreted protein<sup>91</sup>, and (ii) the pRAlta of *C. taiwanensis* LMG19424: symbiotic genes<sup>19</sup>, genes pertaining to plasmid biology (conjugation, replication, partition, based on the annotations<sup>92</sup>), and operons using ProOpDB (<http://operons.ibt.unam.mx/OperonPredictor/>)<sup>93</sup>. Lastly, we also annotated positional orthologs between *R. solanacearum* GMI1000 and *C. taiwanensis* LMG19424 according to specific annotations used for both strains (Supplementary Table 4 and Data 5).

**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). These data were then used to define intervals of confidence around the average values observed in the simulations.

In addition, we used a similar approach to estimate the significance of parallelisms between evolved clones for SNP. We computed 99 random experiments of genome evolution using the mutation spectrum and the total number of SNP of each replicon for the 18 final clones and the three ancestors using a previously published method<sup>17</sup>. Then, we estimated how many nucleotide positions, genes, and COGs were affected in more than one clone for each random experiment. Lastly, we draw the distributions of the expected proportions of parallelisms to define intervals of confidence.

**Statistical analyses.** In order to identify genes that evolved faster in the branch leading to LCA<sup>Ct</sup>, we compared the nucleotide diversity of sequences for LCA<sup>Ct</sup> and LCA<sup>C2</sup> 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 Supplementary Table 11) 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<sup>94</sup> (`{stats}`, `Pz.adjust`).

Statistical analyses with R were done using version 3.1.3 (R: a language and environment 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

- Kawecki, T. J. et al. Experimental evolution. *Trends Ecol. Evol.* **27**, 547–560 (2012).
- Barrick, J. E. & Lenski, R. E. Genome dynamics during experimental evolution. *Nat. Rev. Genet.* **14**, 827–839 (2013).
- 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).
- 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).
- 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).
- Ochman, H. & Moran, N. A. Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science* **292**, 1096–1099 (2001).
- Ben-Gurion, R. & Shafferman, A. Essential virulence determinants of different *Yersinia* species are carried on a common plasmid. *Plasmid* **5**, 183–187 (1981).
- Bernardini, M. L., Mounier, J., d’Hauteville, H., Coquis-Rondon, M. & Sansonetti, P. J. Identification of *icsA*, 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).
- Okinaka, R. T. et al. Sequence and organization of pXO1, the large *Bacillus anthracis* plasmid harboring the anthrax toxin genes. *J. Bacteriol.* **181**, 6509–6515 (1999).
- 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).
- Escobar-Páramo, P. et al. A specific genetic background is required for acquisition and expression of virulence factors in *Escherichia coli*. *Mol. Biol. Evol.* **21**, 1085–1094 (2004).
- 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).
- 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. USA* **92**, 8985–8989 (1995).
- Moulin, L., Béna, 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).
- 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).
- 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).
- 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).
- Marchetti, M. et al. Experimental evolution of a plant pathogen into a legume symbiont. *PLoS Biol.* **8**, e1000280 (2010).
- Amadou, C. et al. Genome sequence of the -rhizobium *Cupriavidus taiwanensis* and comparative genomics of rhizobia. *Genome Res.* **18**, 1472–1483 (2008).
- 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).
- Marchetti, M. et al. Shaping bacterial symbiosis with legumes by experimental evolution. *Mol. Plant Microbe Interact.* **27**, 956–964 (2014).
- 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).
- Guan, S. H. et al. Experimental evolution of nodule intracellular infection in legume symbionts. *Isme J.* **7**, 1367–1377 (2013).
- Remigi, P. et al. Transient hypermutagenesis accelerates the evolution of legume endosymbionts following horizontal gene transfer. *PLoS Biol.* **12**, e1001942 (2014).
- Tenaillon, O. et al. The molecular diversity of adaptive convergence. *Science* **335**, 457–461 (2012).
- 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).
- 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).
- Chen, W.-M. et al. Legume symbiotic nitrogen fixation by  $\beta$ -proteobacteria is widespread in nature. *J. Bacteriol.* **185**, 7266–7272 (2003).
- Andam, C. P., Mondo, S. J. & Parker, M. A. Monophyly of *nodA* and *nifH* genes across Texan and Costa Rican populations of *Cupriavidus* nodule symbionts. *Appl. Environ. Microbiol.* **73**, 4686–4690 (2007).
- 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).
- Rocha, E. P. C. et al. Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J. Theor. Biol.* **239**, 226–235 (2006).
- Kuo, C.-H., Moran, N. A. & Ochman, H. The consequences of genetic drift for bacterial genome complexity. *Genome Res.* **19**, 1450–1454 (2009).
- Paterson, S. et al. Antagonistic coevolution accelerates molecular evolution. *Nature* **464**, 275–278 (2010).
- Beaumont, H. J. E., Gallie, J., Kost, C., Ferguson, G. C. & Rainey, P. B. Experimental evolution of bet hedging. *Nature* **462**, 90–93 (2009).
- Barrick, J. E. et al. Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* **461**, 1243–1247 (2009).
- Perfeito, L., Fernandes, L., Mota, C. & Gordo, I. Adaptive mutations in bacteria: high rate and small effects. *Science* **317**, 813–815 (2007).
- Guidot, A. et al. Multihost experimental evolution of the pathogen *Ralstonia solanacearum* unveils genes involved in adaptation to plants. *Mol. Biol. Evol.* **31**, 2913–2928 (2014).
- Genin, S. & Denny, T. P. Pathogenomics of the *Ralstonia solanacearum* species complex. *Annu. Rev. Phytopathol.* **50**, 67–89 (2012).
- 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).
- 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).
- 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).

42. Bontemps, C. et al. *Burkholderia* species are ancient symbionts of legumes. *Mol. Ecol.* **19**, 44–52 (2010).
43. 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).
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**, e1006044 (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 PhcA-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 natural transformation. *Proc. Natl. Acad. Sci. USA* **111**, 8937–8942 (2014).
47. Daubech, B. et al. Spatio-temporal control of mutualism in legumes helps spread symbiotic nitrogen fixation. *eLife* **6**, e28683 (2017).
48. Fahraeus, G. The infection 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**, 98–108 (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 adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
54. Margulies, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
55. Luo, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18 (2012).
56. Vallenet, D. et al. MicroScope—an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. *Nucleic Acids Res.* **41**, D636–D647 (2013).
57. Bocs, S. AMIGene: annotation of microbial genes. *Nucleic Acids Res.* **31**, 3723–3726 (2003).
58. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 119 (2010).
59. Lagesen, K. et al. RNAMmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
60. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
61. Mariette, J. et al. NG6: integrated next generation sequencing storage and processing environment. *BMC Genom.* **13**, 462 (2012).
62. Rocha, E. P. C. Inference and analysis of the relative stability of bacterial chromosomes. *Mol. Biol. Evol.* **23**, 513–522 (2005).
63. Miele, V., Penel, S. & Duret, L. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinform.* **12**, 116 (2011).
64. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
65. 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).
66. Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* **30**, 1188–1195 (2013).
67. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
68. Nawrocki, E. P. et al. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.* **43**, D130–D137 (2015).
69. Ochman, H. & Wilson, A. C. Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. *J. Mol. Evol.* **26**, 74–86 (1987).
70. Sheelin, S. L., Park, Y., Frith, M. C. & Spouge, J. L. Frameshift alignment: statistics and post-genomic applications. *Bioinformatics* **30**, 3575–3582 (2014).
71. Csuros, M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* **26**, 1910–1912 (2010).
72. Snel, B., Bork, P. & Huynen, M. A. Genome phylogeny based on gene content. *Nat. Genet.* **21**, 108–110 (1999).
73. Richter, M. & Rosselló-Móra, R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. USA* **106**, 19126–19131 (2009).
74. Bruen, T. C. A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172**, 2665–2681 (2005).
75. Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput. Biol.* **11**, e1004041 (2015).
76. 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).
77. 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).
78. 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).
79. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).
80. 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).
81. 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).
82. Barakat, M., Ortet, P. & Whitworth, D. E. P2RP: a web-based framework for the identification and analysis of regulatory proteins in prokaryotic genomes. *BMC Genom.* **14**, 269 (2013).
83. Abby, S. S. et al. Identification of protein secretion systems in bacterial genomes. *Sci. Rep.* **6**, 23080 (2016).
84. Peeters, N. et al. Repertoire, unified nomenclature and evolution of the Type III effector gene set in the *Ralstonia solanacearum* species complex. *BMC Genom.* **14**, 1 (2013).
85. 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).
86. 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).
87. 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).
88. 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).
89. 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).
90. Poueymiro, M. & Genin, S. Secreted proteins from *Ralstonia solanacearum*: a hundred tricks to kill a plant. *Curr. Opin. Microbiol.* **12**, 44–52 (2009).
91. Gonzalez, E. T., Brown, D. G., Swanson, J. K. & Allen, C. Using the *Ralstonia solanacearum* Tat secretome to identify bacterial wilt virulence factors. *Appl. Environ. Microbiol.* **73**, 3779–3786 (2007).
92. Cury, J., Touchon, M. & Rocha, E. P. C. Integrative and conjugative elements and their hosts: composition, distribution and organization. *Nucleic Acids Res.* **45**, 8943–8956 (2017).
93. Taboada, B., Ciria, R., Martinez-Guerrero, C. E. & Merino, E. ProOpDB: Prokaryotic Operon DataBase. *Nucleic Acids Res.* **40**, D627–D631 (2012).
94. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).

## Acknowledgements

This work was supported by funds from the French National Research Agency (ANR-12-ADAP-0014-01 and ANR-16-CE20-0011-01) the “Laboratoire d’Excellence (LABEX) 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 (contrat ANR-10-INBS-09). We thank Olaya Rendueles, Rémi Peyraud, Ludovic Cottret, 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 GeT 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.To., 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.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018