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

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

[1] *Université de Lyon, F-69622 Lyon, France*

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

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

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

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

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

[7] *Direction de la Recherche Fondamentale, 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*

** Author for correspondence: Florent Lassalle, MRC Center for Outbreak Analysis and Modelling, Imperial College London, Praed Street, London W2 1NY, +44 207 594 1379, f.lassalle@imperial.ac.uk*

Abstract

Horizontal gene transfer (HGT) is considered a major source of innovation in bacteria, and as such is expected to drive the adaptation to new ecological niches. However, among the many genes acquired through HGT along the diversification history of genomes, only a fraction may have actively contributed to sustained ecological adaptation. Here, we implement a reverse ecology approach, involving the phylogenetic reconstruction of the evolutionary history of the pangenome within a bacterial clade and the modelling of the HGT process to recognize adaptive gene gains. We apply it to *Agrobacterium* biovar 1, a diverse group of soil and plant-dwelling bacterial species. We identify synapomorphic gene gains for major clades and show that most are organized into blocks of co-transferred genes encoding coherent biochemical pathways. This pattern of gene co-evolution rejects a neutral model of transfer, in which neighbouring genes would be transferred independently of their function. Instead, the conservation of acquired genes appears driven by purifying selection on collectively coded functions. We therefore propose synapomorphic blocks of co-functioning genes as candidate determinants of ecological adaptation of each clade. Their inferred biochemical functions define features of ancestral ecological niches, which consistently hint at the strong selective role of host plant rhizospheres.

Keywords: reverse ecology; ancestral genome; HGT; tree reconciliation; co-transferred genes; *Agrobacterium tumefaciens*.

Introduction

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 organisms. We now also have their genomes, which we believe contain all the information that make their lifestyle possible. However, even if we could describe the molecular function of every base in a genome, it would not necessarily allow us to understand whether this function is significant in its prevalent environment (Doolittle 2013). Another approach consists in recognizing traces of selection for functional adaptation in the histories of genomes. The comparison of genomes reveals a historical signal that can be used to reconstruct genome evolution, their hypothetical ancestral state, and the course of evolutionary events that shaped them in time. Using models of null expectation under neutral evolution, we can discern the events that have been decisive in the adaptive evolution of species. Bacterial genomes are in constant flux, with genes being gained and lost at rates that can exceed the nucleotide substitution rate (Lawrence & Ochman 1997). These 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 those that are found in some strains but not all. In *E. coli,* for example, the core genome comprises between 1,800 and 3,100 genes, depending on the methods and dataset, while the

accessory genome has more than 80,000 genes, with two random strains differing typically by a thousand genes (Touchon et al. 2009; Land et al. 2015). 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 (Young 2016); many are indeed only found in one genome. Among accessory genes, the majority must have ephemeral, if any, adaptive value for the bacteria, and are nothing more than selfish elements caught by the snapshot of genome sequencing (Daubin et al. 2003). 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 (Lassalle et al. 2015).

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. (Costechareyre et al. 2010). We found that genes specific to the species under focus, *i.e.* G8 now called *A. fabrum* (Lassalle et al. 2011), 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 analysing 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 micro-organisms, horizontal transfers. The set of events affecting each homologous gene family in the pangenome under scrutiny can be summarized into an evolutionary scenario, which can be seen as the path of gene evolution within and across branches of the tree of species (Scornavacca et al. 2012). 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 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 deduce 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.

Our reconstructed pangenome 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.

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 (Goodner et al. 2001; Wood et al. 2001; Li et al. 2011; Ruffing et al. 2011; Wibberg et al. 2011; Hao, Lin, et al. 2012; Hao, Xie, et al. 2012). 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 1; Figure S1). The dataset of 47 complete genome sequences (Table 1) was then used to construct a database of homologous gene families using the Hogenom pipeline (Penel et al. 2009) further used to elaborate the Agrogenom database. Bacterial growth was analysed 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 S1). Genome sequences were then assembled with Newbler version 2.6 (Roche Applied Sciences, Basel, Switzerland), using 90% identity and 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 (Kurtz et al. 2004). Genome sequences were then

annotated with the MicroScope platform (Vallenet et al. 2013) and made available through the MaGe web interface (www.genoscope.cns.fr/agc/microscope).

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 S2) extracted from the Agrogenom database were used for 500 jackknife samples (draws without replacement) of 25 gene alignment sets, which were each concatenated and used to infer a maximum-likelihood (ML) tree using PhyML (Guindon & Gascuel 2003) (the same parameters as for gene trees, see S1 Text.). The reference phylogeny was obtained by making the consensus of this sample of trees with CONSENSE algorithm from the Phylip package (Felsenstein 1993), and the branch supports were derived from the frequency of the consensus bipartitions in the sample (Figure S2). Alternative phylogenies were searched using the whole concatenate of 455 universal unicopy families or from a concatenate of 49 ribosomal protein gene families (Table S3) to compute trees with RAxML (version 7.2.8, GTRCAT model, 50 discrete site-heterogeneity categories) (Stamatakis 2006). All three methods yielded very similar results concerning the placement of the different genera and species (Figure S3).

Tree Pattern Matching. The collection of gene trees was searched with TPMS software (Bigot et al. 2013) for the occurrence of particular phylogenetic patterns signing the monophyly of different groups of strains (Figure S4). 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 Text, section 8.

Reconciliation of genome and gene tree histories. We computed a rooted gene tree using PhyML (Guindon & Gascuel 2003) for all the 10,774 gene families containing at least three genes (S1 Text). A pipeline was developed 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. This pipeline combines several methods dedicated to the recognition of different landmarks of duplication and transfers (for a review, see (Doyon et al. 2011)); it is fully detailed in the Supplementary Text Sections 3 and 4, and summarized below. Likely duplication events were first located by looking for multiple gene copies per species in clades of the gene trees, using the 'Unicity' algorithm from TPMS (Bigot et al. 2013) (Figure S5, step 1), identifying 17,569 putative duplications generating 28,343 potential paralogous lineages. We subsequently isolated subtrees from the global gene trees where every species was represented once, i.e. unicopy subtrees. In presence of lineage-specific paralogs ('in-paralogs'), we extracted the several 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 and pruning transferred subtrees (Abby et al. 2010), was run on the unicopy subtrees to detect replacing transfer events at branches with statistically significant (SH-like support > 0.9) topological conflict (Figure S5, step 2). These reconciliations of (potentially overlapping) local subtrees yielded a total of 22,322 high-confidence transfer events, which were integrated into a first set of coherently reconciled gene trees (Figure S5, step 3). To complete the reconciliation provided by Prunier, we used the 'TPMS-XD' algorithm (Bigot et al. 2013) to iteratively search for additional topological incongruences that had lower phylogenetic support but provided a global scenario more parsimonious on duplications and losses: 1,899 conflicting branches were recognized as the place of additional transfer events, allowing a decrease of 10,229 counts of duplication events. 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 (Figure S5, step 5). Finally, we used the Wagner parsimony algorithm implemented in the program Count (Csűrös 2008) to detect 19,553 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 (Figure S5, 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 (Figure S5, step 7). Transfer events are characterized by the location of both donor and receiver ancestor nodes 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. (Williams et al. 2012). Blocks were built by iterative inclusion of genes whose lineages were marked by events with compatible coordinates (as described above), allowing them to be spaced by genes without such signal ('gap' genes) (Figure S6 A, B). Blocks containing gap genes were checked for phylogenetic compatibility of those gap genes with the scenario associated to the block (Figure S6 C). When the gene tree of the gap gene showed weak statistical support (SH-like support < 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 (Figure S6 D). Jointly to its construction, the coordinates of the block event are refined by intersecting the coordinates of its constituent genes (Figure 1B; Figure S6 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 (Figure 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 (Figure 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 (Figure 1 B, step 3), in an analogous way than for the grouping of genes into leaf blocks. Block events were investigated only for origination (O), duplication (D) and transfer (T), not for speciation (S). Block events 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 (older independent neighbour events 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 unspecific aggregation in block of many unrelated events.

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 S1 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 is 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.

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) (Dimmer et al. 2011) 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 (Vallenet et al. 2013) : 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 analysed 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) (Ashburner et al. 2000). 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 (Schlicker et al. 2006; Pesquita et al. 2009). 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 favouring 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 neighbours' genes. *FH* were computed for all windows of neighbour 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 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.*

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 (Figure S7). This web interface for Agrogenom database is accessible at http://phylariane.univ-lyon1.fr/db/agrogenom/3/.

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 contain 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). Following the procedure of (Abby et al. 2012), a species phylogeny was inferred from the unicopy core gene set, using jackknife re-sampling of genes to compute branch supports (Figure S2). Remarkably, significant support is obtained for all clades corresponding to previously described species: *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. Within the *At* clade, groupings of higher order were also highly supported: 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] ([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 S2). In complement to the core gene concatenate, we used the whole set of individual gene phylogenies to test various hypotheses on the species tree topology (see Sup. Text, section 1), which showed that the proposed positions for G2 species and [G6-G8] clade had the best support pangenome-wide (Figure S4). A phylogeny reconstructed based on the genome gene contents show less appropriate to discriminate species, indicating a large quantity of HGT is occurring (Figure S8).

Reconciliation of gene and species histories. To reconstruct the history of HGT and other macro-evolutionary events, we reconciled the topologies of gene trees and species tree, i.e. we assigned an event of either origination, duplication, transfer (ODT), or speciation to each of the 467,528 nodes found in the rooted gene trees of the 10,774 families that contained at least three genes.

Our pipeline 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 2). 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 S4), 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 (Vallenet et al. 2009). Yet, ODT scenarios are generally evaluated for each gene tree independently of its neighbour (Makarova et al. 2006; Kettler et al. 2007). 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 event (Figure 1). We developed a procedure to recognise blocks of genes that co-evolved through the same event, allowing us to minimize the number of convergent ODT events along genome sequences.

First, using a combination of maximum-parsimony and maximum-likelihood criteria (Figure S5; Supplementary Methods), we infer scenarios of ODT events for each gene family independently. However, we leave undetermined the count and location of loss events, as there are often many possible combinations of loss events (in comparison to ODT events), with little information – only gene absence, i.e. missing data – to inform a choice. This results in degrees of freedom on the ODT scenarios (further referred to as the uncertainty of the scenario), with several connected branches of the species tree on which ODT events could equally have happened (Figure 1 A); this uncertainty around the ODT scenarios is recorded as the set of possible branches, referred to as the event coordinates in the species tree (see Methods). We then proceed to recognize blocks of neighbour genes that have ODT events mapped to overlapping coordinate sets. For instance, genes from family 1 and 2 are neighbour in an extant species' genome; when a transfer event occurs in family 1, the reconciled tree of family 2 is scanned for events with similar donor-recipient coordinates. (Figure 1A). If there is a similar event, i.e. the sets of coordinates for each single gene family's events overlap, we hypothesize a common event involving both families. The donor-recipient coordinate set for that joint event is then defined as the intersection of both single events' coordinate sets. Doing so, we reduce the uncertainty on the joint event coordinates (Figure S9). We then proceed to the next neighbour gene, trying to extend the event block.

A joint scenario may not be the most parsimonious in losses for individual gene families (Figure 1 B). However, as this regional pattern is not likely to happen by chance and because ODT events are less frequent than gene loss (Kuo et al. 2009; David & Alm 2011; Szöllősi et al. 2012) and thus less likely to happen in a convergent way, factorizing ODT events for neighbour genes appears a suitable and relevant procedure to minimize the total number of all kinds 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 (Figure S5 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 (Figure S10 A). Moreover, reconstructed blocks of ancestral genes that were hypothetically transferred between ancestral genomes appear to have been even much larger than extant ones (Figure S10 B), showing how frequently rearrangements and partial losses in descendant genomes have dismantled the syntenic blocks involved in ancient transfers. 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 2). 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 is much more parsimonious than the simple sum of all single-gene scenarios (Table 2). 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).

Genome histories reveal lineages with biased fixation of new genes. The reconstructed history of gain and 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 (Figure 2, Table S4). 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 (Figure 2), indicating the presence in contemporary genomes of a large polymorphism of gene presence/absence.

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) (Figure S11 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) (Figure S11 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 (Figure 3). We could recognise 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 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) (Figure S11 F, Figure S12). 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 (Figure S11 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 – all belong to [G1-G5-G13] and [G6-G8] (Figure S12). 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, i.e. having been domesticated.

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 then is expected to spread among close relatives (Cohan & Koeppel 2008). 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 (Lassalle et al. 2015) or the change in relative reliance on a set of resources (Kopac et al. 2014). 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 they are 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 an event can typically happen with the co-transfer of a complete operon. In a previous study focussed on G8 genomes (Lassalle et al. 2011), 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 genes collectively coding a function have been selected amongst 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 (Lassalle et al. 2011).

Yet, it is well known that bacterial genomes are organized in functional units such as operons, super-operons, etc. (Rocha 2008), 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 (Figure 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, Figure 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 (Figure 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 (Lawrence & Roth 1996), as transfer followed by selection for readily functional multi-genic traits would lead to the prevalence in 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, Figure 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* (Homma et al. 2007). 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, which are more likely to carry the sufficient information to encode a new adaptive trait, were under continued selection since their acquisition. It follows that the adaptations that characterize the ecological niche of a clade should be revealed through the identification of the genes specifically conserved in a clade, and notably those grouped in clusters with related functions.

Identification of clade-specific genes in *A. tumefaciens* key clades. We thus explored the histories of gene gain and loss in the clades of *At* to identify 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 of *At*. Listings of clade-specific genes of those key *At* clades can be found in S1

Dataset, or can be browsed on Agrogenom database website http://phylariane.univ-lyon1.fr/db/agrogenom/3/ (Figure 5). Generally, clade-specific genes were often located in relatively large clusters encoding coherent biochemical functions or pathways, which are summarized in Table S5 and hereafter numbered with the AtSp prefix. Those clade-specific gene clusters often match transfer or origination block events identified above (S1 Dataset), although often with limited coverage or with several transfer blocks mapping to a single 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. Full description of the biochemical functions encoded in these clade-specific repertoires can be found in the Supplementary Material (Supplementary Text section 7). A subset of clades of the *At* phylogeny, including species G1, G8, G4 and G7, are represented by several closely related extant genomes, and for this reason were particularly amenable for the accurate definition of clade-specific gene repertoires. For these, chromosomal maps of species-specific genes (Figures S13, S14, S15, S16) 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 plasmid (pAt) (Tables S6, S7).

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 (Harrison et al. 2010), and a variable complement of plasmids of various sizes (Young et al. 2006). More specifically, the chromid of the *Agrobacterium* genus (Mousavi et al. 2015; Ormeño-Orrillo et al. 2015), which includes the *At* clade, is linear (Slater et al. 2009, 2013) as the result of an unique ancestral event of linearization and thus constituting a synapomorphy of this clade (Ramírez-Bahena et al. 2014). Another general feature of *At* genomes is the frequent presence of a pAt, a megaplasmid which was for long referred to as the cryptic plasmid, for its role in the cell biology remains largely unknown. These pAts belong to the larger family of *repABC* (mega-)plasmids, which can conjugate between a broad range of hosts among Rhizobiaceae, as described for symbiotic (Sym) or tumor-inducing (Ti) plasmids in *Rhizobium* and *Agrobacterium*, respectively (González et al. 2003; Gonzalez et al. 2010; Lassalle et al. 2011). In addition, it is well known that pTis can be transferred between strains of different species of *At*, as confirmed by the large similarity observed here between the pTiB6 and pTiTT111 (Figure S15). This suggests that pAts could similarly be transferred amongst species of *At*. However, we found that pAt types are restricted to certain genomic backgrounds, as they host genes that are mostly strain-specific, interspersed with gene clusters specific to the host genome's species (namely G1, G8, G4 and G7 species) or higher groups ([G6-G8] clade) (Figures S13, S12, S13, S16). In particular, 25 G8-specific genes and 11 [G6-G8] clade-specific genes were retrieved on the pAts of the corresponding strains. Our previous study using micro-arrays with G8-C58 genome as a reference did not identify G8-specific genes on plasmids, notably because the G8 strain LMG-46 lacked a pAt (Lassalle et al. 2011). This feature is however unique to the strain within G8 species, which could result from a recent plasmid loss with rare prevalence in the species.

Similarly, the only available isolate of G9 species, strain Hayward 0363, has no detectable plasmid. In both cases, further sampling of wild population is required to test whether the suggested species' status of presence/absence of a pAt is the rule, or if intermediate prevalences can occur. Considering the data presented here, the rule seems to be that pAts are core replicons for most *At* species. In addition, the occurerence of clade-specific genes on the pAt and never on the other plasmids (pTi and smaller ones), in face of its putative ability to transfer widely, suggests a existence of barriers to its transfer. Within Cohan's ecotype framework, we interpret this pattern as the presence of determinants of the species' ecological niche on this particular extra-chromosomal element, which selectively prevents its spread among closely related species (Cohan & Koeppel 2008). This suggests that – for most species of *At* – this third replicon is probably essential in natural environments, which would qualify it as a *bona fide* chromid (Harrison et al. 2010).

Clade-specific gene functions provide 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 (Vogel et al. 2003), dictates under the competitive exclusion principle (Gause 1932) 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 (Costechareyre et al. 2010), and G2 members appear to be specialized towards opportunistic pathogenicity in human (Aujoulat et al. 2011), showing the existence of some kind of niche differentiation among *Agrobacterium* species. Because the clade-specific genes must encode what makes a clade's ecology to be distinct from that of its relatives (Lassalle et al. 2015), we explored the specific functional repertoire of At clades. Strikingly, in most clades, including species or higher-level groups, the sets of clade-specific genes recurrently presented the same classes of functions. These include the transport and metabolism of phenolic compounds, aminoacids and complex sugars, and the production of exopolysaccharides and siderophores, all of which can be related to the life in the plant rhizosphere (Lassalle et al. 2011).

Among these, we can notably report the specific presence of a supernumerary chemotaxis regulation operon *che2* in species G1, which is uniquely linked with an array of genes with predicted functions involved in the catabolism of (possibly aminated) aromatic compounds (Table S5). This suggests G1 strains are able to specifically degrade certain – yet unknown – aromatic compounds, for which they might display specific tropism and/or induction of biofilm formation.

G8 species and [G6-G8] clade present a number of clade-specific gene clusters (Table S5), as previously reported (Lassalle et al. 2011), among which the largest are the ferulic acid degradation and siderophore biosynthesis operons, which have been recently reported to provide a growth advantage and to be expressed in a coordinated manner in a plant rhizosphere environment (Campillo et al. 2014; Baude et al. 2016), showing the joint participation of G8 lineage-specific genes in the adaptation to a plant-related specific ecological niche. Interestingly, the gain of a siderophore biosynthesis locus in the ancestor of [G6-G8] coincided with the loss of the locus coding biosynthesis of another siderophore, agrobactin, which is

15

otherwise ubiquitous in – and unique to – the *At* clade. This conserved switch to a different pathway for iron-scavenging – a crucial function in iron-depleted plant rhizospheres – may provide a competitive advantage with respect to co-occurring agrobacteria.

The species group [G5-G13] specifically presents a phenylacetate degradation pathway operon (Table S5), which biochemical function was demonstrated *in vitro* (Figure S17). This discovery readily provides us with a specific biochemical test for identification of these species, and again hints to the particular affinity of agrobacteria with aromatic compounds that are likely to be found in plant rhizospheres.

Finally, the large cluster that encodes the nitrate respiration (denitrification) pathway, including *nir*, *nor*, *nnr* and *nap* operons is absent from [G1-G5-G13] clade. More recently, this gene cluster was also lost by strains G9-NCPPB925 and G8-ATCC31749, and its presence in strain G3-CFBP6623 seems to result from later transfer from a mosaic of sources within *At*. Considering the absence of this super-operon in close relatives of *At* such as *A. vitis* and *R. leguminosarum*, it appears that it was likely acquired by the ancestor of [G2-G4-G7-G9-G6-G8] clade (node N21 on Figure 1), one of the two large clades that divide *At* complex. These strains possessing the denitrification pathway may be selectively advantaged under certain anaerobic or micro-aerophilic conditions, such as those met in certain soils and rhizospheres.

Species G1 and G8 present a particular case of genomic gene content convergence, as we found that they share 57 synapomorphic genes (Table S6 and S7), in most cases with phylogenetic support for transfer events between respective ancestors. These traits were previously hypothesized to provide key adaptation to the life in the plant rhizosphere of G8 (= *A. fabrum*, (Lassalle et al. 2011)). For instance, they share homologous genes involved in the biosynthesis of curdlan – a cellulose-like polysaccharide – and the biosynthesis of O-antigens of the lipopolysaccharide (LPS), both capsular components that may define attachment properties of the cell to the external environment (Table S5; Supplementary Text S1, section 6.1). Indeed, the LPS synthesized by homologous enzymes in *Brucella spp.* mediate a specific interaction with cells of a eukaryotic host (Vizcaíno et al. 2001). 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 S5).

This overlap of the niche-specifying gene repertoires of G1 and G8 species may cause the convergence of their ideal ecological niches, and thus lead to inter-species competition for resources. However, shared niche-specifying genes are combined to different sets of species-specific genes in each species' core-genome, with which epistatic interaction could induce strong divergence in their phenotype. Typically, even though the loci for biosynthesis of O-antigens of the LPS in G1 and G8 are highly similar (>93% amino acid identity in average for proteins of the AtSp14 locus, Figure S18) and probably produce an equivalent compound, these species' regulation of biofilm production is likely different. Indeed, there are several regulatory genes specific to G1 genomes involved in chemotaxis/biofilm production regulation, such as the *che2* operon (cluster AtSp2) and hub signal-transducing protein HHSS ('hybrid-hybrid' signal-sensing, see S1 Text, section 6.1) found in cluster AtSp14 (Figure S13 and S18), 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; this potentially constitute a whole pathway for response to environmental conditions specific of the niche of G1, such as availability of phenolics as nutrients. Similarly, G8-specific genes of the AtSp26 cluster (Figure S14) are forming a regulatory island involved in the perception and transduction of environmental signals, including mechanosensitive channels and a receptor for phenolic compound related to toluene (Lassalle et al. 2011).

Both species are thus likely to orchestrate the production of similar polysaccharides under different regulation schemes, involving the coordination of their expression with other specific traits – interestingly, in both cases the catabolism of (likely different) phenolics. Similarly, the coordinated expression of several clade-specific genes, resulting in conditional phenotypes has recently been observed in G8-C58 (Baude et al. 2016), strengthening the idea of the existence of an ecological niche to which G8 species is specifically adapted through the expression of a particular *combination* of specific genes. The partial hybridization of G1 and G8 specific genomes thus likely leads each species to tap the same resources in a different way, which should not induce significant competition between them. These species may then form guilds of relatives that exploit partitions of a largely common ecological niche (Lassalle et al. 2015), explaining why we can observe them co-occurring in soils (Vogel et al. 2003; Portier et al. 2006). Regular events of such inter-species exchange of niche-specifying may explain why exploration of the genome based only on the presence/absence pattern of homologs – and not their gain history, as done here – may yield indistinct patterns of clade-specific gene contents, as recently observed among *R. leguminosarum* genomic species (Kumar et al. 2015).

Conclusion

We developed an original method for the reconstruction of the history of all genes in bacterial 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 signatures of purifying selection for genes specifically gained by major clades of *At*. Most of these were organized in large blocks of co-evolving genes that encode coherent pathways, a pattern which constitutes a departure from a neutral model of gene transfer in bacterial genomes. We thus consider these blocks of clade-specific genes as likely determinants of the clades' core ecology. 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 plant roots, which suggest that ecological differentiation of *Agrobacterium* clades occurred through the partitioning of ecological resources available in plant rhizospheres. In the future, sampling of within-

species diversity, coupled with population genomics approaches, could further reveal ecological properties of agrobacteria, including those that may be non-ubiquitous but dynamically maintained by recombination within species (Kashtan et al. 2014; Rosen et al. 2015). Gene co-evolution models, such as the one developed here, may be extend to the investigation of inter-locus linkage in populations of genomes (Cui et al. 2015). Such analyses could reveal complex interactions between molecular pathways that are under ecological selection, opening new ways towards the understanding of bacterial adaptation to the infinite diversity of micro-environments.

Acknowledgements

This project was supported by the French National Research Agency (ANR, http://www.agence-nationale-recherche.fr/) 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 European Research Council (ERC, http://erc.europa.eu/) through grant BIG_IDEA (260801). 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 'Programme Investissement d'Avenir' of the ANR, contract ANR-10-INBS-09) are acknowledged for support within the MicroScope annotation platform.

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. Accession numbers for the corresponding replicon sequences are LN999991-LN999996, LT009714-LT009764 and LT009775-LT009780. Accession numbers of all genomic data used in this dataset are listed Table 2. All other relevant data (output of analyses) are available as Supporting Information files. Original software developed for this work are available at https://github.com/flass/agrogenom.

Competing interests

The authors have declared that no competing interests exist.

Authors' contributions

FL, XN and VD conceived and supervised the study. DC, DMu and LV cultivated the bacteria, prepared the samples and conducted biochemical tests. FL, RP, SP, TG 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, TG 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.

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)

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

Cc – circular chromosomes

Lc – linear chromid

pAt – *At* plasmid

pTi – Tumor-inducing plasmid

Supporting Information

Figure S1. Phylogeny of 131 genomes of Alpha-proteobacteria.

Figure S2. Reference phylogeny of Rhizobiales history.

Figure S3. Alternative reference tree topologies obtained with different methods.

Figure S4. Support for monophyply of groups in all gene trees.

Figure S5. Bioinformatic pipeline for reconciliation of gene and genome histories.

Figure S6. Algorithm for construction of blocks of co-transferred genes.

Figure S7. Schema of Agrogenom relational database.

Figure S8. Hierarchical clustering of *Rhizobiales* genomes according to their gene content.

Figure S9. Gain in reconciliation precision while amalgamating block scenarios.

Figure S10. Distribution of sizes of block events.

Figure S11. Gene gain, loss and conservation within *At* clade ancestors.

Figure S12. Residuals of negative exponential regression of clade age vs. conservation of gained genes.

Figure S13. Historical stratification of gains in the lineage of *A. sp* G1 strain TT111.

Figure S14. Historical stratification of gains in the lineage of *A. sp.* G8 (*A. fabrum*) strain C58.

Figure S15. Historical stratification of gains in the lineage of *A. sp* G4 (*A. radiobacter*) strain B6.

710

20

References

Abby SS, Tannier E, Gouy M, Daubin V. 2010. Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. BMC Bioinformatics.

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

Ashburner M et al. 2000. Gene Ontology: tool for the unification of biology. Nat. Genet. 25:25–29.

Aujoulat F et al. 2011. Multilocus sequence-based analysis delineates a clonal population of *Agrobacterium* (*Rhizobium*) *radiobacter* (*Agrobacterium tumefaciens*) of human origin. J. Bacteriol. 193:2608–2618.

Baude J et al. 2016. Coordinated Regulation of Species-Specific Hydroxycinnamic Acid Degradation and Siderophore Biosynthesis Pathways in *Agrobacterium fabrum*. Appl. Environ. Microbiol. 82:3515–3524.

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

Bruen TC, Philippe H, Bryant D. 2006. A Simple and Robust Statistical Test for Detecting the Presence of Recombination. Genetics. 172:2665–2681.

Campillo T et al. 2014. Analysis of Hydroxycinnamic Acid Degradation in *Agrobacterium fabrum* Reveals a Coenzyme A-Dependent, Beta-Oxidative Deacetylation Pathway. Appl. Environ. Microbiol. 80:3341–3349.

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

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

Costechareyre D et al. 2010. Rapid and Efficient Identification of *Agrobacterium* Species by recA Allele Analysis : *Agrobacterium* recA Diversity. Microb. Ecol. 60:862–72.

Csűrös M. 2008. Ancestral Reconstruction by Asymmetric Wagner Parsimony over Continuous Characters and Squared Parsimony over Distributions. In: Comparative Genomics. Nelson, CE & Vialette, S, editors. Lecture Notes in Computer Science Springer Berlin Heidelberg pp. 72–86.

Cui Y et al. 2015. Epidemic Clones, Oceanic Gene Pools, and Eco-LD in the Free Living Marine Pathogen *Vibrio parahaemolyticus*. Mol. Biol. Evol. 32:1396–1410.

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

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

Dimmer EC et al. 2011. The UniProt-GO Annotation database in 2011. Nucleic Acids Res. 40:D565–D570.

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

Doyon J-P, Ranwez V, Daubin V, Berry V. 2011. Models, Algorithms and Programs for Phylogeny Reconciliation. Brief. Bioinform. 12:392–400.

Felsenstein J. 1993. *PHYLIP (Phylogeny Inference Package) version 3.5c*.

Fitch WM. 1971. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. Syst. Zool. 20:406.

Friedman J, Alm EJ, Shapiro BJ. 2013. Sympatric Speciation: When Is It Possible in Bacteria? PLoS ONE. 8:e53539.

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

Gonzalez V et al. 2010. Conserved Symbiotic Plasmid DNA Sequences in the Multireplicon Pangenomic Structure of *Rhizobium etli*. Appl Env. Microbiol. 76:1604–1614.

González V et al. 2003. The mosaic structure of the symbiotic plasmid of *Rhizobium etli* CFN42 and its relation to other symbiotic genome compartments. Genome Biol. 4:R36–R36.

Goodner B et al. 2001. Genome sequence of the plant pathogen and biotechnology agent Agrobacterium tumefaciens C58. Science. 294:2323–2328.

Guindon S, Gascuel O. 2003. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. Syst. Biol. 52:696–704.

Hao X, Xie P, et al. 2012. Genome Sequence and Mutational Analysis of Plant-Growth-Promoting Bacterium *Agrobacterium tumefaciens* CCNWGS0286 Isolated from a Zinc-Lead Mine Tailing. Appl. Environ. Microbiol. 78:5384–5394.

Hao X, Lin Y, et al. 2012. Genome Sequence of the Arsenite-Oxidizing Strain *Agrobacterium tumefaciens* 5A. J. Bacteriol. 194:903.

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

Homma K, Fukuchi S, Nakamura Y, Gojobori T, Nishikawa K. 2007. 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. 24:805–813.

Kashtan N et al. 2014. Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild *Prochlorococcus*. Science. 344:416–420.

Kettler GC et al. 2007. Patterns and Implications of Gene Gain and Loss in the Evolution of Prochlorococcus. PLoS Genet. 3:e231.

Kopac S et al. 2014. Genomic Heterogeneity and Ecological Speciation within One Subspecies of *Bacillus subtilis*. Appl. Environ. Microbiol. 80:4842–4853.

Kumar N et al. 2015. Bacterial genospecies that are not ecologically coherent: population genomics of *Rhizobium leguminosarum*. Open Biol. 5:140133.

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

Kurtz S et al. 2004. Versatile and open software for comparing large genomes. Genome Biol. 5:R12.

Lassalle F et al. 2011. Genomic Species Are Ecological Species as Revealed by Comparative Genomics in *Agrobacterium tumefaciens*. Genome Biol. Evol. 3:762–781.

Lassalle F, Muller D, Nesme X. 2015. Ecological speciation in bacteria: reverse ecology approaches reveal the adaptive part of bacterial cladogenesis. Res. Microbiol. 166:729–741.

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

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

Li A et al. 2011. Genome Sequence of Agrobacterium tumefaciens Strain F2, a Bioflocculant-Producing Bacterium. J. Bacteriol. 193:5531–5531.

Makarova K et al. 2006. Comparative genomics of the lactic acid bacteria. Proc. Natl. Acad. Sci. 103:15611–15616.

Marri PR, Harris LK, Houmiel K, Slater SC, Ochman H. 2008. The Effect of Chromosome Geometry on Genetic Diversity. Genetics. 179:511–516.

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

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

Ormeño-Orrillo E et al. 2015. Taxonomy of rhizobia and agrobacteria from the Rhizobiaceae family in light of genomics. Syst. Appl. Microbiol. 38:287–291.

Penel S et al. 2009. Databases of homologous gene families for comparative genomics. BMC Bioinformatics. 10:S3.

Pesquita C, Faria D, Falcão AO, Lord P, Couto FM. 2009. Semantic Similarity in Biomedical Ontologies. PLoS Comput Biol. 5:e1000443.

Portier P et al. 2006. Identification of genomic species in *Agrobacterium* biovar 1 by AFLP genomic markers. Appl. Environ. Microbiol. 72:7123–7131.

Ramírez-Bahena MH et al. 2014. 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. 73:202–207.

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

Rosen MJ, Davison M, Bhaya D, Fisher DS. 2015. Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. Science. 348:1019–1023.

Ruffing AM, Castro-Melchor M, Hu W-S, Chen RR. 2011. Genome sequence of the curdlan-producing *Agrobacterium* sp. strain ATCC 31749. J. Bacteriol. 193:4294–4295.

Schlicker A, Domingues FS, Rahnenführer J, Lengauer T. 2006. A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics. 7:302.

Slater S et al. 2013. 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. 79:1414–1417.

Slater SC et al. 2009. Genome sequences of three *Agrobacterium* biovars help elucidate the evolution of multichromosome genomes in bacteria. J. Bacteriol. 191:2501–2511.

Stamatakis A. 2006. RAxML-VI-HPC: Maximum Likelihood-Based Phylogenetic Analyses with Thousands of Taxa and Mixed Models. Bioinformatics. 22:2688–2690.

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

Touchon M et al. 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet. 5:e1000344.

Vallenet D et al. 2009. MicroScope: a platform for microbial genome annotation and comparative genomics. Database J. Biol. Databases Curation. 2009:bap021.

Vallenet D. MicroScope Home - MaGe: Microbial Genome Annotation & Analysis Platform - MicroScope - Web Interface System & Specialized Databases for (re)Annotation and Analysis of Microbial Genomes. MicroScope. https://www.genoscope.cns.fr/agc/microscope/home/index.php (Accessed December 18, 2015).

Vallenet D et al. 2013. MicroScope--an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. Nucleic Acids Res. 41:D636-647.

Vizcaíno N, Cloeckaert A, Zygmunt MS, Fernández-Lago L. 2001. 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. 69:6738–6748.

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

Volff JN, Altenbuchner J. 2000. A new beginning with new ends: linearisation of circular chromosomes during bacterial evolution. FEMS Microbiol. Lett. 186:143–150.

Wibberg D et al. 2011. Complete genome sequencing of *Agrobacterium* sp. H13-3, the former *Rhizobium lupini* H13-3, reveals a tripartite genome consisting of a circular and a linear chromosome and an accessory plasmid but lacking a tumor-inducing Ti-plasmid. J. Biotechnol. In Press, Uncorrected Proof.

Williams D, Gogarten JP, Papke RT. 2012. Quantifying Homologous Replacement of Loci between Haloarchaeal Species. Genome Biol. Evol. 4:1223–1244.

Wood DW et al. 2001. The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. Science. 294:2317–2323.

Young JPW et al. 2006. The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. Genome Biol. 7:R34.

Figure legends and tables

Fig. 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. 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.

Fig. 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.

Fig. 3: Retention of gained genes within *At* genomes follows 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).

Fig. 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 (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.

Fig. 5: 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 indicates 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

25

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.

Table 1: List of 47 Rhizobiales strains used in this study.

Clade/Taxon	Strain name	NCBI Taxid	EMBL Sequence accession number	Nb. of genes
***Agrobacterium biovar 1* / *A. tumefaciens* species complex (*At*)**				
A. sp. G1	H13-3	861208	CP002248-CP002250	5345
	5A	1107544	AGVZ00000000	5518
	CFBP 5771	1183421	LT009762-LT009764	5546
	S56	1183429	LN999991-LN999996	5627
	TT111	1183430	LT009714-LT009717	5856
A. sp. G2 (*A. pusense*)	CFBP 5494	1183436	LT009718-LT009722	6013
A. sp. G3	CFBP 6623	1183432	LT009723-LT009726	5378
A. sp. G4 (*A. radiobacter*)	B6	1183423	LT009758-LT009761	5875
	CFBP 5621	1183422	LT009727-LT009729	5330
	Kerr 14	1183424	LT009730-LT009734	5870
	CCNWGS0286	1082932	AGSM00000000	4979
A. sp. G5	CFBP 6626	1183435	LT009735-LT009738	5332
	F2	1050720	AFSD00000000	5321
A. sp. G6	NCPPB 925	1183431	LT009739-LT009744	6139
A. sp. G7	NCPPB 1641	1183425	LT009775-LT009778	6041
	RV3	1183426	LT009745-LT009747	5182
	Zutra 3/1	1183427	LT009748-LT009751	5685
A. sp. G8 (*A. fabrum*)	C58	176299	AE007869-AE007872	5639
	ATCC 31749	82789	AECL00000000	5535
	J-07	1183433	LT009752-LT009755	5592
A. sp. G9	Hayward 0363	1183434	LT009779-LT009780	4502
A. sp. G13	CFBP 6927	1183428	LT009756-LT009757	4993
Allorhizobium				
Allorhizobium vitis	S4	311402	CP000633-CP000639	5389
Rhizobium sp.	PDO1-076	1125979	AHZC00000000	5340
Rhizobium				
R. rhizogenes	K84	311403	CP000628-CP000632	6684
R. etli	CIAT 652	491916	CP001074-CP001077	6109
	CFN 42	347834	CP000133-CP000138, U80928	6016
	CNPAF512	993047	AEYZ00000000	6544
R. leguminosarum bv. *viciae*	3841	216596	AM236080-AM236086	7263
R. leguminosarum bv. *trifolii*	WSM1325	395491	CP001622-CP001627	7001
	WSM2304	395492	CP001191-CP001195	6415
***Ensifer/Sinorhizob*ium**				
E. meliloti	1021	266834	AL591688, AE006469, AL591985	6234
	BL225C	698936	CP002740-CP002742	6354
	CCNWSX0020	1107881	AGVV01000000	6844
	AK83	693982	CP002781-CP002785	6510
	SM11	707241	CP001830-CP001832	7093
E. medicae	WSM419	366394	CP000738-CP000741	6213
E. fredii	HH103	1117943	HE616890-HE616899	6787
	NGR234	394	CP000874, CP001389, U00090	6366
Mesorhizobium/ Chelativorans				
M. alhagi	CCNWXJ12-2	1107882	AHAM00000000	7184
M. amorphae	CCNWGS0123	1082933	AGSN00000000	7075
M. australicum	WSM2073	7540353	AGIX00000000	5934
M. ciceri bv. *biserrulae*	WSM1271	765698	CP002447, CP002448	6264
M. opportunistum	WSM2075	536019	CP002279	6508
M. loti	MAFF303099	266835	AP003017, BA000012, BA000013	7281
Chelativorans sp.	BNC1	266779	CP000389-CP000392	4543
Parvibaculum				
P. lavamentivorans	DS-1	402881	CP000774	3636

Table 2: Origination, Duplication, Transfer and Speciation events inferred in reconciliations of the Agrogenom database.

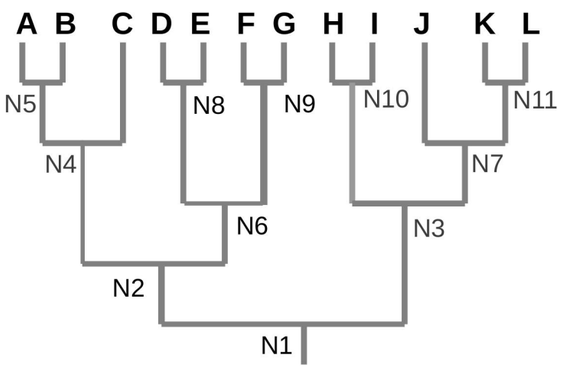
Event type	Single gene events	Block events	(of size >1)	Difference after event integration into blocks
Originations	5,189	4,267	(667)	-922
Duplications	7,340	5,819	(778)	-1,521
Total transfers	43,233	32,255	(5,649)	-10,978
Replacing transfers †	9,271	-	-	
Additive transfers †	33,962	-	-	
Total ODT	55,762	42,341	(7,094)	-13,421
Implied losses	29,843	32,739	-	+2,896
Total ODTL	85,605	75,080	-	-10,525

780

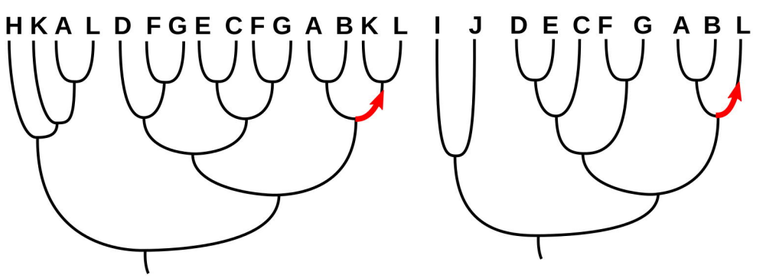
O, origination; D, duplication; T, transfer; L, loss; ODT refers to the combination of all O, D and T events, ODTL also includes losses. †: Replacing and additive transfers are not distinguished in block events.

A.

Reconciliation scenarios projected on the species tree:

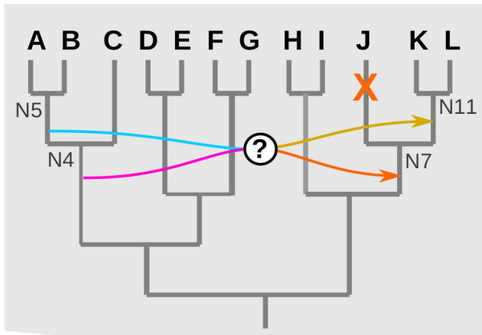


Species tree

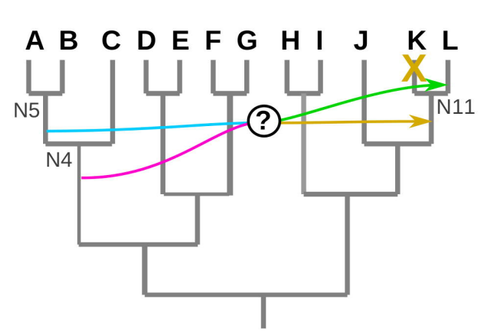


Gene tree 1 **Gene tree 2**

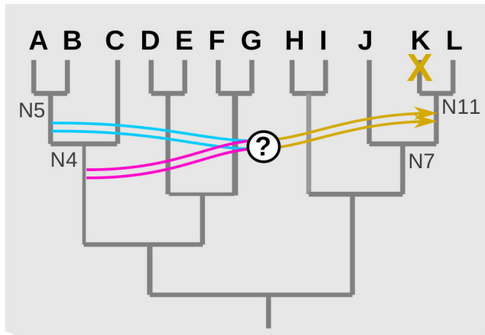
Gene family 1



Gene family 2



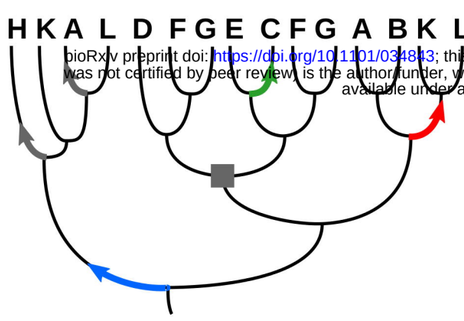
Block (gene fam. 1 & 2)



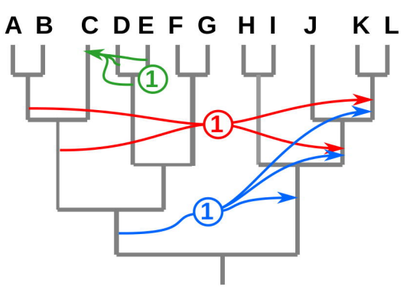
		Single gene 1		Single gene 2		Block (genes 1 & 2)	
Transfer coordinates		Do	Re	Do	Re	Do	Re
	Possible LCA	{N4, N5} N5	{N7, N11} N11	{N4, N5} N5	{N11, L} L	{N4, N5} N5	N11 N11
Local event count							
	Transfers	1		1		1	
	Losses	0		0		1	
Combined event count							
	Transfers	**2**				**1**	
	Losses	**0**				**1**	

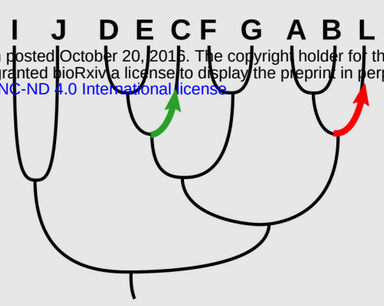
B.

1) Reconstruction evolutionary events in gene families

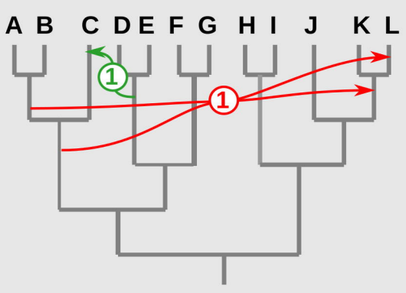


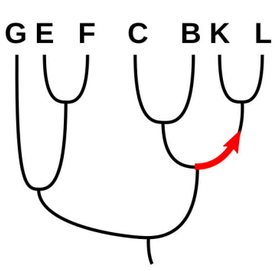
Gene family 1



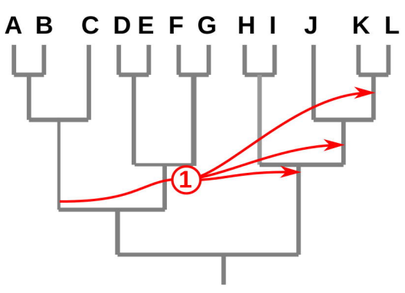


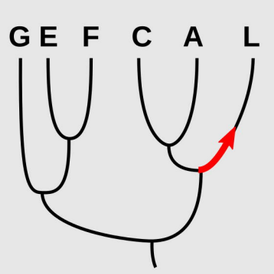
Gene family 2



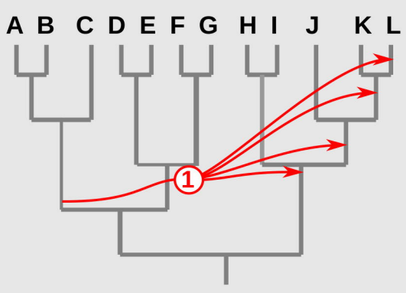


Gene family 3



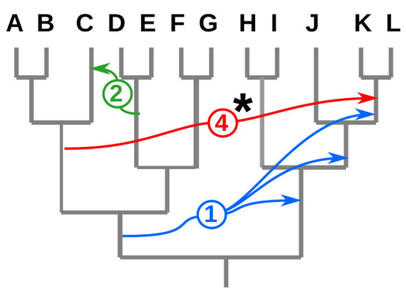


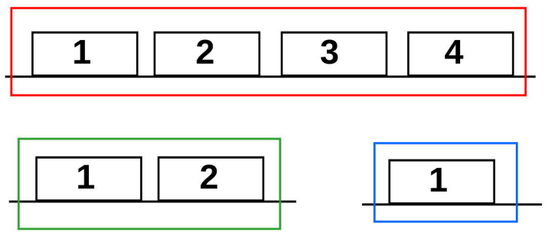
Gene family 4

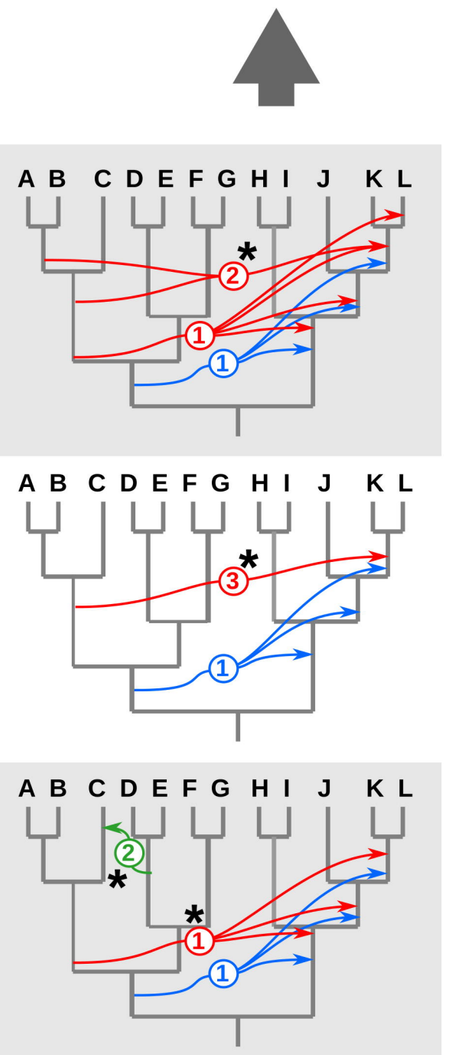


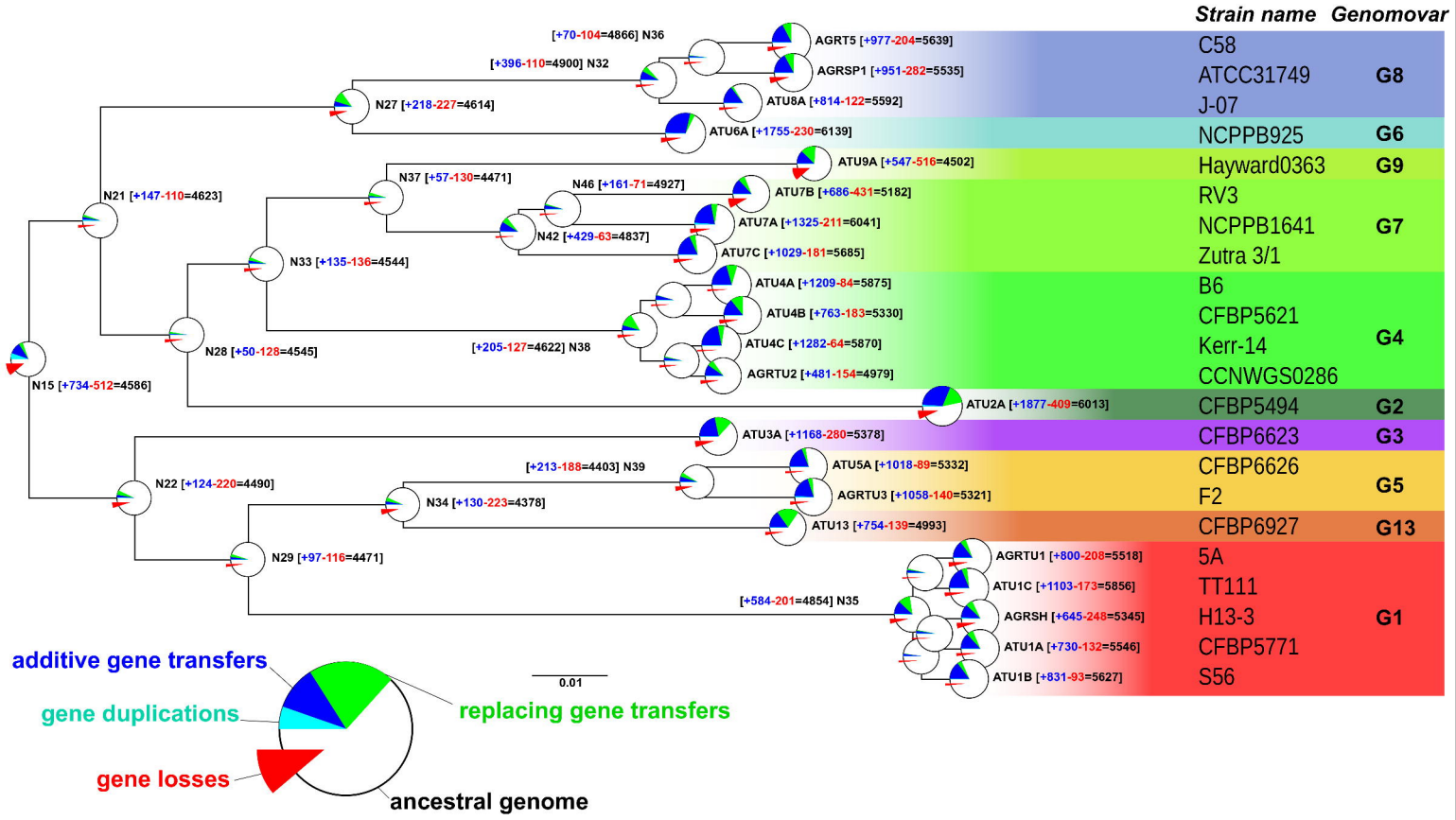
2) Reconstruction of block events in contemporary genomes

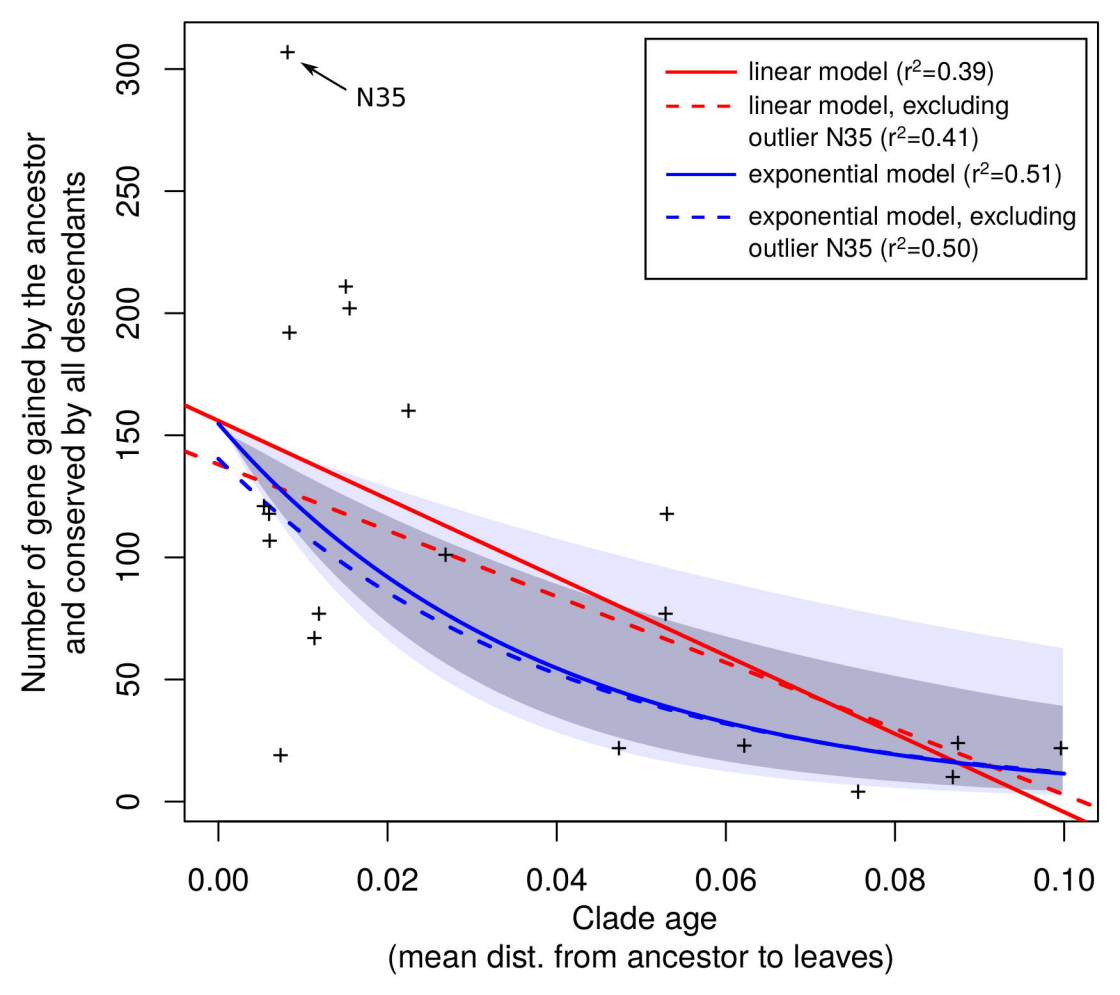
3) Reconstruction of ancestral block events

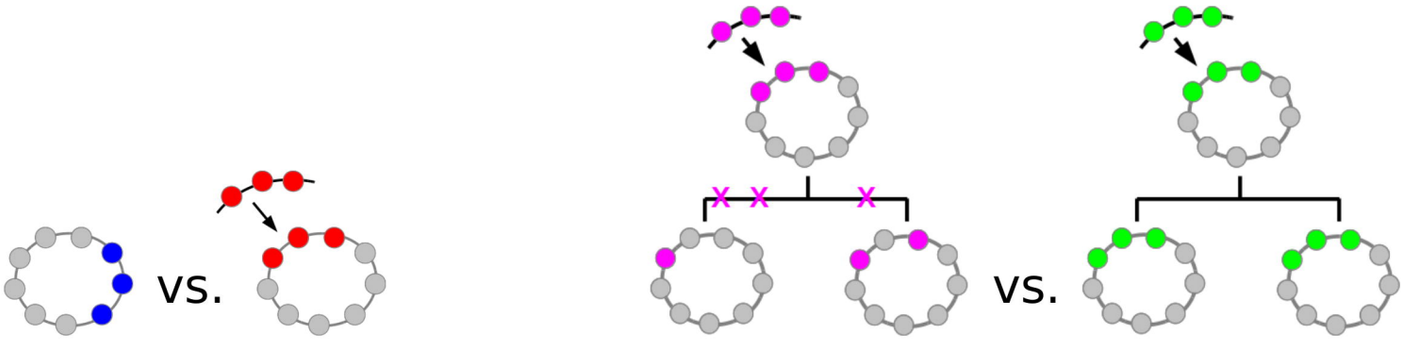







Strain name
Genomovar
[+70-104=4866] N36
AGRT5 [+977-204=5639]
C58
[+396-110=4900] N32
AGRSP1 [+951-282=5535]
ATCC31749
G8
N27 [+218-227=4614]
ATU8A [+814-122=5592]
J-07
ATU6A [+1755-230=6139]
NCPPB925
G6
ATU9A [+547-516=4502]
Hayward0363
G9
N37 [+57-130=4471]
N46 [+161-71=4927]
ATU7B [+686-431=5182]
RV3
N21 [+147-110=4623]
ATU7A [+1325-211=6041]
NCPPB1641
G7
N42 [+429-63=4837]
ATU7C [+1029-181=5685]
Zutra 3/1
N33 [+135-136=4544]
ATU4A [+1209-84=5875]
B6
ATU4B [+763-183=5330]
CFBP5621
N28 [+50-128=4545]
[+205-127=4622] N38
ATU4C [+1282-64=5870]
Kerr-14
G4
N15 [+734-512=4586]
AGRTU2 [+481-154=4979]
CCNWGS0286
ATU2A [+1877-409=6013]
CFBP5494
G2
ATU3A [+1168-280=5378]
CFBP6623
G3
[+213-188=4403] N39
ATU5A [+1018-89=5332]
CFBP6626
N22 [+124-220=4490]
AGRTU3 [+1058-140=5321]
F2
G5
N34 [+130-223=4378]
ATU13 [+754-139=4993]
CFBP6927
G13
N29 [+97-116=4471]
AGRTU1 [+800-208=5518]
5A
ATU1C [+1103-173=5856]
TT111
[+584-201=4854] N35
AGRSH [+645-248=5345]
H13-3
G1
ATU1A [+730-132=5546]
CFBP5771
ATU1B [+831-93=5627]
S56
0.01
additive gene transfers
gene duplications
replacing gene transfers
gene losses
ancestral genome


linear model (r^2=0.39)
linear model, excluding outlier N35 (r^2=0.41)
exponential model (r^2=0.51)
exponential model, excluding outlier N35 (r^2=0.50)
N35
Number of gene gained by the ancestor and conserved by all descendants
0
50
100
150
200
250
300
0.00
0.02
0.04
0.06
0.08
0.10
Clade age
(mean dist. from ancestor to leaves)



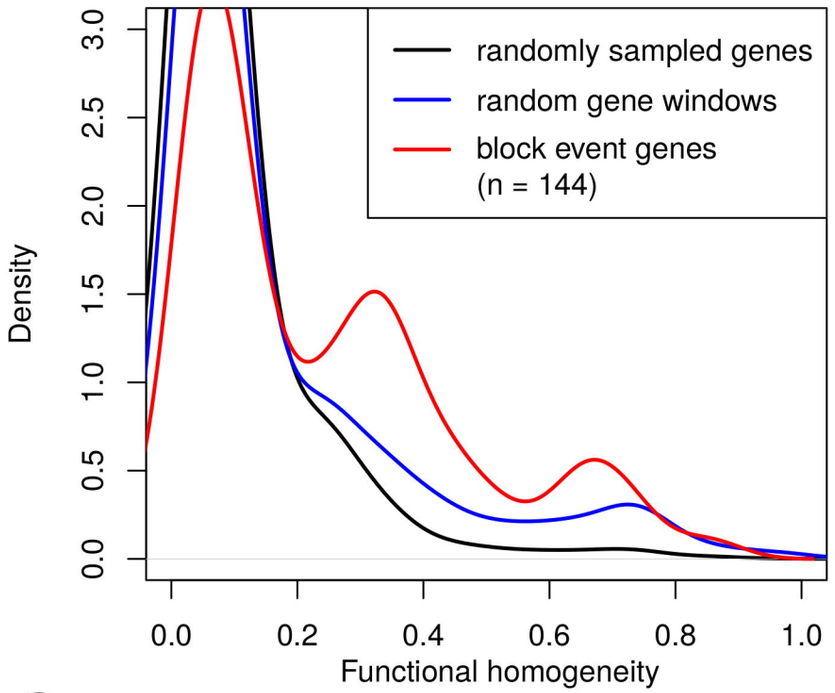
Transferred vs. random gene blocks

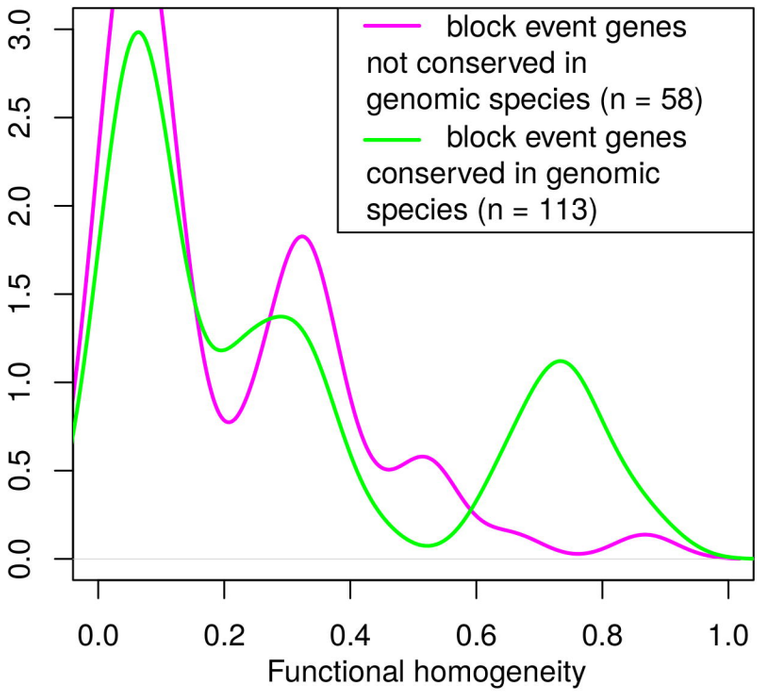
Conserved vs. not conserved transferred gene blocks

A

B

genome of G4-B6, blocks of size 2

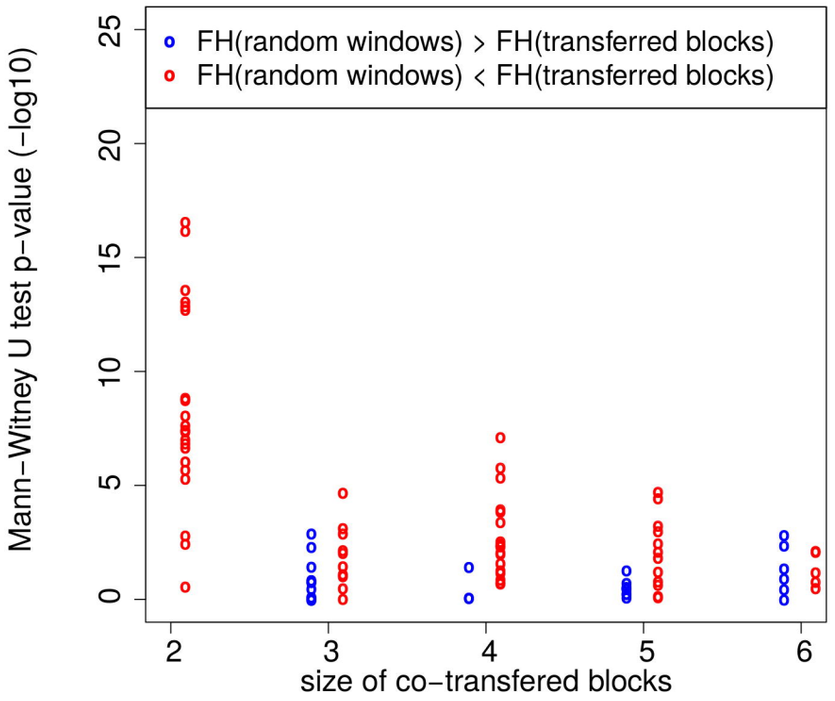


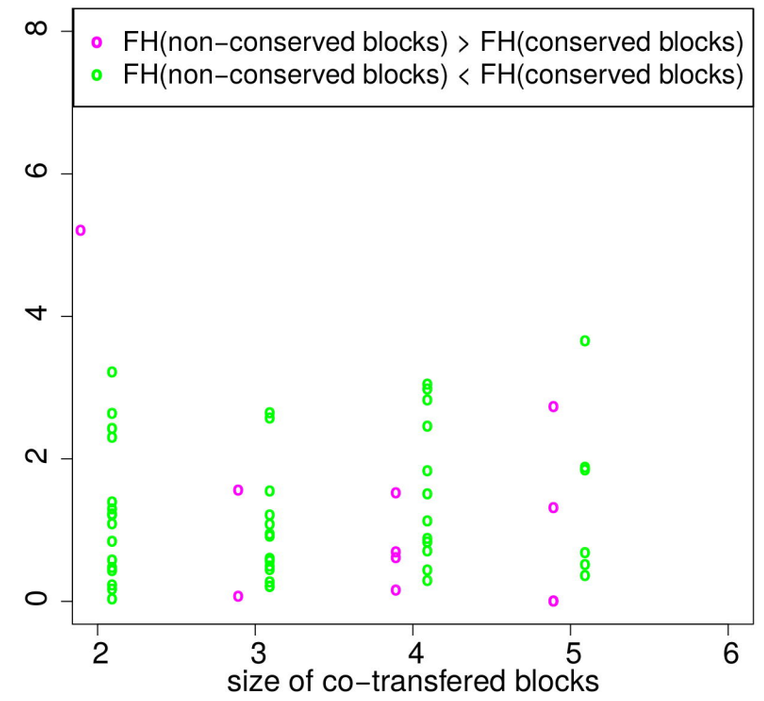


C

D

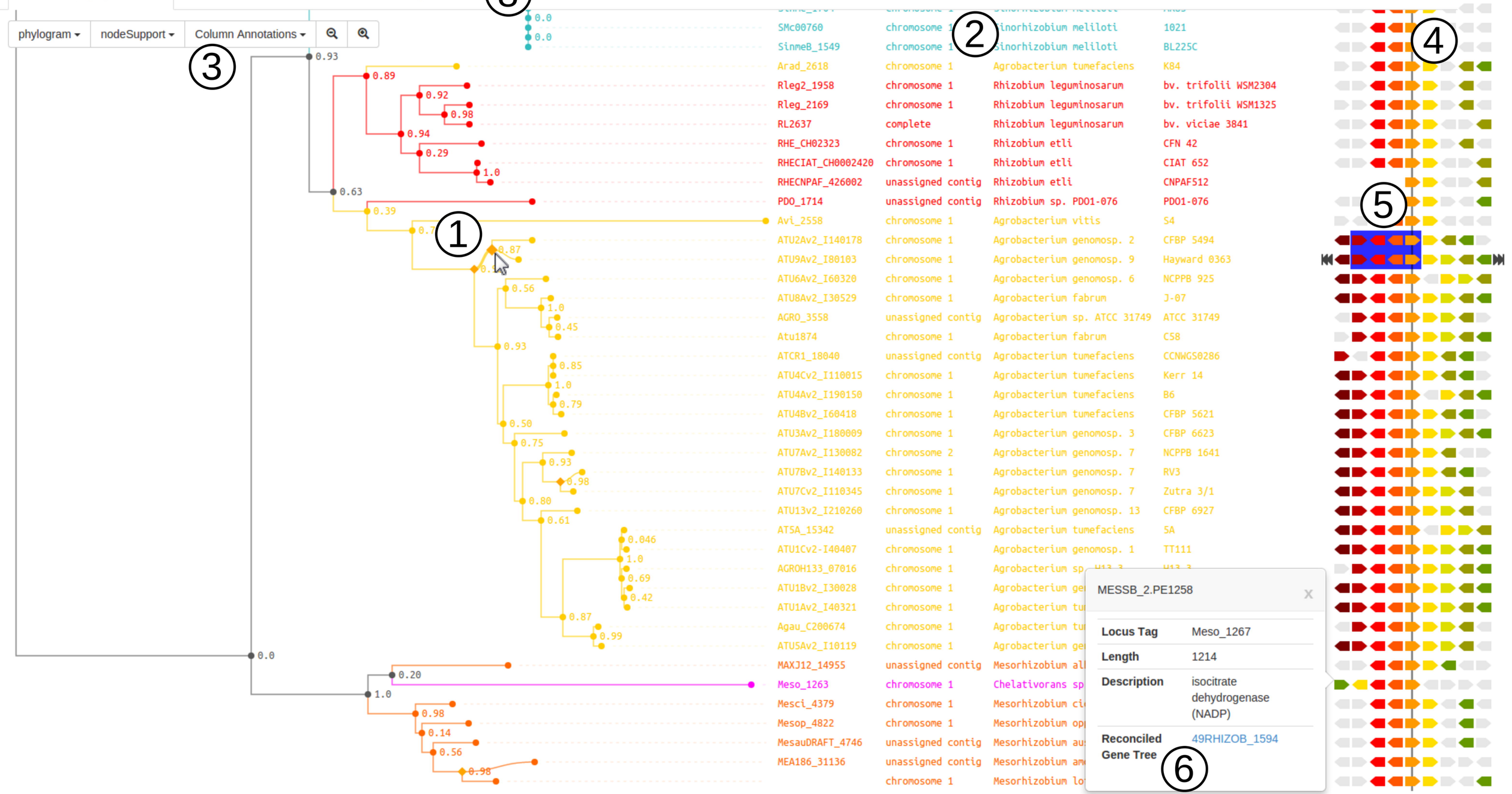
All *At* genomes, blocks of size 2-6





reconciled phylogenetic tree | Taxonomic Tree | Taxonomic Rules | Species Tree | 8

phylogram | nodeSupport | Column Annotations | 3



MESSB_2.PE1258 X

Locus Tag	Meso_1267
Length	1214
Description	isocitrate dehydrogenase (NADP)
Reconciled Gene Tree	49RHIZOB_1594

6