bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Ancestral genome reconstruction reveals the history of ecological diversification in *Agrobacterium*.

Florent Lassalle[1,2,3,4,5,6,a], Rémi Planel[1,2,5], Simon Penel[1,2,5], David Chapulliot[1,2,3,4], Valérie Barbe[7], Audrey Dubost[1,2,3], Alexandra Calteau[7,8,9], David Vallenet[7,8,9], Damien Mornico[7,8,9,b], Laurent Guéguen[1,2,5], Ludovic Vial[1,2,3], Daniel Muller[1,2,3], Vincent Daubin[1,2,5], Xavier Nesme[1,2,3,4].

[1] *Université de Lyon, 69361 Lyon, France*

[2] *Université Lyon 1, 69622 Villeurbanne, France*

[3] *CNRS, UMR5557, Ecologie Microbienne, 69622 Villeurbanne, France*

[4] *INRA, UMR 1418, Ecologie Microbienne, 69622 Villeurbanne, France*

[5] *CNRS, UMR5558 Biométrie et Biologie Evolutive, 69622 Villeurbanne, France*

[6] *Ecole Normale Supérieure de Lyon, 69342 Lyon, France*

[7] *Direction des Sciences du Vivant, Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA), Institut de Génomique, Genoscope, 2 rue Gaston Crémieux, Evry F-91057, France.*

[8] *CNRS, UMR 8030, Laboratoire d'Analyse Bioinformatiques pour la Génomique et le Métabolisme , 2 rue Gaston Crémieux, 91057 Evry, France*

[9] *UEVE, Université d'Evry Val d'Essonne, boulevard François Mitterrand, 91025 Evry, France*

[a] Current address: *University College London, Gower Street, London WC1E 6BT, United Kingdom*

[b] *Current address: Bioinformatics and Biostatistics Hub, Institut Pasteur, 25-28, rue du Dr Roux 75724 Paris, France*

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Both the ecological adaptations of an organism and its evolutionary history are recorded in a genome. We used *Agrobacterium* biovar 1, a diverse group of plant-associated bacteria as a model to search for genomic signatures of ecological adaptation in relation to diversification. We designed a new phylogenetic approach accounting for horizontal transfer and duplication to reconstruct the evolutionary history of *Agrobacterium* genomes and infer ancestral gene contents. We find that allowing for genes to be co-transferred and co-duplicated significantly improves scenarios of genome evolution. Most genes acquired before the diversification of major clades within *Agrobacterium* are organized in blocks of co-evolving genes encoding coherent pathways. This pattern of gene co-evolution rejects a neutral model of transfer, in which neighbouring genes would be transferred independently of their function. In addition, the conservation of acquired genes is driven by purifying selection on collectively coded functions. Based on this criterion, we identify the genomic determinants of ecological niche diversification within this group. The strong selective role of host plant rhizospheres in the history of *Agrobacterium* is evident from the recurrent acquisition of functions involved in the production of secreted secondary metabolites or extracellular matrix, and in the metabolism of plant-derived compounds such as phenolics and amino-acids. Our reconstructed genome history – from single gene trees with transfer and duplication events to blocks of co-evolved genes and functional annotations…– is compiled in an integrative database, Agrogenom, which can be visualized and queried through an interactive web interface accessible at http://phylariane.univ-lyon1.fr/db/agrogenom/3.

Keywords: Reconciliations, ancestral genome, gene transfer, bacterial cladogenesis, reverse ecology, *Agrobacterium tumefaciens*.

Background

Our understanding of the ecology of bacteria is fragmentary. We usually know a subset of the environments from which a species can be sampled, some laboratory conditions in which they can be grown, and sometimes the type of interactions they establish with other organism. We now also have their genomes, which we believe contain all the information that make their lifestyle possible. However, even if we could determine the functional role of every base in a genome, it would not necessarily allow us to understand whether this function is significant in its prevalent environment [1]. What can be learned however, is something about the past. By reconstructing genome evolution, we can discern the events that have been decisive in the evolution of species. Bacterial genomes are in constant flux, with genes being gained and lost at rates that can exceed nucleotide substitution rate [2]. This dynamics leads to the definition of core versus accessory genomes, which respectively gather the genes that are shared by all members of a species and

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

those that are found in some strains but not all. In *E. coli* for example, the core genome comprises 1,800 genes while the accessory genome is more than 20,000 genes, with two random strains differing typically by a thousand genes [3]. Some accessory genes are frequently gained by transfer and then quickly lost, leaving patterns of presence in genome that are inconsistent with the species phylogeny; many are only found in one genome. Among accessory genes, most likely have ephemeral, if any, adaptive value for the bacteria, and are only selfish elements caught by the snapshot of genome sequencing [4]. However, this highly dynamic process also allows accessory genes to settle in genomes, and become part of the core genome of a lineage. Such 'domestication' events constitute the most remarkable deviations from a neutral model where rapid gains and loss prevail. Clade-specific conservation is suggestive of purifying selection acting on the genes, possibly reflecting the adaptation of their bacterial host to a particular ecological niche.

In a previous study, we explored the diversity of gene repertoires among strains of *Agrobacterium* biovar 1 that contains several bona fide but yet unnamed 'genomic' species G1 to G9 and G13, collectively named '*Agrobacterium tumefaciens* species complex' (*At*) according to the proposal of Costechareyre et al. [5]. We found that genes specific to the species under focus, *i.e.* G8 now called *A. fabrum*, [6], were in majority clustered in the genome, and that these clusters gathered genes that encoded coherent biological functions. The conservation of co-functioning genes in genomic clusters appears unlikely in the context of frequent gene turnover. This pattern could be a trace of purifying selection acting to conserve the gene clusters in their wholeness, because the selected unit is the function collectively encoded by the constituent genes. However, it could also result from the neutral process of gene flow, by which neighbour genes that happen to have related functions, such as operons, are transferred together and then maintained by drift. These hypotheses may however be distinguished by analyzing the historical record of evolutionary events that led to the clustering of co-functioning genes.

Most genes have complex histories, marked by many events of gene duplication, loss and, in the case of microorganisms, horizontal transfers. Different genes have different histories that followed the one of species by different paths, i.e. evolutionary scenarios. The reconstruction of the evolutionary scenarios of all homologous gene families present in the genomes under scrutiny is the only way to recognize adaptive events throughout genome histories. Evolutionary scenarios can be inferred by comparing the phylogenetic history of genes and that of species, and by reconciling their discordances through the explicit inference of

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

events of duplication, transfer and loss. This in turn allows to reconstruct the incremental shaping of genome gene contents, from ancestral to contemporary genomes, and to deduct the functional and ecological consequences of these changes.

We used the Rhizobiaceae family as a model taxon, particularly focusing on the *At* clade for which we have an original genome dataset of 22 strains from ten different species, including 16 new genome sequences. We designed a new phylogenetic pipeline for reconstruction of ancestral genomes that accounts for events of horizontal transfer and duplication of genes and makes use of the regional signal in genome histories to increase the confidence and accuracy in reconstructed evolutionary scenarios. Applied to our dataset, this approach identifies blocks of co-transferred and co-duplicated genes, allowing us to test hypotheses on how co-functioning gene clusters were formed. Comparing the level of functional co-operation of genes within blocks of clade-specific genes to the expectation under a neutral model of gene transfer shows that clade-specific genes are more functionally related than expected. This supports the hypothesis by which the domestication of at least some clade-specific genes results from ecological selection.

Results and Discussion

Phylogenomic Database and Reference Species Tree. To reconstruct the histories of genes within Rhizobiaceae, we built the Agrogenom database. It gathers 47 genomes, from genera *Agrobacterium*, *Rhizobium*, *Sinorhizobium*/*Ensifer*, *Mesorhizobium*/*Chelativorans* and *Parvibaculum*. These genomes contains 281,223 coding sequences (CDSs, or genes hereafter) clustered into 42,239 homologous gene families. Out of these families, 27,547 were singletons with no detectable homologs (ORFan families) and 455 were found in exactly one copy in all the 47 genomes (unicopy core gene families). From the jackknife concatenation of the unicopy core gene set, a species phylogeny was inferred [7] (Fig. S1). Remarkably, significant support is obtained for all clades corresponding to species with several strains: *S. melitoti*, *R. etli*, *R. leguminosarum* and indeed for *Agrobacterium* species G1, G8, G4, G5 and G7. In contrast, the support is relatively low for the relative positioning of strains within species, showing conflicting (or lack of) signal among concatenated genes , as expected if members of the same species are frequently recombining. Within the *At* clade, groupings of higher order had also high jackknife support: G8 with G6 (hereafter named [G6-G8] clade), G5 with G13, ([G5-G13] clade), G1 with [G5-G13] ([G1-G5-G13] clade), G3 with [G1-G5-G13]

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

([G3-G1-G5-G13] clade), G7 with G9 ([G7-G9] clade), and G4 with [G7-G9] ([G4-G7-G9] clade). Only some deep splits such as the position of G2 and of [G6-G8] clade relative to the *At* root were not well supported (Fig S1).

Reconciliation of gene and species histories. We computed a gene tree for all the 10,774 families containing at least three genes. Our aim was to reconcile gene tree topologies with the species tree, i.e. that an event of either origination, duplication, transfer (ODT), or speciation be assigned to each of the 467,528 nodes found in the 10,774 gene trees. Horizontal gene transfers (HGT) is the most challenging to correctly assign due to the diversity of signatures it generates [8]. Many HGT events can be recognized from the topological conflict they cause between gene tree and species tree, while others can be identified from the heterogeneous pattern of gene presence they induce. In general these patterns can also be generated by ancient duplication events followed by multiple subsequent losses. A pipeline combining several phylogenetic analysis tools was thus designed to efficiently distinguish each kind of event (Fig. S2). A first search was made for multiple representations of species in gene trees, identifying 17,569 putative duplications generating 28,343 potential paralogous lineages (Fig. S2, step 1). Then, Prunier, a method relying on strongly supported topological conflict between gene and species trees [9], was run to detect transfer events. In presence of lineage-specific paralogs ('in-paralogs'), Prunier test was applied to each overlapping subtree covering the different sets of co-orthologs, i.e. orthologous groups that contain one of the duplicated pair (Fig. S2, step 2). Paralogous gene lineages have evolved independently and may thus have different topologies and branch supports, leading to the inference of potentially different scenarios of reconciliation by parallel Prunier tests at the same gene tree node. This exploration of the reconciliation space by tests of multiple combinations of co-orthologs provided an additional criterion of support for the scenario inferred at these nodes, with repeated inference indicating high support. In total, Prunier yielded 22,322 high-confidence transfer events. Then, TPMS-XD algorithm [10] was used to explore the remaining topological conflicts that corresponded to weakly supported branches in gene trees. Under the objective of minimizing the number of duplications inferred above these conflicting branches, 1,899 of them were recognized as the place of additional replacing transfer events, allowing a parsimonious decrease of 10,229 counts of duplication events. Finally, the Wagner parsimony algorithm from Count package [11] was applied

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

to detect transfers events that did not induce topological conflict between gene trees and species trees but leaved heterogeneous patterns of presence/absence (Fig. S2, step 4 and 6), revealing 19,553 of those additive transfers. This pipeline thus reconstructed a total of 7,340 duplications (1.5% of all gene tree nodes) and 43,233 transfers (9.2%). The remainder of unannotated gene tree nodes correspond to speciation events (where the gene tree topologies locally follow the species tree) and originations (apparition of the gene family in our dataset, mapped at the root of the gene tree) (Table 1). Thanks to the ancestral genome reconstruction, we could distinguish additive transfers that bring new genes from those that replace already present orthologous genes, the latter accounting for a quarter of total transfers (9,271 events). Additive transfers contribute almost five times more than duplications to the total gene input in genomes (Table S1), showing that transfer is the main source of gene content innovation in *At*.

Regional amalgamation of gene histories provides more accurate scenarios. Large-scale comparative genomics analyses have revealed that insertions in genomes typically comprise several consecutive genes, indicating that blocks of genes can evolve in linkage across genomes [12]. Yet, ODT scenarios are generally evaluated for each gene tree independently of its neighbour [13, 14]. This is problematic because a scenario may be optimal (e.g., more parsimonious) for a gene alone, but sub-optimal in a model where genes can be part of the same ODT events (Fig. 1). We developed a procedure to minimize the number of convergent origination, duplication or transfer events along genomes, allowing to recognize blocks of genes that were part of the same event.

Our method of inference provides a way to refine ambiguous scenarios: on each family taken independently, we infer ODT events and record the uncertainty of each scenario. Because we ignore loss events at this point, the sets of possible branches of the species tree on which ODT events could have happened are relatively large (Fig. 1 A). We then proceed to recognize blocks of neighbour genes that have events mapped to overlapping sets of branches in the species tree. For transfers for instance, donor-recipient coordinates of each single-gene event are compared to those of the gene which is a neighbour in a descendent species (Fig. 1B). If the sets of coordinates overlap, we hypothesize a common event and the set of donor and recipient branches is reduced to the intersecting set of the two families. We then proceed to the next neighbour gene. This common scenario may not be the most parsimonious in losses for individual gene families (Fig. 1 B).

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

However, as this regional pattern is not likely to happen by chance and because ODT events are less frequent than gene loss [15–17] and thus less likely to happen convergently, factorizing ODT events for neighbour genes appears a suitable and relevant procedure to minimize the total number of all kind of events (i.e. ODTL events). By amalgamating compatible ODT scenarios of neighbour genes, we reconstructed 'block events', i.e. unique events involving blocks of co-evolved neighbour genes (Fig. S2 step 8-9). Even though the large majority of transfers involve only one gene, we identified several thousands of transfer events involving short blocks of 2 to 6 genes and hundreds of blocks of a dozen or more consecutive genes in extant genomes (Fig. S7 A). Moreover, reconstructed blocks of ancestral genes that were hypothetically transferred between ancestral genomes appear to have been even much larger than extant ones (Fig. S7 B), showing how frequently rearrangements and partial losses in descendant genomes have dismantled the syntenic blocks involved in ancient transfers. Through the procedure of intersection of neighbour event coordinates, we are able to refine the location in the species tree of a large fraction of duplications and transfer events (Fig. S3): the location in the species tree of 83% of duplications and of 69% of transfer receivers could hence be refined, getting from 70 to 93% duplications assigned to a unique species tree node, and from 60 to 83% of transfers with a unique possible receptor. Integrating the reconciliations genome-wide, we found numerous such block events in *At* genomes, with 17.5% of transfers and 13.3% of duplications involving at least two genes (Table 1). Remarkably, block event scenarios resulted in the decrease of 13,421 ODT events relative to scenarios based on single gene histories and an increase of 2,896 of the total amount of losses, thus showing that the integrated genome-wide scenario with block events was much more parsimonious than the simple sum of all single-gene scenarios (Table 1). The count of additional losses is certainly over-estimated, because block events of gene loss must have occurred. This eventuality was however not taken into account in this study due to the too large uncertainty of location of these events in the species tree, preventing the specific aggregation of blocks of common loss events (see Methods).

We thus inferred scenarios that are more parsimonious when taken globally, and at the same time we significantly reduced the uncertainty on event coordinates in the species tree, and thus could confidently use the reconstructed ancestral genomes for our observation linked to the history of diversification of *At*.

Genome histories reveal lineages with biased fixation of new genes. The reconstructed history of gain and

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

loss in ancestral genomes shows heterogeneous dynamics across the tree of *At.* First, we observe that the sizes of genomes are significantly lower in reconstructed ancestral genomes than in extant genomes (Fig. 2, Table S1). For instance, the reconstructed genome of the *At* ancestor is around 4,500-gene large, when extant genomes have an average size of 5,500. This difference of 1,000 genes corresponds approximately to the number of genes recently gained along the terminal branches of the species tree (Fig. 2), indicating the presence in contemporary genomes of a large polymorphism of gene presence/absence. There is also a consistent excess of gene gain by transfer at leaves of the tree compared to inner nodes. There is also a higher number of losses at leaves, which does not compensate excess gains (2.9-fold more gains and 1.4-fold more losses at leaves than at inner nodes; Student's *t*-tests, $p < 10^{-10}$ and $p < 0.02$, respectively), resulting in an increased rate of net gene gain at the terminal branches of the phylogeny. This trend probably illustrates the fact that among the large number of recently acquired genes, the majority will not be fixed [4], because their adaptive value is at best temporary. Similarly, a large fraction of deletions observed in extant genomes are probably deleterious on the long term: for instance, genes involved in adaptation to environments met occasionally are under periodic selection, and their deletion will not cause the immediate death of the bacterium, but will eventually be counter-selected. In fact, ancestral genomes might have been approximately of the same size than extant ones, but the polymorphic fraction of their gene content has not survived in any current genome, and hence could not be taken into account to reconstruct complete ancestral genomes.

Overall, the variations of gene repertoires showed significant relationships with the evolutionary distances in the species tree, indicating that on the long run, rather steady evolutionary processes were operating in these genomes. For instance, the length of the branch leading to the ancestor best explained the quantity of genes gained and lost by an ancestor (linear regression, $r^2 = 0.59$ and 0.32 for gains and losses, respectively), but removing the extreme point of node N35 (i.e. the G1 ancestor) drops the correlations ($r^2 = 0.27$ and 0.28) (Fig. S4 A, B). Interestingly, the quantity of genes gained by an ancestor and subsequently conserved in the descendant clade was robustly explained by the age of the ancestor ($r^2 = 0.39$, or 0.41 when removing N35) (Fig. S4 F). This relationship was better described by a decreasing exponential regression ($r^2 = 0.51$, or 0.50 when removing N35), which reflects a process of 'survival' of genes in genomes through time (Fig. 3). We could recognize outlier genomes in this process of 'gene survival', as the nodes having the largest residuals in the exponential regression (out of the 95% confidence interval). They were, in a decreasing order of excess

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of conservation relative to their age, the ancestors of [G6-G8], G1, G5, [G5-G13], G8 (respectively corresponding to species tree nodes N27, N35, N39, N34 and N32) and subclades of G4 and G7 (nodes N43 and N46) (Fig. S4 F, Fig. S5). These excesses of conservation do not systematically reflect a particular excess of gains in the ancestors: ancestors of G1 and G8 (nodes N35 and N32) have indeed gained more genes than predicted by their respective branch lengths, but on the contrary ancestors of [G6-G8], [G5-G13] and G5 (nodes N27, N34 and N39, respectively) have rather lost genes in excess (Fig. S4 C, D). In the latter cases, the excesses of conserved gains may thus stem from a fixation bias like natural selection for new genes. The outliers that fall above this trend – those clades that conserved more genes than predicted by their age – strikingly all belong to [G1-G5-G13] and [G6-G8] (Fig. S5). The higher rate of conservation in these clades is indicative of a higher proportion of genes having been under purifying selection since their ancestral acquisitions, potentially because these new genes coded functions positively selected in the context of a specific ecology.

Clade-specific genes conserved over long times are likely providing a strong adaptive feature to their host organism. Some adaptive trait can improve the host fitness independently of its ecological niche, and it is expected to spread among close relatives [18]. Conversely, a new trait may prove advantageous as it allows the organism to escape competition from cognate species by increasing the differentiation of its ecological niche, for instance by allowing the exclusive consumption of a resource [19] or the change in relative reliance on a set of resources [20]. Recognizing such niche-specifying determinants among clade-specific gene sets is thus the key to the understanding of the unique ecological properties of a bacterial clade.

Clusters of clade-specific genes are under purifying selection for their collective function. Niche-specifying traits are expected to provide higher differential fitness if less likely to be already present in, or independently acquired by competing relatives. Hence, best candidates for niche-specifying traits consist of novel and complex sets of biochemical functions that do not rely on pre-existing functions. In such a case, it is crucial that the complete set of underlying biochemical functions is gained at once to provide any advantage. Such event can typically happen with the co-transfer of a complete operon. In a previous study focussed on G8 genomes [6], we observed that clade-specific genes tend to occur in clusters of genes with related biochemical function. This apparently non-random pattern suggests that co-transferred groups of

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

genes collectively coding a function have been selected among incoming transferred genes, either by positive selection upon reception, or afterwards by negative (purifying) selection against the destruction of the group by rearrangement or partial deletion. This putative signature of selection made clusters of co-functioning clade-specific genes good candidates as niche-specifying determinants [6].

Though, it is well known that bacterial genomes are organized in functional units such as operons, super-operons, etc. [21], and the co-transfer of cooperating genes could neutrally result from the functional structure of the donor genomes. However, the segments of DNA that are transferred are most probably taken randomly from the donor genomes (apart from the special case of genes coding their own mobility). Thus, under a neutral model, co-transferred genes should not always be co-functioning, and the probability at which a transferred fragment spanned a functional element like an operon would resemble that of any similarly sized fragment of the donor genome.

To test whether clustering of functionally related clade-specific genes results from selection, we designed tests that assess the relation between transfer history and biological function of genes. To do so, we define the degree of functional homogeneity (*FH*) of genes within blocks of neighbour genes based on their Gene Ontology annotation (see Methods). First, we checked that groups made of *n* genes each taken at random in the genome (physically distant genes) had lower *FH* than groups of *n* neighbour genes (taking all possible gene windows of *n* genes), confirming that *FH* captures the functional structure of a genome (Fig. 4 A).

We then compared random groups of neighbour genes without a shared transfer history to blocks of transferred genes of the same size. The distribution of *FH* shows that while in general blocks gather genes that do not encode related functions or for which functional annotations are insufficient (FH ~ 0), transferred blocks of genes presented a minor fraction of intermediate and high functional relatedness (e.g. in G4-B6 genome, *FH* ~ 0.35 and *FH* ~ 0.75, Fig. 4 A). Transferred blocks had significantly higher *FH* than random groups in 45/49 significant tests preformed on independent combination of genomes and block sizes (Fig. 4 A, C). This shows that fixation of transferred blocks of genes in genomes is biased towards blocks that code functional partners in a biological process. This observation supports the hypothesis of positive selection favouring the fixation in a recipient genome of those transferred blocks that can immediately provide a selectable function. It is also compatible with the model of 'selfish operon' proposed by Lawrence and Roth [22], as transfer followed by selection for readily functional multi-genic traits would lead to the prevalence in

10

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

host genomes of genes clustered into tightly linked functional units.

In addition, we observe that among the groups of genes acquired by transfer, those that were conserved in all descendants of the recipient ancestors had more coherent annotated functions than non-conserved ones (11/13 significant tests positive, Fig. 4 B, D). The hypothesis of conserved co-transferred genes encoding more related function than non-conserved ones was previously proposed based on manual inspection of the functional relatedness of a few transferred operons in *E. coli* [23]. The present study presents a quantitative estimation of functional relatedness within transferred blocks of genes, and provides a statistical argument for purifying selection enforcing their collective conservation in genomes. This supports our initial hypothesis that clusters of clade-specific genes participating to a same pathway, more likely to carry the sufficient information to encode a new adaptive trait, were under continued selection since their acquisition. The adaptations that characterize the ecological niche of a clade should thus be revealed through the identification of the genes specific to a clade, and notably those grouped in clusters with related functions.

Clade-specific genes: insights into the possible ecological speciation of clade ancestors. The nature of putative ecological specialization is not obvious for agrobacteria, which are ubiquitous soil-dwellers. The frequent co-occurrence in soil of the different species of *Agrobacterium*, sometimes in the same micro-metric sample [24], dictates under the competitive exclusion principle [25] that they have distinct ecologies, but they only seem to differ by cryptic combinations of environmental factors. Though, some soils and/or host plant show preferential colonization by certain species [5], and G2 members appear to be specialized towards opportunistic pathogenicity in human [26], showing the existence of some kind of niche differentiation among *Agrobacterium* species. Because the clade-specific genes must encode what makes a species' ecology to be distinct from that of its relatives [19], we explored the histories of gene gain and loss in the clades of *At*, looking for synapomorphic presence/absence of genes in *At* clades. We used an automated method for recognition of profiles of contrasted occurrence of genes between related clades, by spotting ancestral gene gains or losses that resulted in conserved presence or absence in the descendant clade (see Methods). Doing so, we could identify parallel gains/losses of orthologous genes in distant clades, notably in case of transfer from one clade ancestor to another. This could reveal the specific sharing of genes between non-sister species

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of *At*. Listings of clade-specific genes of those key *At c*lades can be found in Sup. File S1, or can be browsed on Agrogenom database website [27]. A subset of clades of the *At* phylogeny, namely species G1, G8, G4 and G7, are represented by several closely related extant genomes, and for this reason were particularly amenable for the study of clade-specific gene repertoires. Description of the clade-specific functions encoded in these repertoires can be found in the Supplementary Material (Supplementary Text section 7). Certain classes of functions are recurrently found, including the transport and metabolism of phenolic compounds, and amino-acids and complex sugars, and the production of exopolysaccharide and siderophores, all of which can be related to the life in the plant rhizosphere [6]. Focusing on densely sampled species, chromosomal maps of species-specific genes (Figures 5, 6, S11 and S12) show they were located unevenly on the various replicons of *At* genomes, with a bias towards accumulation on the linear chromid (Lc), and an unexpected presence on the At plasmids (pAt). More generally, clade-specific genes were often located in relatively large clusters encoding coherent biochemical functions or pathways, which are summarized Table S2 and hereafter numbered with the AtSp prefix. Those clade-specific gene clusters often match block transfer or duplication events identified above (Sup. File S1), although often with several transfer blocks covering a clade-specific cluster. This suggests block gain events are likely to cluster at the same loci. Alternatively, it could indicate the limitation of our search procedure in face of the complexity of the gene histories, with different patterns of multiple consecutive transfers in different gene families preventing recognition of their common history.

Ecological adaptations in *Agrobacterium*: a history of variation of shared traits. Here, we try to relate the integrated set of clade-specific functions to a clade-specific ecology [19], using the clustering and functional coherence of clade-specific genes as a signature of purifying selection for their ecological importance. Using our criterion of synapomorphic patterns of gene occurrence – genes gained/lost in a clade ancestor and conserved as present/absent in all extant members of the clade – we could recognize potential adaptive events among a wider set of gene gains/losses: not only those that are strictly specific to a clade, but also those that occurred convergent between distant relatives. For instance a clade ancestor could have acquired a new gene and subsequently transferred it to a distant clade, leading their descendant to specifically share the gene. Our present analysis thus provides a mean to identify what *combination* of genes

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

is unique to a clade. We found that G1 and G8 share 57 synapomorphic genes (Table S3 and S4), in most case with phylogenetic support for transfer events between respective ancestors. In addition, different G1 and G8's clade-specific genes are encoding similar functional pathways, i.e. metabolism of phenolic compounds and production of exopolysaccharides (Table S2). These traits were previously hypothesized to provide key adaptation to the life in the plant rhizosphere of *A. fabrum* (i.e. *Agrobacterium* sp. G8, [6]). These ecological convergences may induce competition for shared resources between the G1 and G8 species.

Is the genetic hybridization of species G1 and G8 leading to their ecological convergence or rather diversification? First, some of the specifically shared genes are deeply diverged and probably result in different phenotypes. For instance, the genomic region AtSp15 encoding the biosynthesis of curdlan is found orthologous in G8, G1 and G2 strains, but their sequences are highly diverged: while the complete genomes of strains G8-C58 and G1-TT111 are 86% identical in average, the AtSp15 loci are 71% identical (among the 2% most diverging fragments between genomes). This distant homology probably confers different functions to the diverged proteins (76% average amino-acid identity over the locus) that may synthesize different form of curdlan. This divergence also explains why this cluster in G1 genomes was not recognized as homologous to G8-C58 genome using comparative genome hybridization, as the hybridization threshold lays around 80-85% nucleotide identity [6]. Second, shared niche-specifying genes are combined to other strictly clade-specific genes in each clade's core-genome. Niches are defined by several factors, and the sharing of identical means of adaptation for some of these factors do not lead to a complete equalization of niches. Typically, species G1 and G8 share the same locus for biosynthesis of O-antigens of the LPS, with highly similar enzymes (>93% amino-acid identity in average for proteins of the AtSp14 locus, Fig. S10) likely producing an equivalent compound. This component of the capsule may mediate a specific interaction with the environment, likely with cells of a eukaryotic host, as it is the case for the LPS synthesized by homologous enzymes in *Brucella spp.* [28]. In the putative case where this LPS would mediate attachment to a specific host plant, both species might compete for the use of resources such as the plant exsudates. However, competition may be reduced by differential access to the resource, as defined by specific epistatic interactions within each clade-specific gene set. For instance, while the LPS produced by G1 and G8 cells must be very similar, their regulation of biofilm production is likely different. There are many regulatory genes specific of G1 genomes, such as the *che2* operon (cluster AtSp2) and hub signal-transducing protein

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

HHSS ('hybrid-hybrid' signal-sensing, see Sup. Text section 7.1) (cluster AtSp14, Fig. 5 and S10) which are involved in chemotaxis regulation, and a sensor protein (cluster AtSp3) modulating c-di-GMP – a secondary messenger involved in the switch from motile to sessile behaviours. Those specific regulators are all found linked to G1-specific genes involved in phenolics catabolism or biofilm production. These latter genes may constitute the downstream regulatory targets of what seems to be a coherent regulation network controlling motility, biofilm production and degradation of phenolics, potentially in response to environmental conditions specific of the niche of G1.

Similarly, G8-specific genes of the AtSp26 cluster (Fig. 6) are involved in the perception and transduction of environmental signals such as mechanical constrains and a phenolic compound related to toluene; this regulatory island may be involved in specific regulation of genes linked to adaptation to G8-specific niche. The partial hybridization of G1 and G8 specific genomes thus likely results in each species tapping the same resources in a different way, which should not lead to a significant competition between them. These species may then form guilds of relatives that exploit partitions of a largely common ecological niche, explaining why we can observe them co-occurring in soils [24, 29]. Regular events of such inter-species hybridisation may explain why exploration of the genome based only on the occurrence of homolog – and not their gain history as done here – may yield indistinct patterns of specific gene gene contents, as recently observed among *R. leguminosarum* genomic species [30].

Role of recombination in species cohesiveness. The biovar 1 species complex of *Agrobacterium* (*At*) is made of ten so-called genomic species [29, 31] that are by definition well-delineated clusters of genome diversity upon which are ultimately defined bona fide bacterial species [32]. The high amount of HGT that occurs quite indistinctly between species blurs their delineation considering their gene content (Fig. S8). In this context, it is almost surprising that we can recover such a clear pattern of differentiation of species from molecular phylogenies (Fig. S1). It may result from different transfer dynamics between core and accessory genes, as previously reported [7, 33]. Indeed replacing transfers of core genes, that mostly reflect homologous recombination (HR), were occurring in large majority within species, accounting for 60%, 67%, 70% and 23% of total transfer events received in G4, G1 and G8, respectively (Fig. 7). A narrow clade like [G7-G9] also shows some preferential recombination with itself (23%), but its lesser density of

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

samplingmight prevent further detection of this phenomenon. Other species samples were not dense enough ($n \leq 2$) to reveal any such pattern. This indicates that HR occurs intensely within species of *At*, and certainly mediates cohesiveness of their genotype. In more details, we can see that in G8 (Fig. 7, blue frame), recombination mapped mostly to the ancestor clade, strongly suggesting that an unsampled diversity branching at the base of G8 significantly contributed to the genome of the species.

Inter-specific recombination in agrobacteria occurs at a lower frequency than within species, consistently with previous experimental findings [34]. It often involves transfers seen as coming from the root of the *At* clade, indicative of relatively frequent recombination of *At* species with unsampled sister lineages such as *A. larrymoorei* [35, 36]. The single representative strains of species G2, G3 and G13 appear to have received many genes from the distant clades [G6-G8] and [G7-G9]. Though one must consider that transfers mapped to these diverged strain lineages are the sum of what happened during large evolutionary times, there is nevertheless a substantial bias in the origin of transferred genes. The dataset of replacing transfers restricted to those confidently inferred by Prunier is similar (data not shown), rejecting the possibility of this bias of transfer routes being caused by the irresolution of deep gene tree nodes.

Additive transfers of non-core gene among genomic species is much more frequent than the replacement of core genes (Fig. S9B). It has long been known that the efficiency of HR in bacteria is linked by a log-linear relationship to the genetic distance of sexual protagonists [34, 37], which may explain why the inter-specific conversion of core genes is rare in comparison to additive HGT. HR machinery may be preserving the genetic cohesiveness of narrow clades such as species by controlling the identity of incoming core genes. This could alternatively reflect the increased rate of encounter of phylogenetically related strains due to closer ecological promiscuity.

Secondary replicons of *Agrobacterium* genomes are the place of genomic innovations. Rhizobiaceae present complex genomic architectures composed of a primary chromosome and a secondary chromosome or megaplasmid bearing essential genes called chromid [38], and a variable complement of plasmids of various sizes [39]. More specifically, the chromid of the *Agrobacterium* genus [35, 36], which includes the *At* clade, is linear [40, 41] as the result of an unique ancestral event of linearization and thus constituting a synapomorphy of this clade [42]. It was previously shown that, at the nucleotide sequence level, genes

15

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

evolved at higher rates on secondary replicons of complex genomes [43], which could be explained by differences in levels of gene expression between replicons impacting the efficacy of selection on genes [44]. Recombination is another phenomenon that make selection act more efficiently by unlinking sites or gene loci evolving under conflicting selective pressures, notably allowing higher nucleotide substitution rates and gene turnover at positively selected loci. We thus wanted to investigate the recombination dynamics of genes borne by the circular chromosome (Cc) vs. a linear chromid (Lc) to compare their different abilities to acquire and maintain adaptive genes. A previous study investigated the prevalence of recombination in Cc vs. Lc of *Agrobacterium* strains based on the linkage of a limited set of marker genes. They observed more inter-genic recombination on Lc but higher intra-genic recombination on Cc, and therefore concluded to no marked difference between replicons [45]. Here, we compare the recombination of both replicons using a much denser sampling, through the inventory in every core gene family of replacing transfer events as well as occurrences of intra-genic recombination. For this purpose, the location of core genes on the replicons of ancestral genomes of *At* was reconstructed using the evolutionary scenarios of the gene families as guides to a parsimony approach.

Gene gains and losses were significantly more frequent on Lc compared to Cc, with an average ratio of 2.59 more gains and 2.66 more losses, taking into account the difference in size of the replicons (paired Student's t-tests, $p < 10^{-5}$ and $p < 10^{-5}$, respectively; Table 2). The ratio of genes gained on Lc per gene gained on Cc is stable among all the genomes (ancestral and extant) of *At* (Pearson's correlation, $r = 0.96$, $p < 10^{-16}$), suggesting it is an intrinsic property of the Lc to integrate new genes more easily than the Cc. The proportions of genes conserved in a clade after acquisition on ancestral Cc or Lc are highly correlated across genomes (Pearson's correlation, $r = 0.95$, $p < 10^{-10}$) and their stable ratio of 0.85 is not statistically different from 1, though it would indicate a slightly greater capacity of Lc to retain the genes it has acquired. This similarity of conservation rates is surprising, given the more elevated rates of loss experienced by Lc, and suggests this replicon may be itself partitioned into stable and dynamic gene compartments. Replacing transfers events were used as indicators of gene-scale events of HR (Table 2). Summing over the history of *At* clade, we found 2,173 and 2,274 replacing transfers were received on Cc and Lc, respectively. Adjusting for the size of the replicons, there is a significant excess of replacing transfers events on Lc (paired Student's t-test, $p < 10^{-5}$). In addition, we searched for signatures of intra-genic events of allelic conversion in

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

alignments of orthologous gene families belonging to the unicopy core-genome of *At* and being found exclusively on either Cc or Lc. The proportions of genes found recombinant on Cc and Lc, are respectively 29% and 35% and are significantly different (χ 2 = 5.65, $p < 0.02$). Altogether, this shows a larger plasticity of gene content and a higher prevalence of HR on the linear chromid compared to the circular chromosome.

The higher rate of HR on the Lc can thus efficiently unlink loci harbouring core genes under strong purifying selection from the remainder of the chromosome where new genes with transient adaptive value can easily be inserted and deleted. This is of particular importance because the majority of clade-specific genes, putatively under strong ecological selection, reside on the Lc (Fig. 5, 6, S11, S12 and Table S2). Linear replicons can exchange fragments as large as the arm of a chromosome through single crossing-over, as observed during meiosis in Eukaryotes, and thus HR in these systems can mobilize large loci encoding entire complex pathways [46], which might explain this replicon's recombinogenic properties. Genotypes resulting from the recombination of large segments of chromosome might indeed rise quickly to fixation as the result of the selective sweep of new adaptive mutations in niche-specifying genes [19]. A high prevalence of single cross-over events might help maintain the linkage between several distant loci participating in the same adaptation. Such a setting could prevent sets of niche-specifying genes to be mixed between species, thus allowing ecological speciation to occur and persist [47]. Transmission of large haplotypes carrying a complete set of niche-specifying genes could also help maintain the set intact in the population even in presence of the high rate of single gene deletions observed on the Lc. However, the largest HR events characterized in this study were of a few 10kb (data not shown). Our ancestral reconstruction procedure may have a limited sensitivity to reconstruct very large single events, notably if the syntenic patterns were interrupted by later gene losses or chromosomal rearrangements. Though, using a cruder analysis, we previously observed chromosomal fragments larger than 1Mb with reduced divergence between pairs of the same *Agrobacterium* species. These likely reflected large events of homogenizing recombination, a pattern which was specific of the Lc [6] and could indeed support the role of this peculiar topology in favouring the hosting of niche-specifying genes.

The large *repABC* plasmids are conjugating to a broad host range among Rhizobiaceae, as described for symbiotic (Sym) or tumorigenic (Ti) plasmids in *Rhizobium* and *Agrobacterium*, respectively [6, 48, 49]. Thus, an unforeseen role for At plasmids (pAt's) was to host clade-specific adaptive genes in G1, G8, G4 and G7 species (Fig. 5, 6, S11 and S12). In particular, 25 G8-specific genes and 11 [G6-G8] clade-specific genes

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

were retrieved on the pAt's of the corresponding strains. Our previous study using micro-arrays with G8-C58 genome as a reference did not identified G8-specific genes on plasmids, notably because the G8 strain LMG-46 lacked a pAt [6]. This feature is however unique to the strain and might be a recent plasmid loss with rare prevalence in the species; a matter that should be further investigated in wild G8 population. The occurrence of clade-specific genes on the pAt and never on the other plasmids (Ti and smaller plasmids) suggests that this particular extrachromosomal element – or least a subset of the genes it carries – has a mobility restricted to the species in which it is found. The presence of core genes of the pAt plasmid thus make them *bona fide* chromids, as was already described for other rhizobial megaplasmids [38]. The highly plasticity of the At plasmids, shown by the little gene content conservation on that replicon between related strains – apart from clade-specific genes – indicate that Lc and pAt are of similar nature. Propensity to recombination was not testable here for the pAt in a similar manner than for Lc, because its very high plasticity renders difficult the assignation of genes to plasmids in ancestral genomes. The fact that genes specific of different clades of *At* could stably settle on their respective pAt reveals those third replicons of *Agrobacterium* genomes as ecologically important and probably essential in natural environments. Some isolates however, like G9-Hayward 0363, have no detectable plasmids, indicating that this genomes architecture with three 'main' replicons is not always the rule.

Conclusion

We developed an original method for the reconstruction of the history of all genes in genomes and applied it to the *Agrobacterium* biovar 1 species complex *(At)*, revealing the dynamics of gene repertoire in this taxon. These dynamics were structured along the tree of species and within genomes, revealing patterns of purifying selection for genes specific to major clades of *At*. Most of these were organized in large blocks of co-evolving genes that encode coherent pathways, which constitutes a departure from a neutral model of gene transfer in bacterial genomes. Genes specific to each species and to the *At* species complex as a whole recurrently encoded functions linked to production of secreted secondary metabolites or extracellular matrix, and to the metabolism of plant-derived compounds such as phenolics, sugars and amino-acids. These clade-specific genes likely constitute parallel adaptations to life in interaction with host plants, which suggest that

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ecological differentiation of *Agrobacterium* clades occurred through partitioning of ecological resources in plant rhizospheres. In addition, we revealed that the particularity of *Agrobacterium* as housing not only one, but actually two chromids, i.e. the linear chromosome and the pAt megaplasmid. These replicons are highly plastic and recombinogenic but nonetheless are the primary places of fixation of genes essential to the ecology of the species.

Methods

Bacteria and genomic sequence dataset. The study focused on the *Agrobacterium* biovar 1 species complex (*At*) with an original dataset of 16 new genomes along to six others publicly released [50–56]. The 22 genomes cover 10 closely related but genomically differentiated species (G1 to G9 and G13), with up to five isolates per species. The sample also includes every genome publicly available for the Rhizobiaceae at the time of the database construction (spring 2012), and more distant relatives from the Phyllobacteriaceae and Rhodobiaceae family (Table 3; Fig. S13). The dataset of 47 complete genome sequences (Table 3) was then used to construct a database of homologous gene families using the Hogenom pipeline [57] further used to elaborate the Agrogenom database.

Bacterial growth was analyzed in the presence of phenylacetate (5mM) using a Microbiology Bioscreen C Reader (Labsystems, Finland) according to the manufacturer's instructions. *Agrobacterium* strains grown overnight in AT medium supplemented with succinate and ammonium sulfate were inoculated at an optical density at 600 nm (OD_{600}) of 0.05 in 200 µl AT medium supplemented with appropriate carbon and nitrogen sources in Bioscreen honeycomb 100-well sterile plates. The cultures were incubated in the dark for 3 days at 28°C with shaking at medium amplitude. Growth measurements (OD_{600}) were obtained at 20-min intervals.

Genome sequencing and assembly. Genomic DNAs of the 16 *At* strains prepared with the phenol-chloroform method were used to prepare libraries with DNA sheared into inserts of median size of 8 kb. Raw sequence data were then generated using 454 GS-FLX sequencer (Roche Applied Sciences, Basel, Switzerland) with a combination of single-read (SR) and mate-pairs (MP) protocols, that yielded coverage ranging from 6.5X to 11X and from 5X to 8X, respectively (Table S5). Genome sequences were then assembled with Newbler version 2.6 (Roche Applied Sciences, Basel, Switzerland), using 90% identity and

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

40-bp thresholds for alignment of reads into contigs and the '--scaffold' option to integrate duplicated contigs into the scaffold assembly. Pseudo-molecules (chromosomes and plasmids) regrouping scaffolds were manually created on the basis of plasmid profiles obtained from Eckhart gels (data not shown) and minimizing rearrangements between closely related genomes considering alignments of obtained with NUCmer program from MUMMER package version 3.0 [58]. Genome sequences were then annotated with the MicroScope platform [59] and made available through the MaGe web interface [60].

Reference species tree. To construct the reference species tree, 455 unicopy core gene families (i.e. families with exactly one copy per genome, listed Table S5) extracted from the Agrogenom database were used for jackknife concatenations with 500 draws without replacement of 25 gene alignments, which were each concatenated and used to infer a maximum-likelihood (ML) tree using PhyML [61] (the same parameters as for gene trees, see Sup. Text.). The reference phylogeny was obtained by making the consensus of this sample of trees with CONSENSE algorithm from the Phylip package [62], and the branch supports were derived from the frequency of the consensus bipartitions in the sample (Figure S1). Alternative phylogenies were searched using the whole concatenate of 455 universal unicopy families or from a concatenate of 49 ribosomal protein gene families (Table S6) to compute trees with RAxML (version 7.2.8, GTRCAT model, 50 discrete site-heterogeneity categories) [63]. All three methods yielded very similar results concerning the placement of the different genera and species (Fig. S14).

Reconciliation of genome and gene tree histories. A pipeline was developed for the reconciliation of multicopy gene trees (potentially including several times the same species) with a species tree by the annotation of events of origination, duplication, transfer or speciation to the nodes of the gene trees. This pipeline combines several methods dedicated to the recognition of different landmarks of duplication and transfers (for a review, see [64]); it is fully detailed in the Supplementary Text Sections 3 and 4, and summarized below. Likely duplication events were first located by looking for mutiple gene copies per species in clades of the gene trees, using the 'Unicity' algorithm from TPMS [10] (Fig. S2, step 1). We subsequently isolated subtrees from the global gene trees where every species were represented once, i.e. unicopy subtrees. In presence of lineage-specific paralogs ('in-paralogs'), we extracted the several

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

overlapping unicopy subtrees that cover the different duplicated gene copies. Prunier, a parsimony-based method that takes into account the phylogenetic support of topological incongruences and iteratively resolves them by identifying transferred subtrees and pruning them [9], was run on the unicopy subtrees to detect replacing transfer events at branches with statistically significant (SH-like support > 0.9) topological conflict (Fig. S2, step 2). These reconciliations of (potentially overlapping) local subtrees were then integrated into a first coherently reconciled gene tree (Fig. S2, step 3). To complete the reconciliation provided by Prunier, we used the 'TPMS-XD' algorithm [10] to iteratively search for additional topological incongruences that had lower phylogenetic support but provided a global scenario more parsimonious on duplications and losses. Having confidently identified duplications and horizontal transfers leading to the emergence of new gene lineages, we could define subfamilies of orthologs nested in homologous gene families (Fig. S2, step 5). Finally, we used the program Count [11] to detect cryptic transfer events from the profile of occurrence of orthologous genes, i.e. transfers that explained heterogeneous profiles of gene occurrence without topological incongruence as evidence, again minimizing the number of inferred losses (Fig. S2, step 6). Each reconciliation of a gene tree corresponds to an evolutionary scenario in the species tree, where presence/absence states and the origination, duplication and transfer events are mapped (Fig. S2, step 7). Transfer events are characterized by the location of both donor and receiver ancestors in the species tree, which specifies the direction of the transfer; duplication are only characterized by their location at an ancestral node in the species tree. These locations in the species tree are referred to as coordinates of the events.

Block event reconstruction. The complete algorithm for block event reconstruction is described in the Supplementary Text, section 5 and summarized here. To detect single evolutionary events (duplication or transfer) involving several consecutive genes, tracks of genes sharing similar evolutionary events were sought using a greedy algorithm similar to that defined by Williams et al. [65]. Blocks were built by iterative inclusion of genes which lineages were marked by events with compatible coordinates (as described above), allowing them to be spaced by genes without such signal ('gap' genes) (Fig. S15 A,B). Blocks containing gap genes were checked for phylogenetic compatibility of those gap genes with the scenario associated to the block (Fig. S15 C). When the gene tree of the gap gene showed weak statistical support (SH-like support <

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

0.9) at the crucial branches supporting the phylogenetic conflict, the transfer event could not be rejected and the block integrity was maintained. Conversely, when the gene tree of the gap gene carried a signal rejecting the transfer event, i.e showing that donor and receptor clades are separated from each other in the gene tree by strongly supported branches, the original block was split into two blocks representing independent transfer events (Fig. S15 D). Jointly to its construction, the coordinates of the block event are refined by intersecting the coordinates of its constituent genes (Fig. 1B; Fig. S15 B,D). To reconstruct the ancestral state of the blocks characterized in contemporary genomes, homologous block events were sought, as block events from different genomes involving homologous genes that descend from the same evolutionary event in the corresponding gene tree (Fig. 1 B, step 2). The integral blocks of genes involved in the events that took place in the ancestral genomes ('ancestral block events') were rebuilt by accretion of homologous block events from several 'leaf' contemporary genomes (Fig. 1 B, step 3). Gene content of homologous leaf block events can differ among contemporary genomes because of partially independent histories of gene neighbourhood evolution (losses, insertions, rearrangements), leading to the disrupted contiguity of genes descending from a same event. However, block events that were disrupted in some leaf genome may appear intact in other genomes. Our accretion procedure links all leaf blocks – disjoint and intact – to one common ancestral block, thus recovering the unity of many block events that appeared as multiple one in individual genomes. In addition, independent gene losses or block disruption lead to varying gene content in homologous leaf block events, leading to potential differences in the inferred set of possible locations of the block events in the species tree. During the accretion of leaf blocks into an ancestral block, these different sets of possible locations of the block event are intersected into a refined set (Fig. 1 B, step 3), in an analogous way than for the grouping of genes into leaf blocks. Block events were investigated only for originations (O), duplications (D) and transfers (T), not for speciations (S). Blocks were not investigated at deep nodes (N1, N2, N3) for O and D (2,586 and 2,934 events discarded, respectively) because of the high risk of false positives (independent neighbour events older that occurred separately in time over these long branches but are annotated with similar coordinates and would thus be spuriously aggregated as block events). Finally, block events of gene loss were not investigated for a similar reason: for a same set of ODT events, many different scenarios of convergent losses are possible, with variable counts and location of loss events in the species tree, which would lead to the unspeciic aggregation in block of manyunrelated events.

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Definition of clade-specific genes from phylogenetic profiles. Clade-specific genes are genes exclusively found in a clade that were gained by the clade ancestor and since then conserved. From the sub-division of homologous gene families into orthologous subfamilies (see above and Sup. Text section 4, step 5.1), we established the phylogenetic profile of presence or absence of each subfamily in extant genomes. From these profiles, we identified contrasting patterns of presence/absence revealing clade-specific genotypes. Contrast is defined relative to a larger background clade in which the focus (foreground) clade in included, where foreground genomes show a pattern consistently opposite to that of all other genomes in the background clade. Background clades were generally chosen as those corresponding to a genus or species complex including the foreground clade, or to the whole tree if the foreground was larger than a genus. Possible subsequent transfer or loss events in the background clade can blur the contrasting pattern in phylogenetic profiles. The search for putative specific presence/absence patterns in the leaf genomes was therefore guided by the identification of unique gain/loss events in the genome of the foreground clade's ancestor , yielding a list of putatively specific gene subfamilies. This list was filtered using a relaxed definition of specificity, i.e. where the presence/absence contrast can be incomplete, with up to two genomes in the background clade sharing the foreground state.

Ancestral location of genes on replicons. Each complete replicons or contigs/scaffolds of genomes in the Agrogenom database was manually assigned to a type of replicon among the following factorial states: i) 'primary' (main chromosomes, i.e. circular chromosomes in *Agrobacterium* members), ii) ’secondary’ (large replicons with core genes restricted to Rhizobiaceae [40], a.k.a. chromids [38], i.e. linear chromids in *At*), iii) 'pAt' in *At* only, conserved megaplasmid identified as non-tumorigenic, iv) 'pTi' in *Agrobacterium* only, megaplasmid identified as tumorigenic, v) 'plasmid' for any other plasmids, including potential pAt or pTi megaplasmids that cannot be firmly identified as such, symbiotic megaplasmids and smaller plasmids and vi) 'unknown' when the assembly of the strain’s genome does not allow to map the contig/scaffold to a replicon. These states were transferred to the genes occupying the molecules and were mapped to the species phylogeny by gene subfamily, using ’absent’ state when the subfamily was not present in a genome.

Following the species phylogeny and the reconstructed occurrence profile of subfamilies in ancestral

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

genomes, the replicon location of subfamilies were propagated to ancestral genomes using Fitch's parsimony algorithm [66]. This method can result in several states being possible at a node, owing to equally parsimonious scenarios. The 'unknown' state is discarded at nodes when another state is proposed at the node, otherwise it is replaced by the next 'known' ancestral state (i.e. in the lineage from the node to the root); this notably allows to infer a location for genes from unfinished genomes (leaf genomes) from the location of the closest homolog. When several 'known' states were possible at a node, we attempted to restrict the set of proposed states by looking at ancestral states (see Sup. text for detailed algorithm).

Tree Pattern Matching. The collection of gene trees was searched with TPMS software [10] for the occurrence of particular phylogenetic patterns signing the monophyly of different groups of strains (Fig. S16). Patterns generally describe the local monophyly of two groups (e.g. G2 and [G4-G7-G9]) within subtrees containing only *At* members and with the external presence of an outgroup ensuring the right rooting of the subtree. For each pattern, another was searched for the occurrence of the same set of leaves but without constrains on their monophyly, to get the number of genes for which the monophyly hypothesis could be tested. Patterns with their translation into TPMS pseudo-Newick formalism referring to reference tree nodes are listed in Supplementary Material.

Detection of recombination in core genes. Core gene families were analyzed for clades within *At* for which we had at least 4 strains, which is the minimum to observe potential homoplasy in sequence alignments. We extracted from the Agrogenom database the sequence alignments of the gene families of their respective core genome, and applied the test PHI [67] to detect signatures of recombination in these alignments. Families were considered recombinant when the permutation test of PHI (1,000 draws) had a p-value ≤ 0.05. When comparing the proportion of recombinant families in the core genome of *At* to core families of younger clades, we had to account for the positive bias of sensitivity of PHI with the number of sequences in alignments. We build comparable datasets by reducing the size of *At* core alignments to the same number of sequences as in the other clade, taking 10 random samples of sequences per *At* core family to smooth the effect of sampling biases.

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Functional homogeneity of gene blocks. To measure to which extant co-transferred genes showed coherence in the functions they encoded, we used measures of semantic similarities of the Gene Ontology (GO) terms annotated to the gene products. First, the GO annotations were retrieved from UniProt-GOA (http://www.ebi.ac.uk/GOA/downloads, last accessed February, 2nd, 2013) [68] for the public genomes, and a similar pipeline of association of GO terms to gene products was used to annotate the genomic sequences produced for this study: results of several automatic annotation methods were retrieved from the PkGDB database [59] : InterProscan, HAMAP, PRIAM and hits of blastp searches on SwissProt and TrEMBL databases (as on the February, 5th, 2013), with a general cut-off e-value of 10e-10. GO annotations were then mapped to gene products using the mappings between those method results and GO terms as provided by Uniprot-GOA for electronic annotation methods (http://www.ebi.ac.uk/GOA/ElectronicAnnotationMethods, last accessed February, 12th, 2013) : InterPro2GO, HAMAP2GO, EC2GO, UniprotKeyword2GO, UniprotSubcellular Location2GO. The annotation dataset was limited to the electronically inferred ones to avoid biases of annotation of certain model strains or genes. The obtained functional annotations of proteomes were analyzed in the frame of Gene Ontology term reference (Full ontology file downloaded at http://www.geneontology.org/GO.downloads.ontology.shtml, last accessed September, 2nd, 2013) [69]. Functional homogeneity (*FH*) within a group of genes is defined as the combination of the pairwise functional similarities between all gene products in the group, each of which is the combination of pairwise similarities between all terms annotated to a pair of genes. Similarities were measured using *Rel* and *funSim* metrics [70, 71]. Computations were done using a custom Python package derived from AIGO package v0.1.0 (https://pypi.python.org/pypi/AIGO).

To assess the potential role of selection in favoring the retention of transferred genes with more coherent functions, the *FH* of transferred gene blocks were compared to that of random groups of genes of same size sampled from the same genome, by uniformly sampling them in replicons or by taking systematic windows of neighbors' genes. *FH* were computed for all windows of neighbor genes around a replicon, but a limited sample of the same size was done for combinations of non-linked genes. Because the size of the group of genes impacts strongly the computation of the similarity metrics, and because the density of annotations can vary among organisms and replicons, the distributions of *FH* were calculated by replicon and by group size. Note that the set of block of transferred genes is included in the set of all gene segments, but that

25

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

independent subsets were considered for statistical comparisons. In turn, to test if functional coherence of a block transferred genes impacted its probability of retention after transfer, we used the information from reconstructed ancestral blocks of transferred genes to compared the same *FH* metric between extant transferred blocks that were integrally conserved in all descendants of the recipient ancestor and extant transferred blocks that were degraded by gene losses in other descendants of the recipient (but were intact in the focal genome). To avoid biases linked to variation in age of the considered transfer events, this comparison was made only for events that occurred in ancestors of species-level clades in *At.*

All scripts used in the present work are available at https://github.com/flass/agrogenom.

Agrogenom database. All the data about genes (functional annotations, gene families), genomes (position of genes, architecture in replicons …), the species tree (nodes, taxonomic information), reconciliations (gene trees, ODTS events), block events, inference analyses (parameters, scores …), and all other data relative to the present work were compiled in a relational database, Agrogenom (Fig. S17). This web interface for Agrogenom database is accessible at http://phylariane.univ-lyon1.fr/db/agrogenom/3/ (Figure 9).

Availability of supporting data

The sixteen new genome sequences used in these projects were submitted to the EBI-ENA (www.ebi.ac.uk/ena) under the BioProjects PRJEB12180-PRJEB12196 and are currently being processed. Accession numbers are not yet available but genome annotation flat files (EMBL format) are provided as supplementary material (Sup. File S2).

List of abbreviations

At – *Agrobacterium tumefaciens* species complex
CDS – coding sequence
FH – functional homogeneity
HGT – horizontal gene transfer
HR – homologous recombination
HHSS – “hybrid-hybrid” signal-sensing protein
ODT – origination, duplication and transfer (events)

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ODTL – origination, duplication, transfer and loss (events)

Cc – circular chromosomes

Lc – linear chromid

pAt – *At* plasmid

pTi – Tumor-inducing plasmid

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

FL, XN and VD conceived and supervised the Agrogenom project. DC, DMu and LV cultivated the bacteria, prepared the samples and conducted biochemical tests. FL, RP, SP and LG contributed code and software. FL and RP designed Agrogenom database. RP conceived Agrogenom web interface. DMo generated functional annotation data. AC and DV integrated data into the Microscope database. VB led the sequencing project. VB, FL and AD participated in the genome assembly. FL and VD conceived the phylogenetic methods. FL implemented the phylogenetic methods and performed the statistical and genomic analyses. FL, LV, DMu, VD and XN participated in writing the manuscript.

Additional files

Additional file 1 (PDF file). Supplemental figures. Figure S1. Reference phylogeny of Rhizobiales history. Obtained by consensus of ML trees built from concatenates of 500 jacknife samplings of 25 genes among the 455 unicopy genes from the core of the 47 genomes. Supports of short branches within *S.meliloti* were all ≥ 0.9. **Figure S2.** Bioinformatic pipeline for reconciliation of gene and genome histories. **Figure S3.** Gain in reconciliation precision while amalgamating block scenarios. **Figure S4.** Gene gain, loss and conservation within *At* clade ancestors. **Figure S5.** Residuals of negative exponential regression of clade age vs. conservation of gained genes. **Figure S6.** Growth curves of representative of *At* genomic species on phenylacetate. **Figure S7.** Distribution of sizes of block events. **Figure S8.** Hierarchical clustering of *Rhizobiales* genomes according to their gene content. **Figure S9.** Intensity of gene transfers among Rhizobiales. **Figure S10.** Syntenic conservation of the AtSp14 cluster in G1, G8 and Brucelaceae. **Figure S11.** Historical stratification of gains in the lineage of *A. sp* G4 (*A. radiobacter*) strain B6. **Figure S12.** Historical stratification of gains in the lineage of *A. sp. G7* strain Zutra 3/1. **Figure S13.** Phylogeny of 131 genomes of Alpha-proteobacteria. **Figure S14.** Alternative reference tree topologies obtained with different methods. **Figure S15.** Algorithm for construction of blocks of co-transferred genes. **Figure S16.** Support for monophyply of groups in all gene trees. **Figure S17.** Schema of Agrogenom relational database.

Additional file 2 (Excel [.xls] multiple spreadsheet). Supplemental tables. Table S1. Statistics of gains and losses per contemporary and ancestral genome and replicon. **Table S2.** Location and functional description of clade-specific gene clusters in *A. tumefaciens* genomes. **Table S3.** Summary of clade-specific genes in TT111 genome. **Table S4.** Summary of clade-specific genes in C58 genome. **Table S5.** List of the 455 universal unicopy gene families. **Table**

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S6. Matrix of presence/absence of the 49 ribosomal gene families in the 47 Rhizobiaceae genomes. **Table S7.** Statistics of the 16 new genome sequences.

Additional file 3 (PDF file). Supplemental text. Section 1. Comparison of several hypotheses for the core-genome reference phylogeny. **Section 2.** Construction of Agrogenom database. **Section 3.** Reconciliation of gene trees with the species tree. **Section 4.** Gene tree reconciliations: detailled procedure. **Section 5.** Block event reconstruction: algorithms. **Section 6.** Of the complexity of interpreting 'highways' of genes transfers. **Section 7.** Clade-specific genes: insights into the ecological properties of clades. **Section 8.** Selected cases of large transfer events. **Section 9.** Bioinformatic scripts, modules and libraries.

Additional file 4 (Excel [.xls] multiple spreadsheet file). Supplementary File S1: Lists of clade-specific genes per clade.

Additional file 5 (.tar.gz archive file). Supplementary File S2: EMBL annotation flat files for the 16 new genome sequences

Acknowledgements

This project was supported by the French National Research Agency (ANR) grant ECOGENOME (ANR-BLAN-08-0090) and ANCESTROME (ANR-10-BINF-01-01) and by AGROMICS grant from the ENVIROMICS challenge of the Interdisciplinary Mission of the French National Centre for Scientific Research (CNRS). FL was supported by a scholarship from Ecole Normale Supérieure de Lyon and by ERC grant BIG_IDEA 260801 (http://erc.europa.eu/). This work was performed using the computing facilities of the CC LBBE/PRABI. The LABGeM (CEA/IG/Genoscope & CNRS UMR8030) and the France Génomique National infrastructure (funded as part of Investissement d'avenir program managed by Agence Nationale pour la Recherche, contract ANR-10-INBS-09) are acknowledged for support within the MicroScope annotation platform.

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Doolittle WF: **Is junk DNA bunk? A critique of ENCODE**. *Proc Natl Acad Sci* 2013, **110**:5294–5300.

2. Lawrence JG, Ochman H: **Amelioration of Bacterial Genomes: Rates of Change and Exchange**. *J Mol Evol* 1997, **44**:383–397.

3. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, Calteau A, Chiapello H, Clermont O, Cruveiller S, Danchin A, Diard M, Dossat C, Karoui ME, Frapy E, Garry L, Ghigo JM, Gilles AM, Johnson J, Le Bouguénec C, Lescat M, Mangenot S, Martinez-Jéhanne V, Matic I, Nassif X, Oztas S, et al.: **Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths**. *PLoS Genet* 2009, **5**:e1000344.

4. Daubin V, Lerat E, Perrière G: **The source of laterally transferred genes in bacterial genomes**. *Genome Biol* 2003, **4**:R57.

5. Costechareyre D, Rhouma A, Lavire C, Portier P, Chapulliot D, Bertolla F, Boubaker A, Dessaux Y, Nesme X: **Rapid and Efficient Identification of Agrobacterium Species by recA Allele Analysis : Agrobacterium recA Diversity**. *Microb Ecol* 2010, **60**:862–72.

6. Lassalle F, Campillo T, Vial L, Baude J, Costechareyre D, Chapulliot D, Shams M, Abrouk D, Lavire C, Oger-Desfeux C, Hommais F, Guéguen L, Daubin V, Muller D, Nesme X: **Genomic Species Are Ecological Species as Revealed by Comparative Genomics in Agrobacterium tumefaciens**. *Genome Biol Evol* 2011, **3**:762 –781.

7. Abby SS, Tannier E, Gouy M, Daubin V: **Lateral gene transfer as a support for the tree of life**. *Proc Natl Acad Sci* 2012, **109**:4962–4967.

8. Ravenhall M, Škunca N, Lassalle F, Dessimoz C: **Inferring Horizontal Gene Transfer**. *PLoS Comput Biol* 2015, **11**:e1004095.

9. Abby SS, Tannier E, Gouy M, Daubin V: **Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests**. *BMC Bioinformatics* 2010, **11**:324–324.

10. Bigot T, Daubin V, Lassalle F, Perrière G: **TPMS: a set of utilities for querying collections of gene trees**. *BMC Bioinformatics* 2013, **14**:109.

11. Csűrös M: **Ancestral Reconstruction by Asymmetric Wagner Parsimony over Continuous Characters and Squared Parsimony over Distributions**. In *Comparative Genomics*. Edited by Nelson CE, Vialette S. Springer Berlin Heidelberg; 2008:72–86. [*Lecture Notes in Computer Science*, vol. 5267]

12. Vallenet D, Engelen S, Mornico D, Cruveiller S, Fleury L, Lajus A, Rouy Z, Roche D, Salvignol G, Scarpelli C, Médigue C: **MicroScope: a platform for microbial genome annotation and comparative genomics**. *Database J Biol Databases Curation* 2009, **2009**:bap021.

13. Makarova K, Slesarev A, Wolf Y, Sorokin A, Mirkin B, Koonin E, Pavlov A, Pavlova N, Karamychev V, Polouchine N, Shakhova V, Grigoriev I, Lou Y, Rohksar D, Lucas S, Huang K, Goodstein DM, Hawkins T, Plengvidhya V, Welker D, Hughes J, Goh Y, Benson A, Baldwin K, Lee

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

J-H, Díaz-Muñiz I, Dosti B, Smeianov V, Wechter W, Barabote R, et al.: **Comparative genomics of the lactic acid bacteria**. *Proc Natl Acad Sci* 2006, **103**:15611 –15616.

14. Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferriera S, Johnson J, Steglich C, Church GM, Richardson P, Chisholm SW: **Patterns and Implications of Gene Gain and Loss in the Evolution of Prochlorococcus**. *PLoS Genet* 2007, **3**:e231.

15. Kuo C-H, Moran NA, Ochman H: **The consequences of genetic drift for bacterial genome complexity**. *Genome Res* 2009, **19**:1450–1454.

16. David LA, Alm EJ: **Rapid evolutionary innovation during an Archaean genetic expansion**. *Nature* 2011, **469**:93–96.

17. Szöllősi GJ, Boussau B, Abby SS, Tannier E, Daubin V: **Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations**. *Proc Natl Acad Sci* 2012, **109**:17513–17518.

18. Cohan FM, Koeppel AF: **The origins of ecological diversity in prokaryotes**. *Curr Biol CB* 2008, **18**:R1024–1034.

19. Lassalle F, Muller D, Nesme X: **Ecological speciation in bacteria: reverse ecology approaches reveal the adaptive part of bacterial cladogenesis**. *Res Microbiol* 2015, **166**:729–741. [*Special Issue on Microbial Diversity, Adaptation and Evolution*]

20. Kopac S, Wang Z, Wiedenbeck J, Sherry J, Wu M, Cohan FM: **Genomic Heterogeneity and Ecological Speciation within One Subspecies of Bacillus subtilis**. *Appl Environ Microbiol* 2014, **80**:4842–4853.

21. Rocha EPC: **The Organization of the Bacterial Genome**. *Annu Rev Genet* 2008, **42**:211–233.

22. Lawrence JG, Roth JR: **Selfish operons: horizontal transfer may drive the evolution of gene clusters**. *Genetics* 1996, **143**:1843–1860.

23. Homma K, Fukuchi S, Nakamura Y, Gojobori T, Nishikawa K: **Gene Cluster Analysis Method Identifies Horizontally Transferred Genes with High Reliability and Indicates that They Provide the Main Mechanism of Operon Gain in 8 Species of Gamma-Proteobacteria**. *Mol Biol Evol* 2007, **24**:805 –813.

24. Vogel J, Normand P, Thioulouse J, Nesme X, Grundmann GL: **Relationship between spatial and genetic distance in Agrobacterium spp. in 1 cubic centimeter of soil**. *Appl Environ Microbiol* 2003, **69**:1482–1487.

25. Gause GF: **Experimental Studies on the Struggle for Existence I. Mixed Population of Two Species of Yeast**. *J Exp Biol* 1932, **9**:389–402.

26. Aujoulat F, Jumas-Bilak E, Masnou A, Sallé F, Faure D, Segonds C, Marchandin H, Teyssier C: **Multilocus sequence-based analysis delineates a clonal population of Agrobacterium (Rhizobium) radiobacter (Agrobacterium tumefaciens) of human origin**. *J Bacteriol* 2011, **193**:2608–2618.

27. **Agrogenom3** [http://phylariane.univ-lyon1.fr/db/agrogenom/3/]

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

28. Vizcaíno N, Cloeckaert A, Zygmunt MS, Fernández-Lago L: **Characterization of a Brucella species 25-kilobase DNA fragment deleted from Brucella abortus reveals a large gene cluster related to the synthesis of a polysaccharide**. *Infect Immun* 2001, **69**:6738–6748.

29. Portier P, Fischer-Le Saux M, Mougel C, Lerondelle C, Chapulliot D, Thioulouse J, Nesme X: **Identification of genomic species in Agrobacterium biovar 1 by AFLP genomic markers**. *Appl Environ Microbiol* 2006, **72**:7123–7131.

30. Kumar N, Lad G, Giuntini E, Kaye ME, Udomwong P, Shamsani NJ, Young JPW, Bailly X: **Bacterial genospecies that are not ecologically coherent: population genomics of Rhizobium leguminosarum**. *Open Biol* 2015, **5**:140133.

31. Popoff MY, Kersters K, Kiredjian M, Miras I, Coynault C: **[Taxonomic position of Agrobacterium strains of hospital origin]**. *Ann Microbiol (Paris)* 1984, **135A**:427–442.

32. Stackebrandt E, Frederiksen W, Garrity GM, Grimont PAD, Kämpfer P, Maiden MCJ, Nesme X, Rosselló-Mora R, Swings J, Trüper HG, Vauterin L, Ward AC, Whitman WB: **Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology**. *Int J Syst Evol Microbiol* 2002, **52**(Pt 3):1043–1047.

33. Choi SC, Rasmussen MD, Hubisz MJ, Gronau I, Stanhope MJ, Siepel A: **Replacing and additive horizontal gene transfer in Streptococcus**. *Mol Biol Evol* 2012, **29**:3309–3320.

34. Costechareyre D, Bertolla F, Nesme X: **Homologous recombination in Agrobacterium: potential implications for the genomic species concept in bacteria**. *Mol Biol Evol* 2009, **26**:167–176.

35. Ormeño-Orrillo E, Servín-Garcidueñas LE, Rogel MA, González V, Peralta H, Mora J, Martínez-Romero J, Martínez-Romero E: **Taxonomy of rhizobia and agrobacteria from the Rhizobiaceae family in light of genomics**. *Syst Appl Microbiol* 2015, **38**:287–291. [*Taxonomy in the Age of Genomics*]

36. Mousavi SA, Willems A, Nesme X, de Lajudie P, Lindström K: **Revised phylogeny of Rhizobiaceae: Proposal of the delineation of Pararhizobium gen. nov., and 13 new species combinations**. *Syst Appl Microbiol* 2015, **38**:84–90.

37. Roberts MS, Cohan FM: **The effect of DNA sequence divergence on sexual isolation in Bacillus**. *Genetics* 1993, **134**:401–408.

38. Harrison PW, Lower RPJ, Kim NKD, Young JPW: **Introducing the bacterial "chromid": not a chromosome, not a plasmid**. *Trends Microbiol* 2010, **18**:141–148.

39. Young JPW, Crossman LC, Johnston AW, Thomson NR, Ghazoui ZF, Hull KH, Wexler M, Curson AR, Todd JD, Poole PS, Mauchline TH, East AK, Quail MA, Churcher C, Arrowsmith C, Cherevach I, Chillingworth T, Clarke K, Cronin A, Davis P, Fraser A, Hance Z, Hauser H, Jagels K, Moule S, Mungall K, Norbertczak H, Rabbinowitsch E, Sanders M, Simmonds M, et al.: **The genome of Rhizobium leguminosarum has recognizable core and accessory components**. *Genome Biol* 2006, **7**:R34.

40. Slater SC, Goldman BS, Goodner B, Setubal JC, Farrand SK, Nester EW, Burr TJ, Banta L, Dickerman AW, Paulsen I, Otten L, Suen G, Welch R, Almeida NF, Arnold F, Burton OT, Du Z, Ewing A, Godsy E, Heisel S, Houmiel KL, Jhaveri J, Lu J, Miller NM, Norton S, Chen Q,

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Phoolcharoen W, Ohlin V, Ondrusek D, Pride N, et al.: **Genome sequences of three Agrobacterium biovars help elucidate the evolution of multichromosome genomes in bacteria**. *J Bacteriol* 2009, **191**:2501–2511.

41. Slater S, Setubal JC, Goodner B, Houmiel K, Sun J, Kaul R, Goldman BS, Farrand SK, Almeida N, Burr T, Nester E, Rhoads DM, Kadoi R, Ostheimer T, Pride N, Sabo A, Henry E, Telepak E, Cromes L, Harkleroad A, Oliphant L, Pratt-Szegila P, Welch R, Wood D: **Reconciliation of Sequence Data and Updated Annotation of the Genome of Agrobacterium tumefaciens C58, and Distribution of a Linear Chromosome in the Genus Agrobacterium**. *Appl Environ Microbiol* 2013, **79**:1414–1417.

42. Ramírez-Bahena MH, Vial L, Lassalle F, Diel B, Chapulliot D, Daubin V, Nesme X, Muller D: **Single acquisition of protelomerase gave rise to speciation of a large and diverse clade within the Agrobacterium/Rhizobium supercluster characterized by the presence of a linear chromid**. *Mol Phylogenet Evol* 2014, **73**:202–207.

43. Cooper VS, Vohr SH, Wrocklage SC, Hatcher PJ: **Why Genes Evolve Faster on Secondary Chromosomes in Bacteria**. *PLoS Comput Biol* 2010, **6**:e1000732.

44. Morrow JD, Cooper VS: **Evolutionary Effects of Translocations in Bacterial Genomes**. *Genome Biol Evol* 2012, **4**:1256–1262.

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1: Origination, Duplication, Transfer and Speciation events inferred in reconciliations of the Agrogenom database.

Event type	Single gene events	Block events	(of size >1)
Originations	5,189	4,267	(667)
Duplications	7,340	5,819	(778)
Total transfers	43,233	32,255	(5,649)
Replacing transfers †	9,271	-	-
Additive transfers †	33,962	-	-
Total ODT	55,762	42,341	(7,094)
Implied losses	29,843	32,739	-
Total ODTL	85,605	75,080	-

O, origination; D, duplication; T, transfer; L, loss; ODT refers to the combination of all O, D and T events, ODTL also includes losses. †: Replacing and additive transfers are not distinguished in block events. Integrating scenarios with block events was more parsimonious in the global count of ODTL events than if each gene family had considered separately, with more loss events but comparatively much less ODT events (L: +2,896; O: -922; D: -1,521; T: -10,978).

Table 2: Gene content plasticity and recombination of Cc vs. Lc.

	Size	# Gains	% Gains	# Losses	% Losses	# Repl. Trans.	% Repl. Trans.	% Rec. Core Fam.
Cc	2073	194	9.3	31	1.4	51	2	29
Lc	1124	229	19.1	39	3.4	53	3.3	35
$p <$ †		10-3	10-9	10-1	10-5	0.93	10-5	0.02[a]

Size: average across all nodes of *At* phylogeny of the number of genes that could be unambiguously mapped to a replicon. Gains, Losses, Repl. Trans.: average across all nodes of *At* phylogeny of gain, loss or replacing transfer events, respectively, in absolute count (#) or percentage relative to genome size (%). % Rec. Core Fam.: percentage of *At* core gene families exclusively located on one replicon that show signatures of recombination (as detected by PHI test (Bruen et al., 2006)). †: *p*-values resulting from Student's t-tests, except (a) which is the result of a Chi-squared test.

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 3: List of 47 Rhizobiales strains used in this study.

Clade/Taxon	Strain name	Code†	NCBI Taxid	Genome size (nb. genes)
Agrobacterium biovar 1 species complex (*At*)				
A. sp. G1	H13-3	AGRSH	861208	5345
	5A	AGRTU1	1107544	5518
	CFBP 5771	ATU1A	1183421	5546
	S56	ATU1B	1183429	5627
	TT111	ATU1C	1183430	5856
A. sp. G2 (*A. pusense*)	CFBP 5494	ATU2A	1183436	6013
A. sp. G3	CFBP 6623	ATU3A	1183432	5378
A. sp. G4 (*A. radiobacter*)	B6	ATU4A	1183423	5875
	CFBP 5621	ATU4B	1183422	5330
	Kerr 14	ATU4C	1183424	5870
	CCNWGS0286	AGRTU2	1082932	4979
A. sp. G5	CFBP 6626	ATU5A	1183435	5332
	F2	AGRTU3	1050720	5321
A. sp. G6	NCPPB 925	ATU6A	1183431	6139
A. sp. G7	NCPPB 1641	ATU7A	1183425	6041
	RV3	ATU7B	1183426	5182
	Zutra 3/1	ATU7C	1183427	5685
A. sp. G8 (*A. fabrum*)	C58	AGRT5	176299	5639
	ATCC 31749	AGRSP1	82789	5535
	J-07	ATU8A	1183433	5592
A. sp. G9	Hayward 0363	ATU9A	1183434	4502
A. sp. G13	CFBP 6927	ATU13	1183428	4993
Allorhizobium				
Allorhizobium vitis	S4	AGRVS	311402	5389
Rhizobium sp.	PDO1-076	RHISP1	1125979	5340
Rhizobium				
R. rhizogenes	K84	AGRRK	311403	6684
R. etli	CIAT 652	RHIE6	491916	6109
	CFN 42	RHIEC	347834	6016
	CNPAF512	RHIET1	993047	6544
R. leguminosarum bv. *viciae*	3841	RHIL3	216596	7263
R. leguminosarum bv. *trifolii*	WSM1325	RHILS	395491	7001
	WSM2304	RHILW	395492	6415
*Ensifer/Sinorhizob*ium				
E. meliloti	1021	RHIME	266834	6234
	BL225C	SINMB	698936	6354
	CCNWSX0020	SINME1	1107881	6844
	AK83	SINMK	693982	6510
	SM11	SINMM	707241	7093
E. medicae	WSM419	SINMW	366394	6213
E. fredii	HH103	SINFR1	1117943	6787
	NGR234	RHISN	394	6366
Mesorhizobium/Chelativorans				
M. alhagi	CCNWXJ12-2	MESAL1	1107882	7184
M. amorphae	CCNWGS0123	MESAM1	1082933	7075
M. australicum	WSM207	MESAU1	7540353	5934
M. ciceri bv. *biserrulae*	WSM1271	MESCW	765698	6264
M. opportunistum	WSM2075	MESOW	536019	6508
M. loti	MAFF303099	RHILO	266835	7281
Chelativorans sp.	BNC1	MESSB	266779	4543
Parvibaculum				
P. lavamentivorans	DS-1	PARL1	402881	3636

†: the genome code used in this study is either the SwissProt 5-letter code for organisms if the genome was already referenced at the time of the study, or an ad-hoc identifier specifically generated for this study.

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

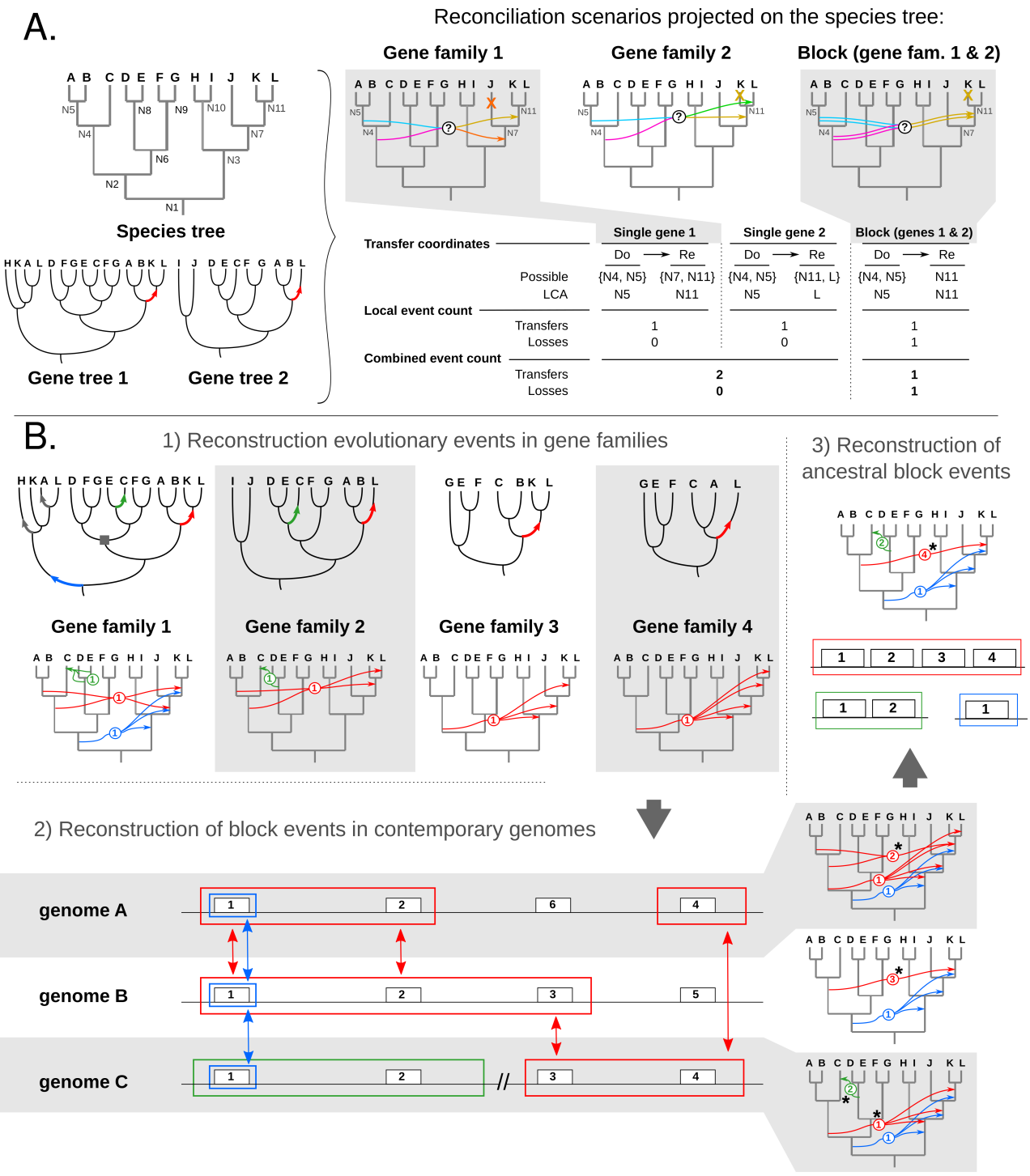
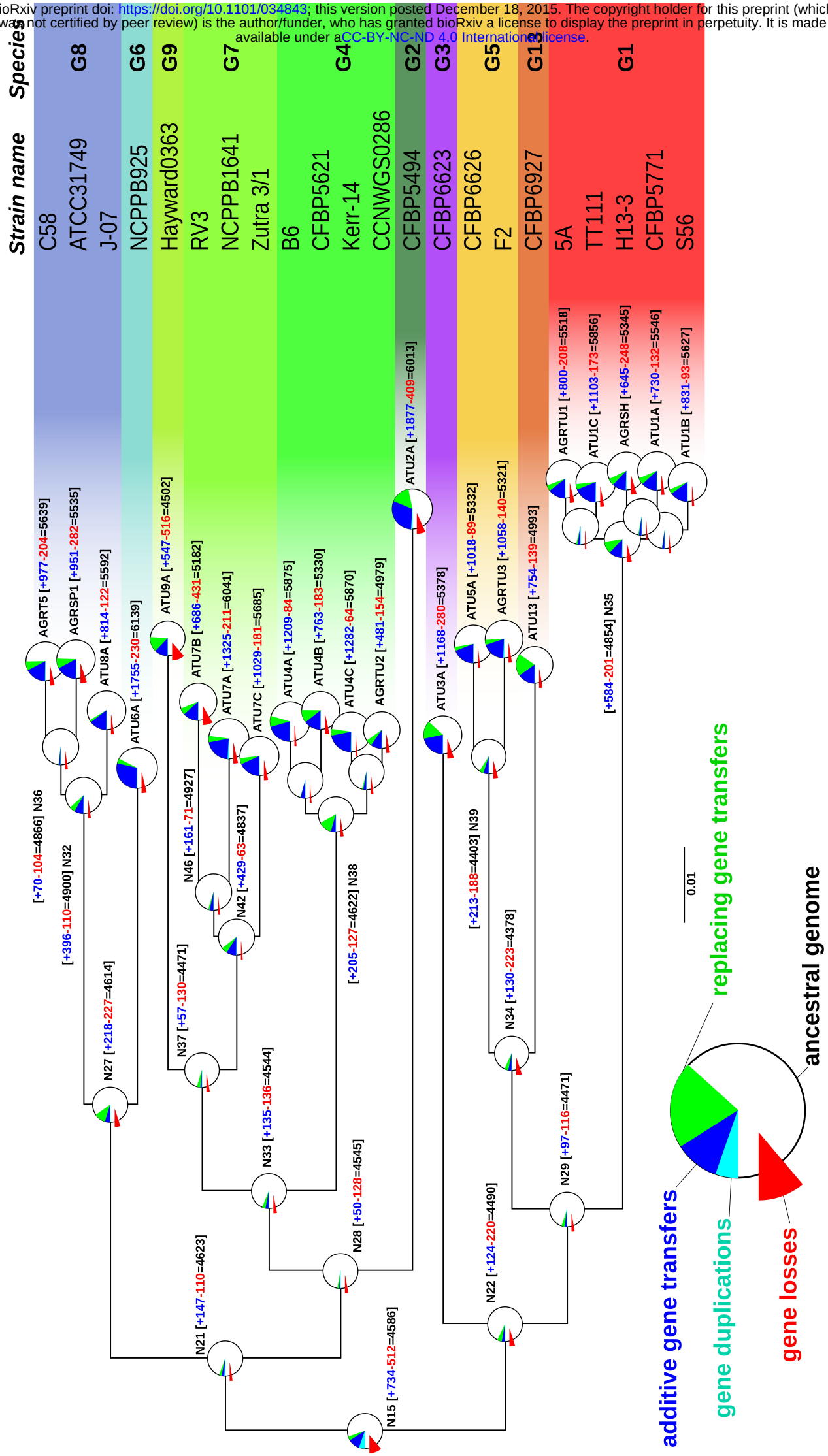


Figure 1: Gene-wise vs. regional reconciliation. A) Transfers inferred in reconciled gene trees 1 and 2 can be translated in several possible scenarios in the species tree that each involves different donor (Do) and receiver (Re) pairs (multiple arrows with question marks, uncertain scenarios). If each gene family is reconciled separately, the scenarios that place the ancestral receiver at the last common ancestor of extant recipient genomes are chosen because they are the most parsimonious in losses (crosses mapped on the species tree and "Local event count" in inset table). That way, the global scenario for the combined loci totalizes two transfers and no subsequent loss (inset table, "Combined event count"). If one considers the possibility of the co-transfer of neighbour genes 1 and 2, a common (Block) transfer event can be found. By minimizing the total number of (co-)transfer events, a scenario can be chosen which is not necessarily the most parsimonious in losses for each gene. In this example, the global scenario for the combined loci which is the most parsimonious in transfer events totalizes one block transfer and one subsequent gene loss. Integrating over all reconciliations in Agrogenom database, considering block scenarios induces 2,896 additional loss events compared to

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

considering independent gene family scenarios (over a total of 32,739 losses), but reduces the count of transfer events by 13,421. B) Refinement of event uncertainties when building block events. Origination, duplication and transfer events are first inferred in each gene family separately (1); for the sake of clarity, the example shows only transfer events as arrows on branches of gene trees (top) and between branches of species trees (bottom). Compatible events affecting genes that are neighbour in at least one extant genome are aggregated into blocks (coloured frames) (2) and this approach is then repeated across genomes (vertical double arrows), thus reconstructing the events that occurred in ancestral genomes (3). Numbers in circles indicate the number of genes combined in a same event, stars indicate when aggregation of an event lead to the refinement of its coordinates.

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 2: Ancestral genome sizes and gain/loss events. The tree is a subtree of that presented Figure S1, focusing on the *At* clade. Net gains (+) and losses (-) and resulting genome sizes (=) are indicated next to nodes. Disc at nodes schematically represent inferred ancestral genomes or actual extant genomes; surfaces are proportional to the genome size. Prevalence of events shaping the gene content are indicated by pie charts indicating the fraction of losses (red), gains by duplication (cyan), gains by transfer (blue) and gene conversions/allelic replacements (green). The relatively high number of event occurring at *At* root is related to the long branch from which it stems in the complete Rhizobiales tree (Fig. S1), which is not represented here.

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

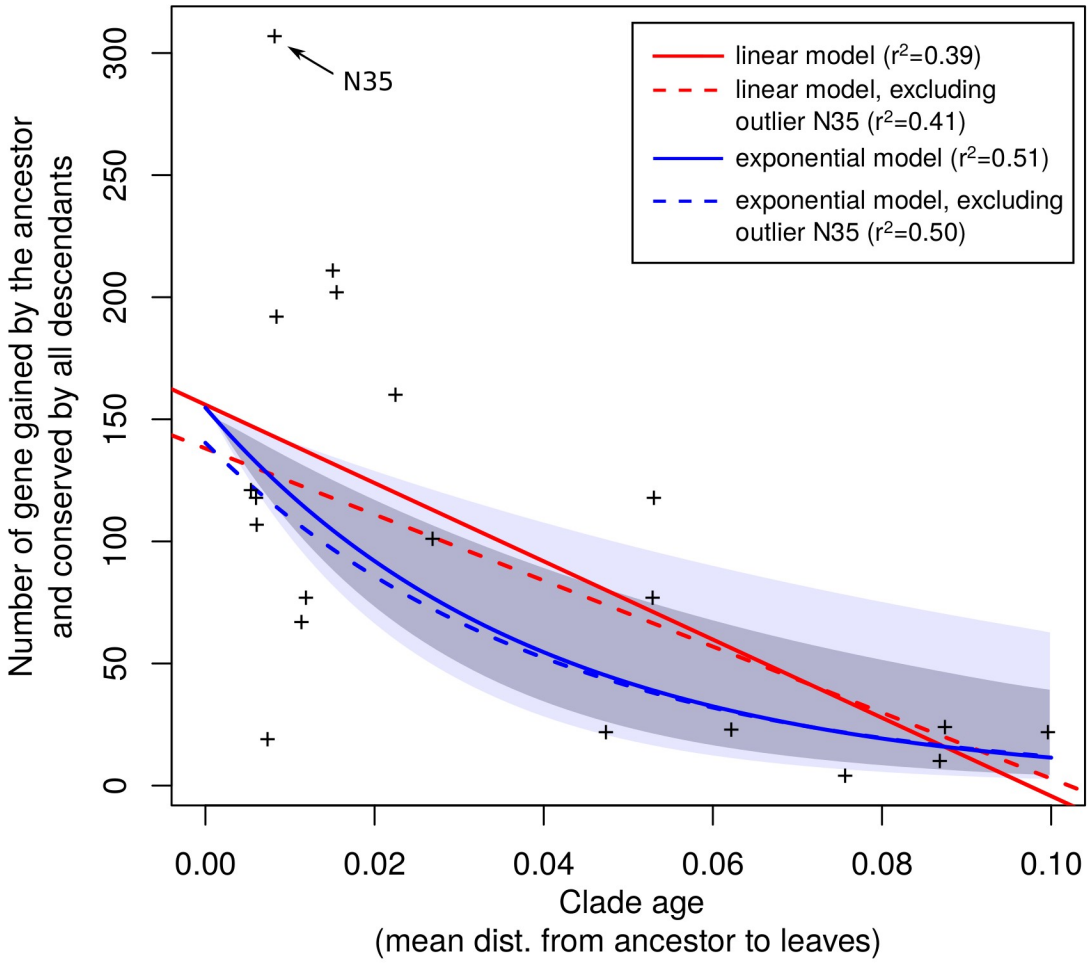


Figure 3: Retention of gained genes within *At* genomes follow a 'survival' model. Node 'N15' (the G1 species ancestor) is the strongest driver in the linear regression. Dark and light shaded areas respectively represent the 95% and 99% confidence intervals of the exponential model (solid blue line).

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

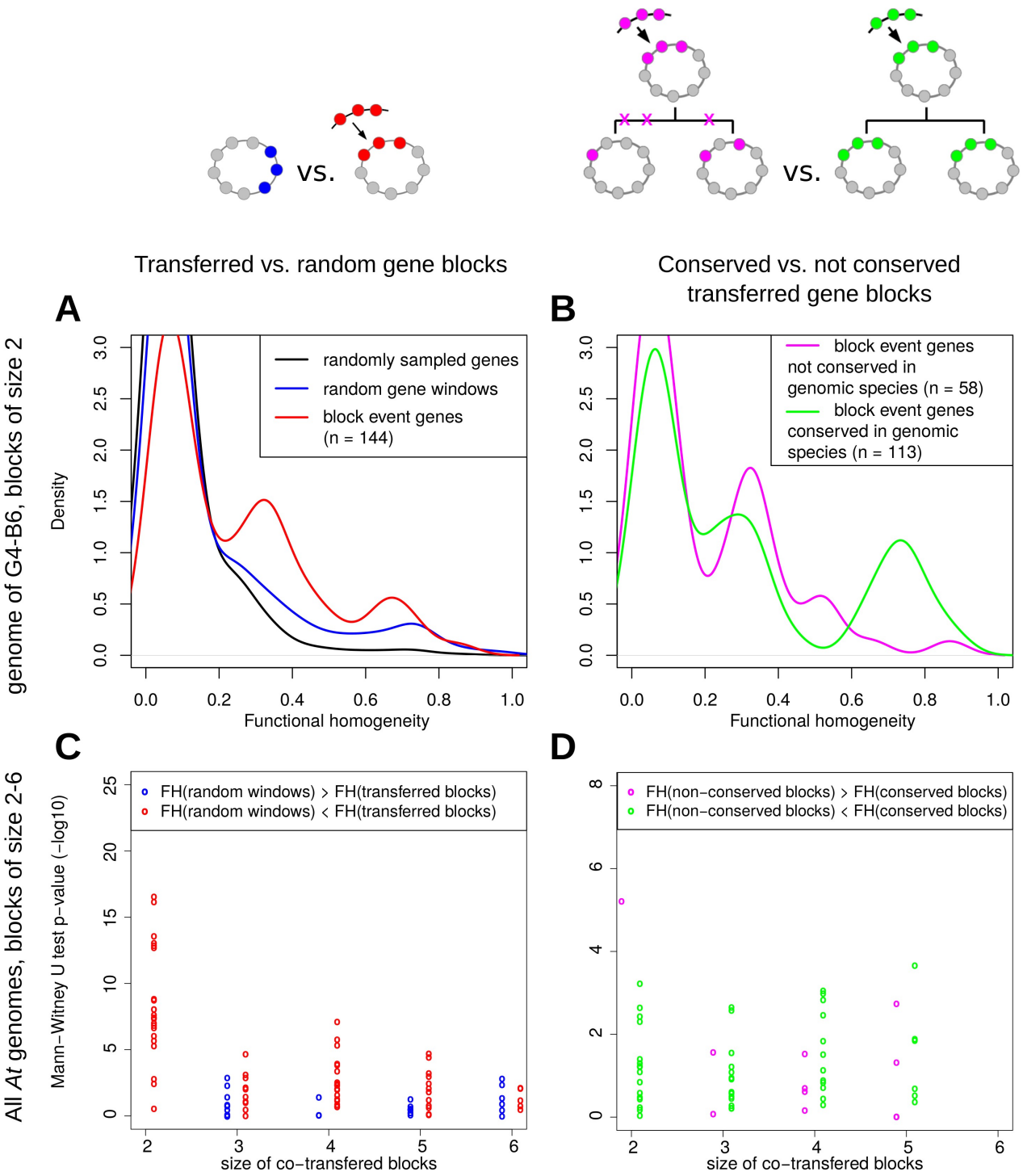


Figure 4: Functional homogeneity of gene clusters. (A, B) Distribution of functional homogeneities (*FH*) of genes within clusters using representative plots comparing clusters of two genes in the B6 genome (a G4 member). (A) Comparison of functional homogeneities of groups of two genes taken in the B6 genome: randomly distant pairs (black), any pair of neighbour genes without common transfer history (blue) or pairs of co-transferred neighbour genes (red). (B) Comparison of functional homogeneities of pairs of co-transferred genes from families conserved in all G4 strains (green) or not conserved (red). (C, D) Distribution of *p*-values of Mann-Whitney-Wilcoxon sum of ranks test comparing the distributions of *FH* (made independently for all *At* genomes at all discrete block sizes) of (C) random windows of non co-transferred genes vs. blocks of co-transferred genes or (D) conserved vs. non-conserved blocks of co-transferred genes. Each point represents an observation from an extant *At* genome for a given size of groups of genes

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(on x-axis). Point colours indicate the higher-*FH* sample (as in A,B): (C) blue, *FH*(random windows) > *FH*(transferred blocks), (n = 29, 4 significant); red, *FH*(random windows) < *FH*(transferred blocks), n = 66 (45 significant) ; (D) purple, *FH*(non-conserved blocks) > *FH*(conserved blocks), n = 11 (2 significant) ; green, *FH* (non-conserved blocks) < *FH* (conserved blocks), n = 49 (11 significant). A test is considered significant at $p < 0.01$. Effective of tests in favour of one hypothesis or the other are counted over all independent tests made for each combination of *At* genomes and discrete block sizes.

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

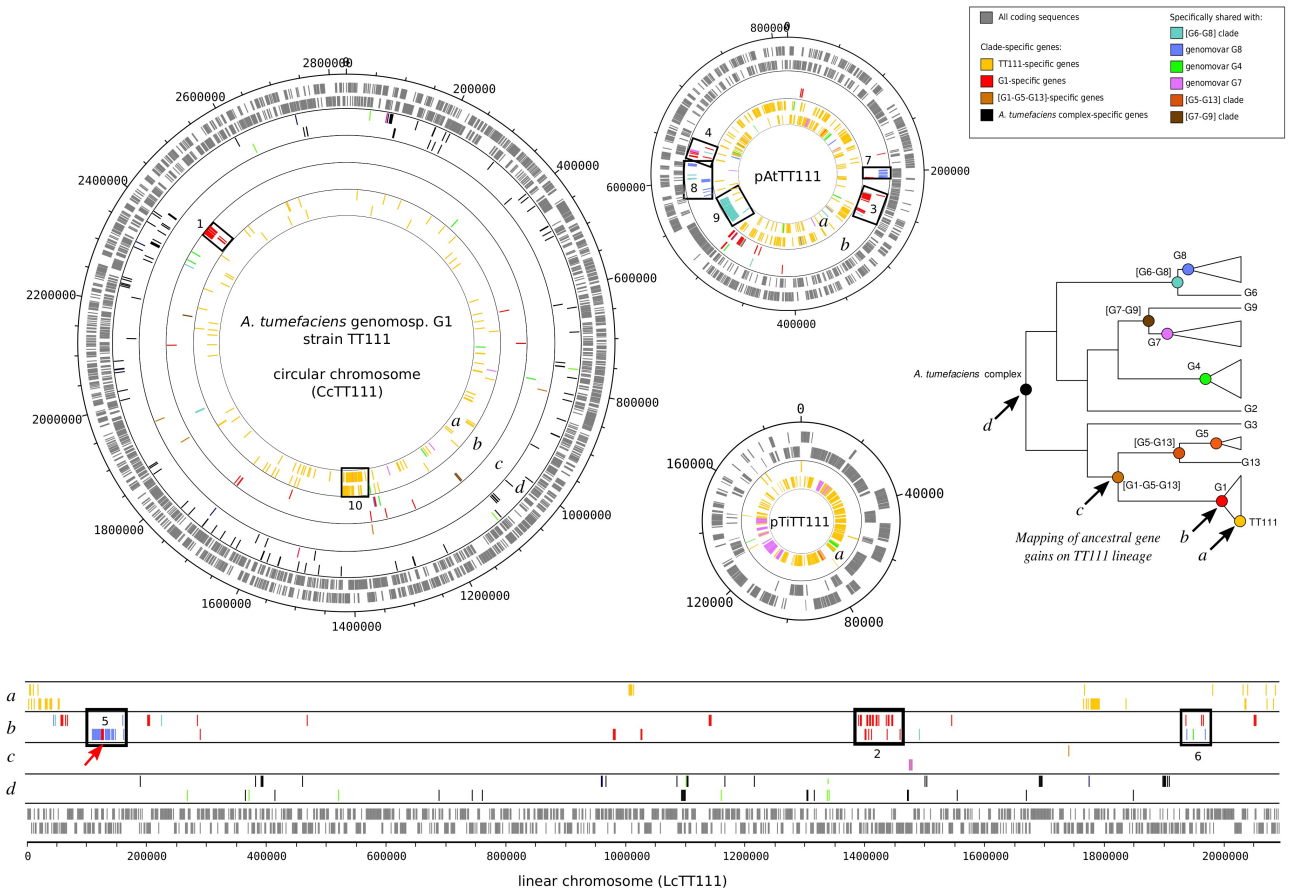


Figure 5: Historical stratification of gains in the lineage of the *Agrobacterium sp.* G1 strain TT111. The four replicons of the genome are represented circularly or linearly according to their molecular topology; replicons are not drawn to scale. Tracks within outermost ring (lowermost layer for linear chromosome) represent location of CDSs on both DNA strands. Other rings (layers) show genes that were acquired along the history of the TT111 lineage, and are labelled (*a-d*) according to the species phylogeny in the right inset.. Colors of genes indicate their specific presence in one of the clade that includes TT111, or their common specific sharing with another clade (see legend box). Outer (lower) vs. inner (upper) tracks in the same rings (layers) distinguish clade-specific genes with strict vs. relaxed specificity criterion. Numbered frames show particular gene clusters within TT111 genome: (1-4) G1-specific clusters: (1) AtSp2: chemotaxis regulation (*che2*) and aromatic compoud metabolism locus; (2) AtSp3: phenolics and amino-acid catabolism; (3,4) AtSp7 and AtSp9: phenolic compounds downstream degradation; (5-8), clusters specifically shared by G1 and G8: (5) AtSp14: lipopolysaccharide O-antigen biosynthesis and neoglucogenesis locus with G1-specific chemotaxis-regulating hybrid sensor (red arrow; red gene in Fig. S10); (6) AtSp12: outer-membrane lipoprotein and sensory protein; (7) AtSp17: deoxyribose uptake and assimilation; (8) AtSp15: exopolysaccharide (curdlan) synthesis, peptidoglycan modification and sensory protein; (9-10): clusters gained by TT111: (9) AtSp29: non-ribosomal peptide synthases involved in siderophore biosynthesis, shared by [G8-G6]; (10) prophage, partially shared by G3-CFBP6623 and G7-Zutra 3/1 (see Fig. S12).

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

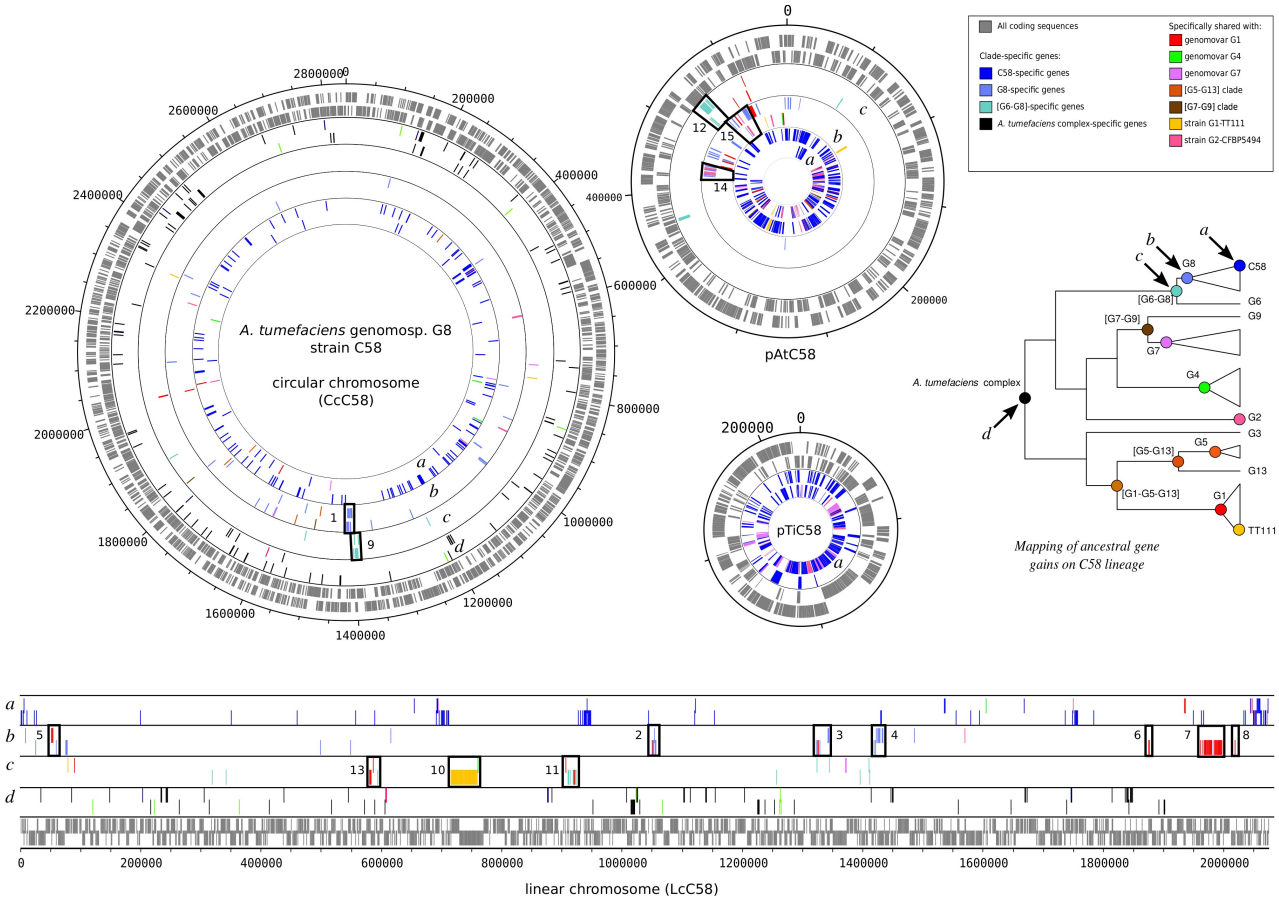


Figure 6: Historical stratification of gains in the lineage of *Agrobacterium sp.* G8 (*A. fabrum*) strain C58. Legend as in Fig. 5. Numbered frames show particular gene clusters within C58 genome: (1-4) G8-specific gene clusters: (1) Atsp21: degradation of hydroxy-cinamic acids (ferulic acid); (2) AtSp23: degradation of complex amino-acids (opine-like compounds); (3) AtSp24 and AtSp25: Drug/toxic resistance (extrusion transporters), sarcosine oxidase; (4) AtSp26: sensing of environmental signals (phenolic compound, mechanical constrains); (5-8) clusters specifically shared by G1 and G8: (5) AtSp15: exopolysaccharide (curdlan) synthesis, peptidoglycan modification and sensory protein; (6) AtSp13: iron-sensing two component system FeuPQ; (7) AtSp14: lipopolysaccharide O-antigen biosynthesis; (8) AtSp12: outer-membrane lipoprotein and sensory protein; (9-12) [G6-G8]-specific gene clusters: (9) AtSp29: sugar (L-sorbose) uptake and catabolism; (10) AtSp30: non-ribosomal peptide synthases involved in siderophore biosynthesis, shared by G1-TT111; (11) AtSp31: sugar metabolism; (12) dipeptide uptake and degradation; cluster specifically shared by G1 and [G6-G8]: (13) AtSp18: D-glucuronate uptake and degradation; (14-15) clusters specifically shared by G2 and G8: (14) AtSp27: Toxic extrusion / secondary metabolite secretion; (15) AtSp28: xanthine/cyclic compound degradation, two-component sensor.

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

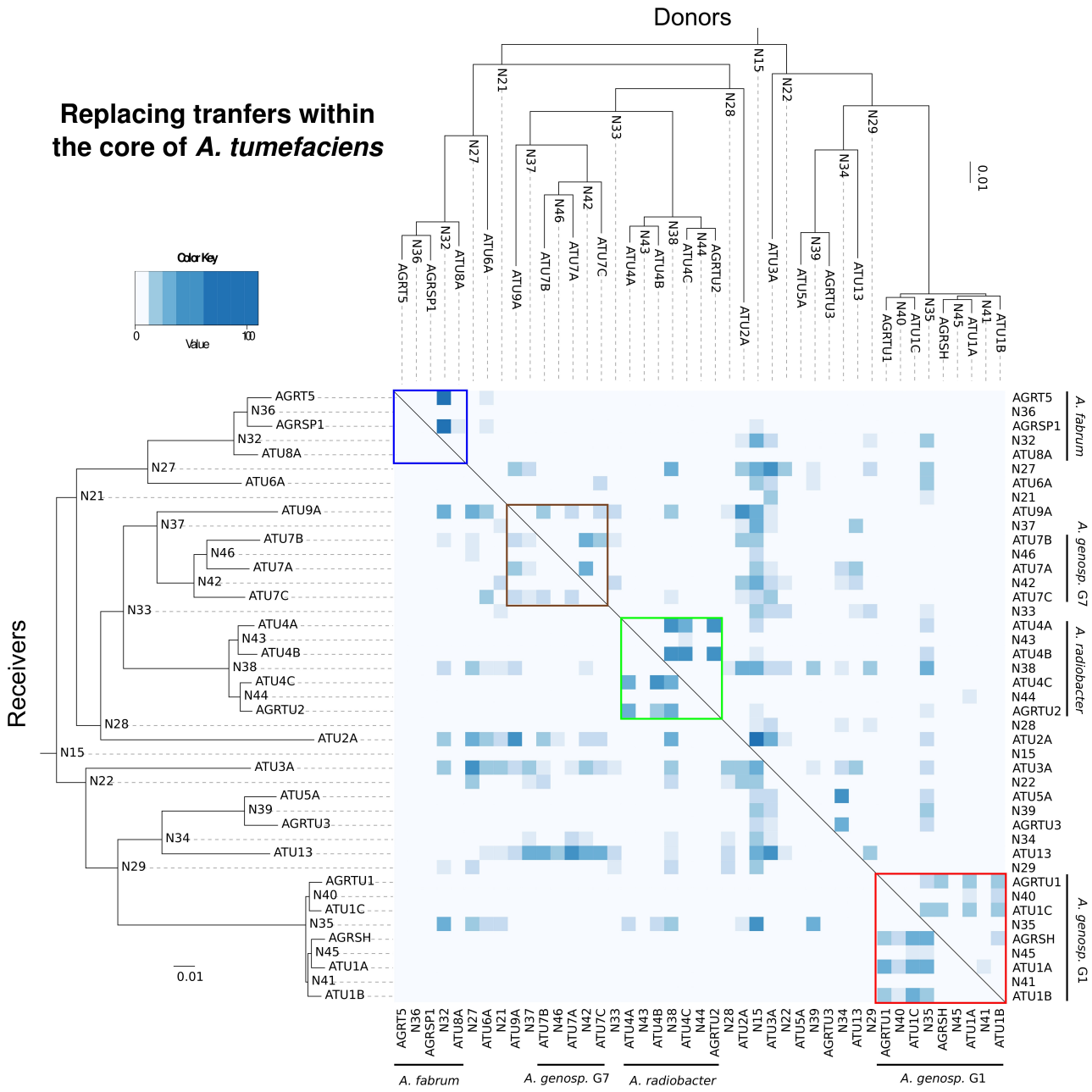


Figure 7: Intensity of gene transfers within the *At* clade. Heatmap of replacing transfers that occurred between pairs of ancestral or contemporary genomes. Each dot of the matrix indicates the count of transferred genes between nodes of the reference phylogeny drawn on left and top. Rows are receivers, columns are donors. Note that to give an insight into the proportion of genomes that recombined, each gene is counted as one transfer, regardless that it was involved in a unique block event. In case of remaining uncertainty on the location of the inferred receiver and donor, the count of the event is uniformly fractioned between possible receivers / donors. Frames highlight transfers whithin clades: blue, *A. fabrum* (= G8) (220 total transfer events [#], with on average 10.6 transfers per branch per 1,000 genes [t/b/kg]); brown [G7-G9] (# 133, 5.3 t/b/kg); green, *A. radiobacter* (= G4) (# 413, 14.6 t/b/kg); red, A. sp. G1 (# 338, 9.3 t/b/kg).

bioRxiv preprint doi: https://doi.org/10.1101/034843; this version posted December 18, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

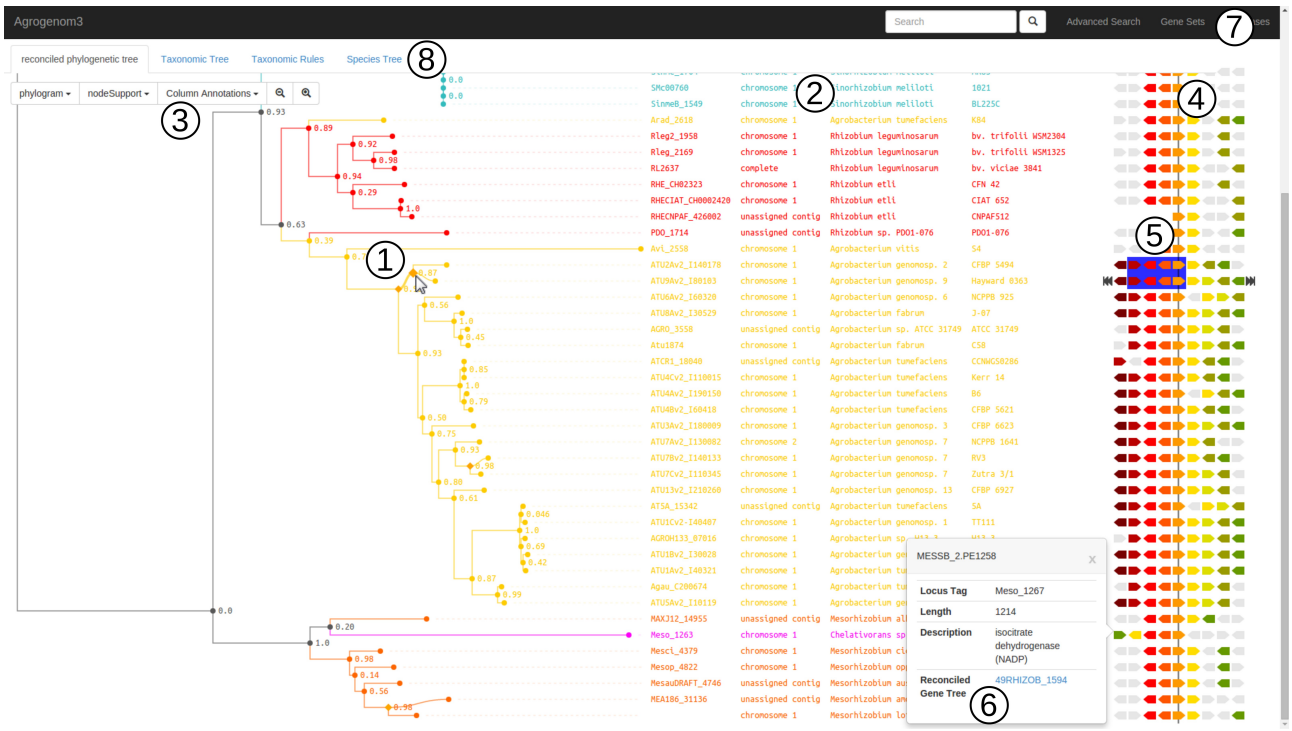


Figure 9: Snapshot of the Agrogenom web interface.

View of the *recA* gene family.(1) Reconciled gene tree; the orange diamond under the mouse cursor indicate a transfer event from G2-CFBP 5494 to G9-Hayward 0363. (2) Detailed annotation of the sequences at the tip of the tree, including locus tag (linking out to MaGe genome browser), chromosomal location, taxon name, database cross-references, etc. (3) Dynamic menu to adapt the level of displayed information. (4) Syntenic view in the genomic neighbourhoods of the focal gene family; homologs are sharing the same colour, defined in reference to a chosen sequence (indicated by the navigation arrows on the sides). (5) Blue frame indicate a block transfer event involving four gene families; this block appears dynamically when hovering the cursor above the transfer node in the gene tree. (6) Pop-up window with functional annotation and characteristics of a gene can be generated by double-clicking on the gene; it contains the link to the gene tree of the gene's family. (7) Search menus: 'Advanced search' to get a gene family by its annotation; 'Gene Sets' to browse lists of genes: clade-specific genes, core genome, ancestral gene content, clade-specific gains/losses. (8) Alternative views of the family with projection on the species tree.