Bacterial genomes, a tale of gene transfer, recombination and cladogenesis

Florent Lassalle

N° 234-2013 Année 2013

Thèse de l'université de Lyon

Présentée

devant L'UNIVERSITÉ CLAUDE BERNARD LYON 1

pour l'obtention

du Diplôme de doctorat
(arrêté du 7 août 2006)

soutenue publiquement le
26 novembre 2013

par

Florent Lassalle

Les génomes bactériens, une histoire de transfert de gènes, de recombinaison et de cladogénèse.

Directeurs de thèse : Xavier Nesme
Vincent Daubin

Jury :		
	Céline Brochier-Armanet	Examinateur
	Vincent Daubin	Directeur de thèse
	Xavier Nesme	Directeur de thèse
	Cécile Neuvéglise	Rapporteur
	Eduardo P. C. Rocha	Examinateur
	J. Peter W. Young	Rapporteur

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. François-Noël GILLY**
Vice-président du Conseil d'Administration	M. le Professeur Hamda BEN HADID
Vice-président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur Philippe LALLE
Vice-président du Conseil Scientifique	M. le Professeur Germain GILLET
Directeur Général des Services	M. Alain HELLEU

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur J. ETIENNE
Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux	Directeur : Mme la Professeure C. BURILLON
Faculté d'Odontologie	Directeur : M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : Mme la Professeure C. VINCIGUERRA
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. le Professeur Y. MATILLON
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : M. le Professeur P. FARGE

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. le Professeur F. DE MARCHI
Département Biologie	Directeur : M. le Professeur F. FLEURY
Département Chimie Biochimie	Directeur : Mme le Professeur H. PARROT
Département GEP	Directeur : M. N. SIAUVE
Département Informatique	Directeur : M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur : M. le Professeur A. GOLDMAN
Département Mécanique	Directeur : M. le Professeur H. BEN HADID
Département Physique	Directeur : Mme S. FLECK
Département Sciences de la Terre	Directeur : Mme la Professeure I. DANIEL
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. C. COLLIGNON
Observatoire des Sciences de l'Univers de Lyon	Directeur : M. B. GUIDERDONI
Polytech Lyon	Directeur : M. P. FOURNIER
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. C. VITON
Institut Universitaire de Formation des Maîtres	Directeur : M. A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Administrateur provisoire : M. N. LEBOISNE

Résumé

Dans les génomes bactériens, les fréquents transferts horizontaux de gènes (HGT) introduisent des innovations génomiques qui peuvent entraîner la diversification des populations bactériennes. À l'inverse, la recombinaison homologue (RH) au sein des populations homogénéise leurs génotypes, et ainsi renforce leur cohésion. Ces processus d'échange génétique, et la fréquence à laquelle ils interviennent au sein et entre les populations, doivent avoir un grand impact sur la cladogénèse bactérienne. Au-delà de la configuration des échanges qui se sont réellement produits entre les bactéries, les traces de RH et de HGT que nous observons dans leurs génomes reflètent les événements qui ont été fixés tout au long de leur histoire. Ce processus de fixation peut être biaisé en ce qui concerne la nature des gènes ou allèles qui ont été introduits. La sélection naturelle peut notamment conduire à la fixation des gènes transférés qui apportent de nouvelles adaptations écologiques. En outre, des biais mécaniques dans le processus de recombinaison lui-même peuvent conduire à la fixation d'allèles non-adaptatifs. Nous avons cherché à caractériser certains de ces processus adaptatifs et non-adaptatifs qui façonnent les génomes bactériens. À cette fin, plusieurs aspects de l'évolution des génomes, comme les variations de leurs répertoires de gènes, de leur architecture et de leur composition en nucléotides ont été examinés à la lumière de leur histoire de transfert et de recombinaison.

Les séquences génomiques des bactéries encodent la totalité de leurs adaptations écologiques et constituent en même temps l'enregistrement de l'histoire de leur diversification. Nous avons utilisé le complexe d'espèces *Agrobacterium tumefaciens* (*At*), un groupe divers de bactéries associées aux plantes comme modèle pour rechercher dans leurs génomes des signatures d'adaptations écologiques associées à leur histoire de diversification. Nous avons exploré la diversité du répertoire de gènes du taxon, d'abord en utilisant la technique d'hybridation génomique comparative (CGH), puis en utilisant les séquences génomiques pour reconstruire l'histoire évolutive des gènes dans les génomes. Pour ce faire, nous avons conçu une nouvelle approche phylogénétique pour la reconstruction de génomes ancestraux tenant compte des événements de transfert horizontal et de duplication des gènes. Cette approche identifie les groupes de gènes co-transférés ou co-dupliqués. L'utilisation de ce signal régional dans les génomes améliore la confiance et la précision de la datation des événements inférés. Les informations sur l'histoire reconstruite des génomes – y compris les arbres de gènes, les événements de transfert et de duplication, les blocs de gènes ayant co-évolué, les synténies, les annotations fonctionnelles ... – sont compilées dans une base de données intégrative, Agrogenom, et peuvent être visualisées et interrogées à travers une interface Web interactive.

A partir des profils CGH et des génomes ancestraux reconstruits, nous avons identifié des gènes spécifiques des principaux clades au sein d'*At*. La plupart d'entre eux sont organisés en grands blocs de gènes ayant co-évolué et qui codent des voies métaboliques cohérentes. Cette organisation constitue une déviation par rapport à un modèle neutre de transfert, indiquant que ces gènes sont sous sélection purificatrice. Les gènes spécifiques de chaque espèce génomique et du complexe d'espèces *At* dans son ensemble codent de façon récurrente des fonctions liées à la production de métabolites secondaires sécrétés ou de matrice extra-cellulaire, ainsi que des fonctions référant au métabolisme de composés d'origine végétale tels que les composés phénoliques et les acides aminés. Ces gènes

spécifiques de clade définissent ainsi des micro-niches spécifiques au sein d'un même macro-environnement où l'interaction avec une plante hôte est primordiale. Ceci suggère que la différenciation écologique des clades d'*At* a eu lieu à travers le partitionnement des ressources écologiques disponibles dans la rhizosphère des plantes.

D'après les histoires reconstituées des gènes, nous avons montré l'intensité des transferts de gènes au sein des espèces génomiques d'*At*, mais aussi l'occurrence de transferts entre espèces génomiques et clades plus anciens. Certains gènes définissant les niche écologiques ont parfois été partagés par des clades éloignés, ce qui a pu induire leur compétition pour des ressources partagées. Cependant, chaque clade est caractérisé par une combinaison spécifique de gènes prédisant des traits écologiques uniques, et dont l'expression semble être coordonnée par des systèmes de régulation impliquant la perception des signaux environnementaux spécifiques. Ce faisant, des clades divergents sont susceptibles de maintenir des niches écologiques différenciées, leur permettant de cohabiter dans l'habitat rhizosphérique.

Une autre caractéristique des organismes cellulaires est le contenu en nucléotides G et C de leur génome, qui est connu pour varier considérablement entre génomes et en leur sein. Chez les mammifères, et probablement plus généralement chez les Eucaryotes, ces différences intra-génomiques du contenu en bases G et C (GC%) se révèlent être fortement influencées par les variations régionales dans les taux de recombinaison à long terme. Ceci est causé par un phénomène appelé conversion génique biaisée vers G/C (gBGC), qui favorise la fixation de mutations vers G ou C dans les régions de forte recombinaison. En revanche, au sein des génomes bactériens, l'hétérogénéité en GC% entre gènes est traditionnellement considérée comme une preuve de la multiplicité de leurs origines, en raison de transferts horizontaux fréquents, mais le mécanisme qui sous-tend la composition biaisée des gènes transférés est encore inconnu. Des études récentes ont suggéré qu'une mystérieuse force sélective favorisant un GC% plus élevé existe chez les bactéries, mais la possibilité qu'il pourrait s'agir de la gBGC a été exclue.

Nous avons montré que la gBGC est probablement à l'œuvre dans la plupart, sinon toutes les espèces bactériennes. D'abord, nous trouvons une relation positive entre le GC% d'un gène et le signe d'événements de recombinaison intra-géniques, et ce dans un ensemble divers de clades bactériens. Deuxièmement, nous montrons que la force évolutive responsable de cette tendance entre en conflit avec la sélection pour l'usage de codons synonymes optimaux, en particulier pour ceux se terminant par A ou U. Nous proposons que la gBGC, précédemment considérée comme spécifique aux eucaryotes sexués, existe également chez les bactéries et pourrait donc être une caractéristique ancestrale des organismes cellulaires. Nous discutons de rôle possible de la gBGC comme cause de nombreuses observations de la génomique bactérienne jusqu'alors inexpliquées, comme le non-équilibre apparent des patrons de substitution de nucléotides, l'hétérogénéité de la composition des gènes dans les génomes, et la corrélation générale entre la taille du génome et le GC%.

Nous avons montré que le transfert de gènes et la recombinaison homologue peuvent contribuer à la diffusion et à l'entretien des adaptations, mais qu'en même temps le transfert de gène peut induire des coûts adaptatifs à travers la concurrence pour les ressources écologiques communes, et que la recombinaison homologue peut interférer avec la sélection via la fixation neutre d'allèles riche en G/C. Cela révèle l'existence de compromis complexes

dans l'évolution des génomes bactériens, qui ne peuvent être mieux compris qu'à la lumière de la reconstruction complète de leur histoire.

Mots-clés : Reconciliations, genome ancestral, transfert de gène, cladogénèse bactérienne, ecologie inverse, *Agrobacterium tumefaciens*, recombinaison homologue, contenu en GC.

Abstract

Bacterial genomes, a tale of gene transfer, recombination and cladogenesis.

In bacterial genomes, the frequent horizontal gene transfers (HGT) introduce genomic novelties that can promote the diversification of bacterial populations. In opposition, homologous recombination (HR) within populations homogenizes their genotypes, enforcing their cohesion. These processes of genetic exchange, and their patterns of occurrence among and within lineages, must have a great impact on bacterial cladogenesis. Beyond the pattern of exchanges actually occurring between bacteria, the traces of HR and HGT we observe in their genomes reflect what events were fixed throughout their history. This fixation process can be biased regarding the nature of genes or alleles that were introduced. Notably, natural selection can drive the fixation of transferred genes that bring new ecological adaptations. In addition, some mechanical biases in the recombination process itself may lead to the fixation of non-adaptive alleles. We aimed to characterize such adaptive and non-adaptive processes that are shaping bacterial genomes. To this end, several aspects of genome evolution, such as variations of their gene repertoires, of their architecture and of their nucleotide composition were examined in the light of their history of transfer and recombination.

Genomic sequences of bacteria code all of their ecological adaptations and at the same time are the record of their diversification history. We used the *Agrobacterium tumefaciens* (*At*) species complex, a diverse group of plant-associated bacteria as a model to search for genomic signatures of ecological adaptation in relation to diversification. We explored the gene repertoire diversity of the taxon, first using micro-array comparative genome hybridization (CGH), and then using genome sequences to reconstruct the evolutionary history of genes in genomes. For this purpose, we designed a new phylogenetic approach for reconstruction of ancestral genomes, accounting for events of horizontal transfer and duplication of genes. This approach identifies groups of co-transferred/duplicated genes. Using this regional signal in genomes provides better confidence and accuracy when dating the events. Informations on reconstructed genome history – including gene trees, transfer and duplication events, blocks of co-evolved genes, syntenies, functional annotations. . . – are compiled in an integrative database, Agrogenom, which can be visualized and queried through an interactive web interface.

From these CGH profiles and reconstructed ancestral genomes, we identified genes specific to major clades within *At*. Most of these were organized in large blocks of co-evolved genes that encoded coherent pathways. This organization constitute a deviation relative to a neutral model of transfer, indicating these genes are under purifying selection. Genes specific to each genomic 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 and amino-acids. These clade-specific genes likely constitute parallel adaptations to life in interaction with host plants. This suggest that ecological differentiation of clades occurred through partitioning of the ecological resources available in plant rhizospheres.

From the reconstructed histories of genes, we showed that intensive mixing occurred within genomic species of *At*, but also that much genetic exchanges occurred between genomic species and higher clades. This sometimes caused putative niche-specifying genes

to be shared by distant clades, potentially inducing competition for shared resources. However, each clade could be characterized by a specific combination of genes predicting unique ecological traits, and whose expression seems to be coordinated by regulation systems involving perception of specific environmental signals. Thereby, diverged clades are likely maintaining differentiated ecological niches enabling their cohabitation in the rhizospheric habitat.

Another characteristic feature of cellular organisms is the GC-content of their genome, which is known to vary widely both among and within genomes. In mammals and probably more generally in Eukaryotes, these intra-genomic differences of GC-content were shown to be strongly influenced by regional variations in long-term recombination rates. This is caused by a phenomenon called GC-biased gene conversion (gBGC), which favours the fixation of G/C mutations in recombining regions. In contrast, in bacterial genomes, the heterogeneity of GC-content among genes is traditionally seen as evidence for their multiplicity of origins, due to frequent HGT, but the mechanism underlying the biased composition of transferred genes is still unknown. Recent studies have suggested that a mysterious selective force favouring higher GC-content exists in Bacteria but the possibility that it could be gBGC has been excluded.

We have shown that gBGC is probably at work in most if not all bacterial species. First we find a consistent positive relationship between the GC-content of a gene and evidence of intra-genic recombination events in a large spectrum of diverse bacterial clades. Second, we show that the evolutionary force responsible for this pattern conflicts with selection for optimal synonymous codons, specifically for AU-ending codons. We propose that gBGC, first thought to be specific to sexual Eukaryotes, exists in Bacteria and could therefore be an ancestral feature of cellular organisms. We discuss the potential of gBGC to account for many previously unexplained observations of bacterial genomics, such as the apparent non-equilibrium of base substitution patterns, the heterogeneity of gene composition within genomes, and the general correlation between genome size and GC-content.

Altogether, we showed that HGT and homologous recombination can contribute to the diffusion and maintenance of adaptations, but at the same time HGT can induce fitness costs through competition for shared ecological resources, and homologous recombination can interfere with selection through neutral fixation of GC-rich alleles. This reveals the existence of complex trade-offs in the evolution of bacterial genomes, which might be better understood under the light of more comprehensive reconstruction of their history.

Keywords: Reconciliations, ancestral genome, gene transfer, bacterial cladogenesis, reverse ecology, *Agrobacterium tumefaciens*, homologous recombinaison, GC-content.

Remerciements

Je veux tout d'abord remercier mes directeurs de thèse pour m'avoir donné ce sujet à traiter et ainsi m'avoir détourné des pratiques de laboratoire telles que la biologie moléculaire, où semble-t-il je ne faisais pas un usage coordonné de mes mains. Grâce à eux, je suis donc devenu un bioinformaticien dédié à l'étude de l'évolution microbienne ; je m'en réjouis car c'est un sujet de conversation qui fait paraître très cultivé en société, mais surtout avec lequel je me suis éclaté et je sais maintenant que cela me plairait bien de me torturer à propos de la vie cachée des petites bêtes pendant encore quelques décennies.

Bien sûr, les côtés plaisants de la thèse ne tiennent pas qu'au contenu scientifique. Je dois aussi à Xavier et Vincent de très bon moments, dont certains parmi tant d'autres me reviennent : à Aussois où Xavier m'a gratifié à l'explication du système de « couches » pour la stabilisation du génépi, ces débats sans fin avec Vincent pour savoir si, oui ou non, Nicolas Cage est un bon acteur, le tour en train de la mine dans les grottes Slovènes en compagnie de Xavier, et ces fois où, après mes errances en festivals métalleux, il me fallait rendre compte à Vincent de comment ce vieux Ozzy tenait-il debout avant qu'il soit question de finir cette revue en retard... Et puis bien sûr ces moments un peu surréalistes de rédaction côte à côte que tout doctorant a du connaître, où le cerveau n'a pas de repos jusqu'à trouver la phrase adéquate et « l'écrire pour pouvoir l'enlever après ».

J'ai partagé ma thèse entre deux laboratoires, le LEM et le LBBE, tous deux situés dans l'auguste bâtiment Mendel de la Doua, ce qui m'a été très profitable. Au delà de l'apprentissage d'une approche mixte de la microbiologie dont je fait l'apologie dans ce document, j'ai recu un double ration de tout ce que la recherche française a à offrir, comme par exemple l'accès aux événements gastronomiques gratuits tels que les repas de Noël ou pots de thèse. Aussi, durant ces cinq années, j'ai fréquenté dans chaque laboratoire de nombreuse personnalités très variées et franchement originales, et j'espère avoir créé un pont supplémentaire entre les « gens du 4ème » (non, pas du 5ème) et ceux « du 2ème » étage (non, pas du premier), tous gagnent à être connus.

Merci d'abord à l'équipe 4 du labo d'Écologie Microbienne (ou « labo des comiques » selon ceux du 2ème) pour son esprit de famille : David, pour m'y a avoir introduit et tout m'avoir montré des pratiques de la recherche ; les grands frères de thèse, Tony et Malek, pour avoir eu avant moi les révélations extralucides sur *Agrobacterium* et m'avoir pavé la voie ; Daniel et Maxime, c'était sympa d'avoir des copains geeks en haut – et ne vous inquiétez pas, Yvan ne saura pas que vous faisiez des infidélités à l'équipe 3 ; Jordan et Seb, la même, mais vous c'est différent, on vous a forcé ; Ludo pour ton oursitude bienveillante qui nous sauvera tous. Et comme la famille s'étend à toute l'UMR, je ne vais pas citer tout le monde, mais je peux dire aux jeunes des soirées à la Feyssine jusqu'au HellFest, aux plus vieux qui ont connu cette fameuse classe de Master en 2008, et puis aux vétérans d'Aussois, ceux qui s'y sont cassé trois membres,

ceux qui ont pu y avoir quatre plats de charcuterie pour la pierrade, ceux qui y ont gagné le Reblochon d'Or avec encore 4g de génépi : on s'est bien marré, et je n'oserai surtout pas tout raconter ici, mais je m 'en souviendrai toujours.

Et je ne mentirai pas, j'avais quand même élu un domicile préférentiel en Biométrie (les geeks, selon ceux du 4ème), au début pour un confort accru (un ordinateur gratuit !), mais j'y ai rapidement pris goût. Il faut dire que tout le monde m'a gentiement adopté. D'abord, on a eu la patience de tout me montrer la bioinfo, alors que je n'y connaissais rien, j'ai apris selon le style des plus grands : merci Laurent G. pour m'avoir donné à retaper dès le début tes scripts mixtes Python/R avec des regexp de 5 lignes, ca m'a fait les pieds mais j'y suis parvenu grâce a tes patientes explications ; idem, merci Simon pour les scripts csh/ACNUC, maintenant je peux parler la langue des Anciens, et puis aussi pour toute la contre/sous-culture improbable qu'on a pu partager ce faisant ; Marc et Nico, pour les bases de sectarisme TeX au service de l'édition du plus beau de tous les manuels de TD ; Thomas pour avoir partagé mes errances algorithmiques ; et puis enfin Rémi pour avoir été mon partenaire dans la construction d'Agrogenom, tu as réussi à transformer mon foutoir originel en un truc qui brille de toutes les couleurs et qui bouge quand on clique dessus, c'est un miracle. Et je n'oublie pas les informaticiens, l'Agence tout risque du labo grâce à qui l'AERES ne trouvera jamais rien de plus à se mettre sous la dent que des pommes pourries pendues au plafond, tout le reste tourne au poil. Et enfin un gros merci collectif à la fine équipe de la pause café, pour toujours avoir pris le temps d'en refaire couler un, et en attendant le café, de se raconter des bêtises plus grosses que nous malgré la bien maigre isolation phonique par rapport au couloir plein de chefs ; on y a bien rigolé, et aussi j'ai trouvé en vous des camarades complices et solidaires de la détresse des autres, c'est franchement ce qui me faisait venir au labo malgré le découragement de certains jours.

Je terminerai sur ceux qui m'ont soutenu et rendu belle la vie hors des labos. D'abord, Manu, Petrov et Antoine avec qui j'ai traîné mes guêtres depuis le lycée Descartes, il y a près de dix ans, jusqu'en Rhône-Alpes et dans pas mal de coins : on en a vu, on en verra !

Un très grand merci à mes parents qui m'ont tout donné pour que j'arrive jusque là, à commencer par l'amour des plantes, de la nature, et ainsi de la biologie. Même avec un doctorat, je dépends encore de vous pour me nommer les petites fleurs !

Et enfin, je te remercie Pauline pour avoir été avec moi tout ce temps là, pour avoir partagé une chouette vie a deux durant ces années dans nos différentes maisons et cabanes, et même quand c'était de loin. Il faut quand même avouer que c'est t'avoir rencontrée la veille qui m'a fait prendre la décision de venir faire la thèse à Lyon ; je ne l'ai pas regretté.

Je dédie ce travail aux professeurs de Biologie que j'ai pu avoir en cours tout au long de mon cursus et qui ont su me donner la passion des Sciences de la Vie : Mr Chardon au collège Choiseul d'Amboise, Mme Merzizen, Mme Buttner et Mme Charrier au lycée L. de Vinci d'Amboise et Mr Latour et Mr Moniaux au lycée Descartes de Tours.

Table des matières

1 Introduction

In the quest for understanding the mechanisms responsible for the huge diversity of bacterial genomes, it is tempting to systematically invoke to the adaptation of bacterial lineages to their environment.

However, one must consider that the extant diversity is the result of a complex evolutionary process combining neutral processes, historical contingency and adaptation.

Thus, to gauge the role of adaptation in the diversification of bacterial genomes, it is crucial to account for both the neutral processes that generate diversity and the sequence of evolutionary events that shaped the genomes.

In the present work, we will explore bacterial genomes in a historical context to decipher the relative part of adaptive versus non-adaptive evolution resulting from cladogenesis, horizontal gene transfer and homologous recombination.

1.1 Bibliographical review

The following text including sections 1.1.1 to 1.1.4 are adapted from a review chapter titled "Evolution of Prokaryotic Pangenomes" that we published in 2012 in a book under the editorial direction of M. Pilar Francino (Francino MP. 2012. *Horizontal Gene Transfer in Microorganisms*. Horizon Scientific Press).

Since Woese and Fox (1977) and the discovery of the two ancient 'kingdoms' of Bacteria and Archaea, molecular biologists have been constantly revising their perception of the abundance and variety of microbial life, and the processes driving their evolution. In recent years, comparison of genome sequences from closely related species and strains of the same species have revealed a yet unforeseen diversity of gene repertoires, much larger than could have been predicted by their morphologies or apparent biochemical capabilities. A puzzling observation is that much of the variation of gene content, and hence of functional repertoires, is specific of strains, owing to frequent horizontal transfer. This questions the genomic coherence and ecological significance of named taxa. However, it is not clear what fraction of genomes are under selective pressure and are actually important in defining the ecology of organisms on the long term.

To tackle the role of selection in shaping the gene content of genomes, one must first understand the processes generating their diversity. Cladogenesis in prokaryotes is intimately associated to their history of horizontal gene transfer and homologous recombination. We will present here a review of the current knowledge on the mechanisms impacting the evolution of gene repertoires in prokaryotes, and present theoretical models that integrate their role during cladogenesis.

1.1.1 Gene repertoires and the organization of prokaryotic genomes

Prokaryotic genomes share global features that make them readily distinguishable from most eukaryotic ones. Eukaryotes are thought to accumulate DNA in their genomes by either duplication, invasion of transposable elements or insertion of endoviruses and usually harbour a high proportion of non-coding DNA. As a result, the coding capacity of a eukaryotic genome can be as low as 5%, for instance for the human genome. Prokaryotic genomes, on the other hand, are rather dense in genes, with usually around 80% of a genome coding for proteins. This compaction goes along with an optimized organization of the genetic information reflecting the action of cellular processes such as genome replication and gene expression. For instance, the asymmetric nature of DNA replication, with a lagging strand and a leading strand, impacts bacterial genome organization (Rocha and Danchin, 2003) as well as compositional and mutational patterns (Lobry and Sueoka, 2002). Similarly, the polarity of replication, from origin (*ori* locus) to terminus (*ter*), also leaves marks on circular bacterial chromosomes, as the composition in A+T nucleotide and evolutionary rates of genes near the terminus are generally higher (Daubin and Perrière, 2003). In the case of fast-growing bacteria, chromosomes

replicate faster than the cell divides, which generates a strong selective pressure for relocating genes requiring high expression rates within the region of the origin of replication (Couturier and Rocha, 2006). Although the replication process in Archaea may differ in many points from the one observed in Bacteria, they share some of their characteristics, notably the strand composition asymmetry (Myllykallio et al., 2000). Bacterial chromosomes also seem to be divided into regions differing in their average gene expression level. Functionally related genes are often co-transcribed in operons and sometimes further clustered in superoperons that can be widely conserved among prokaryotes, like ribosomal superoperons or pilus biosynthesis loci (Lathe et al., 2000). Finally, chromosomal macro-domains, defined following the ori-ter axial symmetry, form units of megabase size that seem to be globally regulated through topological constraints (Esnault et al., 2007).

As these features are conserved among the majority of prokaryotes, the constraints imposed by the replication and expression processes on the structure of bacterial chromosomes must be very strict. Indeed, disruption of the macro-domain symmetry appears to be highly detrimental in *E. coli* (Esnault et al., 2007). The existence of strong selection for chromosomal organization in the whole bacterial kingdom is confirmed by the good conservation of general gene order between pairs of chromosomes of even moderately related strains (Rocha, 2006).

Despite these common organization principles, prokaryotic genomes are extremely diverse. Massive genome sequencing revealed an unexpected diversity of gene sequences. At the evolutionary scale of the whole prokaryotic world, hundreds of thousands of different genes have been inventoried in genomic databases (Penel et al., 2009; Muller et al., 2010). This number is immense, especially in regard to the estimation that fewer than 50 genes are conserved among all prokaryotes, most of which are involved in translation (Charlebois and Doolittle, 2004). All other processes, even those that are deemed essential, such as DNA replication and central metabolism, are apparently not homologous among all prokaryotes. This suggests that, through evolution, most of the ancestral genetic information has been replaced, at least in some lineages.

Although this pattern of loss and gain of genes may not be surprising at the scale of several billion years, the same applies at every phylogenetic depth, from phylum to species (Daubin and Ochman, 2004b; Lerat et al., 2005; van Passel et al., 2008). Recent sequencing of several strains of the same species revealed that the gene content of strains within a species differs markedly while the overall genomic organization is conserved. The comparison of 20 strains of *E. coli* showed that the species' core genome – those genes common to all strains – consists of only ~2000 genes, less than half the average genome size of a strain (Touchon et al., 2009). The remaining genes in each genome are shared by only a fraction of all strains, with the majority being unique to only one. This demonstrates that species are much more diverse than previously suggested by molecular phylogenies based on core genes, and raises the question of what makes the unity of species. The very notion of species is challenged by these facts and a definition for prokaryotic species is still sought and intensely debated (Achtman and Wagner, 2008). The extent of gene diversity within a species can simply be measured by looking at

the size of its pangenome, which corresponds to all genes represented in at least one of the genomes of the considered species. For instance, the pangenome of *E. coli* is approximately 18,000 genes when considering currently sequenced genomes (Touchon et al., 2009). Currently available pangenome data suggest that many species, such as *Streptococcus agalactiae* (Tettelin et al., 2005), *Vibrio cholerae* (Vesth et al., 2010) or *Burkholderia pseudomallei* (Ussery et al., 2009), have an 'open' pangenome where sampling of new genomes invariably reveals the existence of new genes. In contrast, other species seem to have limited gene repertoire diversity, such as *Bacillus anthracis*, for which no new gene has been found after sampling of four strains, indicating a 'closed' pangenome. Similar results were obtained in another recent study using data from almost one hundred genomes of the two sister species *Campylobacter jujeni* and *C. coli*, which established their common pangenome as 'closed' at 3000 genes (Lefébure et al., 2010). A closed pangenome usually means a recent speciation event with no genome diversification, but in the case of *Campylobacter*, it is unsure that the sequenced epidemic strains can be considered a good representative set of the existing species diversity (Lefébure et al., 2010). The population pyramid of genes within a species is rather like a column, with pedestal and capital, i.e. peaks of frequency for genes restrained to a small set of genomes at the basis and for core genes at the top Figure 1.1 (Rocha, 2008; Touchon et al., 2009; Lefébure et al., 2010). This suggests that within a species, genomes are made of a relatively stable set of genes present in almost every genome (the 'stabilome') (Vesth et al., 2010), and of a variable set of genes, which may even be strain specific.

The diversity of bacterial pangenomes adds up to the immensity of the protein universe. In fact, the large set of homologous protein families obtained by merging all prokaryotic pangenomes displays a U-shaped distribution of frequencies similar to that obtained for genes among genomes (Koonin and Wolf, 2008; Lapierre and Gogarten, 2009). This shows that much of the protein diversity consists of sporadic proteins occurring in a very restricted range of genomes, a large fraction of which correspond to ORFans, i.e. putative coding sequences which are unique to a genome.

How can bacterial genomes reconcile their optimized, constrained and highly conserved organization with so much variation in their content? Alignments of genomes of strains from the same species reveal they share a backbone of core genes with conserved order, inter-spaced with sporadic genes. Touchon et al. (2009) showed that in the *E. coli* species, most insertion occurred in a small set of discrete loci. This is in line with the assumption that the disruption of the genomic organization is highly detrimental, and only few sites allow neutral insertion of foreign DNA.

1.1.2 The dynamics of the pangenome

It has long been known that in many bacterial species large fractions of the genome have been acquired 'recently' by horizontal gene transfer (HGT) (Médigue et al., 1991). These fractions are recognizable because the genes they contain show a codon usage which differs markedly from that of the rest of the genome. Such anomalous composition is thought to be a

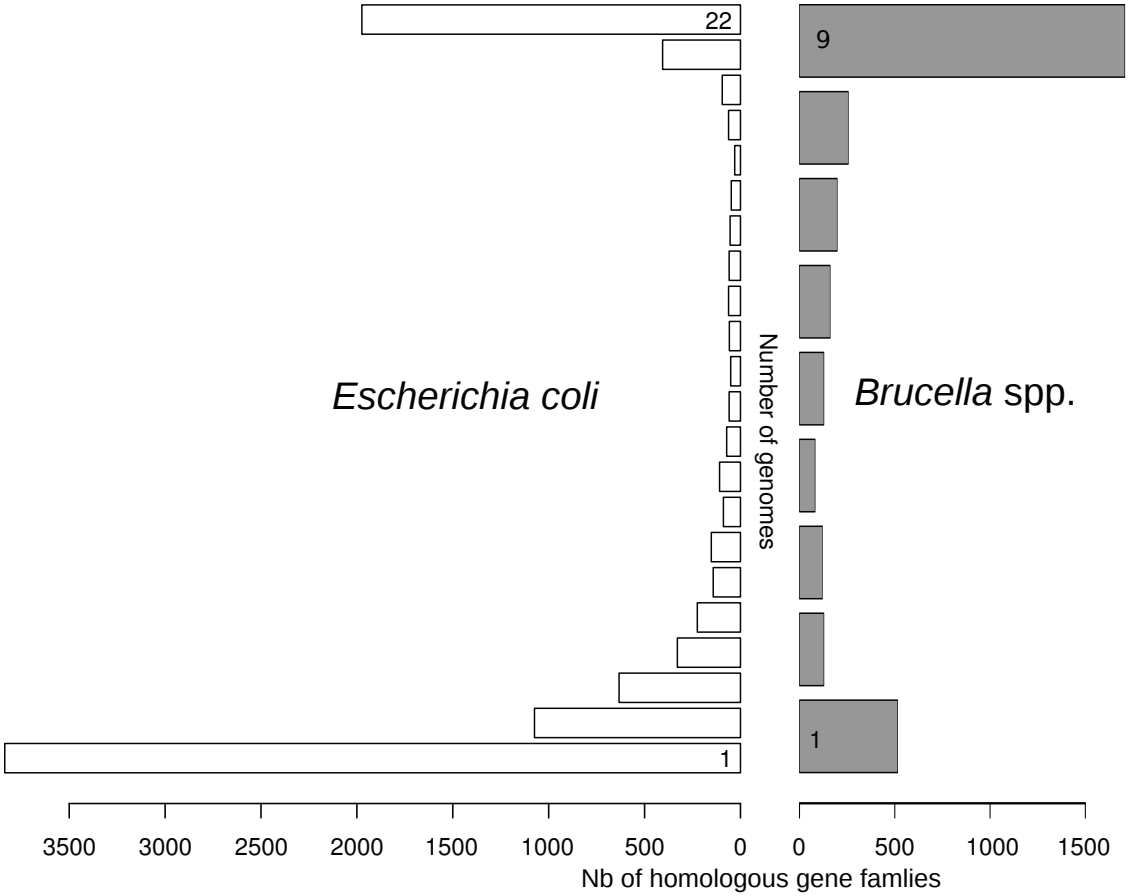


Figure 1.1: **The pangenomes of *Escherichia coli* and *Brucella* spp.**
This plot represents the frequency of genes within 22 sequenced strains of *E. coli* and nine sequenced strains of *Brucella* spp., which includes genomes from five identified species (*B. melitentis*, *B. suis*, *B. abortus*, *B. ovis* and *B. canis*). These two pangenomes have very different dynamics, with a relatively small, closed pan-genome in *Brucella* spp. and a large open one in *E. coli*. Data from Hogenom release 5 (http://pbil.univ-lyon1.fr/databases/hogenom/).

trace of the donor's genome composition, currently in 'amelioration' towards the normal codon use of the new host (Lawrence and Ochman, 1997). The variability of gene repertoires within a species and the elevated number of strain-specific genes suggest high rates of gene gain and loss within a lineage, typically higher than nucleotide substitution (Hao and Golding, 2006). For instance, when comparing two *E. coli* strains, orthologous sequences show less than 5% divergence at the nucleotide level, while gene content can differ by a thousand genes (~1 Mb) corresponding to more than 20% genomic dissimilarity (Touchon et al., 2009). Gene gain and loss thus represent the major source of innovation in prokaryotic genomes, as the modifications of gene repertoires can strongly affect cellular metabolism and morphology. Such changes can efficiently provide prokaryotes with opportunities for adaptation to new ecological niches.

It is yet unclear how much of the observed differences among genomes are adaptive. The insertion of new genes in bacterial genomes is indeed thought to occur at very high frequency, either from duplication of resident genes or via HGT, the latter playing a dominant role (Snel et al., 2002; Mirkin et al., 2003; Kunin et al., 2005; Treangen and Rocha, 2011). These mechanisms are thought to compensate the high deletion rate, which is universal in prokaryotes (Mira et al., 2001; Kuo and Ochman, 2009), and to maintain a homogeneous genome size among strains within a species (Moran and Plague, 2004). Interestingly, obligatory intracellular pathogens or endosymbionts display spectacular reductions of their genomes (Ochman and Moran, 2001), reflecting both the absence of gene input by transfer and the increased effect of drift on slightly deleterious deletions (Ochman and Moran, 2001; Moran and Plague, 2004).

Once they enter the pan-genome, new genes undergo rapid evolution. First, new genes are more prone to be deleted than older ones (Marri et al., 2007; van Passel et al., 2008). In addition, it has been shown that new genes have high evolutionary rates, and that this rate decreases with time spent in the genome (Ochman and Moran, 2001; Daubin and Ochman, 2004a), which has been interpreted by Hao and Golding (2006) as "a mixture of directional selection to adapt in some genes and neutral mutations destroying function in others". Hence, newly acquired genes may not immediately be advantageous for the organism, and therefore be subject to relaxed selection. This can lead to either inactivation followed by rapid deletion or to fixation under a diverged form adapted to the new host environment. Even in free-living bacteria, many predicted open reading frames (ORFs) may be unsuspected decaying genes (Ochman and Davalos, 2006).

1.1.3 The roots of bacterial pangenomes

The consequence of this constant gene turnover is the cohabitation of genes of various origins and ages in a given lineage (Daubin and Ochman, 2004a). Interestingly, within an age class, genes tend to be similar in terms of G+C%, length, and evolutionary rate (Daubin and Ochman, 2004a; Hao and Golding, 2006; Yin and Fischer, 2008). ORFans, those genes that have no known homologues in any sequenced genome, i.e. the youngest genes, are short, fast evolving, and systematically AT-richer than their host genomes. Three major hypothesis are presently discussed for the origin of ORFans: (1) origination de novo from non-coding

sequences; (2) fast evolution of resident genes, to the point of losing any homology signal; (3) transfer from an inaccessible gene pool, consisting probably of mobile DNA elements such as viruses, plasmids, integrons or other integrative elements (IEs).

Origination de novo is not likely to happen in bacterial genomes, as stretches of noncoding DNA are short and mostly under selective constraints for their role in transcriptional regulation. Origin from fast evolution of pre-existing genes – resident or IEs – is difficult to test, as ORFans have no similarity to other sequences in databases. It has been suggested that ORFans have amino-acid compositions reminiscent of translation from random genomic sequences (Yomtovian et al., 2010), but the nucleotide compositional bias inherent to ORFans may explain some of this resemblance. Recent studies show that mutations are universally biased towards AT nucleotides in Bacteria (Hershberg and Petrov, 2010; Hildebrand et al., 2010). These authors concluded that the existence of GC-rich genes in genomes must therefore be explained by selection on GC content, or a selection-like process such as biased gene conversion (Hershberg and Petrov, 2010). Although a selective pressure acting on global genomic GC content is difficult to imagine, it could explain why ORFans, which are under lower purifying selection, have relatively low G+C content.

There are a number of studies that suggest that ORFans come from IEs. IEs are known to be vectors for HGT by transduction/conjugation and frequently carry cellular genes from former hosts, and as such are also good candidates for transmitting new genes. ORFans have been shown to be frequently part of clusters of co-transferred genes that globally show similarities to IEs (Cortez et al., 2009). The viral gene pool is today almost unexplored and first attempts to look into it suggest an incommensurable diversity (Edwards and Rohwer, 2005; Williamson et al., 2008; Angly et al., 2009) that could be an invaluable source of genes for prokaryotic genomes.

1.1.4 Functional role of the pangenome

The question remains of how many of these genes actually contribute to the fitness of their host. Those genes that are sporadically distributed among strains of a bacterial species are expected to have very marginal effects, and indeed genome streamlining experiments demonstrate that up to a fourth of a bacterial genome can be deleted without affecting essential functions. By deleting genes that are poorly conserved among *E. coli* strains (Pósfai et al., 2006), or genes absent in the related enterobacteria *Buchnera* (Mizoguchi et al., 2008), *E. coli* strains can be engineered into fast-growing minimal cell factories (Mizoguchi et al., 2007). Following a different approach, attempts to delete every single gene in genomes were conducted to assess viability and competitiveness of mutant cells (Gerdes et al., 2003). In *E. coli* and *Bacillus subtilis*, respectively 620 and 271 genes were identified as essential, some of them being sporadic within the species (Gerdes et al., 2003).

It thus appears that some newly acquired genes can persist and provide a substantial gain in fitness to their host. For instance, genes that enable rhizobia to have a symbiotic relationship with plants are quite heterogeneously distributed among Proteobacteria (González

et al., 2003), but are fixed in numerous lineages that have adopted this highly beneficial lifestyle. This kind of genes with heterogeneous occurrence in prokaryotes (HOPs) have likely been transferred frequently while rarely settling stably in a species' genome. Indeed, it has been shown that genes providing advantages in specific environmental conditions in *E. coli* have a higher tendency to be transferred and lost across the tree of Proteobacteria (Pal et al., 2005). The tendency of a gene to be transferred also correlates with the situation of its product in the metabolic (Pal et al., 2005) or protein interactions networks (Davids and Zhang, 2008), with genes having few connections being more likely to be transferred. Such peripheral functions might mostly consist of interactions with the environment – like transport or sensing – and to be adaptive only under some specific ecological constraints. HOPs are likely retained when the environment provides sufficient selective advantage for such interactions, but will be lost if such conditions are only transient.

In certain cases, however, such genes accumulate in genomes and provide the cell with environmental versatility. This is the case in the group of the Rhizobiales, where large multi-replicon genomes harbour the highest known diversity of ABC-type transporters (Mauchline et al., 2006), which allows these organisms to live on a high diversity of carbohydrate nutrients. In *Frankia*, host versatility has similarly been related to increased genome size (Normand et al., 2007). In fact, it is possible that genes with high adaptive potential stay in pangenomes without reaching fixation, provided there exists a transient selective advantage, which insures episodic increase of their gene frequency. This would be particularly true if intra-specific recombination occurs frequently. Indeed, condition-specific or niche-specific genes seem to be the most mobile (Pal et al., 2005), and sufficient transfer rates may prevent them from extinction by jumping from an organism to another. For instance, genes determining pathogenicity in *Vibrio cholerae* are associated with mobile elements like CTX phages and TCP pathogenicity islands (Rahman et al., 2008). Pathogenicity in animals is only an accessory ecological niche of *V. cholerae*, which is essentially an aquatic microbe, but the determinants for pathogenicity remain constantly present in the environmental reservoir.

In some instances, transferred adaptive genes become fixed and transmitted vertically to all descendant taxa. The reconstruction of gene acquisition history on the lineage of *Salmonella enterica* sv. Typhimurium LT2 (Porwollik et al., 2002) showed the gradual acquisition of pathogenicity determinants such as fimbriae, lipopolysaccharides or lipoprotein envelope biosynthesis genes during adaptation to the new warm-blood habitat. In this perspective, any clade-specific gene family has potentially been involved in the stable adaptation of the clade to a new ecological niche. Therefore, new genes enabling the exploitation of a new ecological niche in a clade ancestor can be fixed and become core genes of the descendant taxon. Altogether, this suggests that pan- and core genomes are not isolated evolutionary compartments because core genes can be lost and variable genes can become essential. In fact, analysis of cellular networks shows a continuum of connectivity properties correlating with gene ubiquity, suggesting that new genes are progressively integrated into the network and the genome. It has been shown that after its acquisition, a gene slowly acquires more and more cis-regulatory elements, im-

proves its co-expression with functional partners and progressively builds interactions with other gene products (Lercher and Pal, 2008). Notably, as newly acquired genes involved in ecological adaptation form interactions with core genes, they may become essential. In fine, the roles assumed by clade-specific genes involved in niche adaptation could become as essential as the replication or translation machinery in defining a bacterium's biology.

Hence, describing the ecological niche of prokaryotic species, and particularly identifying to what extent genes acquired by HGT can modify its frontiers seems crucial to understand the dynamics of evolution of gene repertoires. While it is likely that species pangenomes contain a majority of genes transiently integrated in genomes with no particular adaptive fate, a fraction may support the ecological innovation of lineages, potentially leading to their ecological speciation (Cohan, 2001). In the 'stable ecotype' model, Cohan (2002b) predicts that genes responsible for niche-specific adaptations will be kept in genomes of members of the descendant clade by periodic selection, thus shifting from the unstable pangenome to the core genome of the new clade. Understanding how diversity can be generated in prokaryotes, and how this is related to changes in their environment and in the selective pressures acting on genes is the key of the pangenome conjecture.

1.1.5 Horizontal gene transfers and its role in speciation

The notion of species in prokaryotes is intensely debated (Achtman and Wagner, 2008). This is intimately linked to the current inability of evolutionary microbiologists to find a unifying model of prokaryotic cladogenesis. The classical 'Biological Species Concept' (BSC) (Mayr, 1942), that was originally defined for animals, places the sexual isolation of clades as the central condition for their divergence. In this model, sexual isolation can mostly arise through appearence of pre-zygotic barriers or geographical separation of lineages. Prokaryotes usually co-occur in the environment, most of times with no geographical structure of populations, and they are clonally reproducing but subject to rampant trans-specific HGT. In this context, it is impossible to understand the emergence of prokaryotic species in the framework of the BSC.

Ecological speciation appears as an alternative, and was proposed to be the major way through which prokaryotes diversify (Cohan, 2002b). Paradoxically, HGT can be the cause of ecological speciation. Acquisition of a gene bearing a new function can lead to the emergence of an ecoytpe, a variant with a specific ecological adaptation allowing the exploitation of an ecological niche different from that of the parental population (Cohan, 2002a) (Fig. 1.2 A). Such ecological isolation induces the escape of the ecotype from genotypic homogeneization with its relatives by homologous recombination. In addition, the ecotype is subject to strong periodic selection for the ecological trait that defines it. This leads to the frequent purge of diversity within the ecotype and thus accelerates its divergence towards a better adaptation to its new niche (Cohan and Koeppel, 2008) (Fig. 1.2 A,B). As the ecotype diverges independently, the potential to recombine with the parent population decreases (Roberts and Cohan, 1993; Majewski et al., 2000) and eventually, when a threshold of recombination over divergence is crossed, the ecotype lineage is no more able to re-hybridize with its relatives, causing the

Figure 1.2: **The stable ecotype model.** (Fig. 1.A,B and 2.A from Cohan and Perry (2007), omitted here for copyright reasons)
(A) Mutation and recombination events that determine ecotype diversity in bacteria. Circles represent different genotypes, and asterisks represent adaptive mutations. (top) Periodic selection mutations. These improve the fitness of an individual such that the mutant and its descendants out-compete all other cells within the ecological niche (ecotype); these mutations do not affect the diversity within other ecotypes because ecological differences prevent direct competition. Periodic selection leads to the distinctness of ecotypes by purging the divergence within but not between ecotypes. (bottom) Ecotype-formation mutations. Here a mutation or recombination event allows the cell to occupy a new ecological niche, founding a new ecotype. The ecotype-formation mutant, as well as its descendants, can now escape periodic selection events from its former ecotype. **(B)** Schematic relationship between history of ecologically distinct populations and diversity DNA sequence clusters. Ecotypes are represented by different colors; periodic selection events are indicated by asterisks; extinct lineages are represented by dashed lines; clades that may be perceived as sequence clusters are marked by a horizontal black line at the top of the phylogeny.
Excerpts from Cohan and Perry (2007).

irremediable separation of descendant clades, i.e. speciation (Fraser et al., 2007; Doroghazi and Buckley, 2011).

Ecological speciation can occur through physical isolation at the micro-scale due to different ecologies (parapatric speciation) (Hunt et al., 2008) or even when maintaining potential contact (sympatric speciation). In the latter case, recombination may allow the transmission of niche-specifying genes among populations and hence erase the ecological barrier between them. However, because periodic selection maintains only the fittest genotypes in each niche, it counter-selects the assortment of niche-specifying genes with genomic backgrounds less adapted to the niche (Cohan and Koeppel, 2008; Hausdorf, 2011), for instance because of inappropriate regulation of the niche-specifying genes. Thus, divergence and speciation of an ecotype may occur even in presence of recombination with its relatives. A simulation study showed that speciation could even occur with the ecological adaptation being coded by several independent loci, but that high levels of recombination could promote mixing of genotypes among ecotype populations, despite the incurred selective cost, and prevent sympatric speciation (Friedman et al., 2013).

It was further proposed that even in presence of high rates of genome-wide recombination, the insertion of transferred genes bearing ecological adpatation could prevent recombination at the insertion locus. Indeed, recombination with members of another population that do not possess the niche-specifying gene at the homologous locus would cause its loss, what would be counter-selected (Retchless and Lawrence, 2007; Lawrence and Retchless, 2009) (Fig. 1.3 A). Population genetics modelling showed that this causes a mechanical arrest of recombination at loci surrounding clade-specific insertions (Vetsigian and Goldenfeld, 2005). The multiplication of such sexually isolated clade-specific loci (Lawrence and Retchless, 2009; Retchless and Lawrence, 2010) or the propagation of the recombination arrest around chromosomes

Figure 1.3: **The fragmented speciation model. (Fig. 33 and 34 from Lawrence and Retchless (2009), omitted here for coyright reasons)**
(A) HGT interferes with recombination between ecologically dissimilar organisms. The gain of ecologically adaptive genes X, Y, and Z in one population can interfere in several ways with recombination between populations. Movement of fragment class A would result in loss of genes X, Y, and Z from population II, what would be counter-selected by periodic selection in the niche; movement of fragment class B would result in potentially problematic gain of genes X, Y, and Z in population I, should their expression there be counterproductive because of innapropriate integration to the cellular networks. Introduction of fragment class C does not result in recombination due to a lack of homologous sequences at the dsDNA end. **(B)** Model for the gradual establishment of genetic isolation. An ancestral population acquires two different adaptive loci. Recombination is inhibited in the regions surrounding those loci, leading to sequence divergence in the flanking genes; this is depicted as gray regions on the chromosomal backbone. The accumulation of additional adaptive loci, as well as growth of the regions with recombination inhibition, eventually leads to the genetic isolation of all genes. Excerpts from Lawrence and Retchless (2009).

(Vetsigian and Goldenfeld, 2005) would eventually lead to the complete separation of lineages (Fig. 1.3 B).

The 'stable ecotype' model (Cohan, 2002b) seems relevant to the study of the first steps of the diversification of a bacterial population, and indeed it was suited to describe the recent divergence of clades of *Bacillus simplex* (Sikorski and Nevo, 2005) or *Mycobacterium tuberculosis* (Smith et al., 2006). However, it cannot describe the evolution of bacterial clades on the long run. The 'fragmented speciation' model does so, notably by considering the successive acquisition of niche-specifying genes along the genome (Retchless and Lawrence, 2007), that can be seen as an iteration of ecotype speciations. Other speciation models exist (Fraser et al., 2009; Achtman and Wagner, 2008), but the following studies will principally rely on the latter as explicative frameworks of our observations on genome evolution.

1.1.6 Reconstructing genomes histories to reveal past and present ecological adaptations

Ecological speciation models consider that periodic selection is going on over the history of a lineage, maintaining the niche-specifying genes originally gained by ecotypes in their descent (Cohan, 2002b; Retchless and Lawrence, 2007). Indeed, genes conserved in genomes since their acquisition by an ecotype ancestor must have been under long-term purifying selection for their ecological function.

Actually, in studies characterizing species pan- and core genomes, many core genes have still no functional annotation, and are thus unlikely to participate to the rather well described bacterial central metabolism. Instead, they may participate in the stable adaptation of the species in its natural environment. Efforts must be made to phenotypically characterize those core genes and better understand their role in a species' biology.

The conservation of genes within all members of a clade constitute a first good indication of purifying selection for their ecological role. Additional observation may support or invalidate the selection hypothesis, notably by the comparison of the processes that introduced genes in genomes to a neutral model of genome evolution. Such model can be empirically derived from the global patterns of gene gain and loss along the history of bacterial genomes.

The history of genomes and the processes that shaped their evolution can be reconstructed. A rigorous approach is to explicitly take into account the history of genes in the reconstruction of the history of genomes. This can be done by reconciling the trees of genes with that of species.

1.2 Reconstruction of ancestral genomes and reconciliation of gene and species histories

1.2.1 Brief history of phylogenetics applied to genes and species

Molecular phylogeny was originally designed to tell the history of macromolecules (Zuckerkandl and Pauling, 1965) but rapidly gene or protein sequences were used as markers of the evolution of their host species, replacing morphological and biochemical characters. A convenient marker was found in the gene coding the RNA component of the small sub-unit of ribosomes (16S rRNA), as it was ubiquitous and well conserved among known species and thus appropriate to compare distant species. The first phylogeny of the three domains of living organisms was hence derived from the analysis of their 16S rRNA sequences (Woese et al., 1985). Other genes in genomes were sequenced and their molecular phylogeny revealed an almost systematic discord among their history. Errors of reconstruction due to saturated signals and to the limited explanatory power of models of molecular evolution could account for a part of these systematic incongruences but soon enough, it became evident that the history of genes were genuinely different (Brown and Doolittle, 1997).

To explain these differences, it was necessary to invoke events of gene duplication followed by differential losses or, most often in gene families sampling prokariotes, events of horizontal gene transfer. However, the question arose of what phylogeny should be considered as the reference of species phylogeny, to which could be compared others to infer potential events of HGT, as even the 16S rRNA could be subject to horizontal transfer (Yap et al., 1999).

With the advent of the genome sequencing era, it became apparent that most genes in genomes disagreed on their history, but paradoxically, it was shown that combining the information from all genes could recover a common vertical signal of descent (Wolf et al., 2001). From this vertical reference could be compared individual gene tree toplogies to model the combination of duplication, transfer and loss (DTL) events that marked the history of genes in genomes. Performing this individual reconstruction of DTL scenarios for all genes and integrating them at the genome scale allowed to reconstruct ancestral genomes (Snel et al., 2002). However, the models used for the reconstruction of these scenario, and the way they integrate

the informations from gene histories has a great impact on the outcome of ancestral genomes.

1.2.2 Concepts of ancestral genome reconstruction and state of the art

Phylogenetic profile mapping methods Several method have been developed through years, and a popular approach uses the mapping of profiles of presence/absence of genes in extant genomes (i.e. phylogenetic profiles) on a phylogeny of species to propagate the presence/absence states to ancestral genomes. This can be done under different models of gene evolution, the simpler being a birth and death process corresponding to events of gene gain and loss along lineages. These models can be subdivided in classes depending on the way best solution of reconstruction are found: using a parsimony approaches (Mirkin et al., 2003; Makarova and Koonin, 2005; Boussau et al., 2004), that are computationally efficient, or in a probabilistic framework (Pagel, 1997; Yang et al., 2012; Viklund et al., 2012), that is more satisfying theoretically but also more computationally demanding.

Other methods distinguish the process by which a gene can be gained in a lineage either by duplication, by horizontal transfer or by apparent origination – which can reflect true gene genesis or HGT from an unsampled lineage. Again, these models are available as either using parsimonious (Snel et al., 2002; Csűrös, 2008) or probabilistic (Csűrös and Miklós, 2009) criteria to find the best solution. These methods based on phylogenetic profiles are efficient in recognizing variation of gene content, but fail in detecting events of gene replacement by HGT. In addition, most are only indicated to deal with profiles of distribution of clusters of orthologous genes (Mirkin et al., 2003), which poses problem of the definition of orthology (Kristensen et al., 2011), and in addition loses the information about the origin of gene lineages. When they deal with multi-copy homologous gene families (Csűrös, 2008; Csűrös and Miklós, 2009), mapping methods lose sensitivity when working on more than a few paralogs, because the parallel gains and losses in paralogous lineages can be missed if they do not change the apparent number of homologs in genomes.

Reconciliation methods A more rigorous way to describe the history of gene families is to consider their phylogenetic trees. Reconciliation of gene and species histories consists in mapping evolutionary events on both gene and species tree, in a way that make their respective histories concordant (Fig. 1.4). Different kinds of reconciliation methods exist, notably differing in the nature of event they try to infer to explain the incongruences between gene and species tree topologies.

Many programs designed for the reconciliation of Eukaryotic phylogenies only consider the duplication and loss (DL) events. This is the case of "TreeBest" (Vilella et al., 2009) and "PhylDog" (Boussau et al., 2013), among many others. On the other hand, some programs, like "Prunier" (Abby et al., 2010), were designed specifically to tackle the problem of HGT in prokaryotic gene phylogenies and to provide a phylogenetically-sound alternative to programs based on

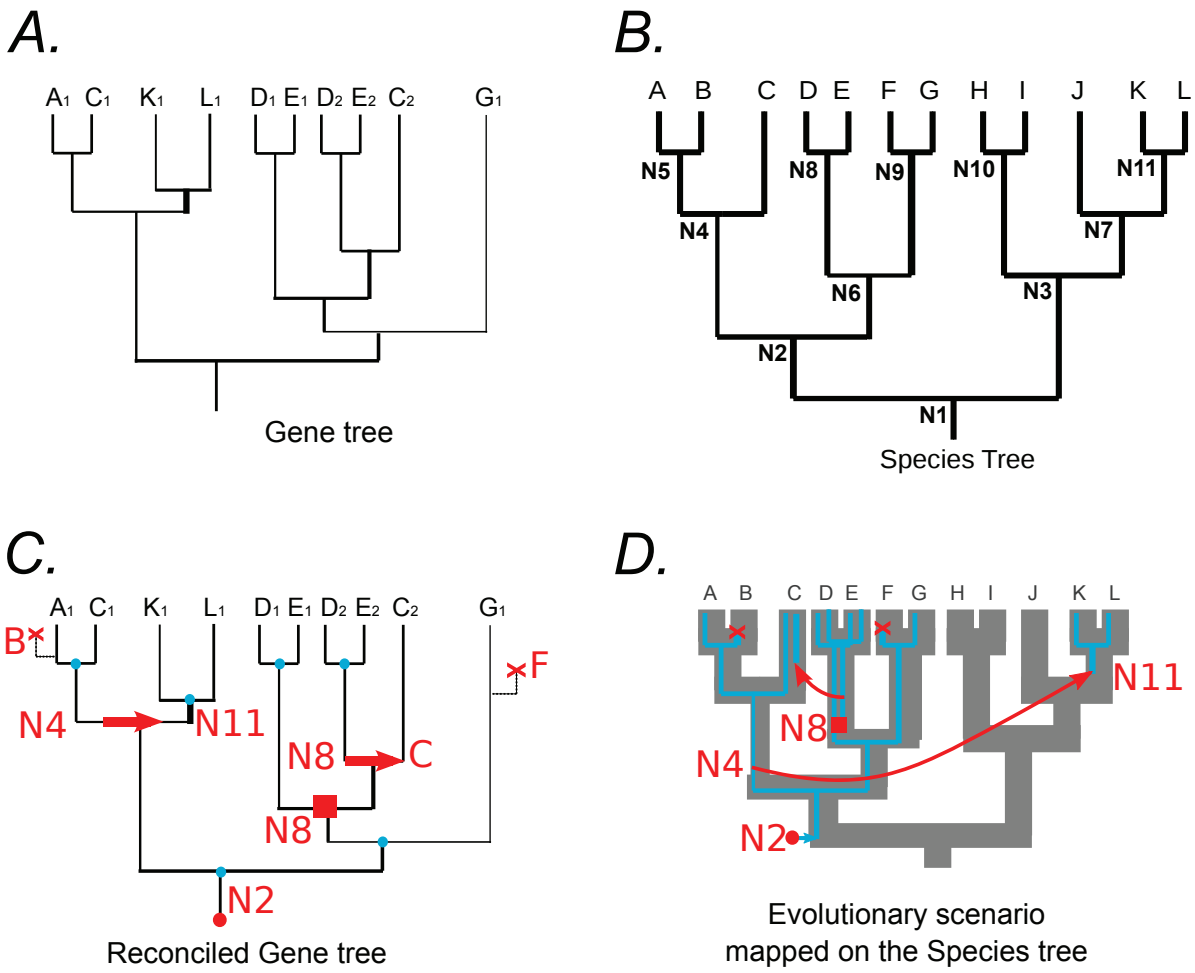


Figure 1.4: **Schematic example of a reconciliation.**
A gene tree (A) and the reference species tree (B) are reconciled so that the pattern of occurrence of genes in extant species and their evolutionary relationships are congruent with the history of species. It leads to the inference along the history of the gene of events of origination (red points), duplication (red squares), horizontal transfer (red arrows) and losses (red crosses), and of a complement of speciations (blue points) (C). These events are located in the species tree (red tags in (C) referring to nodes in (B)) and the scenario depicting the evolutionary history of the gene family is contained in the tree of species (D). It follows that the presence and absence of the gene family in ancestral genomes is reconstructed (D).

the detection of anomalous nucleotide composition as signatures of HGT. Other programs deal with the more general, but harder problem of a duplication, transfer and loss (DTL) model.

Similarly to mapping methods, reconciliation methods can rely on a parsimonious or a probabilistic framework. As the complexity of the task of reconciliation can be very high in large multi-copy gene families marked by numerous DTL events, and because these approaches are generally intended to be applied to complete genome datasets, computationally efficient parsimony methods were often preferred. Among them can be cited "Mowgli" (Doyon et al., 2008), that efficiently explores the space of DTL parsimony costs through a dynamic programming approach. This dynamic programming was then used to implement a reasonably rapidly performing program, "ODT", that implement the DTL reconciliation in a probabilistic framework (Szöllosi and Daubin, 2012; Szöllősi et al., 2012).

One drawback of reconciliation approaches is there sensitivity to errors in reconstruction in gene and species phylogenies. This problem can be dealt with by considering in priority the robust signal in gene trees to orientate the reconciliation. Following this principle, the program Prunier identifies the major statistical conflicts and removes them in an ordered fashion (Abby et al., 2010). In this procedure, the removal of key conflicts induce the resolution of other conflicts with lower support, which leads to conservative HGT reconciliations.

Another approach consists in changing the gene tree, for instance by spotting nodes of gene trees where the phylogenetic support is low and that induce many evolutionary events, and trying to change the local gene tree topology. This is interesting since it provides a guide to the exploration of the gene tree space that is independent of the sequence alignment. Maximum-likelihood methods for reconstruction of gene trees are popular because their are relatively fast, but one of their major issues is the inefficient exploration of the topological space, which is extremely vast and strewn with local optima. Thus, modifying the gene tree to find better reconciliations might lead to find better gene trees overall.

A first kind of these approaches consist of exploring a gene tree space through a gene tree sample given in input. The "TreeBest" pipeline (Vilella et al., 2009) compares a set of input gene trees computed with several phylogenetic programs and evolutionary models, and keeps the one minimizing the reconciliation costs. While this allows to cover the variability of evolutionary rates among families, the consideration of a small number of 'majority' (maximum-likelihood) results may not escape local optima of the topological space. The program "AngST" (David and Alm, 2011) uses larger samples of replicates of a gene tree, such as those generated by bootstrapping or bayesian analyses, in a procedure to amalgamate the information from replicates. This method explores the sample of of trees to find local topologies that betters the reconciliation . A related procedure was implemented to complete the "ODT" program by amalgamating subtrees of trees sampled in bayesian analyses according to their posterior probabilities to provide an estimation of the joint likelihood of the possible gene tree topologies with the possible reconciliations (Szöllősi et al., 2013).

Another method is the active exploration of the space of gene tree topologies around the

input topology to find more parsimonious/probable reconciliations. "MowgliNNI" (Nguyen et al., 2012, 2013), an extension of "Mowgli", performs nearest neighbor interchanges (NNI) of gene tree branches that are not well supported, but this does not guarantees that a topology that is found optimal under the reconciliation parsimony criterion will be a good tree given the original alignment. "PhylDog" may provide the most satisfying solution by jointly estimating the reconciliation, the gene trees and the species tree under a global probabilistic model. This approach is however extremely demanding in computation time and space, and restricted to a duplication-loss (DL) model. Under the more complete DTL model, the "ODT" program can however use the property of time consistency of transfer, i.e. the necessity for lineages that exchange genes to have co-existed in time, to provide a relative time order for speciations in the reference tree (Szöllősi et al., 2012). This idea was pushed forward to take into account the possible transit of genes in extinct or unsampled lineages during their evolution, what revealed that indeed most of the evolution of genes happened in lineages we never saw (Szöllosi et al., 2013).

Finally, recent development explored more complex models of gene evolution than the simple gene-centered DTL model. For instance, the "DLCoal" model – which only tackles the DL problem – (Rasmussen and Kellis, 2012) integrates the process of incomplete lineage sorting (ILS), that is the perpetuation of polymorphisms after speciation, which can be resolved differently in sister lineages and lead to topological incongruences with the species tree. In addition, "DLCoal" models the non-independence of neighbouring genes by introducing a 'locus' layer between the gene and genome levels of evolution. The DTL model of "AngST" was extended in "Ranger-DTL" with the explicit modelling of the loss of efficiency of homologous recombination with phylogenetic distance (Bansal et al., 2012).

These refinements of phylogenetic reconstruction of ancestral genomes and gene histories – and hopefully their integration in a unique framework – will greatly profit to the community, as one will be able to locate any genomic innovation in the history of species and to understand their origin in the evolutionary dynamics and selective pressures acting at the level of genes, loci, genomes/organism, clades and maybe communities of organisms.

1.3 Glossary

In the section 2.4, I will use several terms referring to the lexicon of genes, genomes, reconciliation and ancestral reconstruction. I will consider several biological concepts, that correspond to objects that have (somehow) a material existence but need nonetheless to be defined (genome, gene, species . . .), and others that are purely conceptual (gene tree, species tree, ancestral genome . . .).

A *clade* is defined in a *tree* as a monophyletic group of extant organisms or genes and their ancestors. Many clades found in *species trees* and *gene trees* can be equated to *taxa* following the

official nomenclature, though many taxa are in debate for their definition considering which organisms it should include, and these debates are notably rooted in the the observation of varried grouping of organisms as clades depending on the phylogenetic tree.

A *species* is a clade regrouping fairly similar organisms. Finer definitions invoking cohesive reproduction, ecological selection and other factors are subject to a long-lasting debate on the nature and even the possible existence of prokaryotic species, that I would not bring in this manuscript; a glimpse of it may however be taken in the following literature: Lan and Reeves (2000, 2001); Cohan (2002b); de Queiroz (2005); Konstantinidis and Tiedje (2005); Konstantinidis et al. (2006); Achtman and Wagner (2008); Fraser et al. (2009); Doolittle and Zhaxybayeva (2009); Vos (2011). If used in this study, the term *species* refer to *genomic species*, a practical taxonomic level defined by clusters of genomic diversity (Stackebrandt et al., 2002). In much cases, though, I will prefer the use of *clade* for *bona-fide* genomic species as for other monophyletic clusters, because the monophyly of species seems to be the only criterion on which evolutionary microbiologists agree.

Mention of a *biovar* in a taxon name indicate a subdivision of a larger taxon which name it follows (as for instance in "*Agrobacterium* biovar 1", which indicates a subdivision of *Agrobacterium* genus) to which isolates are assigned based on biochemical testing of their metabolic properties, usually the ability to degrade a range of compounds and/or use them as a carbon, nitrogen or energy source. Similarly, *pathovar* and *symbiovar* refer to classifications that are made regarding the pathogenic or symbiotic phenotype of isolates, usually referring to their host tropism. Finally, *genomovar* refer to a classification based on genotypic traits rather than on phenotypic ones, and particularly to a set of traits that are representative of the genome-wide variation, as opposed to single-locus variant classifications. Indeed, genome-wide and single-locus variants can be completely decorrelated, especially when the investigated single locus is linked to a phenotype under strong environmental selection, like outer-membrane proteins of pathogens that would be recognized differently by the host's immune system - in this particular case one can use the term *serovar*, that refers to a classification based on the recognition of antigens by type antibodies. Note that the term *genomovar* is used to name genome-based subdivisions of a taxon, but that the concept of a genomic variant is designated by the term *genomic species* or *genomospecies*. For instance, "*Agrobacterium tumefaciens* genomovar G4" is the name of one of the genomic species of the *A. tumefaciens* species complex.

A *genome* is the sum of all genetic information coded in the deoxy-ribonucleic acid (DNA) polymers found in one cell. It can correspond to several distinct molecules, the *replicons*. A genome is documented by its sequence (a *genomic sequence*), which is a sentence using the 'ATGC' alphabet representing the succession of nucleotides that constitutes the DNA molecules. The sequences that make the datasets of the following studies do not represent the genome of a single cell, but that of a clonal colony that was used for the sequencing experiment. These genomic sequences are usually considered as representative of the genomes of all individual cells of a strain – a presumably genotypically homogeneous clone – and sometime of a species, which is certainly abusive, as can be seen in the present work. A genome can also be an *ancestral*

genome, that is the complete genetic information of the putative common ancestor of a clade. In the present evolutionary study, a genome can generally be confounded with the organism that bears it, because the only characters we used to infer the history of organisms are contained in their genome.

A *replicon* is a DNA molecule present in the cell that can be replicated over the cell cycle. This term includes chromosomes, plasmids and chromids – that are an intermediate between the two former, as proposed by Harrison et al. (2010). A replicon can be present in one or several copies per cell, have various size (from a few kb for large-copy number plasmids to a few Mb for chromosomes in Bacteria) and be conjugative or mobilizable. In our perspective, a replicon in a genome can be found homologous to another replicon in another genome, but this relationship can be complex, as the the genes that make up replicons can be distributed to different among replicons in different genomes. Homology of replicons can be recognized by their homologous systems of replication and by their large-scale *synteny*, i.e. the collinearity of homologous segments of DNA at a scale larger than a gene.

A *gene* is here first defined as a coding sequence (CDS), i.e. an open reading frame in a genomic sequence that was annotated as (putatively) coding for a protein product. Annotation of CDSs in genomes can vary depending on the pipeline, but usually rely on signature of biased codon usage and mostly on sequence conservation with homologous loci in other annotated genomes. This leads to the second potential meaning of a *gene*, that is the entity that evolves in genomes: it corresponds to an homologous *gene family* in extant genomes, and to the ancestor of this gene family in the genome of the putative common ancestor of extant species.

A *gene family* is a set of extant gene sequences whose similarity was attributed to a shared common ancestry. Homologous gene families used in this work are defined using the Hogenom procedure (Penel et al., 2009), based on a transitive similarity search. The evolutionary relatedness of homologous genes can be represented by the alignment of gene sequences, and by the reconstruction of a *gene tree*.

The *homology* is the relationship between biological objects, here extant genes or gene lineages, that share a common ancestor. Homologous genes, or *homologs*, can be *orthologs* if the path separating them in the *gene tree* does not involve duplications, otherwise they are *paralogs*. I also use the term *co-orthology* to design the relationship of orthology of a gene with a pair of paralogs (Kristensen et al., 2011). The term *xenolog*, meant to refer to genes related by a history of transfer, is not used here, because many orthologs would in fact be hidden xenologs, due to frequent *homologous recombination* in closely related lineages.

A *phylogenetic tree*, or an *evolutionary tree*, or simply a *tree*, is a dendrogram depicting the evolutionary relatedness of objects situated at its leaves. Here, we mostly consider trees representing molecular phylogenies, that represent the evolutionary relatedness of sets of homologous sequences. The dendrogram is made of *branches* connecting *nodes*, one particular type of node being the *leaves*. Branches may have lengths proportional to the quantity of evolution that separates two nodes of the tree, as computed from the sequence alignment under a model of evolution of biological macromolecules (DNA for our concern in the present work),

and in this case the dendrogram is a phylogram. The branches can moreover be annotated with *supports*, which can be computed in several ways and thus can have several meanings, but in general refer to the confidence one can have in the bipartition of leaves represented by the branch. Because the tree is depicting an evolutionary process, *internal nodes* (by opposition to leaves) represent an ancestral sequence. When the tree is *rooted*, i.e. when is defined a start point for the evolutionary process, an internal node can be considered as the ancestor of all nodes (internal and leaves) below it, which together form a *clade*.

A *gene tree* is a tree depicting the evolutionary history of sequences of a *gene family*, which are represented at the tree leaves (Fig. 1.4 A). The trees produced by reconstruction methods is generally not rooted, and rooting a gene tree can be a complicated problem in the face of a complex evolutionary history of the gene family. A gene tree can be annotated with evolutionary events at its nodes, in which case it is a *reconciled gene tree* (Fig. 1.4 C).

A *species tree* is a tree depicting the evolutionary history of organisms, which is equivalent to the evolutionary history of their genomes. This equivalence is more evident when one uses genetic information sampled from the largest possible fraction of the species' genomes to compute the species tree. Rooting a species tree is in general easier than for a gene tree, provided there is in the dataset an *outgroup*, i.e. a distant relative who will group as a sister clade of the dataset of interest. A species tree can be used as a *reference tree* for the reconciliation of histories of gene and genomes (Fig. 1.4 B).

An *ancestor* and its *ancestral genome* are a hypothetical genotype that is inferred at a node of the species phylogeny, as the common ancestor of the clade of organisms below. It must be noted that a branch of the species tree represent a continuum of many common ancestors of the clade to which leads the branch, and that the node at its end represents the last common ancestor (LCA) of the clade. In the present work, we assign *evolutionary events* to ancestors, who are referred to by the label of a node in the species tree, i.e. the LCA of the clade below. When not explicitely mentioned, using the LCA is a convenience of language to refer to one among all the ancestors in the lineage, as our phylogenetic analyses cannot distinguish them. This dicrete representation of evolution can lead to amalgamate many successive lineages that may have varied significantly of genome and ecology. This is especially true for long branches leading to isolated clades in the species phylogeny, which would need sampling of related clades (if they still exist) to add interleaving ancestors along the branch.

A *reconciliation* (of a gene tree with a species tree) is a set of *evolutionary events* associated to nodes in a the gene tree; the set of nodes and the set of events form a bijection. The sum of events forms a global *scenario* for the evolutionary history of the gene family, and this scenario can be mapped on the species tree (Fig. 1.4 D). A reconciliation is generally chosen among a large set of possible ones, and reconstructing it properly is one of the main matters of section 2.4.

An *evolutionary event* or *event* is invoked at the node of a reconciled gene tree to explain the distribution of species in the clade of genes below the node, in the context of surrounding clades, and given the reference species tree. It can be (in the present study) an origination, a speciation, a duplication, a *horizontal transfer* or a loss. An event is *located* in the species

tree (Fig. 1.4 C), what makes possible the projection of all gene tree events in the species tree, providing an evolutionary scenario for the gene family(Fig. 1.4 D). A *block event* is an event that spans several neighbouring genes, whose co-evolution at the occasion of the event leaves a common pattern in gene phylogenies. It is important to stress that the events that are annotated in gene trees are not the exhaustive inventory of what occurred during the diversification of lineages; these events are those that were followed by the successful survival of the gene lineage. The recorded event is thus the product of the transfer, duplication or speciation event itself and of the process of drift or selection that led to the fixation of the gene lineage in one of the genome lineages, whose extant representatives were sequenced.

The *location* of an evolutionary event is the mapping of the event in the species tree, i.e. the identification of the ancestor in which the event occurred. It consists of the label of the node representing the LCA of the ancestral lineage. I also use the term *coordinates* as an equivalent to *location*; *coordinates* is mostly used for horizontal transfers because they are characterized by two locations: that of the donor and that of the receiver lineage. In a complete reconciliation, the location/coordinates of an event in the species tree should correspond to only one node (a pair for transfers). When the reconciliation is not complete, i.e. when some events are not determined in the global scenario, I choosed to represent the location of the event with a set of nodes (see section 2.4.3.1 and Fig. 2.15), thus keeping their location *uncertain*.

A *horizontal transfer*, is the event of transmission of genetic material between two lineages, in a way decoupled from reproduction. A *horizontal gene transfer* or *HGT* is this process happening at the scale of a gene, or several genes for *block transfer events*. HGT can take place in many different ways physically, that can first be differentiated by the mechanism of uptake of incoming DNA into the recipient cell: conjugation, transduction and transformation. It can then be distinguished by the way the foreign DNA is integrated in the genome, notably via site-specific (non-homologous) recombination relying on integrases, that are generally associated to a viral vector; or via *homologous recombination*, that relies on the cellular machinery to identify regions of high sequence similarity and stretches of identical DNA and eventually to replace (*convert*) the native DNA. All those notions are equated in the present study, because the only way HGT are characterized is by the recognition in gene trees of phylogenetic incongruences with the species tree. What we can distinguish is the fact that a transfer adds a gene lineage to the clade (an *additive transfer*), or is replacing an orthologous lineage (a *replacing transfer*, that can be noted *replT*). These kinds of transfers can intuitively be thought to occur through non-homologous and homologous recombination, respectively, but this can be misleading: homologous recombination can lead to the addition (or substraction) of genes to a genome, by recombining fragments that differ in their median content (Fig. 1.3). Orthologous replacement can occur by the introduction of a gene orthologous to a resident one, but at a different locus, followed by the subsequent loss of the native copy; this will leave the same pattern in gene trees than an event of homologous recombination. The latter kind of replacement can be differentiated based on the syntenic context of genes; though this kind of analysis is not systematically used in the manuscript, it an easily be done through the **Agrogenom** interface

at `http://phylariane.univ-lyon1.fr/db/agrogenom/3/`.

Homologous recombination (HR), as defined above, is a kind of HGT. However, it can be convenient to differentiate HR from HGT, because of their different connotations in the literature: a genetic exchange that occurs via homologous recombination between fairly differentiated genotypes, so that it can be recognized in gene trees, will be called a HGT. HR rather refers to a process occurring mostly within genotypically homogeneous populations, because of the requirement of a high similarity between recombining alleles, leading to a log-linear decrease of HR efficiency with phylogenetic distance (Roberts and Cohan, 1993). Its frequency is relatively high compared to HGT, but its prevalence in nature is even harder to quantify than that of HGT, because most of the events of HR occur between virtually identical genotypes, and results in the supression of differences in their sequences. The nature of the process of *conversion* of alleles thus hinders its recognition by phylogenetic means at the micro-diversity scale. However, population genomic approaches can detect the presence of HR at such a scale, because they can recognize 'in negative' the absence of diversity at a locus compared to another. Because HR homogenizes the genotype of close conspecifics, it is considered the major mechanism of genetic cohesion of 'species'.

1.4 Outline

In this manuscript, I will explore the processes of diversification of bacterial lineages. I will particularly focus on how horizontal gene transfer and recombination shape the genomes and what are the adaptive consequences of their action.

In a first chapter, the diversity of gene repertoires among closely related species will be studied in a phylogenetic framework to recover the history of variation of gene contents along lineages. The identification of clade-specific traits in this history will be used to unravel past ecological adaptations, in a 'reverse ecology' approach. Empiric characterization of the main processes that shape gene repertoires – duplication, horizontal transfer and loss of genes – will serve to contrast patterns of gene conservation within clades and test the role of ecological adaptation in driving cladogenesis.

In a second chapter, the role of homologous recombination (HR) in shaping bacterial genomes will be revisited. While HR is known to maintain the genotypic cohesion of species and promote the propagation of adaptive mutation within populations, the possibility of the recombination process itself shaping the genome was unforeseen. In the light of the GC-biased gene conversion (gBGC) hypothesis, the long-term recombination history of genes and genomes and its effect on the nucleotide composition landscape will be explored. Notably, the possibility of recombination to be neutrally interfering with selection on codon usage will be tested. Finally, we will assay the potential of gBGC as a new evolutionary force to explain genome-wide variations of GC-content in a population genetics framework.

2

Evolution of gene repertoires in genomes of *A. tumefaciens* reveal the role of ecological adaptation in bacterial cladogenesis

2.1 *Agrobacterium tumefaciens*, a model organism for the study of bacterial cladogenesis and the quest for ecological adaptations

Agrobacterium spp. is a genus belonging to the Rhizobiaceae family of the Alpha-Proteobacteria class, which is primarily known for its pathogeny on plants (mostly woody dicotyledones). Indeed, agrobacteria can carry a plasmid that is able to transfer a segment of its DNA to a host plant genome. This transferred DNA (T-DNA) bears on one hand oncogenic genes, which expression in plant induce tumors in the form of crown galls (for tumor-inducing, Ti plasmids) or hairy roots (for root-inducing, Ri plasmids), and on the other hand genes for biosynthesis of opines (Otten et al., 2008). These molecules are condensates of amino-acids and sugars or keto-carbohydrates which are released by the diseased plant and then used by agrobacteria as carbon and nitrogen source (Moore et al., 1997). This ability to transfer DNA made agrobacteria a tool of choice in biotechnology to build genetically modified plants.

This pathogenic status originally defined the *Agrobacterium* genus, which gathered a heterogeneous set of organisms. They could be classified into three groups based on biochemical properties: biovar 1 strains were able to produce 3-keto-sugars, and grouped most strains of named species *A. tumefaciens* and *A. radiobacter*, while biovar 2 strains could not produce such

sugars, and grouped most strains *A. rhizogenes* with a root-inducing phenotype (Keane et al., 1970; Kersters et al., 1973). In addition, a third group (biovar 3) was defined that gathered strains inducing crown galls on grapevine and having a preferential use of L-tartaric acid, which were afterwards called *A. vitis* (Ophel and Kerr, 1990).

However, under the light of classification techniques using molecular markers, the *Agrobacterium* genus appeared polyphyletic (Young et al., 2001), and a debate opened on whether *Agrobacterium* should be integrated to the related genus *Rhizobium* (Young et al., 2001, 2003) or let as an independent taxonomic unit (Farrand et al., 2003). This polyphyly was resolved by transferring biovar 2 strains to the genus *Rhizobium*, forming the new species *R. rhizogenes* (Lindström and Young, 2011). This question is one of those still in debate in the field of taxonomic classification of Rhizobiaceae (Lindstrom and Martinez-Romero, 2002), and highlights the difficulty of differentiating closely related groups of bacteria on phenotypical characters that would justify assignation of a latin binomial species name (Stackebrandt et al., 2002). In particular, the original classification of agrobacteria regarding their pathogenic character – whereas this trait is coded by genes borne by accessory plasmids – certainly accounts for the current confusion on the matter.

Indeed, Agrobacteria were for long considered to be only plant pathogens, which ecology would be limited to crown gall and root tumors. However, they can be isolated from soils and from rhizospheres of healthy plants (Mougel et al., 2001) and the Ti plasmid – the causative agent of crown gall (Watson et al., 1975) – is absent from the majority of soil isolates (Bouzar et al., 1993). As a matter of fact, agrobacteria are primarily plant commensals, as they are usually found at much greater density in rhizospheres than soils (Mougel, 2000), and some were even shown to be able to promote plant growth (Hao et al., 2012b). In fact, there was a amalgamation between the true plant pathogens (the Ti/Ri plasmids) and their vectors (agrobacteria).

It has been shown that pathogenic agrobacteria could persist in soils outside of pathogenic outbreaks for long times (Krimi et al., 2002). To understand how these reservoirs of disease can be maintained, it is crucial to know the primary ecology of agrobacteria. Notably, determining the fitness of agrobacteria in their primary niche relative to that in their pathogenic (secondary) niche would help to understand the advantage of bearing Ti/Ri plasmids and how they fluctuate in populations of agrobacteria. Moreover, the association of the pTi/pRi with their agrobacterial hosts could involve complex interactions of genotypes, with certain plasmid types being more adapted to certain host genomes (Bouzar et al., 1993). Defining which agrobacterial population is susceptible of bearing a certain (plant-specific) type of plasmid would help to prevent epidemics, notably by diagnostic of resident agrobacterial populations prior to bed plants in nurseries, but also to design biocontrol strategies.

While some agrobacteria show marked association with host plants (at least regarding infection), as for instance *A. larrymoorei* on *Ficus* spp. (Bouzar and Jones, 2001) and *A. vitis* on grapevine (Ophel and Kerr, 1990), the association of *A. tumefaciens* with a particular niche is not documented. While the majority of known diversity consist of isolates that were recovered from crown gall tumors – and thus are not informative about the primary niche of the taxon –

systematic sampling efforts revealed the occurrence of *A. tumefaciens* in soils, ditch waters, and most importantly in (healthy) plant rhizopheres (Mougel et al., 2001; Portier et al., 2006; Shams et al., 2013), but without clear association to a plant taxon.

Figure 2.1: (Fig. 1 from from Shams et al. (2012), omitted here for copyright reasons) Maximum-likelihood phylogeny of the *recA* gene of type-strains of all *bona-fide* genomic species of *Agrobacterium* spp. known to date and related Rhizobiales using the revised nomenclature proposed by Costechareyre et al. (2010). Only significant support (SH-like) values (> 0.95) are given. The branch length unit is the number of substitutions per nucleotidic site. *B. Bradyrhizobium*, *Rh. Rhodopseudomonas*, *Az. Azorhizobium*, *M. Mesorhizobium*, *E. Ensifer*, *R. Rhizobium*, *A. Agrobacterium*.
Excerpt from Shams et al. (2012).

In fact, *A.tumefaciens*, corresponds to several species based on the study of their genomic diversity (i.e. genomic species). So far, eleven genomic species have been described, that are called genomovar G1 to G9 (Ley et al., 1973; Popoff et al., 1984) and G13 (Portier et al., 2006), and more recently *Rhizobium nepotum* (Puławska et al., 2012). *A.tumefaciens* thus forms a complex of genomic species, that are all closely related but distinguishable based on whole-genome analyses (Portier et al., 2006) and molecular marker phylogenies (Costechareyre et al., 2010; Shams et al., 2013). This genomic diversity is likely to reflect divergent ecological adaptations, especially because different species can be found co-occurring in the same micro-sample of soil (Vogel et al., 2003): under the competitive exclusion principle (Hardin, 1960), related species must share different niches to avoid competition that would result in the extinction of the less fit species.

Though, it is not evident for the moment what could be the nature of this ecological differences. Rather than marked associations with a host, genomic species of *A. tumefaciens* seem to show preferential association with some plant rhizospheres (Mougel, 2000). The host plant may "softly" select for some species, thus biasing the species diversity found in their rhizospheres compared to neighbour soils or other rhizospheres, but not to the point of restricting the species inventory. Such quantitative difference in the diversity of agrobacterial populations associated to plants may stem from subtle differences in the composition of plant root exudates. In addition, other environments can provide secondary niches to *A. tumefaciens*, that might participate in their ecological isolation. Indeed, it appeared in the past decades that *A. tumefaciens* could be responsible of nosocomial infections, notably on immuno-depressed human patients, such as HIV-positive and cystic fibrosis patients and sometimes following the introduction of a catheter in the bloodstream. A survey of the diversity of strains causing such infections again showed no strict association to a genomic species, but a prevalence of genomovar G2 (Aujoulat et al., 2011).

Altogether, these observations suggest the existence of ecological niches for each genomic group of *A. tumefaciens* that would be defined by complex associations of ecological factors. Without an *a priori* to test hypotheses of ecological adaptations, an alternative reside in an

approach of reverse ecology. From the comparison of genomes, one could define genomic characters that are specific to each genomic species, and could encode specific phenotypes. Studying the functions of these species-specific traits might help elucidate the cryptic ecology of *A. tumefaciens* species. In addition, the diversity within *A. tumefaciens* species complex is structured, with some of the species of *A. tumefaciens* being genomically closely related, as for instance genomovars G6 and G8, or genomovars G7 and G9 (Costechareyre et al., 2009). The comparison of their genome should provide a good resolution when searching for recent events of genomic diversification that could have initiated ecological differentiation. In general, reconstructing the history of genomes of the taxon in a phylogenetic framework would allow to replace key evolutionary event that shaped agrobacterial genomes at every step of their diversification. Among those events, some might have brought in genomes some characters that provided new ecological adaptations and were selected for. Recognizing such events might help to assign ecologies to present day species, and at the same time to understand how ecological adaptation can influence the process of cladogenesis in bacteria.

We thus undertook the reconstruction of the diversification history of *A. tumefaciens* by comparing the genomes of strains of several isolates all genomic species and closely related outgroup species. We first used an approach of comparative genome hybridization (CGH), taking advantage of the public and well-annotated sequence of the complete genome of strain C58 (Goodner et al., 2001; Wood et al., 2001) and of several Ti plasmids to design a micro-array that covered most of the genomic diversity of *A. tumefaciens* known at the time. This allowed us to probe 25 strains of *A. tumefaciens* and cognate species for the presence of gene homologous to that of strain C58 2.3.2.

We then gathered a dataset of genomic sequences of Rhizobiales, including sixteen new genomes of *A. tumefaciens* (sequenced for this study), plus that of strain C58 and five others that were publicly released in the meantime (Wibberg et al., 2011; Li et al., 2011; Ruffing et al., 2011; Hao et al., 2012a,b), and other Rhizobiales genomes, among which those of representative agrobacteria of biovar 2 and 3 (Slater et al., 2009). We reconstructed the history of genes in these genomes, in a phylogenetic framework (section 2.4).

Figure 2.2: (Fig. 4 from Slater et al. (2009), omitted here for copyright reasons) Reconstruction of the origin of secondary chromosomes and related large replicons within the Rhizobiales through transfers of gene clusters from the primordial chromosome to what originally was a repABC-type plasmid (called here the ITR plasmid, corresponding to the chromid introduced by Harrison et al. (2010)). LGT, lateral gene transfer.
Excerpt from Slater et al. (2009).

An interesting feature of *A. tumefaciens* is the complexity of its genome, that contains, in addition of the classic bacterial circular chromosome and of the facultative Ti plasmid, a linear chromosome and another facultative megaplasmid, the pAt. The linear chromosome is of particular interest because it belongs to a family of secondary replicons specific of Rhizobiaceae

(Slater et al., 2009), called chromids that are of plasmid descent but have compositional and evolutionary properties of chromosomes (Harrison et al., 2010). A comparative genomic analysis retracing the history of the genome structure of Rhizobiales was previously done by Slater et al. (2009), but at a low phylogenetic resolution, with only one representative genome per genus or biovar. The linear topology of the chromid is specific to a clade of strains belonging to the biovar 1, and its potentially peculiar evolutionary dynamics have not yet been investigated by a comparison of several homologous linear chromids. We notably looked at the dynamics of recombination and genomic plasticity of each replicon in *A. tumefaciens* genomes, in order to characterize the impact of the replicon structure of the genome on its evolution and its adaptive potential.

2.2 Preamble to the comparative genomic studies

Finding traces of adaptation in genomes is usually achieved by identifying patterns of sequence conservation that depart from a neutral expectation. Notably, in the context of the study of bacterial pan-genomes, one can recognize the action of purifying selection in the excessive conservation of patterns of gene presence/absence in clades. Conservation of gene presence/absence states can readily be revealed by comparing core genomes of inclusive clades. However, to recognize *excessive* conservation, these patterns must be contrasted with a null hypothesis of genome evolution. Defining such neutral model is a hard task, notably for mechanistic models that should account for many different processes that participate in shaping genomes, for which most parameters are unknown. Alternatively, one can rely on empirical models based on measures of characteristic parameters, where the impact of selection is recognized as statistical departures from general trends. In that case, it is particularly important to consider the potential biases of the method used to generate the data in the definition of neutral expectations.

For instance, in our first study presented section 2.3.2, we used micro-array CGH to characterize the profile of gene presence/absence in genomes of the *A. tumefaciens* species complex, in reference to the genome of C58, a strain that belongs to genomovar G8. We were interested in finding genes specific of genomovar G8, because their specific conservation in this species make them good candidates for being under selective constraints, potentially for having supported past adaptations of the species to its ecological niche. However, the mere observation of the specific presence of these genes in G8 genomes was not sufficient to assert that these G8-specific genes were under purifying selection, because of the biased design of the experiment. Indeed, the DNA chips used to probe the gene content diversity of agrobacterial genomes represented the genome of C58, which is a G8 strain, so it was likely that we would find more C58 gene homologs in other G8 strains than in other species. Because of this methodological bias, the distribution of genes across species could not be used *per se* to reveal signatures of selection.

In that perspective, it appeared necessary to find another test of the selective hypothesis

that was independent of the methods used to reconstruct the evolution of gene repertoires. We found that G8-specific genes were often found in clusters in C58 genome. This regional distribution was neither expected from the experiment design, nor from a model of independent acquisition of genes in genomes. In addition, these species-specific gene clusters appeared to gather genes with high functional relatedness. This suggested that genes in these clusters were gained together in the genome of G8 ancestor and conserved since. The global conservation of the gene clusters could be explained by the component genes encoding together a selectable function, and thus being collectively under purifying selection. We thus wanted to recover the largest number of informations about the co-evolution of neighbour genes in genomes and their functional properties, to be able to compare evolutionary scenarios involving selection to neutral alternatives.

This led us to develop the pipeline presented in section 2.4 for the analysis of gene histories from complete genomic sequences. At the time of the begin of the study, there was no bioinformatic program to our knowledge (notably among those presented section 1.2.2) that would have allowed us to reconstruct ancestral genomes while considering the co-evolution of neighbouring genes through transfer and duplication. We thus innovated a method of reconciliation of history of genes and genomes that takes into account the regional phylogenetic signal, and uses it to reconstruct horizontal transfers or duplications involving blocks of genes. We then could exhaustively describe the patterns of (co-)transfer of genes in genomes. This allowed us to test the neutrality of observing co-transferred genes with related function to be specifically conserved in clades.

2.3 Probing the ubiquity of genes in *A. tumefaciens* genomes to characterize species-specific genes: insights into the specific ecology of *A. fabrum*

2.3.1 Introduction

The following article was published in 2011 in the journal *Genome Biology and Evolution* (Lassalle et al., 2011). It presents the exploration of the diversity of gene contents in genomes of the *A. tumefaciens* species complex, based on genome hybridization experiments with the genome of strain C58 as a reference. Strain C58 belongs to genomovar G8, one of the ten genomic species of the complex, for which every available strain was tested. This focus on genomovar G8 let us identify several genomic islands specifically conserved in this species. The evolutionary characteristics and functional annotations of these species-specific genes provided insights in the history of ecological adaptation of the taxon.

This work gathers many collaborators, and my personal contribution in the experiments is the following: I adapted the pipeline for signal processing of CGH raw data into profiles of presence/absence of genes, originally developed by Laurent Guéguen, for its efficient use on the dataset. I additionally developed a method for the inference of sequence conservation from

CGH raw data, and used it to observe the regional patterns of conservation across replicons and genomes. I performed the evolutionary studies, i.e. the analysis of the patterns of presence/absence in the context of the evolutionary history of the taxon, leading to the definition of clade-specific genes, and the analysis of the codon usage of genes along a gradient of evolutionary conservation. Finally, I manually re-annotated the biochemical functions of species-specific gene products from the results of multiple bioinformatic programs and similarity searches, to integrate the knowledge on the function of these species-specific genes into a model of putative ecology of the G8 clade. Strong predictions on molecular function of genes could be tested and validated by collaborators in the laboratory, revealing unforeseen aspects of the physiology of agrobacteria.

2.3.2 Manuscript

Genome Biology and Evolution

SMBE

Genomic Species Are Ecological Species as Revealed by Comparative Genomics in *Agrobacterium tumefaciens*

Florent Lassalle[1,2], Tony Campillo[1,3], Ludovic Vial[1], Jessica Baude[3], Denis Costechareyre[1], David Chapulliot[1], Malek Shams[1], Danis Abrouk[1], Céline Lavire[1], Christine Oger-Desfeux[4], Florence Hommais[3], Laurent Guéguen[2], Vincent Daubin[2], Daniel Muller[1], and Xavier Nesme*[,1]

[1]Université de Lyon; Université Lyon 1; CNRS; INRA; Laboratoire Ecologie Microbienne Lyon, UMR 5557, USC 1193, Villeurbanne, France

[2]Université de Lyon; Université Lyon 1; CNRS; Laboratoire de Biométrie et Biologie Evolutive, UMR 5558, Villeurbanne, France

[3]Université de Lyon; Université Lyon 1; CNRS; INSA de Lyon; Bayer Crop Science; UMR 5240, Laboratoire Microbiologie, Adaptation et Pathogénie, Villeurbanne, France

[4]Université de Lyon; Université Lyon 1; SFR Bio-Environnement et Santé; PRABI Pôle Rhône-Alpes de Bio-Informatique, Villeurbanne, France

*Corresponding author: E-mail: nesme@univ-lyon1.fr.

Accepted: 4 July 2011

Abstract

The definition of bacterial species is based on genomic similarities, giving rise to the operational concept of genomic species, but the reasons of the occurrence of differentiated genomic species remain largely unknown. We used the *Agrobacterium tumefaciens* species complex and particularly the genomic species presently called genomovar G8, which includes the sequenced strain C58, to test the hypothesis of genomic species having specific ecological adaptations possibly involved in the speciation process. We analyzed the gene repertoire specific to G8 to identify potential adaptive genes. By hybridizing 25 strains of *A. tumefaciens* on DNA microarrays spanning the C58 genome, we highlighted the presence and absence of genes homologous to C58 in the taxon. We found 196 genes specific to genomovar G8 that were mostly clustered into seven genomic islands on the C58 genome—one on the circular chromosome and six on the linear chromosome—suggesting higher plasticity and a major adaptive role of the latter. Clusters encoded putative functional units, four of which had been verified experimentally. The combination of G8-specific functions defines a hypothetical species primary niche for G8 related to commensal interaction with a host plant. This supports that the G8 ancestor was able to exploit a new ecological niche, maybe initiating ecological isolation and thus speciation. Searching genomic data for synapomorphic traits is a powerful way to describe bacterial species. This procedure allowed us to find such phenotypic traits specific to genomovar G8 and thus propose a Latin binomial, *Agrobacterium fabrum*, for this bona fide genomic species.

Key words: bacterial species, *Agrobacterium*, ecological niche, bacterial evolution, linear chromosome.

Introduction

The species as basic taxonomic unit dates back to Carl Linnaeus and has since been universally used to describe all living organisms, including microbes. In superior Eukaria, the separation of distinct species relies on the occurrence of sexual barriers, as summed up in the famous biological species concept proposed by Mayr (1942). However, in asexually reproducing organisms, species are defined upon similarities of their members contrasted by interspecies genetic discontinuities.

In Bacteria, similarity discontinuities were first revealed through phenotypic traits and used to classify strains in different species by numerical taxonomy (Sneath and Sokal 1973). It was soon discovered that discontinuities also occur at the genomic level, leading to the current genomic species definition. Indeed, empirical studies revealed a gap in the distribution of genomic DNA hybridization ratio for pairwise comparisons of numerous strains around 70% (or around 5 °C for ΔTm) that matched previous phenotype-based distinction of species. Strains displaying genomic similarities

above this level are thus considered to belong to the same species (Wayne et al. 1987; Stackebrandt et al. 2002), that are called genomic species. Alternatively, based on sequence data, genomic species can be distinguished through multilocus sequence analysis (Gevers et al. 2005). This is in line with the phylogenetic species concept based on the evolutionary relatedness among organisms that applies to all organisms, including Bacteria and Archaea, as pinpointed by Staley (2004). Although this definition is operational, efficiently leading to the delineation of readily distinguishable genomic species in most taxa, we still need to understand what mechanisms lead to differentiation of such genomic species (Fraser et al. 2009).

In our view, a genomic species is likely descending from a single ancestor that speciated a long time ago consecutively to adaptations to a novel ecological niche. Adaptations of the ancestor to its ecological niche were determined by adaptive mutations that should have been conserved in the progeny as long as they continued to exploit the same primary niche. Traces of adaptation could thus be found in progeny genomes, namely species-specific genes present in genomes of all members of a given species but not in closely related species. Species-specific genes inherited from the ancestor may still be responsible for the adaptation of present species members to a species-specific ecological niche. This hypothesis can be tested using comparative genomics to reveal species-specific genes that likely encode species-specific ecological functions.

Some studies intended to characterize the genomic specificities of bacterial species and understand their evolutionary history (Porwollik et al. 2002; Cai et al. 2009; Touchon et al. 2009; Lefébure et al. 2010; Zhao et al. 2010), other studies aimed to characterize the differences in ecology of ecotypes (or ecovars) among a taxon (Cohan 2002; Sikorski and Nevo 2005; Johnson et al. 2006; Sanjay et al. 2008; Cai et al. 2009; Connor et al. 2010; Zhao et al. 2010). We aimed to combine these approaches to test the hypothesis of genomic species arising from specific ecological adaptations. Good candidate species to test this hypothesis should preferentially display high within-species diversity, so as to capture the most common species characters, with the least possible divergence from their closest neighbors, thus maximizing the chance of detecting specific determinants. The bacterial taxon *Agrobacterium tumefaciens* fulfils these requirements. According to the current genomic species definition, this taxon displays a too large genomic divergence to be a single species and must be considered as a complex of ten distinct genomic species, currently named genomovar G1 to G9 and G13 (Mougel et al. 2002; Costechareyre et al. 2009). Although, they clearly belong to distinct genomic/genetic lineages, these species have not yet received Latin binomials essentially because they are not easily distinguishable by usual biochemical identification systems. They are, however, bona fide species to test our hypothesis because they are closely related and also have large infraspecies diversity. In addition, agrobacteria are common inhabitants of soils and rhizospheres, with several strains and genomic species commonly found in the same soil samples (Vogel et al. 2003; Costechareyre et al. 2010). Because complete competitor cannot coexist, according to the competitive exclusion principle (Hardin 1960), co-occurring species must be adapted to partly different ecological niches. Hence, often co-occurring and highly diverse *Agrobacterium* species are choice candidates for testing whether genomic species harbor presumptive determinants of a species-specific ecology. In addition, strain C58 of genomovar G8 is completely sequenced (Goodner et al. 2001; Wood et al. 2001), so a set of reference genes is available for classification according to their level of ubiquity within the entire taxon. The genomic sequence of strain H13-3 from *A. tumefaciens* genomovar G1 (Wibberg et al. 2011) has been published at the time of submission of this work, providing another reference to validate our results.

In the present work, we looked for genes that could be involved in the ecological specificity of bacterial species. Because we were able to experimentally determine the set of genes specific to genomovar G8, we focused particularly on genomovar G8 as a model species. We then: 1) manually annotated the functions of genes putatively determining ecological specificities, 2) inferred cellular pathways that may be involved in the adaptation of G8 agrobacteria to their ecological niches, and 3) experimentally validated most of the predicted functions and checked that they specifically occurred within all G8 members but not elsewhere. We used this information to precise our representation of the ecological niches of genomic species of the *A. tumefaciens* complex and develop a scenario for ecology-driven speciation of genomovar G8.

Materials and Methods

Bacterial Strains and Culture Conditions

In the present paper, we used a homogenous nomenclature defined according to the literature as follows: *A. tumefaciens* for members of the species complex with reference to genomic species, as explained by Costechareyre et al. (2010), *A. larrymoorei* for strain AF3.10 (Bouzar and Jones 2001), *A. vitis* for the sequenced strain S4 (Ophel and Kerr 1990; Slater et al. 2009), *Rhizobium rhizogenes* for the sequenced strain K84 (Slater et al. 2009; Velázquez et al. 2010), and *Ensifer meliloti* for the sequenced strain 1021 (Galibert et al. 2001; Martens et al. 2007). Strains of the species complex *A. tumefaciens* and *A. larrymoorei* tested in the study (table 1) are available at the *Collection Française de Bactéries Phytopathogènes* (CFBP, INRA, Angers, France). They were routinely grown at 28 °C on YPG medium (yeast extract, 5 g/l; Bacto Peptone, 5 g/l; glucose, 10 g/l, pH 7.2). Genomic DNAs were extracted and purified from 50-ml liquid YPG cultures using the standard phenol–chloroform method (Sambrook and Russell 2001).

Table 1
Agrobacterium Strains Used in This Study

Strain Name	CFBP Code	Nb of Detected C58 CDS Homologs			
		CcC58	LcC58	pAtC58	pTiC58
Agrobacterium tumefaciens genomovar G1					
CFBP 5771		2493	1392	205	14
ICPB TT111	5767	2461	1394	101	30
A. tumefaciens genomovar G2					
CFBP 5495		2371	1185	53	0
CFBP 5494		2168	944	36	0
A. tumefaciens genomovar G3					
CFBP 6624		2501	1104	0	0
CFBP 6623		2586	1388	32	0
A. tumefaciens genomovar G4 (bona fide *A. radiobacter*)					
B6	2413	2584	1389	131	32
DC07-012	7273	2503	1283	36	0
Kerr 14	5761	2557	1376	3	81
A. tumefaciens genomovar G5					
CFBP 6625		1751	626	1	0
CFBP 6626		2378	1164	9	1
A. tumefaciens genomovar G6					
NCPPB 925	5499	2533	1507	50	36
A. tumefaciens genomovar G7					
DC07-042	7274	2370	1184	5	0
RV3	5500	2516	1266	1	0
Zutra 3/1	6999	2529	1349	22	169
A. tumefaciens genomovar G8 (*A. fabrum* nov. sp.)					
Mushin 6	6550	2686	1693	273	132
C58T	1903	2765	1851	542	197
DC04-004	7272	2757	1851	542	197
J-07	5773	2683	1677	200	0
LMG 46	6554	2674	1669	0	172
LMG 75	6549	2681	1736	198	117
T37	5503	2663	1678	270	134
A. tumefaciens genomovar G9					
Hayward 0362	5507	2565	1195	7	197
Hayward 0363	5506	2524	1213	17	0
A. tumefaciens genomovar G13					
CFBP 6927		2517	1215	10	0
A. larrymoorei					
AF 3.10T	5473	1032	228	8	0

NOTE.—CFBP, Collection Française de Bactéries Phytopathogènes, INRA, Angers, France (http://www.angers.inra.fr/cfbp/). CcC58, circular chromosome of C58; LcC58, linear chromosome of C58; pAtC58, At plasmid of C58; pTiC58, Ti plasmid of C58.

Comparative Genome Hybridization Array Design

Comparative genome hybridizations (CGHs) were performed with DNA microarrays specifically designed for this experiment. Microarrays were made of 389,307 spots of 50-nt probes. All the four replicons of *A. tumefaciens* str. C58 (GenBank accessions NC_003302, NC_003303, NC_003304, NC_003305) were covered with a probe every 50 nt on each strand and a shift of 25 nt between strands. To obtain a set of control genes known to be absent from the tested strains, the microarrays also included probes designed in the same way to cover some plasmids from diverse Rhizobiaceae members, corresponding to the following GenBank accessions: NC_002377 (pTiA6), DQ058764 (pTiBo542), NC_002575 (pRi1724), NC_006277 (pAgK84), AJ271050 (pRi2659), AF065242 (pTiChry5), and plasmids from *R. etli* str. CFN42: NC_007 762, NC_007763, NC_007764, NC_004041, NC_007765, NC_007766. In order to model hybridization intensities as a function of levels of DNA base pairing between probe and target DNAs, the microarray contained 50-nt probes designed on the direct strand of all alleles of *mutS*, *recA*, and *gyrB* genes known at that time in *A. tumefaciens*, *A. rubi*, *A. larrymoorei*, *A. vitis*, and in some remote Rhizobiaceae species including *R. rhizogenes* and *E. meliloti*. The microarray also included 39,746 constructor-designed random probes for hybridization control. The C58 whole-genome microarray was constructed by NimbleGen Systems Inc. (Madison, WI), which also performed DNA labeling, hybridization, image capture, and raw data extraction steps according to internal company procedures. Hybridization intensities considered hereafter are $\log_2$ transformations of the raw data delivered by the company. Microarray design and experimental raw data are available at http://www.ebi.ac.uk/arrayexpress/ under the accessions A-MEXP-1977 and E-MTAB-558, respectively.

Modeling of Probe Hybridization Behaviors

Hybridization intensity (I) ranged, approximately, from 6 to 16 arbitrary units, including a long range (6–9) for background noise. Even in case of a perfect match, I spanned over a long range (e.g., 8–16 with C58). This complicated the determination of a single presence/absence threshold value indistinctly valid for all probes, especially for strains distantly related to C58. Instead, we used the fact that lacking genes are characterized by long stretches of successive probes mostly delivering a low background signal. Thus, to detect C58 coding DNA sequences (CDSs) homologs in the tested strains, we developed a method to classify segments of C58 replicons according to the homogenous presence or absence of homologous segments in each tested strain by comparison with perfectly matching C58 DNA probes as positive control. For each replicon a, plots of probe hybridization intensities of tested strain i (denoted $I_{a.i}$) and reference C58 DNA (denoted $I_{a.C58}$) revealed the presence of two populations of points: one, which displayed an almost linear relationship between $I_{a.i}$ and $I_{a.C58}$, corresponding to probed regions that were "present" in the tested strain; and another, that displayed no correlation between $I_{a.i}$ and $I_{a.C58}$, corresponding to regions that were "absent" (supplementary fig. S3, Supplementary Material online). A model (M) fitting these conditions was constructed using a mixture of two linear models, that is, (A) (absent) and (P) (present):

$$(\mathrm{A}) : I_{a.i} \text{ follows } a \text{ law } N(mA; sdA)$$

$$(P): I_{a\cdot i} \text{ follows a law } N\left(mP + \rho \cdot I_{a\cdot C58}; sdP\right)$$

$$(M): I_{a\cdot i} \text{ follows a law } p\cdot N\left(mA; sdA\right) + (1-p)\cdot N\left(mP + \rho \cdot I_{a\cdot C58}; sdP\right)$$

where parameters {*mA*, *sdA*} and {*mP*, *sdP*} are means and standard deviations for normal models (A) and (P), ρ is the slope of the linear relationship between present $I_{a\cdot i}$ and $I_{a\cdot C58}$, and *p* is the weight of model (A) in (M), which reflects the proportion of probes belonging to the absent population.

For each strain *i*, given the set of probes representing a C58 replicon *a* on the microarray, it is straightforward to compute a likelihood function on the basis of this modeling. Then, we looked for the maximum likelihood given this data using the method of Nelder and Mead (1965) as implemented in the "stats" R package (R Development Core Team 2009). This process optimized values for the parameters in three steps:

1. First, parameters {*mA*, *sdA*} were analytically computed from the means and standard deviations of two sets of control probes: a) on probes covering NC_004041 (p42d) which was absent from every tested strain, giving values {mA_a, sdA_a} or b) on constructor-designed random probes, giving values {mA_b, sdA_b}.
2. Secondly, parameters were optimized from both start points {$p = 0.5$, $mA_0 = mA_x$, $sdA_0 = sdA_x$, $mP = 1$, $sdP = 1$, $\rho = 1$}, with $x = a$ or $x = b$. During this first optimization step, {*mA*, *sdA*} were fixed in order to find the (P) mode. The posterior likelihood of models was calculated for each set of optimized parameters and the best fit between both sets of optimized parameters *a* and *b* was kept.
3. To adjust the proportions of points recruited by each mode, parameters were again optimized with {*mA*, *mP*, ρ} fixed and with a constraint on *sdA*: $sdA \leq (1.05\cdot sdA_0)$. To maintain the exclusivity of (A) and (P) modes, an additional constraint was set in the case of plasmids NC_003604 and NC_003605: $mA + 1.5.\ sdA + 11.\ \rho \leq mP + 2\cdot sdP$.

{*mA*, *sdA*} parameters were constrained during steps 2) and 3) to avoid a side effect of optimization due to the non-exclusivity of both modes, which may lead to overrecruitment of present points in absent mode (A) in some instances. When mode (A) should recruit very few points, that is, for tested strains very closely related to C58, a greedy optimization algorithm was tented to fill mode (A) by enlarging its boundaries (i.e., increasing *sdA*) or shifting its mean *mA* toward present points or conversely to fill mode (P) when plasmids NC_003604 and NC_003605 were completely absent from a strain. Sets of parameters for each microarray are listed in supplementary table S7 (Supplementary Material online).

Segmentation of C58 Replicons into Regions Present/Absent in Tested Strains

Multiple prediction partitioning was performed using Sarment Python modules (Guéguen 2005) to build an incremental partitioning of the sequence of replicon *a* when hybridized with strain *i* into segments of consecutive probes of common Absent or Present state given the likelihoods of each probe by models (A) and (P) nested in the optimized model (M). The segmentation process was independent of the sequence annotation; however, it appeared that partitions generally occurred between CDSs. As our interest was to screen for C58 CDSs present or absent in other strains, the incremental process was stopped when the number of CDSs which 100% probes mapped in absent segments was stabilized, typically after a few hundred segmentation iterations. Note that as a result of the segmentation procedure, some probes with high hybridization values surrounded by large number of low hybridization value probes can occur within absent segments. All CDSs located in absent segments are nevertheless considered absent.

Estimate of Genomic DNA-Probe Similarities

Similarity between microarray probes and probed DNA was estimated via probes of alleles of marker genes *gyrB*, *mutS*, and *recA* spotted on the microarray. The actual nucleotidic similarities between probes and known sequences of marker genes of the probed strain were computed using BlastN. The results were imported and parsed using Biopython libraries (Cock et al. 2009). Linear regressions between hybridization intensities and actual nucleotide similarities were done for each microarray while excluding nonhybridized probe noises (empirically determined to be below 80% genomic DNA-probe match) by using the stats R package (R Development Core Team 2009). Linear models (supplementary table S8, Supplementary Material online) were used to estimate the similarity between hybridized DNA and microarray probes (estimated nucleotidic similarity [ES]), thus allowing the calculation of average ESs of all CDSs covered by probes on the microarray (supplementary table S2, Supplementary Material online). For C58 replicons, it was also possible to cope with intensity heterogeneities among CDSs by calculating weighted estimated nucleotidic similarities (WESs). ESs recorded with a given strain were thus divided by the corresponding values obtained with C58, then adjusted according to the actual similarities of sequenced polymerase chain reaction (PCR) products of the tested strain with the C58 genome (supplementary table S3, Supplementary Material online).

Codon Usage Analysis

Effective counts of the 64 codons of the 5,355 CDSs of C58 were calculated using the "seqinR" package from R project (Charif and Lobry 2007). Correspondence

analyses were performed and projected using "dudi.coa" and "s.class" functions from the "ade4" package (Dufour and Dray 2007).

Strain Clustering

Agrobacterium tumefaciens strains were clustered on the basis of gene presence/absence characters, as described in Lake (1994), by using logdet distances with C58 as conditioning genome. Logdet/paralinear distances (Lake 1994) were computed using the "binary.dist" function from R project (R Development Core Team 2009). Trees were built using the NEIGHBOR algorithm from PHYLIP package (Felsenstein 1993).

Functional Annotation of Specific Genes

The functional annotation of CDSs included in G8-specific clusters was manually curated using a relational database, that is, AgrobacterScope (in open access at https://www.genoscope.cns.fr/agc/microscope), with the MaGe web interface (Vallenet et al. 2009).

Construction of Deletion Mutants of SpG8-Specific Clusters

Mutants were constructed by mutagenic PCR as described by Choi and Schweizer (2005). Briefly, mutagenic PCR fragments were created by joining three fragments corresponding to the two regions flanking the sequence to be deleted of C58 (ca. 1 kb each) and a fragment encoding the *nptII* kanamycin resistance gene amplified from plasmid pKD4 (Datsenko and Wanner 2000) by using 70-nt primers consisting of 20 nt priming the kanamycin-resistance gene (3′ segment of the primer) and 50 (±3) nucleotides corresponding to flanking sequence ends of targeted sites (5′ segment of the primer) (supplementary table S9, Supplementary Material online). First and second round PCRs were performed as in Choi and Schweizer (2005), and then PCR fragments were cloned into the pGEM-Teasy vector (Promega, Madison, WI) according to manufacturer's instructions. After digestion of the resulting plasmids with *Apa*I and *Spe*I, fragments were subcloned into pJQ200SK, a plasmid carrying the *sacB* gene conferring sucrose sensitivity (Quandt and Hynes 1993) digested with the same enzymes. To generate deleted mutants, PCR fragments cloned into pJQ200SK were inserted in C58 by electroporation. Single recombinants were selected on YPG media containing 25 μg/ml kanamycin and 25 μg/ml neomycin. Double crossover events were identified by sucrose resistance on YPG media supplemented with 5% sucrose. Deletion mutants were verified by diagnostic PCR with appropriate primers.

Experimental Validation Assays

Ferulic acid catabolism was tested using the two-step procedure described by Civolani et al. (2000). In a first step, cells were induced for 24 h at 28 °C (optical density $[OD]_{600\ nm} = 1$) in AT minimal medium (Petit et al. 1978) supplemented with 10 mM $(NH_4)_2SO_4$ and 10 mM succinic acid, and 0.52 mM (0.1 mg·mL^{-1}) ferulic acid (Sigma-Aldrich, St Louis, MO). Cells harvested by centrifugation were then suspended at $OD_{600\ nm} = 0.1$ into AT medium containing 10 mM $(NH_4)_2SO_4$ and 0.52 mM ferulic acid as sole carbon source. Ferulic acid disappearance was monitored by high-performance liquid chromatography (HPLC) performed on an Agilent 1200 series (Agilent Technologies, Santa Carla, CA) liquid chromatograph associated with a diode array detector. Data acquisition and processing were controlled via Agilent Chemstation software. The separations were carried out on a Kromasil 100-5C18 column (250 × 4.6 mm). Compounds were eluted with a methanol–water gradient (0.4% formic acid) in which the methanol concentration was varied over time as follows: from 0 to 5 min, 20%; 5 to 22 min, increased to 62%; 22 to 25 min, increased to 100%; 25 to 30 min, 100%; 30 to 31 min, decreased to 20%. The flow rate was 1 ml.min^{-1}. Ferulic acid was detected at a wavelength of 320 nm after injection of 5-μl sample. UV spectra and retention time (12.43 min) of ferulic acid were determined by injection of a methanolic suspension of ferulic acid. Identification of ferulic acid in bacterial cultures was confirmed by comparison with this standard.

Curdlan production was assessed by streaking bacteria onto plates containing a modified Congo red medium adapted from Kneen and LaRue (1983) with glucose as sole carbon source, incubated at 28 °C for 48 h and then kept at room temperature for 48 h.

Results

Presence of Homologs of CDSs from C58 Replicons and Other Rhizobiaceae Plasmids

CGH results obtained with an original C58 genome–based microarray were used to detect the presence or absence of genes homologous to C58 in 25 agrobacterial strains. These strains included seven G8 members, one to three for each of the nine other genomic species of the *A. tumefaciens* complex and one for *A. larrymoorei*, a sister species of *A. tumefaciens* (table 1). An original probabilistic method was used to segment C58 replicon sequences into regions that were absent or present in the tested strains, thus allowing us to detect the presence of homologs of C58 CDSs in the tested strains (supplementary table S1, Supplementary Material online).

The absence of detection, however, might have been due to the absence of a real locus or to a weak hybridization signal caused by high sequence divergence between the target genome and C58 DNAs. We thus calculated the ES for CDSs of all replicons probed by the microarray in reference to internal control probes of known mismatch values with tested DNAs (supplementary table S2, Supplementary Material

online). In addition, we observed strong intensity heterogeneities among CDSs, even in case of a perfect match with C58. A better similarity estimate was reached by weighting ESs of tested strains by ESs recorded with C58 to provide WESs (supplementary table S3, Supplementary Material online). We used the segmentation method to detect the presence of CDSs displaying WESs as small as 80–86% in G1-ICPB TT111 and G7-DC07-042, respectively, in spite of high hybridization background noise (supplementary table S4, Supplementary Material online). This demonstrated the higher sensitivity of our probabilistic approach over threshold methods.

Beyond the analysis in terms of CDS presence, WES allowed us to estimate whether C58 and the tested strains had divergent or identical alleles. WES measures were plotted against CDS positioned along replicons for G8 members (fig. 1). G8-DC07-004 was expected to belong to the same clone or at least the same clonal complex as C58 because both strains have identical alleles for marker genes located on circular and linear chromosomes (*recA*, *mutS*, *gyrB*, *chvA*, *ampC*, *glgE*, *gltD*, ...) (data not shown). G8-DC07-004 displayed an average WES of 100% with little dispersion (average ΔWES = $\pm$1%) over all four replicons and thus appeared to be identical to C58, except for eight genes that were lacking on the circular chromosome. In contrast, G8-T37, which had different alleles for most marker genes, displayed an average WES clearly below 100% with greater dispersion (average ΔWES = $\pm$4%). In contrast, this allowed us to discover that large regions of the C58 linear chromosome were likely identical in other G8 strains. Remarkably, a region of more than 1 Mb encompassing the left arm of the linear chromosome was identical between C58 and G8-LMG 75 or G8-Mushin 6 (fig. 1). This strongly suggests recent transfer of half of the linear chromosome between members of this species. Although we found such long regions of identity with C58 in several G8 members, such long stretches of genome identity with C58 were not found outside G8 (supplementary fig. S1, Supplementary Material online), suggesting that large transfer events may essentially concern members of the same species.

Strains were clustered according to their C58 CDS homolog content. As expected, this clearly allowed significant clustering of all G8 members (fig. 2). Differences in gene content similarities according to chromosomes were observed between genomic species. For six genomic species, the different members significantly grouped when considering CDSs of the C58 linear chromosome, compared with only four grouped genomic species when considering CDSs of the C58 circular chromosome. This suggests that the linear chromosome content better characterizes genomic species than the circular chromosome. Remarkably, G2 members as well as G5-CFBP 6625 were located at a basal position (i.e., far from G8), indicating that they had the lowest number of CDS in common with C58. The remaining species branched together at the same distance from the G8 groups (forking branches), indicating that they had comparable numbers but different sets of C58 CDS homologs.

The presence of C58 CDS homologs according to their location in C58 replicons confirmed the presence of C58-circular and -linear chromosome CDS homologs in all tested strains (table 1). In contrast, this revealed the complete lack of pTiC58 or pAtC58 homologs in several strains such as for G8-J-07 and G8-LMG 46, which respectively lacked Ti and At plasmids or the absence of large regions of C58 plasmids in numerous strains (supplementary table S1 and fig. S1, Supplementary Material online), which highlights the mosaic nature of these replicons. The segmentation method was not applicable for replicons outside C58, thus hampering detection of barely similar CDS homologs in these cases. Nevertheless, high CDS homologies of around 100% ES were recorded for all CDSs of pTiA6 for both G4-B6 and G1-ICPB TT111, indicating the likely presence of the same Ti plasmid in both strains (supplementary table S2 and fig. S2, Supplementary Material online). The results of the CDS presence/absence analysis were, however, related to the high hybridization stringency conditions used, which may not allow detection of barely similar homologs. For instance, *A. larrymoorei* AF 3-10 was found to have no detectable CDS homology with pTiC58 (table 1), although this strain is known to be pathogenic with a chrysopine type Ti plasmid (Vaudequin-Dransart et al. 1995). As expected, however, AF3.10 exhibited significant estimated similarity values (ca. 93%) with more similar CDSs of the chrysopine type Ti plasmid pTiChry5 (supplementary table S2 and fig. S2, Supplementary Material online).

Ubiquity Level of C58 CDSs in the *A. tumefaciens* Complex

Homologs of the 5,355 CDSs of C58 were classed according to their level of ubiquity in *Agrobacterium* strains and grouped in six classes: only in C58 ("specific to C58," 166 genes); only in G8 strains but not all ("sporadic in G8," 151 genes); in all G8 and only G8 strains ("specific to G8," 196 genes); with no specific presence pattern in *A. tumefaciens* ("sporadic in *At*," 2,846 genes); in all *A. tumefaciens* strains but not in *A. larrymoorei* ("specific to *At*," 976 genes); or in both *A. tumefaciens* and *A. larrymoorei* ("*At-Al* core genome," 1,020 genes) (supplementary table S1, Supplementary Material online).

The core genome of *A. tumefaciens* ("*At* core genome," sum of the "specific of *At*" and *At-Al* core genome classes) consists of 1,996 genes (37% of the genome). Seventy-five percent and 25% of the *At* core genome are located on circular and linear chromosomes, respectively (accounting for 56% and 25% of these replicons, respectively), showing clear core genome enrichment on the circular chromosome (fig. 3). As expected, no part of the core genome was found on plasmids because these accessory replicons were lacking in some strains (table 1).

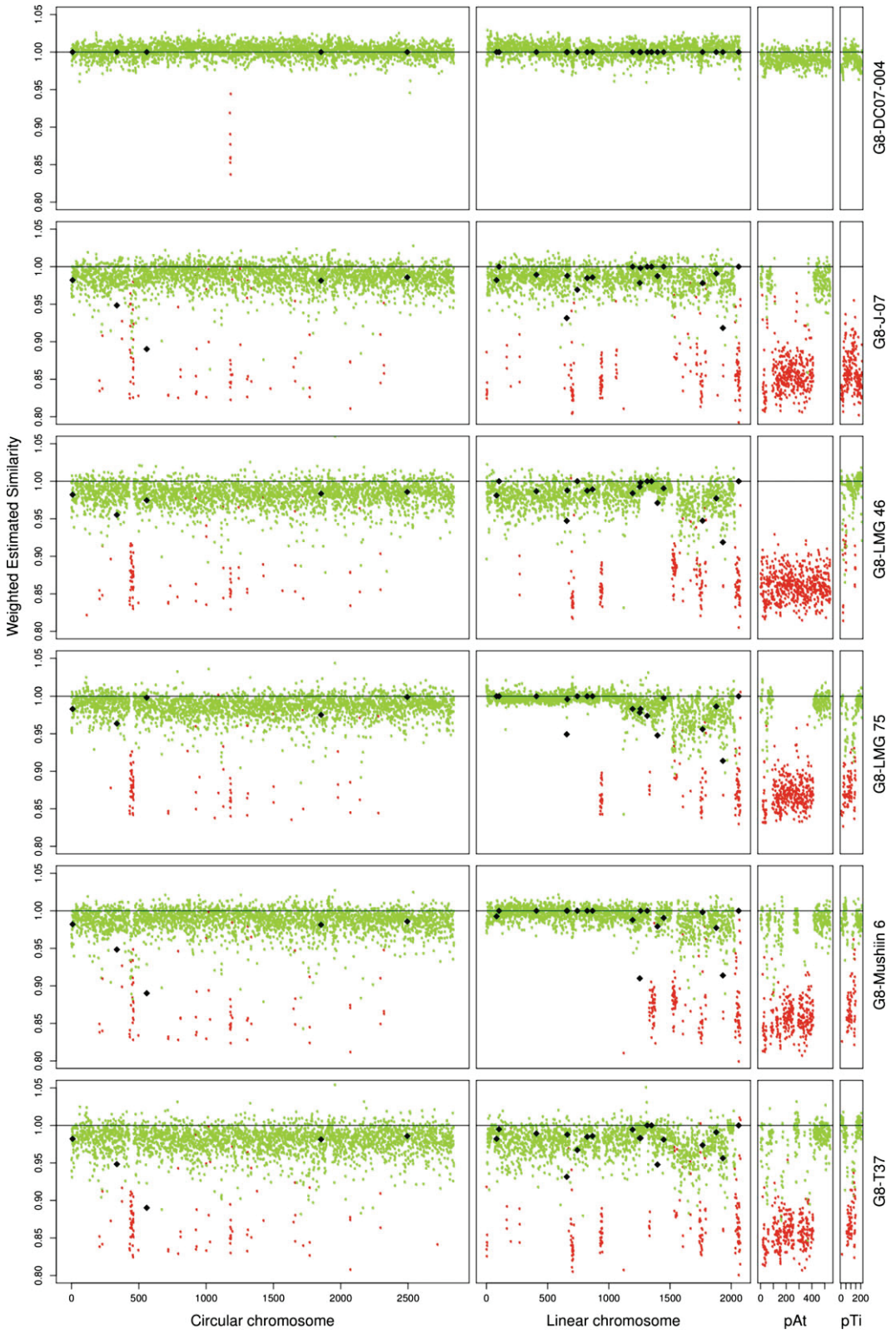


Fig. 1.—Presence and estimated similarity of C58 CDS homologs in other genomovar G8 members. Percentage of WES of C58 CDS homologs were plotted against their coordinates on the four C58 replicons. Dot colors indicate the presence (green) or absence (red) of C58 CDS homologs. Diamonds indicate actual similarity values of sequenced PCR products.

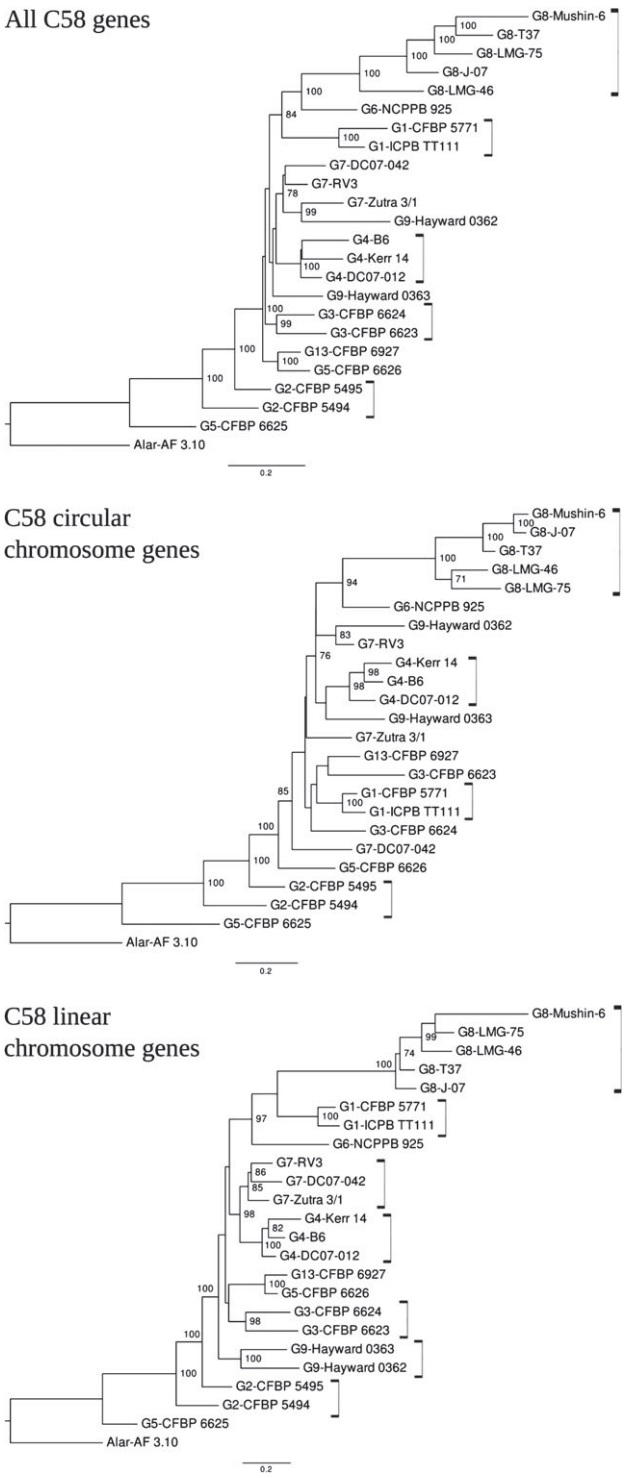


Fig. 2.—Clustering of *Agrobacterium tumefaciens* strains based on absence/presence of C58 CDS homologs. Neighbor-joining trees were constructed using the paralinear distances of Lake (1994) calculated from the presence/absence of C58 CDSs in other strains. With the reference genome being C58, this strain and its nearly identical relative DC07-004 were excluded from the analysis.

Genes Specific to C58 and Sporadic in G8. G8-DC07-004 was found to be the same as C58 except for eight CDSs (Atu1183–Atu1190), which were absent in G8-DC07-004 (table 1). These eight CDSs were, strictly speaking, the real C58-specific genes, whereas the remaining 158 genes specific to C58 were specific to both C58 and G8-DC07-004. These genes were mainly grouped in three clusters: Atu1183-Atu1194, Atu4606-Atu4615, and Atu4864-Atu4896. The Atu1183–Atu1194 region, which included the deleted region described above, was located on the circular chromosome. It constituted a prophage (ATPP-2, see below). The last two clusters were located at the right extremity of the linear chromosome close to the telomeric region. They harbored transposase genes, suggesting that they were recombinogenic. The Atu4606–Atu4615 region was annotated as being involved in lipopolysaccharide biosynthesis, a function likely gained by transfer and specific to C58 and G8-DC07-004.

In addition, four regions sporadic in G8 were also identified as probable mobile genomic elements. Three were evidently prophages, which we named *A. tumefaciens* prophages (ATPP): ATPP-1 (Atu0436–Atu0471), ATPP-2 (Atu1183–Atu1194), and ATPP-3 (Atu3831–Atu3858), which contained genes encoding proteins characteristic of prophages, such as: integrase, excisionase, resolvase, DNA methyl-transferase, phage tail structural and assembly proteins, and DNA-dependent RNA polymerase. By analyzing similarities in their integrase gene sequences with those available in databanks, prophages were assigned to known prophage families: ATPP-1 and ATPP-2 were related to *Podoviridae* of P22 and T-7 families, respectively, whereas ATPP-3 was related to *Myoviridae* of the P4 family. The very recent publication of the genomic sequence of strain H13-3 of genomovar G1 (Wibberg et al. 2011) showed that prophages ATPP-1 and ATPP-2 were absent from this strain, but that traces of their past presence could be found at the corresponding loci. The fourth region, ranging from Atu3636 to Atu3665 next to tRNA genes, had a less clear nature. It was apparently undergoing a process of genetic decay and was referred to in this study as a decaying mobile DNA region (DMR). Many CDSs in this region were short CDSs that coded for hypothetical proteins, pseudogenes, and gene remnants, indicating that this region might no longer be under selection pressure.

Genes Specific to G8. In fact, 51 CDSs were strictly specific to the genomovar G8, but 145 CDSs found in all G8 members were also detected in one or two other non-G8 strains. In several instances, those latter genes were contiguous to strictly G8-specific genes (supplementary table S1, Supplementary Material online), suggesting that they cooperate with strict G8-specific genes for their functions. Thus, in order to capture more complete functions, we decided to use a loose definition of species-specific genes by merging the two gene classes for a total of 196 G8-specific CDSs (SpG8) (supplementary table S5, Supplementary Material online). No SpG8 genes were

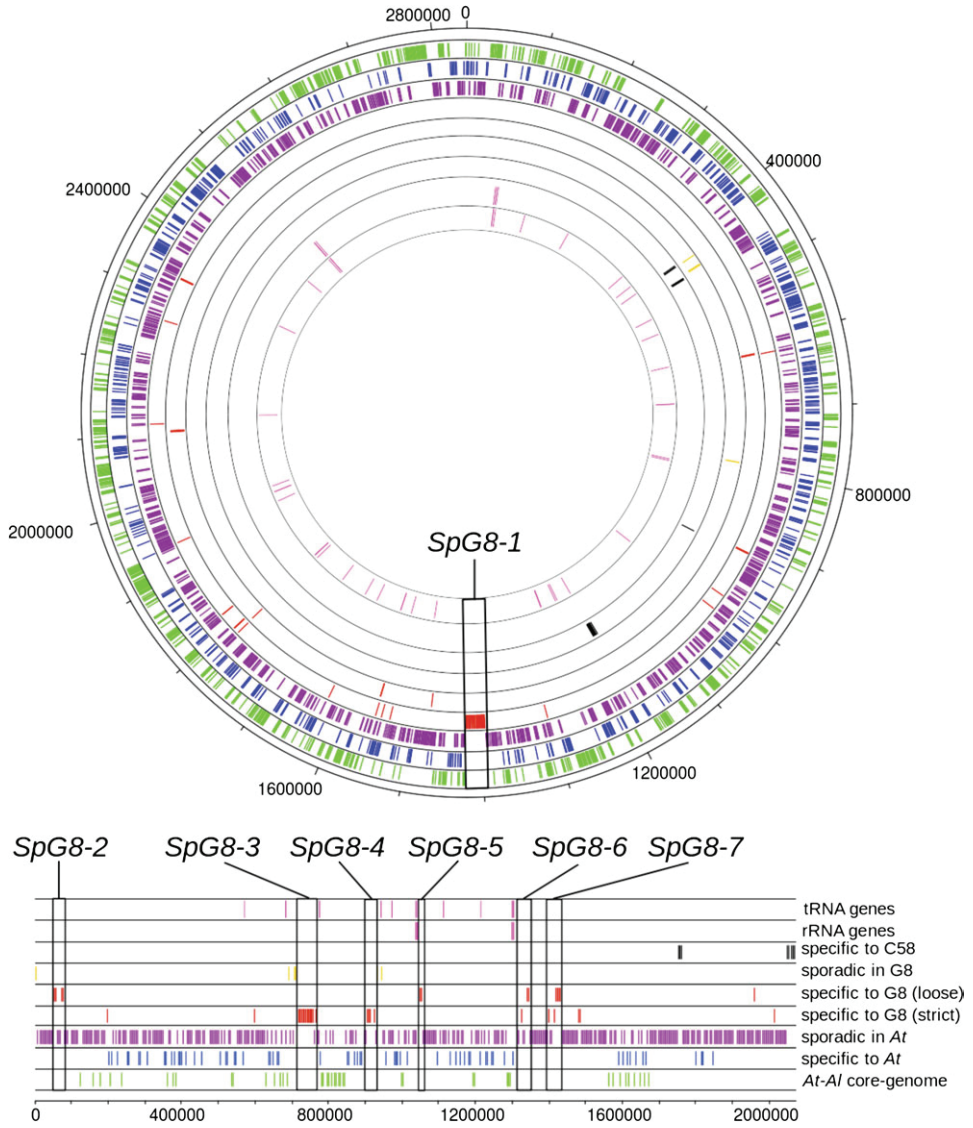


Fig. 3.—Ubiquity in the *Agrobacterium tumefaciens* species complex of C58 CDSs according to their localization on C58 chromosomes. Tracks are numbered from inner to outer track (circular chromosome) or top to bottom track (linear chromosome). tRNA and rRNA genes are represented track 1 and 2, respectively (pink). CDSs are represented according to their levels of ubiquity: track 3, specific to C58 (black); track 4, sporadic in G8 (yellow); track 5, strictly specific to G8, and track 6, specific to G8 with a loose criterion (red); track 7, sporadic in *At* (purple); track 8, specific to *At* (blue); track 9, "*At-Al* core-genome" (green). Boxes indicate G8-specific (SpG8) gene clusters.

found on plasmids, but they were unevenly dispersed on the two chromosomes: 72% on the linear and 28% on the circular chromosomes, respectively. Remarkably, 61% of SpG8 genes were organized into clusters of five or more contiguous CDSs, whereas others were interspersed within the C58 genome (supplementary table S5, Supplementary Material online). Seven large SpG8 clusters, numbered SpG8-1 to SpG8-7, were located either on the circular chromosome (SpG8-1) or on the linear chromosome for the six others (table 2 and fig. 3). As explained below, some SpG8 clusters were subsequently divided into subclusters encoding homogeneous functions. Sequence data validated the presence of SpG8 clusters in G8 members and the absence of SpG8 genes with a similarity above 70% outside G8 (data not shown).

SpG8 regions seem to occur in hotspots of gene insertions. Cluster SpG8-3 adjoins a region containing different types of putative mobile elements referred to here as DMR. SpG8-4 was located next to the putative prophage ATPP-3. Blocks made of SpG8-3 and DMR and SpG8-4 and ATPP-3 are next to tRNA genes. SpG8-5 and SpG8-6 are next to rRNA operons containing tRNA genes (fig. 3).

We performed a correspondence analysis on the codon usage of CDSs in C58 to determine whether genes of different ubiquity classes could be differentiated on the basis of

Table 2
Characteristics of SpG8 Gene Clusters

G8-Specific Regions	C58 CDSs	Region Occurrence Outside G8	Main Predicted Functions	Experimental Validation[a]
SpG8-1a	Atu1398–Atu1408	G6-NCCPB 925	Sugar and amino acid transport; sugar metabolism	Not done
SpG8-1b	Atu1409–Atu1423	G9-Hayward 0362	Ferulic acid uptake and catabolism	Present work
SpG8-2a	Atu3054–Atu3059	*r*	Curdlan EPS biosynthesis	Present work
SpG8-2b	Atu3069–Atu3073	*r*	Secondary metabolite biosynthesis	Not done
SpG8-3	Atu3663–Atu3691	G1-ICPPB TT111 G6-NCCPB 925	Siderophore biosynthesis; iron-siderophore uptake	Rondon et al. (2004)
SpG8-4	Atu3808–Atu3830	G6-NCCPB 925	Ribose transport; monosaccharide catabolism and carbohydrate metabolism	Not done
SpG8-5	Atu3947–Atu3952	*r*	Opine-like compounds catabolism	Not done
SpG8-6a	Atu4196–Atu4206	G1-CFBP 5771	Drug/toxic (tetracycline) resistance	Luo and Farrand (1999)
SpG8-6b	Atu4213–Atu4221	*r*	Drug/toxic resistance	Not done
SpG8-7a	Atu4285–Atu4294	G6-NCCPB 925	Environmental signal sensing/transduction	Not done
SpG8-7b	Atu4295–Atu4307	Not present outside G8	Environmental signal sensing/transduction	Not done

NOTE.—*r*, rare occurrence of some CDSs outside G8.
[a]Deleted mutants of C58 were obtained for all regions.

their DNA sequence composition. Inertia ellipses of ubiquity classes were found to be dispersed along the first axis of the codon usage space over a gradient reproducing the ubiquity class order, which extends from core-genome genes to sporadic/strain-specific genes (fig. 4). Within the ubiquity class corresponding to SpG8 genes, we could sort SpG8 gene clusters along the "core versus sporadic" axis, from the group of isolated SpG8 genes (i.e., not located in clusters, fig. 4, box #0) at the "sporadic-like" extremity, to loci SpG8-1, SpG8-4, SpG8-7 at the "core gene" extremity. Notably, subclusters of large SpG8 gene clusters with different occurrence patterns in *A. tumefaciens* displayed different genome signatures. In fact, SpG8-1a and SpG8-7a, which shared genes with the closest relative of G8, that is, G6-NCPPB 925 (Costechareyre et al. 2010), displayed a more marked "core-like" code usage signature than SpG8-1b and SpG8-7b (fig. 4).

SpG8 Functions

We were able to infer global functions for most SpG8 clusters, strengthening the hypothesis that they correspond to coherent functional units. Our expert manual annotations available in the AgrobacterScope database revealed that SpG8 clusters encoded functional units related to environmental sensing (SpG8-7), secreted metabolite production (SpG8-2a, SpG8-3), detoxification (SpG8-6), and metabolite catabolism (SpG8-1, SpG8-4, SpG8-5) (table 2).

SpG8-7: Environmental Signal Sensing. SpG8-7 encoded functions that could be related to environmental signal sensing and transduction: two mechanosensitive channels of the MscS family and a two-component transduction system with a receptor histidine kinase containing a PAS sensory box and two putative response regulators. Genes encoding a two-component system (Atu4300 and Atu4305) were homologous to *nwsAB* from *Bradyrhizobium japonicum* USDA 110, whose proteins are involved in plant host recognition during the nodulation process (Lang et al. 2008) and to *todST* from *Pseudomonas putida* F1 and *styRS* from *Pseudomonas* sp. VLB120, whose proteins recognize toluene and styrene, respectively, and activate related degradation pathways (Lau et al. 1997; Panke et al. 1998). A more comprehensive block was conserved in synteny with genes from *Parvibaculum lavamentivorans* DS-1 (4 genes, 60.2% amino acid identity on average) and *E. meliloti* 1021 (7 genes, 64% amino-acid identity on average).

SpG8-2a: Curdlan Biosynthesis. Atu3056 in SpG8-2a codes for a putative beta-1,3-glucan synthase (curdlan synthase, CrdS) that is involved in the synthesis of curdlan, an exopolysaccharide. We experimentally verified this function by deleting the whole locus in C58 (i.e., Atu3054–Atu3059). As a result, colonies formed by the mutant C58ΔSpG8-2a did not bind Congo Red, whereas colonies formed by wild-type C58 were red, indicating that the mutant was affected in polysaccharide production (fig. 5*B*). In addition, we found that all G8 members similarly accumulated red dye, in contrast with members of other genomic species (data not shown), which demonstrates that this function is specific to G8.

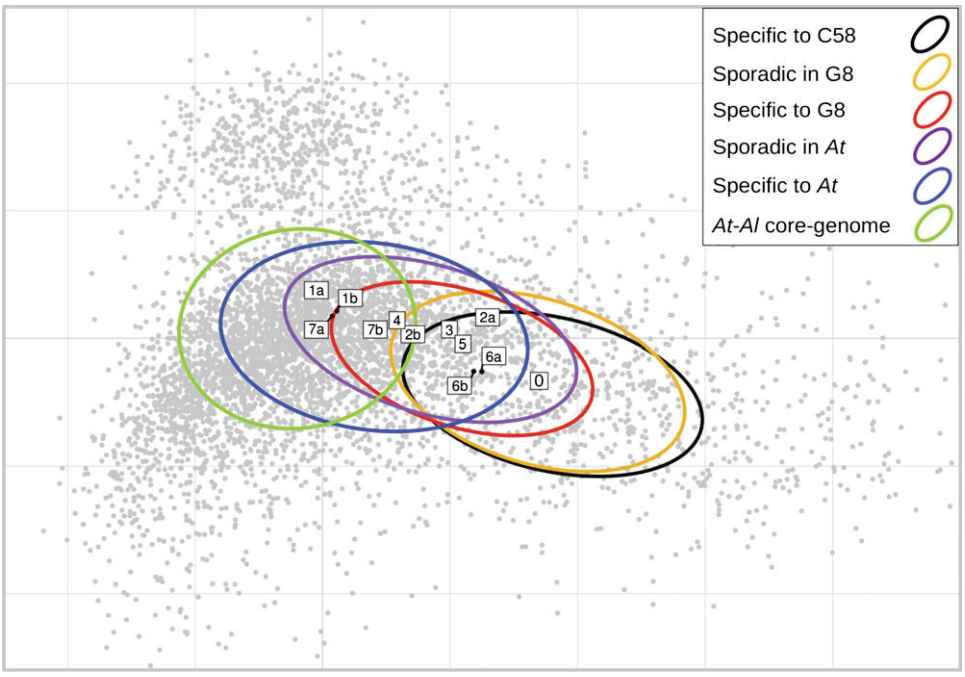


Fig. 4.—Codon usage signatures of SpG8 genes. First factorial plan in the correspondence analysis of C58 CDSs according to their codon usage. First and second axes explain 4.5% and 2.2% of total variance, respectively. Grey dots represent CDSs, ellipses represent the inertia of ubiquity classes, and boxes represent barycenters of SpG8 loci named as detailed in table 2 and supplementary table S6 (Supplementary Material online) (and 0 for interspersed SpG8 CDSs). Codon usage of ubiquity classes were found to gradually vary from core to sporadic genes revealing, in turn, that SpG8 clusters can be distinguished by this criterion from core-like ones to sporadic-like ones. Interestingly, SpG8-1a, SpG8-4, and SpG8-7a—which are shared by the most closely related non-G8 strain G6-NCPPB 925 (table 2)—displayed a core-like codon usage.

SpG8-3: Siderophore Biosynthesis, Release, and Reuptake. The largest SpG8 gene cluster, that is, SpG8-3, ranged from Atu3663 to Atu3693. Nearly, all SpG8-3 genes were also shared by G6-NCPPB 925 and G1-ICPB TT111. This region has already been characterized as coding for functions involved in siderophore biosynthesis in C58 (Rondon et al. 2004). It includes eight polypeptides that may form a mega-enzyme complex corresponding to seven nonribosomal peptide synthase (NRPS) modules and three polyketide synthase modules. An isolated NRPS gene (Atu3072) located at the remote locus SpG8-2 may also interact with this mega-enzyme complex. Genes for transporter proteins were also located in SpG8-3: Atu3669 coding for a transporter of the multidrug extrusion transporter family MATE, that includes proteins involved in secondary metabolite transport (Moriyama et al. 2008), and thus is perhaps involved in siderophore release in the medium; and Atu3684–Atu3691 that are homologous to *fecABCDE* genes involved in TonB-dependent reuptake of the siderophore when it is chelated to iron (Braun et al. 2006). Finally, Atu3684, Atu3692, and Atu3693 (*fecAIR*) seemed to form a cell surface signaling system, whose homolog in *Escherichia coli* was proposed to regulate the whole system of biosynthesis, release, and reuptake of the siderophore (Braun et al. 2006). We noted that the whole region was conserved in synteny with *A. vitis* S4 (30 genes, 51.2% amino-acid identity on average) on its larger plasmid.

SpG8-6: Detoxification. SpG8-6 (Atu4196–4221) contained three putative multidrug transporter systems. Interestingly, one of them (*tetR-tetA*, Atu4205-Atu4206) was experimentally characterized for tetracycline resistance in G8-C58 and G8-T37 (Luo and Farrand 1999). These authors did not detect this resistance in several other agrobacteria, and some of their genomic species assignments are now known: G4-B6, G4-ATCC15955, and G1-Bo542, thus confirming the G8 specificity of this genomic region. Tetracycline is, however, not the natural inducer of these genes (Luo and Farrand 1999). The TetR-TetA efflux pump system might allow for detoxification of other unknown compounds.

SpG8-4, SpG8-1a, SpG8-5: Carbohydate Catabolism. Among regions involved in carbohydate catabolism, SpG8-4 (Atu3808–3830) seems to constitute a functional unit dedicated to monosaccharide uptake, via the putative ribose-specific ABC transporter encoded by $rbsAC_1C_2B$ genes, and sugar metabolism involving putative enzymatic functions such as rhamnose mutarotase or D-galactarate dehydrogenase. Four LysR-type transcriptional regulators were found within this locus, which could be involved in substrate-

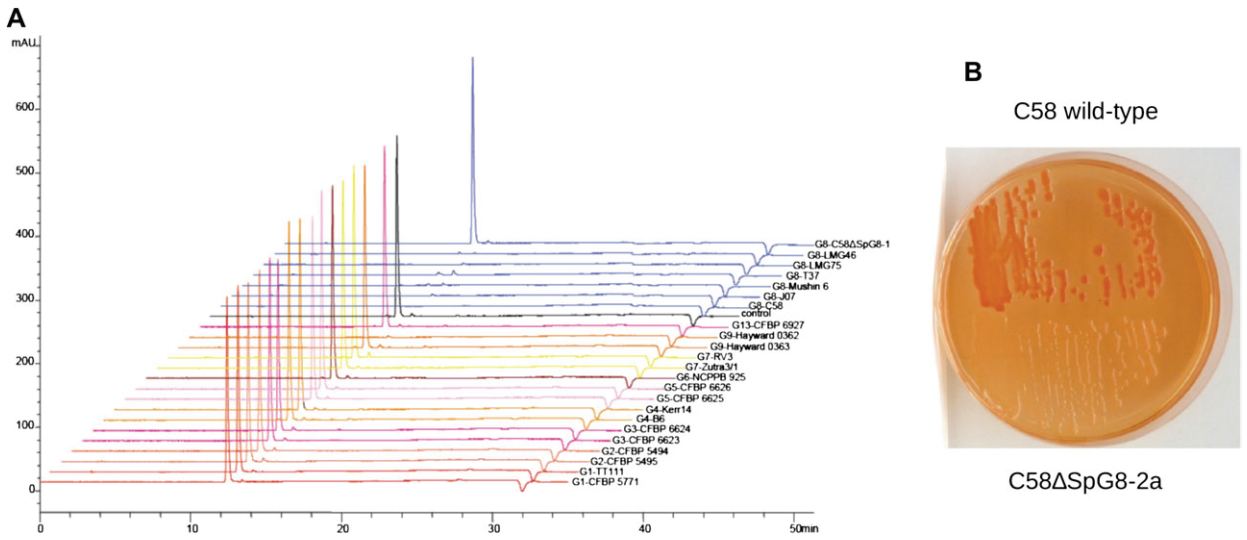


Fig. 5.—Experimental evidences of G8-specific phenotypes determined by SpG8 loci. (*A*) Ferulic acid degradation by *A. tumefaciens* strains determined by HPLC and UV spectrum at 320 nm. mAU, milli absorbance units. All genomovar G8 members were able to catabolize all ferulic acid in 12 h, contrary to non-G8 strains lacking SpG8-1b. (B) Curdlan production revealed by red dye on Congo red medium. C58: Agrobacterium tumefaciens wild-type strain C58 (red colonies), C58ΔSpG8-2a: SpG8-2a-deleted mutant (white colonies).

dependant regulation of metabolic pathways (Maddocks and Oyston 2008).

SpG8-1a (Atu1398–1409) also seemed to be involved in sugar metabolism with an ABC transporter operon putatively specific to monosaccharides and genes for glycolate catabolism enzymes. Two other ABC transporter operons were located alongside within SpG8-1, which homologs have been described to specifically import amino acids. These include genes homologous to *braBCDEF* genes from *R. leguminosarum bv. viciae* 8401 that are involved in branched-chain amino acid uptake (Hosie et al. 2002).

SpG8-5 (Atu3947–Atu3952) encodes enzymes similar to sarcosine oxidase, ornithine cyclodeaminase, and an alanine racemase with a lectine-like sugar-binding domain likely involved in the catabolism of opine-like compounds. Opines are condensates of an amino acid and a sugar or a cetonic acid that are well known to be involved in the ecology of plant pathogenic *Agrobacterium* (Vaudequin-Dransart et al. 1995). However, the annotation is not precise enough to ascertain the substrate molecule class. It is thus possible that the concerned substrates belong to another class of condensates of amino acids and sugars called Amadori compounds—a class of molecules produced in decaying plant material and thus common in humic soil.

All functions found in SpG8-1, SpG8-4, SpG8-5 are thus likely to confer G8 agrobacteria a general ability to metabolize sugar and/or opine/Amadori-like compounds. However, many functional annotations are made on the basis of protein similarities with databases. In the case of sugar-binding proteins and ABC transporters, protein families contain many sequences, only a few of which are characterized. Incidentally, although we obtained deletion mutants of these three regions, we could not yet assign precise candidate substrates to improve the annotation or to experimentally verify the predicted functions.

SpG8-1b: Phenolic Catabolism. SpG8-1b (Atu1409–Atu1423) was shared by all G8 strains and also by G9-Hayward 0362. This locus was involved in phenolic catabolism (fig. 6, supplementary table S6, Supplementary Material online for details on homology relationships). Indeed, SpG8-1b includes a gene homologous to *fcs* (Atu1416), which is involved in a pathway for CoA-dependent, non-beta-oxidative degradation of ferulic acid in *Pseudomonas* (Overhage et al. 1999; Plaggenborg et al. 2003; Calisti et al. 2008), and other putative enzymatic functions that could be related to the same metabolic pathway, including: an enoyl-CoA hydratase (Ech) (Atu1417), a feruloyl-CoA dehydratase (Fcd) (Atu1414), a tetrahydrofolate-dependent vanillate O-demethylase (LigM) (Atu1420), and a methylenetetrahydrofolate reductase (MetF) (Atu1418), as well as substrate-binding regulators VanR (Atu1419) and FerR (Atu1422). Indeed, we were able to reconstruct a complete ferulic acid degradation pathway (fig. 6*B*) and to propose a transcriptional regulation scheme (fig. 6*C*) in C58—and in other G8 strains as well—thanks to the presence of a gene nonspecific to G8 in the linear chromosome, that is, Atu4645 (*vdh*), encoding vanillin oxidase. The final product of this putative pathway was protocathechuic acid, which can be degraded into metabolites suitable for complete oxidation through the tri-carboxylic acid cycle (Parke 1995). In addition, the SpG8-1 gene Atu1415 encoded a putative n-phenylalkanoyl-CoA dehydratase. This

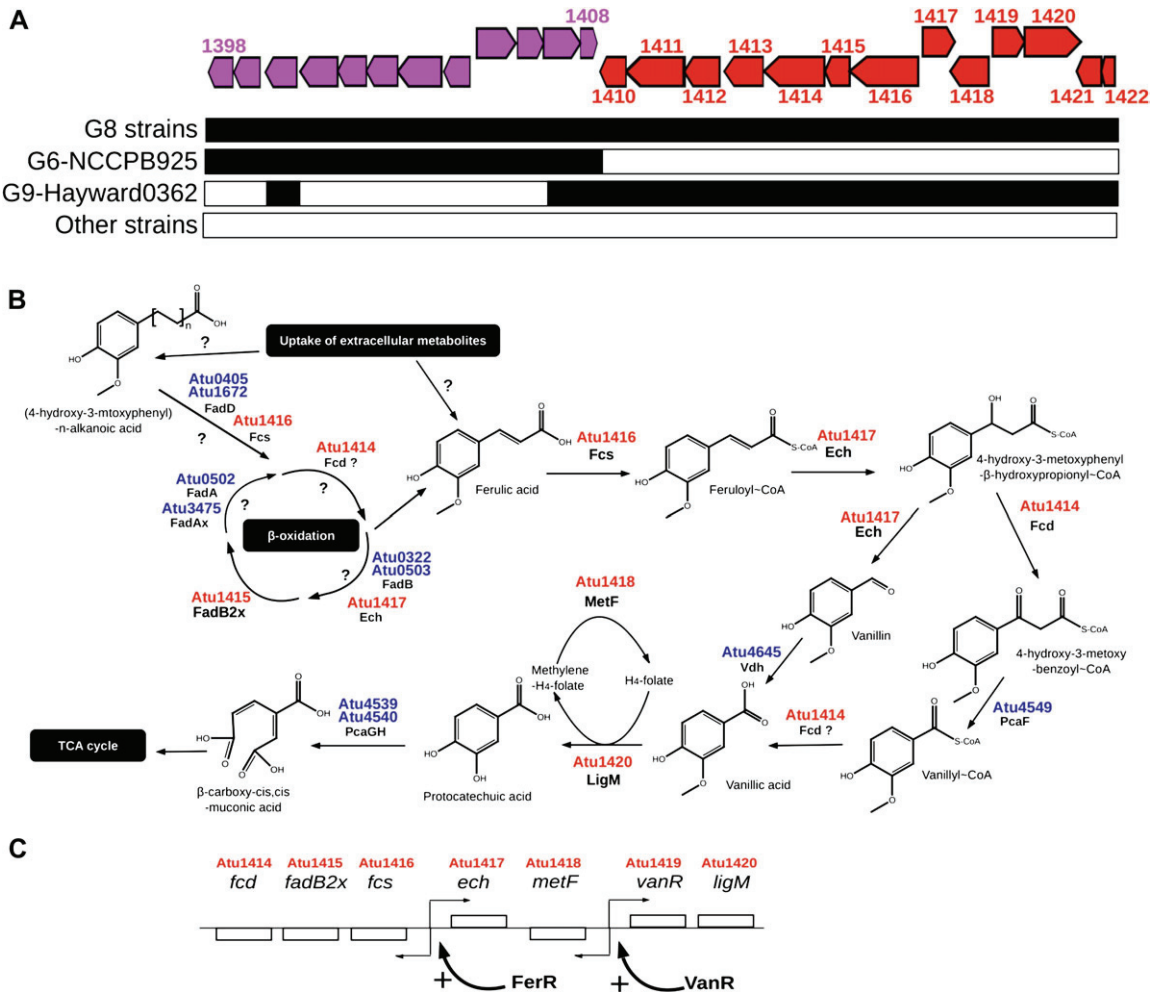


Fig. 6.—Putative ferulic acid catabolism pathway encoded by SpG8-1b. (*A*) SpG8-1 CDSs organization in C58 (top): subregions SpG8-1a and SpG8-1b are colored in purple and red, respectively. Presence in other *Agrobacterium tumefaciens* strains (bottom): presence, black; absence, white. (*B*) Reconstructed ferulic acid catabolism pathway encoded by SpG8-1b according to similarities to sequences in databases and associated literature: Fcs, feruloyl-CoA synthetase (Overhage et al. 1999; Plaggenborg et al. 2003); Ech, enoyl-CoA hydratase (Pelletier and Harwood 1998); Fcd, feruloyl-CoA dehydratase; LigM, tetrahydrofolate-dependent vanillate O-demethylase (Nishikawa et al. 1998); MetF, methylenetetrahydrofolate reductase (Nishikawa et al. 1998). (*C*) Putative transcriptional regulation of SpG8-1b genes inferred from sequence similarities in databases: VanR, vanillate catabolism repressor (Morawski et al. 2000); FerR, ferulate catabolism regulator (Breese and Fuchs 1998; Calisti et al. 2008).

enzyme is involved in a beta-oxidative pathway of long chain substituted phenolic degradation in *Pseudomonas* (Olivera et al. 2001), yielding short-chain phenylalkanoyl-CoAs such as cinnamoyl-CoA. This suggests alternative entries for this putative G8-specific phenolic degradation pathway: either by uptake of ferulic acid (as one of the cognate transporters could provide this ability) or by transformation of more complex phenolics, for example, by iterative oxidation of long chain–substituted cinnamic acids. G8 strains could thus likely degrade ferulic acid into protocathechuic acid and then assimilate it as a carbon source.

We verified this possibility by testing strains for their ability to degrade ferulic acid. After 12 h incubation, strains bearing SpG8-1b (G8 strains and G9-Hayward 0362) degraded all the ferulic acid in a comparable manner, whereas other strains did not (fig. 5*A*). In addition, a C58 mutant deleted for the whole SpG8-1b (C58ΔSpG8-1b, fig. 5*A*) was unable to degrade ferulic acid. This locus is therefore clearly involved in ferulic acid degradation.

As a generalization, other genes involved in aromatic compounds catabolism were found in other SpG8 gene clusters: a putative mandelate racemase, in SpG8-1a (Atu1406) and a putative shikimate dehydrogenase in SpG8-7b (Atu4295). These enzymatic reactions were parts of pathways leading to protocathechuic acid production, mandelate degradation, or shikimate degradation, respectively, suggesting that degradation of aromatic compounds into protocathechuic acid may be a crucial synapomorphic trait of genomovar G8.

Occurrence of SpG8 Genes in Other G8 Members

A question could be asked about the overall relevance of the present work. Within the chosen species, that is, the *A. tumefaciens* genomovar G8, we used the largest set of markedly different strains available when the array was designed. The results were thus valid for this set of strains, but are SpG8 genes also present in other G8 members? Indeed, curdlan biosynthesis genes were already described in the industrial agrobacterial strain ATCC 31749 (Portilho et al. 2006), that we found to be very likely another G8 member based on the high sequence similarities of several genes of ATCC 31749 and C58 as compared with other genomic species (data not shown). Similar observations could also apply to *tetA-tetR* (Luo and Farrand 1999). In addition, we tracked G8 members in the many agrobacterial strains that can be isolated from various environments. We succeeded in isolating a new one, that is, MKS.01 (CFBP 7336), which differed from all other G8 strains by core gene markers, and we verified that MKS.01 had all the G8-specific genes and phenotypes defined in the present work, including curdlan production and ferulic acid degradation (data not shown). Conversely, the SpG8 genomic islands appear to be absent from the recently published genomic sequence of the genomovar G1 strain H13-3 (Wibberg et al. 2011). All of these a posteriori verifications confirmed the hypothesis of genomic species characterized by common species-specific genes inherited from a common ancestor adapted to a specific primary ecological niche.

Discussion

The aim of the present study was to disclose the speciation mechanism leading to differentiation of genomic species by assessing the presence of species-specific genes in genomes. As this is amenable by comparative genomics, the present study was geared toward detecting C58 gene homologs in other genomes by using a microarray constructed with probes spanning the entire C58 genome. However, although it is easy to detect present genes with CGH arrays when hybridized DNAs are highly similar to C58 DNA (i.e., in genomovar G8), there is a dramatic decrease in the signal over background ratio for more divergent genomes from other genomic species. This difficulty was overcome by taking into account the regional organization of the hybridization signal along the genome because, in spite of their weakness, successive signals of present loci are generally more intense than noises of absent ones. Thanks to this partition procedure, we were able to confidently detect the presence of genes not only in all other species of the *A. tumefaciens* complex but also in the remote species *A. larrymoorei*. We, however, failed to obtain accurate results in *A. vitis* CFBP 5523^T, *R. rhizogenes* K84, or *E. meliloti* 1021 (data not shown), likely because those bacteria generally diverged beyond the estimated detection threshold (80% similarity at best) of the procedure (supplementary table S4, Supplementary Material online).

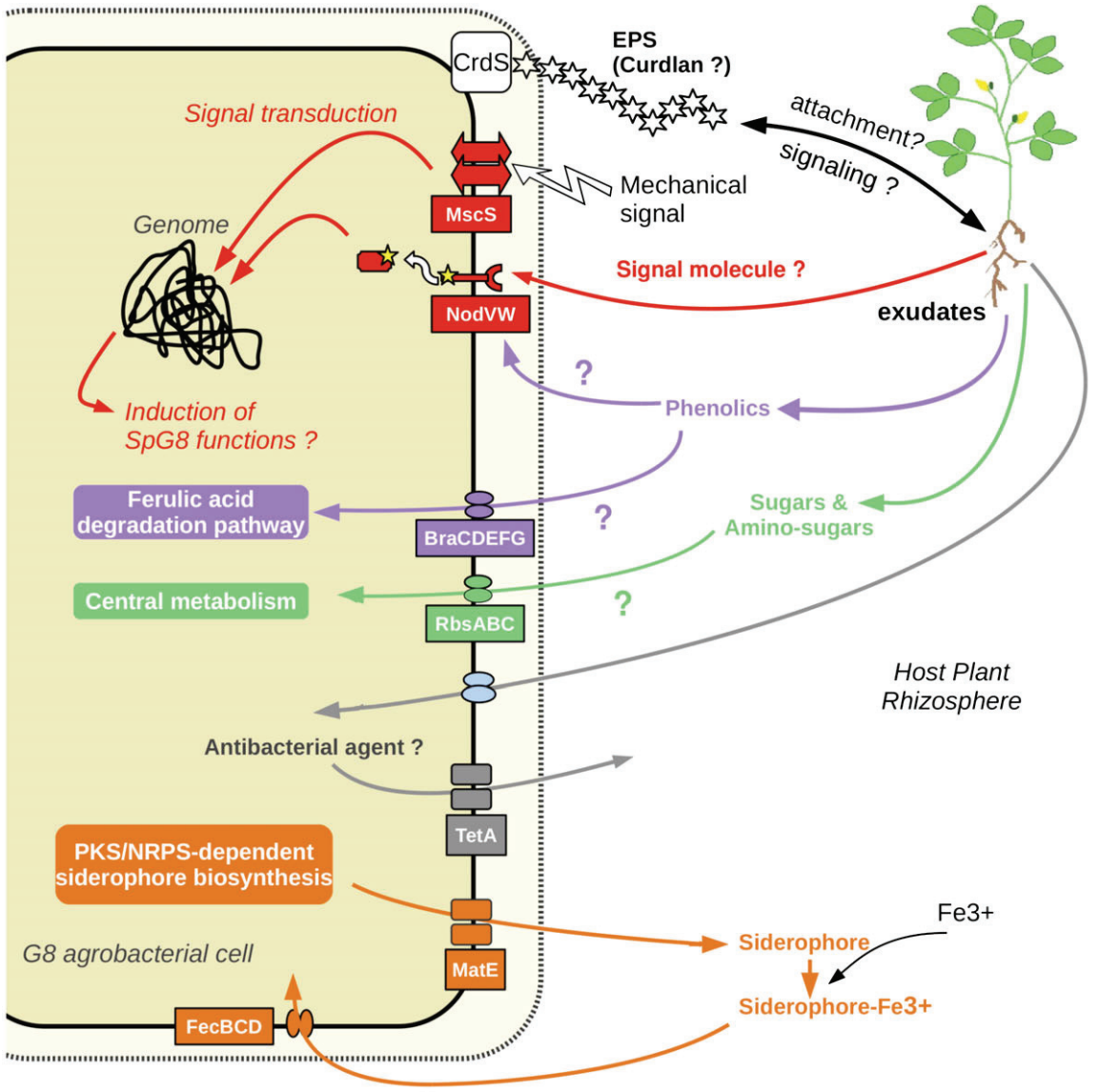
G8-Specific Functions Useful for Life in Plant Rhizospheres

Remarkably, SpG8 functions seemed to collectively define an ecological niche of G8 agrobacteria related to commensal interactions with plants (fig. 7). Although agrobacteria are well known to be pathogenic to plants by inducing crown gall disease, this is a secondary ecological trait related to the availability of a dispensable plasmid (i.e., Ti plasmid which harbors the pathogenicity determinants). Agrobacteria are generally Ti plasmid free and are primarily common soilborne organisms able to live commensally in plant rhizospheres (Savka et al. 2002; Hartmann et al. 2009). As we were looking for general adaptive determinants of the species regardless of its pathogenic status, we assumed that species-specific adaptations were more likely related to life in soils or rhizospheres rather than to crown galls.

SpG8 loci code for numerous catabolic pathways of carbohydrates, namely ferulic acid (SpG8-1b), diverse sugars (SpG8-1a, SpG8-4), amino acids (SpG8-1a, SpG8-5), and opine-like/Amadori compounds (SpG8-5). All are typical molecules that can be found in plant rhizospheres, exuded by plants (complex sugars) or derived from plant degradation products (phenolics, opine/Amadori products). Clearly, ferulic acid is a plant compound involved in the lignin biosynthesis present at plant wounds (Humphreys et al. 1999). Degradation products are, however, released in soil and may thus facilitate the survival of agrobacteria in soil as well. The SpG8 loci involvement in life adaptation to plants or soil are therefore not exclusive alternatives. Moreover, sugar, phenolics, and opines are also known to play an important role in the pathogenicity of Ti plasmids harboring agrobacteria. In that sense, the present results support the exaptation hypothesis of Dessaux's team (Vaudequin-Dransart et al. 1995), who proposed that the ability of agrobacteria to use opines selectively arose from a more general ability of this taxon to use opine-like compounds, including Amadori products and other related substrates.

In addition to carbon resources available in the rhizosphere, other factors are important in the bacterial niche definition. For instance, bacterial cells may be able to recognize a favorable environment, reach it (e.g., via positive tropism) and stay inside it (e.g., via physical attachment), or modify it (e.g., by secreting extracellular products or stimulating a plant to modify its exudation spectrum). These functions involve molecular signaling that can be distant (by diffusion of a signal molecule) or by contact between the bacterium and its specific habitat (including a partner plant).

Indeed, production of insoluble β-1,3-glucan exopolysaccharide (curdlan) encoded by SpG8-2a may play a role in attachment (Matthysse and McMahan 1998; Rodríguez-Navarro



Locus	Predicted functions	Locus	Predicted functions
SpG8-1a	Sugar & amino-acid transport; sugar metabolism	SpG8-2	Curdlan EPS biosynthesis
SpG8-4	Monosaccharide transport & catabolism; carbohydrate metabolism	SpG8-3	Siderophore biosynthesis; iron-siderophore uptake
SpG8-5	Opine-like compound catabolism	SpG8-6	Drug / toxic resistance
SpG8-1b	Ferulic acid uptake and catabolism	SpG8-7	Environmental signal sensing / transduction

Fig. 7.—Hypothetical integrated functioning of SpG8 genes allowing G8 member adaptation to their specific ecological niche. EPS, exopolysaccharide; CrdS, curdlan synthase; McsS, mechano-sensitive channel; NodVW, two-component system sensor kinase and response regulator; BraCDEFG, branched-chain amino acid transporter; RbsABC, ribose transporter; TetA, tetracycline extrusion pump; MatE, multidrug transporter; FecBCD, iron-siderophore transporter.

et al. 2007) and contact signaling. Annotated functions may, however, have pleiotropic effects, and curdlan production may also be important for passive resistance to toxics, especially in plant rhizospheres where antimicrobial agents-like flavonoids are secreted (Palumbo et al. 1998). Other putative defense mechanism of G8 agrobacteria may be provided by SpG8 loci by action of multidrug exporters (SpG8-6), whereas the siderophore biosynthesis locus (SpG8-3) might provide another general fitness gain in competition with other bacteria present in the biotope. Scavenging of limiting resources like iron is

known to be a very potent means to outperform competitors, especially in habitats like rhizospheres with dense and diverse populations, as described for plant growth-promoting rhizobacteria like *Pseudomonas fluorescens* or *R. rhizogenes* (Penyalver et al. 2001; Siddiqui 2006).

Finally, locus SpG8-7b, which encodes membrane proteins involved in the perception of mechanical and chemical signals, is a candidate to facilitate recognition of favorable environments. Interestingly, those putative environment-sensing genes are homologous to systems of perception of toluene and styrene in *Pseudomonas* sp. (Lau et al. 1997; Panke et al. 1998) and are conserved in synteny with those from the chromosome of *P. lavamentivorans* SD-1. This latter species is known to switch from motile to sessile behavior in the presence of phenyl-substituted long-chain fatty acids (Schleheck et al. 2004) that share structural features with ferulic acid. These homology relationships strongly suggest that two-component signaling systems of this family are activated by the presence of some phenolics in the medium. Moreover, these genes code proteins also homologous to NodVW/NwsAB from *B. japonicum*, which mediate host recognition during the nodulation process. Considering these relationships, we hypothesized that locus SpG8-7b is involved in the perception of signals from the environment that may be responsible for activation of other functions, including, perhaps, SpG8 functions such as phenolic metabolism. The frequent reference to phenolics in annotation of SpG8 genes suggests that these compounds could be of primary importance in the biology of G8 agrobacteria, being both metabolites and signals released by the host plant.

Evolutionary History of SpG8 Genes

The presence of species-specific genes can be understood as due to the conservation of ancestral genes lost in other species or to the acquisition of foreign genes by the most recent common ancestor. Several SpG8 regions (SpG8-1a, SpG8-4, and SpG8-7a) contained genes that were also found in G6-NCCPB925, the closest outgroup of G8. This suggests that these specific regions may have been present in the common ancestor of G8 and G6. Remarkably, these regions tended toward the codon usage signature of core-genome genes (supplementary table S5, Supplementary Material online, fig. 4). Based on these elements, SpG8-1, SpG8-4, and SpG8-7 may be clusters of ancestral genes already present in the genome of an ancient ancestor of agrobacteria, specifically retained in the [G6,G8] clade but lost in other clades. This is especially probable for region SpG8-1. This region was possibly present as an entire cluster in ancestors of G6-NCCPB925 and G9-Hayward0362 and may then have been partially lost, leading to differential retention of subregions SpG8-1a and SpG8-1b in G6-NCCPB925 and G9-Hayward0362, respectively (fig. 6*A*). In contrast, transfers may be more likely for SpG8 genes shared with more distant genomic species such as G1 (SpG8-3 and SpG8-6a gene clusters) or G4 (Atu4215-4218 in cluster SpG8-6b), which have sporadic-like codon usage signatures. This suggests that lateral gene transfer as well as gene retention contribute to the establishment of a species-specific gene repertoire.

Gene Content Flexibility of the Linear Chromosome

As previously observed at higher taxonomic level by comparing *Agrobacterium* "biovars" (i.e., *A. tumefaciens* C58 to *A. vitis* S4 and *R. rhizogenes* K84; Slater et al. 2009) and other bacteria such as *Vibrio* (Chen et al. 2003; Vesth et al. 2010), the second chromosome of *A. tumefaciens* genomic species also appears as the major spot for innovation in the gene repertoire (fig. 3). This high genomic flexibility of the second chromosome was moreover likely facilitated by a linear architecture as illustrated by the transfer of half of the linear chromosome between C58 and other G8 members (fig. 1). Actually, a bacterial linear chromosome could behave as a standard eukaryotic chromosome requiring a single crossover to exchange almost a complete chromosome branch. Linear chromosomes, which are rare genomic features in bacteria, may facilitate the spread of adaptive genomic innovations and likely played a key role in speciation in the *A. tumefaciens* complex as suspected in *Streptomyces* or *Borrelia* spp. (Volff and Altenbuchner 2000; Chen et al. 2010).

Parapatric Speciation Gives Rise to Genomic Species

We found genes coding ecologically relevant functions present in the genomes of members of a given genomic species but not in its closest relatives. They were likely present in the most recent common ancestor of its members likely allowing ecological isolation. We may in turn speculate this isolation initiated the speciation process. We chose to work with a genomic species with high known diversity (Mougel et al. 2002; Portier et al. 2006; Costechareyre et al. 2009, 2010) and also because this species has very closely related sister species often co-inhabiting the same soil (Costechareyre et al. 2010), even at the very microscale (Vogel et al. 2003). Agrobacteria are moreover common rhizospheric bacteria (Krimi et al. 2002, Costechareyre et al. 2010) which genomic species are differentially trapped according to plant host (Lavire C, unpublished data). Agrobacteria in soils thus form ecological guilds where every species likely taps the same resources (e.g., rhizospheres) in a similar way, except for a few specific traits. As agrobacterial species are not geographically isolated and because they have determinants for species-specific ecological niche, we assume that these species have arisen by parapatric speciation. It is likely that speciations in the same habitat occurred as a consequence of local adaptations to host plants, as suggested by annotations of G8-specific functions.

Adaptations to plants might be related to host specificity as already suggested by the known preferential occurrence of *A. rubi*, *A. vitis*, and *A. larrymoorei* in tumors of *Rubus* sp., *Vitis* sp., and *Ficus benjamina*, respectively. Determination of specific adaptations of *A. tumefaciens* species may also improve our knowledge about crown gall epidemiology, including preferential spread by some hosts. In the case of G8, we suspected preferential trapping by *Medicago truncatula* (Lavire C, unpublished data), echoing the homology of SpG8-7 with sensors of *E. meliloti*—the symbiont of *M. truncatula*. Adaptation to plant is possibly not confined to commensal adaptation to the root biotope but more generally to ecological features encountered in the whole plant, including tumors. Consequently, it is possible that *A. tumefaciens* species adaptations to plants may also modulate the epidemiology of pathogenic agrobacteria.

Interest of Ecological Species Concept Investigation for Taxonomy

Agrobacterium tumefaciens genomic species are valid species but they are still awaiting a valid Latin binomial because they were lacking well-characterized distinctive phenotypic traits. Novel G8 members could be identified by phylogenetic analysis of core genes such as *recA* or *chvA*, as previously described (Costechareyre et al. 2009, 2010) or by looking for G8-specific genes via CGH microarrays or PCR. However, the present work actually emphasizes several traits such as curdlan production, ferulic acid degradation, resistance to tetracyclin for genomovar G8 that, when combined, would be valuable traits for species distinction. This is why, in agreement with the latest recommendations of Stackebrandt et al. (2002), we propose to valid the status of genomovar G8 as a recognized bacterial species by giving it a Latin binomial and a type strain, C58. We thus propose this novel species be named *Agrobacterium fabrum*, from the Latin plural genitive of *smith*, in reference to the use of C58 to construct genetically modified plants, while also honoring the pioneer isolator of *Agrobacterium* (Smith) as well as other scientists with a Faber-related name in different languages, for example, Smith, Schmidt, Smet, Faivre, Farrand, Faure, Herrera, etc., who studied various aspects of *Agrobacterium* biology.

Generalization of the Concept of Bacterial Genomic Species as Ecological Species

The question of ecological speciation of bacteria is still in debate (Achtman and Wagner 2008) partly because the bacterial species definition is at the center of this debate. Here, we only consider the genomically based species definition still acknowledged by international taxonomic committees (Wayne et al. 1987; Stackebrandt et al. 2002), even though if there is still named bacterial species—especially in anciently described human pathogens—that do not fit the genomic species criterion. We thus chose a taxon level relevant for the current taxonomy and intended to verify that this taxon level could have specific ecological features that scheme a potential primary niche. This was usually achieved by investigating differential ecological properties of species as for instance within the genus *Prochlorococcus* (Johnson et al. 2006). We showed here that the discovery of the specific ecological niche of a species is amenable by comparative genomic, when it is performed with several strains within this species compared with strains belonging to closely related species. This was done with *Salmonella enterica* (Porwollik et al. 2002), *Lactobacillus casei* (Cai et al. 2009), and *Campylobacter coli* versus *C. jejuni* (Lefébure et al. 2010). This should be generalized in future taxonomic investigations in order to improve the biological information attached to novel species. Of course, it is also possible to infer primary ecology of other taxonomic levels such as strain clusters within a species as shown in *E. coli* (Touchon et al. 2009), genera, or still higher taxa (Philippot et al. 2010). Interestingly, these latter authors showed that the broader the clade, the less defined is the associated ecology.

The present study highlighted the relevance of looking for species-specific genes by assessing genome features. We showed that—at least for the present model—species-specific genes were involved in ecological adaptations to the species primary niche. Consequently, it is likely that in this instance the genomic species was of ecovar descent. Our study benefits from the synergy between bioinformatic treatments of high throughput data and bench works. Both approaches are essential for reconstructing—without a priori knowledge—a reliable ecological niche model for further investigations on bacterial speciation and evolution.

Supplementary Material

Supplementary tables S1–S9 and figures S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

F.L. received a doctoral grant from ENS-Lyon, T.C. from Ministère de l'Éducation nationale de l'Enseignement Supérieur et de la Recherche, and M.S. from Ministère des Affaires Etrangères et Européennes. D.M. was supported by an INRA postdoctoral fellowship. The authors would like to thank P. Oger who allowed the inclusion of pTiBo542 in the microarray design, G. Meiffren at CESN for assistance in HPLC analyses, P. Portier at CFBP (http://www-intranet.angers.inra.fr/cfbp/) for strain repository, A. Calteau at Genoscope (https://www.genoscope.cns.fr/agc/microscope/) for AgrobacterScope platform management, M.K. Lhommé at Lyon 2 University for Latin advice, and translator Dr D. Manley for reading the manuscript and providing suggestions. This work was supported by the EcoGenome project of Agence

Nationale de la Recherche (grant number BLAN08-1_335186); Lyon 1 University (grant numbers IFR41-2006_Nesme, BQR-2006_Nesme); Bureau des Ressources Génétiques (grant number BRG-2005_Nesme); and the Département Santé des Plantes et Environnement of INRA (grant number DSPE-2005_Nesme).

Literature Cited

Achtman M, Wagner M. 2008. Microbial diversity and the genetic nature of microbial species. Nat Rev Microbiol. 6:439.

Bouzar H, Jones J. 2001. *Agrobacterium larrymoorei* sp. nov., a pathogen isolated from aerial tumours of *Ficus benjamina*. Int J Syst Evol Microbiol. 51:1023–1026.

Braun V, Mahren S, Sauter A. 2006. Gene regulation by transmembrane signaling. Biometals. 19:103–113.

Breese K, Fuchs G. 1998. 4-Hydroxybenzoyl-CoA reductase (dehydroxylating) from the denitrifying bacterium *Thauera aromatica*—prosthetic groups, electron donor, and genes of a member of the molybdenum-flavin-iron-sulfur proteins. Eur J Biochem. 251:916–923.

Cai H, et al. 2009. Genome sequence and comparative genome analysis of *Lactobacillus casei*: insights into their niche-associated evolution. Genome Biol Evol. 1:239–257.

Calisti C, Ficca AG, Barghini P, Ruzzi M. 2008. Regulation of ferulic catabolic genes in *Pseudomonas fluorescens* BF13: involvement of a MarR family regulator. Appl Microbiol Biotechnol. 80:475–483.

Charif D, Lobry JR. 2007. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U, Porto M, Roman E, Vendruscolo M, editors. Structural approaches to sequence evolution: Molecules networks populations. Heidelberg (Germany): Springer-Verlag. p. 207–232.

Chen CY, et al. 2003. Comparative genome analysis of *Vibrio vulnificus*, a marine pathogen. Genome Res. 13:2577–2587.

Chen W, et al. 2010. Chromosomal instability in *Streptomyces avermitilis*: major deletion in the central region and stable circularized chromosome. BMC Microbiol. 10:198.

Choi K, Schweizer HP. 2005. An improved method for rapid generation of unmarked *Pseudomonas aeruginosa* deletion mutants. BMC Microbiol. 5:30.

Civolani C, Barghini P, Roncetti AR, Ruzzi M, Schiesser A. 2000. Bioconversion of ferulic acid into vanillic acid by means of a vanillate-negative mutant of *Pseudomonas fluorescens* strain BF13. Appl Environ Microbiol. 66:2311–2317.

Cock PJA, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 25:1422–1423.

Cohan FM. 2002. Sexual isolation and speciation in bacteria. Genetica. 116:359–370.

Connor N, et al. 2010. Ecology of speciation in the genus *Bacillus*. Appl Environ Microbiol. 76(5):1349–1358.

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

Costechareyre D, et al. 2010. Rapid and efficient identification of *Agrobacterium* species by *recA* allele analysis: *Agrobacterium recA* diversity. Microb Ecol. 60(4):862–872.

Datsenko KA, Wanner BL. 2000. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. Proc Natl Acad Sci U S A. 97:6640–6645.

Dufour A, Dray S. 2007. The ade4 package: implementing the duality diagram for ecologists. J Stat Softw. [Internet] 22(i04) [cited 2011 Jul 28].

Felsenstein J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Seattle (WA): Department of Genome Sciences, University of Washington.

Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. 2009. The bacterial species challenge: making sense of genetic and ecological diversity. Science. 323:741–746.

Galibert F, et al. 2001. The composite genome of the legume symbiont *Sinorhizobium meliloti*. Science. 293:668–672.

Gevers D, et al. 2005. Re-evaluating prokaryotic species. Nat Rev Microbiol. 3:733–739.

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

Guéguen L. 2005. Sarment: Python modules for HMM analysis and partitioning of sequences. Bioinformatics. 21:3427–3428.

Hardin G. 1960. The competitive exclusion principle. Science. 131:1292–1297.

Hartmann A, Schmid M, Tuinen D, Berg G. 2009. Plant-driven selection of microbes. Plant Soil. 321:235–257.

Hosie AHF, Allaway D, Galloway CS, Dunsby HA, Poole PS. 2002. *Rhizobium leguminosarum* has a second general amino acid permease with unusually broad substrate specificity and high similarity to branched-chain amino acid transporters (Bra/LIV) of the ABC family. J Bacteriol. 184:4071–4080.

Humphreys JM, Hemm MR, Chapple C. 1999. New routes for lignin biosynthesis defined by biochemical characterization of recombinant ferulate 5-hydroxylase, a multifunctional cytochrome P450-dependent monooxygenase. Proc Natl Acad Sci U S A. 96:10045–10050.

Johnson ZI, et al. 2006. Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. Science 311:1737–1740.

Kneen BE, LaRue TA. 1983. Congo red absorption by *Rhizobium leguminosarum*. Appl Environ Microbiol. 45:340–342.

Krimi Z, Petit A, Mougel C, Dessaux Y, Nesme X. 2002. Seasonal fluctuations and long-term persistence of pathogenic populations of *Agrobacterium spp*. in soils. Appl Environ Microbiol. 68:3358–33654.

Lake JA. 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc Natl Acad Sci U S A. 91:1455–1459.

Lang K, Lindemann A, Hauser F, Göttfert M. 2008. The genistein stimulon of *Bradyrhizobium japonicum*. Mol Genet Genomics. 279:203–211.

Lau PC, et al. 1997. A bacterial basic region leucine zipper histidine kinase regulating toluene degradation. Proc Natl Acad Sci U S A. 94:1453–1458.

Lefébure T, Pavinski Bitar PD, Suzuki H, Stanhope MJ. 2010. Evolutionary dynamics of complete *Campylobacter* pan-genomes and the bacterial species concept. Genome Biol Evol. 2:646–655.

Luo ZQ, Farrand SK. 1999. Cloning and characterization of a tetracycline resistance determinant present in *Agrobacterium tumefaciens* C58. J Bacteriol. 181:618–626.

Maddocks SE, Oyston PCF. 2008. Structure and function of the LysR-type transcriptional regulator (LTTR) family proteins. Microbiology. 154:3609–3623.

Martens M, et al. 2007. Multilocus sequence analysis of *Ensifer* and related taxa. Int J Syst Evol Microbiol. 57:489–503.

Matthysse AG, McMahan S. 1998. Root colonization by *Agrobacterium tumefaciens* is reduced in *cel, attB, attD*, and *attR* mutants. Appl Environ Microbiol. 64:2341–2345.

Mayr E. 1942. Systematics and the origin of species, from the viewpoint of a zoologist. New York (NY): Columbia University Press.
Morawski B, Segura A, Ornston LN. 2000. Repression of *Acinetobacter* vanillate demethylase synthesis by VanR, a member of the GntR family of transcriptional regulators. FEMS Microbiol Lett. 187:65–68.
Moriyama Y, Hiasa M, Matsumoto T, Omote H. 2008. Multidrug and toxic compound extrusion (MATE)-type proteins as anchor transporters for the excretion of metabolic waste products and xenobiotics. Xenobiotica. 38(7–8):1107–1118.
Mougel C, Thioulouse J, Perrière G, Nesme X. 2002. A mathematical method for determining genome divergence and species delineation using AFLP. Int J Syst Evol Microbiol. 52:573–586.
Nelder JA, Mead R. 1965. A simplex method for function minimization. Comput J. 7:308–313.
Nishikawa S, et al. 1998. Cloning and sequencing of the *Sphingomonas (Pseudomonas) paucimobilis* gene essential for the O demethylation of vanillate and syringate. Appl Environ Microbiol. 64:836–842.
Olivera ER, et al. 2001. Two different pathways are involved in the beta-oxidation of n-alkanoic and n-phenylalkanoic acids in *Pseudomonas putida* U: genetic studies and biotechnological applications. Mol Microbiol. 39:863–874.
Ophel K, Kerr A. 1990. *Agrobacterium vitis sp. nov.* for strains of *Agrobacterium* biovar 3 from grapevines. Int J Syst Bacteriol. 40:236–241.
Overhage J, Priefert H, Steinbuchel A. 1999. Biochemical and genetic analyses of ferulic acid catabolism in *Pseudomonas sp*. strain HR199. Appl Environ Microbiol. 65:4837–4847.
Palumbo JD, Kado CI, Phillips DA. 1998. An isoflavonoid-inducible efflux pump in *Agrobacterium tumefaciens* is involved in competitive colonization of roots. J Bacteriol. 180:3107–3113.
Panke S, Witholt B, Schmid A, Wubbolts MG. 1998. Towards a biocatalyst for (S)-styrene oxide production: characterization of the styrene degradation pathway of *Pseudomonas sp*. strain VLB120. Appl Environ Microbiol. 64:2032–2043.
Parke D. 1995. Supraoperonic clustering of *pca* genes for catabolism of the phenolic compound protocatechuate in *Agrobacterium tumefaciens*. J Bacteriol. 177:3808–3817.
Pelletier DA, Harwood CS. 1998. 2-Ketocyclohexanecarboxyl coenzyme A hydrolase, the ring cleavage enzyme required for anaerobic benzoate degradation by *Rhodopseudomonas palustris*. J Bacteriol. 180:2330–2336.
Penyalver R, Oger P, López MM, Farrand SK. 2001. Iron-binding compounds from *Agrobacterium spp.*: biological control strain *Agrobacterium rhizogenes* K84 produces a hydroxamate siderophore. Appl Environ Microbiol. 67:654–664.
Petit A, et al. 1978. Substrate induction of conjugative activity of *Agrobacterium tumefaciens* Ti plasmids. Nature 271:570–572.
Philippot L, et al. 2010. The ecological coherence of high bacterial taxonomic ranks. Nat Rev Microbiol. 8:523–529.
Plaggenborg R, Overhage J, Steinbüchel A, Priefert H. 2003. Functional analyses of genes involved in the metabolism of ferulic acid in *Pseudomonas putida* KT2440. Appl Microbiol Biotechnol. 61:528–535.
Portier P, et al. 2006. Identification of genomic species in *Agrobacterium* biovar 1 by AFLP genomic markers. Appl Environ Microbiol. 72:7123–7131.
Portilho M, Matioli G, Zanin GM, de Moraes FF, Scamparini ARP. 2006. Production of insoluble exopolysaccharide of *Agrobacterium sp*. (ATCC 31749 and IFO 13140). Appl Biochem Biotechnol. 131:864–869.
Porwollik S, Wong RM, McClelland M. 2002. Evolutionary genomics of *Salmonella*: gene acquisitions revealed by microarray analysis. Proc Natl Acad Sci U S A. 99:8956–8961.
Quandt J, Hynes MF. 1993. Versatile suicide vectors which allow direct selection for gene replacement in gram-negative bacteria. Gene. 127:15–21.
R Development Core Team. 2009. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for statistical computing.
Rodríguez-Navarro DN, Dardanelli MS, Ruíz-Saínz JE. 2007. Attachment of bacteria to the roots of higher plants. FEMS Microbiol Lett. 272:127–136.
Rondon MR, Ballering KS, Thomas MG. 2004. Identification and analysis of a siderophore biosynthetic gene cluster from *Agrobacterium tumefaciens* C58. Microbiology. 150:3857–3866.
Sambrook J, Russell DW. 2001. Molecular cloning: a laboratory manual. Cold Spring Harbor (NY): CSHL Press.
Sanjay AB, Stach JEM, Goodfellow M. 2008. Genetic and phenotypic evidence for *Streptomyces griseus* ecovars isolated from a beach and dune sand system. Antonie Van Leeuwenhoekss. 94:63–74.
Savka MA, Dessaux Y, Oger P, Rossbach S. 2002. Engineering bacterial competitiveness and persistence in the phytosphere. Mol Plant Microbe Interact. 15:866–874.
Schleheck D, Tindall BJ, Rosselló-Mora R, Cook AM. 2004. *Parvibaculum lavamentivorans gen. nov., sp. nov.*, a novel heterotroph that initiates catabolism of linear alkylbenzenesulfonate. Int J Syst Evol Microbiol. 54:1489–1497.
Siddiqui ZA. 2006. *PGPR: biocontrol and biofertilization*. Dordrecht (The Netherlands): Springer.
Sikorski J, Nevo E. 2005. Adaptation and incipient sympatric speciation of *Bacillus simplex* under microclimatic contrast at "Evolution Canyons" I and II, Israel. Proc Natl Acad Sci U S A. 102:15924–15929.
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.
Sneath PHA, Sokal RR. 1973. Numerical taxonomy. San Fransisco (CA): WH Freeman and Co.
Stackebrandt E, et al. 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol. 52:1043–1047.
Staley JT. 2004. Speciation and bacterial phylospecies. In: Bull AT, editor. Microbial diversity and bioprospecting. Washington (DC): ASM Press. pp. 40–47.
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. 2009:bap021.
Vaudequin-Dransart V, et al. 1995. Novel Ti plasmids in *Agrobacterium* strains isolated from fig tree and chrysanthemum tumors and their opine-like molecules. Mol Plant Microbe Interact. 8:311–321.
Velázquez E, et al. 2010. Analysis of core genes supports the reclassification of strains *Agrobacterium radiobacter* K84 and *Agrobacterium tumefaciens* AKE10 into the species *Rhizobium rhizogenes*. Syst Appl Microbiol. 33:247–251.
Vesth T, et al. 2010. On the origins of a *Vibrio* species. Microb Ecol. 59:1–13.
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.
Wayne LG, et al. 1987. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. Int J Syst Bacteriol. 37: 463–464.
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. Forthcoming doi:10.1016/j.jbiotec.2011.01.010
Wood DW, et al. 2001. The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. Science. 294:2317–2323.
Zhao J, Deng Y, Manno D, Hawari J. 2010. *Shewanella spp.* genomic evolution for a cold marine lifestyle and in-situ explosive biodegradation. PLoS One. 5:e9109.

Associate editor: Bill Martin

2.4 Reconstructing ancestral genomes of *A. tumefaciens* reveals ecological adaptations along their diversification

2.4.1 Introduction

The ecology of bacteria is poorly understood. Though we have some idea of which environment certain species can be sampled from, what compounds they can metabolize or what interactions they can establish with other organism, we know very little of what are the selective pressures acting on bacteria in their everyday life. The genomes of living organisms contain all the informations that make their lifestyle possible, and for this reason genomes appear as a gold mine for biologists in quest of understanding. Though, even if we knew how to decode the complex genomic language and determine the functional role of every base in one genome, that would not help us to understand what portion contributes significantly to the fitness of the organism in its current environment (Doolittle, 2013).

What we can learn however, is something about their past. Indeed, by comparing genome sequences, we can recognize patterns of divergence and conservation of the genetic information, which can tell us much about the selective pressures that were acting on genomes on the long term. For instance, observing a gene sequence highly conserved between two divergent genomes is indicative of purifying selection that acted on this gene, for a significant time since the split between the two genomes. This comparison can be done with multiple genomes, in a phylogenetic framework that models what happened in a lineage and when, thus reconstructing ancestral genomes to tell their history.

Most interestingly, some methods for reconstruction of ancestral genomes can exhaustively describe the processes that participated to the evolution of genomes, and notably the history of genes in genomes. Most genes have complex histories, marked by many events of gene duplication, loss and, in the case of micro-organisms, horizontal transfers. Different genes have different histories that followed the one of species by different paths, i.e. evolutionary scenarios. The reconciliation of genome and gene histories through the inference of events of duplication, transfer and loss of genes allow to reconstruct the gene content of ancestral genomes. This can be very valuable to understand what were the selective pressures that marked the evolution of a lineage, because (1) the individual histories of genes can tell us what gene function was brought to which ancestor, how, and from where – either by duplication of a native gene or by transfer from a close or distant relative, etc., every scenario potentially implying a different selective advantage; (2) the ancestral gene contents give a functional context in which the gain or loss of a gene function can be interpreted, notably when taking into account the syntenic relationships and functional interactions between genes; (3) the inference of all the individual events allows the estimation of the global process of genome evolution, which can be used as an empiric reference for detecting deviations from the norms, as a potential signature of the action of selection.

The bacterial genomes are known to be in constant flux, with genes getting in and out at a rate that can exceed the nucleotide substitution rate (Lawrence and Ochman, 1997), leading

to differences of more than a thousand genes between the closest strains of the same species (Touchon et al., 2009). This dynamics leads to the definition of 'core' versus 'accessory' genomes, which respectively gather the genes that are shared by all member strains of a species and those that are found in some strains but not all. Some accessory genes are frequently gained by transfer and then quickly lost, leaving patterns of presence in genome that are patchy regarding the species phylogeny; others only appear in one genome. Among those accessory genes, most have likely ephemeral, if any, adaptive value, and are only caught in our analysis by the snapshot of genome sequencing. Though, in this apparent chaos of gene content variations, some accessory genes stably settle in genomes, and become part of the core genome of a lineage. These 'domestication' events constitute the most remarkable deviations from a neutral model where rapid gains and loss prevail. These clade-specific genes could thus have been under purifying selection since their acquisition in the corresponding clade common ancestor. This could for instance reflect their contribution to the adaptation of their host to its ecological niche. However, not all genes gained by the ancestor of a clade and conserved afterwards were necessarily under purifying selection ; drift or selfish replication and transmission can maintain non-adaptive genes in genomes. Can other patterns of evolution of genes in genomes be used to distinguish the action of selection from that of drift?

In a previous study where we explored the diversity of gene repertoires among strains of the *Agrobacterium tumefaciens* species complex, we found genes specific to the species under focus, *A. fabrum* (Lassalle et al., 2011). These species-specific genes were in majority organized as clusters in genomes, and these clusters gathered genes that encoded coherent biological functions. The clustering of co-functioning clade-specific genes may be a trace of purifying selection acting to conserve the clusters in their wholeness, because the selected unit is the function collectively encoded by the constituent genes.

We thus used the Rhizobiaceae family as a model taxon to reconstruct the histories of genes in genomes. We particularly focused on the *Agrobacterium tumefaciens* species complex (*At*) for which we have an original dataset of 22 strains from ten different species, allowing a fine resolution of inference for events such as gene duplication, loss or transfers. We inferred the history of all genes in the genome dataset, and combined these scenarios of gene evolution to reconstruct ancestral genomes and their dynamics of gene content evolution. For this purpose, we designed a new phylogenetic approach for reconstruction of ancestral genomes, accounting for events of horizontal transfer and duplication of genes. This approach identifies blocks of co-transferred and co-duplicated genes. Using this regional signal in genomes provided better confidence and accuracy when dating the events in the reference species phylogeny.

When looking at the evolutionary events that marked the emergence of key clades, we observed that, as for *A. fabrum* (Lassalle et al., 2011), some transferred genes were retained in descendants and constitute genomic synapomorphies of clades. A large portion of these clade-specific genes are systematically found as clusters of co-transferred genes encoding coherent functions. We tested the hypothesis of these co-transferred clade-specific genes being domesticated and under selection for a collective function, by 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. Clade-specific genes were indeed more co-functionning than expected, supporting that purifying selection is maintaining clade-specific genes in genomes.

Altogether, these observations indicate that in the midst of the large turnover of genes in and out genomes, some gains or losses were retained by selection, potentially in relation to the adaptation to changing ecological niches. Following the history of clade diversification, these adaptive synapomorphies shaped the core genome of clades and their ecological niches.

2.4.2 Results

2.4.2.1 Genomic sequence dataset

We defined a broad sample of genomes that would reveal long-range gene exchanges and at the same time that would be dense enough to distinguish the variations that exist between close relatives. The focus was made on the *Agrobacterium tumefaciens* species complex, since we had an original dataset of sixteen new genomes plus six publicly released genomes (Wood et al., 2001; Goodner et al., 2001; Wibberg et al., 2011; Li et al., 2011; Ruffing et al., 2011; Hao et al., 2012a,b) (Table 2.1). These twenty-two genomes cover ten closely related but genomically differentiated species within the *A. tumefaciens* complex (Popoff et al., 1984; Portier et al., 2006; Costechareyre et al., 2009), with up to five isolates per species. To chose distant relatives to include to the database, we relied on phylogenies of the Alpha-Proteobacteria division made by Williams et al. (2007) and Viklund et al. (2012), and by ourselves with a more complete dataset of 131 genomes (Sup. Fig. 2.17). Our sample covers several genera of the Rhizobiales order, including every genome publicly available for the Rhizobiaceae family at the time of the database construction (spring 2012), among which *Agrobacterium*, *Rhizobium* and *Ensifer* (*Sinorhizobium*), that are thought to be frequently exchanging since they all are inhabitants of soils and rhizospheres and because many isolates harbour conjugative megaplasmids which can recombine across genera (Young et al., 2006). We excluded the genomes of intra-cellular plant pathogens Ca. *Liberibacter*, which are reduced and highly diverged, to avoid biases in reconstructions of gene trees and genome gene content. We added all available genomes of *Mesorhizobium* from the close family Phyllobacteriaceae, to complete the set of related rhizobia with genomes with different genome architectures (Slater et al., 2009). Finally, we added as an outgroup the genome of *Parvibaculum lavamentivorans* strain DS-1, from the distant family Rhodobiaceae, to ensure the outgroup rooting of the trees, but also to look for potentially interesting exchanges between *A. tumefaciens* and this particular lineage, as observed in a previous work (Lassalle et al., 2011). We thus gathered a dataset of 47 complete genome sequences that allowed to focus on the on-going divergence of *A. tumefaciens* lineages, and to capture the contribution of their rhizobial cousins with whom they may share adaptations to life in soils and plant rhizospheres.

Table 2.1: **List of 47 Rhizobiales strains used in this study** (continued next page).

Clade	Code	NCBI Taxid	Strain name	Genome size (nb. genes)
Agrobacterium tumefaciens				
Genomovar G1	AGRSH	861208	H13-3	5345
	AGRTU1	1107544	5A	5518
	ATU1A	1183421	CFBP 5771	5546
	ATU1B	1183429	S56	5627
	ATU1C	1183430	TT111	5856
Genomovar G2	ATU2A	1183436	CFBP 5494	6013
Genomovar G3	ATU3A	1183432	CFBP 6623	5378
Genomovar G4	ATU4A	1183423	B6	5875
	ATU4B	1183422	CFBP 5621	5330
	ATU4C	1183424	Kerr 14	5870
	AGRTU2	1082932	CCNWGS0286	4979
Genomovar G5	ATU5A	1183435	CFBP 6626	5332
	AGRTU3	1050720	F2	5321
Genomovar G6	ATU6A	1183431	NCPPB 925	6139
Genomovar G7	ATU7A	1183425	NCPPB 1641	6041
	ATU7B	1183426	RV3	5182
	ATU7C	1183427	Zutra 3/1	5685
Genomovar G8	AGRT5	176299	C58	5639
	AGRSP1	82789	ATCC 31749	5535
	ATU8A	1183433	J-07	5592
Genomovar G9	ATU9A	1183434	Hayward 0363	4502
Genomovar G13	ATU13	1183428	CFBP 6927	4993
Agrobacterium biovar 2	AGRVS	311402	Agrobacterium vitis S4	5389
	RHISP1	1125979	Rhizobium sp. PDO1-076	5340
Rhizobium	AGRRK	311403	R. rhizogenes K84	6684
	RHIE6	491916	R. etli CIAT 652	6109
	RHIEC	347834	R. etli CFN 42	6016
	RHIET1	993047	R. etli CNPAF512	6544
	RHIL3	216596	R. leguminosarum bv. viciae 3841	7263
	RHILS	395491	R. leguminosarum bv. trifolii WSM1325	7001
	RHILW	395492	R. leguminosarum bv. trifolii WSM2304	6415
Ensifer/Sinorhizobium	RHIME	266834	S. meliloti 1021	6234
	SINMB	698936	S. meliloti BL225C	6354
	SINME1	1107881	S. meliloti CCNWSX0020	6844
	SINMK	693982	S. meliloti AK83	6510
	SINMM	707241	S. meliloti SM11	7093
	SINMW	366394	S. medicae WSM419	6213
	SINFR1	1117943	S. fredii HH103	6787
	RHISN	394	S. fredii NGR234	6366

Table 2.1: (continued) **List of 47 Rhizobiales strains used in this study.**

	Code	NCBI Taxid	Strain name	Genome size (nb. genes)
Mesorhizobium	MESAL1	1107882	M. alhagi CCNWXJ12-2	7184
	MESAM1	1082933	M. amorphae CCNWGS0123	7075
	MESAU1	754035	M. australicum WSM2073	5934
	MESCW	765698	M. ciceri biovar biserrulae WSM1271	6264
	MESOW	536019	M. opportunistum WSM2075	6508
	RHILO	266835	M. loti MAFF303099	7281
	MESSB	266779	Chelativorans sp. BNC1	4543
Parvibaculum	PARL1	402881	P. lavamentivorans DS-1	3636

2.4.2.2 Species phylogeny

Previous works showed that using concatenation of core genes was a good way to recover a vertical signal (Wolf et al., 2001), because the accumulation of signal of independent genes make the vertical history emerge even in the face of multiple horizontal transfers. Indeed, comparison of many different gene trees showed HGT are mostly randomly distributed and do not hide the underlying vertical signal (Puigbò et al., 2009). Jacknife is an intermediate approach between the concatenation of core genes and the comparison of different gene trees, that allows to recover the basal vertical signal of descent, while accounting for gene variability. Indeed, it was shown that a jacknife approach allowed to recover the tree that minimizes the number of convergent events of horizontal gene transfers in genomes (Abby et al., 2012). Using an original approach based on the phylogenetic framework of Abby et al. (2012), we used the phylogeny of species based on the core genome as a reference for the reconciliation of genome and gene histories.

From the 47 genome sequence dataset, we built a database of homologous gene families using the Hogenom procedure (Penel et al., 2009). We defined the unicopy core genome of this dataset as those gene families found in exactly one copy in our 47 genomes, and found 455 homologous gene families matching this criterion. The phylogeny of the genomes of the 47 Rhizobiales was obtained using an approach of jacknife sampling of the unicopy core genome. From the set of unicopy core 455 gene families, we made 500 draws without replacement of 25 gene alignments, which were each concatenated and used to infer a ML tree. The reference phylogeny was obtained by making the consensus of this sample of trees, and the branch supports were derived from the frequency of bipartitions in the sample. The jacknife supports are thus indicative of (but not proportional to) the fraction of genes supporting the main topological signal caught in concatenates, presented as a reference (Fig. 2.3). This reference tree recovered the monophyly of all previously described species (Popoff et al., 1984; Portier et al., 2006; Costechareyre et al., 2010; Shams et al., 2013) and compared to alternative topologies, it appeared the best representative of the histories of all gene families in the pangenome of our dataset (Sup. Fig. 2.19).

Focusing on the *Agrobacterium tumefaciens* species complex (*At*), the jacknife support indicates that monophyly of all genomic species is well supported. In addition, some grouping of higher order are found with high support: genomovar G8 with genomovar G6 (hereafter named [G6-G8] clade), G5 with G13, ([G5-G13] clade), G1 with [G5-G13] ([G1-G5-G13] clade), G7 with G9 ([G7-G9] clade), and G4 with [G7-G9] ([G4-G7-G9] clade). Other deep splits are not stable among genes, namely the position of genomovar G2 and of [G6-G8] clade relative to *At* root. Also, no high support is recovered for the relative positioning of strains within species, indicating that members of a same species are frequently recombining.

However, one must know that this support based on gene sampling is very conservative compared to those based on site sampling such as classing bootstrap or SH-like supports. If genes tend to tell different stories but a robust genomic signal is nonetheless present interspersed in genes, a simple concatenate approach will recover high bootstrap supports. On the contrary, jacknife supports will reveal the existence of discord between genes, even if the background signal is unambiguously supporting the main topology. This is the case here, as the supports obtained by bootstrapping sites on the full 455 core gene concatenate are high (Sup. Fig. 2.18), even for the position of G2 and of [G6-G8] clade which are found flipped around the root of *At* compared to the consensus of gene samples (Fig. 2.3).

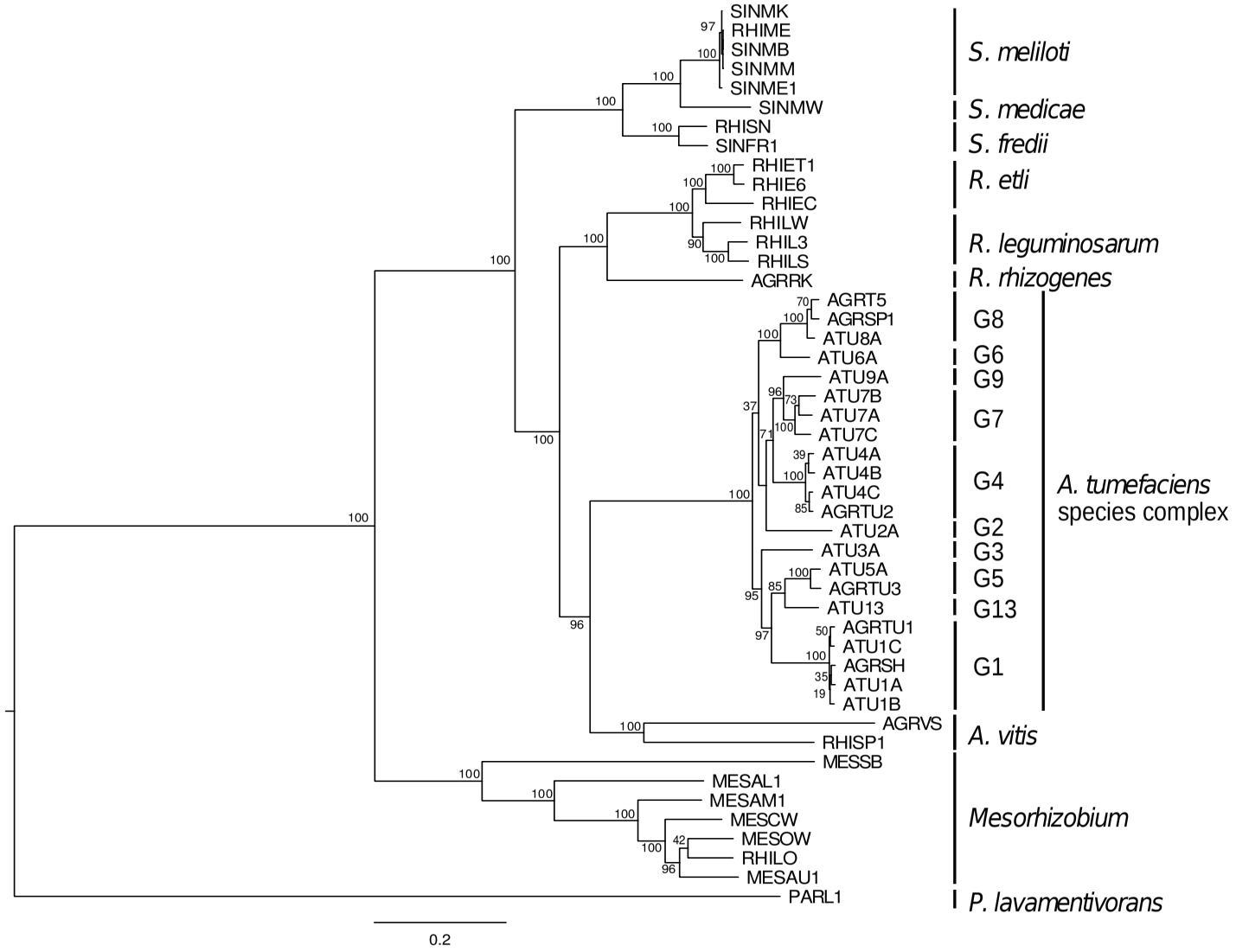


Figure 2.3: **Rereference phylogeny of Rhizobiales history.**
Obtained by consensus of ML trees built from concatenates of 500 jacknife sampling of 25 genes among the 455 unicopy genes from the core of the 47 genomes. Supports of short branches within *S.meliloti* that were omitted on the figure were all ≥ 0.9.

2.4.2.3 Reconstruction of ancestral genomes

Reconciliation of genome and gene tree histories The genes that constitute bacterial genomes can be considered as independent evolving units, whose histories are multiple and diverse. The multiple gene histories can be integrated to tell the global history of genomes and of the organisms bearing them. This corresponds with finding a tree representing the history of vertical descent in genomes and using it as a reference to be compared to individual gene histories. The incongruences between reference and gene histories can then be interpreted as events of duplication, loss and horizontal gene transfer (HGT) of genes, so that the inferred scenarios of gene evolution reconcile gene trees with the reference species tree. The sum of reconciliations of all genes can then be translated into a scenario of evolution of genes in genomes.

To describe the diversity of genes in genomes, we used gene families defined with the Hogenom procedure (Penel et al., 2009), which are homologous gene clusters containing potentially many paralogs, and can hence be very large and have complex histories of duplication, transfer and loss. Reconstructing such histories can be challenging, because many scenarios are possible in the face of the gene tree topology, notably when the latter contains errors of reconstruction. Reconstructing the history of orthologous gene families can thus appear more pragmatic (Makarova et al., 2006). However, homologous gene families provide more complete information about the nature and origin of evolutionary events – origination, duplications or transfers (ODT) – that generated gene lineages. Because we were interested in these processes that shape genomes, we developed an original method for the reconstruction of histories of all homologous gene families in genomes.

This pipeline combines several phylogenetic methods to detect horizontal gene transfers in gene trees (Fig. 2.4, step 1-4), while accounting for the complexity of multi-copy homologous gene families. We first used Prunier (Abby et al., 2010), a program that detects topological incongruences between gene and species trees characterized by strong phylogenetic support, and reconciles them by inferring a parsimonious scenario of HGT events. This method, however, was originally designed to deal with unicopy gene families. Because we tackled the more challenging problem of reconciling all – uni- or multi-copy – gene families, we used Prunier on multiple unicopy subtrees extracted from the complete gene trees, searching to reconcile the history of each gene subtree with the species tree by inferring gene transfer events (Fig. 2.4, step 2). The local subtree reconciliations were then integrated to yield a first coherently reconciled gene tree (Fig. 2.4, step 3).

Prunier provides a very conservative detection of HGT (Abby et al., 2010), but does not take into account losses. Indeed, horizontal transfers can be recognized from topological conflict between the gene and species trees, but also from a heterogeneous distribution of genes in genomes that would induce an aberrant pattern of duplication and/or losses under a vertical model of evolution. Therefore, to complete the backbone of reconciliation provided by Prunier, we used the 'TPMS-XD' algorithm developed by Bigot et al. (2013) to iteratively search for additional topological incongruences that had lower phylogenetic support but whose

recognition as transfer events provided a global scenario more parsimonious on duplications and losses.

Having confidently identified duplications and horizontal transfers leading to the emergence of new gene lineages, we could define subfamilies of orthologs nested in homologous gene families (see Methods, Fig. 2.4, step 5). Finally, we used the program Count (Csűrös, 2008) to detect cryptic transfer events from the profile of occurrence of orthologous genes, i.e. transfers that explained heterogeneous profiles of gene occurrence without topological incongruence as evidence, again minimizing the number of inferred losses (Fig. 2.4, step 6).

Each reconciliation of a gene trees corresponds to an evolutionary scenario in the species tree, where presence/absence states and the duplication, transfer and speciation events are mapped (Fig. 2.4, step 7). In particular, transfer events are characterized by the identification of both a donor and a receiver, which specifies the direction of the transfer. However, because we intended to reconcile histories of all gene families in the genome, we placed ourselves in the general case of gene histories with multiple copies punctuated by multiple duplication, horizontal transfer and loss events. In these cases, the combination of ODT events can be interpreted with several alternative evolutionary scenarios that differ by the number of hidden loss events and by the location of ODT events in the reference tree (Fig. 2.5). A reasonable way to choose a scenario among all possible ones for a gene tree is to take the most parsimonious in convergent events of loss. It is however difficult to estimate a relevant cost for each event, notably that of losses relative to transfer and duplications. In addition, arbitrary costs would not account for the variability of selective pressures experienced by diverse gene families and lineages within families.

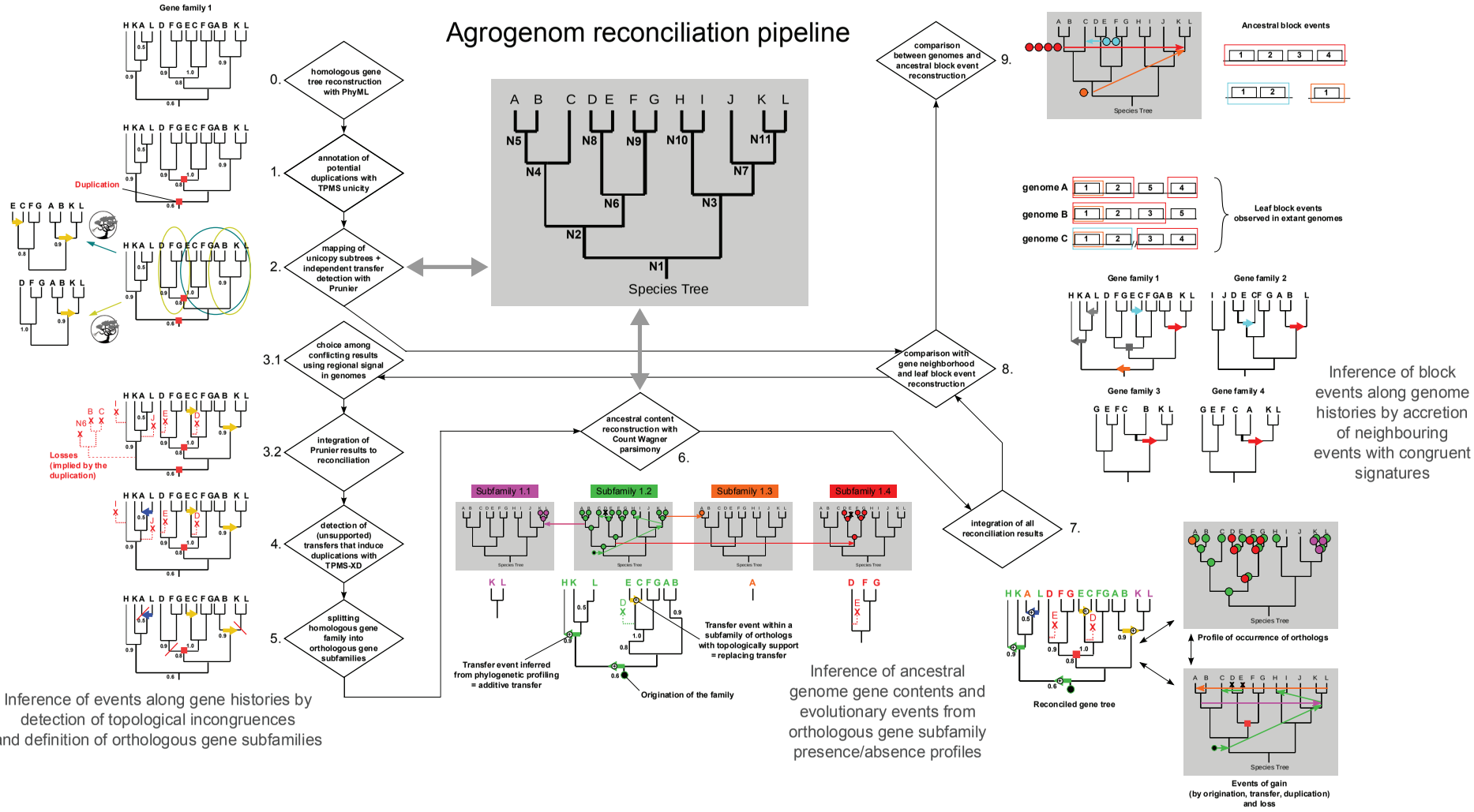


Figure 2.4: **Bioinformatic pipeline for reconciliation of gene and genome histories.**
(1-5) Detection of horizontal transfer events; (6,7) Inference of gene content of ancestral genomes and integration of reconciliations with evolutionary scenarios; (8,9) Reconstruction of blocks of co-evolved genes. For more details, see Materials and Methods

Regional amalgamation of gene histories refines the precision of reconciliations Another source of phylogenetic signal for the reconstruction of scenarios of reconciliation lies in the regional organization of genomes. Indeed, several genes can be co-transferred or co-duplicated in an unique event. Recognizing neighbouring gene presenting the same origination, duplication or transfer pattern as a unique event allows to minimize the global number of inferred ODT events (Fig. 2.5). Doing so, a common scenario is inferred for the block event. This common scenario may not be the most parsimonious in losses for individual gene families (Fig. 2.5). However, it seems more reasonable to minimize the number of ODT events by factorizing neighbour events, because this regional pattern is not likely to happen by chance (see Discussion).

We thus scanned the genomes and the database of reconciled gene trees to find neighbour genes linked to duplication or transfer events with compatible scenarios among the set of possible ones (Fig. 2.5). We used a greedy algorithm that amalgamates the scenarios of neighbouring genes when they show compatible coordinates in the reference tree (see Sup. Mat. 2.30), and thus allows to significantly refine a large number of scenarios (Sup. Fig. 2.23).

Doing so, we reconstructed block events, i.e. unique events involving blocks of co-evolved neighbour genes (Fig. 2.4, step 8-9). We found numerous block events in genomes 2.2, with 17.3% of transfer events and 24.5% of duplication events involving at least two genes. This demonstrates the necessity of considering as unique events those involving several genes in order to correctly estimate the dynamics of gene flow in genomes. Though the large majority of transfers involve only one gene (> 90%), we identified several thousands of transfer events involving short fragments of 2 to 6 genes, and hundreds of block transfers of a dozen genes or more. Some block events could involve as many as 25 consecutive genes in an extant genome (Sup. Fig. 2.21). We reconstructed the corresponding blocks of ancestral genes that were hypothetically transferred between ancestral genomes. It appeared that ancestrally transferred blocks could have been much larger than extant ones, showing how frequently rearrangements and partial losses has dismantled syntenic blocks in descendant genomes resulting from ancient transfers.

Indeed, we found many large blocks of transferred genes, that could span more than a hundred genes in ancestral or extant genomes (see selected cases in Sup. Mat. 2.7.6). There is a notable case of transfer of a 125-kb integrative and conjugative element (ICE) between strains G4-Kerr14 and G7-Zutra 3/1. It contains several cargo genes, many of which are involved in the uptake and degradation of phenolic and sugar derivatives, suggesting that this genomic island brings potential selective advantages to these two strains.

There is a long list of events involving transfers of large DNA segments, especially recent exchange among strains. More ancient large transfer events were much more rarely identified; one particular event of interest is the transfer between the ancestors of genomovars G1 and

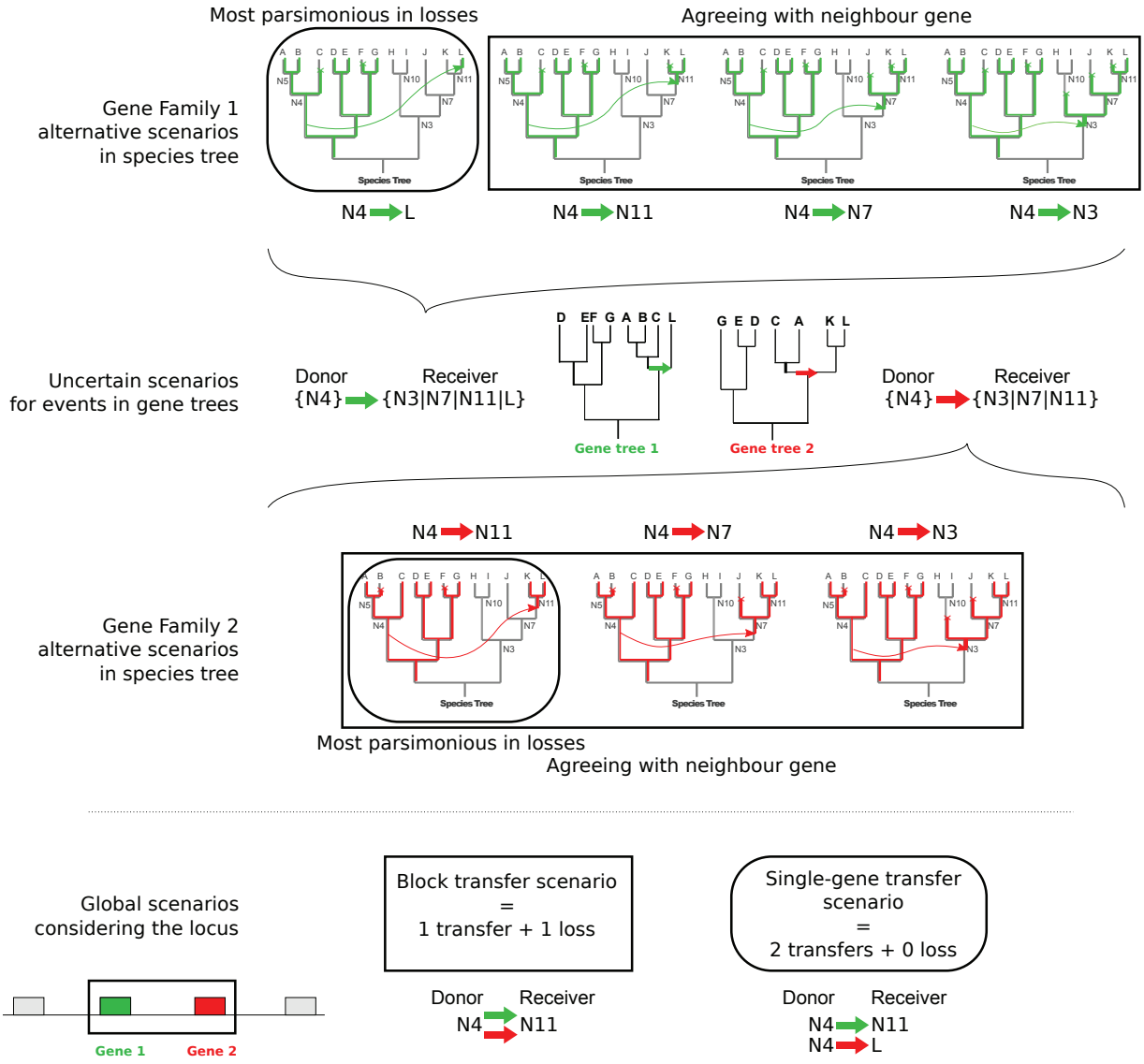


Figure 2.5: **Gene-wise vs. regional reconciliation.**
The transfers inferred in reconciled gene trees 1 and 2 can be translated in several possible scenarios in the species tree that each involve different (donor, receiver) pairs. If we reconcile each gene family separately, the scenarios that place the receiver at the LCA of genomes positive for the gene (round frames) are chosen because they are the most parsimonious in losses. That way, the global scenario for the locus totalizes two transfers and no subsequent loss.
If one considers the possibility of the co-transfer of neighbour genes 1 and 2, a common transfer scenario (square frames) can be found, which is not necessarily the most parsimonious in loss for each gene. In this case, the global scenario for the locus totalizes one block transfer and one subsequent loss, which is the most parsimonious in transfer events.
Integrating over all reconciliations in Agrogenom database, considering block scenarios induces 2,896 additional loss events compared to considering independent gene family scenarios (over a total of 32,739 losses).

Event type	Single gene events	Single gene events eligible for block reconstruction *	Block events **	Average number of gene per block event
Origination	5189	2603	1681	1.55
Duplication	7340	4406	3035	1.45
Transfer	43233	43131	32139	1.34
(Replacing)	9271	9249	8288	1.12
(Additive)	33962	33882	25350	1.34
Total ODT	55762	50140	36855	1.36
Speciation	411766	-	-	-
Total ODTS	467528	-	-	-

Table 2.2: **Origination, Duplication, Transfer and Speciation events inferred in reconciliations of the Agrogenom database.**
The Agrogenom database gathers 281,223 genes from 47 genomes, that were clustered into 42,239 homologous gene families. Out of these, 27,547 were ORFan families – i.e. genes with no homologs – and there were 10,774 families with at least three genes for which a gene tree was computed. Through the reconciliation of the latter gene tree collection, ODTS events were assigned to a total of 467,528 gene tree nodes. We annotated duplications at 7,340 nodes (1.5%) and transfers at 43,233 nodes (9.2%), the remainder corresponding to speciations and originations – i.e. apparition of the gene family in our dataset, mapped at the root of the gene tree.
* : block events were investigated only for originations (O), duplications (D) and transfers (T), but not for speciations (S). Blocks were not investigated for O and D at deep nodes (N1, N2, N3) because of the high risk of false positives (independent neighbour events older than the dataset annotated similarly that would be spuriously recognized as block events).
** : some block transfer events can unite single gene transfers that were classified as replacing or additive individually, so the block event counts of each type do not add up to the total.

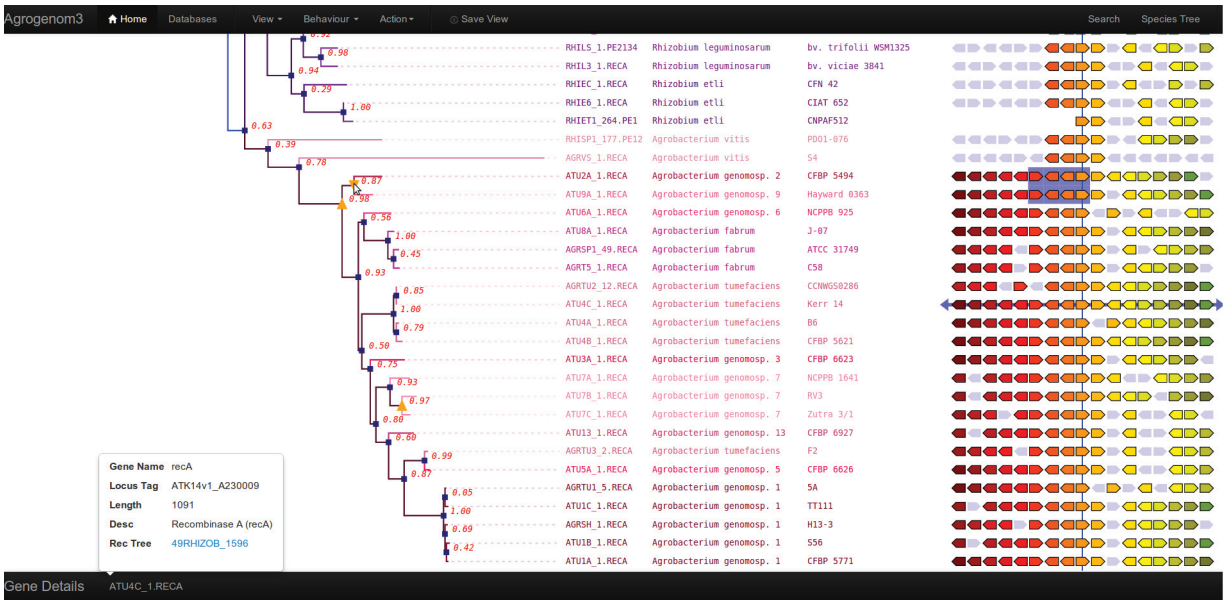


Figure 2.6: **Snapshot of the Agrogenom web interface.** View of the *recA* gene family, highlighting a block transfer from G2-CFBP 5494 to G9-Hayward 0363.

G8 of a 45-kb DNA segment encoding lipopolysaccharide (LPS) O-antigen biosynthesis, which was conserved in both clades and thus constitutes a specific character of those clades (see result section 2.4.2.8 exposing clade-specific genes).

2.4.2.4 Agrogenom database

Agrogenom is a relational database which combines all data on genes (functional annotations, gene families), genomes (position of genes, architecture in replicons . . .), the block events, the species tree (nodes, taxonomic information), reconciled gene trees (nodes, branches, ODTS events), inference analyses (parameters, scores . . .), and all other data relative to the current work.

This database is provided with a web interface (developped by Rémi Planel) that can be queried and browsed. This graphical interface presents an user-friendly, responsive graphical environment whose richness in informations is adaptable to the user demand. Notably, an interactive gene tree-centred view (Fig. 2.6) allows to browse gene histories in relation to gene syntenies in genomes. Gene trees can be manipulated (collapsing and isolation of subtrees, colouring and selection of subtrees using taxonomical queries) and many data are accessible dynamically by clicking on objects modelling biological items (gene, node with evolutionary event . . .). The block duplication and transfer events that gather several gene histories are represented on syntenies in relation to gene tree events. Gene histories can be compared by navigating between gene families using intuitive links.
This web interface for Agrogenom database is accessible at `http://phylariane.univ-lyon1.fr/db/agrogenom/3/`.

2.4.2.5 Genome histories reveal selective pressures that shaped gene contents

Dynamics of gains and losses in ancestral genomes The reconstructed history of gain and loss in ancestral genomes is not homogeneous across the tree of *A. tumefaciens*. Two apparent dynamics can be opposed: that of ancestral nodes, for which the amount of gains and losses are pretty similar, and that of leaf nodes, i.e. extant genomes, for which the number of gains, in average 1,200 genes, is much larger than the number of losses. In fact, there were 2.9-fold more gains and 1.4-fold more losses at leaves than on internal nodes (Student's *t*-tests, $p < 10^{-10}$ and $p < 0.02$, respectively). This results in larger sizes of extant genomes compared to those of inferred ancestral genomes. This tendency is in fact continuous along the phylogeny of species, though much more pronounced at the tips: there is a significant negative linear relationship between sizes of ancestral genomes and their distances to leaves (Pearson's correlation coefficient for the full dataset: $r = -0.50, p < 10^{-3}$; for *At* clade ancestors: $r = -0.65, p < 2 \cdot 10^{-3}$).

About the nature of the events that introduced new genes in genomes, we saw that it largely consisted of gene transfers. Thanks to the ancestral genome reconstruction, we could distinguish those transfers that bring new genes from those that replace already present orthologous genes (see Methods), such as those resulting from homologous recombination. A quarter of total transfers (9,271 events) were orthologous replacements, the other three quarters of transfers were additive events. Additive transfers contribute almost five times more than duplications to the total gene input in genomes, and this bias is even more accentuated in recent nodes of the species phylogeny (Fig. 2.7).

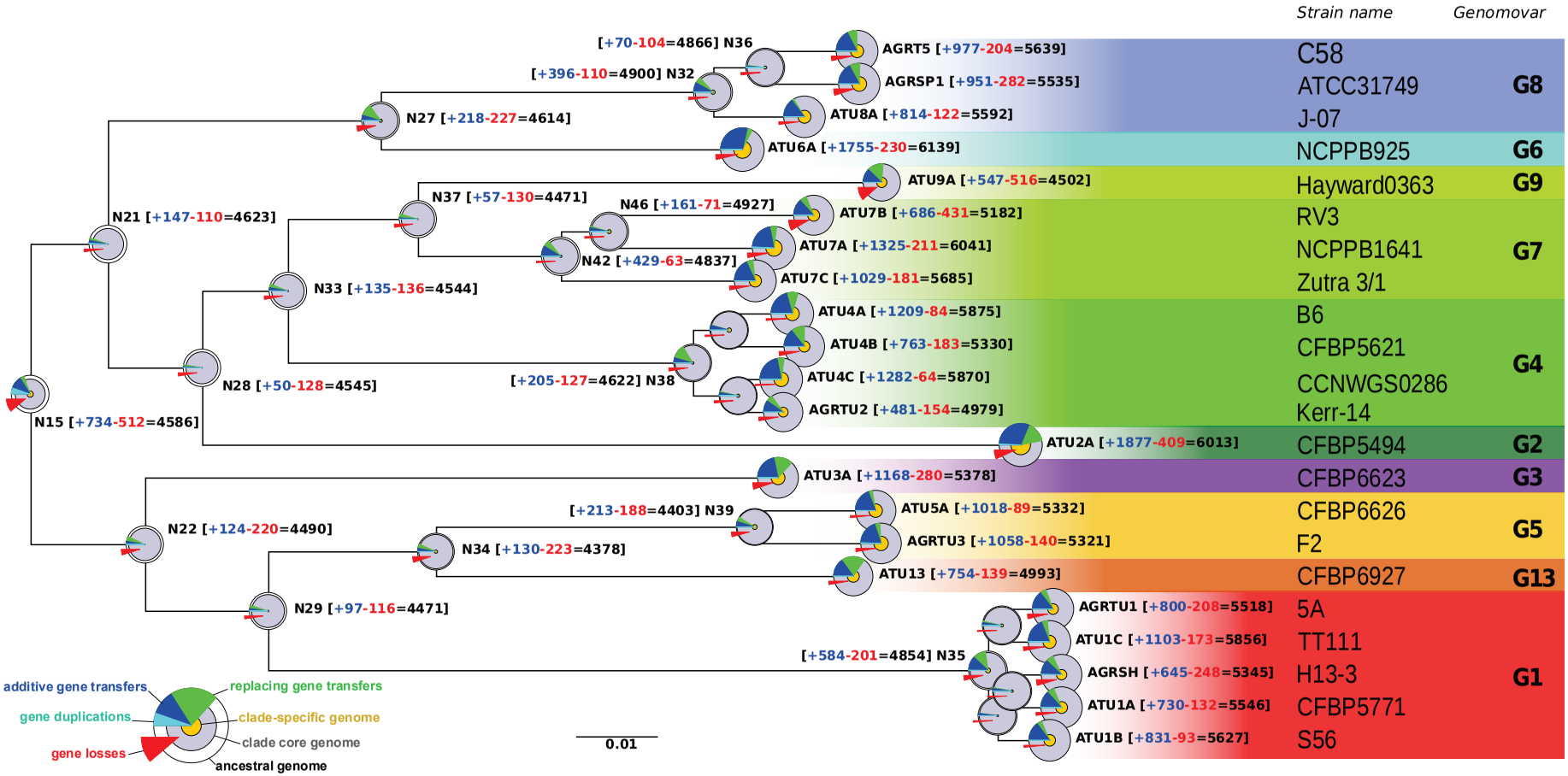


Figure 2.7: **Ancestral genome sizes and gain/loss events.**
The tree is a subtree of that presented Figure 2.3, focusing on *A. tumefaciens* clade. Neat gains (+) and losses (-) and resulting sizes (=) are indicated next to nodes. Disc at nodes schematically represent inferred ancestral genomes or actual 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).

We looked at the global patterns of genome evolution along *At* history, searching for potential general patterns in dynamics of gain and loss. If we can observe trends in the processes of genome evolution, the occurrence of outliers relatively to these trends may be indicative of fixation biases, as natural selection can cause. We thus looked at the explanatory power of the evolutionary distances on the species tree relative to the dynamics of gene repertoires. It is important to stress that these distances are not used during the ancestral reconstruction process (as reconciliation methods used here only rely on the topology of the species tree) and hence these variables are completely independent from the reconstructed gene contents and evolutionary events.

The quantity of genes gained and lost by an ancestor was best explained by the length of the branch leading to the ancestor (Fig. 2.8 A,B) (linear regression, r^2 = 0.59 and 0.32 for gains and losses, respectively), but removing the extreme point of 'N35' (genomovar G1 ancestor) makes the correlation drop (r^2 = 0.27 and 0.28). We can nonetheless observe that nodes 'N32', 'N35' and 'N42' (ancestors of genomovars G1, G8, G7, respectively) have gained more genes than predicted by the lengths of their respective branches, and that on the contrary, nodes 'N27', 'N22', 'N34' have excessively lost genes. Again, we have a limited confidence in these observations, and the excess of gains and losses may stem from complex biases linked to the shape of the species tree.

When looking at the quantity of genes gained by an ancestor and subsequently conserved in the descendant clade, we found this trait was robustly explained by the age of its ancestor (linear regression, r^2 = 0.39 or 0.41 when removing 'N35'). This relationship was better described by a decreasing exponential regression (r^2 = 0.51 or 0.50 when removing 'N35'), which reflects a process of 'survival' of genes in genomes through time (Fig. 2.8 F). We could recognize outlier genomes in this process of 'gene survival', as the nodes having the largest residuals in the exponential regression. These were, in a decreasing order of excess of conservation relative to their age, the nodes 'N27', 'N35', 'N39', 'N34' and 'N32' (Sup. Fig. 2.32). These excesses of conservation do not systematically reflect a particular excess of gains in the ancestors: 'N32' and 'N35' have indeed gained more genes than predicted by their respective branch lengths, but on the contrary 'N27', 'N39', 'N34' have not gained more genes than other – they rather have lost genes in excess (Fig. 2.8 C,D). In the latter cases, the excesses of conserved gains may thus stem from a fixation bias. Interestingly, these outlier ancestral genomes all belong to the same two clades, i.e. [G1-G5-G13] and [G6-G8]. These genes gained by ancestors of a clade and conserved afterwards characterize the clade. The nature of these potentially selected clade-specific genes is discussed further (result section 2.4.2.8).

Patterns of gene transfer: prevalence within species and among rhizobia These intense flux of genes in and out genomes certainly impact the fitness of bacterial lineages in their ecological niche. Indeed, gain of gene by HGT is an efficient way to acquire a new function that can contribute to the adaptation to the organism's niche. Notably, many bacterial lineages cohabit

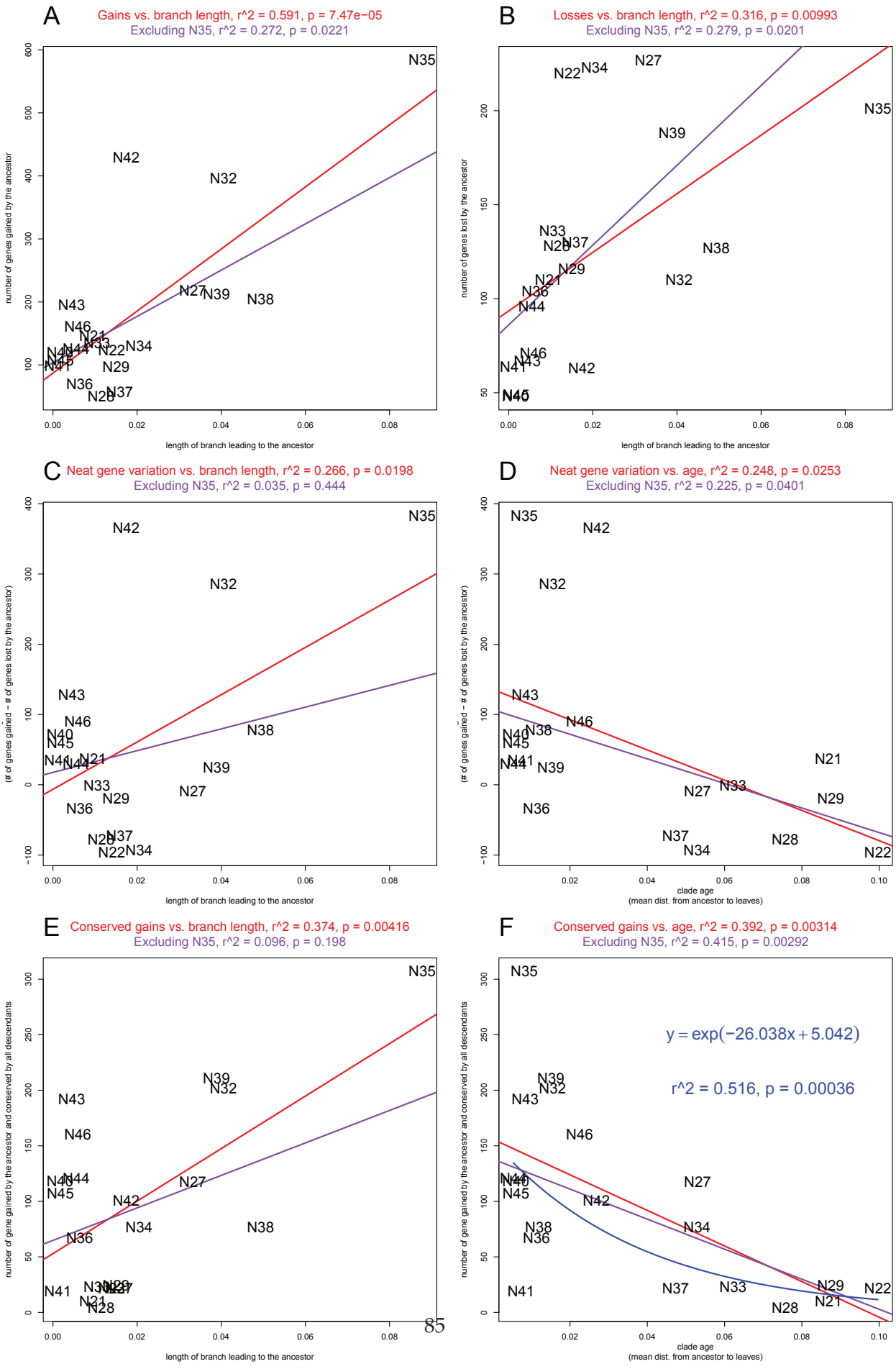


Figure 2.8: **Gene gain, loss and conservation within *A. tumefaciens* clade ancestors.**

in natural habitats as communities, and their members may share some aspects of their ecology. HGT between lineages sharing ecological traits should be more often retained, because genes exchanged between such pairs are more likely to be adaptive upon acquisition. Thus, it is tempting to look for 'highways' of gene sharing as a clue for particular ecological association of taxa.

Several recent works intended to characterize horizontal gene transfers in prokaryotes to detect possible highways of gene exchange, and to relate them to possible explanatory variables such as taxonomy, nucleotide content or lifestyles (Dagan et al., 2008; Popa et al., 2011). However, many bias can be introduced in such work, first because of the sampling of genomes and the shape of the species tree (see supplementary discussion, Sup. Mat. 2.7.4), but also for methodological reasons. The identification of transferred genes from sequence similarity only does not model the history of the exchange (i.e. does not invoke putative ancestral donor and receiver genomes) and for this reason they may misidentify the true protagonists of HGT. Also, gene transfer can involve several genes and considering the regional phylogenetic information can help to better recognize a transfer event (Williams et al., 2012). Methods that do not consider this possibility may overlook some events and at the same time may over-estimate the frequency of exchanges by counting several times a same event that involved several genes. We took into account these factors to reconstruct more accurately the patterns of gene transfers across the history of Rhizobiales.

The most evident pattern in the transfer map is the occurrence of clusters of intense exchange along the diagonal of the transfer matrix (Sup. Fig. 2.25). These intensely-exchanging clusters correspond more or less to transfers among members of the same genera: *Ensifer*/*Sinorhizobium*, *Rhizobium*, the *A. tumefaciens* complex and *Mesorhizobium*. In addition, there were some outstanding exchanges out of the diagonal, notably frequent exchanges between unrelated rhizobia, in particular between *Rhizobium leguminosarum/etli* group and *Mesorhizobium*, and among the triad formed by *Agrobacterium* biovar 3 (*A. vitis*), former *Agrobacterium* biovar 2 (*R. rhizogenes*) and *Mesorhizobium*. These cross-genus exchanges fit well with the similarity of ecologies of these groups of clades as nitrogen-fixing symbionts for *R. leguminosarum/etli* and *Mesorhizobium* or as (competing) inhabitants of the rhizosphere. Patterns were similar when considering separately the transfers inferred with Prunier, TPMS-XD or Count (data not shown), comforting the robustness of our observations.

We also obtain similar patterns of exchange between genera when looking separately at replacing (*replT*) vs. additive transfers (*addT*) (Sup. Fig. 2.26; 2.27), what could indicate that adaptation can be brought by new genes with completely new functions or by divergent alleles that can be functionally differentiated. Interestingly, there is an exception within the *A. tumefaciens* complex, with different distributions of transfers depending on their nature: *replT* occurred within narrow foci while *addT* occurred in a more diffuse way within the taxon.

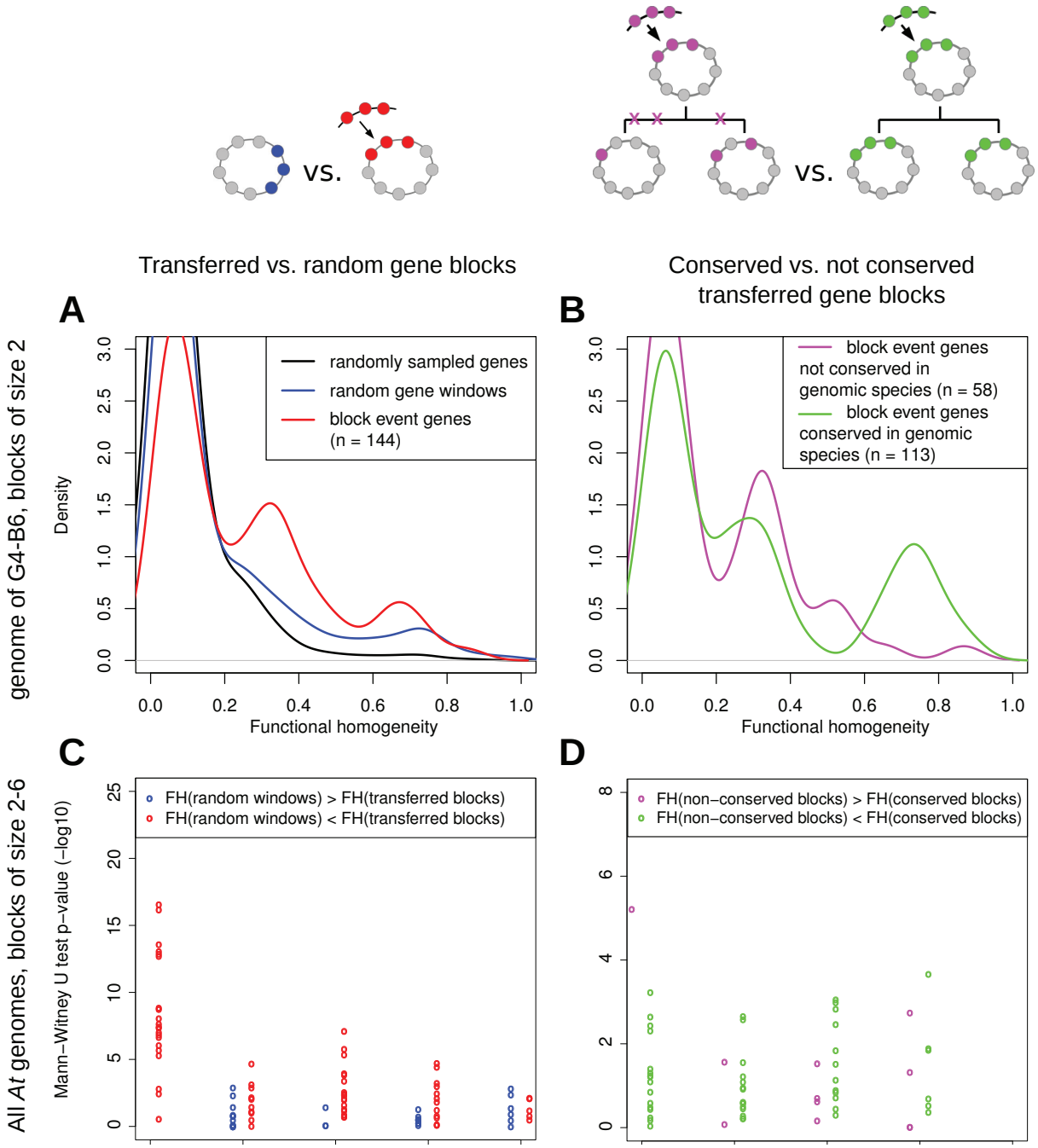


Figure 2.9: **Functional homogeneity of clusters of genes.**
(A, B) Distribution of functional homogeneities (FH) of genes within clusters; representative plots comparing clusters of two genes in the genome of strain B6 (genomovar G4) are shown. **(A)** Comparison of functional homogeneities of groups of two genes taken randomly in B6 genome (black), of any pairs of neighbour genes (blue) or of pairs of co-transferred genes (red). **(B)** Comparison of functional homogeneities of pairs of co-transferred genes from families conserved in all strains of the genomic species to which belongs the reference (green, e.g. the species above reference strain B6 is genomovar G4) or not conserved (red). Because the quality of functional annotation varies among genomes and the size of gene clusters impacts the measure of FH, statistical difference between these population of clusters was made independently for all *At* genomes at all size of blocks.
(C, D) Distribution of p-values of Mann-Whitney-Wilcoxon sum of ranks test comparing the distributions of functional homogeneities of **(C)** systematic gene windows 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 genome of *A. tumefaciens* for a given size of groups of genes (on x-axis). The colour of the point indicate the higher-FH sample (as in A,B): (C) FH(random windows) > FH(transferred blocks), blue, $n = 29$; FH(random windows) < FH(transferred blocks), red, $n = 66$; (D) FH(non-conserved blocks) > FH(conserved blocks)

Evaluation of the selection pressures acting on transferred blocks of genes Highways of gene sharing might reveal exchanges of genes that evolved for a long time in the same lineage, undergone similar selective pressures and are thus more likely to be adaptive in the same general ecological context. Although, there are thousand of genes in a genome, and not all participate to the adaptation to the same aspect of the ecological niche. It is therefore not straightforward to use the origin of a transfer as a predictor of the ecological function and of the selective pressures characterizing the genes.

However, a finer indication of the adaptive role of genes may reside in the neighbourhood of genes of the same origin. The 'selfish operon' model postulates that the neighbourhood of co-functioning genes is maintained by selection at the gene and organismal level, because their close linkage favours their co-transfer, which guaranties that the collective gene function can be selected for upon acquisition (Lawrence and Roth, 1996; Lawrence, 1999), which in turn promotes the survival of genes. There is a couple of predictions of that model that can be tested: (1) if co-functioning of co-transferred genes is the cause for their successful survival across genomes, identified blocks of transferred genes should be in average more coherent in function than groups of genes with independent histories ; (2) blocks of transferred genes that brought a selectable function should be more frequently conserved than those that code non-functional assemblies of genes.

To assess the relation between the transfer history of genes and their biological function, we computed the degree of functional homogeneity (FH) of blocks of transferred genes based on their Gene Ontology annotation (see Methods).

Plots of the distribution of FH of all transferred blocks show that most transferred blocks contain genes that are not encoding related function (FH ~0) but there are minor peaks representing transferred blocks of intermediate and high functional relatedness (e.g. in G4-B6 genome, FH ~0.35 and FH ~0.75, Fig. 2.9 A). We then compared this distribution to the distribution of FH of non-transferred blocks of genes across the genomes.

First, we found that considering groups of n genes sampled from a genome, groups made of n genes each taken at random (distant genes) were slightly less homogeneous in their functions that groups of n neighbouring genes (taking all possible gene windows of n genes), confirming we can capture the functional structure of genomes (Fig. 2.9 A). Moreover, groups of genes in random gene windows were significantly less homogeneous in their functions than blocks of transferred genes of the same size (Fig. 2.9 A,C). This does indicate a role of transfer in shaping the functional lattice of genomes.

Finally, blocks of co-transferred genes conserved in all members of a genomic species had in general higher FH than those not conserved in all members of the species (Fig. 2.9 B,D). Even though this was marginally significant (only 11/60 tests have p-values $< 10^{-2}$, among which 9/11 support FH(conserved transfers) > FH(non-conserved transfers) (Fig. 2.9 D), this indicates that at least a portion of genes conserved following acquisition by transfer correspond to groups of genes collectively coding functions under purifying selection.

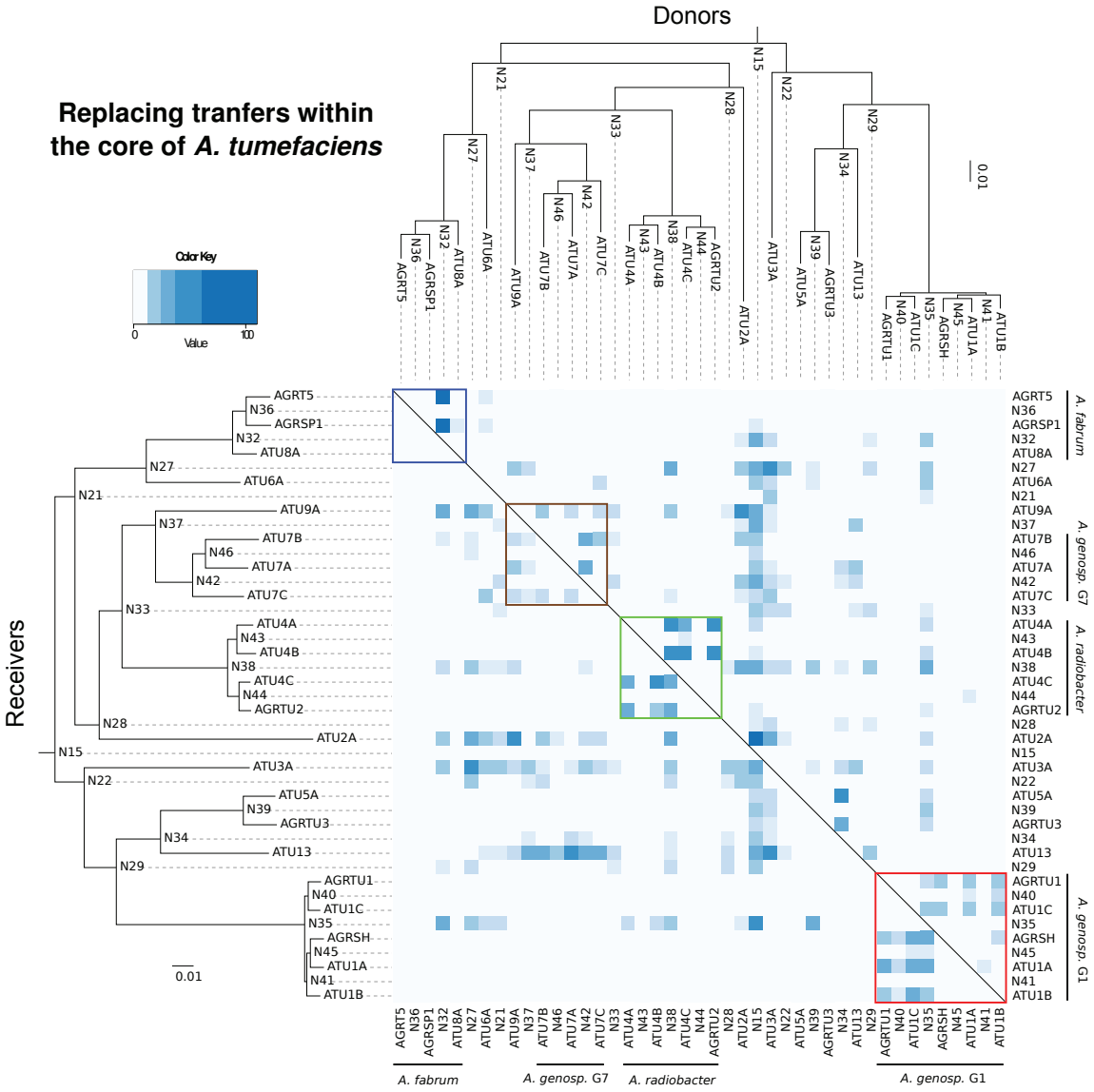


Figure 2.10: **Intensity of gene transfers whithin *A. tumefaciens*.**
Heatmap of replacing transfers that occurred between pairs of ancestral or contemporary genomes. Each dot of the matrix indicates the count of transferred genes between nodes of the reference phylogeny drawn on left and top. Rows are receivers, columns are donors. Note that to give an insight into the proportion of genomes that recombined, each gene is counted as one transfer, regardless that it was involved in an unique block event. In case of remaining uncertainty on the location of the inferred receiver and donor, the count of the event is uniformly fractioned between possible receivers / donors. Frames highlight transfers whithin clades: blue, genomovar G8 (220 total transfer events [#], with on average 10.6 transfers per branch per 1,000 genes [t/b/kg]); brown [G7-G9 clade] (# 133, 5.3 t/b/kg); green, genomovar G4 (# 413, 14.6 t/b/kg); red, genomovar G1 (# 338, 9.3 t/b/kg)

2.4.2.6 Homologous recombination maintains cohesion of species

We wanted to assess the potential role of homologous recombination (HR) as a cohesive force in during cladogenesis in *At*. We thus looked at the distribution of event of HR that marked

the core of *A. tumefaciens* genomes, using as a proxy the transfer events in our reconciliation scenarios that led to replacement of orthologous genes, i.e. replacing transfers (*replT*). As seen for the global patterns of transfers, the core genome seems to recombine predominantly among members of the same species. This is evident for genomovar G1 and G4, which are the best sampled ones (Fig. 2.10, red and green frames, respectively), and in genomovar G8 (blue frame) where there is mostly a signal for recombination with unsampled G8 strains, which are mapped to their ancestor ('N32'). The last well-sampled clade, [G7-G9], show less evidence of replacing transfers.

Out of the diagonal, many *replT* are seen coming from the ancestor of all *A. tumefaciens*, indicating an intense recombination with unsampled basal lineages or close relatives such as *A. larymoorei*. The only representative strains of genomovars G2, G3 and G13 are all seen as receiving many genes from the distant clades [G6-G8] and [G7-G9]. Though one must consider that transfers mapped to these isolated strains are the sum of what happened during a large evolutionary time, this shows a substantial bias in the origin of gene transfers. The dataset of *replT* restricted to those confidently inferred by Prunier is similar (data not shown), rejecting the possibility of this being caused by the irresolutions of deep gene tree nodes. However, the prevalence of recombination outside species is generally minor, since *replT* events inside clades accounted for 23%, 60%, 67%, 70% and of total *replT* event received in [G7-G9], G4, G1, and G8 clades, respectively. This pattern is consistent with previous experiments of Costechareyre et al. (2009), who showed that there was a log-linear decrease of recombination efficacy with distance between genomes, but that there was no specific barrier to recombination between genomic species of *At*.

We tested if the global pattern of *replT* that we observed was congruent with a model of efficacy of homologous recombination that decreases with phylogenetic distance (Roberts and Cohan, 1993; Costechareyre et al., 2009). Comparing the frequency of *replT* events with the distance of (donor, receiver) pairs in the species tree showed that this was not the case, as neither linear nor loglinear regressions were found explicative ($r^2 < 0.03$), suggesting that this measure of recombination we used here is impacted by more complex factors.

2.4.2.7 Effect of the particular architecture of *A. tumefaciens* genomes on gene evolution

The linear chromid is more recombinogenic than the circular chromosome The genomes of *A. tumefaciens* are characterized by the occurrence of two main replicons, the circular chromosome and the linear chromid. We wanted to investigate the dynamics of the gene content of each replicon to understand the potential effect of chromosomal location on the evolution of genes. Notably, it has long been suspected that the linear topology of the chromid of *A. tumefaciens* could confer it a higher plasticity (Marri et al., 2008). We thus took advantage of our whole-genome analysis of gene transfers in *A. tumefaciens* to re-investigate this possible feature of linear chromosome.

	Size	# Gains	% Gains	# Losses	% Losses	# Repl. Trans.	% Repl. Trans.	% Rec. Core Fam.
Cc	2073	194	9.3	31	1.4	51	2.0	29
Lc	1124	229	19.1	39	3.4	53	3.3	35
$p <$		10^{-3}	10^{-9}	10^{-1}	10^{-5}	0.93	10^{-5}	0.02[a]

Table 2.3: Gene content plasticity and recombination of Cc vs. Lc.

Size: average across all nodes of *At* phylogeny of the number of genes that could be unambiguously mapped to a replicon. Gains, Losses, Repl. Trans.: average across all nodes of *At* phylogeny of gain, loss or replacing transfer events, respectively, in absolute count (#) or percentage relative to genome size (%). % Rec. Core Fam. : percentage of *At* core gene families exclusively located on one replicon that show signatures of recombination (as detected by PHI (Bruen et al., 2006)).

p, p-values resulting from Student's *t*-tests, except (a) which is the result of a Chi-squered test.

To characterize potential differences in genomic plasticity, we measured the frequencies of recombination of the circular chromosome (Cc) versus linear chromosome (Lc), considering various signatures associated to homologous recombination (HR) events. For this purpose, the location of genes on the replicons of ancestral genomes of the *At* species complex was reconstructed using the evolutionary scenarios of the gene families as guides to a parsimony approach.

First, we looked for variation in gene content in relation to the position of genes on replicons (Table 2.3). We found that gene gains and losses were in absolute counts more frequent on the linear chromid compared to the circular chromosome, with an average ratio of 1.40-fold more gains and 1.44-fold more losses (paired Student's *t*-tests, $p < 10^{-3}$ for gains and $p < 10^{-1}$ for losses). This effect was even stronger when taking into account the difference in size of the replicons (Lc size was 62% of Cc size in average): the linear chromid was the place of 2.59-fold more gains per gene family and 2.66-fold more losses per gene family (paired Student's *t*-tests, $p < 10^{-9}$ for gains and $p < 10^{-5}$ for losses). This shows that the linear chromid has indeed a more rapid gene turnover, potentially due to more frequent large HR events. This ratio of 1.40 genes gained on the Lc per gene gained on the Cc is stable among all genomes (ancestral and extant) of *At* (Pearson's correlation, $r = 0.96$, $p < 10^{-16}$), suggesting there is an intrinsic property of the Lc to integrate new genes more easily than the Cc (Sup. Fig. 2.31).

Considering the tendency of genomes to retain genes after their acquisitions, the two replicons are remarkably similar: the proportions of genes gained by clade ancestors on the Cc or Lc and conserved in the progeny are highly correlated (Pearson's correlation, $r = 0.95$, $p < 10^{-10}$) and their ratio of 0.85 is not statistically different from 1, though it would indicate a slight greater capacity of the linear chromosome to retain the genes it acquired. This is surprising, given the more elevated rates of losses experienced by the Lc, and suggests this replicon may be itself partitioned into stable and dynamic compartments.

We then looked at the distribution among replicons of gene-scale events of homologous recombination, as indicated by *replT* events in our reconciliation scenarios (Table 2.3). Summing

over the history of *A. tumefaciens* clade, there were 2,173 and 2,274 replacing transfers that took place on Cc and Lc, respectively. Adjusting for the size of the replicons, there is a significant excess of *replT* events on the linear chromid (0.020 and 0.033 *replT*/gene/reference tree node for Cc and Lc, respectively; paired Student's *t*-test, $p < 10^{-5}$).

In addition, we searched for signatures of intra-genic events of allelic conversion in alignments of orthologs belonging to the unicopy core-genome of *A. tumefaciens* and found exclusively on either the Cc or the Lc. 29% and 35% of genes were found recombinant on the Cc and Lc, respectively; this difference is significant ($\chi^2 = 5.6456, p < 0.02$).

Altogether, this shows a larger plasticity of gene content and a higher prevalence of HR on the linear chromid compared to the circular chromosome.

Migration of genes between replicons accross *A. tumefaciens* history Location on one or the other replicon seems to have a significant impact on the evolutionary dynamics of genes, and this probably reflects different regimes of selection experienced by different replicons, as previously suggested by Cooper et al. (2010). Translocation of genes among replicons may then have an effect on their evolutionary fate . We thus looked for changes in gene location throughout the history of *At*. We notably distinguished translocation events that occurred in the common ancestor of all *At*. Indeed, this ancestor acquired the *telA* gene coding a pro-telomerase, what led to the linearization of the chromid which was ancestrally circular (Ramirez-Bahena et al., in press; Annex 5.2.2). This change of topology was already shown to coincide with many gene translocations to the chromid (Slater et al., 2009).

When looking at the global pattern of exchange among replicons, it is striking that there are two main routes: between the circular chromosome (Cc) and the linear chromid (Lc) and between the Lc and the At plasmid (pAt). This suggests location of genes may evolve to fit a gradient of evolutionary properties found in this molecule series: from the stable Cc to the dynamic and conjugative pAt, with the Lc as a staging intermediate point.

The translocations that happened on the branch leading to the ancestor of *A. tumefaciens* complex account for most of this effect: 267 genes moved from Cc to Lc when the chromid linearized (Table 2.4). Most of these genes are located in several large gene clusters of (Sup. Tab. 2.7). The translocated gene clusters match in general those previously defined by Slater et al. (2009); the differences are most probably reflecting our use of several genomic sequences to reconstruct ancestral genomes when Slater et al. (2009) used only that of strain C58. These clusters include many genes encoding essential proteins involved in the central cell machinery, such as several amino-acyl-tRNA synthetases, ribosomal proteins, the enzymes of the cobalamin biosynthesis pathway and the secretion protein SecA. There were also genes encoding peripheral metabolism functions, among which xylose transport and catabolism and succinoglycan biosynthesis. The translocation of functions essential to the cell or potentially

		To Cc	Lc	pAt	pTi	p
From	Cc	-	267 / 61	1 / 22	0 / 0	0 / 13
	Lc	21 / 97	-	5 / 182	1 / 5	0 / 45
	pAt	0 / 3	0 / 51	-	0 / 1	0 / 22
	pTi	0 / 0	0 / 0	0 / 3	-	0 / 4
	p	1 / 3	2 / 8	3 / 26	1 / 23	-

Table 2.4: **Counts of gene translocations among replicons over the history of *A. tumefaciens*.** Figures on the left indicate translocations that happened on the branch leading to *At* common ancestor (code N15), figures on the right indicate events that happened afterwards. pTi are defined as megaplasmids of strains experimentally shown as tumorigenic (data not shown) that can be distinguished from other megaplasmids by the presence of pathogeny-associated genes (like T-DNA, auxine synthesis genes, characterized opine synthesis/degradation genes). pAt are defined as those megaplasmids that do not show signs of being tumorigenic. p design other plasmids (smaller ones or megaplasmids with no clear identity).

ecologically important were likely seminal in fixing irremediably the chromid as an indispensable element of the genome of the common ancestor of *A. tumefaciens.*

Excluding these major translocations at the origin of *At* clade, and accounting for the size of the replicon (Sup. Table 2.6), the pattern of gene migration along the history of *A. tumefaciens* appears to be directed toward larger replicons, as there are in 4.1-fold more translocations from the Lc to the Cc than the inverse, and 1.6-fold more from the pAt to the Lc than the inverse (Sup. Table 2.6). This might correspond to the course of domestication of genes toward stable replicons, but this observation must be considered with caution, because an explicit modeling of gene migrations taking into account the underlying phylogenetic tree might give different results.

In opposition to what happened in the ancestor of *At*, there were few gene translocations during the evolution of *At* subclades that characterized them. Indeed, most gene families that migrated from a replicon to another in one lineage were also translocated in other lineages. For instance, the locus containing the (*cfa*) operon (Atu1974-Atu1979 in C58 genome) has a heterogeneous location among genomic species of *A. tumefaciens*, signing multiple independent translocations from the Cc to the Lc in G1, G2 and [G7-G9] clade (Sup. Fig.2.28). The tRNA genes flanking this locus on one side and a hypothetical protein similar to recombinases on the other side suggests this mobile cassette is an integron.

2.4.2.8 Clade-specific genes: insights into the possible ecological speciation of clade ancestors

We defined clade-specific presence/absence of genes using an automated method for recognition of contrasted occurrence profiles between related clades. This was done by spotting ancestral gene gains or losses that resulted in conserved presence or absence in the descendant clade. Such pattern distinguishes a clade from its sister clade, i.e. clade-specific presence/absence of genes constitute synapomorphies. We could identify parallel gains/losses of homologous genes in distant clades, notably in case of transfer from one clade ancestor to another. This revealed the specific sharing of genes between non-sister genomic species of *A. tumefaciens* complex. A subset of internal nodes of *At* phylogeny are represented by several closely related extant genomes, and for this reason were particularly amenable for the study of clade-specific gene repertoires. Functional description of gene repertoires specific to genomovars G1 and G8 , and to the [G6-G8], [G5-G13] and *At* clades are described shortly in the following section, and are further detailed in the Supplementary Material (section 2.7.5). Clade-specific genes were often located in relatively large clusters encoding coherent functions, that are summarized Table 2.5.

Cluster label	Reference organism name	Locus	Non-specific occurrence	Functional description	Figure reference, # box
G1-specific	G1-H13-3				
AtSp1		AGROH133_06050–AGROH133_06053		two-component transduction system x2	
AtSp2		AGROH133_08627–AGROH133_08661		chemotaxis regulation, phenolic compound catabolism, amino-acid transporter	2.11, 1
AtSp3		AGROH133_12531–AGROH133_12572	shared by G2-CFBP 5494	agmatinase, amino-acid ABC transporter, regulators	2.11, 2
AtSp4		AGROH133_13101–AGROH133_13154	shared by G8-ATCC31749 and G7-Zutra3/1	aromatic / phenolic compound uptake catabolism	
AtSp5		AGROH133_13179–AGROH133_13208	shared by G2-CFBP 5494	monooxygenase, diguanylate cyclase/phosphodiesterase, proline/glycine/betaine ABC transporter	
AtSp6		AGROH133_13327–AGROH133_13329		tetrameric sarcosine-oxidase	
AtSp7		AGROH133_14153–AGROH133_14157	shared by G4-CFBP 5621 and G7-RV3	aromatic compound downstream degradation	2.11, 3
AtSp8		AGROH133_14082–AGROH133_14097		antibiotic / toxic compound resistance	
AtSp9		AGROH133_15088–AGROH133_15100		Alkylated aromatic compound degradation	2.11, 4
AtSp10		AGROH133_15094–AGROH133_15098		alkylhydroperoxidase	
AtSp11		AGROH133_15345–AGROH133_15356	shared by G9-Hayward0363	Type I secretion system, putative adhesin	
G1+G8-specific	G1-H13-3				
AtSp12		AGROH133_09434–AGROH133_09437		outer membrane lipoprotein	2.11, 6
AtSp13		AGROH133_09517–AGROH133_09523		two-component transduction system FeuPQ	
AtSp14		AGROH133_10151–AGROH133_10254		O-antigen biosynthesis (related to Brucella sp.)	2.11, 5
AtSp15		AGROH133_15086–AGROH133_15086		curdlan synthesis	2.11, 8
AtSp16 *		AGROH133_15041–AGROH133_15058		two-component transduction system KaiC	
AtSp17		AGROH133_15108–AGROH133_15138	missing in G8-ATCC31749	deoxyribose uptake and assimilation	2.11, 7
G1+[G6-G8]-specific	G1-H13-3				
AtSp18		AGROH133_13355–AGROH133_13364		D-glucuronate uptake and degradation	
AtSp19		AGROH133_15061–AGROH133_15064		Methionine/cysteine metabolism	

* : homologs in G8 genomes are not clusterred

Table 2.5: **Location and functional description of clade-specific gene clusters in *A. tumefaciens* genomes** (continued p97)

Cluster label	Reference organism name	Locus	Former label (in Lassalle et al., 2011)	Non-specific occurrence	Functional description	Figure reference, # box
G8-specific	G8-C58					
AtSp20		Atu0628–Atu0630			Two-component sensor	
AtSp21		Atu1410–Atu1422	SpG8-1b		ferulic acid uptake & catabolism	2.12, 1
AtSp22		Atu3072	SpG8-2b		non-ribosomal peptide synthase	
AtSp23		Atu3946–Atu3952	SpG8-5		Opine-like compounds catabolism	2.12, 2
AtSp24		Atu4199–Atu4206	SpG8-6a		Drug/toxic (tetracycline) resistance	2.12, 3
AtSp25		Atu4219–Atu4221	SpG8-6b		Drug/toxic (acriflavine) resistance, sarcosine oxidase	2.12, 3
AtSp26		Atu4295–Atu4307	SpG8-7b		Environmental signal sensing/transduction	2.12, 4
AtSp27		Atu5418–Atu5430		shared by G2-CFBP 5494	Toxic extrusion / secondary metabolite secretion	2.12, 14
AtSp28		Atu5497–Atu5506		partially shared by G2-CFBP 5494	xanthine/cyclic compound degradation, two-component sensor	2.12, 15
[G6-G8]--specific	G8-C58					
AtSp29		Atu1399–Atu1406	SpG8-1a		Sugar uptake and metabolism	2.12, 9
AtSp30		Atu3666–Atu3686	SpG8-3	shared by G1-TT111	NRPS enzymes, siderophore biosynthesis	2.12, 10
AtSp31		Atu3808–Atu3828	SpG8-4		undefined sugar transformation pathway	2.12, 11
AtSp32		Atu5473–Atu5483			Dipeptide uptake and degradation	2.12, 12
G8+G1 -specific	G8-C58					
AtSp15		Atu3054–Atu3059	SpG8-2a		curdlan synthesis	2.12, 5
AtSp13		Atu4711–Atu4713			Two-component transduction system FeuPQ	2.12, 6
AtSp14		Atu4788–Atu4817			O-antigen biosynthesis (related to Brucella sp.)	2.12, 7
AtSp12		Atu4837–Atu4839			Outer membrane lipoprotein	2.12, 8
AtSp17		Atu4841–Atu4849		missing in G8-ATCC31749	deoxyribose uptake and assimilation	
[G6-G8]+G1 -specific	G8-C58					
AtSp18		Atu3527–Atu3536			D-glucuronate uptake and degradation	2.12, 13
AtSp19		Atu3823–Atu3824			Methionine/cysteine metabolism	
Lost in [G1-G5-G13]	G8-C58					
AtSp33		Atu4381–Atu4410		Lost in [G1-G5-G13], G9-Hayward0363 and G8-ATCC31749	nitrate respiration	

Table 2.5: **Location and functional description of clade-specific gene clusters in *A. tumefaciens* genomes** (continued p101)

Genomic synapomorphies of genomovar G1 There are 75 genes present in all genomovar G1 and in no other *A. tumefaciens* strains, and 60 other genes present in all G1 and found in a heterogeneous set of other strains, totalizing 135 G1-specific genes (Sup. Table 2.9). Other genes are found specifically shared with other clades, notably with genomovar G8 (57 genes) and [G6-G8] clade (17 genes) (Sup. Table 2.9). The large majority of those specific genes are located in relatively large clusters (84/135 G1-specic genes, and 99/108 of genes specifically shared with other genomic species). Genes in these clusters are in general annotated with coherent functions (Table 2.5, Figure 2.11). Briefly, clustered G1-specific gene are annotated with functions that can be linked to a restricted set of cellular pathways: chemotaxis regulation, with a supplementary *che* operon (*che2*) in cluster AtSp2, HHSS gene in AtSp14 and several diguanylate cyclases/esterases, notably in cluster AtSp3; phenolic compound catabolism, notably in clusters AtSp2, AtSp3, AtSp7 and AtSp9; amino-acid uptake and catabolism; biofilm production and secretion with LPS O-antigen biosynthesis in AtSp14, lipoprotein in AtSp12, curdlan biosynthesis in AtSp15 and T1SS-mediated secretion of a putative adhesin in AtSp11.

Genomic synapomorphies of genomovar G8 and [G6-G8] clade The genomovar G8, for which the binomial *A. fabrum* has recently been proposed (Lassalle et al., 2011), is characterized by 57 genes present in all its representative strains and in no other *A. tumefaciens* strains, and 34 other genes present in all G8 and found in a heterogeneous set of other strains, totalizing 91 G8-specific genes (Table 2.10). Other genes are found specifically shared with other clades, notably with genomovar G1 (Table 2.10). A large fraction of those specific genes (62/91 of G8-specific genes, 69/71 genes specifically shared with other genomic species) are located in clusters of two to more than thirty contiguous genes (Table 2.5, Fig. 2.12). This specific gene repertoire and its organization in clusters generally match previous findings based on microarray hibridization experiments that focused on genomovar G8 (Lassalle et al., 2011). For the sake of coherence of denomination throughout this manuscript, these clusters will be renamed with the AtSp nomenclature used above; correspondence with former cluster denomination from Lassalle et al. (2011) is indicated 2.5.

AtSp21, is the largest G8-specific gene cluster located on the circular chromosome. It contains the operon *braCDEFG*, that encodes an ABC transporter of amino-acids with broad-range specificity, and genes coding enzymes of the ferulic acid degradation pathway, that were recently characterized for their expression and molecular functions (Campillo et al., submitted). AtSp29 cluster, which is found immediately upstream, is specific to [G6-G8] clade and is dedicated to transport and catabolism of sugars and amino-acids. It includes an enzyme (encoded by Atu1408 in C58 genome) whose predicted activity of transformation of L-sorbose into L-iditol echoes the specific ability of G8 strains to degrade L-sorbose (Vial L., Bourri M., personal communications).

On the linear chromosome, a cluster encoding curdlan exopolysaccharide biosynthesis, AtSp15, was formerly thought to be specific to genomovar G8 (Lassalle et al., 2011) but appears specifically shared by the only representative strain of genomovar G2 and all strains of

genomovar G1. Several other G8-specific or [G6-G8]-specific genes are found on the linear chromosome (Table 2.5) consistently with previous results (Lassalle et al., 2011), which encode the uptake and degradation of complex amino-acids (AtSp23), the catabolism (AtSp21, AtSp26, AtSp28) or the modifcation and extrusion of aromatic compounds (AtSp24, AtSp25), and the perception of multiple chemical and mechanic environmental signals (AtSp26). In addition, we found clade-specific gene clusters on the At plasmid of G8 strains: G8-specific gene cluster AtSp28 codes the degradation of xanthine or another related cyclic compound and [G6-G8]-specific gene cluster AtSp32 putatively codes the uptake and degradation of dipeptides that include an aromatic amino-acid.

Genomic synapomorphies of [G5-G13] clade Thirty-five genes are specific to the clade grouping genomovars G5 and G13 (63 when including genes occurring heterogeneoously in other *At* strains). Among those, there are three gene clusters (AtSp34-36) encoding potential ecologically important functions, namely a transporter of oligopeptides, a peptide methionine sulfoxide reductase (involved in response to oxidative stress), and the complete pathway for degradation of phenylacetate. The latter metabolic ability has been verified experimentally in G5 and G13 strains and shown to provide a growth advantage in presence of phenylacetate (Vial L., personal communications).

Genomic synapomorphies of [G1-G5-G13] clade The large cluster conserved in agrobacteria (AtSp33, Atu4381-Atu4410 in C58 genome) that encodes nitrate respiration (denitrification) pathway, including *nir*, *nor*, *nnr* and *nap* operons, was lost in this clade. This gene cluster was parallelly lost by strains G9-NCPPB925 and G8-ATCC31749. These strains devoid of the denitrification pathway may be selectively disadvantaged under certain anaerobic conditions. This is not certain however, because other anaerobic respiration pathways are predicted to be conserved in all *A. tumefaciens*, notably the fumarate respiration pathway.

Genomic synapomorphies of the *A. tumefaciens* complex There are 120 genes present exclusively in all *A. tumefaciens* strains, and 165 genes that are also found present in at most two other distant Rhizobiales, i.e. not directly related like *A. vitis*. It appears there is no group of genes encoding concerted functions, and the larger clusters of *At*-specific genes are six genes long. They are distributed among diverse functional categories that span the ones represented in the ancestral genome (no specific enrichment in GO terms, data not shown). Notably, several *At*-specific genes are involved in the central intermediary metabolism, the biosynthesis of cell wall and outer membrane and the replication and translation machineries. In fact, specific gene gains related to these essential processes corresponded in some case to non-homologous replacement of enzymatic activities, such as glutamate synthase and glycerol-3-phosphate dehydrogenase. This suggests that the synapomorphies of *At* of ancestor that lasted until now

are modifications of central processes that did not involve major re-wiring of the metabolic networks, but rather subtle modifications of activities and specificities of key enzymes. In addition, several genes encode more peripheral functions, likely involved in the interaction of agrobacteria of *At* clade with their environment, such as sensing of external signals, metabolism of iron and detoxification.

Cluster label	Reference organism name	Locus	Non-specific occurrence**	Functional description	Figure reference, # box
[G5-G13]-specific	G5-F2				
AtSp34		Agau_C201287–Agau_C201293	shared by G7-NCPPB 1641	oligopeptide transporter	
AtSp35		Agau_L100709–Agau_L100243		Complete phenylacetate degradation pathway	
AtSp36		Agau_L101258–Agau_L101263		peptide methionine sulfoxide reductase	
G4-specific	G4-B6				
AtSp37		ATU4Av2_II10139, ATU4Av2_II10142-ATU4Av2_II10149	shared by G9-Hayward0363 and G1-H13-3	aromatic compound (acriflavine) perception, degradation and efflux	2.33, 1
AtSp38		ATU4Av2_II60047-ATU4Av2_II60061	shared by [G5-G13] clade	uptake and catabolism of sugars (sorbose, dehydro-fructose)	2.33, 2
AtSp39		ATU4Av2_pI10294-ATU4Av2_pI10301	shared by G2-CFBP5494 and G1-TT111	ABC transporter and periplasmic gamma-glutamyltransferase (detoxification gamma-glutamyl cycle)	2.33, 3
[G4-G7-G9]-specific	G4-B6				
AtSp40		ATU4Av2_II40560-ATU4Av2_II40570	shared by 3 G1 strains and G6-NCPPB925	uptake and degradation of a (sulfated) polygalacturonide/polyglucuronide	2.33, 4
AtSp41		ATU4Av2_II110018-ATU4Av2_II110020		ferrichrome-iron sensing and uptake	2.33, 5
G7-specific	G7-Zutra 3-1				
AtSp42		ATU7Cv2_I30002-ATU7Cv2_I30006	shared by G4-B6, G3-CFBP6623 and [G5-G13] clade	QS-like two-component system regulators, periplasmic protein-disulfide isomerase DsbABG	2.34, 1
AtSp43		ATU7Cv2_I160081-ATU7Cv2_I160082		predicted addiction module	2.34, 2
AtSp44		ATU7Cv2_I180027-ATU7Cv2_I180028		putative DNA helicase and protein of unknown function	2.34, 3
AtSp45		ATU7Cv2_II100295-ATU7Cv2_II100305, ATU7Cv2_II100311-ATU7Cv2_II100318	shared by genomovars G1 and G3 and [G6-G8] clade	carbohydrate (polyamine) uptake and catabolism, methylamine assimlation	2.34, 4
AtSp46		ATU7Cv2_II120015-ATU7Cv2_II120019	shared by G8-J-07, G1-CFBP5771, G3-CFBP6623	prohibitin (DNA synthesis represession), phage shock protein (repression of sigma54-dependent transcription)	2.34, 5
AtSp47		ATU7Cv2_II150009-ATU7Cv2_II150014	shared by G13-CFBP6927	system for sensing (FecIR), TonB-dependent import (FhuACD) and utilization (ViuB) of iron(3+)-hydroxamate siderophore	2.34, 6
AtSp48		ATU7Cv2_pI0195-ATU7Cv2_pI0205	shared by G1-TT111 and G6-NCPPB925	enzymes and regulators of unknown functions	2.36, 7

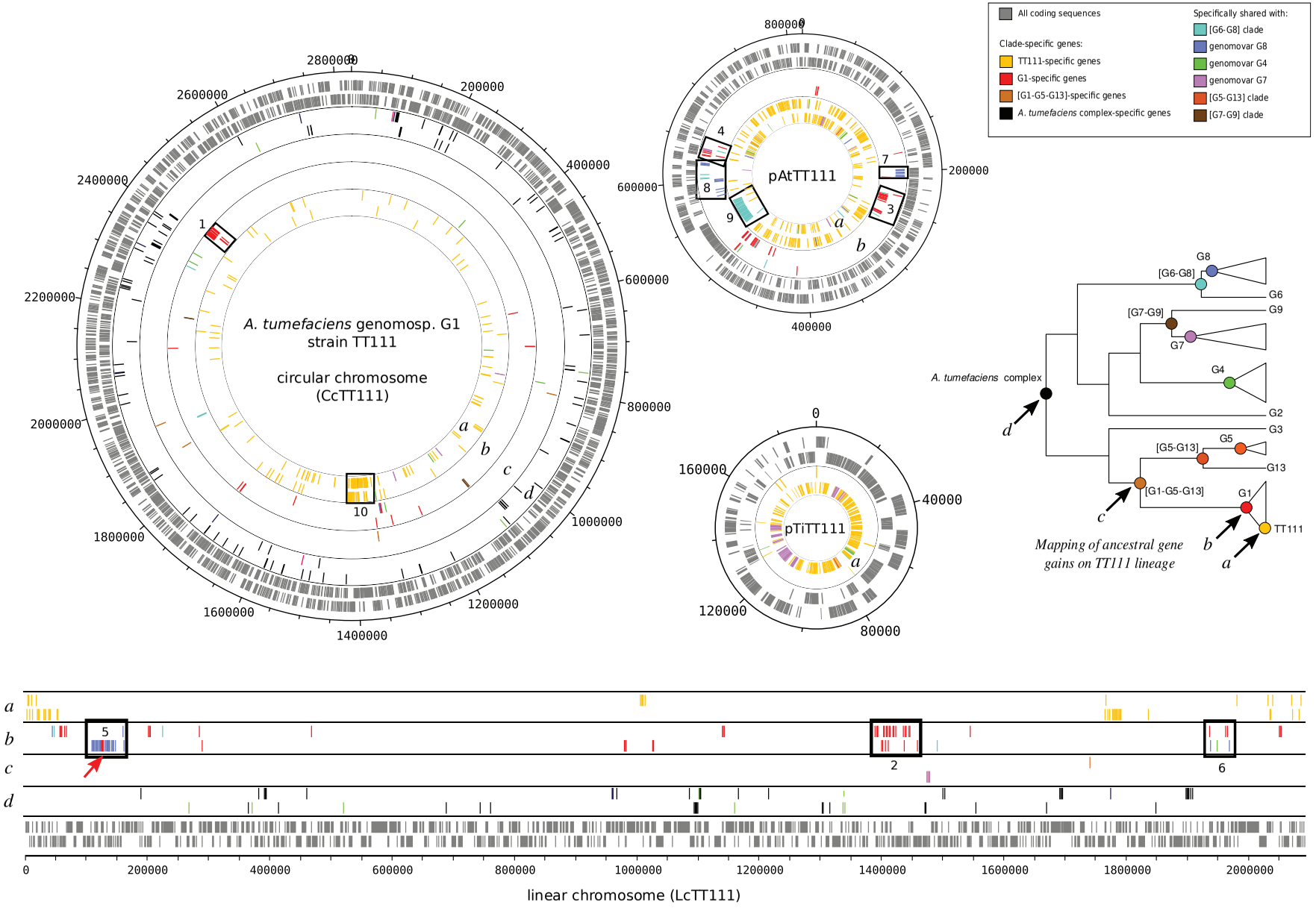


Figure 2.11: **Historical stratification of gains in the lineage of *A. tumefaciens* strain TT111.** (legend next page)

Figure 2.11: **Historical stratification of gains in the lineage of *A. tumefaciens* strain TT111.** The four replicons of the genome are represented circularly or linearly according to their molecular topology; replicons are not drawn to scale. Tracks within outermost ring (lowermost layer for linear chromosome) represent location of CDS on both DNA strands. Other rings (layers) show genes that were acquired along the history of TT111 lineage, and are labelled (a-d) according to the phylogeny on the right. Because pAt plasmid harbors almost only genes gained since genomovar G1 ancestor and pTi plasmid was gained by TT111 strain ancestor, tracks representing older genes are omitted.
Colors of genes indicate their specific presence in one of the clade that includes TT111, or their specific sharing with another clade (see legend box). Outer (lower) vs. inner (upper) tracks in the same rings (layers) distinguish clade-specific genes with strict vs. relaxed specificity criterion.
Numbered frames show particular gene clusters within TT111 genome: (1-4) G1-specific clusters: (1) AtSp2: chemotaxis regulation (che2) and aromatic compoud metabolism locus; (2) AtSp3: phenolics and amino-acid catabolism; (3,4) AtSp7,9: phenolic compounds downstream degradation; (5-8), clusters specifically shared by genomovar G1 and G8: (5) AtSp14: lipopolysaccharide O-antigen biosynthesis and neoglucogenesis locus with G1-specific chemotaxis-regulating hybrid sensor (red arrow, see figure XXX1); (6) AtSp12: outer-membrane lipoprotein and sensory protein; (7) AtSp17: deoxyribose uptake and assimilation; (8) AtSp15: exopolysaccharide (curdlan) synthesis, peptidoglycan modification and sensory protein; (9-10): clusters gained by TT111: (9) AtSp29: non-ribosomal peptide synthases involved in siderophore biosynthesis, shared by [G8-G6]; (10) prophage, partially shared by G3-CFBP6623 and G7-Zutra 3/1 (see figure XXX2).

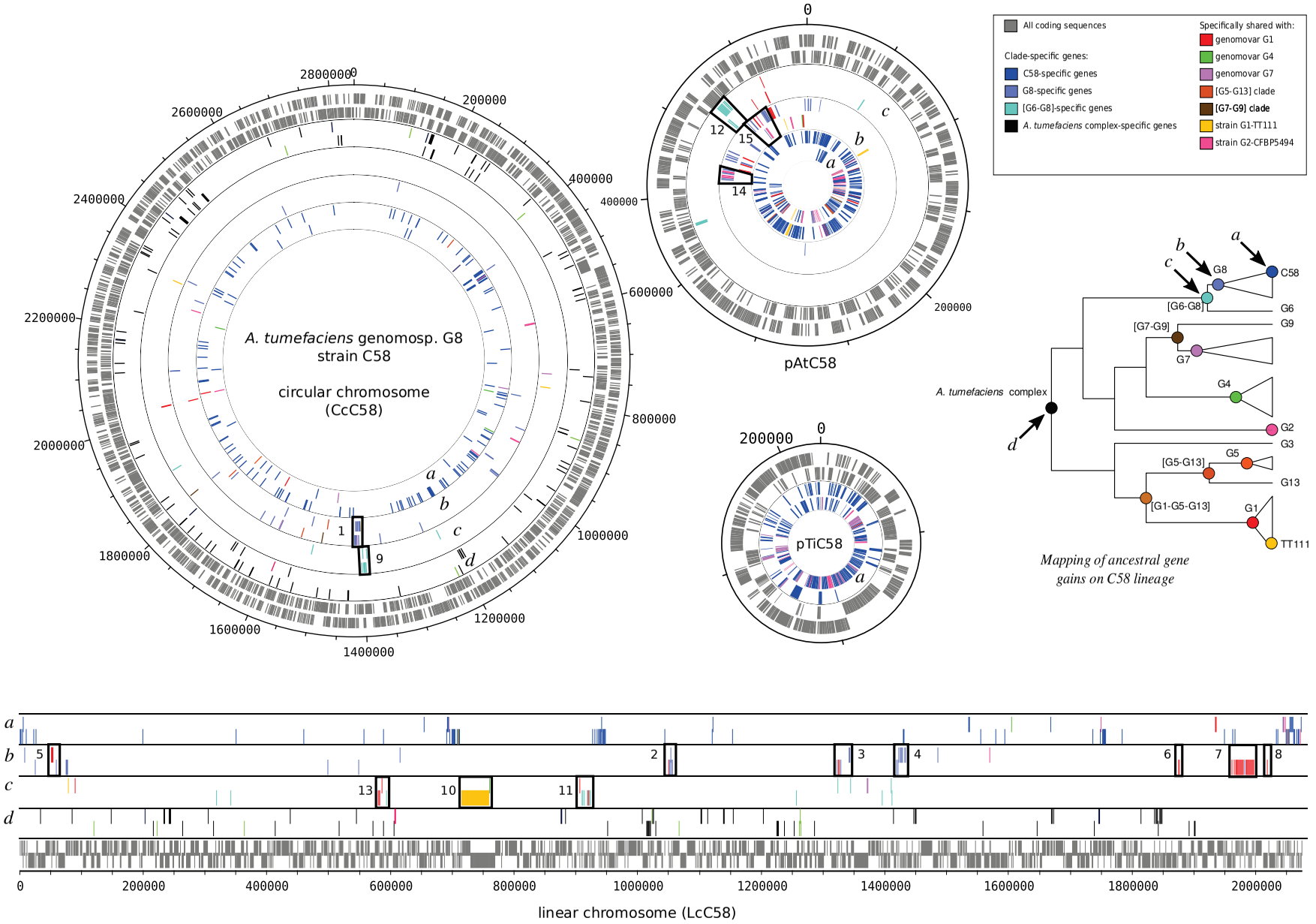


Figure 2.12: Historical stratification of gains in the lineage of *A. tumefaciens* strain C58. (legend next page)

Figure 2.12: (continued) Legend as in Fig. 2.11.
Numbered frames show particular gene clusters within C58 genome: (1-4) G8-specific gene clusters: (1) Atsp21: degradation of hydroxy-cinamic acids (ferulic acid); (2) AtSp23: degradation of complex amino-acids (opine-like compounds); (3) AtSp24 and AtSp25: Drug/toxic resistance (extrusion transporters), sarcosine oxidase; (4) AtSp26: sensing of environmental signals (phenolic compound, mechanical constrains); (5-8) clusters specifically shared by genomovar G1 and G8: (5) AtSp15: exopolysaccharide (curdlan) synthesis, peptidoglycan modification and sensory protein; (6) AtSp13: iron-sensing two component system FeuPQ; (7) AtSp14: lipopolysaccharide O-antigen biosynthesis; (8) AtSp12: outer-membrane lipoprotein and sensory protein; (9-12) [G6-G8]-specific gene clusters: (9) AtSp29: sugar (L-sorbose) uptake and catabolism; (10) AtSp30: non-ribosomal peptide synthases involved in siderophore biosynthesis, shared by G1-TT111; (11) AtSp31: sugar metabolism; (12) dipeptide uptake and degradation; (13) AtSp18: D-glucuronate uptake and degradation, specifically shared by genomovar G1 and G8; (14-15) clusters specifically shared by genomovar G2 and [G6- G8] clade: (14) AtSp27: Toxic extrusion / secondary metabolite secretion; (15) AtSp28: xanthine/-cyclic compound degradation, two-component sensor.

2.4.3 Discussion

This work intended to describe the history of genomes in a group of bacterial species and to understand the evolutionary processes that participated in shaping their gene content. For this, we designed an original procedure for reconciliation of histories of genes and genomes, that notably accounts for the non-independence of histories of neighbour genes. We obtained scenarios depicting the events of duplication and horizontal transfer that explained the discordance between histories of genes and species. We jointly inferred the profiles of presence and absence of orthologous genes in ancestral genomes, and the events of gain, loss and allelic replacement that happened on each branch of the phylogeny of species. A relational database, Agrogenom, was built to integrate the many aspects of genome evolution we studied: scenarios of gene evolution from the point-of-views of the gene trees and of the species tree, gene syntenies, genome architecture, and functional annotations of genes.

We combined these informations to recognize patterns indicative of adaptation throughout the history of diversification of genomes. To confidently characterize possible signatures of natural selection, and to be able to confront them to an expectation under a neutral model, we needed to reconstruct accurate and precise evolutionary scenarios.

2.4.3.1 Precise reconciliations using regional signal in genomes

Local reconciliation of histories of orthologs in multicopy gene trees A reconciliation method primarily aims to assign an event of either horizontal transfer, duplication, or speciation (DTS) to each node of a gene tree. This task is hard, since the size of the enumeration of possible scenarios grows quickly with the number of genes and species, even with simple reconciliation models (for a review, see Doyon et al. (2011)). The exercise of reconciling gene trees with species trees is especially complicated in prokaryotes because horizontal gene transfer can leave several different patterns. One common kind of HGT leads to the replacement of a resident gene (e.g. via homologous recombination). This leaves incongruences in gene trees that are easily recognizable, because a species is 'badly' placed in the tree. It can however get tricky when there are multiple nested incongruences, which can be interpreted in several different sequences of transfer events. Another more cryptic kind of transfer, which is more frequent between distant lineages, is the introduction of a new gene, or of a new copy of an already existing gene. The latter case is particularly problematic, as it can easily be confounded with a duplication event. Conversely, hidden paralogy – duplication followed by many independent losses that erase the multiplicity of duplicated genes in genomes – can be misinterpreted as transfer events.

For this reason, we searched for HGT using Prunier, a parcimony-based method that takes into account the phylogenetic support of topological incongruences and iteratively resolves them by identifying transferred subtrees and pruning them (Abby et al., 2010). To correctly

identify horizontal transfers in presence of duplication events, we first made a search for multiple representation of species in gene trees, to recognize any potential paralogous lineage (Fig. 2.4, step 1). We then applied Prunier to several different subtrees of orthologs. In presence of lineage-specific paralogs ('in-paralogs'), we applied Prunier to each (overlapping) sets of co-orthologs, i.e. orthologous groups that contain one of the a duplicated pair. (Fig. 2.4, step 2). Paralogous gene lineages have evolved independently and may thus have different topologies and branch supports, leading to the inference of potentially different scenarios of reconciliation by Prunier at the same gene tree node. This exploration of the reconciliation space by tests of multiple combinations of co-orthologs allowed us to choose the best scenario as the most frequently inferred one. In addition, the fraction of Prunier tests that agreed on the chosen scenario gave us a measure of confidence on the reconciliation. Finally, cross-checking of scenarios inferred for partially overlapping subtrees ensured the coherence of local scenarios in the global history of the gene family.

Reconciliation of gene blocks provide more accurate scenarios Another important feature of reconciliations is the coherent localization of the origination, duplication and transfer (ODT) events in the species tree. For a gene tree with a fixed set of ODT events, there are potentially many scenarios locating the ODT events at different nodes of the species tree, and these come with different number of convergent losses following the gene acquisition (Fig. 2.5).

While one could assume that the more reasonable scenario is the one that minimizes such convergent loss patterns within a gene family, another perspective could be to try to minimize the number of convergent transfer, duplication or origination events that are inferred across the genomes by recognizing those that are in fact the same event. Indeed, the occurrence in genomes of many gene cassettes – i.e. insertions of several consecutive genes, sometimes with operonic structures – indicates that blocks of genes can be transferred at once (Markowitz et al., 2009). Even though gene histories are practically all different when regarded globally, ODT events that have involved several neighbouring genes must have left analogous local patterns in the respective gene trees. Such multi-gene events can be recognized in neighbouring genes based on the ODT scenarios we obtained for each gene family (Fig. 2.5).

However, histories of neighbouring gene families can have different census of extant species indicative of different independent patterns of losses. In order to recognize an evolutionary event that involved several gene families – a block event – the (donor, recipient) coordinates of each single-gene event need to be matching the one of its neighbour. If one has to choose a single scenario for each gene family, for instance the one implied by a parsimonious scenario with a minimum number of losses, the independent losses that occurred in each gene family will likely lead to locate each event to close, but different nodes of the species tree. This would prevent the recognition of unique events spanning several gene families, leading to the artefactual count of many convergent ODT events involving single genes that are neighbouring in genomes. To avoid this, we kept uncertain the (donor, recipient) coordinates of ODT events

and the number and location of subsequent losses (Fig. 2.15 B,C).

Conversely, if we consider losses as a noise around the signal of an evolutionary event, taking several independent samples (i.e. gene families) with different noises lets us better estimate the true characteristics of the multi-gene event: it can be characterized in each neighbouring gene family by different 'noisy' uncertain scenarios, but each must contain the 'true' signal. Using this principle, we can refine the precision of a common evolutionary scenario by intersecting the scenarios of all co-evolved genes (Fig. 2.16).

We scanned the genomes and the reconciled gene tree collection to find neighbour gene events with compatible coordinates in the reference tree, using a greedy algorithm that intends to cover the genomes with blocks of genes that have congruent histories (Sup. Mat. 2.7.3; Sup. Fig. 2.30). We hence reconstructed block events, i.e. unique evolutionary events involving blocks of co-evolved neighbour genes, and simultaneously refined the scenarios of these block events. Doing so, we reconstructed scenarios more parsimonious in ODT events than if we had considered gene family evolution separately: the global reconciliation scenario involved more loss events but less origination, duplication and transfer events (L: +2,896; O: -922; D: -1,371; T: -10,992) (Fig. 2.5, Table 2.2). We note than the count of additional losses is certainly over-estimated, because block events of gene loss could have occurred, but were not taken into account in this study. At the same time, we could refine the location in the species tree of 83% of duplications and of 69% of transfer receivers (Sup. Fig. 2.23), getting from 70 to 93% duplications assigned to a unique node, and from 60 to 83% of transfers with a unique possible receiver. Most donors of transfer events were precisely defined at the first steps of reconciliation (Sup. Fig. 2.22), so there was no significant gain in precision for the location of donors.

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

Our method succeeds in joining many single-gene events into blocks with coherent scenarios, that can span up to a hundred genes. Can this be, on the contrary, an artefactual grouping of independently evolving genes (i.e. false positives)? This is hard to verify, as there is no reference to reconstruct past evolutionary events apart from the genomes we used. However, we can reason that our method is performing properly. First, because we have a dense dataset, the number of nodes of the species tree to which can be mapped an event is high (93) and even more when it comes to a transfer, which is defined by a pair of coordinates (93^2). As the majority of single events are located to a single species node before amalgamation in blocks (Sup. Fig. 2.22), the probability of neighbour genes having similar event coordinates by chance should be small. This might not stand if the events, in particular transfers, do not occur uniformly in

the (receiver) genomes. This is a known fact that some regions are hotspots of integration of transferred genes, because they correspond to integrons or other mobile elements (Juhas et al., 2009) – as we observed some in the present study – or because of organizational constrains of the genome (Touchon et al., 2009). In this case, successive integration of genes coming from potentially closely related – and thus undistinguishable – donors might happen. Indeed, some gene cassettes are known to be restricted to a small taxonomical range (Mutreja et al., 2011; Holden et al., 2013), though they are often related to fitness as host-specialized pathogens.

A way to assess *a posteriori* if the genes we grouped in the same block did co-evolved or not, is to look if the distribution of the supposedly co-evolving blocks differs from a random sampling of neighbour genes, and if this difference can be explained by the evolutionary events we inferred. We looked at the functional annotation of genes and its homogeneity within blocks of genes in genomes. It appeared that co-transferred genes are in average more similar in function than genes taken in random windows in genomes. This indicates that there must have been a selective pressure for gathering the genes at the same locus – either in the donor genome before their co-transfer, or in the recipient genome after their convergent transfer. The latter hypothesis is less likely since it postulates that genes independently transferred from a same lineage are more likely to co-function than when they come from different taxa, i.e. that the distribution of existing functions is biased according to taxonomy. This is plausible, but the inventory of cellular functions existing in versatile genomes of Rhizobiales is very large, so the hypothesis of unlinked genes with similar function to be convergently transferred from the same lineage and then gatherd to the same very locus by genomic rearrangement is highly improbable. Thus, the fact that we can observe traces of selection for a collective function in the inferred blocks of co-transferred genes suggests that these blocks are genuine.

Altogether this indicates that our method conservatively and accurately characterizes events of duplication, transfer or speciation and their location in the reference tree of species.

2.4.3.2 A history of Rhizobiales, from the point-of-view of the entire genome

The definition of a history of species is central in the reconciliation procedure (Fig. 2.4). One could oppose to this method that there is no tree-like pattern of descent for species (Doolittle, 1999), and that this approach is fundamentally biased. However, the tree we reconstructed for our dataset of 47 genomes of Rhizobiales strains appears to be representative of the histories of the majority of genes. First, the jacknife support of branches, that reflects the part of core genes that support the consensus tree topology, is high ($\geq$ 70%) for all deep nodes, except for those characterizing the placement of [G6-G8] clade and strain G2-CFBP 5494. Second, while we detect a large number of horizontal transfers (with an average of ~4 HGT/gene tree in our database), the large majority (~90%) of splits in gene trees were assigned to speciations, i.e. showing the gene followed the history of species.

The position of [G6-G8] clade and strain G2-CFBP 5494 in the reference tree (Fig. 2.3) is the major ambiguity of the reference tree. Exploration of alternative reconstruction methods

showed that the unrooted topology of the tree of *At* clade was robust, but that its root was unstable (see Sup. Mat. 2.7.1). Considering alternative rootings (Sup. Fig. 2.18) leading to different grouping of clades (Sup. Fig. 2.19) would however not change the sense of most of our observations. Notably, the particular pattern of gene sharing between genomovar G1 and [G6-G8] clade, that is evident when clustering strains by gene content (Sup. Fig. 2.20), is likely not due to an error in phylogenetic reconstruction missing the proximity of this clades, because their distant positioning in the tree is always well supported, thus implying HGT events.

2.4.3.3 Role of recombination in species cohesion

A. tumefaciens species complex is made of ten genomic species (Popoff et al., 1984; Portier et al., 2006) that are by definition well-delineated clusters of diversity (Stackebrandt et al., 2002). The high amount of HGT that occurs quite undistinctly between genomic species (Sup. Fig. 2.26), blur the delineation of species considering their gene content (Sup. Fig. 2.20). In this context, it is almost surprising that we can recover such a clear pattern of differentiation of genomic species from molecular phylogenies (Fig. 2.3). It may result from different transfer dynamics between core and accessory genes, as previously reported (Abby et al., 2012; Choi et al., 2012).

We indeed saw that replacing transfers of core genes, that mostly reflect homologous recombination (HR), were occurring in large majority within genomic species, as seen for genomovars G1, G4 and G8, and within narrow clades like [G7-G9]. Other genomic species were not enough densely sampled ($n \leq 2$) to reveal any such pattern. This clearly indicates that HR is intensely occurring within genomic species of *At*, in which it certainly mediates cohesiveness of their genotype.

HGT is not prohibited between species, as seen for exchange of accessory genes, but the nature of the HR process limits the integration of divergent core alleles. It has long been known that the efficiency of HR in bacteria is linked by a log-linear relationship to the genetic distance of sexual protagonists (Roberts and Cohan, 1993), what inhibits the hybridization with distant relatives and thus accelerates the process of their divergence. HR may thus be preserving the genetic cohesiveness of narrow clades such as genomic species by controlling the identity of incoming core genes. This can be done with core genes because there always exists a homologous DNA template in the cell that allows the mismatch-repair pathway to validate or not its integration. In opposition, additive gene transfers consist by definition in the introduction of a gene with no close homologue in the recipient genome, and therefore should escape this control. In fact, non-homologous DNA is often introduced by homologous recombination of neighbour genes, so inter-specific additive transfer may only occur in regions of high conservation within the species complex. It could be interresting to test this prediction by reconstructing the ancestral gene orders and gene sequences in addition to the history of transfers.

2.4.3.4 Ancestral genome content and evolutionary dynamics of genes

In the tree of genomes, we observed that the sizes of genomes and the amounts of gene gains and losses were much larger at leaf nodes, i.e. extant genomes, than those inferred for internal nodes, i.e. reconstructed ancestral genomes. Indeed, we find a small genome for the common ancestor of our 47 Rhizobiales (ca. 3,800 genes) when extant genomes can be larger than 7,000 genes, and there is a significant negative linear relationship between sizes of ancestral genomes and their age. There is, however, no particular reason to think that the ancestral genomes were always smaller than their descendants. It has been argued that such result was the sign of an over-estimation of the number of horizontal transfers that occurred in bacterial genomes (Dagan and Martin, 2007).

These apparent discrepancies can however be interpreted as the presence in contemporary genomes of a large polymorphism of gene presence/absence. Indeed, among the large number of recently acquired genes, the majority will not be fixed, because their adaptive value is null or ephemeral. Similarly, a large fraction of deletions are probably deleterious on the long term. For instance, genes involved in adaptation to environments met occasionally are under periodic selection, and their deletion will not cause the immediate death of the bacterium, but will be counter-selected eventually. In fact, ancestral genomes might have been approximately of the same size than extant ones, but the 'volatile' fraction of their gene content has not lasted in any current genome, and hence could not be reconstructed in ancestral ones.

This high turnover of genes was thus probably constant in history, with new transferred genes continually taking the place of genes that were lost (Lerat et al., 2005; Marri et al., 2007). If transient genes are bringing functions with mostly environment-specific adaptive values, this implies that the constant renewal of the functional repertoire of genomes may reflect, to a certain extant, the variations of the ecological niches of organisms. Each HGT variant can thus be considered an ecotype (Cohan and Koeppel, 2008), whose ecological specialization might often prove useless, and thus is deemed to extinction. However, some ecotype lineages may successfully exploit their new functional abilities provided by HGT. Indeed, many genes can be acquired in a small time by a lineage of genomes in response to positive selection for adaptation to an new environment. This is well illustrated by the the drastic shift of ecology of the arsenic-resistant bacterium *Rhizobium* sp. NT-26. This bacterium, deriving from a soil-dwelling rhizobial ancestor, fixed a host of genes related to the resistance to heavy metals, that allowed it to colonize the arsenic-rich environment of a gold mine (Andres et al., 2013) (see annex 5.1.2). This kind of evolutionary paths towards extreme ecotypes may be dead-ends, because lineages whose genome got too specialized in a particular environment cannot cope with the following environment shift. This is the probable fate of most host-restricted obligatory pathogens. However, in a few cases, the evolutionary bet may be paying on the long term, and lead to the funding of a new clade with specialized functions integrated to its core genome. This is the case for obligatory symbionts like *Buchnera* that pushed the evolution of their genome over a point of no return, but succesfully colonized the niche of aphid cells.

There hence exists a gradient of selective pressures acting on genes that should match their age in genomes, from recent genes under episodic positive selection in an ephemeral ecological niche, to old core genes under strong purifying selection for their essential function in the life of the cell. The reduction of the set of genes conserved in a clade with its age is thus associated to a skew towards functions under strong selective constrains, i.e. genes involved in central cellular functions rather than in adaptation to an ecological niche. Incidentally, the partial gene content inferred for clade ancestors are also biased, and for too ancient genomes it may miss niche-specifying genes. At the scale of *A. tumefaciens* species complex, the ancestral genome of the *At* ancestor is around 4,500 genes, when extant genomes have an average size of 5,500. This difference of 1,000 genes matches approximately the number of volatile genes recently gained by contemporary genomes. Under the hypothesis of a similar volatile gene fraction in the ancestor of *At* clade than in its descendants, we likely have obtained a good estimate of its gene repertoire, including genes coding the adaptation to long-lasting niches.

We looked at the global pattern of evolution of gene repertoires along *At* history to identify trends and potential exceptions in them, that could sign adaptive processes. First, we found heterogeneous quantities of gene gain and loss for the different clade ancestors within *At*, that partially correlated with the evolutionary time separating ancestors. This revealed the ancestors of genomovar G1, G8 and G7 as outliers who excessively gained genes. This could be a signature of past positive selection for diversification of the gene repertoire of these ancestors, but the amount of variance explained by the regression was too low to be very confident in this observation. It would be interesting to compare these results with those from methods that explicitly model the processes of gene gain and loss along the tree of species (Csűrös and Miklós, 2009; Szöllosi and Daubin, 2012) to see if they observe such large variance and what underlies it.

More interestingly, we found a robust correlation between the age of a clade and the degree of conservation of genes after their acquisition by the clade ancestor. This relationship could be well modelled by a survival process (fit to a decreasing exponential: $r^2 = 0.51$) (Fig. 2.8 F). The outliers that fall above this trend – those clades that conserved more genes than predicted by the age of the clade – strikingly all belong to the clades [G1-G5-G13] and [G6-G8] (Sup. Fig. 2.32). This tendency of genome of these clades to retain more gene than other *A. tumefaciens* might indicate their ancestors have acquired genes that since were under purifying selection, potentially for their ecological role.

This accumulation through time of genomic synapomorphies under selection must had a strong impact on the ecological niche of [G1-G5-G13] and [G6-G8] clades, as was seen for *Salmonella enterica* becoming a specialized vertebrate pathogen (Porwollik et al., 2002). Ecological specialization is, however, not as obvious for agrobacteria, from the point of view of field ecology. Indeed, agrobacteria are all found in soils and rhizospheres and often co-occurring in the same micro-metric sample (Vogel et al., 2003), and their ecology can thus appear indistinct. Although, some soils and/or host plant show preferential colonization by certain species

(Costechareyre et al., 2010) and G2 species appears to be a specialized opportunistic human pathogen (Aujoulat et al., 2011), showing the existence of niche differentiation among species of *At*. This is likely only the emerged part of the iceberg and cryptic multi-factorial ecologies may distinguish the ten species of the complex. We thus propose to explore the specific gene repertoires of the clades of *At*, and notably of [G1-G5-G13] and [G6-G8] clades, in search for potential ecological adaptations.

2.4.3.5 Clusters of clade-specific are under purifying selection for their collective function

In the present work, we hypothesized that genes specifically acquired by an ancestor and subsequently conserved in the descendant lineages, i.e. clade-specific genes, were under purifying selection, supposedly because they are involved in adaptation of the clade to its ecological niche. However, the observation of conserved genes in extant genomes could also be explained under a neutral model of drift, because appearance of deletion mutants and fixation of their genotypes are stochastic processes that can take a long time. So, to what extent selection is responsible in the long-term maintenance of transferred genes? And conversely, can we consider genes that were conserved after transfer as under purifying selection?

We thus searched for signatures of selection on transferred genes in the patterns of occurrence of gene families and their conservation as syntenic blocks in genomes. A recurrent observation of the present study is that clade-specific genes tend to occur in clusters, corresponding to blocks of co-transferred genes. Most interestingly, those clade-specific genes found in blocks often appear related in their function, and clusters spanning up to thirty genes can encode only one cellular pathway (Table 2.5, AtSp14, AtSp29). The fact that large gene clusters have not been dismantled by rearrangements and deletions may suggest that a function collectively coded by the gene cluster is selected. However, 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 simply results 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 that a transferred fragment spanned a functional element like an operon would depend on the size of the transferred fragment and on the scale of the functional arrangement of genomes.

We found that the history of transfers in genomes accounted for a part of their functional structure (Fig. 2.9 A,C). Indeed, our observation reflects a biased retention of transferred blocks of genes in genomes towards blocks that code functional partners in a biological process. This can be explained by natural selection favouring the fixation in a recipient genome of those transferred blocks that can immediately provide a selectable function. This provides support

to the model of 'selfish operon' proposed by Lawrence and Roth (1996), as repetitive transfer and selection for a function in the host can gather genes into coherent functional units able to travel across genomes. Efficient transfer incidentally promote the survival of genes that would not be adaptive enough – in average over time and environmental conditions – to be maintained in one lineage by purifying selection. For this reason, this clustering process has been called 'selfish' from the point-of-view of the genes (Lawrence and Roth, 1996).

However, the 'selfishness' of co-transferred genes is not a necessity; rather, this process of clustering of co-functioning genes should be even more efficient in presence of strong selection at the level of the organism for the acquisition of the transferred gene function. Indeed, we found that among the groups of genes acquired by transfer, those which were conserved in all descendants had more coherent annotated functions than non-conserved ones (Fig. 2.9 B,D). This shows that genes conserved in genomes are more likely to provide an adaptive function to their host than transient genes. This bias cannot be explained by a drift model, and indeed supports our initial hypothesis according to which clade-specific genes are retained by their host genome for the adaptive advantage they provided. This 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 conservation in genomes.

By scrutinizing the functional repertoires encoded by these clusters of genes specific of *A. tumefaciens* clades, we found that the function they encode are linked with the adaptation of agrobacteria to rhizospheres.

2.4.3.6 Clade-specific genes in the light of speciation models

We were interested in finding the genes that contributed on the long term to the adaptation of species to their ecological niche. At the scale of ecotypes – emergent clades with a differentiated ecology (Cohan and Koeppel, 2008) – niche-specifying genes can be recognized as characters unique to these clades. Indeed, it has been argued that natural selection was prohibiting the transmission of these niche-specifying genes to close relatives of the adapted ecotype (Cohan and Perry, 2007; Hausdorf, 2011; Friedman et al., 2013). The 'stable ecotype' and related models has been shown to be relevant in practical cases of recent divergence (Sikorski and Nevo, 2005; Smith et al., 2006), however it only accounts for the very first times of speciation, and do not describe the genome-wide dynamics of cladogenesis on the longer term (Shapiro et al., 2012).

The model of 'fragmented speciation' proposed by Retchless and Lawrence (2007) has a more historical perspective and describes the process by which genomes gradually diverge. This model supposes that genetic exchange – notably by homologous recombination – continues over time after the initial split of lineages, but goes on decreasing as more and more adaptive genes are specifically acquired by each lineage, where they insulate genomic regions from

the possibility of recombination (Lawrence and Retchless, 2009). This model has been tested in a few cases of bacterial clades diverged for different timespans, and there were evidences supporting it (Retchless and Lawrence, 2010) and against it (Luo et al., 2011; Cadillo-Quiroz et al., 2012). One notable limitation of this model is to not consider the possibility of re-hybridaton of lineages after a long isolation and divergence. This hybridation process was observed with a substantial introgression in *Campylobacter coli* of genes from *C. jejuni*, that could have a role in the ecological adaptation of agricultural lineages (Sheppard et al., 2013).

This hybridization process is interesting in our perspective because it can be the source of many adaptations. Indeed, genetic exchange between distant relatives will provide genes that 'made their proofs' in a genomic context which is essentially the same and for this reason may integrate more easily in the recipient genome – this would be the case in *A. tumefaciens*, where ~60% of genes in genomes belong to the core genome of the species complex. Moreover, in the case of *A. tumefaciens*, the different species cohabit in the environment (Vogel et al., 2003) and have apparently undistinguishable ecologies. Though this is certainly not true in the detail, it suggests that they are exposed to similar biochemical environments, that maybe only vary in degrees. Thus, certain adaptations on which one of these species relies to make a living, such as the degradation of a particular sugar or attachment to a particular plant, would probably provide a substantial selective advantage to a cognate species. The transfer of genes serving the adaptation to the rhizosphere environment may thus be promoted between species of *At*. This idea of possible pooling of ecological resources between relatives rejoins that of the model of 'nano-niches' proposed by Cohan and Koeppel (2008). In this model, cognate ecotype lineages share a pool of resources but use them in qualitatively or quantitatively different ways, thus partitioning the whole ecological niche. Such ecotypes, however, have a high risk of being outcompeted by a successful generalist. This is not the case at a larger evolutionary scale, as pairs of clades with advanced divergence, such as between genomic species of *A. tumefaciens* (~10-15%) (Lassalle et al., 2011) or between *C. coli* and *C. jejuni* (15%) (Sheppard et al., 2013), can hybridize without risking complete overlap of their ecological niches and competition to extinction.

In the present study, we took a historical perspective of genome evolution, and focused on the potential of clade-specific genes to provide ecological adaptations to their hosts. We were mostly interested in the variations of gene repertoires and in the pattern of conservation of these variations as a signature of selection. Therefore, we searched for clade-specific patterns of gene occurrence, defined as genes gained/lost in a clade ancestor and conserved as present/absent in all extant members of the clade. Patterns of gene occurrence in a clade are thus contrasted with patterns of occurrence in its closest relative, regardless of what is found in more distant lineages, i.e. we looked for synapomorphic gene presence/absence. Using this criterion, we could recognize gene presence/absence that are strictly specific to a clade, but also those that are specifically shared among distant relatives, as for instance between two non-sister genomic species of *A. tumefaciens* complex.

This leads to a more comprehensive description of patterns of gene distribution along the phylogeny of organisms, which can be linked to scenarios of evolution more complex than the simple 'stable ecotype' model. Our present analysis provides a mean to identify what *combination of* genes is unique to the genomes of a clade, and try to relate the integrated set of clade-specific functions to a clade-specific ecology.

2.4.3.7 Ecological adaptations in *A. tumefaciens*: a history of variation of shared traits.

Hybridization of genomovar G1 and G8: ecological convergence or diversification? The sets of functions found in G1-specific and G8-specific genes both sketch an ecology orientated towards metabolism of phenolics and production of exopolysaccharides (Lassalle et al., 2011) (section 2.3.2 and 2.4.2.8). This can obviously be explained by the set of genes they specifically share (clusters AtSp12-19), but also because unrelated – non-orthologous – G1-specific and G8-specific genes are similar in their general functional annotations. Is this leading to their ecological convergence? And in that case, can we predict that a species will eventually prevail in a competition for shared ecological resources? This is not likely, for three reasons.

First, some of the specifically shared genes may encode products with diverged functions and thus have different impacts on the species' ecology. For instance, the genomic region AtSp15 encoding the biosynthesis of curdlan EPS is found orthologous in G8, G1 and G2 strains, but in this case orthology may not be a good proxy of similar function, because their sequences have diverged more than expected given the global genomic distance. Nucleotide sequences of AtSp15 are 71% identical between C58 and TT111, compared to 86% average nucleotide identity of the complete genomes, and belong to the 2nd more diverged centile of identities of genome fragments. This may indicate two parallel HGT events from two distant relatives of a foreign taxon that introduced diverged alleles in each species, or that both clades once shared identical alleles of these genes and that they evolved independently faster than the rest of the genome, potentially because of positive selection for phenotypic diversification. Indeed, the branch leading to the G8 cluster in trees of these gene families were always longer than branches leading to G1 and G2. Moreover, mapping of substitutions along the branches (Romiguier et al., 2012) showed that dN/dS ratio is two to three-fold more elevated on branches leading to G8 than on those leading to G1 or G2, a signature of past positive selection for these proteins to diverge in G8 ancestor. Another possibility would be that these genes were ancestrally present in the common ancestor of all *A. tumefaciens* and diverged precociously between G1 and G8 lineages, as proposed in the 'fragmented speciation' model (Retchless and Lawrence, 2007). This is not likely, because (1) this scenario would have needed many losses in other lineages and (2) the high codon usage adaptation predicted for long-residing genes under this model (Retchless and Lawrence, 2007) is not observed for AtSp15 genes: in G8, these genes have an average CAI of 0.45, which is in the lower third of CAI values in C58 linear chromosome, corresponding to the codon usage of recently acquired genes (Figure 4, box 2a

in Lassalle et al. (2011)). Whatever its cause is, this distant homology suggests a potential functional differentiation of the diverged proteins (76% amino-acid id. over the locus). This large divergence also explains why this cluster was not seen in G1 genomes by micro-array hybridization Lassalle et al. (2011), as the detection threshold lies around 80-85% nucleotide identity.

Second, shared niche-specifying genes are combined to other strictly clade-specific genes in each clade. These globally different genotypes must code different phenotypes. Indeed, niches are defined by several factors and the sharing of identical means of adaptation for one factor do not lead to a complete equalization of niches. For instance, genomovar G1 and G8 share the same locus for biosynthesis of a O-antigens of the lipo-polysaccharide. This 30-kb region likely encode the synthesis of a complex component of these species' capsules, which may mediate a specific interaction with the environment, probably with an eukaryotic host as it is the case for the LPS synthesized by homologous enzymes in *Brucella spp.* (Vizcaíno et al., 2001). Such interaction may be of great importance in those species biology, because it could specify the attachment to a particular host plant. In this case, both species might compete for the use of resources such as the plant exudates. However, other major traits differentiate markedly genomovar G1 and G8, as for instance the ability to respire nitrate. This anaerobic respiratory pathway may allow G8 strains to colonize rhizospheres of plants in flooded soils, while G1 strains might be restricted to drier terrains.

Finally, specific epistatic interactions between genes of each clade-specific set may participate in phenotypic differenciation. For instance, while the O-antigens produced by G1 and G8 cells must be very similar given the high similarity of the biosynthetic proteins (>93% amino-acid identity in average for proteins of the AtSp14 locus), the regulation of biofilm production in G1 and G8 cells is likely different.

In G1 genomes, there are many G1-specific regulatory genes, such as the *che2* operon (cluster AtSp2) and hub signal-transducting protein HHSS (cluster AtSp14) which are involved in chemotaxis regulation, and a sensor protein (cluster AtSp3) modulating c-di-GMP – a secondary messenger involved in the switch from motile to sessile behavious. Those specific regulators are all found linked to G1-specific genes involved in phenolics catabolism or biofilm production. These latter genes may constitute the downstream regulatory targets of what seems to be a coherent regulation network controlling motility, biofilm production and degradation of phenolics, potentially in response to environmental conditions specific of the niche of genomovar G1. Similarly, G8-specific genes of the AtSp26 cluster are involved in the perception and transduction of environmental signals such as mechanical constrains and a phenolic compound related to toluene; this regulatory island may be involved in specific regulation of genes linked to adpatation to G8-specific niche.

The hybridization of G1 and G8 clades thus results in each species tapping the same resources in a different way, which should not lead to a significant competition between them. These species may then form guilds of relatives that exploit partitions of a common ecological niche, explaining why we can observe them co-occurring in soils (Vogel et al., 2003; Portier et al., 2006).

Guilds of *A. tumefaciens* species co-exist by partitioning common resources These diverse functions coded by genes specific to the whole *A. tumefaciens* complex are likely to reflect important dimensions of the ecology of the group. Interestingly, the functional modules to which participate *At*-specific genes are all reminiscent of those encoded by genes specific to lower clades, such as iron scavenging, metabolism of phenolics, resistance to environmental toxics, metabolism of complex amino-acids and sugars and production of EPS. This would indicate that *A. tumefaciens* species have ecological niches not much different from the one of their common ancestor. If the general environmental factors defining the ecological niche of agrobacteria are similar among lineages, they evolved differently while trying to fit this common adaptive landscape. In many cases, genes associated to the same metabolic pathways were acquired independently. Those genes are in general functionally equivalent, but may differ in efficiency in some environmental set-up. In some species ancestral genes were lost and simultaneously replaced by ones serving the same purpose in a different manner. This might prove very efficient in insulating one species' niche from the one shared by the whole group.

For instance, the loss of the locus coding biosynthesis of the siderophore agrobactin in the ancestor of [G6-G8] clade coincided with the gain of another siderophore biosynthesis locus (table 2.5, AtSp30). It was shown in *Vibrio* that producers of siderophores incur a cost of supplying for cheaters that do not produce it (Cordero et al., 2012). This functional replacement must have allowed them to continue to scavenge iron, while escaping the competition for a good common to all agrobacteria and avoiding the potential cost due to cheater species that would not express their own siderophore. Strikingly, [G6-G8] clade also lost genes involved in agrobactin uptake in other species, indicating G6 and G8 strains are not even cheating in scavenging agrobactin that could be available in the environment. This shows that species can differentiate by playing differently on the same field. These differences could be qualitative or quantitative. In the case of the replacement of siderophore, the [G6-G8]-specific molecule, though not characterized yet, is likely much more complex than agrobactin, given the locus encoding its biosynthesis is almost three times as large. Molecules produced by distant homologous enzymes in cyanobacteria are indeed more complex than the three-catechol condensate of agrobactin (Hoffmann et al., 2003). This could provide better affinity for iron in certain media, or also better affinity for re-uptake.

A few functions are found as a leitmotiv in specific genomes of *At* clade – production and uptake of siderophores, uptake and metabolism of phenolics, extrusion and degradation of

toxic molecules – may be several facets of the same ecology. Indeed, siderophores and antibiotics are often complex polycyclic molecules that can be derived from simple phenols by particular pathways of secondary metabolism. Interestingly, it is possible that transporters and enzymes involved in metabolism of such compounds are not very specific of their substrate and can participate in several different pathways linked to one or another biological function. The so-called multi-drug transporters and ring-cleavage mono- and di-oxygenases commonly involved in these pathways are indeed known to have low specificity (Campillo et al., unpublished results). A similar observation can be done on the collusion between enzymes involved in degradation of complex sugars found in plant exudates and those involved in the biosynthesis of exo-polysaccharides, both functions found frequently in *At* clade-specific genes. Either we are unable to distinguish the true molecular function of these genes, or they belong to a versatile pool of proteins involved in a complex metabolic interaction between agrobacteria and their environment.

If these pathways degrade and produce ranges of structurally related molecules, the coupling of these pathways should be facilitated, as each 'metabolic module' can plug into another by transforming at least one of the many products of the other. Transfer of genes coding such metabolic modules would thus more probably provide a selective advantage when arriving in a genome containing other such modules. Recombination between distant species of metabolic modules serving in similar biological processes would then constitute a propitious context for ecological innovation.

2.4.3.8 Secondary (and third) replicon of *A. tumefaciens* genomes are the place of genomic innovation

The genomes of Rhizobiaceae have complex architectures, containing a primary chromosome and a chromid, i.e. a secondary chromosome or a large megaplasmide bearing essential genes (Harrison et al., 2010), and a variable complement of plasmids of different sizes (Young et al., 2006). More particularly, a subclade of Rhizobium/Agrobacterium supergroup that includes *A. tumefaciens* has a linear chromid (Slater et al., 2009, 2013), which is the result of a unique ancestral event of linearization and thus constitutes a synapomorphy of this clade (Ramirez-Bahena et al., in press; Annex 5.2.2).

We showed that most of the genes coding putative ecological adaptations are born by the linear chromid and the pAt plasmid of *At* (Fig. 2.11 and 2.12; Sup. Tables 2.9 and 2.10). This might results from particular evolutionary dynamics determined by these replicons. Notably, different replicons of *At* genome may be subject to different replicon-wide levels of selection and/or recombination.

The linear chromid is highly plastic and recombinogenic but stabilizes adaptive genes. It was shown that genes had higher evolutionary rates on smaller chromosomes of complex genomes (Cooper et al., 2010), which could be explained by differences in levels of gene expression between chromosomes impacting the efficacy of selection on genes (Morrow and

Cooper, 2012). Another factor that can modulate the evolutionary rate of genes is homologous recombination. Indeed, if a region of the genome recombines more than another, this will reduce the overall diversity of sequences, but promote the rapid fixation of advantageous mutations, thus accelerating the divergence of positively selected sites. In addition, HR can promote the more rapid variation of gene content by propagating gene insertions and deletions located between endpoints of a recombination track. In particular, the linear topology of chromosomes theoretically favours very large exchanges, because it only needs a single crossover to recombine a fragment, which can lead to the exchange of fragments as large as the arm of a chromosome, as observed during meiosis in Eukaryotes.

We previously observed large tracks of reduced divergence between pairs of the same species of *At*, and this pattern was specific of the chromid (Lassalle et al., 2011). This could indeed reflect large events of homogenizing recombination, but it could also stem from large selective sweeps, as was proposed by Epstein et al. (2012) regarding a similar observation on the chromosome of *S. meliloti*. The latter hypothesis could also apply to the chromids of *A. tumefaciens* species, since they bear most of the species-specific genes that are likely under strong purifying selection. The hypothesis of selective sweeps and that of intensive HR on the chromid are mutually exclusive, because HR would break the linkage between selected genes and other genes, thus preventing large sweeps. However, if HR is not more frequent on the chromid, but involves much longer segments of DNA, as hypothetically rendered possible by a linear topology, large selective sweeps could be favoured by HR.

A previous study of Marri et al. (2008) investigated the prevalence of recombination in circular vs. linear chromosomes of *Agrobacterium* biovar 1 strains based on the linkage of marker genes. They observed more inter-genic recombination on the linear chromosome but higher intra-genic recombination on the circular chromosome, and therefore concluded to no marked difference between them (Marri et al., 2008).

Using our inventory of replacing transfer events as well as a test of occurrence of intra-genic recombination, we unambiguously demonstrated the larger prevalence of HR on the linear than on the circular chromosome. In addition we saw a more intense variation of gene content through gain and loss on the chromid, which is probably the consequence of the higher HR rate. However, we failed to characterize events of recombination at a scale larger than several 10kb, whereas we previously showed events mobilizing fragment larger than 1Mb could happen within species (Lassalle et al. (2011), Fig. 1). This observation is thus more in favour of more frequent HR, and not selection large-scale event, to the cause of dynamic of the linear chromid.

Interestingly, the propensity of circular and linear chromosome to retain genes after acquisition were similar, with a slight excess of conservation on the linear chromosome. This is at odds with the fact that the linear chromosome has a higher rates of gain and loss of genes. However, it corresponds to the observation of the majority of clade specific genes on this chromosome rather than on the circular one (Fig. 2.11 and 2.12). This suggests that the linear

chomosome is partitioned into two compartments: a dynamic one where new genes are easily inserted and deleted, and one which provides stable residence to core genes. As suggested by the occurrence of many species-specific gene in large clusters of genes, including several blocks of co-transferred genes (Sup. Tables 2.8), stable components may arise by insertion of a large HGT fragment, followed by accretion of other clade-specific genes. Such process may be mediated by arrest of recombination with other clades at non-homologous loci (Vetsigian and Goldenfeld, 2005; Lawrence and Retchless, 2009).

An unforeseen role for pAt plasmids as host of clade-specific adaptations The pAt plasmids of genomovar G1 genomes are the hosts of 37 species-specific genes, and similarly 25 G8-specific genes and 11 [G6-G8]-specific genes are born by the pAt plasmids of the corresponding strains. The occurrence of species-specific genes on a plasmid may seem paradoxal because of the mobile nature of plasmids and their usually broad host range in Rhizobiaceae, as seen by frequent transfers of parts or entire symbiotic (pSym) or tumorigenic (pTi) plasmids in *Rhizobum* and *Agrobacteium*, respectively (González et al., 2003; Gonzalez et al., 2010; Lassalle et al., 2011).

In a previous study based on using micro-arrays based on G8-C58 genome to explore the gene repertoire diversity of *At*, we showed there was neither G8-specific, nor G8 core genes located on plasmids of the reference strain C58 (Lassalle et al., 2011). This was interpreted as the obligatory location of ecologically important genes on the chromosomes rather than on facultative replicons. Here, we find that there are 25 G8-specific and 11 [G6-G8]-specific genes located on the strains' pAt, but no on the pTi. This nuances the former view, notably because every of the four tested strains of G8 and G6 appear to have a pAt (or at least genes homologous to it in the case of the draft genome of ATCC 31749). In fact, in our previous work using micro-array CGH, every tested strain possessed a 147-gene locus of the pAtC58 (located from Atu5419 to Atu5549 and over the orgin from Atu5000 to Atu5015), but strain LMG-46, for which no gene of pAtC58 was found present (Lassalle et al. (2011), Fig. 1 and Table S1), indicating a loss of the pAt by this strain.

A recent study shown that lineages of C58 strains maintained in laboratory collection for long time or obtained in mutation-accumulation experiment had accumulated very few mutations but independently had large deletions in pAtC58 that conferred a selective advantage because of the reduced cost of pAt carriage (Morton et al., 2013). Most interestingly, the observed large deletions together covered about two-third of the mega-plasmid, but never occurred in the G8-conserved region (Atu5419-Atu5015), while the direct-repeat motifs that mediated the large deletions were also present within and around this locus (Morton et al., 2013). This suggests that deletions in this regions could be counter-selected, maybe because of the presence on the pAt of functions essential to the bacterium's life. This is however mostly speculative because the laboratory growth conditions do not match the true ecology of the organism and the few cases investigated in this study do not provide material for the statistical assessment of the essentiality of this region.

The occurrence of so much conserved genes on the pAt – and never on the other plasmids – rather suggest that this particular extrachromosomal element may harbor genes fixed in the species (core genes) with essential functions. This 'chromid' behaviour (Harrison et al., 2010) was already described for other rhizobial megaplasmids, and in fact the linear chromosome of *A. tumefaciens* is of plasmid origin and is itself a bona fide chromid (Harrison et al., 2010). In some cases, clade-specific genes are born by different replicon depending on strains and most of this variation indeed occurs between linear and pAt chromids, suggesting their similar role in housing niche-adaptive genes. In fact, the Lc and pAt are of similar nature, in that they can harbor clade-specific genes while being highly plastic and recombinogenic. The latter trait was not testable here for the pAt in a similar manner than for the Lc, because their very high plasticity renders difficult the assignation of genes to plasmids in ancestral genomes; nonetheless, pAt are known to be conjugative and hence to recombine intensely between genomes (**?**).

The fact that genes specific of different clades of *At* could stably settle on the pAt reveals it as an ecologically important, probably not dispensable third replicon of *A. tumefaciens* genomes. Some isolates however, like G9-Hayward 0363, have no detectable plasmids, indicating that this genomes architecture with three 'main' replicons is not the rule.

2.4.4 Conclusion

We developed an original method for the reconstruction of the history of all genes in genomes and applied it to *A. tumefaciens*, revealing the dynamics of gene repertoire in this taxon. These dynamics were structured along the tree of species and within genomes, revealing patterns of purifying selection for genes specific to major clades of *At*. Most of these were organized in large blocks of co-evolved genes that encoded coherent pathways, which constitutes a departure from a neutral model of gene transfer in bacterial genomes. Genes specific to each genomic 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 and amino-acids. These clade-specific genes likely constitute parallel adaptations to life in interaction with host plants, which suggest that ecological differentiation of *At* clades occurred through partitioning of the ecological resources of plant rhizospheres. In addition we revealed the particularity of *At* as housing not one, but two chromids, i.e. the linear chromosome and the pAt plasmid. These replicons are highly plastic and recombinogenic plastic but nonetheless are the place of fixation of genes essential to the ecology of species

2.4.5 Material and Methods

2.4.5.1 Genome sequencing and assembly

Genomic DNA of the sixteen strains was prepared with the phenol-chloroform method. Libraries were prepared from genomic DNA sheared into inserts of median size of 8 kb. Raw

Strain code	Strain Name	Clones	MP_Coverage	SR_Coverage	Insert_size
ATU1A	CFBP 5771	AHY_B_CFBP5771	6X	10X	8kb
ATU1B	S56	AHY_C_S56	2.4X	11X	8kb
ATU1C	ICPPB TT111	AHY_A_TT111	3X	11X	8kb
ATU2A	CFBP 5494 / CIP 497-74	AHY_R_CIP497-74	8.6X	7X	8kb
ATU3A	CFBP 6623 / CIP 107443	AHY_D_CIP107443	5X	7X	8kb
ATU4A	B6	AHY_E_B6	8X	11X	8kb
ATU4B	CFBP 5621	AHY_G_CFBP5621	5X	9X	8kb
ATU4C	Kerr 14 / CFBP 5761	AHY_F_Kerr14	6.2X	7.4X	8kb
ATU5A	CFBP 6626 / CIP 107444	AHY_H_CIP107444	7X	10X	8kb
ATU6A	NCPPB 925	AHY_I_NCPPB925	7X	7.5X	8kb
ATU7A	NCPPB 1641	AHY_M_NCPPB1641	7X	10X	8kb
ATU7B	RV3	AHY_K_RV3	6X	8.7X	8kb
ATU7C	Zutra 3/1	AHY_L_Zutra3/1	6X	6.5X	8kb
ATU8A	J0-7	AHY_N_J0-7	6X	9.1X	8kb
ATU9A	Hayward 0363	AHY_P_Hayward0363	6.6X	8X	8kb
ATU13	CFBP 6927	AHY_Q_CFBP6927	5X	7X	8kb

Figure 2.13: **Statistics of the 16 new genome sequences.**
MP: mate-pair reads, SR: single reads.

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 ref genome sequencing). 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 sequnces were then annotated with the MicroScope pipeline (Vallenet et al., 2009; **?**) and made available through the MaGe web interface (`https://www.genoscope.cns.fr/agc/microscope/about/collabprojects.php?P_id=51`).

2.4.5.2 Construction of Phylogenomic Database

Our pipeline for database construction is mostly that of Hogenom databases (Penel et al., 2009) up to the gene tree building and is summarized here.

Genomes were retrieved from ENA (`http://www.ebi.ac.uk/ena/`) and Microscope (`https://www.genoscope.cns.fr/agc/microscope/about/collabprojects.php?P_id=51`) databases. The genome sequences of the dataset and details about new sequences from this study are listed Table 1. EMBL flat files were used to build an ACNUC database (Gouy et al., 1985) subsequently used for sequence retrieval from keyword-based queries. This database named 'agrogenom' is accessible through client tool Query (`http://pbil.univ-lyon1.fr/databases/acnuc/acnuc.html`).

CDS sub-sequences were extracted and translated to protein sequences that were compared through an all-against-all BLASTP run (Altschul et al., 1990). The BLAST hits matrix was used to build a similarity network with SiLiX program (version 1.2.5-p1, default parameters: proteins are linked if they present a hit of $\geq$ 35% over $\geq$ 80% length) (Miele et al., 2011) which

was in turn used to infer families of homologous proteins with HiFiX program (version 1.0.3, default parameters) (Miele et al., 2012). Gene family information was stored in the 'agrogenom' ACNUC database and protein sequences were retrieved by families in Fasta format.

Protein families were aligned using MUSCLE (version 3.8.31, default parameters) (Edgar, 2004) and then retro-translated with pal2nal program (version 14) (Suyama et al., 2006). Nucleic alignments were restricted to conserved blocks with Gblocks (version 0.91b, minimum 50% of sequences in conserved and flank positions and all gaps allowed, DNA mode) (Castresana, 2000) and gene trees were computed from these alignment with PhyML (version 3.0, GTR+Γ8+I model of evolution, best of SPR and NNI moves, SH-like supports) (Guindon and Gascuel, 2003).

A PostgreSQL relational database was built by an extension of the Phylarianne schema (`http://phylariane.univ-lyon1.fr/`) to house data relative to genomes, gene and species trees, reconciliations, block events and functional annotations. A glimpse of the database shema is presented Fig. 2.14. This database is accessible through an interactive graphical web interface at `http://phylariane.univ-lyon1.fr/db/agrogenom/3/`

2.4.5.3 Reference species tree

455 universal and unicopy gene families (i.e. families with exactly one copy per genome, listed Sup. Table **??**) were used to infer the reference phylogeny. 500 independent jacknife sampling were done, with each replicate being a concatenate of alignments (as described above) of 25 families randomly sampled without replacement. Trees were computed with PhyML (same parameters as for gene trees) (Guindon and Gascuel, 2003)). Their consensus tree (extended majority majority rule) was computed with Consense algorithm from the Phylip package (Felsenstein, 1993) (Figure 2.3).

Alternative phylogenies were searched using the whole concatenate of 455 universal unicopy families or from a concatenate of 49 ribosomal protein gene families (Sup. Table **??**) 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, except for relative placements of genomic species within the *A. tumefaciens* complex (Fig. 2.18).

2.4.5.4 Reconciliation of gene tree with the species tree

Gene trees were first rooted with TPMS program (Bigot et al., 2013) using the combo criterion minimizing both unicity of subtrees (implicit minimization of ancient duplication events) and their taxonomic depth (implicit minimization of ancient transfer events, knowing the species phylogeny). Then, using a custom Python library (alfacinha.tree2, Sup. Mat. 2.7.7), subtree pruning and regrafting (SPR) moves were attempted on non-supported branches of gene trees (SH-like support < 0.9) in order to minimize the number of duplication events (merging events) or the number of tree leaves involved (bring events closer to the tips). Branch

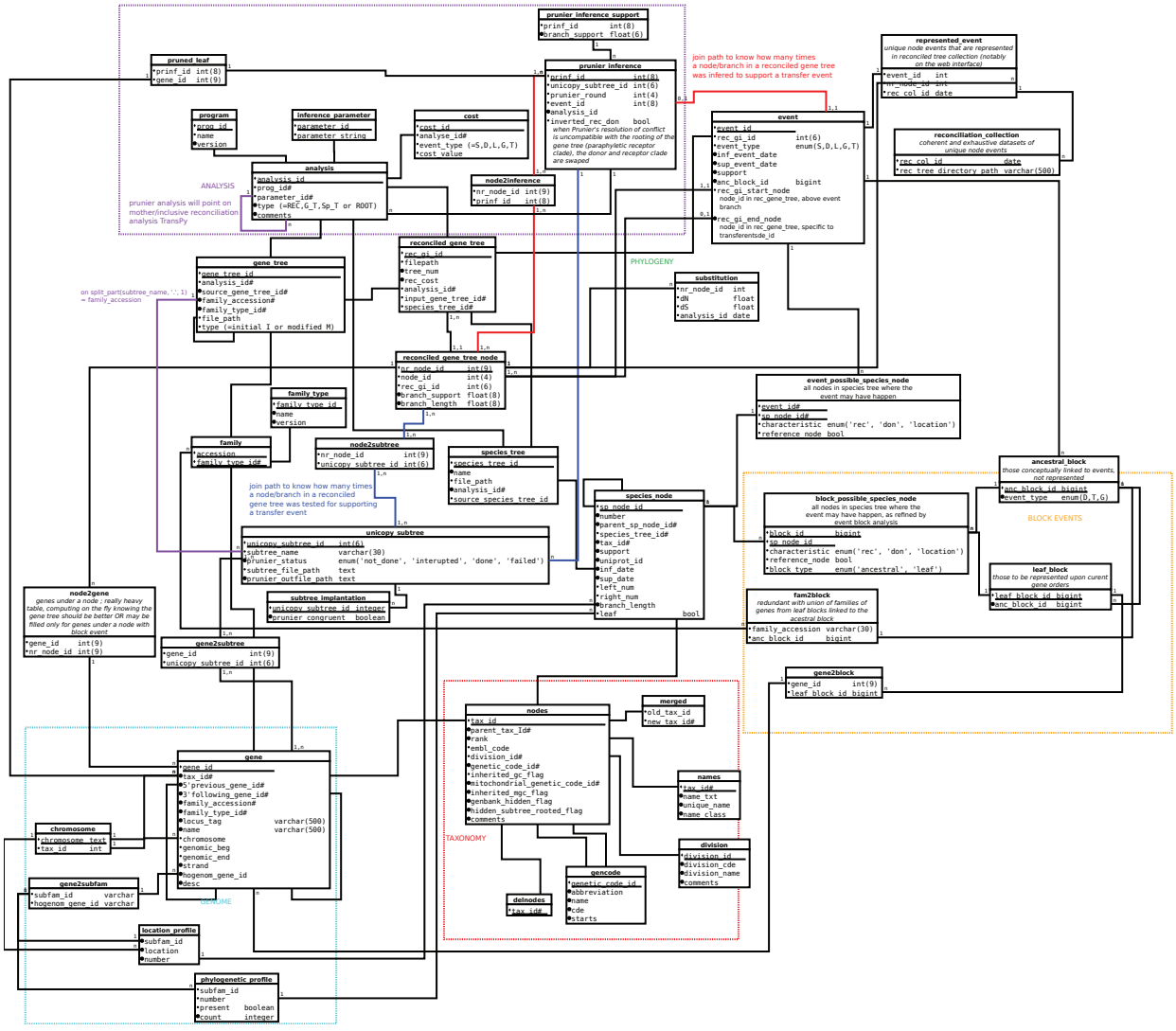


Figure 2.14: Schema of Agrogenom relational database

lengths and supports were re-estimated with PhyML (Guindon and Gascuel, 2003) using the same model as previously.

Reconciliation was then done using a specially developed Python library that allow manipulation of phylogenetic trees and providing an API to interact with diverse phylogenetic programs and with a dedicated relational database, Agrogenom, where all reconcilation data were stored. This was done in a multi-step procedure as follows:

(1) Identification in a full gene tree of subtrees of (co-)orthologous (unicopy) genes to be tested for transfers with Prunier (Abby et al., 2010). All possible combination of (co-)orthologs were explored, with each gene being potentially represented in several gene combination sets ; the majority of nodes in the full gene tree are thus covered by several unicopy subtrees that will be tested independently (Fig. 2.4, step 2).

(2) Detection of transfers with Prunier on unicopy subtrees and mapping of independent transfer inferences to nodes of the full gene tree ; nodes covered by several independent Prunier tests may have been or not detected as a transfer depending on the test, and several inferences of transfer may yield different scenarios (transfer direction and/or precision of the mapping to the reference species tree, see detailed procedure in Sup. Mat. 2.7.2).

(3) For such nodes with several possibilities of events, choice of the event that was inferred in the majority of cases (potentially rejecting minority reports of transfers) (Fig. 2.4, step 3). The transfer events inferred here are highly conservative because supported by a reconciliation method that models the cost of inferring speciations vs. HGT given the gene tree topology, branch lengths and supports in 'maximum statistical agreement forest' framework (Abby et al., 2010). These transfer inferences are supported themselves by replication of the transfer inference with independent leaf sets.

(4) Completion of the reconciliation by resolution of unicity conflict (representation of a species under both children of a node) either by inference of transfers in presence of taxonomic incongruence (as defined in TPMS-XD algorithm (Bigot et al., 2013)) or by inference of duplications (Fig. 2.4, step 4). The transfer events inferred here are moderately reliable because they are supported by topological incongruences in the gene tree that may lack branch support.

(5) Finally, ancestral gene contents are inferred by reconstructing the history of gene gain and loss independently for each sub-family of orthologs (obtained by pruning subtrees from the full gene tree at every node annotated as a duplication or transfer) (Fig. 2.4, step 5). Inference of gain/loss scenarios was assayed with two methods:

    1. Dollo parsimony, that map a unique gain to the last common ancestor (LCA) of all the species represented in the subtree, and losses in every lineage above it missing the gene;
    2. Wagner parsimony, that returns the most parsimonious combination of gene gains and losses given a gain/loss penalty ratio. The latter parameter was set to 1 for

a best compromise between keeping coherent ancestral genome sizes (penalizing losses enough to have the genome size of an ancestor smaller or equal to that of its children) and coherent gene histories (penalizing gains enough to avoid multiple gains happening among groups of closely related strains);

The latter method was used as implemented in Count software (Csűrös, 2008). This method infers several gains that are not mapped to the LCA of the species represented in the subtree, and these additional gains needed to be interpreted as transfers that were mapped to nodes in the full gene trees. The transfer events inferred here are the least reliable since they are not supported by topological incongruences between the gene and species tree. Nevertheless, they greatly improves the reconciliation in case of clade-specific groups of genes by avoiding to map their acquisition too high in the species phylogeny and inferring excessive counts of losses (Fig. 2.4, step 6).

The results of the Wagner parsimony method were chosen because it reconstructs ancestral genomes with minor variation of size when going back in time compared to Dollo parsimony.

The reconciliation of gene trees can be translated into an evolutionary scenario in the species tree, where presence/absence states and the duplication, transfer and speciation events are mapped on the reference tree (Fig. 2.4, step 7). In particular, transfer events are characterized by the identification of both a donor and a receiver, which specifies the direction of the transfer. However, because we did not limited our analysis to core genes, extant genomes can be heterogeneously represented in the gene tree. In these cases, the combination of DTS events can be interpreted with several alternative evolutionary scenarios that differ by the number of hidden loss events and by the location of DTS events in the reference tree (Fig. 2.15 A,B). Among these, one can be chosen as the most parsimonious in losses. This approach is however sensitive to the relative costs assigned to each events (Doyon et al., 2011), and determination of the costs best describing the evolution of each gene family and subfamily lineages is a tricky problem.

To orientate this choice, we used the regional signal in genomes: we searched neighbour genes that could agree on possible scenarios of transfer, origination or duplication inferred in their respective gene trees (Fig. 2.4, step 8). If such compatible neighbour events were to be found, the hypothesis of a unique common event would be more parsimonious in transfers, duplications or originations in the global genome scenario. Therefore, rather than immediately determining a simple location for an event – or a simple couple of donor/receiver for a transfer – we kept the secenarios uncertain by mapping the event to a set of possible nodes in the reference tree (Fig. 2.15 C).

The complete reconciliation scenario were then integrated into the Agrogenom database, that notably stored the possible coordinates in the species tree of the various transfer, duplication and speciation events inferred in the gene trees.

This whole reconciliation procedure is detailed in Supplementary Materials section 2.7.2.

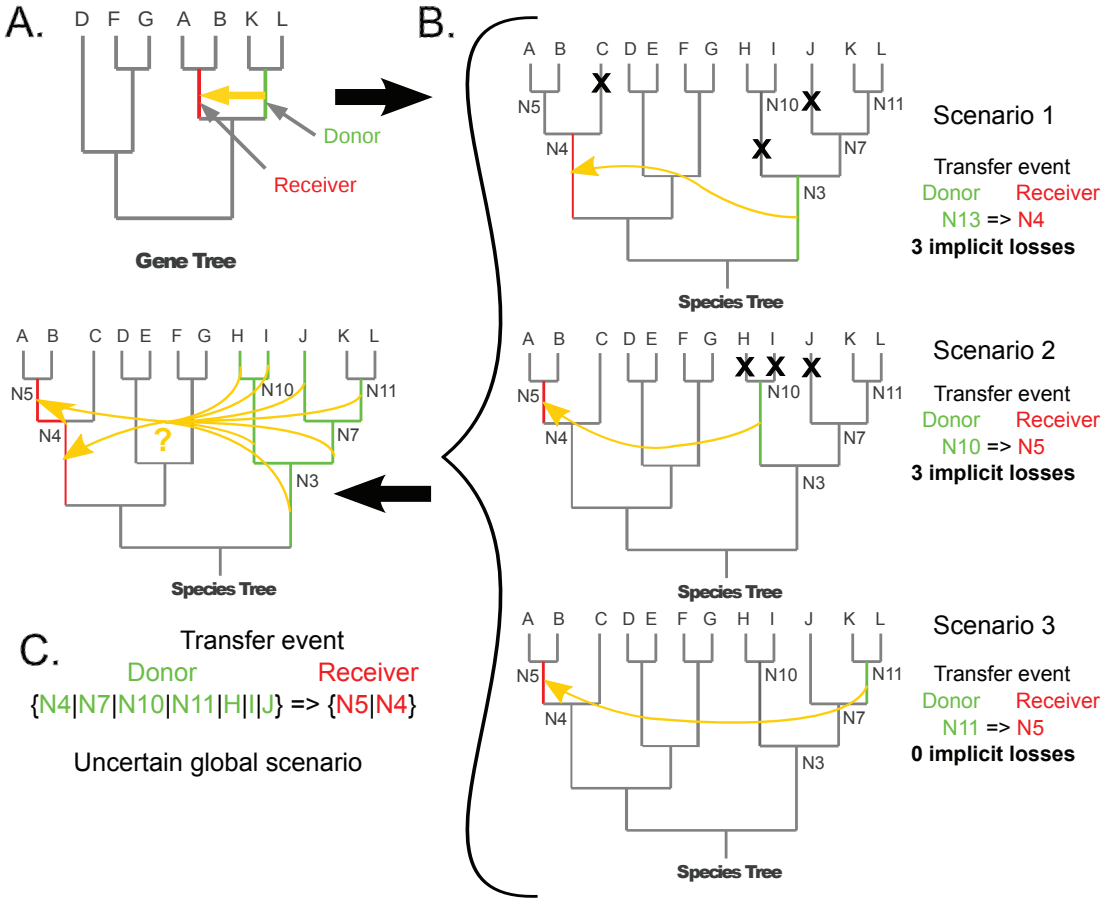


Figure 2.15: Uncertainties in inference of transfer event coordinates in the species trees. Given the gene tree (A), several scenarios involving different donor and receiver pairs are possible (B). In the three examples of scenarios shown in (B), scenarios 1 and 2 induce the same cost of three convergent losses; the scenario 3 has the donor and receiver mapped to the LCA of genomes positive for the gene, and thus is the most parsimonious, with no losses. However this choice may not be the true one, as can be revealed by neighbouring genes with a stronger signal for one of the other possible scenarios. Until the transfer scenario can be confronted to the reconciled histories of neighbour genes, the uncertainty on its coordinates is preserved (C).

2.4.5.5 Block events reconstruction

Duplication or transfer events can involve several consecutive genes on a replicon. To have a more realistic view of frequency and extent of those events, recognition of such blocks is necessary. This is also a way to explain the occurrence of many convergent single-gene transfers between species, and thus to obtain evolutionary scenarios more parsimonious on the number of unique transfer, duplication and origination events.

We seek for tracks of genes whose lineages (path from the gene leaf to the root of the gene tree) share similar events, given the event coordinate system described above. However, different genes involved in a same block event may have very different histories precluding or following the focused event. Therefore, while a 'core' signal for the event shall be conserved

across genes, the uncertainties of the mappings of each gene event in the species tree may be different (Fig. 2.16). Thus, we define as a block event a block of consecutive genes having a non-null intersection of their sets of possible locations in the species tree, and this intersection is the refined location of the block event.

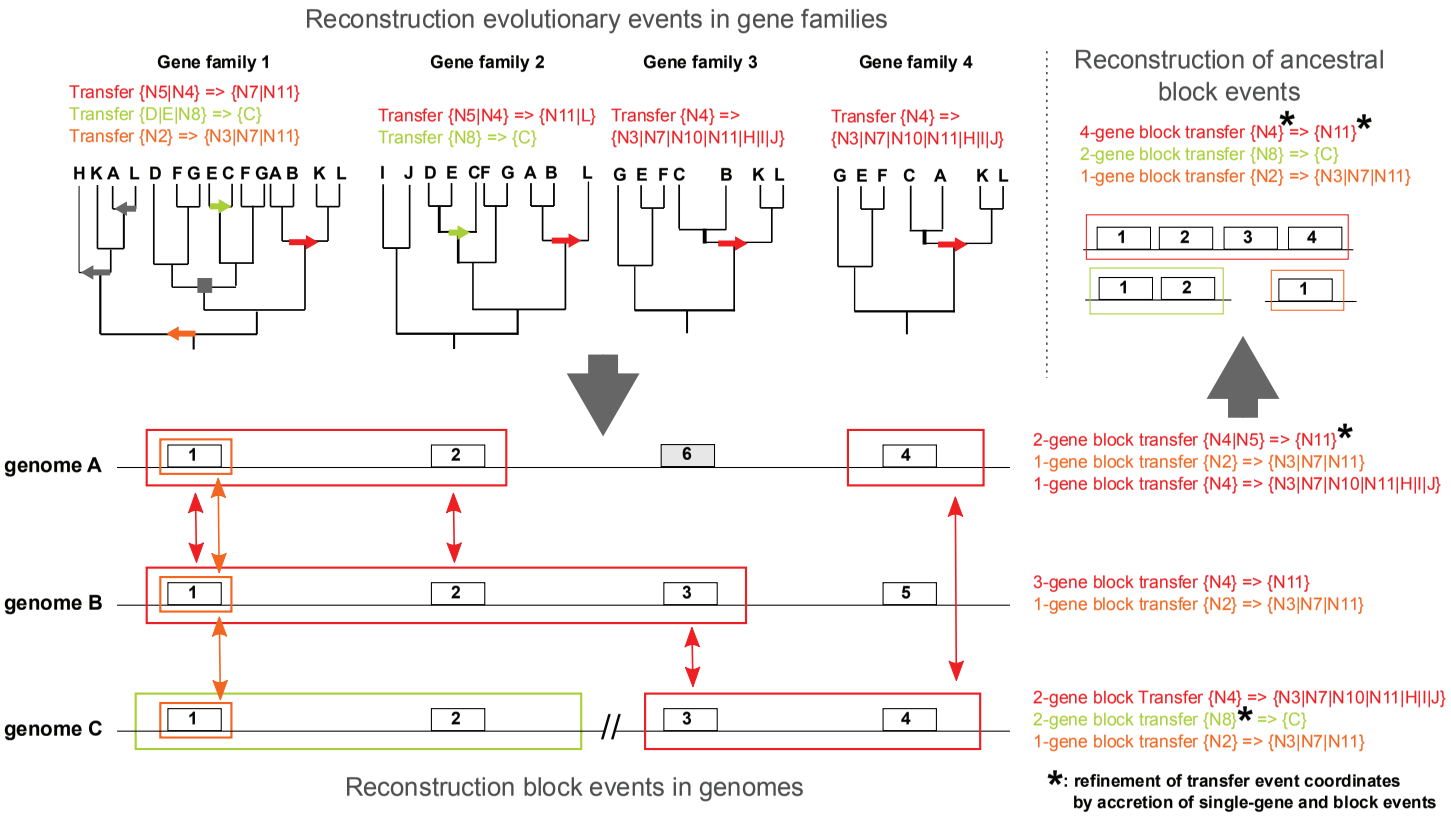


Figure 2.16: Refinement of event uncertainties when building block events.

To avoid redundancy of blocks initiated on successive genes with compatible events, gene events that were already allocated to a block are not used to initiate another block. The algorithm for block reconstruction is thus 'greedy', meaning than it intends to build the largest non-redundant blocks of compatible events.

In case of transfer block events, blocks containing gap genes are then checked for phylogenetic compatibility of those gap genes with the scenario associated to the block. Indeed, a gene tree may not have signal for the transfer scenario (i.e., weak statistical support at the crucial branches supporting the phylogenetic conflict signing the transfer), and could still not reject the possibility of this transfer event. At the opposite, the gene tree of the gap gene may bear signal rejecting the transfer event, in the form of strongly supported bipartitions showing that donor and receptor clades are separated from each other in the gene tree. In such a case, the gap gene showing signal against the block event scenario, it must be excluded and the original block split into son blocks, each one representing an independent transfer event. This controls for the risk of artefactual agglomeration of independent events ; however, it may also part genes that were affected by a same event as ancient neighbours, but were separated by the insertion of a gene with a different history.

Those cases of artefactually dissociated gene events are often recovered as one same ancestral event through the reconstruction of ancestral block events.

These blocks of co-evolved genes found contiguous in a genome are the extant form of an event that happened in the genome of an ancestral species. We thus expect to find homologous blocks of co-evolved genes in other descendants of this ancestor. Those homologous blocks are linked by the events their homologous gene parts share. Thus, the ancestral block event that took place in the ancestral genome can be rebuild by accretion of homologous block events from several 'leaf' genomes. As inferences of leaf block events can differ among leaf genomes because of partially independent histories of gene neighbourhood evolution (losses, insertions, rearrangements), blocks that were dissociated in some leaf genome may appear intact in other. In this case, this accretion procedure links disjoint leaf blocks to one ancestral block, allowing us to recover the unity of the block event. Similarly, the possible locations of the events in the species tree inferred in each of the leaf blocks can differ given the pattern of homologs in presence in each leaf genome. Thus, the accretion of leaf blocks into a ancestral block can refine the set of possible locations of the block event, in an analogous way than for the grouping of genes into leaf blocks.

2.4.5.6 Definition of clade-specific genes from phylogenetic profiles

Species-specific genes are genes exclusively found in a clade that are thought to have been gained by the clade ancestor and since then conserved, and are therefore good candidates to have contributed to the ancestor ecological adaptation and reproductive success.

From the sub-division of homologous gene families into orthologous subfamilies (see above and Sup. Mat. 2.7.2, step 5.1), we can establish the phylogenetic profile of presence or absence of each subfamily in extant genomes. From these profiles, we wanted to identify contrasting

patterns of presence/absence revealing genotypes specific of clades. Searching clade-specific state is tricky since it needs first to define what is the contrasting clade (a parent clade of the focus clade in which the genomes of the focus clade show a consistent pattern opposite to that of all other external genomes). Because of possible transfer events blurring the pattern and of taxonomic sampling bias in the genome dataset, it is hard to define a strict and general rule for finding specific patterns of gene presence. We thus took advantage of the reconciliation and ancestral genome history data to to search gene subfamilies specifically present/absent in a clade from the set of subfamilies gained/lost by the clade ancestor. This allowed us to have a fuzzy definition of specificity, where the presence/absence contrast between focus and external clades can be blurred by subsequent transfer to, or independent lost in an external group. We thus obtain a list of gene gains and losses corresponding to genomic synapomorphies of focus clades.

2.4.5.7 Ancestral location of genes on replicons

Each molecule of contemporary genomes in the Agrogenom database (complete replicons or contigs/scaffolds) was manually assigned to a type of replicon among the following factorial states: 'primary' (main chromosomes, including circular chromosomes in *A. tumefaciens*), 'secondary' (large replicons with core genes restricted to Rhizobiaceae [Slater et al., 2009], a.k.a. chromids [Harisson et al., 2010], including linear chromosomes in *A. tumefaciens*), 'pAt' (in *A. tumefaciens* only, conserved megaplasmid identified as not tumorigenic), 'pTi' (in *A. tumefaciens* only, megaplasmid identified as tumorigenic), 'plasmid' (any other plasmids, including potential pAt or pTi megaplasmids that cannot be firmly identified as such, symbiotic megaplasmids and smaller plasmids) and 'unknown' (when the assembly of the strain's genome do not allow to map the contig/scaffold to a replicon). These states were transferred to the genes occupying the molecules and were mapped to the species phylogeny by gene subfamily, using 'absent' state when the subfamily was not present in a genome.

Following the species phylogeny and the reconstructed occurrence profile of sufamilies in ancestral genomes, the replicon location of subfamilies were propagated to ancestral genomes using Fitch's parsimony algorithm (Fitch, 1971). This method can result in several states being possible at a node, owing to equally parsimonious scenarios. The 'unknown' state is discarded at nodes when another state is proposed at the node, otherwise it is replaced by the next 'known' ancestral state (i.e. in the lineage from the node to the root); this notably allow to infer a location for genes from unfinished genomes (leaf genomes) from the closest homolog location. When several 'known' states were possible at a node, we attempted to restrict the set of proposed states by looking at ancestral states (see Supplementary Material for detailed algorithm).

2.4.5.8 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. Patterns generally describe the local monophyly of two groups (e.g. G2 and [G4-G7-G9]) within subtrees containing only *A. tumefaciens* 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 whithout constrains on their monophyly, to get the number of genes for which the monophyly hypothesis could be tested. Patterns with their translation into TPMS pseudo-Newick formalism referring to reference tree nodes are listed in Supplementary Material.

2.4.5.9 Detection of recombination in core genes

Core gene families were analysed for clades within *At* for which we had at least 4 strains, which is the minimum to observe potential homoplasy in sequence alignments. We extracted from the Agrogenom database the sequence alignments of the gene families of their respective core genome, and applied the test PHI (Bruen et al., 2006) to detect signatures of recombination in these alignments. Families were considered recombinant when the p-value of the permutation test of PHI was ≤ 0.05. When comparing the proportion of recombinant families in the core genome of *At* to core families of younger clades, we had to account for the positive bias of sensitivity of PHI with the number of sequences in alignments. We build comparable datasets by reducing the size of *At* core alignments to the same number of sequences as in the other clade, taking 10 random samples of sequences per *At* core family to smooth the effect of sampling biases.

2.4.5.10 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., 2009) : 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 annoatations were then mapped to gene products using the mapings 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. We note that we excluded manual annotations of of the

annotation dataset. Though manual annotations are certainly the most accurate relative to the true function of genes, they may be biased in their distribution along replicons, because some studies have focused on the functional characterization of an operon or a locus identified as transferred. This would have evidently biased our analysis of the functional clustering of genes, so we limited the annotation dataset to the electronicaly inferred ones – which are mostly based on the recognition of conserved sequence motifs, thus independent of their position on replicons. In addition, the annotations based on such homology search is expected to be more homogeneous accross genomes, would they belong to a model organism or not.

The obtained functionnal annotations of proteomes were analysed in the frame of Gene Ontology term reference (Full ontology file dowloaded 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 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). Computation were done using a custom Python package derived from AIGO 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, we compared the FH of transferred gene blocks 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 of 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 considerred for statistical comparisons.

2.5 Supplementary Figures

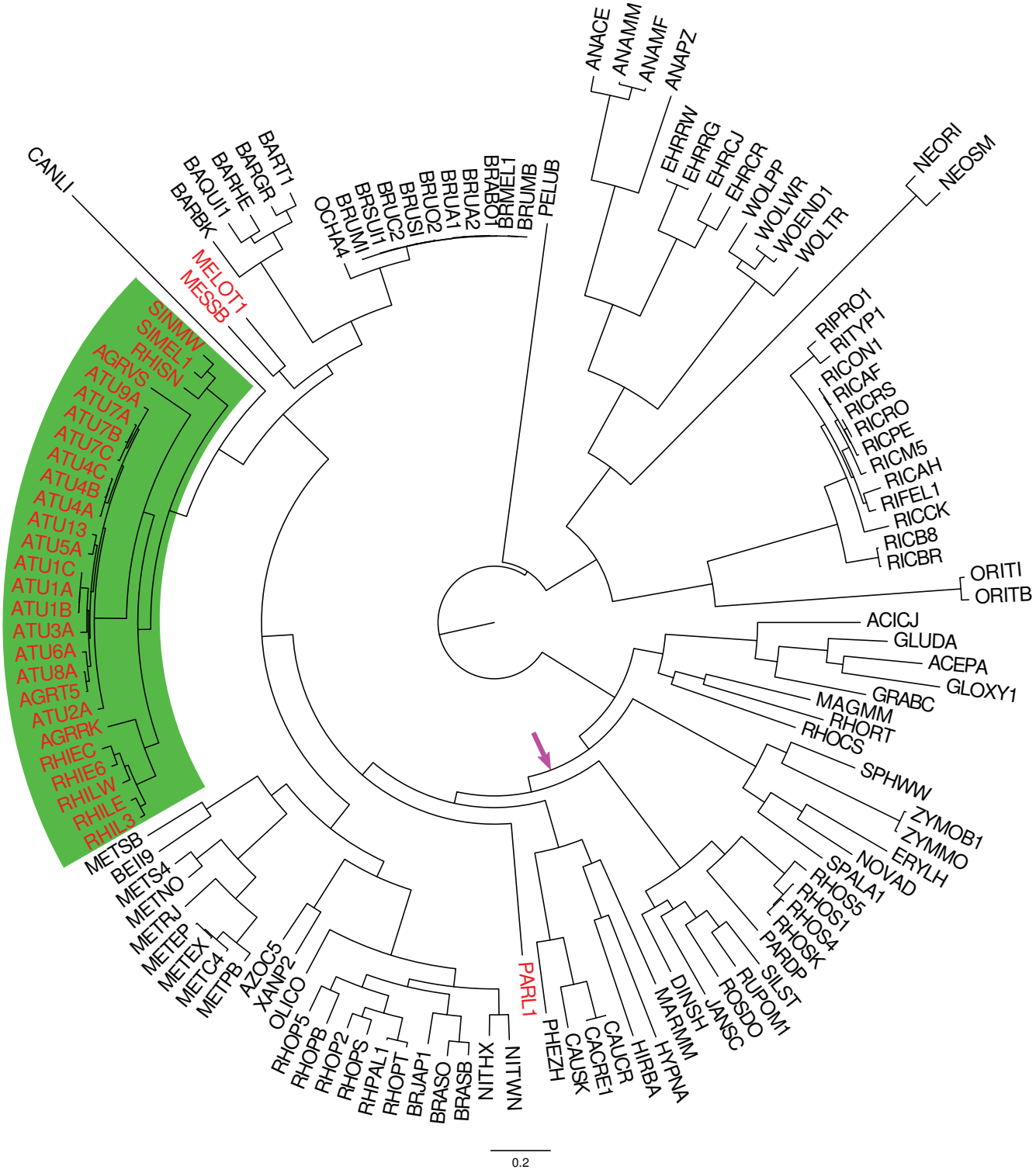


Figure 2.17: **Phylogram representing a phylogeny of 131 genomes of Alpha-proteobacteria.** Tip names correspond to those defined in Table 2.1 and to the Uniprot 5-letter code of organisms (`http://www.uniprot.org/docs/speclist`). The Rhizobiaceae clade is highlighted in green. Species that were integrated in the 47-Rhizobiales dataset are coloured in red.
This tree was obtained by maximum-likelihood (ML) from the concatenated alignment of 61 universal unicopy gene families using RAxML (version 7.0.4) (Stamatakis, 2006) with PROTMIXWAGF model (with 25 rate categroris) and with a start tree obtained from the consensus (using CONSENSE program from PHYLIP package (Felsenstein, 1993), MRE rule) of the 61 individual ML trees obtained with PhyML (Guindon and Gascuel, 2003) under a LG+4G model. The tree is rooted as proposed in Williams et al. (2007) based on the branching of outgroups; the purple arrow indicates an alternative root proposed in Abby et al. (2012) that minimized the number of horizontal transfer events in a reconciliation approach.

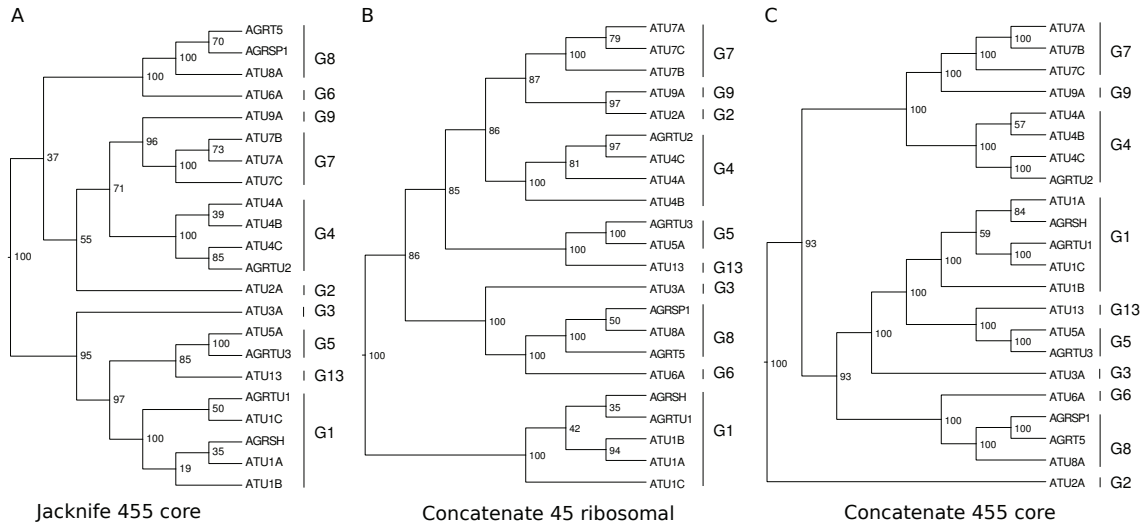


Figure 2.18: **Alternative reference tree topologies obtained with different methods:** (A) Consensus of ML trees obtained from concatenated alignments of jacknife gene samples (500 draws of 25 genes among 455 core genes), branch supports are bases on the 500 jacknife trees; (B) ML tree obtained from concatenated alignments of ribosomal protein genes, branch supports are based on 1,000 random bootstrap replicates ; (C) ML tree obtained from concatenated alignments of 455 core genes, branch supports are based on 200 random bootstrap replicates.

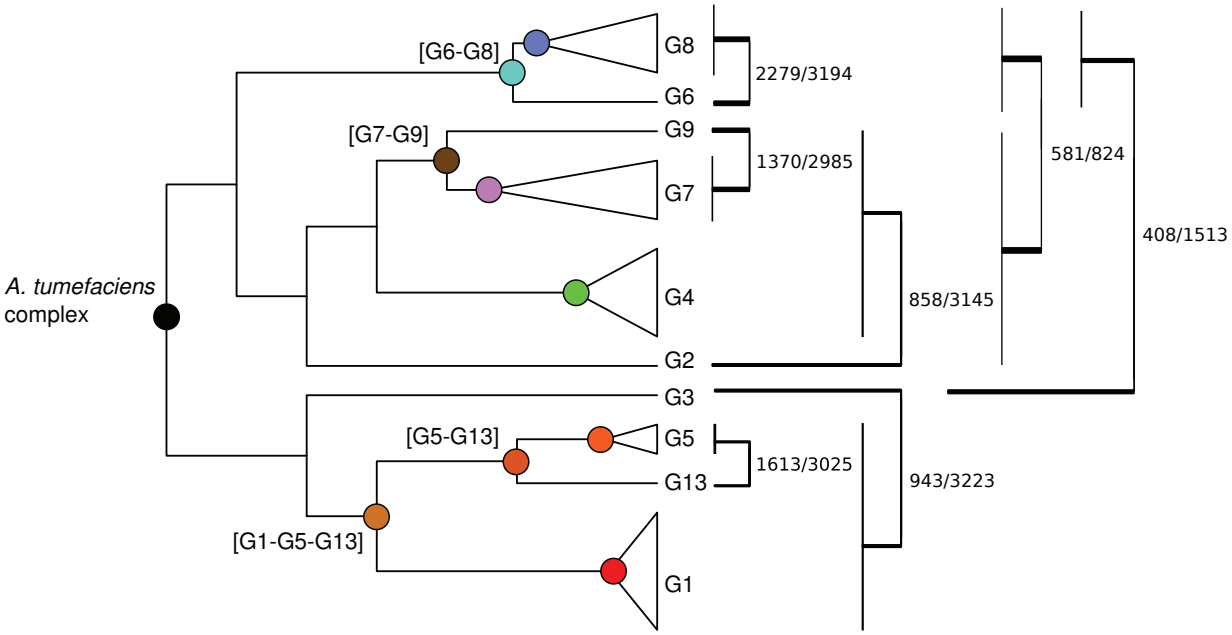


Figure 2.19: **Support for monophyply of pairs of groups in all gene family trees.** Tested pairs are indicated by brackets on the right. Numbers indicate the count of gene families in the Agrogenom database that matched a pattern of monophyly par the pair over the number of families for which the monophyly could be tested (i.e. given the set of represented species).

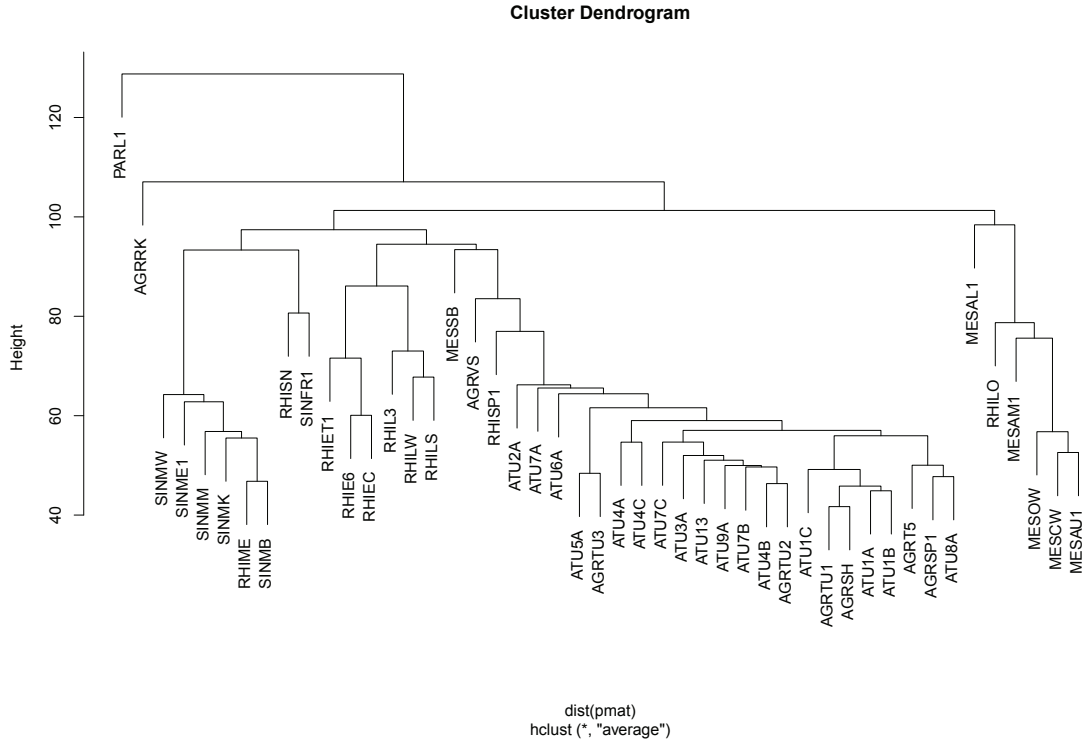


Figure 2.20: **Hierarchical clustering of *Agrobacterium* strains on genome gene content.** Hierarchical clustering was performed using presence/absence profiles of 41,664 gene families, using euclidean distances and UPGMA algorithm ("average" method in hclust() function from R environment (Team, 2009))

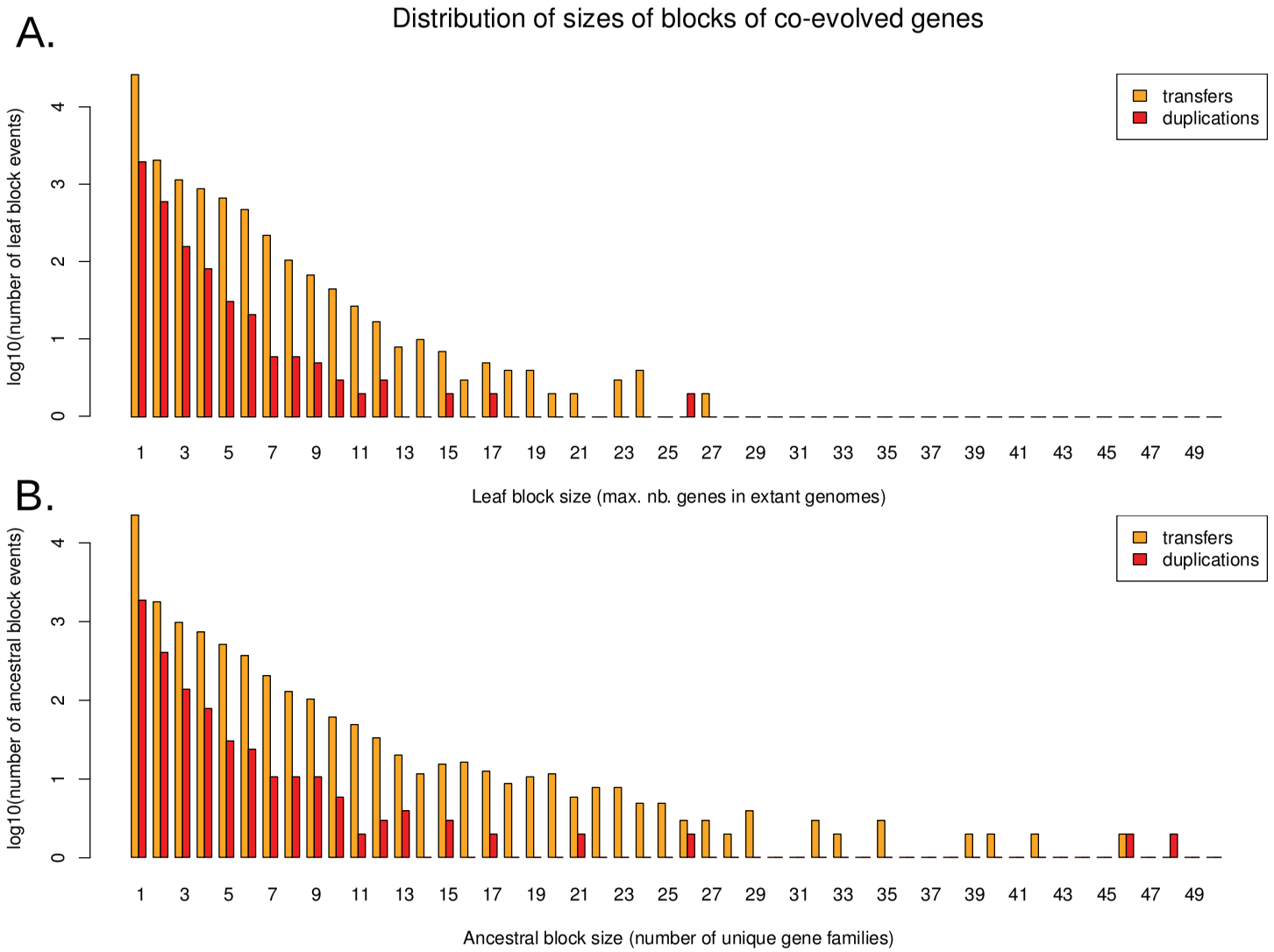


Figure 2.21: **Distribution of sizes of block events.**

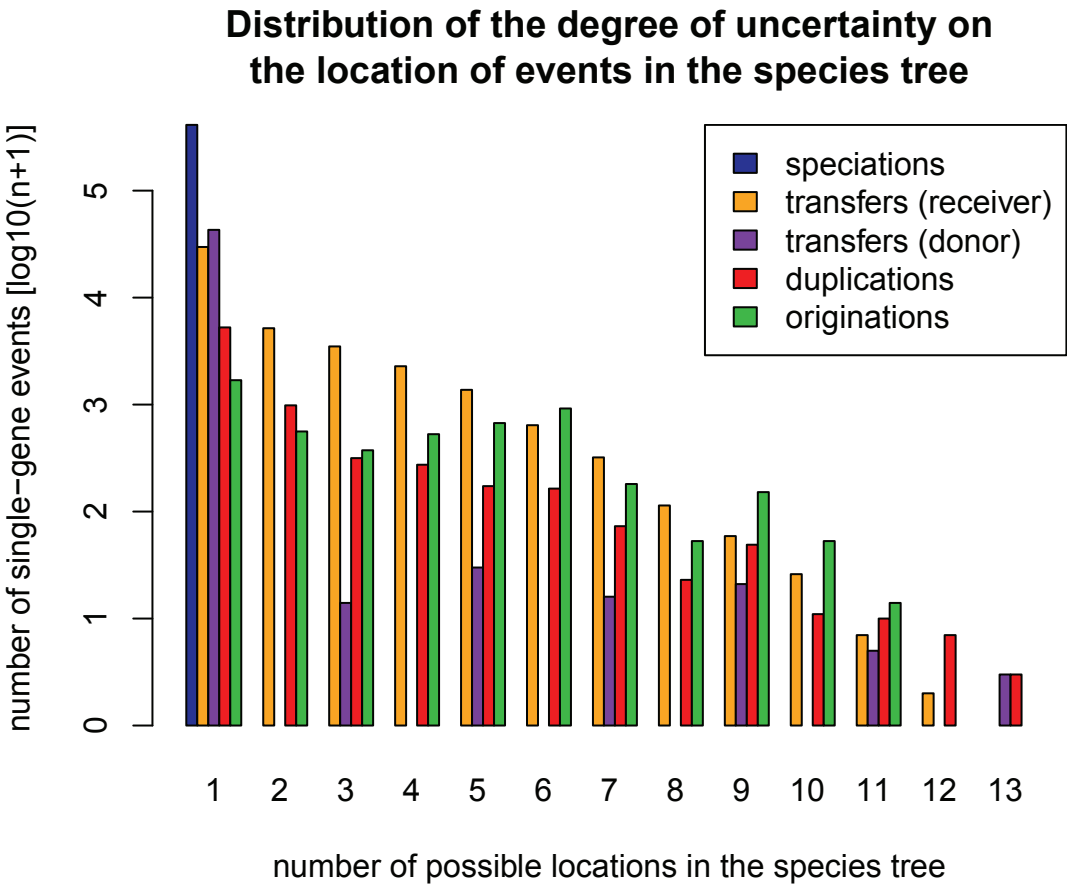


Figure 2.22: **Distribution of the degree of uncertainty on the location of single-gene events**, i.e. before any refinement by amalgamation of scenarios in blocks.

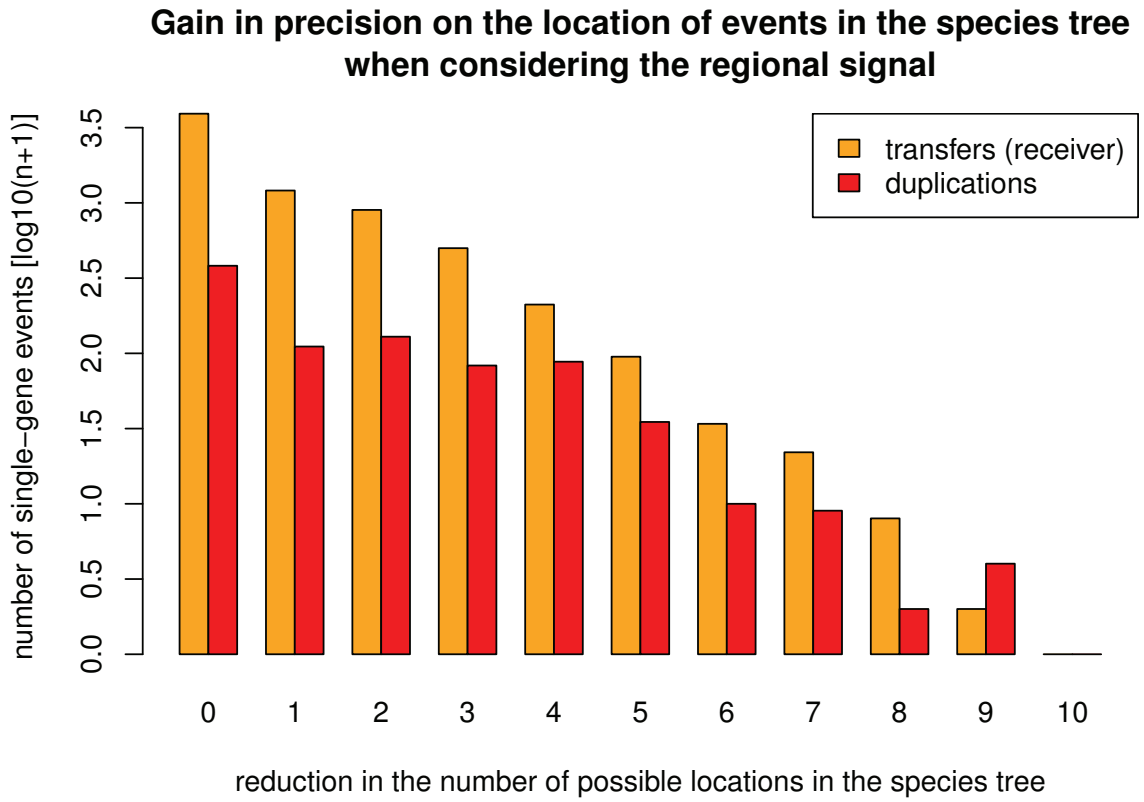


Figure 2.23: **Gain in reconciliation precision between scenarios** of single gene events and amalgamated scenarios of corresponding block events. Gain in precision corresponds to the difference between the numbers of possible nodes of the reference tree before and after amalgamation of duplication or transfer scenarios; zero point means no amelioration of the signal. For instance, for toy events presented Fig. 2.16, the red event in gene family 1 has two possible receivers, N7 and N11, and only one (N11) remains after event amalgamation; the gain on precision is thus of $2 - 1 = 1$ point. Similarly, the orange event and the green event in family 1 are both refined of zero point. Plot represents gain in precision for blocks involving ≥ 2 single-gene events which could be refined (i.e. whose location was uncertain).

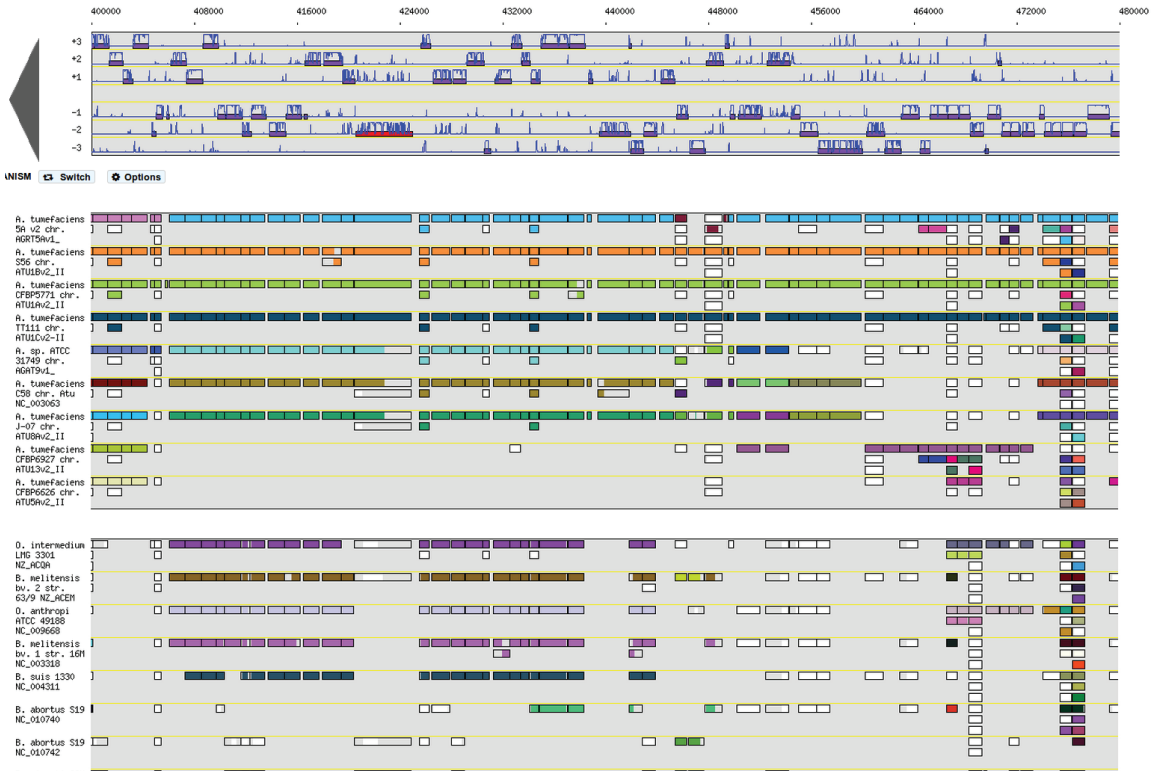


Figure 2.24: **Syntenic conservation of AtSp14 cluster in G1, G8 and Brucelaceae.**
On top, the reference locus on H13-3 linear chromosome, and its syntenic conservation is projected on genomes of strains of genomovar G1 and G8 and Brucellaceae (view of MaGe interface).
These genes code the biosynthesis of an O-antigen decoration of the lipopolysaccharide (LPS) ; HHSS gene is coloured in red.

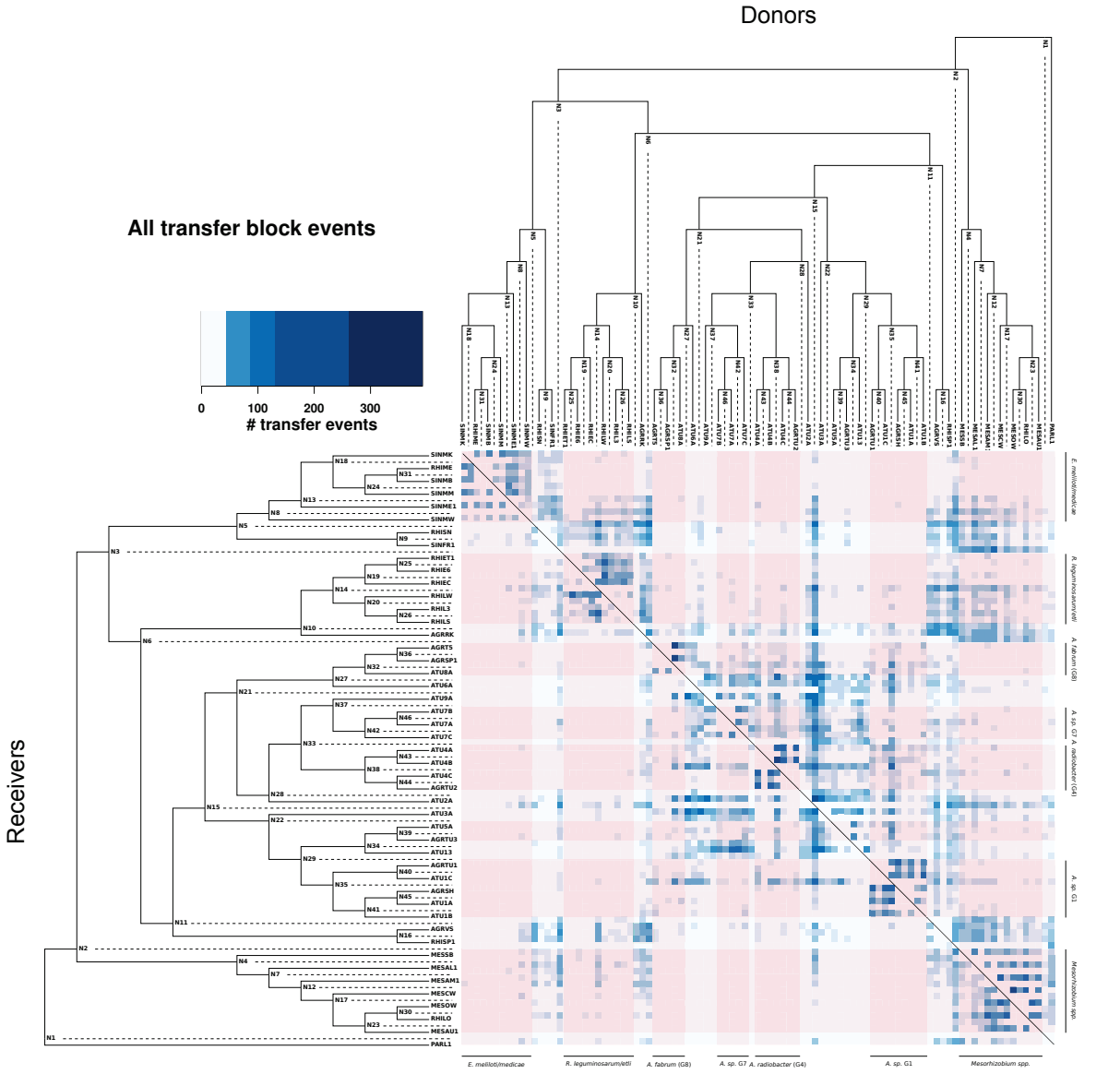


Figure 2.25: **Intensity of gene transfers among Rhizobiales.**
Heatmap of transfers that occurred between pairs of ancestral or contemporary genomes. Each dot of the matrix indicate the intensity of transfer between nodes of the reference phylogeny drawn on left and top (trees are cladogram, branch length have no meaning). Rows are receivers, columns are donors. Pink shading highlight coherent species or groups of species (corresponding taxon names on the right and bottom). Note that a block event is counted as one event. In case of remaining uncertainty on the location of the inferred receiver and donor, the count of the event id uniformly fractioned between possible receivers / donors. This matrix is the sum of those presented 2.27 and 2.26.

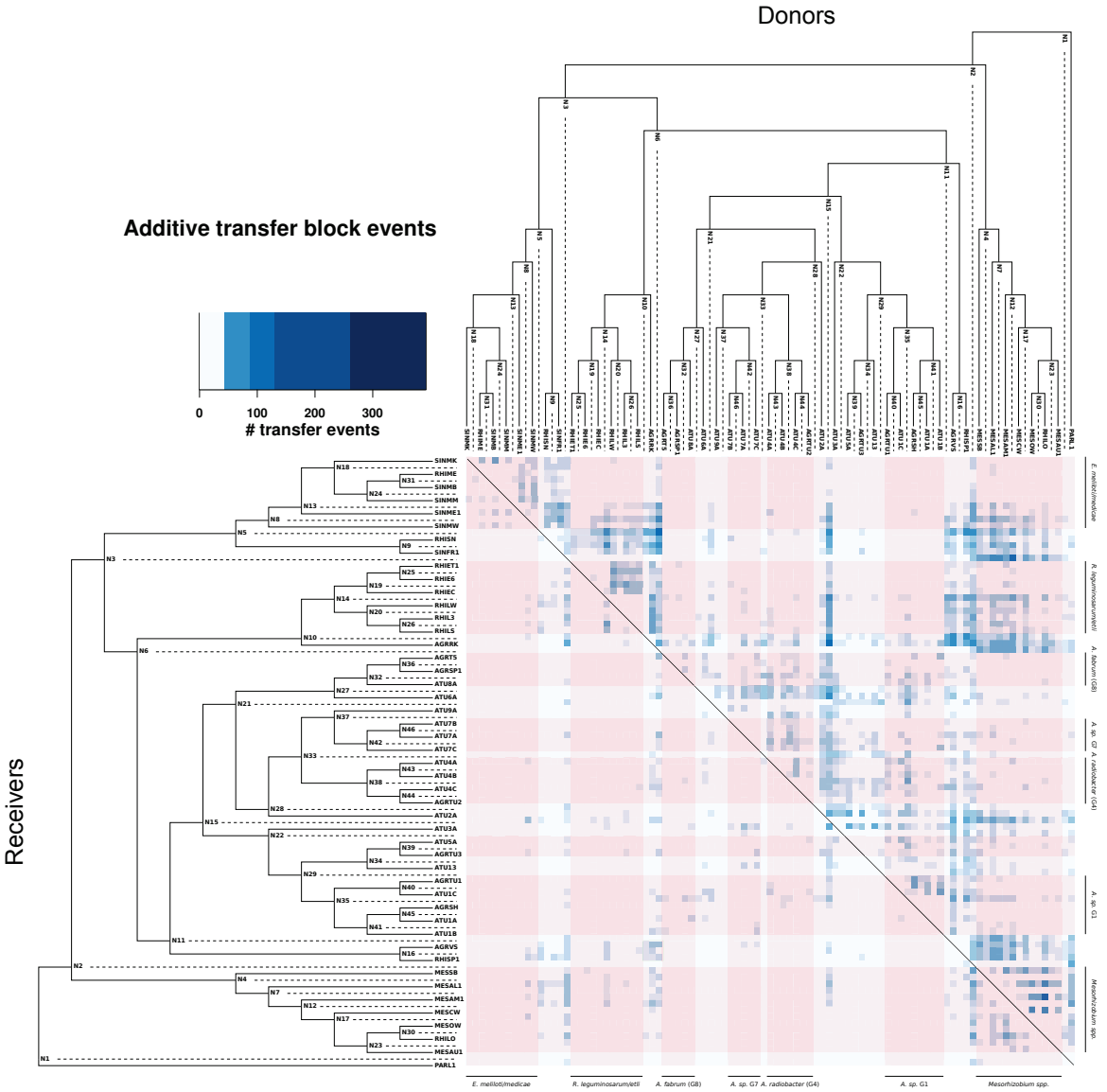


Figure 2.26: **Intensity of additive gene transfers among Rhizobiales.**
Legend as in 2.25. When there is no signal for orientating the additive transfer, i.e. choosing the donor, an arbitrary choice is made, that always take the same donor in a given pair. This artefact of the method accounts for the apparent asymmetry of the map.

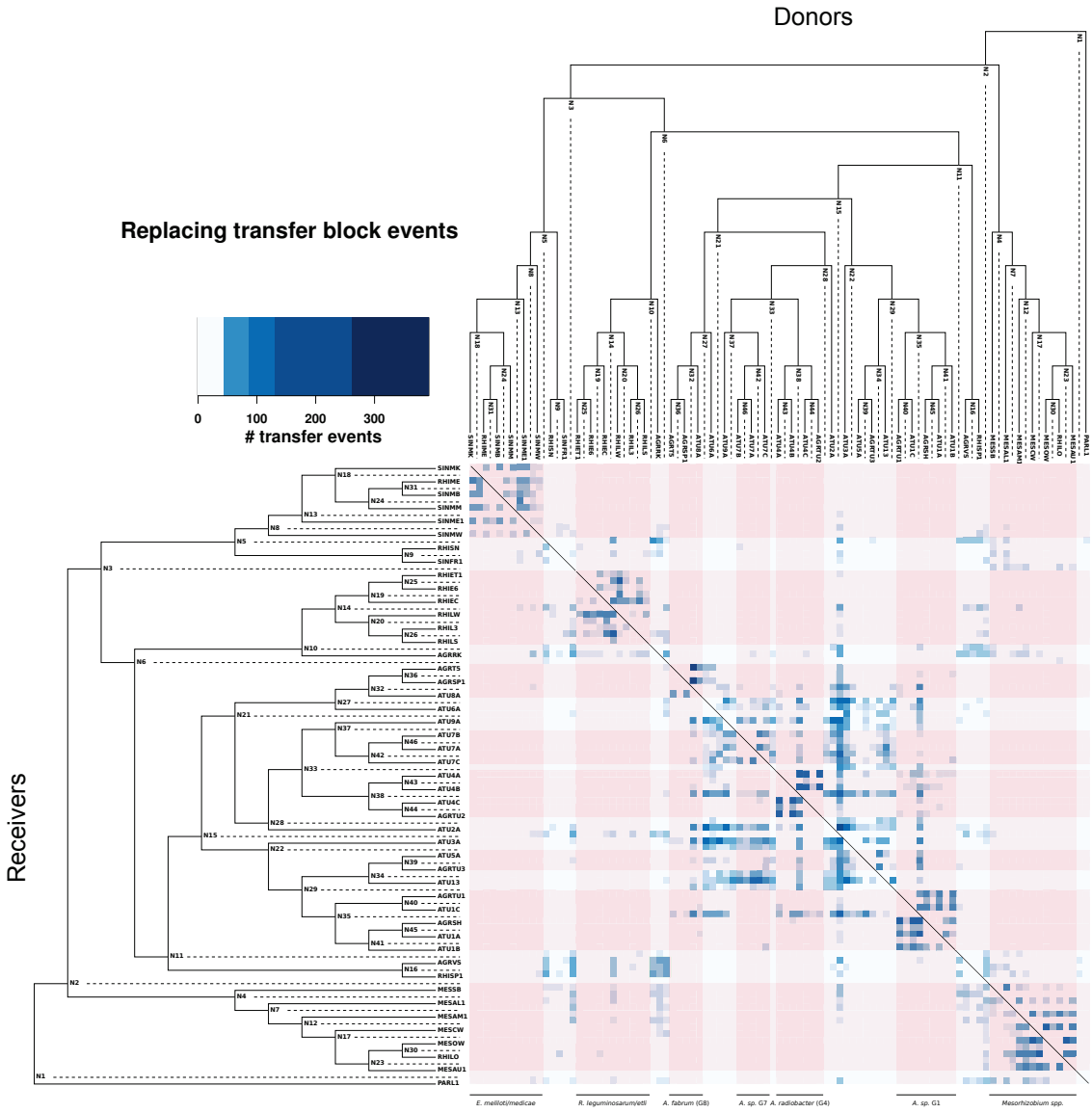


Figure 2.27: **Intensity of replacing gene transfers among Rhizobiales.** Legend as in 2.25.

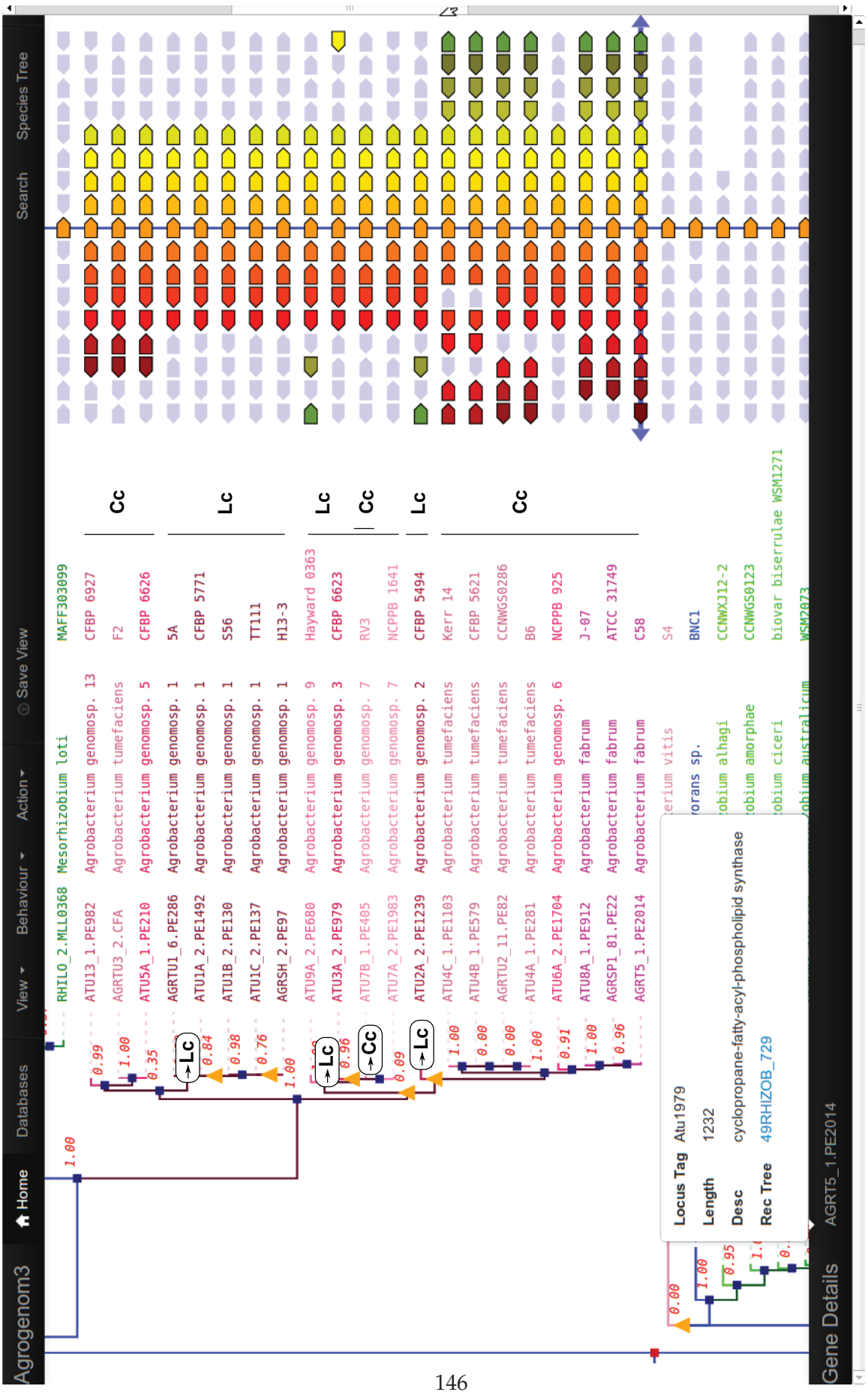


Figure 2.28: **Multiple migration of *cfa* gene between replicons of *A. tumefaciens*.**

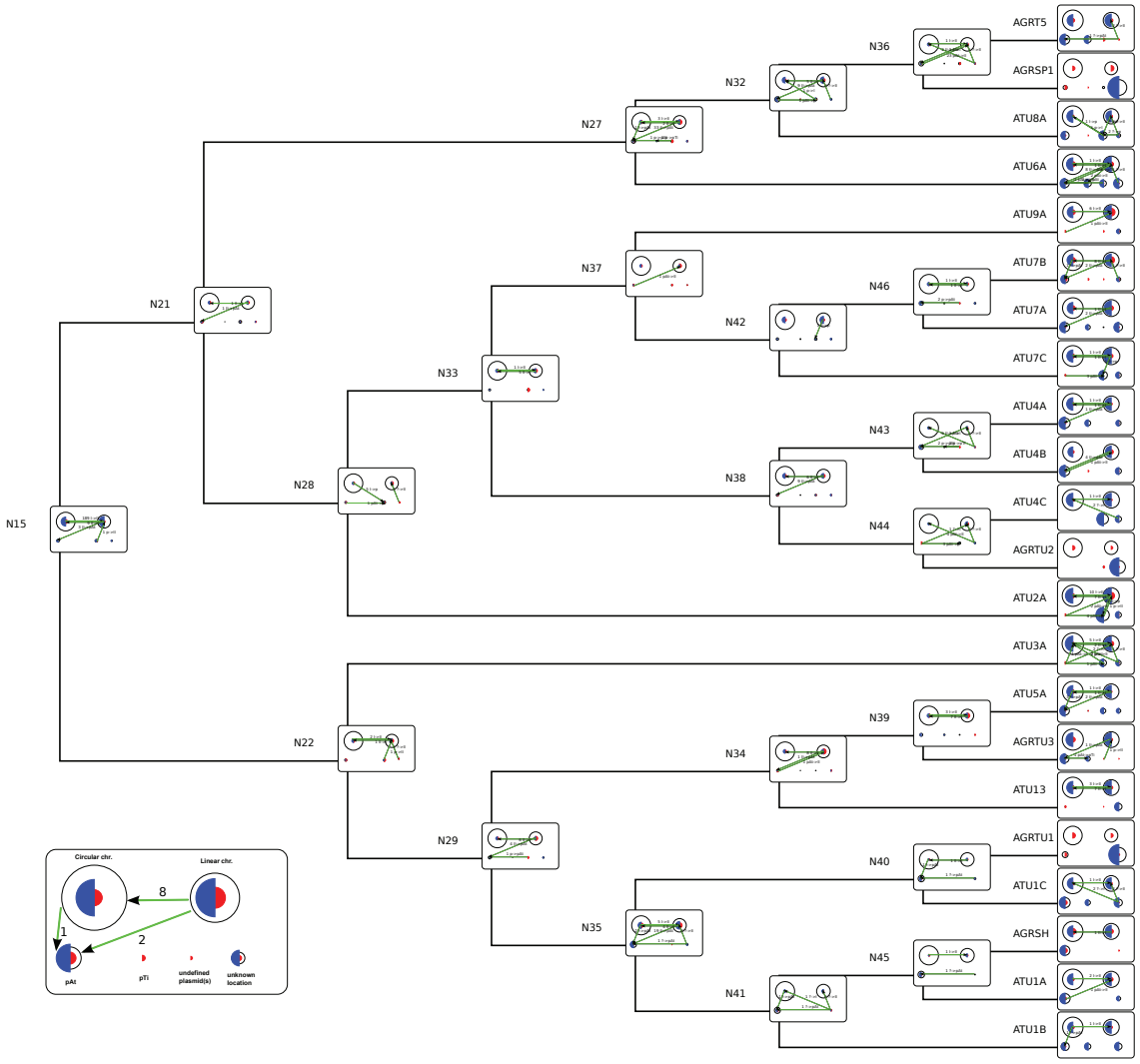


Figure 2.29: **Dynamics of the location of genes on ancestral replicons.**
Genomes are schematically represented at each node of the phylogeny, with up to six genome compartments: the circular chromosome, the linear chromid, the At plasmid, Ti plasmid and undefined plasmid(s), and an unknown location. Those compartments are represented by black-lined disks with a surface which is proportional to their size. Blue and red half-disks indicate the amount of gains and losses on each replicon, respectively (surface proportional to the number of gained/lost genes). Translocations between replicons are indicated by green arrows with the number of translocated genes aside.

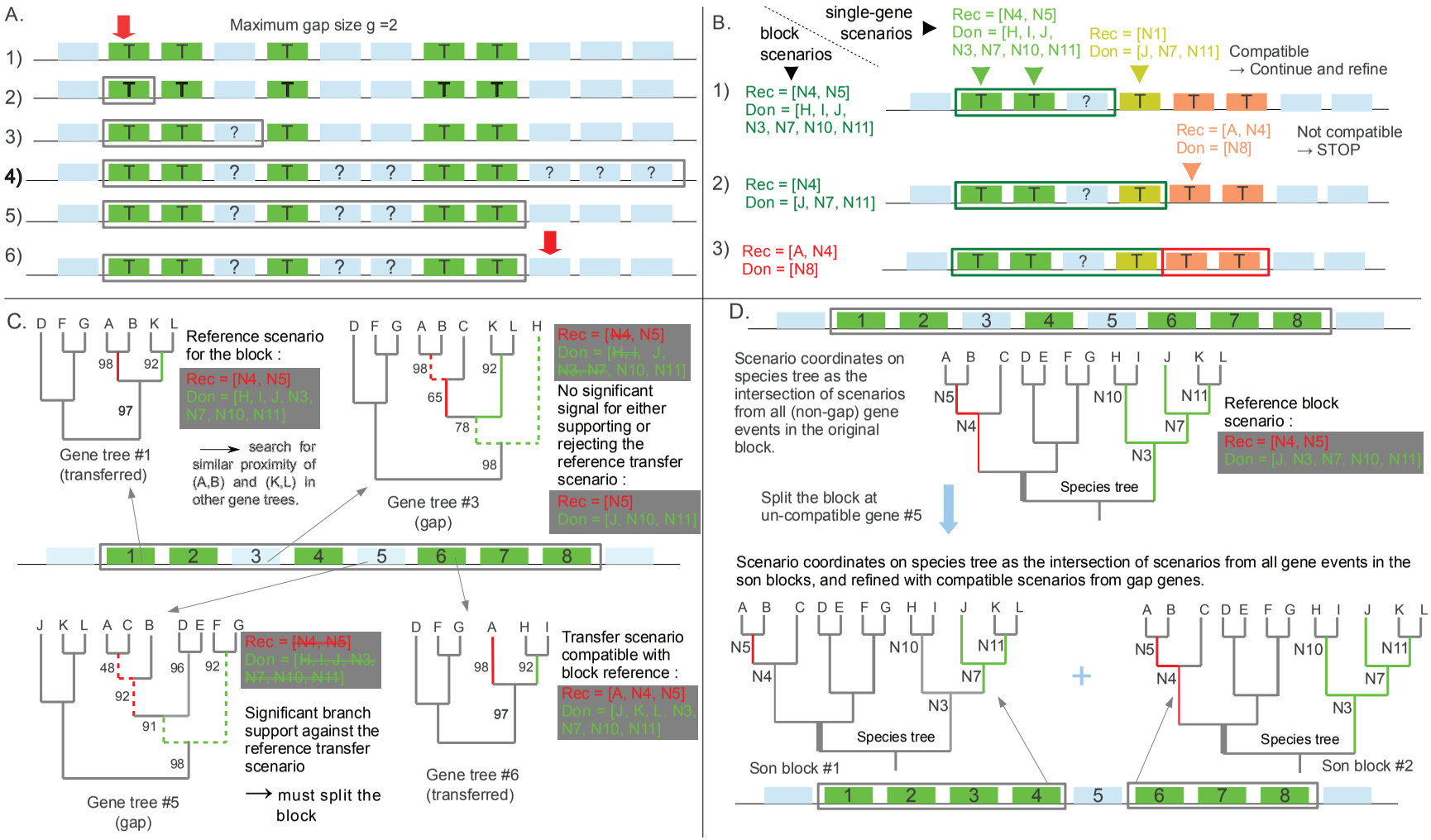


Figure 2.30: **Algorithm for construction of blocks of co-transferred genes.**

A. Course of the extension of a block. A gene detected transferred (green 'T' boxes) is encountered during the course of the replicon (1). The block is initiated (2) and extended to gene having a consistent signal transfer (other 'T' boxes) or no signal of transfer (boxes '?'), In which case a gap is opened (3). When the extension leads to create a gap greater than size $g = 2$, the block is stopped (4), then extremal genes without signal for a transfer are removed from the block (5). The search for block events resumes after the last gene block (6).

B. Concordant or discordant scenarios of transfer. A block of events has an overall scenario corresponding to the intersection of those of its constituent genes. The scenarios are described as sets of possible receiver (Rec) and donor (Don) nodes, as exemplified Sup. Fig. 2.15. The dark green block consists of genes whose transfer scenarios are identical (light green boxes, light green scenario, corresponding to the case presented in Sup. Fig. 2.15) and gap genes without signal: the block has an overall scenario identical to light green genes (1). The block is extended to the brown gene for which the sets of possible receivers and donors are smaller: the global scenario sees its coordinates refined accordingly (2). When continuing the march on the chromosome to extend the dark green block, a pink gene is met; the donor set of the dark green block and that of the pink gene have an empty intersection: the scenarios are considered discordant, and the block is stopped (2); a new block is initiated (3).

C-D. Checking the compatibility of 'gap' genes with the scenario of the block. C. Test if the scenario of the transferred genes (green boxes) is rejected by the 'gap' genes (blue boxes). D. Breaking of an inconsistent block into several blocks and re-computation of their respective scenarios.

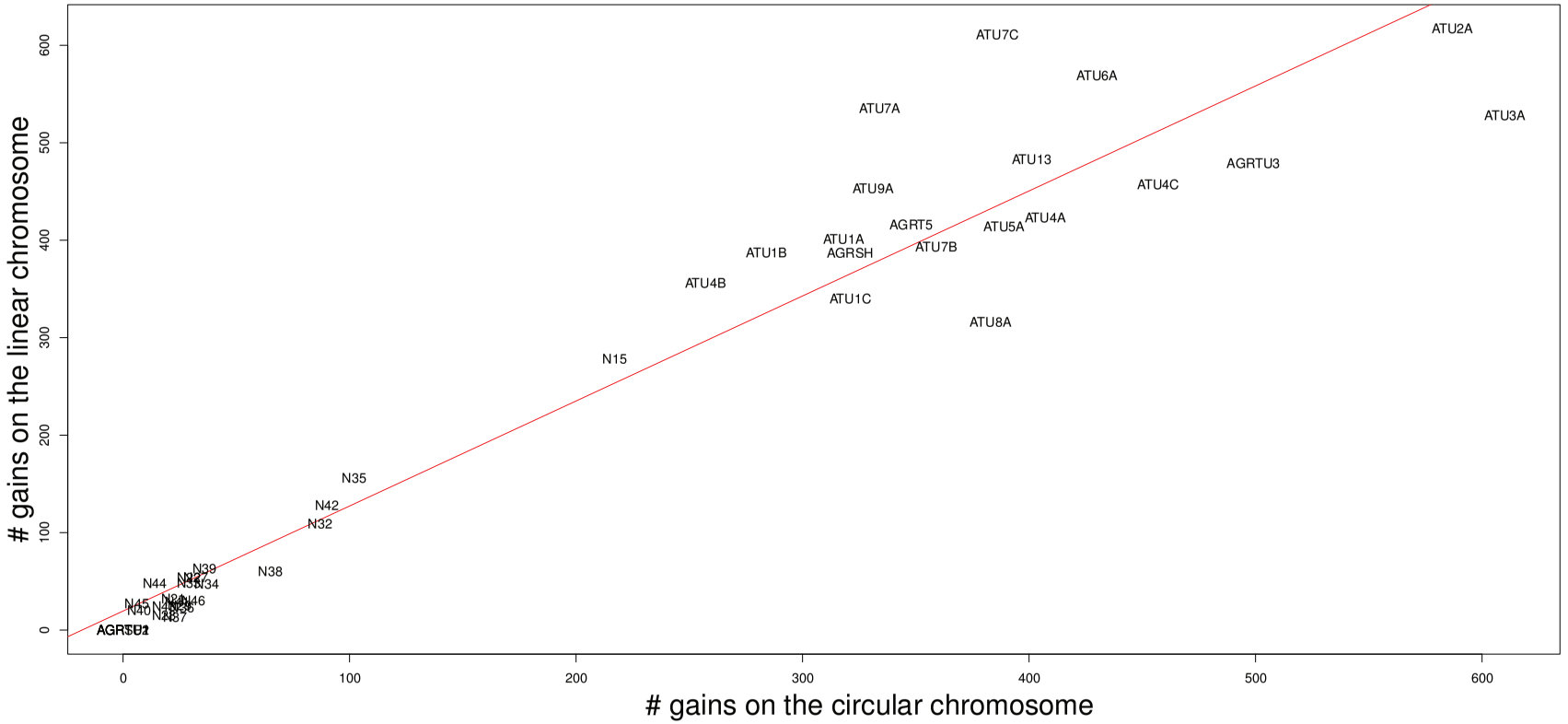


Figure 2.31: **Gene gains on the circular vs. linear chromosomes of *A. tumefaciens* ancestral and extant genomes.**
X-axis represent the number of gene gains located to the primary chromosome, i.e. circular chromosome, Y-axis represent the number of gene gains located to the secondary chromosome, i.e. linear chromid.

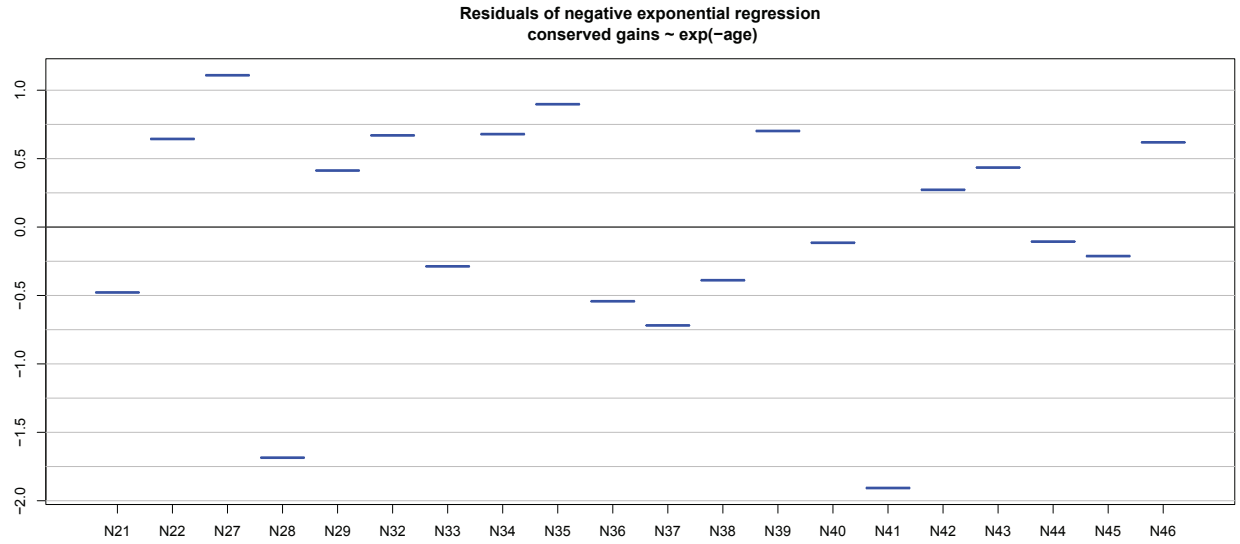


Figure 2.32: **Residuals of negative exponential regression of clade age vs. conservation of gained genes.**
The regression is define as $cg \sim \exp(-26.038 \cdot age + 5.042)$, with the following summary statistics: Cov(log(cg), age) = -0.0269, sd(age) = 0.0321, sd(log(cg)) = 1.165. cg, conserved gains.

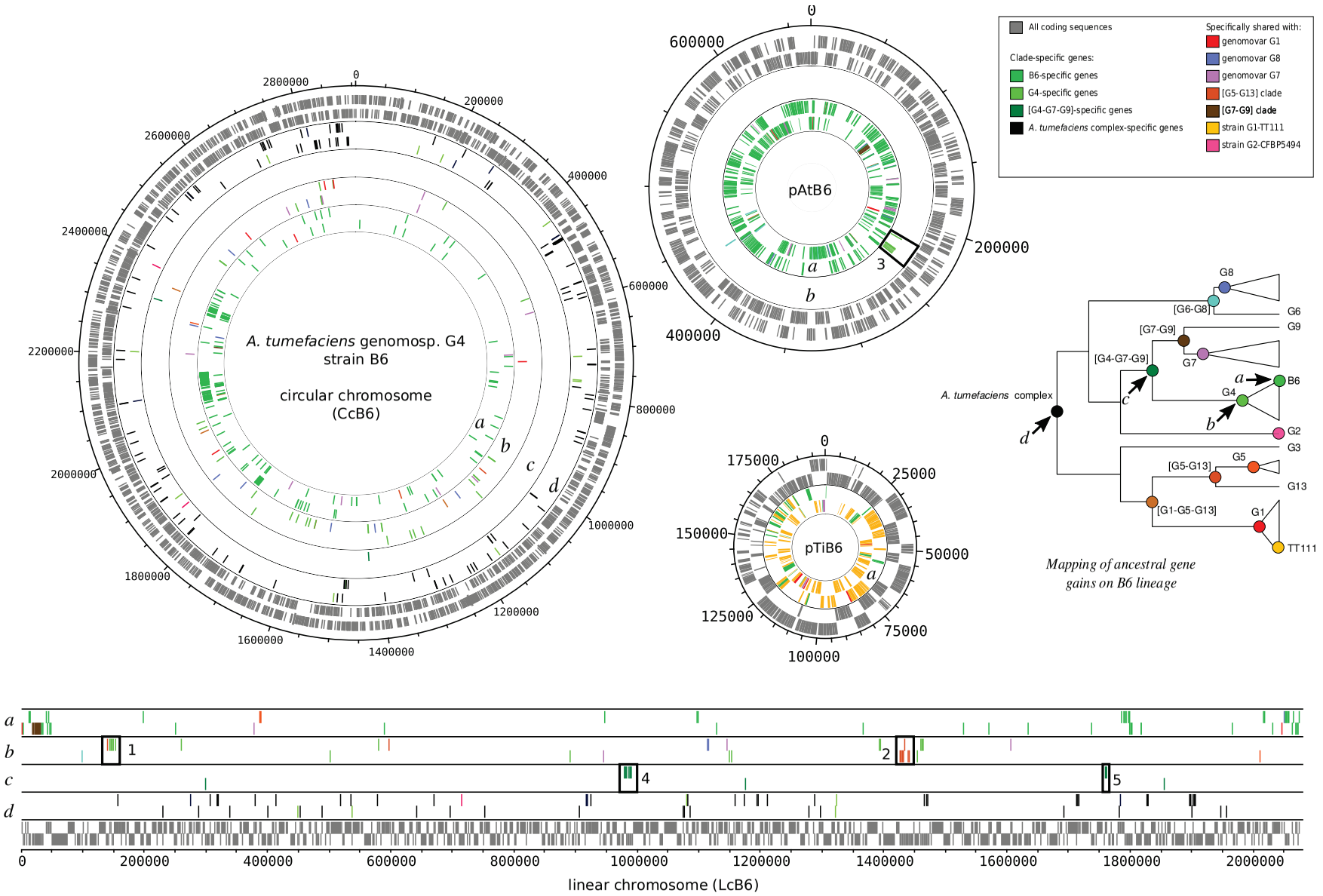


Figure 2.33: **Historical stratification of gains in the lineage of *A. tumefaciens* strain B6.** (legend next page)

Figure 2.33: **Historical stratification of gains in the lineage of *A. tumefaciens* strain B6** (continued). Legend as in Fig. 2.11.
Numbered frames show particular gene clusters within B6 genome: (1-3) G4-specific gene clusters: (1) Atsp37 : aromatic compound (acriflavine) perception, degradation and efflux; (2) Atsp38 : uptake and catabolism of sugars (sorbose, dehydro-fructose) (shared by [G5-G13] clade); (3) Atsp39 : gamma-glutamyl cycle for detoxification of periplasmic compounds; (4-5) [G7-G7-G9]-specific gene clusters: (4) Atsp40 : uptake and degradation of a (sulfated) polygalacturonide/polyglucuronide; (5) Atsp41 : ferrichrome-iron sensing and uptake. Note that the major part of pTiB6 is shared by G1-TT111, as was previously shown (Lassalle et al., 2011), suggesting that pTiB6 and pTiTT111 are related by a recent transfer event.

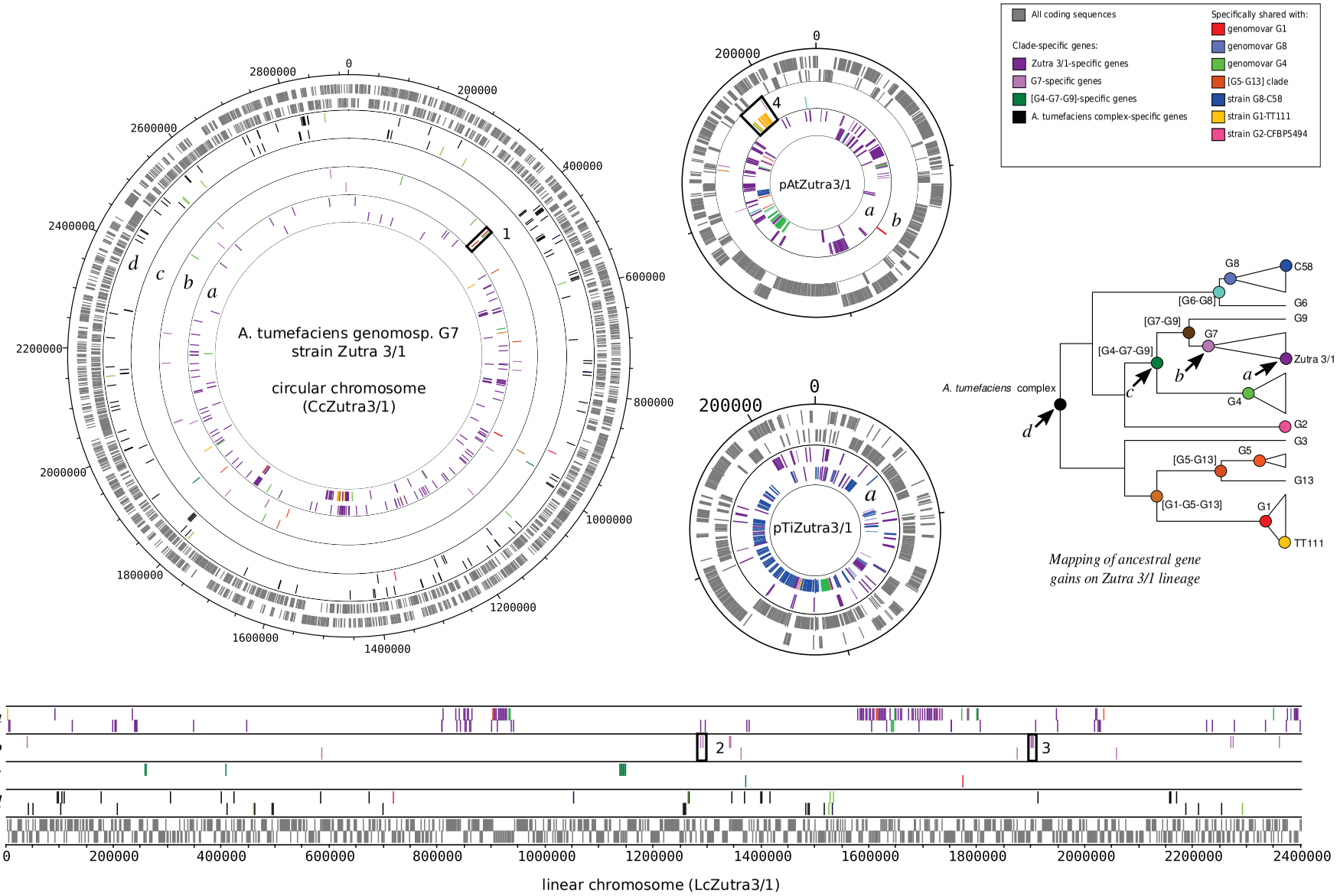


Figure 2.34: **Historical stratification of gains in the lineage of *A. tumefaciens* strain Zutra 3/1.** (legend next page)

Figure 2.34: **Historical stratification of gains in the lineage of *A. tumefaciens* strain Zutra 3/1.** (continued). Legend as in Fig. 2.11.
Numbered frames show particular gene clusters within Zutra 3/1 genome: (1-4) G7-specific gene clusters: (1) Atsp42: Quorum sensing- regulated periplasmic protein-disulfide isomerase DsbABG (ensures proper conformation of disulfide-bonded proteins like c-type cytochromes for covalent haem attachment); (2) AtSp46: prohibitin (DNA synthesis represession), phage shock prot (represses the expression of sigma54-dependent operons); (3) AtSp47: system for sensing (FecIR), TonB-dependent import (FhuACD) and utilization (ViuB) of iron(3+)-hydroxamate siderophore; (4) AtSp48: enzymes and regulators of unknown function.

2.6 Supplementary Tables

Supplementary tables too large for display in manuscript are available at `http://pbil.univ-lyon1.fr/pub/agrogenom/xxxxxxx`.

		To				
		Cc	Lc	pAt	pTi	p
From	Cc	-	0.112 / 0.018	0.0 / 0.009	0.0 / 0.0	0.0 / 0.005
	Lc	0.014 / 0.070	-	0.003 / 0.116	0.001 / 0.003	0.0 / 0.029
	pAt	0.0 / 0.011	0.0 / 0.195	-	0.0 / 0.004	0.0 / 0.084
	pTi	0.0 / 0.0	0.0 / 0.0	0.0 / 0.063	-	0.0 / 0.085
	p	0.011 / 0.019	0.021 / 0.054	0.032 / 0.187	0.011 / 0.170	-

Table 2.6: Proportion of gene translocation among replicons along the history of *A. tumefaciens*, as total counts of translocated genes divided by the mean across *At* history of donor replicon size. Figures on the left indicate translocations happened on the branch leading to *At* common ancestor (code N15), figures on the right indicate translocations happened afterwards.

Table 2.7: Inventory of gene translocation among replicons during the evolution of *A. tumefaciens*. Table available at `http://pbil.univ-lyon1.fr/pub/agrogenom/xxxxxxx`.

Table 2.8: Inventory of genes specific to clades of *A. tumefaciens*. Table available at `http://pbil.univ-lyon1.fr/pub/agrogenom/xxxxxxx`.

2.7 Supplementary Material

2.7.1 Comparison of several hypotheses for the core-genome reference phylogeny

Studies using data from the whole genome of bacteria showed in many case that almost each locus of the genome has a distinct history, i.e. supports a unique tree topology (Abby et al., 2012; Shapiro et al., 2012). Several methods exist to summarize the main phylogenetic signal of a genome-wide dataset, among which the most commonly used is the concatenation of alignments of conserved genes. Such "super-matrix" approaches can recover a majority signal from a large dataset, even if it is hidden by a large amount of noise due to phylogenetic irresolution and HGT. This global approach is powerful, but has a notable drawback: it does not inform about the fraction of genome supporting the tree inferred from the main signal. In fact, the "majority signal" is often just supported by the largest of a collection of many small fractions of the genome. It has been shown that the main tree can be representative of only a handful of gene trees among the hundreds that are reconstructed for individual loci, even though the main tree is strongly supported by high bootstrap or likelihood supports (Abby et al., 2012). This is due to a smoothing of effect of large matrices, because even a slightly higher proportion of a genome corresponds to a difference of thousands of sites, which yields a strong statistical support for the main topology.

The supermatrix approach is then too extremely opposed to the observation of all possibly different gene-specific topologies. To satisfy this trade-off, we used a methodology that allowed us to recover the main signal from genome-wide data, but to put it in perspective with the amount of genes supporting it. Some sophisticated statistical methods exist to both infer a species tree and evaluate its concordance with individual gene trees, but they rely on some prior on the structure of data, such as the limited size of the gene tree space topology (Ané et al., 2007) and use heavy computation that seem hardly scalable to datasets with large number of taxa (ans thus possible tree topologies). We preferred a simpler method without such an a priori, and thus used a jacknife sampling approach.

This consisted in sampling several times relatively small sets of genes for which the alignments were concatenated and trees computed, that reflected the variability of the phylogenetic signal genome-wide, while still being based on a large set of characters. Then the majority-rule consensus of these trees was taken to reflect the main phylogenetic signal of the dataset, while the frequency of its bipartitions in the jacknife sample reflected the genome fraction in support of this main signal.

From the jacknife sampling of 455 unicopy genes in the core of the 47 genomes, we obtained a consensus tree presented Fig. 2.3 that grouped together all named genera and species, exception made of strain *Rhizobium sp*. PD01-76 which groups with *A. vitis* S4 and must belong to a clade related to *Agrobacterium* biovar III. Focusing on *A. tumefaciens*, all genomic species were found monophyletic with high support, and grouping of some genomic species based on *recA* or *telA* marker gene phylogenies (Costechareyre et al. (2010); Shams et al. (2013); Ramirez-Bahena et al., in press, see Annex 5.2.2) were recovered with good support: genomovars G8 and

G6 (hereafter named [G6-G8] clade), G1 with G5 and G13 ([G1-G5-G13] clade), G7 with G9 ([G7-G9] clade), G4 with the latter ([G4-G7-G9] clade). Also in accordance to some studies but not all, the genomovar G2 groups with [G4-G7-G9] clade. Altogether, comparisons of the jacknife core tree to phylogenies previously built with marker genes or obtained from simple concatenation of ribosomal protein genes or core genes (Sup. Fig. 2.18) yield one robust feature: the separation of [G4-G7-G9] and [G1-G5-G13] clades marking a bifurcation at the root of *A. tumefaciens* complex. The remaining groups, genomovar G2 and G3 and [G6-G8] clade, have a labile positioning within *A. tumefaciens*. As we are looking for a reference tree for reconciliation of gene and genome histories, we had to find the topology which was the most representative of the history of the genome, since any deviation from it in a gene tree will constitute an incongruence to be explained by an evolutionary event. Bad reference choice may then yield evolutionary scenarios for genes very far from parsimony. We thus looked for what could cause such an unstable topology and if there were better alternatives.

Comparing all those phylogenies yield one robust feature: the separation of [G4-G7-G9] and [G1-G5-G13] clades marking a bifurcation at the root of *A. tumefaciens* complex. The remaining groups, genomovar G2 and G3 and [G6-G8] clade, are found in different branching combinations. Notably, according to the phylogenies based on the 455 core genes (either all concatenated or jacknife-sampled) the three labile groups are found close to the basis of the two robust groups (Sup. Fig. 2.18 A,C). In, fact, simple concatenation of the 455 unicopy core genes yielded the same topology for *A. tumefaciens* complex as for the tree jacknife-sampled dataset, but with a different rooting within Rhizobiaceae: the root of the concatenate tree id located between [G4-G7-G9] clade and G2 strain in the jacknife tree, making G2 strain and [G6-G8] clade to flip sides (Sup. Fig. 2.18 A,C).

The basal position of genomovar G2 and G3 in these core phylogenies is linked to long branches leading to them (Fig. 2.3), suggesting a possible artifact of long branch attraction. These species have only one representative strain each, and the long branches are likely due to an inaccurate reconstruction of their history in absence of cognate strains. The basal position of these two species and their long branches are observed neither in *recA* and *gyrB* phylogenies with two strains per species (Costechareyre D., PhD thesis; Ramirez-Bahena et al., in press, Annex 5.2.2), nor in that based on the concatenate of slowly evolving ribosomal protein genes (Sup. Fig. 2.18 B), comforting the hypothesis of reconstruction artifact due to insufficient sampling and saturation of phylogenetic signal. The strong jacknife support for G3 grouping with [G1-G5-G13] clade is likely due to the fact that all sampled concatenates contain rapidly evolving genes that lead to systematic saturation of signal.

Alternative topologies found in *recA*, *gyrB* and ribosomal protein gene concatenate phylogenies (Sup. Fig. 2.18 B) share common features: the grouping of G2 strains directly with G9, and the grouping or at least proximity of G3 strains with [G6-G8] clade. These hypotheses of grouping based on slowly evolving markers could be good candidates for settling a reference tree.

To test which hypothesis of grouping seemed the most relevant, we looked in the whole

set of gene phylogenies for congruent patterns of branching. This allows us to exploit the information from the whole genome – as opposed to restricted to unicopy conserved genes – and to make no assumption on the preferred qualities of the genes we use. For this, we used the program TPMS (Bigot et al., 2013) to find subtrees of possibly multicopy gene trees that matched particular patterns, and counted unique genes of the target species that contributed to these potentially overlapping subtrees (see Methods section 2.4.5.8). We searched for the monophyletic grouping of different clades of *A. tumefaciens*, as summarized in Figure 2.19. Branching of G2 with [G4-G7-G9] clade is indeed recognized for only 27% of unique genes, but this is higher than the occurrence of G2 branching directly with G9 (12%). Grouping of G3 with [G1-G5-G13] clade is also quite infrequent (29%) but also slightly higher than the occurrence of the monophyly of G3 and [G6-G8] clade with (26%). These low fractions of trees supporting the patterns could be due to the reduced probability of observing genes of a single strain to have matching the pattern compared to testing a similar hypothesis when several strains from the same species are represented: if a gene from a single strain originally matched the pattern but was recently converted by transfer from a outgroup of *A. tumefaciens*, the pattern would not be observed, however if several strain are available in the dataset, the pattern would be recovered. To test for such an effect, we searched patterns matching well-supported monophylies involving species with a single sample strain. We were able to find that 45% of single G9 strain genes were found supporting its monophyly with G7, 53% of G13 genes supporting grouping with G5, and 71% of G6 genes supporting grouping with G8. In these cases, this method found that half the genome or more recorded the close relatedness of those groups. Interestingly, we found that [G6-G8] clade grouped with [G4-G7-G9] in 71% of cases, though this rely on a small number of testable genes (824) because of the constrain of finding both clades monophyletic in gene trees. This figure is a good argument to keep a reference tree where [G6-G8] and [G4-G7-G9] cades group closely as in jacknife core tree (Sup. Fig. 2.18 A) rather than in its tilted version in the core concatenate tree (Sup. Fig. 2.18 C).

Even though the jacknife core tree is not representative of a majority of genes about some nodes of the history of *A. tumefaciens*, it appears the best compromise in representing the main signal for all internal nodes. We thus chose it as a reference phylogeny for reconciliation of genome and gene tree histories.

2.7.2 Gene tree reconciliations: detailled procedure

(1) Nodes with unicity conflict, i.e. where a species is represented in both children nodes, were listed in each tree. To decide if this unicity conflict originated from duplication or rather additive transfer events, we computed the duplication-induced loss rate (DLR) as the ratio of the number of loss necessary to complete a duplication scenario to the number of ancestral species nodes below the ancestor to which the duplication would map. Manual examination of trees led to choose an arbitrary threshold of DLR = 0.2 above which the duplication hypothesis seems too costly and the additive transfer hypothesis is preferred. Use of a alternate criterion was tried with the duplication consistency score (DC

score) (Vilella et al., 2009), but showed poorer discrimination of duplication or transfer-like patterns, certainly because it was first implemented to assess duplication confidence in a vertebrate genome evolution model not considering transfer events (Vilella et al., 2009). Note that this arbitrary decision is not definitive, as further detection of transfer events can lead to resolution of a certain amount of unicity conflict. Unicopy subsets of leaves were sampled in the full gene tree so that they gather the maximal set of species without putting together paralogs, i.e. gathering only orthologs or co-orthologs (sensu Kristensen et al. (2011)). All possible combination of orthologs were explored, with each leaf being potentially represented in several leaf combination sets. The corresponding subtrees (same topology as the full gene tree, but restricted to an unicopy leaf set) where then searched for transfer events

(2) Transfer were searched using Prunier software (version 2.0, fast algorithm, forward search depth = 2, branch support threshold = 0.9) that detects supported topological incongruence with the reference tree of species (Abby et al., 2010). Unicopy subtrees were mapped back on the full gene tree, associating each node of the unicopy subtree to one or several collapsed nodes of the full gene tree, and transfer events found by Prunier were annotated on a node of the full gene tree. Doing so with every unicopy subtree, a count is made by node of the number of times a node of the full gene tree is tested for transfer (i.e. covered by an unicopy subtree submitted to Prunier ananlysis, Prunier coverage) and of the number of times is is detected as a particular transfer scenario. The latter count gives the support for each transfer scenario. The total Prunier coverage minus the sum of supports for all proposed transfer scenario gives the support for an absence of transfer at this node (support for a speciation or duplication, depending on the further completion of the reconciliation).

(3) To choose between several events proposed at he same node, we used two criteria: first, inclusion in the largest block event, and second, best coverage by Prunier replicate tests.

    1. At this step, a preliminary search for block of co-transferred genes was performed (see section 2.7.3). Several conflicting transfer scenarios could be proposed at the same node, especially when a sparse sample of species in the tested unicopy subtree gave poor context to orientate the transfer. This often result in inferring a scenario of transfer from species A to species B and another scenario of transfer from B to A. Gene families locally co-evolve (for instance events of co-transfer), but most of their history should be independent, giving potentially different patterns of losses in respective unicopy subtrees of neighbouring genes. Although it can be hard to decide on a scenario of a single transfer event, a series of neighbouring genes with compatible scenarios gives good confidence in the shared scenario. Therefore, when several events are considered, the one with the longest block event is chosen.

    2. If no choice can be done considering block events (a majority of transfer events involve only one gene), the event supported by the most tests of Prunier is retained.

This can notably retain an absence of transfer, when a majority of Prunier tests did not infer any transfer event at the node (often by explaining the potential phylogenetic conflict by transfers in another part of the tree).

(4) Transferred subtrees were pruned from the full gene tree to yield a forest of subtrees free of transfer events. Search of unicity conflict was performed again on each tree of this forest. Nodes bearing unicity conflict were explored in a post-order traversal. Under each conflicting node, groups of leaves representing monophyletic group of species and present in multiple copies were searched for phylogenetic incongruence with the species tree using the algorithm for detection of taxonomic perturbation from (Bigot et al., 2013). When taxonomic perturbation was found, a transfer was inferred and the transferred subtree was pruned, potentially resolving the unicity conflict. After each transfer detection, search of unicity conflict was redone on the pruned tree. When this iterative search reached the root of the tree and no more transfer could be inferred, remaining uncity conflict was explained by annotating duplications at conflicting nodes. On one hand, this procedure allowed us to avoid over-annotating duplications in cases of additive transfers where the topological incongruence was not enough supported to be detected by Prunier. On the other hand, applying this transfer detection method regardless of supports could be hazardous if done all the whole gene tree, given the numerous errors in the reconstruction of the gene tree topology; applying it only in case of unicity conflict is thus more conservative, as it considers the signature of additive transfers.

(5) Ancestral content inference was done using either Dollo parsimony, or asymmetric Wagner parsimony using Count software (Csűrös, 2008). In the latter cases, this method was adapted to be applied to orthologous subtrees of a gene tree.

1. The gene trees were first atomised into a forest of orthologous subtrees by systematically pruning the recipient subtree under a transfer or duplication event (choosing arbitrarily the duplication 'recipient' as the smaller subtree). This was done by exploring the node events in a post-order traversal. Each leaf (gene) set of these subtrees were considered as orthologous subfamilies for which the gain/loss history and ancestral presence states were reconstructed independently.
2. Xenologous replacement consist of the entry of a transferred gene in the genome followed by coincidental loss of the resident gene. This can be instantaneous in case of gene conversion by homologous recombination. This kind of transfer does not result in a new gene copy and if it was given by a close relative, the replacing gene is expected to be quite similar in function to the native gene it was substituted to. Therefore, in order to keep subfamilies as (functionally) coherent groups of leaves rather than true orthologous groups, subtrees containing only speciations and *transfers that do not disturb the monophyly of the represented species* were kept as unique subfamilies punctuated of allelic replacements.

3. Having previously detected transfer and duplication events, inferring the ancestral states and events should be straightforward, by inferring a gain at the last common ancestor (LCA) of the represented species, and completing the history with losses (Dollo parsimony). However, subfamilies with patchy distribution of genes in the species tree suggests conspicuous transfer events (not in topological conflict with the species tree) rather than gain at an ancient LCA followed by several losses. Asymmetric Wagner parsimony method correct for this bias by inferring additional transfers that are annotated on the gene tree. Since these transfers do not induce phylogenetic or unicity conflict, the genes in the subfamily are still considered orthologous, though they do not match exactly the definition of a group of leaves only related by speciation events. Again, this allow to keep subfamilies as coherent clusters of genes with probably similar functions – a goal for which orthology is often a proxy.

4. Then, gain/loss histories of all subfamilies are integrated to the global family history. It implies to correct the ancestor to which some gain have been mapped. Indeed, the ancestor in which happened a duplication is the LCA of the merged set of species represented in paralogous children subtrees. This global species set can be larger than each individual orthologous species set, because of independent parallel losses. The ancestor of each orthologous may then be inferred lower in the species tree that it should, what is corrected.

5. Global integration of gain/loss/duplication/transfer reception/emission/allelic replacement history was done across all families to describe the history of the whole genomes.

Procedure steps are implemented in several python scripts : 1: 'find_acestral_duplication.py'; 2: 'rec_to_db.py'; 3.1: 'getblockevents.py' for preliminary block search; 3 and 4: 'refine_transfer_annotations.py'; 5: 'ancestral_content.py'.

2.7.3 Block event reconstruction: algorithms

The leaf blocks are built following a greedy algorithm of systematic exploration of gene histories along replicons:

```
for each gene in a replicon do
    for each event above this gene in its gene tree do
        a block event is initiated with this seed gene event
        while compatible neighbour is found or gap length < max. gap length do
            neighbour gene on the right is explored;
            for each event in the lineage of this neighbour gene do
                if the neighbour event is compatible to the block event then
                    the neighbour is added to the block;
                    the block coordinates are refined (to the intersection of block and
                    neighbour event);
                    stop event exploration;
                end
                if no compatible event found then
                    add the gene to the block as a gap gene
                end
            end
        end
        if no more neighbour to add then
            the last trail of gap genes is trimmed from the block
        end
    end
end
```

Algorithm 1: Construction of leaf blocks

The maximum gap length parameter was set to 0 for duplication, 4 for transfers and originations.

To avoid to gather in blocks events with no true co-evolution relationship, we avoided to build blocks for old duplications (those that are ancestral to our dataset of species), because any gene in a multi-copy family is expected to present such event above it in the gene tree, and neighbourhood of those genes is not likely to reflect their co-evolution. In our datasets, events of duplication located at the ancestors of the whole dataset (N1), of the Rhizobiaceae and the Phyllobacteriaceae (N2) and of the Rhizobiaceae (N3) were discarded for block event reconstruction.

The ancestral blocks are built following the following algorithm:

$d_{events\_ancblocks}$ is a map of gene tree events to the list of ancestral blocks that are linked to them **for** *each leaf block b* **do**

$l_{putative\_ancblocks}$ is a list of putative ancestral blocks to which it could be assigned, given they share gene tree events;

for *each gene tree event e recorded in b* **do**

append $d_{events\_ancblocks}[e]$ to $l_{putative\_ancblocks}$;

end

if $l_{putative\_ancblocks}$ *contains no ancestral block* **then**

create a new one containing the leaf block ;

end here

else if $l_{putative\_ancblocks}$ *contains only one ancestral block a* **then**

if *b is compatible to a* **then**

assign the leaf block to *a*;

end here

else

a part of *b* and leaf block parts of *a* have some common gene tree events : find their compatible and incompatible parts and split them accordingly into two compatible leaf blocks (to be assigned to the same ancestral block) and other incompatible blocks that correspond to independent block events.

else if $l_{putative\_ancblocks}$ *contains* $n \geq 1$ *ancestral blocks* **then**

a_1 is the first ancestral block of $l_{putative\_ancblocks}$ **for** *other ancestral block* a_i *in* $l_{putative\_ancblocks}$, *with* $i \in [2, n]$ **do**

if a_1 *and* a_i *are compatible* **then**

merge them

else

the pair of ancestral blocks are globally incompatible

if *their leaf block parts have no common gene tree events* **then**

they were hooked by events from distinct partitions of *b*: *b* must be an artefactual fusion of independent block events, split it accordingly. 2) their leaf block parts have some common gene tree events : find their compatible and incompatible parts and split them accordingly into two compatible ancestral blocks to be merged and other incompatible blocks that correspond to independent block events.

end

when all inconsistencies of relations between incompatible blocks are resolved (by splitting incompatible parts and merging compatible parts), start again the procedure for assignation of *b*.

end

end

Algorithm 2: Construction of ancestral blocks

2.7.4 Of the complexity of interpreting 'highways' of genes transfers

The shape of the reference phylogeny has thus a strong influence on the patterns of diversification of genomes, but it may also strongly impact the results of the reconstruction. This is notably the case when we infer the location of evolutionary events, and particularly when identifying the donor and recipient of transfer events. Indeed, due to the uneven sampling of lineages and of the various levels of diversity observed in sampled clade, our ability to detect and accurately assign duplication or transfer events to an ancestor was certainly biased. First, the length of the branch leading to a node represents the quantity of evolution from its nearest ancestor, which is related to time, but also to the number of hidden sister lineages that speciated from this branch and went extinct or were not sampled in the study. If a transfer occurred from one of these hidden lineages to one represented in our study, the donor will be mapped to their closest representative in the dataset. One must recall that a node represent the last common ancestor of a clade, but that DTS events assigned to it may in fact have occurred in any older ancestor along the branch or, when tracking the donor of a transfer, it may come from potentially unsampled cousins (Sup Fig. 2.30). Nodes with long branches will then be seen as donors of HGT frequently, because they summarize a long history of preceding ancestors and of many other relatives. As the strength of this 'donor-pooling' effect depends on the diversity and lifespan of unknown sister lineages, it is very difficult to account for it quantitatively. Importantly, this effect is not symmetric, as transfer received by unsampled lineages will not be seen at all; this is well illustrated by the apparent importance of the ancestor of all *A. tumefaciens* as a universal donor.

Second, because some gene families are transferred and lost at very high rate, they occur in genomes with almost no dependency on the reference phylogeny. For this reason, clades that are densely sampled are more likely to have these genes with rapid turn-over represented in at least one of their genomes. It results that the nodes corresponding to ancestors of densely sampled clades are more likely to be seen as donor or recipient of a transfer event involving these mobile genes, just because they happen to be the closest taxon represented in the gene tree. Though the inferred ancestor might have indeed involved as a donor in a transfer event, it may hide many interleaving transfer events from clades which were not represented in the dataset but which would branch closer of the recipient in the gene tree. Because these and other artefacts linked to the shape of the reference tree are mixed in a complex manner, it is impossible to account for their effect without a strong modelling background, for instance using a coalescent framework (Didelot et al., 2010), that the present study does not provides. It is important to stress, however, that the inferences of event locations are not wrong in essence, but that their quantitative analysis can be misleading. The heat maps of transfer presented Figure 2.10 (and Sup. Figs. 2.25, 2.26, 2.27) must then be used cautiously when comparing the frequency of exchanges between different pairs of ancestors, especially when it involves clades with long branches.

		CcTT111	LcTT111	pAtTT111	pTiTT111	subtotal	total
Specific G1	strict	28	37	13	0	78	165
	relaxed*	12	48	26	1	87	
Specifically shared G1+G8	strict	0	33	10	0	43	57
	relaxed*	4	4	6	0	14	
Specifically shared G1+[G6-G8]	strict	0	0	0	0	0	26
	relaxed*	3	17	6	0	26	

Table 2.9: Summary of clade-specific genes in TT111 genome.
*:relaxed G1-specific genes or relaxed specifically shared genes are present in all G1 or G1+X genomes and in addition in some other strain genomes that do not constitute a coherent clade at least as large as a species.

Knowing these biases, we can nonetheless try to interpret the pattern in the map of gene exchange (Fig. 2.25). Apart from exchange along the symmetry axis, which signs sex within coherent populations, we saw a large amount of gen transfers between the *R. leguminosarum etli* group and the genus *Mesorhizobium.*, and to a lesser extant with the *Ensifer* (*Sinorhizobium*) genus, which are all nodule-forming rhizobia but belong to different families (Rhizobiaceae and Phylobacteriaceae). Such pattern had already been described with sequencing of the first genomes of these clades (Young et al., 2006) and is here confirmed as a frequent route of transfers. Indeed, it involves many genes involved in the metabolism of nitrogen fixation and establishment of symbiosis, as well as genes coding the mobility of these functions (data not shown).

2.7.5 Clade-specific genes: insights into the ecological properties of clades

In the following text, we will describe what gene sets are specific to clades of *A. tumefaciens* and present the main functions they encode. Those clade-specific genes are mostly located in relatively large clusters which often encode coherent metabolic or physiological pathways. These gene clusters will be systematically labelled with the prefix AtSp followed by a numeric suffix. This nomenclature aims to name homologous clusters uniquely across the study and transversally among different genomic context (Table 2.5).

2.7.5.1 Genomic synapomorphies of genomovar G1

There are 78 genes present in all genomovar G1 and in no other *A. tumefaciens* strains, and 87 other genes present in all G1 and found in a heterogeneous set of other strains, totalizing 165 G1-specific genes (Sup. Table 2.9). Other genes are found specifically shared with other clades, notable with genomovar G8 (57 genes) and [G6-G8] clade (26 genes) (Sup. Table 2.9). A notable fraction of those specific genes are located in relatively large clusters in which are represented some few recurrent functional categories (Table 2.5). These main functions are detailed below and figure 2.11.

Chemotaxis and phenolic/aromatic compound degradation pathways One main gene cluster occupies a 24-kb locus on the circular chromosome next to a rRNA operon (AtSp2, frame 1 in Fig. 2.11). It contains several oxydases, mono-oxygenases and amino-transferases which predicted functions together seem to participate in the degradation of one – or several diverse - aromatic compounds that may be aminated. This last feature is further supported by the cognate presence of a predicted amino-acid ABC transporter. Lastly, this cluster encodes and a complete chemotaxis operon including several methyl-accepting chemotaxis proteins (MCPs) and histidine kinases (HK) (Wibberg et al., 2011) that must mediate the transduction of signal for the presence of chemoattractant(s) potentially related to the aromatic compounds degraded by linked enzymes. All Rhizobiaceae have one primary chemotaxis regulation operon *che1* and this secondary one, *che2*, is found only in genomovar G1 among *A. tumefaciens*. This nine-gene chemotaxis cluster *che2* is shared by *Rhizobium etli, R. leguminosarum* and *A. vitis*/*R. sp.* PDO1-076 clade, but not the rest of the locus which is rather distantly related to *Mesorhizobium ciceri*. This assemblage of chemotaxis and catabolic genes is thus unique to genomovar G1.

A second large locus of the linear chromosome (AtSp4, frame 2 in Fig. 2.11) is found specific to genomovar G1 when relaxing the stringency of specificity definition. While several other *A. tumefaciens* strains bear the genes of this locus, the pattern of occurrence outside G1 is heterogeneous across the 50-kb locus and the relations of non-G1 and G1 genes in phylogenetic trees are rarely direct, or in a way indicating a transfer from G1 to the other agrobacteria. Among the functions encoded in this locus, there are notably several mentions of isochorismatase hydrolase, and multiple occurrences of monooxygenase, dioxygenase and peroxygenase enzymes that are often involved in ring-cleavage reactions. In particular, the locus encodes a perhydrolase (non-heme chloroperoxidase) and a multimeric aldehyde oxidase whose homologs are documented as non selective on their substrate (Song et al., 2006; Yasuhara et al., 2005). These genes likely code generalist degrading enzymes that may be involved in detoxication of complex aromatic compounds and/or their catabolism for growth. Interestingly, the locus also harbours a sensory diguanylate cyclase/phosphodiesterase which can regulate the motile/sessile behaviour of the cell through control of cytoplasmic c-di-GMP concentration.

Two other loci on the pAt (AtSp7 and AtSp9) also contain several genes that code enzymes involved in downstream degradation of possibly alkylated phenolic compounds.

It happens that some gene among species-specific gene clusters are found present in other isolated strains of *A. tumefaciens*, showing that gene transfer is occurring frequently, but most of times they result in a very partial sharing of biochemical pathways encoded by clade-specific gene clusters, and thus the potential niche-specifying functions were likely not transmitted. However, some species-specific gene clusters are consistently shared by unrelated strains and sometimes species, showing that gene transfer among species may result in propagation of potentially niche-specifying functions. Here are some cases of gene clusters specifically shared by genomovar G1 and other distant species, classified by the general function they confer.

Amino-acid catabolism There are two loci that are specifically shared by all G1 strains and the only sampled strain of genomovar G2 (AtSp3 and AtSp5). They encode two amino-acid ABC transporters and each one is associated to an enzyme gene: an agmatinase, involved in degradation of polyamines, and a monooxygenase, involved in degradation of aromatic compounds. In addition, there are G1-specific genes located further on the linear chromosome that code a tetrameric enzyme resembling a sarcosine oxidase. Together, these features indicate an ability of genomovar G1 to import and degrade specific amino-acids and polyamines.

G1+G8: Exopolysaccharide biosynthesis A 45-kb locus of the linear chromosome of G1 strains (AtSp14) is closely related to a locus found in genomovar G8 (>93% amino-acid identity averaging over 29 proteins shared by H13-3 and C58) and more distantly to one found in Brucellaceae (>51% amino-acid identity averaging over 25 proteins shared by H13-3 and *B. melitensis* 16M), which was identified as encoding a pathway for biosynthesis of a lipo-polysaccharide (LPS) O-antigen (Vizcaíno et al., 2001). Brucellar LPS are extensively described in the literature (reviewed in Cardoso et al. (2006)) because they can confer smooth phenotype which help the pathogen to escape recognition by the immune system of the host. Though, the role of these particular genes in the biosynthetic process of LPS is not documented. Since this locus is absent from *B. abortus* genomes compared to those of *B. melitensis* and *B. suis*, one could relate the occurrence of these genes to the differences in O-substitutions of the core LPS between *B. abortus* and *B. melitensis*, the so-called the A and M antigens (Meikle et al., 1989; Kubler-Kielb and Vinogradov, 2013).

The parsimony approach for ancestral genome reconstruction (see Methods section 2.7.2) inferred AtSp14 locus was gained twice in *A. tumefaciens* history, potentially by transfer from one species to the other, but there are no strong argument to exclude the possibility of the common ancestor of G1 and G8 (the ancestor of all *At*) to have possessed these genes and that they were then lost multiple times. There are conserved parts of this [G1-G8]-specific locus that are not found in Brucellaceae. First, the locus contains genes involved in neo-glucogenesis from storage polysacchraride (like glycogen), that potentially participate in mobilizing the cell's resources to support exoploysaccharide production. This specific sharing of genes by G1 and G8 is also found elsewhere in the genome, and it includes genes with functions that echoes the supply in sugars and the LPS biosynthesis: deoxyribose uptake and assimilation (AtSp17) and exopolysaccharide (curdlan) biosynthesis (AtSp15) – the latter being previously thought to be only found in G8 genomes based on an array-CGH study (Lassalle et al., 2011). Both clusters (AtSp15, AtSp17) are found on the At plasmids (pAt) in G1 strains but on the linear chromosomes in G8 strains.

One particular gene of the AtSp14 locus is specific to [G1-G8] compared to Bucellaceae: AGROH133_10197 (in H13-3 genome) codes a sensory protein-coding gene that has a different structure in G1 and G8. In G8 genomes, this gene encodes a hybrid sensory histidine kinase-response regulator, whose homolog in G1 appears to be fused to another signal transduction protein: a hybrid CheR-CheB methyl-esterase/methyl-transferease. The G1-encoded 'hybrid-

hybrid' signal-sensing protein (HHSS) should be able to sense an environmental (cytoplasmic) signal, to be auto-regulated and to transduce the signal to other regulatory proteins by phosphorylation and by methylation – notably the chemotaxis regulation proteins – and also to bind DNA to regulate transcription of genes. The surrounding LPS locus is likely part of the target regulated genes, either by direct DNA binding or by signal transduction mediated by the two LuxR-like response regulators encoded in the locus, one being located right next to the HHSS gene.

G1+G9: Extra-cellular secretion One feature specifically found in G1 At plasmids (AtSp11) and shared by strain G9-Hayward0363, the unique representative of genomovar G9, is the presence of a complete type 1 secretion system (T1SS). The canonical T1SS components are accompanied by a putative periplasmic hydrolase and a hybrid enzyme whose first moiety is related to MurF and catalyses the last step of murein biosynthesis and the second moiety is related to glycosylases, suggesting that enzymes modifies the peptidoglycan to help the proper inclusion of T1SS complex in it. There is also one uncharacterized protein with repetitive structure suggesting possible adhesion properties, which might constitute the secreted product. This locus may have an important role in secreting biofilm constituents as it has been shown in *Acinetobacter* (Loehfelm et al., 2008) or to establish interaction with a host extra-cellular matrix.

The common regulation of exo-polysaccharide biosynthesis and motility has already been observed in *A. tumefaciens* G8 strain C58 (Xu et al., 2013), in the related organism *Rhizobium* sp. NT-26 (Andres et al. (2013), see annex) and in other bacteria (Barraud et al., 2009; Marchal et al., 2010). This has to do with the switch between motile ans sessile lifestyles, polar flagellum-mediated swimming being opposed to adhesion and formation of biofilm. This transition is in general commanded by the intra-cellular amount of cyclic di-guanylate (c-di-GMP) which can bind the PilZ domain (Schirmer and Jenal, 2009; Hickman and Harwood, 2008; Krasteva et al., 2010) notably found in cellulose synthase subunit A common to all *A. tumefaciens*. This second messenger is regulated by the relative activity of diguanylate cyclases and phosphodieasterases present in the cell, with high c-di-GMP levels leading to a sedentary lifestyle (Kolter and Greenberg, 2006; Xu et al., 2013). In fact, this secondary messenger is much likely involved in regulating local c-di-GMP concentration in compartments of the cell to promote anisomorphous phenotypes like unipolar polysaccharide production (Xu et al., 2013).

In G1, there seems to exist a system of regulation of chemotaxis/EPS production that integrates sensing of more environmental stimuli. Indeed, among the two enzymes active on c-di-GMP coded by G1-specific genes, one (AGROH133_13181) has a sensory PAS domain and is located in the locus AtSp3 that encodes a phenolic compound degradation pathway, suggesting a regulation of c-di-GMP relative to the perception of a substrate for the neighbour gene-coded enzymes.

Altogether, it seems to exists in genomovar G1 a specific regulon integrating regulation of motility, biofilm synthesis and phenolic compound trophism. These cellular processes would be regulated, on one hand through modulation of c-di-GMP concentration and on the other

	Specificity *	CcC58	LcC58	pAtC58	pTiC58	subtotal	total
Specific G8	strict	18	19	21	0	58	101
	relaxed	20	8	15	0	43	
Specific [G6-G8]	strict	8	15	10	0	33	77
	relaxed	7	28	9	0	44	
Specifically shared G8+G1	strict	0	29	3	0	32	45
	relaxed	2	6	5	0	13	

Table 2.10: Summary of clade-specific genes in C58 genome.
*:relaxed G8-specific genes or relaxed specifically shared genes are present in all G8 or G8+X genomes and in addition in some other strain genomes that do not constitute a coherent clade at least as large as a species.

hand, via the activity of a multi-functional protein, HHSS. The role of this hub protein in G1 cell physiology is certainly central and investigating it is certainly of great interest. For instance, targeted deletion mutagenesis of this gene and of regions coding its sensory, DNA-binding or several signal transduction domains would likely have great influence the cell transcritome and interactome and probably yield strong macroscopic phenotypes such as impaired swimming or modified biofilm production. In fact, phenotypic differences have been observed between G1 and G8 strains that were not predicted given the gene occurrence profiles: curdlan production genes and deoxyribose assimilation genes are specifically shared by genomovars G1 and G8, but curdlan production is a specific phenotype of G8 in in rich-medium culture conditions (Lassalle et al., 2011) and degradation of 2-deoxyribose was observed to be specific to G1 strains (Vial L., Bourri M., personal communication). This suggests that from a common set of genes and encoded functions, genomovars G1 and G8 could have different response to their environment through differential regulation of gene expression.

2.7.5.2 Genomic synapomorphies of genomovar G8 and [G6-G8] clade

There are 58 genes present in all genomovar G8 strains and in no other *A. tumefaciens* strains, and 34 other genes present in all G8 and found in a heterogeneous set of other strains, totalizing 101 G8-specific genes (Sup. Table 2.10). Other genes are found specifically shared with other clades, notably with genomovar G1 (Sup. Table 2.10). A notable fraction of those specific genes (50/101) are located in clusters of two to more than thirty contiguous genes in which are represented some few recurrent functions.

Considering genes specific to genomovar G8 and those specific to clade [G6-G8], we mostly found the same genes than in our previous comparative genomics analysis focused on genomovar G8 (Lassalle et al., 2011) and their arrangement in clusters is very similar. For the sake of coherence of denomination throughout this manuscript, these clusters will be re-named with the AtSp nomenclature used above; correspondence with former cluster names from Lassalle et al. (2011) is indicated Table 2.5.

AtSp21, is the largest G8-specific gene cluster located on the circular chromosome. It

contains the operon *braCDEFG*, that encodes an ABC transporter of amino-acids with broad-range specificity, and genes coding enzymes of the ferulic acid degradation pathway, that were recently characterized for their expression and molecular functions (Campillo et al., in prep.). AtSp29 cluster, which is found immediately upstream, is specific to [G6-G8] clade and is dedicated to transport and catabolism of sugars and amino-acids. It includes an enzyme (encoded by Atu1408 in C58 genome) which predicted activity of transformation of L-sorbose into L-iditol echoes the specific ability of G8 strains to degrade L-sorbose (Vial L., Bourri M., personal communications).

On the linear chromosome, a cluster encoding curdlan exopolysaccharide biosynthesis, AtSp15, was formerly thought to be specific to genomovar G8 (Lassalle et al., 2011) but appears specifically shared by the only representative strain of genomovar G2 and all strains of genomovar G1. Several other G8-specific or [G6-G8]-specific gene clusters are found on the linear chromosome (Table 2.5), consistently with previous results (Lassalle et al., 2011). In addition, we found clade-specific gene clusters on the At plasmid of G8 strains: G8-specific gene cluster AtSp28 codes the degradation of xanthine or another related cyclic compound and [G6-G8]-specific gene cluster AtSp32 putatively codes the uptake and degradation of dipeptides that include an aromatic amino-acid.

2.7.5.3 Genomic synapomorphies of [G1-G5-G13]

The large cluster conserved in agrobacteria (AtSp33, Atu4381-Atu4410 in G8-C58 genome) which encodes nitrate respiration (denitrification) pathway, including *nir*, *nor*, *nnr* and *nap* operons was lost in this clade. This gene cluster was parallely lost by strains G9-NCPPB925 and G8-ATCC31749 (for this one this could be an artefact of draft assembly having missed this genomic region, though this is not probable given its size). These strains devoid of the denitrification pathway may be selectively disadvantaged under certain anaerobic conditions. This is not certain however, because other anaerobic respiration pathways are predicted to be conserved in all *A. tumefaciens*, notably the fumarate respiration pathway.

2.7.5.4 Genomic synapomorphies of the *A. tumefaciens* complex

There are 120 genes present exclusively in all *A. tumefaciens* strains, and 165 genes that are also found present in at most two other distant Rhizobiales, i.e. not directly related like *A. vitis*. It appear there are no large group of genes encoding concerted functions, and neither there are clusters of *At*-specific genes larger than an operon of six genes. They are distributed among diverse functional categories that span the ones represented in the ancestral genome (no specific enrichment in GO terms, data not shown), among which some are redundantly found:

Central metabolism Gain of a methylenetetrahydrofolate reductase MetF, needed to regenerate tetrahydrofolate, related to that specifically present in G8 (AtSp21) (G8 has 2 copies).

Non-homologous gene displacements enriching the functional repertoire Loss of a NADPH-dependent glutamate synthase (49RHIZOB_2054) is concomitant to gain of one of another type (gltB1, 49RHIZOB_3984) together with a sulphite reductase (cysJ) involved in cysteine biosynthesis, both putatively associated to Calvin cycle anabolism.

Phopho-gycerate mutase is an essential enzyme of the glycolysis coded by two genes in Rhizobiaceae, *gpmA* and *gpmB*. In *A. tumefaciens*, *gpmA* is lost but functionally replaced by non-homologous gene of rhodobacterial origin, *gpmI* which codes an iso-enzyme. GmpA and GmpB belong to the family of phopho-gycerate mutase needing 2,3-bis-phospho-glycerate as a cofactor and show high specificity to their substrate mono-phopho-glycerate. GmpI family enzymes do not need a cofactor and catalyse less efficiently the interconversion of phopho-glycerate. They are however more promiscuous in the substrate they can phosphorylate (Rgden, 2002), and could therefore provide new kind of phosphorylated compounds, and notably sugars that are integrated to complex polysaccharides, or membrane lipids. This can be linked to the presence of an extra diverged copy of glpD, coding for glycerol-3-phosphate dehydrogenase which is involed in biosynthesis of phospholipids.

Cell wall and outer membrane Gain of cell wall teichoic acid biosynthesis enzyme (*tagA*, Atu0587) and two succinoglycan biosynthesis transporters (Atu0588 and *mdoC*, Atu3522). Gain of putative scaffold protein for murein sacculus septation (MipA). Presence of specific succino-glyclan EPS biosynthesis genes: *exoU* and a specific extra copy of *exoQ*.

Informational processes Gain of two rRNA modification enzymes (RsmJ and RluA). Loss of a second copy of rpoH. Gain of protelomerase TelA. Complex history of DNA polymerase III subunit alpha gene family: the copy of *dnaE* shared by all *A. tumefaciens* and mostly borne on the linear chromosome is of distant origin compared to other Rhizobiales, and many strains bear one or several extra copies of more distant type which is more prone to transfer and most often plasmid-borne.

Sensing multiple gains of two-component systems, dyguanylate cyclases/esterases and methyl-accepting proteins coupled to sensor domain

carbohydrate metabolism:

many transporter and catabolic enzymes for amino-acids/poly-amines.

Iron metabolism One free iron transporter (FbpAB), 2 iron-siderophore ABC transporter, including iron-hydroxamate transporter FhuABCD used for uptake of iron-complexed agrobactin, an iron storage protein (Dps), an heme oxygenase and precorin hydrolase (ChiG) involed in heme degradation/transformation.

Genes for biosynthesis of secondary metabolites, including phosphopantetheinyl transferase (*agbD*) catalyzing the fixation of phosphopantethein prosthetic group to catalytic modules of NRPS/PKS mega-enzymes, among which the ones responsible for biosynthesis of

agrobactin and G6-G8-specific siderophore.

Detoxification Two P-type ATPases for extrusion of copper and heavy metals, respectively, putative nickel ABC transporters and a MFS permease of unknown specificity.

2.7.6 Selected cases of large transfer events

An integrative and conjugative element (ICE) was transferred between strains G4-Kerr14 and G7-Zutra 3/1. The circular chromosome of Kerr14 and the linear chromosome of Zutra 3/1 are found almost 100% identical over 125 kb (coordinates on CcKerr14: from 2,490kb to 2,615kb). This shared locus is unique to these two strains among *A. tumefaciens*. It contains genes encoding a complete type IV secretion system, partition enzymes and a resolvase-type recombinase at an extremity sign the nature of this locus as an ICE. Right next to the recombinase gene is located a tRNA gene and a rRNA operon, which must have been the substrate of the recombinase for the integration of the ICE.

It bears cargo genes, among which several catabolic genes involved in the uptake and degradation of several compounds, among others: phenolics and degradation products (benzaldehyde, benzoylformate, toluate, catechol, muconate, mandelate), nitrilated compounds (acrylonitrile), sugars and derivated polyols (xylulose, tagatose, sorbitol, xylitol, ribitol). There are also genes for a lipoprotein and an outer membrane protein of RopB family.

A large block event highlighted the relatedness of 70-kb plasmid pIV from strain G6-NCPPB925 and p79 from *A. vitis* strain S4, which appears to be circularized ICEs, as indicated by the presence of resolvase genes. The common ancestor of these elements must be quite ancient, since their average nucleotide similarity is of 75%, and because most genes they share are associated to their mobility while the metabolic genes they harbor are different.

A 12-kb mobile element bearing mercuric resistance genes was transferred between G3-CFBP6623 and G5-6626 very recently (genes almost 100% identical).

A Mu-like prophage was transferred several times between strains of *A. tumefaciens* and related strains S4 and PDO1-076, with variable sizes of conserved parts.

2.7.7 Bioinformatic scripts, modules and libraries

All custom scripts and Python modules/libraries developed in this work are available at `http://pbil.univ-lyon1.fr/pub/lassalle/lib/`.

2.8 Comparative genomics of *A. tumefaciens*: Synthesis and perspecitves

2.8.1 Comparison of outcomes from two different studies of the pangenome of *A. tumefaciens*

In the study presented section 2.3.2, we used microarray comparative genome hybridization (microarray-CGH) to seek the occurrence of close homologs of G8 strain C58 genes in 25 other strains of the *Agrobacterium tumefaciens* (*At*) species complex, (Lassalle et al., 2011). We could define sets of genes specific to inclusive groups on the C58 lineage, among which genomovar G8. In fact, this study was designed to focus on specificities of G8 genomes, therefore the dataset was biased toward a dense sampling of G8 strains (seven G8 strains representative of the known diversity of the species vs. 2 strains per other genomic species), yielding good ability to identify genes conserved in the whole G8 genomovar. However, poor sampling of other species, added to the intrinsic loss of sensitivity of microarrays with genetic distance did not allowed to detect systematically the occurrence of close homologs in genomes of other genomovars. For this reason, the contrast of presence in G8 vs. absence in other *At* may have been poor, possibly leading to false positives in the designation of G8-specific genes. In addition, we found that relaxing the specificity criterion to recruit genes which occurred rarely in strains of other species permitted to define (almost) G8-specific genomic regions with coherent functions. We hypothesized that genes in these regions had a common history since their acquisition in the ancestor of genomovar G8, but that later transfer of individual or short blocks of genes blurred the specificity profile. However, we had no mean to test these hypotheses of evolutionary scenario, apart from codon usage signatures that revealed the older integration in the genome of genes that indeed were likely acquired by the older common ancestor of [G6-G8] clade. In the work presented section 2.4, we compared genome sequences of 22 strains of *At* and focused on variations of their gene repertoires across the history of the taxon. The sampling was done around multiple foci of diversity, with five genomic species for which several strains were sampled (2 for genomovar G5, 3 for G8 and G7, 4 for G4 and 5 for G1), favouring the sensitive detection of rare genes in these lineages, and at the same time providing more confidence in designating species core genomes. We took a particular care to recognize when genes were gained or lost in lineages, and related this information with the patterns of conservation in extant genomes. This led to the identification of clade-specific presence or absence of genes, that constituted genomic synapomorphies of clades. This definition of clade-specific genes was notoriously different from the previous one in the way we treated heterogeneous occurrence profiles. Rather than ignoring those complex patterns or using a threshold of relaxed specificity, we conceived an automated method for recognition of coherent occurrence in clades that contrasted with sister clades. This notably allowed to spot parallel gains/losses of homologous genes in distant clades.

These studies were based on different datasets and used different methodology, so one could have expected quite different results. That is not the case, as the G8-specific genes defined in

the present phylogeny-driven genome comparison are for the large majority the same than those defined previously using microarray-based comparative genomics (Lassalle et al., 2011), showing the robustness of our biological findings. We notably could recover in the genome sequence-based study the 'relaxed' G8-specific genes defined in the microarray-based study, but in this case we could explicitly model the history that led to their heterogeneous presence in the taxon, helping us to understand the underlying process of cladogenesis in relation to ecological adaptation.

2.8.2 A more complete model for the diversification of *A. tumefaciens*

In the first study comparing genomes of *At*, we observed a set of genes specific to genomovar G8 and the [G6-G8] clade that were providing these taxa with unique phenotypic traits, and probably specific ecological abilities (Lassalle et al., 2011) (section 2.4). We concluded that these clades were the progeny of an ecovar ancestor that acquired those niche-specifying traits.

However, G8-specific genes were located in C58 genome in several gene islands, suggesting the G8-specific gene set was acquired in several times, at least as many as gene clusters. This pattern does not fit properly the model of ecotypes as defined by Cohan (2002a), because formation of an ecotype involves one single adaptive mutation. In fact, this is purely rhetorical, because the ecotype model describes the ecological differentiation of lineages that is happening in the smallest differential of time, and the evolution of a lineage as that of genomovar G8 is certainly the sum of several successive ecotype speciations. Of these many ecological speciation event, the sole that is evident is that which led to the split of lineages that are still presented today (or at least in our sample). Indeed, the many sister lineages of the successful probably went extinct, and the absence of current trace of their branching point leads us to confound the successive common ancestors of G8 lineage with their last common ancestor. This highlights the interest of speciation models which have a more historical perspective, like that of 'fragmented speciation' (Retchless and Lawrence, 2007).

We indeed had the intuition of the asynchrony of G8-specific gene acquisition, because the several G8specific gene clusters had different codon usage signatures, ranging from that resembling the *At* core to that resembling C58-specific genes (Lassalle et al., 2011). We concluded that core-like G8-specific gene clusters SpG8-1, SpG8-4, and SpG8-7 (AtSp29+AtSp21, AtSp37 and AtSp26, respectively, in section 2.5) were likely genes ancestral to the whole *At* clade and lost in other lineages. The reconstruction of ancestral genomes could have validated this hypothesis, notably by linking these gene clusters to distant homologs found in other *At* genomes, that we could not have spotted by CGH. This is however not the case, indicating that all G8-specific and [G6-G8]-specific gene clusters were indeed acquired in the respective clade ancestors.

The diversity of codon usage adaptation seen among them (Fig. 4 in Lassalle et al. (2011)) may rather reflect their age in this lineage, as suggested by Retchless and Lawrence (2007), with G8-specific genes with core-like codon usage being probably the oldest acquisitions of the G8 lineage. This ordination in time of the events of acquisitions of adaptive genes might tell us what was the primary ecological adaptations that occurred in these lineage. In the

case of [G6-G8] clade, gene clusters AtSp29 seem the oldest acquisition, followed by cluster AtSp31. Interestingly, both encode proteins involved in uptake and metabolism of sugars and other carbohydrates, which may correspond to a first seminal adaptation followed by a second enforcing the same trait. In G8 clade, cluster AtSp21, which encodes the degradation pathway of para-hydroxycinnamic acids, may preclude the acquisition of other clusters such as AtSp26, which encodes a two-component system for sensing of phenolic compound. These sequences of gene gains are defining adaptive paths toward a specific ecology.

Lastly, it is remarkable that both gene clusters AtSp29 and AtSp21, which gains are seen as the first ecological adaptation of [G6-G8] and G8 clades, respectively, are located next to each other near the terminus of the circular chromosome of G8 strains, while all other younger acquisitions are located on the linear chromid. In the light of the higher plasticity of the linear chromid discussed section 2.4.2.7 and of its probable higher evolutionary rates (Cooper et al., 2010), this indicates that these genes coding the original adaptation, had not only the time to ameliorate their codon usage towards the average of the core genome, but also to settle at a stably evolving locus of the genome.

2.8.3 Perspectives

The variations of codon usage of genes of the same "age" may also be the result of positive selection for a codon usage better adapted to their level of expression. Similarly, the excessive divergence of orthologs shared by species living in sympatry (see section 2.4.3.7), may be the result of positive selection for functional differentiation of the enzymes product and incidentally of the EPS produced by enzymes to mediate different interaction with the environment. Such selective pressures on the sequences of genes and proteins were not explored in the present work, but may be the source of many insights. Actually, we are currently developing methods for analysis of the pattern of substitutions in gene trees in a way that allows the exploration of our database of gene trees without *a priori* on the hypotheses to be tested. Indeed, usual means for research of purifying/positive selection on coding sequences involve test of a hypothesis on the nature of the selection regime occurring on a particular branch/lineage/site of the alignment. However, we do not have – and do not want to have – any *a priori* on which particular part of a gene history was the place of selection, and the test of all (numerous) hypotheses poses the problem of multiple testing revealing mostly false positives. For this reason, we explore the possibility to first *document* the patterns of substitutions using the substitution mapping method developed by Romiguier et al. (2012) and only then to try to test the presence of selection on subsets of gene trees/branches in relation their reconciled history. Other integrated methods for detection of functional differentiation (Caffrey et al., 2012) also appear promising.

A point that was particularly developed in the present work is the modelling of co-evolution of genes, as we recognized genes that were ancestrally linked during horizontal transfer and duplication events. This approach does not not model the evolution of gene order, which can prove interesting in understanding the functional organization of ancestral genomes, as well as

their replicon structure. Several programs were developped to do so and assayed on datasets of complex bacterial genomes such as *Burkholderia* (Jones et al., 2012) and the Rhizobiaceae (Yang et al., 2012). Methods were proposed to perform this gene order reconstruction on reconciled gene tree collections (Bérard et al., 2012), thus linking the gene adjacency information to the evolutionary event. An improvement of such methods could be done using the present results, as the gene neighborhood is an information that can be used *per se* to reconcile gene trees.

As a final perspective, moving to population genomics would add great power to the approach of reconstructing ancestral genomes and the processes driving their evolution. As the new sequencing era allows the facilitated and cheap sequencing of any new isolate, the integration of these data into population genetics (coalescent) frameworks might allow to describe more completely the evolution of genes and genome. Notably, the presence of genes maintained at low frequency in populations can reveal unforeseen selective pressures for accessory ecologies (Coleman and Chisholm, 2010).

3

GC-biased gene conversion shapes the bacterial genomic landscape

The following section corresponds to a manuscript to be submitted soon to a peer-reviewed journal.

3.1 GC-content evolution in bacterial genomes: the biased gene conversion hypothesis expands

GC-content evolution in bacterial genomes: the biased gene conversion hypothesis expands.

Authors and Affiliations

Florent Lassalle[1,2], Séverine Périan[1], Thomas Bataillon[3], Xavier Nesme[2], Laurent Duret[1] and Vincent Daubin[1].

[1]Université de Lyon; Université Lyon 1; CNRS; Laboratoire de Biométrie et Biologie Evolutive, UMR 5558, Villeurbanne, France

[2]Université de Lyon; Université Lyon 1; CNRS; INRA; Laboratoire Ecologie Microbienne Lyon, UMR 5557, USC 1193, Villeurbanne, France

[3]Aarhus University, Bioinformatics Research Center, Århus C, Denmark

Abstract

The characterization of functional elements in genomes relies on the identification of the footprints of natural selection. In this quest, taking into account neutral evolutionary processes such as mutation and genetic drift is crucial because these forces can generate patterns that may obscure or mimic signatures of selection. In mammals, and probably in many eukaryotes, another such confounding factor, called GC-Biased Gene Conversion (gBGC) has been documented. This mechanism reproduces the patterns expected under selection for higher GC-content, specifically in highly recombining genomic regions. Recent results have suggested that a mysterious selective force favouring higher GC-content exists in Bacteria but the possibility that it could merely be gBGC has been excluded. Here, we show that gBGC is probably at work in most if not all bacterial species. First we find a consistent positive relationship between the GC-content of a gene and evidence of intra-genic recombination events throughout a broad spectrum of bacterial clades. Second, we show that the evolutionary force responsible for this pattern also conflicts with selection for optimal codons, specifically for AU-ending codons. We propose that gBGC, first thought to be specific

to sexual Eukaryotes, exists in Bacteria and could therefore be an ancestral feature of cellular organisms. We argue that if gBGC occurs in bacteria, it can account for many previously unexplained observations, such as the apparent non-equilibrium of base substitution patterns, the heterogeneity of gene composition within bacterial genomes, and the strong positive correlation between genome size and GC-content. Because gBGC produces patterns similar to positive selection, it is essential to take this process into account when studying which evolutionary forces are acting on genes in Bacteria.

Author Summary

Classical population genetics models indicate that the efficiency of selection, and hence adaptation, depends on a number of non-selective factors, such as the size of a population or the intensity of mutation. In the last 10 years, evidence has accumulated that another mechanism called Biased Gene Conversion (BGC) can interfere with selection and even mimic its effects. This phenomenon, which arises from a particularity of the recombination machinery, was first thought to be restricted to sexual eukaryotic organisms. Here, we show that this mechanism probably exists in Bacteria and has a strong impact on the evolution of genomes. This discovery not only explains many previously unconnected features of bacterial genome evolution, but also highlights the importance of non-adaptive evolutionary processes in Bacteria.

Blurb

Does biased gene conversion, a process previously thought to be restricted to eukaryotes, have an unforeseen role in bacterial genome evolution?

Introduction

Comparative genomics is a fundamental key to the inner workings of genomes. The identification of genes and other functional elements such as regulatory regions, as well as the understanding of their influence on the fitness of organisms rely essentially on the detection of signatures of natural selection within genomes [1]. In that respect, devising a model of sequence evolution in the absence of selective constraints (aka a neutral model) is critical for the detection of functional sequences. Indeed, to explain the features of a given genomic segment, comparing the fit of a neutral model to a model that also invokes of selection (either purifying or positive) is the operational way to infer evolutionary constraint and hence function.

The base composition of genomic sequences varies widely, both across species and along chromosomes [2,3]). For instance, the genomic GC-content of cellular organisms ranges from 13% to about 75% [4,5], with vast intra-genomic heterogeneity. These large-scale variations in base composition affect all parts of genomes, intergenic regions and genes – including all three codon positions [6] - and hence cannot be simply explained by selective constraints on the encoded proteins. Determining the underlying causes (selective or neutral) of these variations in GC-content is a major issue in genetics: if they result from selection, it implies that the genomic base composition *per se* is an important trait that contributes to the fitness of organisms; conversely, if these "genomic landscapes" are largely shaped by non-adaptive molecular processes, then characterising these processes is essential for the reliable detection of selection (see e.g. [7]).

In mammals, the analysis of polymorphism data and substitution patterns along genomes demonstrated that the evolution of GC-content is driven by recombination, which tends to increase the probability of fixation of AT->GC mutations [8,9]. The impact of recombination on base composition in these genomes is most probably due to a phenomenon known as GC-biased gene conversion (gBGC), which favours G/C nucleotides at polymorphic sites in the conversion of intermediates of recombination (see review in [10]). Although gBGC as a process is unrelated to natural selection, it affects the probability of fixation of alleles in patterns similar to selection [11]. It has been shown to be an important confounding factor, which can mimic some marks of positive selection [7,12] and interfere with selection by actively promoting the fixation of deleterious alleles [13,14]. The process of gBGC has been observed directly in yeast meiosis products [15], and there is ample evidence, based on the analysis of relationships between recombination rate and substitution patterns within genomes, that this process affects many other eukaryotes [16–18]

In Bacteria and Archaea, several environmental factors potentially affecting genomic GC-content have been proposed (such as the availability of oxygen or nitrogen in the environment, growth

temperature, or the variety of environments encountered by an organism, see for instance [19] and ref. therein). Because these effects are weak and the nature of the selective pressures remain elusive, the major force driving genomic GC has long been considered to be mutational bias [20]. Recently however, two independent analyses have shown that in virtually all Bacteria, independently of their genomic GC, there is an excess of G/C->A/T mutations [21,22]. This suggests that an unknown process, selective or neutral, is opposing this universal mutational bias by favouring the fixation of G/C alleles. Hildebrand et al. [22] observed that this bias was still present after removing datasets with evidence of recombination. Moreover they found no correlation between GC-content and recombination rate across bacterial species. They therefore concluded that this force could not be gBGC and hence that selection was driving an increase of genomic GC in Bacteria. The nature of this selective advantage remains however mysterious.

Here we argue that the analyses performed by Hildebrand et al. [22] are not conclusive regarding the gBGC hypothesis, and we present evidence that variations in GC-content observed in Bacteria are influenced by gBGC. One pervasive signature of gBGC is that genomic regions undergoing high recombination rates will have also acquire a high GC-content [6]. We thus studied the relationship between recombination and GC content in 20 groups of Bacteria and one group of Archaea. This dataset covers a wide range of clades representative of the bacterial diversity. To avoid problems inherent to comparisons of recombination rates among species (such as differences in polymorphism, genome samples, population size, mutation rates, an other life history factors), we examined the intragenomic variability for both recombination and GC-content.

We show that in a wide variety of bacterial species, genes with evidence of recombination have a higher GC-content. We further show that this bias towards GC nucleotides in recombining genes interferes with selection on codon usage: while recombination favours the fixation of GC-ending optimal codons, it seems to almost systematically impair the selection on AT-ending optimal codons. These two observations strongly suggest that homologous recombination, *via* gBGC, is a crucial factor universally influencing the nucleotide content of genes and genomes. If confirmed, gBGC could account for several pervasive yet unexplained features of bacterial genomes. Finally, we emphasize that because gBGC has the ability to both mimic and interfere with natural selection, gBGC must be considered by future studies geared at understanding processes driving bacterial genome evolution

Results

A universal relationship between recombination and GC% in Bacteria. In Bacteria, recombination occurs in the form of gene conversion (i.e. unidirectional transfer of genetic material from a donor sequence towards a homologous recipient sequence). To detect past gene conversion events in bacterial species, it is necessary to compare closely related genomes. We therefore selected in the

database of homologous gene families HOGENOM (release 6) [23] all groups of closely related species or strains encompassing at least 6 sequenced genomes. This dataset contains 20 bacterial groups and one archaeal group. For each gene family represented in these groups, we computed i) the average GC% and ii) the index of recombination provided by PHI [24]. PHI is a method for detecting recombination in multiple alignments at the scale of the gene, which has been shown to be more robust than most methods to variations in recombination rates, sequence divergence and population dynamics [24]. We used this test to determine if homologous gene families had experienced gene conversion events among members of the taxa of interest. One important feature of this test is that it measures whether there is sufficient phylogenetic signal in an alignment to tell if recombination has occurred. Only those alignments with sufficient signal, whether recombinant or non-recombinant were retained for tests in the remaining of this study.

In eukaryotes, a general relationship between various estimates of recombination rate and the GC% of genes has been documented and provides indirect evidence for gBGC. Our first goal was to test this prediction in Bacteria and Archaea. To exclude a potential effect of the number of genes in the alignment on our estimates of recombination (because alignments with more sequences are expected to give more power to detect recombination), we focused on single-copy genes of the core genome (i.e. genes that are present in only one copy and found in each genome of a group). In 7 of the 21 groups, the proportion of single-copy genes of the core genome with evidence for recombination was very low (<2.5%), suggesting that these species are clonal or nearly so (Table 1; shaded datasets in Fig. 1). In 13 of the 14 remaining groups, we found a significant positive difference in average GC-content at all and/or at the third position of codons (GC3) between recombinant and non-recombinant genes (Figure 1). The difference in GC3 is generally larger than that at all positions (in 12 out of 13 comparisons), suggesting that the effect of recombination on gene composition is stronger at synonymous positions.

GC and AU ending optimal codons show opposite association with recombination. Recombination is known to enhance the efficacy of selection by breaking linkage between neighbouring selected sites. It is therefore possible that selection is more efficient in recombining genes. This effect should theoretically be more pronounced in the case of selection on codon usage [25], which is relatively weak compared to selection on amino acid sequences. Thus recombination –in the absence of any gBGC – can potentially explain the pronounced effect observed on GC3: if optimal codons tend to be GC-rich, this effect could explain a relationship between GC-content and recombination. The "selection model" sketched above predicts that the frequency of all optimal codons (both GC-ending and AU-ending) should increase with recombination. In contrast, a model invoking gBGC predicts that GC-ending optimal codons should be enriched in recombining regions, but that AU-ending optimal codons should display a weaker or, if gBGC is strong enough to override selection

on codon usage, the opposite pattern. We therefore looked specifically at the frequency of the different optimal codons in recombining and non-recombining genes.

As expected in both models, recombining genes are enriched in GC-ending optimal codons in most datasets (12/14 testable datasets, 9/12 significant). However, in a majority of the groups (12/14 testable datasets, 11/12 significant), recombining genes are depleted in AU-ending optimal codons (Figure 2). In the 2 cases (*Campylobacter* and *Streptococcus*) where the frequency of AU-ending optimal codons increases with recombination, this increase is much weaker than that of GC-ending optimal codons, as expected if gBGC interferes with selection (Figure 2). This asymmetry between GC-ending and AU-ending optimal codons excludes the possibility of pervasive selection for codon usage promoting a better adaptation to the pool of tRNA for genes in regions of high recombination, but is compatible with the predictions of gBGC.

Discussion

Selection for higher genomic GC%?

Our results suggest that recombination affects the GC-content of genes in most bacterial phyla. We analysed genes of the core genome to compare the base composition of genes with or without evidence of recombination. In our sample, seven bacterial species showed very little evidence of recombination (less than 2.5% of gene alignments with detectable traces of recombination): the *Burkholderia pseudomalei* group, *Chlamydia trachomatis, Francisella tularensis, Mycobacterium tuberculosis* and *Yersinia pestis* which are species known to be pathogenic clonal complexes with low polymorphism and probably very low recombination [26–30] while *Brucella spp.* and *Sulfolobus spp.* are composed of ecologically isolated clades, likely because of their respective lifestyle as obligate intracellular pathogen or ecotypes endemic of hot springs [31–33]. The 14 other bacterial species contain clear signal of recombination (14% to 57% of genes with evidence of recombination). In 13 of these 14 species, we observed that the GC-content (measured at the third codon position or along the entire coding region) is higher among recombining genes compared to other (hereafter labelled as “non-recombining”) core genes.

Several hypotheses have been proposed to explain the variations in GC-content among bacterial genomes [34]. Recently, two studies have revealed that the genomic GC-content of bacterial genomes is always higher than what would be predicted from mutational bias [21,22]. Hence, it seems inescapable that some other evolutionary force is driving the genomic GC-content towards higher values in virtually all bacterial species, except maybe for the most AT-rich genomes [21]. Recombination is known to enhance the efficiency of selection by breaking linkage among sites. It is therefore conceivable that our results merely reveal a universal selective pressure favouring GC-

rich alleles. But the mechanism underlying such selection would have to be acting more efficiently on synonymous sites than non-synonymous sites because the difference of GC% between recombining and non-recombining genes is higher at the third position of codons. This excludes potential selection on amino-acid content. One selectable trait that may influence synonymous positions is codon usage. If optimal codons tend to be GC-rich, recombination could drive GC% higher by favouring the adaptation of genes to better translation efficiency. However, we observed a higher GC-content in recombining genes even in species favouring A/U-ending codons (fig. 2). Moreover, A/U-ending and G/C-ending optimal codons show opposite relationships with recombination in most species. These observations suggest that the evolutionary force explaining our results is also largely independent from selection on codon usage.

In fact, as suggested by Hershberg and Petrov [21], who observed that the intergenic regions of bacterial genomes also have higher GC% than expected from their mutational pattern, it seems likely that the process is unlinked to gene expression or function. Hence, either there is selection acting simultaneously on each nucleotide of a bacterial genome to become G or C, or GC-biased gene conversion, which has now been observed in a variety of Eukaryotes is also at work in Bacteria.

gBGC and genome nucleotide composition

The hypothesis that gBGC plays a role in bacterial genome evolution, had been considered previously [22]. Hildebrand et al. analysed the correlation between genome-wide measures of recombination rate (scaled by effective population size) with genomic GC-content among 34 species, covering different bacterial phyla. As they did not find any significant correlation, they concluded that there was no evidence of gBGC in Bacteria [22]. However, the intensity of gBGC depends not only on the recombination rate and on the effective population size, but also on the strength of the mismatch repair bias causing the conversion bias (for review, see [6]). The biases induced by the DNA repair machinery are likely to differ among species. Moreover, even though the mutation patterns are universally AT-biased, they do show significant differences among species [21]. Hence, the gBGC model does not necessarily predict a good correlation between genome-wide average recombination rate and genomic GC-content among species, particularly for the large evolutionary scale considered here. To illustrate this point, one can compare two eukaryotic species known to be subject to gBGC: human and budding yeast [6]. The recombination rate is more than 300 times higher in budding yeast than in human (375 cM/Mb vs 1.1 cM/Mb) [15], and the effective population size is probably several orders of magnitude larger in yeasts than in mammals [35]. Yet the genomic GC-content is lower in yeast (38%) than in human (41%). Thus, at this evolutionary scale, the difference in genomic GC-content is not positively correlated to differences

in recombination rate and effective population size. In fact, to test expectation from the gBGC model, it is more appropriate to investigate correlations between base composition and recombination rate within genomes, so that the other parameters (effective population size, mutation pattern and DNA repair machinery) can be controlled for. Indeed, as predicted by the gBGC model, in both humans and budding yeast, there is a positive correlation between recombination and intra-genomic variations in GC-content [18].

A dynamic model of the evolution of nucleotide composition: long-term equilibrium* vs. *short-term disequilibrium

If the base composition of a genome is at evolutionary equilibrium then, by definition, the number of AT to GC substitutions must be equal to the number of GC to AT substitutions. Hildebrandt and colleagues [22] noted that in a large majority of bacterial genomes (94/149), the number of GC to AT changes (inferred from the comparison of closely related organisms) exceeds the number of AT to GC changes. Given that genomic base composition strongly fluctuates over evolutionary times (as demonstrated by the wide distribution of GC-content across bacterial species), it is not surprising that many genomes are not at equilibrium. However, what is unexpected is that this non-stationarity predominantly leads to loosing GC-content: *a priori*, at the scale of the entire bacterial biodiversity, one would expect to observe as many GC-increasing genomes as GC-decreasing genomes. One potential explanation is that the observed excess of GC to AT changes among closely related genomes corresponds to polymorphic mutations, which eventually do not reach fixation (because of selection or gBGC) [22]. Here we would like to emphasize that polymorphism is not the only possible explanation: in fact, the gBGC model predicts that non-stationarity should be a common feature of genomes. Indeed, the strength of gBGC is expected to fluctuate rapidly over time, notably as a consequence of variations in effective population size and recombination rate. Population genetic models predict that when gBGC is effective, it can very rapidly increase the GC-content of genomes, whereas in absence of gBGC the evolution of base composition through mutational bias is a very slow process [9]. Hence, under this model of fluctuating gBGC, genomes should be subject to short episodes of rapid increase in GC-content followed by longer periods of slow decrease driven by the universal mutational bias towards AT. Thus, even under the assumption that the base composition of genomes remains stable on the long term, the gBGC model predicts that at any given time point, a majority of genomes should be apparently loosing GC-content. It is therefore not surprising that an excess of GC to AT changes is observed even in bacterial genomes that show no evidence of recent recombination [22].

Is gBGC a universally conserved mechanism?

Altogether our results are consistent with the hypothesis that the base composition of

bacterial genomes is affected by gBGC. There is clear evidence that gBGC occurs in eukaryotic genomes [10] and the analysis of conversion tracts in yeast meiotic product indicates that the conversion bias is most probably due to the mismatch repair machinery (MMR) [36]. Given that the MMR components involved in homologous recombination (MutS, MutL) are conserved between Bacteria and Eukaryotes, it is *a priori* not surprising that gBGC also affects Bacteria. However, the association of gBGC to MutSL genes is not straightforward. These genes are absent of three of our genome datasets, *C. jejuni, H. pylori* and *Bifidobacterium longum,* resulting from ancestral losses in Delta-Proteobacteria and Actinobacteridae, respectively [37]. In *H. pylori* indeed, we do not find evidence of gBGC while the genomes are recombining at high frequency [38,39]. In *C. jejuni* and *B. longum*, however, we observe patterns similar to the other bacterial datasets that are in support of the existence of gBGC, indicating that it does not depend on the presence of a typical MutSL complex. The existence of gBGC in Bacteria and Eukaryotes suggests that it may have been present in the last universal common ancestor of all cellular life forms (LUCA). Unfortunately, the only archaeal dataset matching our criteria was a group of *Sulfolobus* sp. genomes (*Susp*) for which our analysis showed few evidence of recombination (Table 1), in agreement with the previously described isolation of endemic clades in this group [33].

gBGC could explain the relationships between bacterial genome size and GC-content

The model we propose provides a simple explanation for several important results of comparative bacterial genomics. First, gBGC can explain why bacterial genomes can maintain a high GC-content, even though the pattern of mutation is universally AT-biased [21,22]. Second, gBGC can explain some of the intragenomic heterogeneity in GC-content observed in bacterial genomes. Indeed, we observe that genes with evidence of recombination display on average substantially higher GC% than other genes. This observation also suggests that the probability of recombination is variable among genes in the genome, possibly for adaptive reasons. Furthermore, because bacterial genomes have a tendency to incorporate genes from distant organisms, gBGC could explain why native and anciently acquired genes tend to have higher GC% than recently transferred ones [40,41].

Interestingly, gBGC could also explain the strong relationship observed between genome size and genomic GC (Fig. 3). This pattern, which was first reported by Heddi et al. [42], has remained unexplained up to now [43–45]. It is known that the genomes of endosymbiotic bacteria, which are generally very reduced, tend to be very AT-rich, possibly because of the loss of key repair genes [43]. However, even when bacteria with such lifestyles are excluded, the correlation between GC-content and genome size remain almost unchanged ($r^2 = 0.29$, fig. 3).

We propose that this relationship could be the result of different requirements for recombination. Indeed, population genetic models predict that maintaining a large number of genes requires efficient recombination. In the absence of recombination, all genes are irremediably linked and deleterious mutations can be efficiently eliminated only on a small number of loci [46,47]. Recombination, by breaking associations among genes, is allowing selection to act more independently on each locus. Hence, large number of functional genes can only be maintained in species with relatively high recombination rates. In the presence of gBGC, a secondary consequence of this requirement should be an increase in genomic GC. This hypothesis fits with the observation that selection is in fact more efficient in large genomes [48]. At the other end of the genome size spectrum, some bacterial species, and in particular endosymbiotic Bacteria such as *Buchnera aphidicola*, are driven toward extremely low GC-content: because they are generally transmitted vertically from parent host to offspring, they typically have a very small effective population size and lack the opportunity to recombine with other lineages, so much so that they tend to lose some key genes of the recombination machinery [49]. As a result, they can only maintain small genomes and cannot counteract the universal mutational bias toward AT with gBGC.

gBGC, a new component of the neutral theory of evolution

The variations of GC-content in Bacteria have long sought explanation. The results presented here highlight a strong relationship between the GC-content of genes and their history of recombination. This result, and the observation that bacterial genomes are generally above the GC-content predicted from their mutational bias towards AT can only be explained by complex, *ad-hoc* scenarios invoking pervasive selection for higher GC throughout the genome, but are fully consistent with the existence of gBGC. This mechanism, which favours the fixation of G/C alleles at polymorphic sites in recombining regions, was previously thought to be present only in Eukaryotes. Our discovery is important because gBGC has been shown to interfere with the efficiency of selection in Eukaryotes, and to lead to false positives in the search for regions under positive selection in a genome. GC-Biased Gene Conversion could explain many features of bacterial genomes, such as the intragenomic heterogeneity of GC-content or the tendency of large genomes to be GC-rich. The prevalence – if not universality – of this phenomenon underlines the importance of incorporating gBGC in the set of evolutionary forces to be considered when searching for signature of adaption in genomes.

Genomic datasets

We used the HOGENOM database [23] to select sets of genome sequences comprising at least six closely related strains or species. This criterion left 17 groups of species representing a variety of bacterial and archeal species (table 1). For each gene family, CDSs were extracted and re-aligned with MUSCLE [50] using default parameters. G+C% and genes frequencies in genomes were computed using custom Python scripts.

Detection of recombinant genes

We used the software PHI [24] to test if a gene family alignment contained evidence for recombination. Only families for which the site permutation test could be performed were considered, i.e. where the phylogenetic signal was sufficient to accept or reject the hypothesis of recombination. An alignment was determined to be "recombinant" if the p-value of the permutation test was lower than 0.05.

Frequencies of optimal codons

Optimal codons for each amino-acid were computed by comparing synonymous codon frequencies within CDSs encoding ribosomal proteins (as defined from HOGENOM family annotations) versus all other CDSs. The codons statistically enriched in ribosomal proteins (with Chi-squared p-value < 0.05) were considered as optimal under the hypothesis of selection for highly expressed genes to be adapted to the tRNA pool (table S2). We then computed the frequency of optimal codons (Fop) by making the ratio of the count of the optimal codons over the count of the corresponding amino-acid. Fop were calculated for each optimal codon separately and for optimal codons pooled by composition at the third position. As there is a debate on whether this method is appropriate to define optimal codons [51], we also used an alternative definition and took optimal codons datasets from a previous exhaustive survey of Hershberg and Petrov [52] (see Figure S1 and table S3). This alternative approach yielded similar conclusions regarding the relative frequency of GC-ending and AU-ending optimal codons.

References:

1. Doolittle WF (2013) Is junk DNA bunk? A critique of ENCODE. Proc Natl Acad Sci 110: 5294–5300. doi:10.1073/pnas.1221376110.

2. Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, et al. (1985) The mosaic genome of warm-blooded vertebrates. Science 228: 953–958.

3. Sueoka N (1962) ON THE GENETIC BASIS OF VARIATION AND HETEROGENEITY OF DNA BASE COMPOSITION. Proc Natl Acad Sci 48: 582–592.

4. McCutcheon JP, Moran NA (2010) Functional convergence in reduced genomes of bacterial symbionts spanning 200 My of evolution. Genome Biol Evol 2: 708–718. doi:10.1093/gbe/evq055.

5. Pagani I, Liolios K, Jansson J, Chen I-MA, Smirnova T, et al. (2012) The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 40: D571–579. doi:10.1093/nar/gkr1100.

6. Duret L, Galtier N (2009) Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet 10: 285–311. doi:10.1146/annurev-genom-082908-150001.

7. Ratnakumar A, Mousset S, Glemin S, Berglund J, Galtier N, et al. (2010) Detecting positive selection within genomes: the problem of biased gene conversion. Philos Trans R Soc B Biol Sci 365: 2571–2580. doi:10.1098/rstb.2010.0007.

8. Meunier J, Duret L (2004) Recombination drives the evolution of GC-content in the human genome. Mol Biol Evol 21: 984–990. doi:10.1093/molbev/msh070.

9. Duret L, Arndt PF (2008) The Impact of Recombination on Nucleotide Substitutions in the Human Genome. PLoS Genet 4: e1000071. doi:10.1371/journal.pgen.1000071.

10. Webster MT, Hurst LD (2012) Direct and indirect consequences of meiotic recombination: implications for genome evolution. Trends Genet TIG 28: 101–109. doi:10.1016/j.tig.2011.11.002.

11. Nagylaki T (1983) Evolution of a finite population under gene conversion. Proc Natl Acad Sci U S A 80: 6278–6281.

12. Galtier N, Duret L (2007) Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends Genet TIG 23: 273–277. doi:10.1016/j.tig.2007.03.011.

13. Galtier N, Duret L, Glémin S, Ranwez V (2009) GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. Trends Genet 25: 1–5. doi:10.1016/j.tig.2008.10.011.

14. Necşulea A, Popa A, Cooper DN, Stenson PD, Mouchiroud D, et al. (2011) Meiotic recombination favors the spreading of deleterious mutations in human populations. Hum Mutat 32: 198–206. doi:10.1002/humu.21407.

15. Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM (2008) High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature 454: 479–485. doi:10.1038/nature07135.

16. Capra JA, Pollard KS (2011) Substitution patterns are GC-biased in divergent sequences across the metazoans. Genome Biol Evol 3: 516–527. doi:10.1093/gbe/evr051.

17. Escobar JS, Glémin S, Galtier N (2011) GC-biased gene conversion impacts ribosomal DNA evolution in vertebrates, angiosperms, and other eukaryotes. Mol Biol Evol 28: 2561–2575. doi:10.1093/molbev/msr079.

18. Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, et al. (2012) Evidence for widespread GC-biased gene conversion in eukaryotes. Genome Biol Evol 4: 675–682. doi:10.1093/gbe/evs052.

19. Foerstner KU, von Mering C, Hooper SD, Bork P (2005) Environments shape the nucleotide composition of genomes. EMBO Rep 6: 1208–1213. doi:10.1038/sj.embor.7400538.

20. Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci U S A 85: 2653–2657.

21. Hershberg R, Petrov DA (2010) Evidence That Mutation Is Universally Biased towards AT in Bacteria. PLoS Genet 6: e1001115. doi:10.1371/journal.pgen.1001115.

22. Hildebrand F, Meyer A, Eyre-Walker A (2010) Evidence of Selection upon Genomic GC-Content in Bacteria. PLoS Genet 6: e1001107. doi:10.1371/journal.pgen.1001107.

23. Penel S, Arigon A-M, Dufayard J-F, Sertier A-S, Daubin V, et al. (2009) Databases of homologous gene families for comparative genomics. BMC Bioinformatics 10: S3. doi:10.1186/1471-2105-10-S6-S3.

24. Bruen TC, Philippe H, Bryant D (2006) A Simple and Robust Statistical Test for Detecting the Presence of Recombination. Genetics 172: 2665–2681. doi:10.1534/genetics.105.048975.

25. McVean GA, Charlesworth B (2000) The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics 155: 929–944.

26. Ussery DW, Kiil K, Lagesen K, Sicheritz-Pontén T, Bohlin J, et al. (2009) The genus burkholderia: analysis of 56 genomic sequences. Genome Dyn 6: 140–157. doi:10.1159/000235768.

27. Joseph SJ, Didelot X, Gandhi K, Dean D, Read TD (2011) Interplay of recombination and selection in the genomes of Chlamydia trachomatis. Biol Direct 6: 28. doi:10.1186/1745-6150-6-28.

28. Achtman M, Zurth K, Morelli G, Torrea G, Guiyoule A, et al. (1999) Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis. Proc Natl Acad Sci 96: 14043–14048. doi:10.1073/pnas.96.24.14043.

29. Keim P, Johansson A, Wagner DM (2007) Molecular Epidemiology, Evolution, and Ecology of Francisella. Ann N Y Acad Sci 1105: 30–66. doi:10.1196/annals.1409.011.

30. Supply P, Marceau M, Mangenot S, Roche D, Rouanet C, et al. (2013) Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of Mycobacterium tuberculosis. Nat Genet 45: 172–179. doi:10.1038/ng.2517.

31. Wattam AR, Williams KP, Snyder EE, Almeida NF Jr, Shukla M, et al. (2009) Analysis of ten Brucella genomes reveals evidence for horizontal gene transfer despite a preferred intracellular lifestyle. J Bacteriol 191: 3569–3579. doi:10.1128/JB.01767-08.

32. Whitaker RJ, Grogan DW, Taylor JW (2003) Geographic Barriers Isolate Endemic Populations of Hyperthermophilic Archaea. Science 301: 976–978. doi:10.1126/science.1086909.

33. Reno ML, Held NL, Fields CJ, Burke PV, Whitaker RJ (2009) Biogeography of the Sulfolobus islandicus pan-genome. Proc Natl Acad Sci U S A 106: 8605–8610. doi:10.1073/pnas.0808945106.

34. Rocha EPC, Feil EJ (2010) Mutational Patterns Cannot Explain Genome Composition: Are There Any Neutral Sites in the Genomes of Bacteria? PLoS Genet 6: e1001104. doi:10.1371/journal.pgen.1001104.

35. Lynch M (2010) Evolution of the mutation rate. Trends Genet TIG 26: 345–352. doi:10.1016/j.tig.2010.05.003.

36. Lesecque Y, Mouchiroud D, Duret L (2013) GC-Biased Gene Conversion in Yeast Is Specifically Associated with Crossovers: Molecular Mechanisms and Evolutionary Significance. Mol Biol Evol. doi:10.1093/molbev/mst056.

37. Lin Z, Nei M, Ma H (2007) The origins and early evolution of DNA mismatch repair genes—multiple horizontal gene transfers and co-evolution. Nucleic Acids Res 35: 7591–7603. doi:10.1093/nar/gkm921.

38. Suerbaum S, Smith JM, Bapumia K, Morelli G, Smith NH, et al. (1998) Free recombination within Helicobacter pylori. Proc Natl Acad Sci 95: 12619–12624. doi:10.1073/pnas.95.21.12619.

39. Falush D, Kraft C, Taylor NS, Correa P, Fox JG, et al. (2001) Recombination and mutation during long-term gastric colonization by Helicobacter pylori: Estimates of clock rates, recombination size, and minimal age. Proc Natl Acad Sci 98: 15056–15061. doi:10.1073/pnas.251396098.

40. Daubin V, Lerat E, Perrière G (2003) The source of laterally transferred genes in bacterial genomes. Genome Biol 4: R57. doi:10.1186/gb-2003-4-9-r57.

41. Daubin V, Ochman H (2004) Bacterial Genomes as New Gene Homes: The Genealogy of ORFans in E. coli. Genome Res 14: 1036 –1042. doi:10.1101/gr.2231904.

42. Heddi A, Charles H, Khatchadourian C, Bonnot G, Nardon P (1998) Molecular characterization of the principal symbiotic bacteria of the weevil Sitophilus oryzae: a peculiar G + C content of an endocytobiotic DNA. J Mol Evol 47: 52–61.

43. Moran NA (2002) Microbial minimalism: genome reduction in bacterial pathogens. Cell 108: 583–586.

44. Bastolla U, Moya A, Viguera E, van Ham RCHJ (2004) Genomic determinants of protein folding thermodynamics in prokaryotic organisms. J Mol Biol 343: 1451–1466. doi:10.1016/j.jmb.2004.08.086.

45. Musto H, Naya H, Zavala A, Romero H, Alvarez-Valín F, et al. (2006) Genomic GC level, optimal growth temperature, and genome size in prokaryotes. Biochem Biophys Res Commun

347: 1–3. doi:10.1016/j.bbrc.2006.06.054.

46. MULLER HJ (1964) THE RELATION OF RECOMBINATION TO MUTATIONAL ADVANCE. Mutat Res 106: 2–9.

47. Felsenstein J (1974) The evolutionary advantage of recombination. Genetics 78: 737–756.

48. Kuo C-H, Ochman H (2009) Deletional Bias across the Three Domains of Life. Genome Biol Evol 2009: 145–152. doi:10.1093/gbe/evp016.

49. McCutcheon JP, Moran NA (2012) Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol 10: 13–26. doi:10.1038/nrmicro2670.

50. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113. doi:10.1186/1471-2105-5-113.

51. Hershberg R, Petrov DA (2012) On the Limitations of Using Ribosomal Genes as References for the Study of Codon Usage: A Rebuttal. PLoS ONE 7: e49060. doi:10.1371/journal.pone.0049060.

52. Hershberg R, Petrov DA (2009) General Rules for Optimal Codon Choice. PLoS Genet 5: e1000556. doi:10.1371/journal.pgen.1000556.

53. Smith JM (1992) Analyzing the mosaic structure of genes. J Mol Evol 34: 126–129. doi:10.1007/BF00182389.

54. Jakobsen IB, Easteal S (1996) A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. Comput Appl Biosci CABIOS 12: 291–295.

55. Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet: 54–78.

56. Miyashita N, Langley CH (1988) Molecular and phenotypic variation of the white locus region in Drosophila melanogaster. Genetics 120: 199–212.

57. McVean G, Awadalla P, Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics 160: 1231–1241.

58. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW (2006) GARD: a

genetic algorithm for recombination detection. Bioinformatics 22: 3096 –3098. doi:10.1093/bioinformatics/btl474.

59. Didelot X, Falush D (2007) Inference of bacterial microevolution using multilocus sequence data. Genetics 175: 1251–1266. doi:10.1534/genetics.106.063305.

60. Didelot X, Lawson D, Darling A, Falush D (2010) Inference of Homologous Recombination in Bacteria Using Whole Genome Sequences. Genetics 186: 1435–1449. doi:10.1534/genetics.110.120121.

61. Choi SC, Rasmussen MD, Hubisz MJ, Gronau I, Stanhope MJ (2012) Replacing and Additive Horizontal Gene Transfer in Streptococcus. Mol Biol Evol. Available: http://mbe.oxfordjournals.org.gate1.inist.fr/content/early/2012/07/12/molbev.mss138. Accessed 18 July 2012.

Dataset	Taxon name	# genomes	# core genes	# recombining core genes	# non-recombining core genes	Mean GC%	Mean GC1	Mean GC2	Mean GC3
Brsp	*Brucella* spp.	9	1675	0	902	0.58	0.63	0.44	0.67
Ftul	*Francisella tularensis*	8	1015	0	731	0.33	0.45	0.33	0.22
Mtub	*Mycobacterium tuberculosis* complex	7	2222	0	240	0.66	0.68	0.49	0.79
Ypes	*Yersinia pestis*	11	2017	18	1197	0.49	0.57	0.4	0.49
Bmal	*Burkholderia* Pseudomallei group	9	1482	16	1093	0.68	0.68	0.47	0.88
Susp	*Sulfolobus* spp.	8	1386	17	1041	0.35	0.42	0.33	0.29
Ctra	*Clamydia trachomatis*	13	772	17	704	0.42	0.52	0.38	0.35
Bcen	*Burkholderia cenocepacia* complex (BCC)	8	1939	247	1687	0.67	0.68	0.47	0.87
Saur	*Staphylococcus aureus*	15	1464	188	1137	0.33	0.46	0.32	0.22
Abau	*Acinetobacter* spp.	6	1429	233	1059	0.4	0.54	0.37	0.3
Cjej	*Campylobacter jejunii*	6	1048	182	826	0.31	0.43	0.31	0.19
Blon	*Bifidobacterium longum*	6	1006	193	523	0.61	0.63	0.44	0.77
Spyo	*Streptococcus pyogenes*	12	1051	224	738	0.39	0.5	0.35	0.32
Cbot	*Clostridium botulinum*	8	1715	401	1246	0.29	0.39	0.29	0.17
Spne	*Streptococcus pneumoniae*	13	1090	256	728	0.41	0.52	0.34	0.37
Sent	*Salmonella enterica*	14	2121	589	1432	0.53	0.6	0.41	0.59
Lisp	*Listeria* spp.	8	1631	642	973	0.38	0.5	0.34	0.3
Nmen	*Nesseiria meningitidis*	8	1156	580	571	0.54	0.59	0.41	0.64
Bant	*Bacillus* spp.	17	1730	868	846	0.36	0.5	0.33	0.26
Ecol	*Escherichia coli*	35	1357	749	577	0.52	0.61	0.4	0.56
Hpyl	*Helicobacter pylori*	14	995	650	345	0.4	0.45	0.32	0.43

Table 3.1: Statistics of the dataset used in this study.
The number of non-recombinant, recombinant, and other (i.e., with insufficient phylogenetic signal) core-genome families used for figure 1 are described. Both core genome GC% and its decomposition at the three codon positions are shown.

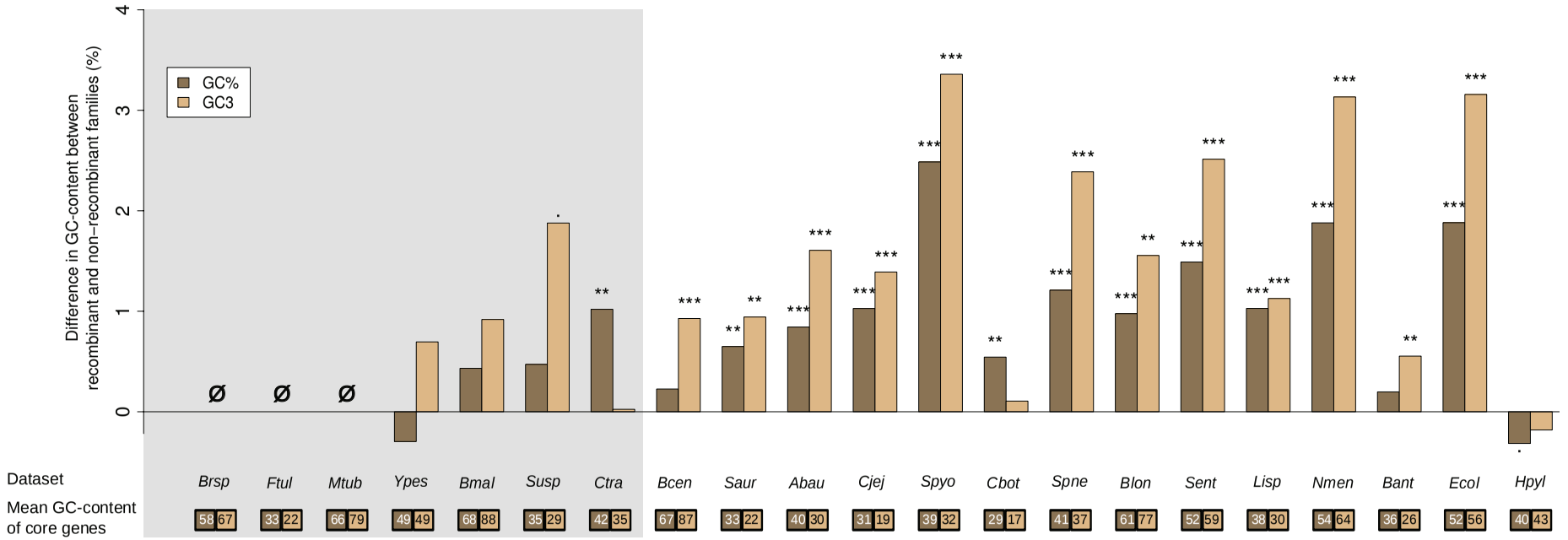


Figure 3.1: **Effect of recombination on core gene family G+C content.**
Difference in average G+C of recombinant and non-recombinant gene families, on total CDS length (GC%, dark brown) or at third codon position only (GC3, light brown) in the core genome of each dataset. A positive difference means recombinant families are richer in G+C. Stars indicate significance of a Student's t-test ('.', $p < 0.1$; '*', $p < 0.05$; '**', $p < 0.01$; '***', $p < 0.001$). Statistical tests are detailed Table S1. Dataset abbreviations are explained Table 1. Boxes under dataset names indicate the mean GC% and GC3 values of core genes. Shading in background marks datasets with very few recombinant gene families (detailed Table 1 [Table 3.1]).

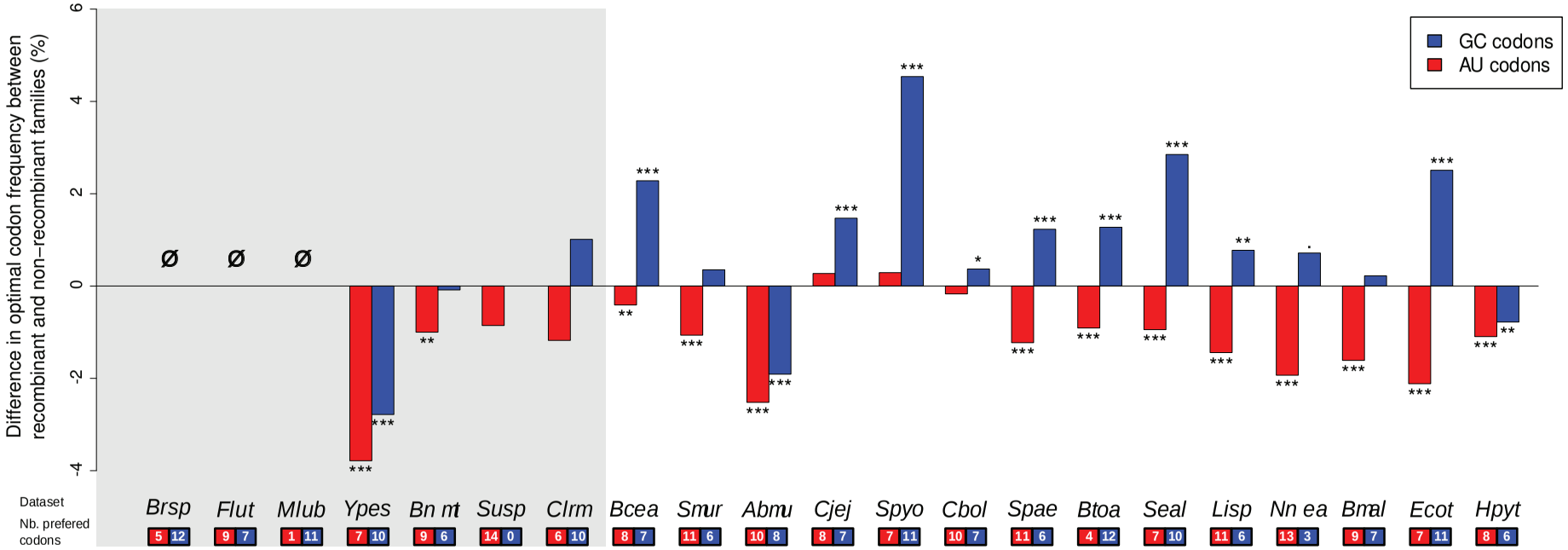


Figure 3.2: **Effect of recombination on gene family codon usage bias.**
Difference in frequency of optimal codons (as determined by ribosomal comparison method) of recombinant and non-recombinant gene families in each dataset for AU-rich (red) and GC-rich (blue) codons. A positive difference means recombinant families are richer in optimal codons. Stars indicate significance of a chi-squared test for independence of the distribution of total counts of optimal and non-optimal codons between recombinant and non recombinant families. Boxes under dataset names indicate the numbers of AU-rich and GC-rich preferred codons used by the taxon (detailed Table S2 [Sup. Table 3.2]). Symbols, shading and dataset abbreviations as in Fig. 1 [Fig 3.1].

3.1.1 Supplementary Figures

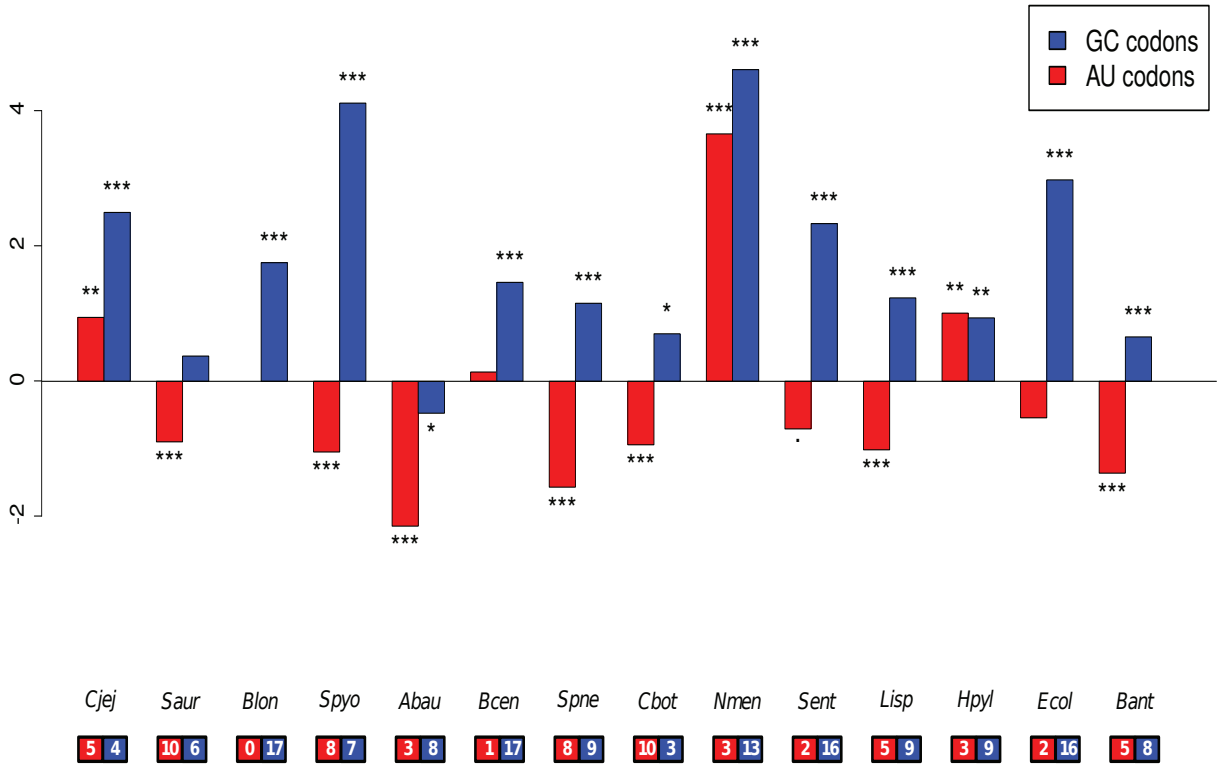


Figure 3.3: **Effect of recombination on gene family codon usage bias when defining optimal codons from the dataset from Hershberg and Petrov (2009).**
Difference in frequency of optimal codons (as determined by Hershberg & Petrov, 2009) of recombinant and non-recombinant gene families in each dataset for AU-rich (red) and GC-rich (blue) codons. A positive difference means recombinant families are richer in optimal codons. Stars indicate significance of a chi-squared test for independence of the distribution of total counts of optimal and non-optimal codons between recombinant and non recombinant families. Boxes under dataset names indicate the numbers of AU-rich and GC-rich preferred codons used by the taxon (detailed Table S3 [Sup. Table 3.3]). Symbols, shading and dataset abbreviations as in Fig. 1 [Fig 3.1].

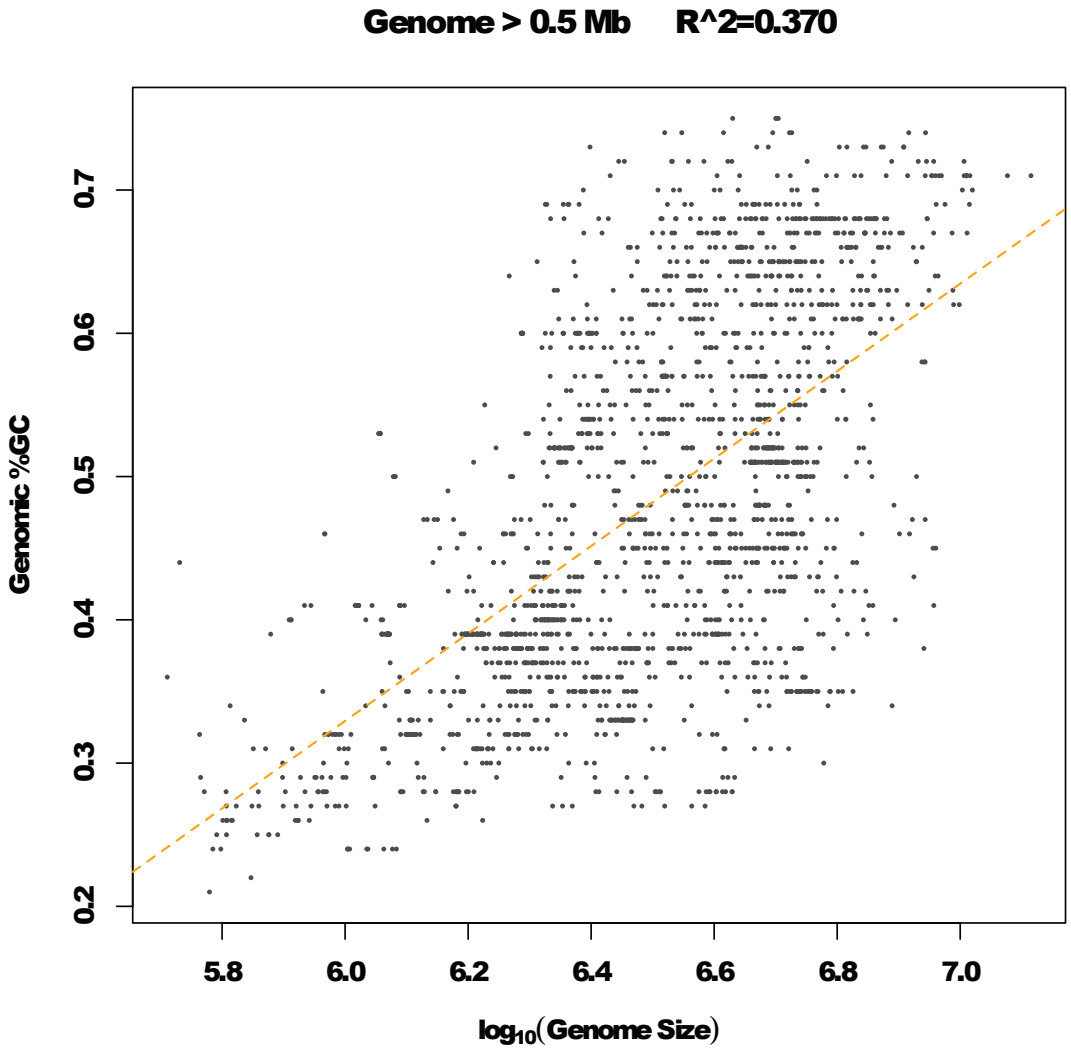


Figure 3.4: **The relationships of genome size and GC%.** The plot is based on 1795 complete genome sequences from Bacteria and Archaea. Source: http://img.jgi.doe.gov

3.1.2 Supplementary Tables

Table 3.2: **Table S2: Preferred codons of prokaryotic datasets as by comparison of ribosomal protein genes to the rest of the genome.**

Dataset	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Abau	GCT	NA	GAC	GAA	TTC	GGT	CAC	ATC	NA	CTT	NA	AAC	CCA	CAA	CGT	TCT	ACT	GTT	NA	TAC
Bant	GCT	TGC	GAC	NA	TTC	GGT	CAC	ATC	NA	CTT	NA	AAC	CCT	CAA	CGT	TCT	ACT	GTT	NA	TAC
Bcen	GCA	NA	NA	GAA	NA	GGT	CAC	ATT	AAG	CTG	NA	AAC	CCG	CAA	CGT	TCG	ACT	GTT	NA	TAC
Blon	GCT	NA	GAC	GAG	NA	GGT	CAC	ATC	AAG	CTC	NA	AAC	CCG	CAG	CGT	TCC	ACC	GTT	NA	TAC
Bmal	GCT	NA	NA	GAA	NA	GGT	CAC	ATT	AAG	CTG	NA	AAC	CCG	CAA	CGT	TCT	ACT	GTT	NA	TAC
Brsp	GCT	NA	GAC	GAA	TTC	GGT	CAC	ATC	AAG	CTC	NA	AAC	CCG	CAG	CGT	TCG	ACC	GTT	NA	TAC
Cbot	GCT	TGC	GAC	GAA	TTC	GGT	CAC	ATC	NA	TTA	NA	AAC	CCA	CAA	AGA	TCA	ACA	GTT	NA	TAC
Cjej	GCA	TGC	GAC	NA	TTC	GGT	CAC	ATC	NA	CTA	NA	AAC	CCA	NA	AGA	TCA	ACT	GTA	NA	TAC
Ctra	GCT	TGC	NA	GAG	TTC	GGT	CAC	ATC	AAG	TTG	NA	NA	CCT	CAG	AGA	AGC	ACT	GTT	NA	TAC
Ecol	GCT	TGC	GAC	GAA	TTC	GGT	CAC	ATC	AAG	CTG	NA	AAC	CCG	CAG	CGT	TCT	ACT	GTT	NA	TAC
Ftul	GCT	NA	GAC	GAA	TTC	GGT	CAC	ATC	AAG	CTT	NA	AAC	CCT	NA	CGT	TCT	ACT	GTT	NA	TAC
Hpyl	GCA	NA	GAC	NA	TTC	GGT	NA	ATT	AAG	TTG	NA	NA	CCA	CAG	AGA	TCA	ACT	GTA	NA	TAC
Lisp	GCT	NA	GAC	GAA	TTC	GGT	CAC	ATC	AAA	CTT	NA	AAC	CCT	CAA	CGT	TCT	ACT	GTA	NA	TAC
Mtub	GCC	NA	GAC	GAG	NA	GGC	NA	ATC	AAG	CTC	NA	AAC	NA	CAG	CGT	NA	ACC	GTC	NA	NA
Nmen	GCT	TGT	GAT	NA	TTC	GGT	NA	ATT	AAA	TTG	NA	AAT	CCT	CAA	CGT	TCT	ACT	GTA	NA	TAC
Saur	GCT	NA	GAC	GAA	TTC	GGT	CAC	ATC	AAA	TTA	NA	AAC	CCA	CAA	CGT	TCA	ACT	GTA	NA	TAC
Sent	GCT	TGC	GAC	GAA	TTC	GGT	CAC	ATC	AAG	CTG	NA	AAC	CCG	NA	CGT	TCT	ACT	GTT	NA	TAC
Spne	GCA	NA	GAC	GAA	TTC	GGT	CAC	ATC	AAA	CTT	NA	AAC	CCA	CAA	CGT	TCA	ACT	GTA	NA	TAC
Spyo	GCA	TGC	GAC	GAA	TTC	GGT	CAC	ATC	AAA	CTT	NA	AAC	CCA	CAA	CGT	TCA	ACT	GTT	NA	TAC
Susp	GCA	NA	GAT	GAA	TTT	GGT	CAT	ATA	NA	TTA	NA	NA	CCA	NA	AGA	AGT	ACA	GTA	NA	TAT
Ypes	GCT	TGC	GAC	GAA	TTC	GGT	CAC	ATC	AAG	CTG	NA	AAC	CCG	NA	CGT	TCT	ACT	GTT	NA	TAC

Table 3.3: **Table S3: Preferred codons of prokaryotic datasets as the consensus of strain sets taken from Hershberg and Petrov (2010) [21].**

Dataset	A	C	D	E	F	G	H	I	K	L	N	P	Q	R	S	T	V	Y
Abau	GCG	NA	GAC	GAA	TTC	GGC	CAC	ATC	AAA	NA	AAC	CCA	NA	NA	NA	NA	NA	TAC
Bant	NA	NA	GAC	GAA	TTC	GGC	CAC	ATC	AAA	TTA	AAC	CCA	CAA	CGC	NA	NA	NA	TAC
Bcen	GCG	TGC	GAC	GAA	TTC	GGC	CAC	ATC	AAG	CTG	AAC	CCG	CAG	CGC	TCG	ACG	GTG	TAC
Blon	GCC	NA	GAC	GAG	TTC	GGC	CAC	ATC	AAG	CTG	AAC	CCG	CAG	CGC	TCC	ACC	GTG	TAC
Bmal	GCG	TGC	GAC	NA	TTC	GGC	CAC	ATC	AAG	CTG	AAC	CCG	CAG	CGC	TCG	ACG	GTG	TAC
Brsp	GCC	NA	GAC	GAA	TTC	GGC	NA	ATC	AAG	CTG	AAC	CCG	CAG	CGC	TCG	ACC	GTG	NA
Cbot	GCT	NA	NA	GAA	TTC	GGT	NA	NA	NA	TTA	AAC	CCA	CAA	AGA	TCT	ACT	GTT	TAC
Cjej	NA	NA	NA	GAA	NA	GGC	NA	ATC	AAA	CTT	NA	CCT	NA	CGC	AGC	NA	GTA	NA
Ctra	GCT	NA	NA	GAA	TTC	GGA	NA	NA	NA	NA	NA	CCT	NA	CGT	TCT	NA	NA	NA
Ecol	GCG	TGC	GAC	GAA	TTC	GGC	CAC	ATC	AAA	CTG	AAC	CCG	CAG	CGC	AGC	ACC	GTG	TAC
Ftul	NA	NA	NA	NA	TTC	GGT	NA	ATC	NA	CTT	AAC	NA	NA	NA	TCT	ACT	GTT	NA
Hpyl	GCG	NA	NA	GAA	TTT	GGG	NA	ATC	AAA	NA	AAC	CCC	NA	CGC	AGC	ACG	GTG	NA
Lisp	GCT	NA	GAC	GAA	TTC	GGC	CAC	ATC	AAA	CTT	AAC	CCA	CAA	CGC	TCC	ACA	GTT	TAC
Mtub	GCC	TGC	GAC	GAG	TTC	GGC	CAC	ATC	AAG	CTG	AAC	CCG	CAG	CGG	TCG	ACC	GTG	TAC
Nmen	GCC	NA	GAC	GAA	TTC	GGC	CAC	ATC	AAA	NA	AAC	CCG	CAA	CGC	AGC	ACC	GTC	TAC
Saur	GCA	NA	GAC	GAA	TTC	GGT	NA	ATC	AAA	TTA	AAC	CCA	CAA	CGC	TCT	ACA	GTT	TAC
Sent	GCG	TGC	GAC	GAA	TTC	GGC	CAC	ATC	AAA	CTG	AAC	CCG	CAG	CGC	AGC	ACC	GTG	TAC
Spne	GCT	NA	GAC	GAA	TTC	GGT	CAC	ATC	AAA	TTG	AAC	CCA	CAA	CGC	AGC	ACT	GTT	TAC
Spyo	GCT	NA	GAC	GAA	TTC	GGT	CAC	ATC	AAA	NA	AAC	CCA	CAA	CGC	TCT	NA	GTT	TAC
Susp	GCT	NA	GAC	GAG	TTC	GGG	NA	ATC	AAG	TTG	AAC	CCC	NA	AGG	TCC	ACC	GTG	TAC
Ypes	GCG	NA	GAC	GAA	TTC	GGC	CAC	ATC	AAA	CTG	AAC	CCG	CAG	CGC	AGC	ACC	GTG	TAC

3.1.3 Supplementary Material

Recombination detection methods Since most of the present work rely on relating recombination to other sequence parameters, it was crucial to use an accurate method for recombination detection from gene alignments. Several different methods have been developed in the last two decades, with a large spectrum of complexity in modelling and testing the presence of recombination in sequences. PHI [29] was chosen first because Bruen and colleagues compared the PHI statistics to a large panel of other measures and statistics existing at the time of their publication: Max χ^2 [41], NSS [42], measures of correlation of linkage desequilibrium (r^2 and |D'| with distance [43–45] and results obtained from a coalescent-based likelihood permutation test (LPT) from LDHat [46]. In this benchmark, PHI appeared in general to be equally or more sensitive and specific than all other methods [29], and was extremely faster to compute, which was a notable advantage given the size of the datasets we planned to analyse. We also tried two other methods not discussed in Bruen et al.: the SBP/GARD algorithm from the HyPhy package [47] and the ClonalFrame/ClonalOrigin program [48,49]. Both methods use explicit modelling of a coalescent-based framework.

GARD is a genetic algorithm using simulated populations to sample possible scenarios of sequence evolution under a model including recombination [47]. While proving very sensitive on punctual analyses, GARD method have drawbacks when considering its use at large scale. As its sensitivity depends on a good sampling of the virtual populations that are simulated, one needs to take large samples to obtain robust results, which induces very intensive computation. Comparison of results from GARD and PHI showed general agreement of the methods (data not shown), so PHI was preferred for its rapidity and robustness.

ClonalOrigin is a Bayesian method for describing the history of recombination in datasets of complete genomes [48]. It has the advantage to provide quantitative estimates of the the coalescent model parameters, including the locus-specific recombination rate, rather than just testing the presence of recombination. Using this method could have allowed us to correlate the recombination rate to the GC% of genes. However, ClonalOrigin is designed to estimates the recombination history from alignment of syntenic blocks of genomes and its model may be over-parameterized for gene alignments. Indeed, some studies reported 'biases from boundary effects in short alignment blocks' when working on alignments shorter than 1.5kb [50]. We also observed difficulties to reach convergence of model parameters when estimating them on gene alignments, even when running the MCMC algorithm for more than one million iterations, which correspond to several weeks of computation for each gene alignment (data not shown). Considering these results were not satisfying to perform parametric correlations, the computation time involved made us to prefer PHI statistics and to perform comparison tests of GC% of recombining vs. non-recombining genes rather than correlations.

3.2 Heterogeneity of genome GC-content and gene population sizes

3.2.1 Hypothesis: large gene population size enhances gBGC

The following section deals with the idea of gBGC being the cause of the variations of GC-content within prokaryotic genomes. In short, gBGC could be affecting pangenomes in a heterogeneous way because of varied gene-specific population sizes. This is in fact the original intuition that led us to tackle the existence of gBGC in prokaryotic genomes. However, the test of this particular hypothesis – the link between pangenome structure and composition biases stemming from gBGC – revealed complex phenomena that gBGC alone could not account for. For this reason, this was not integrated to the manuscript to be submitted. However, it constitutes an intriguing result that we propose to present here.

The dynamics of bacterial genomes is particular in that there are wide variations in gene content between members of the same species: some genes are shared by all members of the species (core genome), whereas many others are only present in a subset of strains (and constitute the so-called accessory genome). The concept of pan-genome has been introduced to describe this variability (Tettelin et al., 2005).

The pangenome is structured in two main compartments regrouping the majority of gene families: the core genome and the strain-specific genome, in which can be found ORFans and other recently acquired genes (Fig. 1.1). It has been documented that this spectrum of gene frequencies is related to the variation of GC-content within genomes: there is as a systematic bias of base composition opposing young genes, including ORFans, to older genes in a AT-rich to GC-rich gradient (Daubin et al., 2003; Daubin and Ochman, 2004a; Lassalle et al., 2011).

The anomalous composition of new genes can be explained by the composition of their genome of origin, which is likely different from their new host genome, given the existence of large inter-genomic variations (Sueoka, 1988). It was proposed by Lawrence and Ochman (1997) that the age of the gene was a structuring factor of base composition because anomalous genes 'ameliorate' with time towards the mutational equilibrium specified by the local mutational bias (Lawrence and Ochman, 1997). However, this model does not explain the polarity of the compositional bias: why should donor genomes always be AT-richer than the recipient?

In fact, the age of genes may be a confounding factor with another variable, which is the population size of genes within a species: young genes are rare and old genes are ubiquitous. The frequency of genes in the pan-genome can be seen as a proxy of their population size. The structure of the pan-genome may reveal the dynamic of gene gain and loss in bacterial populations, with genes experiencing phases of intermediate frequencies in the species before complete loss or fixation. Alternatively, some genes could be maintained at relatively low frequency due to local adaptations. In any case, it implies that, in the long run, genes evolving in the same species have different specific population size.

It appears there is a significant positive relationship between GC-content and frequency of genes in most the taxa studied above (Table 3.4). Interestingly, when removing extremes of the

pangenome – i.e. ORFans and core genes – from the dataset, these correlations still hold in many cases, though to a lesser extant, (Table 3.4), showing that this relation does not only characterize the opposition of the two major pangenome component but a true continuous effect of gene frequency. This relation of GC-content with gene population size might be explained in the light of GC-biased gene conversion (gBGC) by a gene population model.

First, genes of the core genome have a higher population size than recent genes, and hence they are expected to be less subject to the impact of genetic drift. Drift delays the fixation of favoured alleles and thus attenuates the effect of gBGC as it does with selection, and for this reason core genes are more prone to fix the GC-rich alleles than rare genes.

Moreover, rare genes should also have fewer opportunities to recombine, because sex with a conspecific is less likely to bring a homolog in the host cell. As the strength of gBGC is proportional to the recombination rate, rare genes with lower gene-specific recombination rate are less subject to this fixation bias. This effect must be multiplied by the time of residence of genes. Core genes which have been 'there' at high frequency for a long time, i.e. long before the origin of the observed species, had much time to recombine and reach high GC-content value through gBGC. In opposition, recent genes and notably ORFans may have been replicating for a long time in autonomous elements such as plasmids, viruses and other ICEs ((Cortez et al., 2009)) that paradoxically must not recombine much (by homologous recombination) because they are most often rare in the population they infect.

The combination of these properties linked to higher gene frequency – reduced drift and higher recombination rate – makes gBGC to be stronger in frequent genes. Different classes of genes would thus tend toward different substitutional equilibriums: rare genes would only undergo the local mutational bias, that was shown to be in general biased towards AT-rich compositions in Prokaryotes (Hershberg and Petrov, 2010; Hildebrand et al., 2010), and would thus decrease their GC-content towards the mutational equilibrium. In opposition, core genes undergo both the mutational bias and a strong opposing effect of gBGC. Depending on the relative strength of these forces, the GC-content of core genes will increase or at least not decrease as much as for rare genes.

Overall, the observation of a positive relation between GC-content of genes and their frequency in pangenomes (Table 3.4) in most tested clades is a support of such a model.

Dataset	Taxon full name	nb. sampled genomes	nb. recombinant / total gene families in taxon pangenome		Spearman's rho G+C% ~gene freq.	
				(core genome)	full pangenome	without ORFans & core
Brsp	*Brucella* spp.	9	13 / 2875	(0 / 902)	0.267***	nd
Ftul	*Francisella tularensis*	8	8 / 1740	(0 / 1015)	0.188***	nd
Mtub	*Mycobacterium tuberculosis* complex	7	19 / 2991	(0 / 2222)	0.037	nd
Bmal	*Burkholderia* Pseudomallei group	9	153 / 3172	(17 / 1109)	-0.013	0.117***
Ctra	*Clamydia trachomatis*	13	20 / 800	(17 / 721)	0.044	0.134
Susp	*Sulfolobus* spp.	8	100 / 1760	(17 / 1058)	-0.040	-0.147**
Ypes	*Yersini pestis*	11	81 / 1962	(18 / 1215)	0.114***	-0.040
Cjej	*Campylobacter jejunii*	6	227 / 1314	(186 / 1008)	0.123***	0.047
Saur	*Staphylococcus aureus*	15	342 / 2071	(188 / 1325)	0.205***	0.018
Blon	*Bifidobacterium longum*	6	237 / 964	(194 / 716)	0.166***	0.122
Spyo	*Streptococcus pyogenes*	12	336 / 1499	(224 / 962)	0.111***	0.078
Abau	*Acinetobacter* spp.	6	392 / 2399	(235 / 1292)	0.112***	0.096**
Bcen	*Burkholderia cenocepacia* complex (BCC)	8	534 / 4398	(254 / 1934)	0.090***	0.242***
Spne	*Streptococcus pneumoniae*	13	371 / 1579	(258 / 984)	0.391***	0.179***
Cbot	*Clostridium botulinum*	8	620 / 2698	(402 / 1647)	0.053**	0.063
Nmen	*Nesseiria meningitidis*	8	744 / 1624	(580 / 1151)	0.261***	0.237***
Sent	*Salmonella enterica*	14	986 / 3655	(591 / 2021)	0.235***	0.198***
Lisp	*Listeria* spp.	8	781 / 2264	(645 / 1615)	0.234***	0.308***
Ecol	*Escherichia coli*	35	1872 / 4265	(750 / 1326)	0.303***	0.265***
Bant	*Bacillus* spp.	17	1864 / 4529	(869 / 1714)	0.358***	0.168***
Hpyl	*Helicobacter pylori*	14	856 / 1759	(650 / 955)	0.495***	nd

Table 3.4: **Correlation of gene population sizes and GC%**
** : p-value < 10^{-2}, ***: p-value < 10^{-3}, nd: non determined.

3.2.2 A complex interplay of mutation, selection and recombination

In more detail, the models predicts the neutral fixation of G/C alleles to be enhanced in genes with a large population. This fixation bias is however expected to be counteracted by selection, notably on the sequence of proteins coded by genes. It implies that non-synonymous sites are expected to be less marked by the effect of gBGC than synonymous sites, as seen for many mutational patterns (Sueoka, 1988; Lawrence and Ochman, 1997; Daubin and Perrière, 2003). This dichotomy of coding sequences is often approximated by the oppostion of GC-content at the third codon position, mostly synonymous (GC3) to GC-contents at the first and second codon positions, always and mostly non-synonymous, respectively (GC1 and GC2). GC3 is thus expected to display a higher correlation with gene frequency than GC1 and GC2 would do.

Surprisingly, it is not always the case. As expected, in the species *E. coli*, which has an intermediate genomic GC-content, the GC3 accounts for most of the total GC-content variation with frequency. An opposite pattern is seen in the low-GC species *Staphylococcus aureus* (Fig. 3.5). In the latter species, GC3 is negatively associated to the frequency of genes, in opposition to GC1 and GC2. This reversal of trends seems to be associated with the genomic GC-content of genomes, as in the extremely GC-rich genomes of the *Burkholderia* Paseudomalei group, the GC3 is strongly increasing with population size. One could expect that when the effect of the mutational bias is stronger than the effect of gBGC, as it is possible in a genome that is AT-rich and moderately recombining like in *S. aureus*, GC3 would not be correlated to frequency. However, a corollary under our model would be that non-synonymous sites, that are more constrained, should display an even lower correlation than synonymous positions. Fig. 3.5 shows that is not the case, as in *S. aureus*, GC1 and GC2 covary positively with gene population size. The pattern is complex and it is very difficult to find an unifying process that would explain the trends observed in the different tested clades (Fig. 3.6). Hence, the intra-genomic variations of GC-content cannot be explained under our gene population model, at least not with gBGC as a major driving force of GC-content.

This model may need to include other forces, notably selection for codon usage and for proteome composition, but as their respective effect can be confounding or opposing to that of gBGC depending on genome characteristics – the genomic GC-content, the composition of the set of optimal codons, etc. – it is impossible to interpret the observed variations of GC1,2,3 out of the verbal model proposed here. Explicit modelling of forces in action, estimation of their relevant parameters and test of models with increasing complexity may answer this conjecture. This could be done by building upon already existing coalescent models of bacterial genome evolution that consider homologous recombination throughgout the history of core genes (Didelot et al., 2010), or models already integrating the evolution of core and accessory genes in pangenomes (Szöllősi et al., 2012; Szöllosi et al., 2013; Bansal et al., 2012).

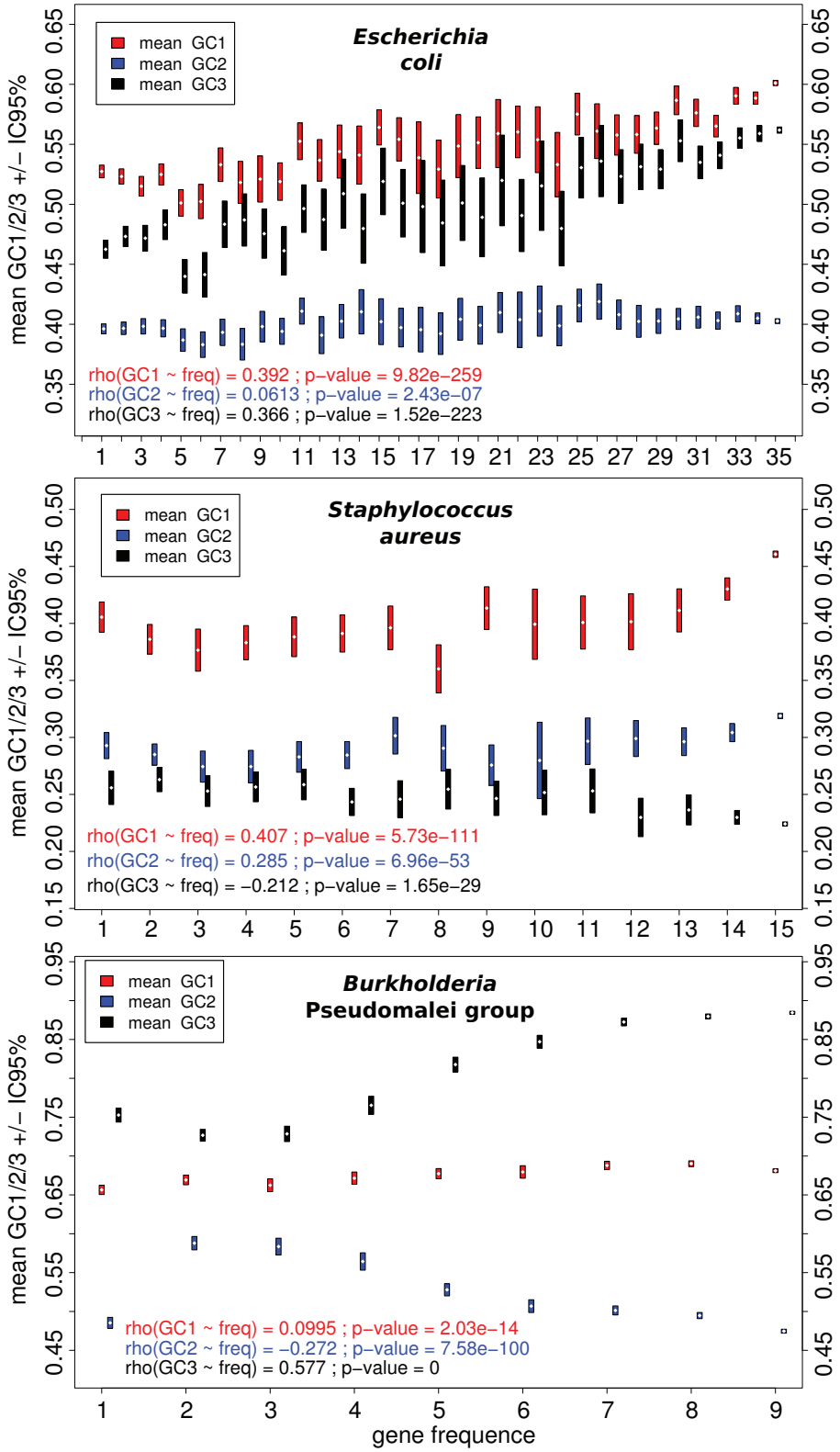


Figure 3.5: **Variations of gene GC1,2,3 with their frequency in the pangenome.** Focus on three representative plots: medium-GC *E. coli*, low-GC *S. aureus* and high-GC *B.* Pseudomalei group.

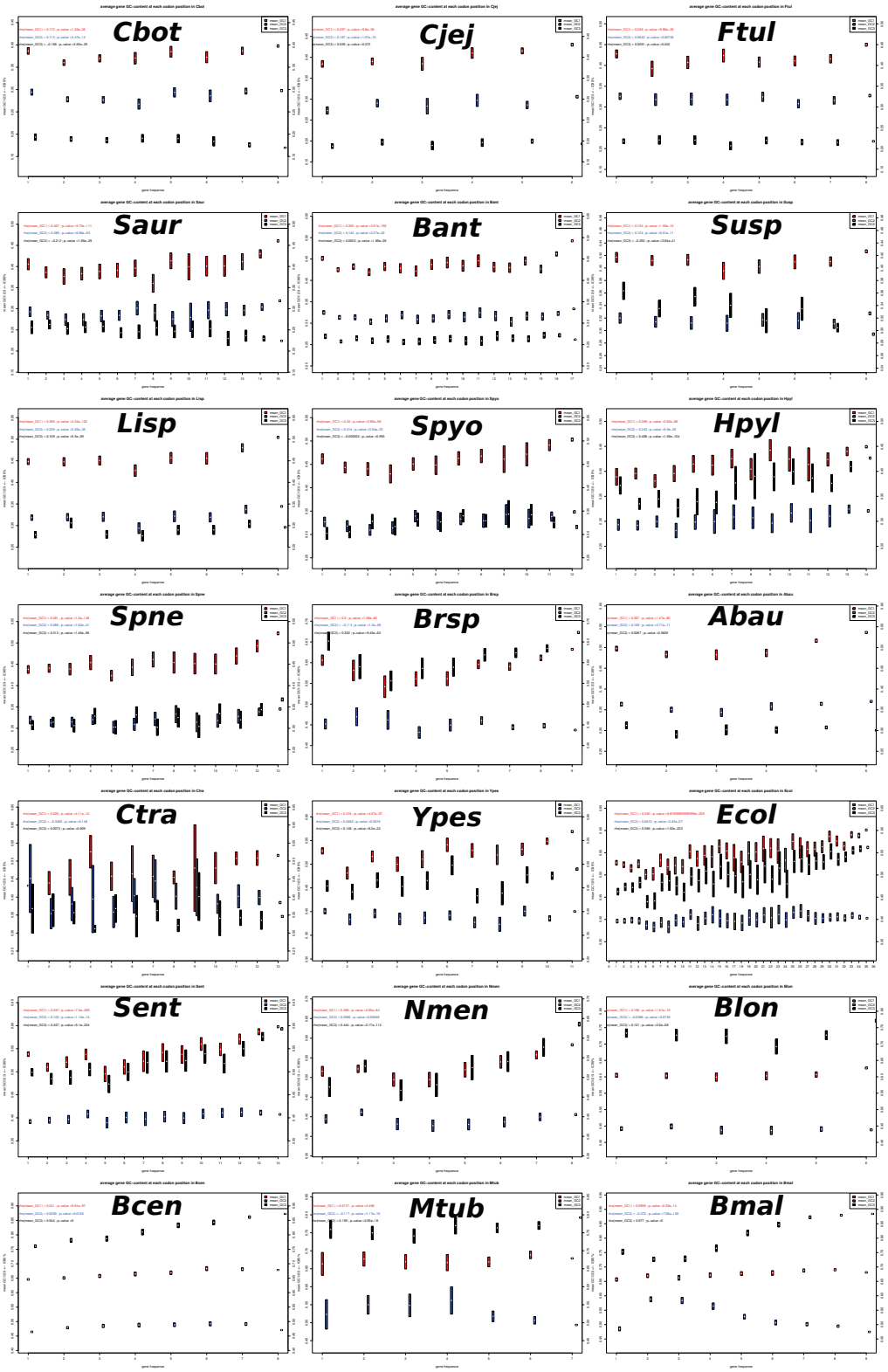


Figure 3.6: **Variations of gene GC1,2,3 with their frequency in the pangenome.**
All tested datasets are represented classified from left to right and then top to bottom by increasing genomic GC-content. Acronyms are as in table 3.4.

3.3 Validation of results in *A. tumefaciens*

A. tumefaciens genome dataset used in section 2.4 was not publicly released at the time of submission of the article presented section 3.1 and for this reason it was not included to the set of bacterial genome dataset used in the present chapter. However, it is interesting to see if the results we obtained in this chapter for a diverse set of bacterial clades – support for the existence of GC-biased gene conversion and interference of this process with selection for optimal codon usage – were also valid for *A. tumefaciens*.

The effect of recombination on genome composition was thus investigated in *A. tumefaciens* complex core-genome and in the dataset restricted to the genomic species for which we had the more genomes available, genomovar G1. For the the complete species complex, more than half core families were detected as recombining (1503 rec vs. 1288 norec families), in agreement with the frequent homologous recombination observed in the whole taxon (Fig. 2.10). However, only a small fraction of genomovar G1 core families were seen as recombining (368 rec vs. 3197 norec families), suggesting the test used here failed to recover the intense homologous recombination history observed for genomovar G1 with phylogenetic reconcilition methods (Fig. 2.10), probably because of the low sequence diversity of the taxon (Fig. 2.3) or the small size of the sample ($n_{genomes} = 5$).

Genomic GC-content was positively biased in recombining gene families of both core-genomes of *A. tumefaciens* complex and genomovar G1 (Fig. 3.7 A). This suggests gBGC is at work in *A. tumefaciens* as well, and its effect was evident even in genomovar G1 dataset where we had little sensitivity to detect recombination. In addition, GC-ending and AU-ending optimal codons showed opposite variations of frequency with recombination in *A. tumefaciens* complex, confirming that gBGC is acting and conflicts with selection for an optimal usage of codons (Fig. 3.7 B). This is however not the case in genomovar G1, where both GC- and AU-ending optimal codons were depleted in recombining families; this is consistent nor with a gBGC model, neither with a model of recombination enhancing selection for optimal codons, and rather suggests that the classification of recombining vs. non-recombining families in this subset of data was not accurate.

Finally, variations of GC-content at the three codon positions with the frequency of genes in the pangenome of *A. tumefaciens* complex displayed trends intermediate to what was observed for *E. coli* and *Burkholderia* Pseudomalei group, i.e. positive correlation of GC1 and GC3, but low negative correlation of GC2 with gene frequency, consistently with the intermediate GC-content of *A. tumefaciens*.

Hence, *A. tumefaciens* adds another stone to the demonstration of the existence of gBGC in Bacteria. Moreover, it shows that its effects can be characterized when looking at a quite diverse clade: *A. tumefaciens* is more diverse than any other dataset tested in section 3.1 and displays substantial variations of GC-content between species – genomovar G1 has 0.9% less G/C in total (1.8% at the third codon position) than the average of all *A. tumefaciens* (Fig. 3.7). This

introduces the possibility of integration of the predictions of a gBGC model to the observation made previously on the histories of genes in genomes of *A. tumefaciens*.

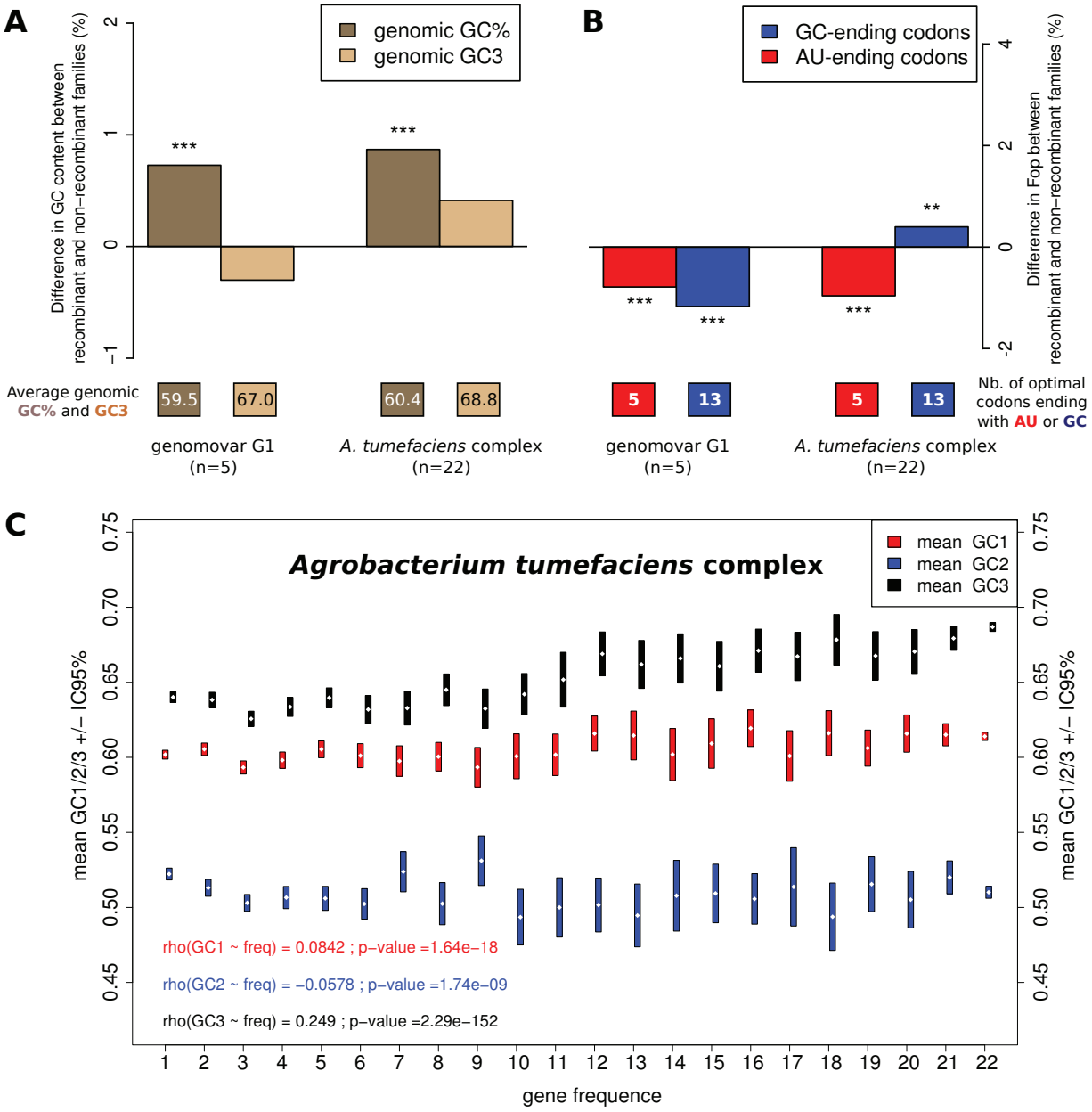


Figure 3.7: **Evidence of gBGC in *A. tumefaciens* species complex.**
(A,B) Effect of recombination on core gene families. Recombinant (rec) and non-recombinant (norec) sets of core-genome families are compared: 368 rec vs. 3197 norec families in genomovar G1; 1503 rec vs. 1288 norec families in *A. tumefaciens* complex.
(A) Effect of recombination on core gene family G+C content. Legend as in Fig. 3.1.
(B) Effect of recombination on core gene family codon usage bias. Legend as in Fig. 3.2.
(C) Variations of gene GC1,2,3 with their frequency in the pangenome.

4

Final Discussion & Perspectives

In this work, we took two different approaches to characterize the role of genetic exchange in the form of horizontal gene transfer (HGT) and homologous recombination (HR) in the diversification of bacterial taxa, and their potential impact on the fitness of organisms in their environment. We showed that along the history of genomes, HGT was providing a large source of genotypic innovation for emerging bacterial clades, and that the interplay of selection with this constant gene input was durably shaping the gene content of genomes. In addition, we showed that homologous recombination was occurring at various global rates amongst replicons of a genome, certainly causing differences of plasticity of their gene content, and incidentally causing differences in the selective pressures experienced by the genes they bear. Finally we explored the hypothesis of HR being responsible for the neutral increase of GC-content in bacterial genomes, and showed it could mimic or interfere with selection for an optimal codon usage.

Here, we propose to revisit, in the context of the results presented above, a still pending question of evolutionary microbiology: what are the cause and consequences of the intra-genomic heterogeneity of nucleotide composition?

In several bacterial taxa (Daubin et al., 2003; Daubin and Ochman, 2004a; Lefébure et al., 2010; Lassalle et al., 2011), it was observed that the distribution of frequencies of genes in the pangenome followed a gradient of GC-content, with the more recent genes being AT-rich and the more ancient being GC-rich. Most intriguingly, this bias has always the same polarity, irrespective of the richness of the genome in G/C nucleotides (Daubin et al., 2003).

We explored the capacity of gBGC to explain this feature in a model of gene populations. We found no clear answer, because different categories of sites – synonymous and non-synonymous sites – seem to evolve across the pangenome in a different manner depending on the studied taxa (section 3.2.2). This inter-genomic variations of compositional patterns predominantly points at differences of histories and lifestyles among these clades. This results in very heterogeneous selective pressures acting on genomes on the long term. This heterogeneity must be even stronger amongst genes, as we showed that genes of the pangenome had each unique histories and varied adaptive roles.

Indeed, essential genes – that are mostly parts of the central cell machinery – tend to evolve more slowly than ones with peripheral roles (Rocha and Danchin, 2004). Essential genes are almost always part of the core of genomes, whose conservation witnesses the purifying selection they experience. This certainly reflects the frequency at which selection operates on those genes: by definition, the inactivation of an essential gene is immediately lethal, and even subtle changes that would impact its expression can be rapidly counter-selected. In contrast, genes involved in the adaptation to an ecological niche are only subject to periodic selection, as specified in the 'stable ecotype' model (Cohan, 2001). Periodic does not necessarily mean weak, as episodes of competition for rare resources in oligotrophic environments may readily lead to the extinction of the less fit. In the interval, however, slightly deleterious mutations – like synonymous changes that induce non-optimal codon usage – can get fixed freely.

In addition, these selection regimes are not constant throughout the history of genes. Notably, those genes that were part of nomadic elements with effectively null adaptive values in most of the genomes they visit, can be domesticated by one genome and then undergo a drastic shift of selective pressure. This can even impact 'third-party' linked genes. The 'stable ecotype' model predicts that the complete genome can be swept with niche-specifying genes during periodic selection events (Cohan and Koeppel, 2008), what can lead to the fixation of many (slightly) deleterious linked mutations. A less strict model where not all the genome is linked can be considered. For instance, niche-specifying genes can be borne by mobile genetic elements such as plasmids. If a plasmid happen to be domesticated for its role in the ecological adaptation of an ecotype, as it was shown for *Yersinia pestis* (Chain et al., 2004) or in the present study with the pAt of *A. tumefaciens* genomic species, periodic selection on niche-specifying genes can enforce conservation of the plasmid, but let other genes borne by the plasmid to evolve under little constrain, as seen for non-conserved parts of the pAt of G8, which consist mostly of strain-specific genes.

All these mechanism that induce to the relaxation of selective constrains on gene sequence lead the gene to accumulate mutations, whose nature is biased. Mutational biases can occur genome-wide (Sueoka, 1988), on one of the DNA strand Lobry and Sueoka (2002), and in complex interdependency between sites (Sung W., Lynch M. et al., unpublished). These biases could be genome-specific (Sueoka, 1988) and thus the successive dwelling of genes in genomes is expected to yield a complex pattern. However, mutational biases were recently shown to be similar amongst bacterial taxa in that they are all directed towards AT-richness(Hershberg and

Petrov, 2010; Hildebrand et al., 2010). It follows that any relaxation of selective constraints on a gene sequence would theoretically result in enrichment in A/T nucleotides, which is the same outcome than the reduction in the effective recombination rate under a gBGC model, revealing the difficulty to decipher the role of each process.

Such historical variations of selective pressures coupled to variations of mutational pressures in different genomic context cannot be untangled so easily. Given this abundance of interplaying factors that can impact gene GC-content, it is almost surprising we could even bring out the association of GC-content with recombination in prokaryotic core genomes (section 3.1). However, as we mentioned above, core and essential genes tend to have more stable history of selection and to be less subject to transfers (Pál et al., 2005), what could lead to a relative stability of the mutational and substitutional processes to which they are subject.

Can we define other categories of genes that would undergo coherent patterns of mutation and substitutions? It was pointed out that replicons of the same genome could present different patterns of codon usage and/or nucleotide composition (Harrison et al., 2010). As mentioned by Harrison et al. (2010), in *A. tumefaciens* as in many other bacteria bearing a chromid, the overall GC-content and codon usage of the chromosome are similar to those of the chromid, and are nearly identical when accounting for the presence of genomic islands on the chromid. This would indicate similar global mutational and selective pressures acting on both replicons. However, we showed that the circular chromosome and the linear chromid differed markedly by their homologous recombination frequency, and in their dynamics of gene content. How can the linear chromid maintain an GC-content equal to that of the chromosome while recombining more – and thus putatively accumulating G/C alleles via gBGC – and housing more transient genes which are subject to low selective pressures?

Such global comparison of molecules may thus be inappropriate, because of many confounding factors: the circular chromosome houses the majority of the core genome (Lassalle et al., 2011), among which essential genes under constant selection, while the linear chromid houses the majority of clade-specific genes (Lassalle et al. (2011) and section 2.4.2.8), which might rather be under periodic selection. The existence of supra-genic trends in the evolution of nucleotide composition of genomes should rather be sought a smaller scales, looking for genes, like genomic islands (GIs), that co-evolved long enough to acquire common composition. Indeed, it has long been known that GIs have remarkable composition, and are often identified by this criterion. However, many such islands might exist without apparent composition bias, because the succession of evolutionary event that built them up in several different genomes might have left nothing but a heterogeneous composition as a signature.

In this perspective, the present works provides an significant progress, because it allows to recognize blocks of co-evolved genes independently of their composition. Indeed, it was shown previously that methods using anomalous composition criteria to detect HGT could be mistaken quite often Ragan (2001), and this might stem from the different selective pressures accompanying the particular mode of evolution of GIs. In addition, these intrinsic methods fail to identify ancient transfers, simply because gene sequences ameliorate relatively rapidly

to their host genome (Lawrence and Ochman, 1997). In the present work, we could efficiently identify genes that were brought together in genomes and link this to the past evolutionary context of each gene, as represented in their respective reconciled gene trees. In this framework it might be possible to reconstruct the succession of mutational and selective pressures that occurred along the individual and common histories of genes. Notably, new methods for the reconstruction of the substitutional processes long trees under complex models of molecular evolution could be of great help (Romiguier et al., 2012).

Altogether, the intermingling of historical, ecological and neutral causes in the determination of the evolution of gene sequences claims the necessity of defining new frameworks of genome analysis that combine all these aspects. The reconstruction of the history of genomes at the gene level to infer the multiple events that shaped them must integrate models of population genetics that would describe the relative contribution of population-level processes such as drift, selection and gBGC in driving the evolution of genes. Such global models should be the key to the understanding of the inner workings of genomes and of the private life of prokaryotes.

5 Annexes

5.1 Ecophysiology of the arsenite-oxidizing bacterium *Rhizobium* sp. NT-26

5.1.1 Introduction

The following manuscript was published in 2013 to the journal *Genome Biology and Evolution*. It follows a long-term collaboration with research teams in Strasbourg, France and London, UK, that started with internship in these laboratoris during may Master degree.

This work present the multi-omic characterization and physiological analysis of the arsenite-oxidizing bacterium *Rhizobium* sp. This bacterium, though related to rhizosphere commensals of the genus *Agrobacterium*, was found in the nutrient-poor and extremely toxic habitat of anancient gold mine contaminated by arsenic, where it appears particularly well adapted. Indeed, strain NT-26 is extremely resistant to many heavy-metals and metaloids, including the different arsenic oxides, that are prevalent in this environment. In addition, this strain is chemoautolithotroph via the oxidation of arsenite and thiosulphate, both minerals that derive from the solubilization of the environing arsenopyrite [FeAsS] rock.

These adaptations are supported by many specific gene gains linked to an extreme genomic differentiation compared to its rhizobial ancestor, notably through the replacement of the secondary replicon that is typical of the family by another megaplasmid that bears many heavy-metal resistance determinants. Strain NT-26 however still displays ecological characteristics of its genus, as it can promote plant growth.

The emergence of such an extremely derived lineage was driven by positive selection for resistance to the environing toxics, and then their trophic exploitation. This adaptive path might have been triggered by the acquisition of an arsenite-oxidase operon cassette which is commonly present at low frequency in Rhizobiaceae, what would have conferred a pre-adaptation to the arsenic-rich environment.

My personal contribution in this work consists in the phylogenetic comparison of the strains NT-26 with other sequenced Rhizobiaceae in order to establish its taxonomical status, and the analysis of the evolution of its genome architecture.

5.1.2 Manuscript

Genome Biology and Evolution

Life in an Arsenic-Containing Gold Mine: Genome and Physiology of the Autotrophic Arsenite-Oxidizing Bacterium *Rhizobium* sp. NT-26

Jérémy Andres[1], Florence Arsène-Ploetze[1], Valérie Barbe[2], Céline Brochier-Armanet[3], Jessica Cleiss-Arnold[1], Jean-Yves Coppée[4], Marie-Agnès Dillies[4], Lucie Geist[1], Aurélie Joublin[1], Sandrine Koechler[1], Florent Lassalle[3,5,6], Marie Marchal[1], Claudine Médigue[7], Daniel Muller[6], Xavier Nesme[6], Frédéric Plewniak[1], Caroline Proux[4], Martha Helena Ramírez-Bahena[6,8], Chantal Schenowitz[2], Odile Sismeiro[4], David Vallenet[7], Joanne M. Santini[5,*], and Philippe N. Bertin[1,*]

[1]Laboratoire Génétique Moléculaire, Génomique et Microbiologie, UMR7156 CNRS Université de Strasbourg, Strasbourg, France

[2]Laboratoire de Finition, CEA-IG-Genoscope, Evry, France

[3]Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive, UMR5558, Villeurbanne, France

[4]Plate-forme Technologique Transcriptome et Epigénome, Institut Pasteur, Paris, France

[5]Institute of Structural and Molecular Biology, University College London, United Kingdom

[6]Université de Lyon, Université Lyon 1, CNRS, INRA, Laboratoire Ecologie Microbienne Lyon, UMR5557, USC1193, Villeurbanne, France

[7]Laboratoire Analyses Bioinformatiques pour la Génomique et le Métabolisme, Genoscope-IG-CEA, Evry, France

[8]Instituto de Recursos Naturales y Agrobiologia, IRNASA-CSIC, Salamanca, Spain

*Corresponding authors: E-mail: philippe.bertin@unistra.fr; j.santini@ucl.ac.uk.

Accepted: April 9, 2013

Data deposition: This project has been deposited at NCBI and ArrayExpress under the accession numbers 1125847 and E-MEXP-3021, respectively.

Abstract

Arsenic is widespread in the environment and its presence is a result of natural or anthropogenic activities. Microbes have developed different mechanisms to deal with toxic compounds such as arsenic and this is to resist or metabolize the compound. Here, we present the first reference set of genomic, transcriptomic and proteomic data of an *Alphaproteobacterium* isolated from an arsenic-containing goldmine: *Rhizobium* sp. NT-26. Although phylogenetically related to the plant-associated bacteria, this organism has lost the major colonizing capabilities needed for symbiosis with legumes. In contrast, the genome of *Rhizobium* sp. NT-26 comprises a megaplasmid containing the various genes, which enable it to metabolize arsenite. Remarkably, although the genes required for arsenite oxidation and flagellar motility/biofilm formation are carried by the megaplasmid and the chromosome, respectively, a coordinate regulation of these two mechanisms was observed. Taken together, these processes illustrate the impact environmental pressure can have on the evolution of bacterial genomes, improving the fitness of bacterial strains by the acquisition of novel functions.

Key words: arsenic metabolism, motility/biofilm, *Rhizobium/Agrobacterium*, transcriptomics/proteomics, phylogeny, rhizosphere.

Introduction

To deal with high concentrations of toxic metals, micro-organisms have evolved various strategies, which enable them to detoxify their environment. These processes involve physicochemical reactions, for example, precipitation or solubilization, adsorption or desorption (Borch et al. 2010), and metabolic oxido-reduction reactions (Gadd 2010). In addition, most of the metallic elements found in the periodic table may play a crucial role in microbial physiology, for example, as components of metalloproteins, or as electron

donors or acceptors in energy metabolism (Stolz 2011). Such a metabolism may have been important in the early stages of life, due to a high concentration of metals, including arsenic, in the primordial planet (reviewed in van Lis et al. 2012).

In recent years, the various "omics" methods, which include genome sequencing, comparative genomics, and transcriptome or proteome analysis, have allowed to address the physiology of organisms in a global way. Such approaches have therefore greatly improved the understanding of microbial metabolism (Bertin et al. 2008; Holmes et al. 2009; Wilkins et al. 2009), including the global functioning of ecosystems, as recently demonstrated for an arsenic-rich microbial community (Bertin et al. 2011). To date, the genomes of more than 20 arsenic-metabolizing strains have been sequenced. They originate from various environments, belong to unrelated taxonomic groups, and have different carbon and energy requirements (reviewed in Bertin et al. 2012; van Lis et al. 2013).

Regarding arsenic, which is mainly present in two oxidation states in aquatic environments, that is, arsenite [As(III)] and arsenate [As(V)], microorganisms have acquired various metabolic capacities. These include As(V) reduction, which is usually part of the resistance mechanism, but also functions involved in As(III) oxidation or methylation (reviewed in Stolz 2011). Unlike arsenite methyltransferase genes, which are not often found in bacterial genomes, genes encoding arsenite oxidase are widespread in Bacteria and Archaea (Heinrich-Salmeron et al. 2011 and reviewed in Osborne and Santini 2012). In *Herminiimonas arsenicoxidans*, the arsenite oxidase *aioBA* genes are located in an arsenic genomic island, which also contains genes involved in arsenic resistance and biosynthesis of a molybdenum cofactor of the Aio enzyme (Muller et al. 2007). Such a genetic organization has also been observed in *Thiomonas arsenitoxydans* (Arsène-Ploetze et al. 2010; Bertin et al. 2012) and the presence of *aio* genes on a plasmid has been reported in *Nitrobacter hamburgensis* and *Thermus thermophilus* str. HB8 (Bertin et al. 2012). These observations suggest that *aioBA* genes may be acquired by horizontal gene transfer.

The adaptative response to arsenic has been recently shown as occurring in two steps (Cleiss-Arnold et al. 2010). First, bacterial cells express various genes involved in defence mechanisms, for example, oxidative stress and arsenic efflux. Next, several metabolic activities are induced, including arsenite oxidation which, in heterotrophic bacteria like *H. arsenicoxydans*, may be principally considered as a detoxification mechanism (Muller et al. 2007). In contrast, in bacteria that can grow autotrophically such as *T. arsenitoxydans* (Arsène-Ploetze et al. 2010) arsenite oxidation is part of a bioenergetic mechanism involved in energy generation. Despite some similarities, the genome organization of these two bacteria and their arsenic response, including biofilm formation, have been shown to differ markedly (Marchal et al. 2010, 2011).

In their natural environment, bacteria usually grow in biofilms, which are structured microbial communities embedded in extracellular polymeric substances (EPS) composed of sugars, proteins, and DNA (Hall-Stoodley et al. 2004; McDougald et al. 2012). Even though biofilm formation can be a problem in the field of human health, it allows bacteria to survive and thrive in highly toxic environments, including those characterized by high concentrations of heavy metals or metalloids such as arsenic (Guibaud et al. 2006; Muller et al. 2007). Unlike *H. arsenicoxydans* (Marchal et al. 2010), *T. arsenitoxydans* has been shown to induce biofilm formation in the presence of As(III) (Marchal et al. 2011). In addition, after biofilm development, the induction of cell motility has led to accelerated cell dispersion, an important process in the colonization of alternative ecological niches.

To gain a better understanding of the genetic determinants involved in the metabolism of arsenic, we have investigated the response to As(III) in *Rhizobium* sp. NT-26, a motile, chemolithoautotrophic arsenite oxidizer isolated from a gold mine in Australia (Santini et al. 2000). This strain belongs to the *Rhizobiaceae* family of the *Alphaproteobacteria*, which includes many species living in association with plants, such as plant mutualists of the *Rhizobium* and *Ensifer* (formerly *Sinorhizobium*) genera (Martens et al. 2007) and plant pathogens or plant growth-promoting rhizobacteria of the *Agrobacterium* genus (Hao et al. 2011). The *Rhizobium* sp. NT-26 genome was sequenced and annotated, and its physiology was investigated using differential transcriptomics and proteomics, and random mutagenesis. Remarkably, the synthesis of flagella was shown to be controlled by arsenite, suggesting a possible coordinate regulation between clusters located on two genetic elements. Indeed, proteins involved in As(III) oxidation were shown to be encoded by genes present on a megaplasmid, whereas flagellar genes are located on the chromosome.

Materials and Methods

Bacterial Strains, Plasmid, and Growth Conditions

Rhizobium sp. NT-26 and its mutant strains were cultivated at 28 °C in minimal salts medium (MSM) containing 0.04% yeast extract (Santini et al. 2000) and supplemented with As(III) and agar when required. *Escherichia coli* S17.1 λ*pir* was cultivated at 28 °C in Luria–Bertani (LB) (MP Biomedicals) medium supplemented with 20 mg/l kanamycin (Sigma) for the maintenance of the pTGN/mini-Tn*5* *gfp-km* plasmid (Tang et al. 1999).

Random Mutagenesis and Screening

Using the suicide vector pTGN carrying the mini-Tn*5* transposon, random mutagenesis was performed to construct a mutant library and to identify genes involved either in arsenite oxidation or in motility. Mobilization of the plasmid was performed using *E. coli* S17.1 λ*pir* carrying plasmid pTGN as the

donor and *Rhizobium* sp. NT-26 as the recipient. For conjugation, both strains in exponential phase, respectively, corresponding to an optical density (OD) of 0.6 and 0.135 at 600 nm, were superposed on LB plates at 28 °C for 24 h. As the *Rhizobium* sp. NT-26 strain used in this study is rifampicin resistant (Santini and vanden Hoven 2004), mutants were then selected on LB plates supplemented with 20 mg/l kanamycin and rifampicin.

Colonies from the library were screened for the loss of arsenite oxidation or the loss of motility. Briefly, mutants were individually inoculated into 96-well microtiter plates containing MSM with 0.04% yeast extract and 8 mM As(III) and incubated at 28 °C in 1.5% agar for 48 h or 0.3% agar for 24 h, respectively. The library was screened in the following two ways: 1) the silver nitrate method was used to detect arsenite oxidation (Lett et al. 2001; Muller et al. 2003) and 2) the diameter of the swarming ring was used to determine whether the cells were motile (Muller et al. 2007). Each phenotype was subsequently confirmed on Petri dishes in the corresponding medium. Mutants unable to oxidize arsenite were also tested for motility and vice versa.

To identify the disrupted gene in each mutant, the genomic region close to the mini-Tn*5* insertion site was amplified by inverse polymerase chain reaction (PCR). Total DNA was extracted with the Wizard Genomic DNA purification kit according to the manufacturer's instructions (Promega). One microgram of DNA was digested with 50 U of restriction enzymes that do not cut the transposon sequence (ClaI or PstI) in a 50 µl reaction volume at 37 °C for 2 h. After precipitation by ethanol and sodium acetate, digested DNA was ligated with 10 U of DNA ligase (Fermentas) in a volume of 20 µl overnight at 16 °C. PCR was carried out on 25 ng of this template in a 25 µl volume reaction with iProof DNA Polymerase (Bio-Rad) and Oend (ACTTGTGTATAAGAGTCAG) and Iend (AGATCTGA TCAAGAGACAG) primers. The program used involved a denaturation step at 98 °C for 30 s, followed by 35 cycles of denaturation at 98 °C for 10 s, annealing at 52 °C for 30 s and elongation at 72 °C for 3 min, and a final elongation step at 72 °C for 10 min. Amplification products were checked on an agarose gel and sequenced with Oend by MilleGen (http://www.millegen.com/, last accessed April 30, 2013). The Blastn tool on the MaGe interface (Vallenet et al. 2006) was used to align the sequences with that of the *Rhizobium* sp. NT-26 genome allowing for identification of the disrupted gene. For each mutant, the precise insertion site and orientation of the mini-Tn*5* was determined by PCR, combining the Oend and Iend primers with new specific primers (supplementary table S1, Supplementary Material online) designed around each probable insertion site.

Biofilm Quantification

Biofilm formation by *Rhizobium* sp. NT-26 wild-type and mutant strains grown in the presence or absence of arsenite was measured by the crystal violet method. Cultures were grown in MSM medium containing 0.04% yeast extract supplemented with and without 8 mM As(III) and incubated at 28 °C overnight with shaking (120 rpm). The cultures were then diluted with fresh medium to an OD of 0.1 at 600 nm. Each strain was tested with six replicates of 200 µl in two flat-bottomed polystyrene 96-well microtiter plates (Nunc). Cultures were incubated at 28 °C for 24 h and 48 h without agitation. Biofilm formation was quantified using crystal violet, as previously described (Hommais et al. 2002). Briefly, after removing the culture medium, wells were gently rinsed three times with 0.1 M phosphate-buffered saline (PBS). To fix biofilms, plates were dried at 55 °C for 25 min, then 200 µl of 0.1% [w/v] crystal violet solution (Merck) was added to the wells and the plates were incubated at 30 °C for 30 min. Free crystal violet was removed and wells were washed three times with PBS. Plates were dried at room temperature and the biofilm was subsequently dissolved in 200 µl of 95% [v/v] ethanol over 30 min. Finally, the absorbance was read at 595 nm with a microplate reader (Synergy HT).

Pulsed-Field Gel Electrophoresis

Plasmid profiles were determined by a modified Eckhardt agarose gel electrophoresis technique, as described previously (Hynes and McGregor 1990). *Rhizobium* sp. NT-26 was grown in LB until an OD of 0.5 at 600 nm was reached, and 150 µl of culture were used per well. Electrophoresis was carried out at 4 °C, 5 V for 30 min and 85 V for 7 h on a 0.7% agarose gel containing 1% [w/v] sodium dodecyl sulfate (SDS) (Ramírez-Bahena et al. 2012). Plasmid size was estimated by comparison with those from *Agrobacterium tumefaciens* C58 (Wood et al. 2001).

Plant Trapping Tests

Nodulation experiments were performed under gnotobiotic conditions. Seeds of *Macroptilium atropurpureum, Vicia faba, Phaseolus vulgaris*, and *Pisum sativum* were surface sterilized for 2 min in 95% ethyl alcohol and then three times for 3 min in 1% sodium hypochlorite, each time washed with sterile water. Germination was carried out at 28 °C in dark conditions on glass plates covered with sterile filter paper moistened with sterile water. Pots with a capacity of 1.5 l were filled with sterile vermiculite, and 200 ml of nutrient sterile solution (Rigaud and Puppo 1975) was added per pot. Two seedlings were sown in each pot and plants were inoculated with a suspension of 10^5 CFU/ml 1 week after their transfer to hydroponic growth. *Rhizobium* sp. NT-26 was grown on YMB medium (Mannitol 0.7%, Yeast extract 0.2%, KH_2PO_4 0.02%, $MgSO_4$ 0.02%) and 1 ml of inoculum was applied to each seedling. Plants were regularly observed for nodule formation, and nodulation was quantified after inoculation as described in Gremaud and Harper (1989).

Genome Sequencing

The complete genome sequence of *Rhizobium* sp. NT-26 was obtained by combining Sanger and 454 sequencing methods. Sanger reads were obtained from a 10 kb insert library constructed after mechanical shearing of the genomic DNA and cloning of the generated inserts into the plasmid pCNS, as described previously (Muller et al. 2007). Plasmid DNA was purified and end-sequenced (26,888 reads) by dye-terminator chemistry with ABI3730 sequencers (Applied Biosystems) leading approximately to a 4× coverage. Reads were assembled by Newbler with around 20× coverage of 454 GS FLX reads (Roche) and validated via the Consed interface. Finishing steps were performed using primer walking of clones, PCR and in vitro transposition technology with the Template Generation System II Kit (Finnzyme), corresponding to 252, 32 and 8,404 additional reads, respectively. Approximately 70× coverage of 36 bp Illumina reads were mapped in the polishing phase, using SOAP (http://soap.genomics.org.cn/, last accessed April 30, 2013), as previously described (Aury et al. 2008).

Comparative Analysis of 24 *Rhizobiaceae* Genomes

The 23 genomes of *Rhizobiaceae* publicly available at the time of experiments (supplementary table S2, Supplementary Material online) were retrieved from ENA database (http://www.ebi.ac.uk/ena/, last accessed April 30, 2013). A homologous gene family database was built under the HOGENOM procedure (Penel et al. 2009) based on these 23 genomes and the one of *Rhizobium* sp. NT-26. Homologous protein sequences were aligned using MUSCLE (v.3.8.31, default parameters) (Edgar 2004) and then retro-translated with the pal2nal program (v.14) (Suyama et al. 2006). Nucleic acid alignments were restricted to conserved blocks with Gblocks (v.0.91b, minimum 50% of sequences in conserved and flank positions and all gaps allowed, codon mode) (Castresana 2000) and gene trees were computed from these alignments with PhyML (v.3.0, GTR + G8 + I model of evolution, best of SPR and NNI moves, SH-like branch supports) (Guindon and Gascuel 2003). All alignments and phylogenetic trees are shown in supplementary methods S1 and S2, Supplementary Material online. Replicon mapping and gene content comparison were done with custom Python scripts.

Species Phylogenies

The "core" set contained 822 gene families present in every 24 strains in only one copy. The "ribosomal" set contained 51 gene families whose products were annotated as "ribosomal protein" or related terms in at least one genome and were present in at least 22 strains. Full alignments of both family sets, and third codon-removed version of core family set were concatenated and used for species tree construction with RaxML (version 7.2.8-ALPHA, GTRCAT model with 50 categories, branch supports from 200 and 1,000 rapid bootstrap trees for "core" and "ribosomal" alignments, respectively) (Stamatakis 2006) (species trees are stored in supplementary methods S3, Supplementary Material online).

Tree Pattern Matching

Phylogenetic trees of gene families were searched for particular phylogenetic patterns, that is, subtrees with specific arrangement of relative branching leaves representing taxa, with TPMS software (Bigot et al. 2012): "NT26outAgro", that is *Rhizobium* sp. NT-26 as a direct outgroup of *Agrobacterium* genus; "NT26inAgro", that is *Rhizobium* sp. NT-26 as an ingroup of *Agrobacterium* and sister group of *A. tumefaciens*. Both searches were made first without considering branch support and then matching only with >0.9 SH-like branch support at nodes of interest (supplementary methods S4, Supplementary Material online).

Aio Phylogenies

Homologs of AioA, AioB, AioR, AioS, and AioX were retrieved from the nr database at the NCBI (http://www.ncbi.nlm.nih.gov/, last accessed April 30, 2013) using the BlastP program (Altschul et al. 1997) with the protein sequences of *Rhizobium* sp. NT-26 as queries and default parameters except the "Max target sequences" parameter which was set to 1,000. For each Aio protein, the 500 homologs displaying the highest similarity with the sequence of *Rhizobium* sp. NT-26 were retrieved and aligned using MAFFT (version 6, default parameters) (Katoh and Toh 2008). The resulting alignments were trimmed using BMGE (default parameters) (Criscuolo and Gribaldo 2010). Preliminary phylogenies were inferred using the Neighbor-Joining method implemented in SeaView (Poisson evolutionary distance) (Gouy et al. 2010). The robustness of the resulting trees was estimated with the nonparametric bootstrap procedure implemented in SeaView (100 replicates of the original alignments). Based on the resulting trees, the closest relatives of *Rhizobium* sp. NT-26 sequences were identified and used for more detailed phylogenetic analyses. The corresponding sequences were realigned and the resulting alignments trimmed using the same procedure. Final phylogenetic analyses were performed using the maximum likelihood and Bayesian approaches implemented in PhyML (version 3) (Guindon et al. 2009) and MrBayes (version 3.2) (Ronquist et al. 2012), respectively. PhyML was run with the LG evolutionary model (Le and Gascuel 2008) and a gamma distribution with four categories of substitution rates (Γ4) and an estimated alpha parameter. The robustness of the maximum likelihood trees was estimated by the nonparametric procedure implemented in PhyML (100 replicates of the original alignments). MrBayes was run with a mixed substitution model and a Γ4 distribution. Four chains were run in parallel for 1,000,000 generations. The first 2,000 generations were discarded as "burnin." The remaining trees were sampled every 100 generations to build the consensus tree.

Total RNA Extraction, Microarrays, and Data Analysis

A custom 15 K microarray with a probe length of 60 mer was manufactured by Agilent Technologies following the protocol used for *H. arsenicoxydans* (Weiss et al. 2009). Total RNA was extracted from *Rhizobium* sp. NT-26 strain grown heterotrophically in MSM containing 0.04% yeast extract in the absence and presence of 5.3 mM As(III) until late log phase (OD at 600 nm of 0.115 and 0.152, respectively) as described previously (Santini et al. 2007). RNA quality was checked using an Agilent Bioanalyzer. Ten micrograms of total RNA was reverse transcribed using the Fairplay III Microarray labeling kit (Agilent Technologies) and cDNA were indirectly labeled using Cy3 or Cy5 Mono reactive dyes (GE Healthcare). Labeled cDNA quality and quantity were determined by spectroscopy at 260, 280, 550, and 650 nm. The labeled Cy3 and Cy5 target quantities were adjusted to 250 pmol, mixed together and concentrated with Microcon YM-30 (Millipore). Hybridization was performed for 17 h at 65 °C. Three distinct biological RNA samples as well as dye swap experiments were performed for each culture condition. Arrays were scanned as described previously (Weiss et al. 2009). Data were acquired by Genepix Pro 6.0 (Axon Instrument) and statistically analyzed as described previously (Koechler et al. 2010). Genes having a BH adjusted *P* value lower than 0.05 were considered as differentially expressed between the two conditions and were retained for further study. Microarray data were deposited in ArrayExpress (E-MEXP-3021).

Preparation of Proteins Extracts and 2D Gel Electrophoresis

Experiments were performed with four protein extracts from four replicates for each growth condition. Strain NT-26 was grown heterotrophically in MSM containing 0.04% yeast extract in the absence or presence of 5.3 mM As(III). Exponential phase cultures were harvested by centrifugation at 6,000 × g for 10 min at 4 °C. Pellets were suspended in 400 µl of distilled water supplemented with 1 µl Benzonase Nuclease (Sigma) and 4 µl of Protease Inhibitor Mix (GE Healthcare). Cell suspensions were sonicated on ice with 10 pulses of 30 s at 28% of amplitude with 30 s intervals using a VC 750 sonicator (Bioblock Scientific). Cellular debris were removed by two centrifugations, the first at 6,000 × g for 5 min and the second at 16,000 × g for 90 min. Protein concentrations were measured using the Bradford method (Bradford 1976).

Differential accumulation of proteins was either monitored by Colloidal Brilliant Blue staining or DIGE (Marouga et al. 2005). For Colloidal Brilliant Blue staining experiments, 300 µg of protein extract were diluted to a final volume of 350 µl with rehydration buffer (8 M urea, 2% [w/v] CHAPS, 0.5% [v/v] IPG buffer pH 3-10, 40 mM DTT, and 0.01% [w/v] bromophenol blue). For DIGE experiments, 50 µg of protein was adjusted to pH 8.8 by adding 50 mM Tris–HCl final concentration, and either stained with 400 pmol of Cy3 or Cy5 (GE Healthcare). In addition, 25 µg of each 8 extracts (4 replicates for 2 conditions) were pooled and stained with 1,600 pmol of Cy2 to serve as an internal standard. For the staining, each CyDye DIGE fluor stock solution was diluted in high grade dimethylformamide to a final concentration of 400 pmol/µl. One microliter of the dilution was added to 50 µg of protein (Cy3 and Cy5) or 4 µl to 200 µg of the internal standard pool (Cy2) and kept in the dark and on ice for 30 min. The reaction was stopped by adding 1 µl of a 10 mM solution of lysine to 50 µg of protein (4 µl to 200 µg) and then 1 volume of 2× sample buffer (9 M urea, 3 M thiourea, 130 mM DTT, 4% [w/v] CHAPS, 2% [v/v] IPG buffer Pharmalyte 3-10) was added prior to incubation on ice for 10 min. One Cy3-labeled sample (condition 1) and one Cy5-labeled sample (condition 2) were mixed with one-fourth of the Cy2-labeled pool and rehydration buffer was added to reach a volume of 350 µl. Dye swap experiments were performed for each culture condition.

For protein separation, samples were first loaded onto an 18 cm pH 4–7 IPG strip. IEF was conducted using the Ettan IPGphor system (GE Healthcare), as previously described (Weiss et al. 2009). The strips were equilibrated in SDS equilibration buffer (30 mM Tris–HCl pH 8.8, 6 M urea, 34.5% [v/v] glycerol, 2% [w/v] SDS, 0.01% [w/v] bromophenol blue) supplemented with 1% [w/v] DTT for 15 min and then with 2.5% [w/v] iodoacetamide for 15 min. SDS polyacrylamide gel electrophoresis was subsequently performed using 11.5% SDS gels, using the Ettan DALT*six* system (GE Healthcare) with the following steps: 1 h at 60 mA, 80 V, 4 W and 1 h at 240 mA, 500 V, 52 W. Gels were stained with Colloidal Brilliant Blue or digitized using a Typhoon Scanner (GE Healthcare).

Differential protein expression analysis was performed as previously described (Bryan et al. 2009; Weiss et al. 2009). Spots were selected and identified by MALDI-TOF and Nano LC-MS/MS, and data analysis were performed with Mascot (Matrix Science Ltd.) as described previously (Bryan et al. 2009) against a *Rhizobium* sp. NT-26 protein database. All identifications were incorporated into the "InPact" proteomic database developed previously (http://inpact.u-strasbg.fr/~db/, last accessed April 30, 2013) (Bertin et al. 2008).

Transmission Electron Microscopy

Rhizobium sp. NT-26 or the *aioR* mutant were grown in MSM containing 0.04% yeast extract in the presence or absence of 8 mM As(III) for 24 h. A drop of culture was deposited onto Formvar-coated nickel grids and after cell decantation, the liquid excess was removed. Uranyl acetate 2% was added to negatively stain bacteria and flagella and these samples were dried. Grids were observed with a Hitachi H-600 transmission electron microscope (TEM) at 75 kV and photographed with a Hamamatsu ORCA-HR camera using the AMT software (Advanced Microscopy Techniques).

Results and Discussion

General Genome Features

Chromosome, Plasmids, and Genomic Plasticity

The *Rhizobium* sp. NT-26 genome includes a single 4.2 Mbp chromosome and two plasmids. The circular chromosome consists of 4,239,731 bp with 4,380 coding sequences, including 4,303 coding DNA sequences and 59 RNA genes, and representing 90.28% of the whole genome (fig. 1). Among these coding sequences (CDS), 34.40% are of unknown function.

The mean G + C content of the chromosome is 61.97% but its distribution is not homogenous (fig. 1), and the *Rhizobium* sp. NT-26 chromosome exhibits 65 regions of genomic plasticity (RGP, supplementary table S3, Supplementary Material online) (Vallenet et al. 2009) in comparison with that of *A. tumefaciens* 5A. The G + C content of these regions, their size (5–207 kb) and the codon adaptation index lower than the average are characteristic of genomic islands (GEI) (Juhas et al. 2009). Moreover, transposable elements and tRNA encoding genes are present in several of these regions, which further support the lateral transfer of these potential genomic islands (Daubin et al. 2003). Such genetic events are known to promote bacterial adaptation under environmental stresses by the acquisition of various capacities through horizontal gene transfer, an important mechanism of microbial genome evolution (Juhas et al. 2009). In agreement with this, more than 15 loci coding for metabolic functions that may improve the fitness of the strain to its environment were found among the 65 RGP identified in the genome of *Rhizobium* sp. NT-26. These include amino acids and carbon sources transport, inorganic carbon fixation, nitrogen metabolism, and sulfur oxidation (supplementary table S3, Supplementary Material online).

The genomes of bacteria in the *Agrobacterium* and *Rhizobium* genera are known to include several extrachromosomal replicons, which encode various functions required for the adaptation to specific niches (López-Guerrero et al. 2012). The *Rhizobium* sp. NT-26 genome comprises two plasmids, including a megaplasmid of 322,264 bp containing 367 coding sequences (CDS) (fig. 1). The presence and the size of the megaplasmid were confirmed experimentally by a modified Eckhardt gel electrophoresis method (supplementary fig. S1, Supplementary Material online). The second plasmid is 15,430 bp and more than half of its CDS encode proteins with unknown functions.

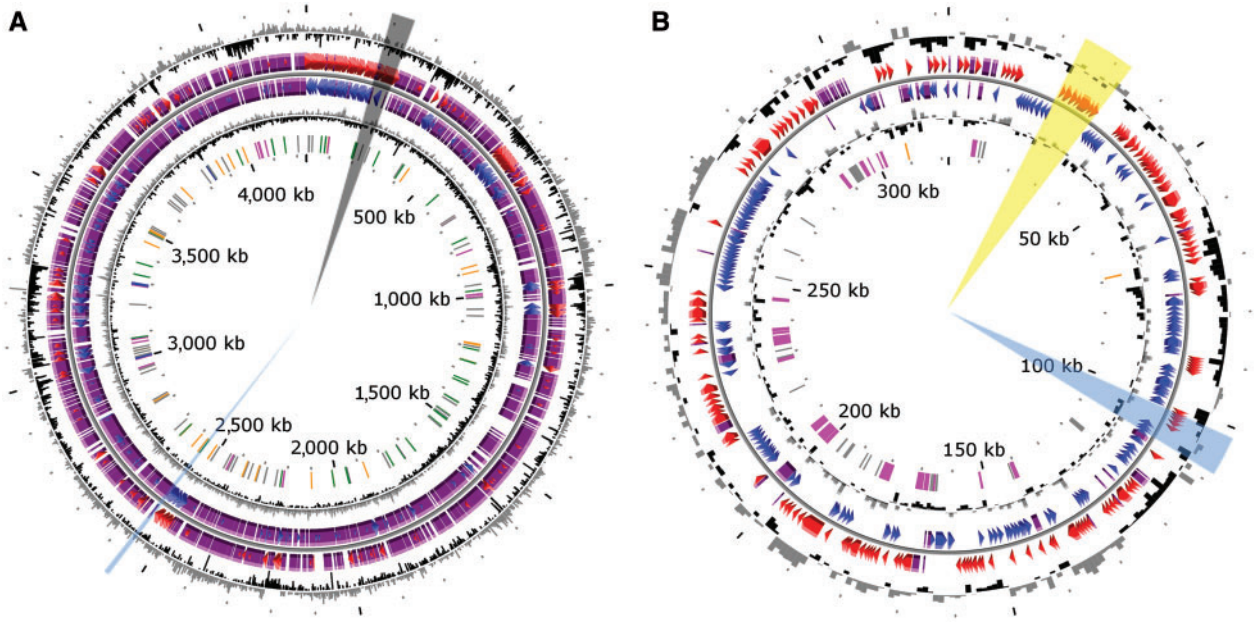


Fig. 1.—Circular representation of the *Rhizobium* sp. NT-26 genome. The chromosomal (*A*) and plasmidic (*B*) characteristics are 4,239-Mb long, 61.97% GC, 9 16S-23S-5S rRNA, 50 tRNA, and 4,294 CDSs; and 322-kb long, 60.19% GC, 0 16S-23S-5S rRNA, 0 tRNA, and 367 CDSs, respectively. From outside, circles display 1) the GC percent deviation in a 1,000 bp window (GC window – mean GC); 2) and 3) predicted CDSs transcribed in the clockwise and counterclockwise direction, respectively; red and blue colors correspond to validated annotations, orange to automatic annotation and purple to primary automatic annotation; 4) GC skew (G + C/G – C) in a 1,000 bp window; 5) rRNA are shown in blue, tRNA in green, miscRNA in orange, transposable elements in pink, and pseudogenes in gray. The regions with the genes coding for proteins involved in motility, reduction of arsenate or oxidation of arsenite are highlighted in black, blue, or yellow, respectively. The figure does not represent the p2 plasmid and the scale between the two genetic determinants is not respected (https://www.genoscope.cns.fr/agc/microscope/home/index.php, last accessed April 30, 2013).

The large plasmid harbors four different replication systems, in particular three *repABC* operons. The first one is duplicated. These two *repABC* operons are related to *repABC* of *Dinoroseobacter* (>90% identity) of the *Rhodobacteraceae,* a family of the *Rhodobacterales*. The third *repABC* operon is related to *repABC* of *Rhizobiales*. The fourth replication system is composed of a ParB/RepC replication system homolog of *A. tumefaciens* NCPPB925 plasmid origin of replication. Such a redundancy is not rare in *Rhizobiales*. For example, two replicons in *R. etli* CFN42, one in *R. leguminosarum* 3841 and one in *Ruegeria* sp. PR1b contain two *repABC* operons (Zhong et al. 2003; González et al. 2006; Young et al. 2006).

The smaller plasmid replication system is different from the canonical *repABC* replication system. It is constructed as the replication system described in pTAR of *A. vitis* (Gallie and Kado 1988), that is, the origin region carries a *repA*-like gene, a *parA* gene and a putative regulator locus coding for a putative segregation protein. Besides these replication transfer genes, the 15 kb plasmid harbors a toxin antitoxin system, which may explain its maintenance in *Rhizobium* sp. NT-26.

Plasmidic Adaptive Traits

In symbiotic bacteria, plasmids are known to play a role in their interaction with plants (López-Guerrero et al. 2012). In addition to multiple transposases and insertion sequences, the megaplasmid identified in *Rhizobium* sp. NT-26 encodes a putative type IV secretion system known to be involved in conjugal DNA transfer, including between bacteria and plants. Indeed, two complete *tra* clusters were found on the 322 kb plasmid: the first one is related to the type IV secretion system found in the *Rhizobium/Agrobacterium* genus, whereas the second one is related to the type IV secretion system of marine bacteria members of the *Rhodobacteraceae* family, that is, *Oceanibulbus indolifex* or *Ruegeria* sp. PR1b plasmid pSD25. However, canonical *nodABC* genes coding for proteins NodA (acetyl transferase), NodC (oligomerization of *N*-acetyl-glucosamine), and NodB (chitooligosaccharide deacatylase) that are required for the synthesis of the core structure of lipo-chitooligosaccharide (i.e., Nod factor) (Dénarié et al. 1996) were not identified in the genome of *Rhizobium* sp. NT-26. CDS displaying some similarities with other *nod* genes, that is, encoding enzymes that control specific substitutions on the chitooligosaccharide backbone, or the *fix* operon are present (supplementary table S4, Supplementary Material online), but these genes are also well conserved in nonsymbiotic prokaryotes. Nevertheless, it has been demonstrated that the nodulation of some *Fabaceae* by rhizobia occurs in the absence of the *nodABC* genes and lipo-chitooligosaccharidic Nod factors (Giraud et al. 2007). This indicates that other signaling strategies can trigger nodule organogenesis in some legumes. Nodulation assays on various *Fabaceae* plants were therefore performed as previously described (Gremaud and Harper 1989), but no nodules were observed at 3 weeks after inoculation or later at 4 weeks (data not shown).

Despite a lack of plant nodulation, root inoculation by *Rhizobium* sp. NT-26 suggested a potential phytobeneficial effect (fig. 2). Direct plant-growth promotion can be derived from phosphorus solubilization (Richardson et al. 2009), production of plant growth regulators (phytohormones) such as auxins, gibberellins, and cytokinins (Spaepen et al. 2009), NO production and/or by supplying biologically fixed nitrogen (Creus et al. 2005). Increasing the bioavailability of phosphate as micronutrient is mediated by bacterial phosphatase activity, and a phosphatase homolog, that is, NT26v4_0651, is present in the *Rhizobium* sp. NT-26 genome. Moreover, two main classes of dissimilatory nitrite reductase (Nir) involved in NO production exist among denitrifying bacteria: the heme-cytochrome cd_1 type encoded by *nirS* genes and the copper-containing type encoded by *nirK* genes (Zumft 1997). A *nirK* homolog but no *nirS* homolog was identified in the *Rhizobium* sp. NT-26 genome. Finally, no other homolog of classical phytobeneficial functions was identified when analyzing the genome, for example, phytohormone synthesis such as auxin by *ipdC/ppdC* or acetoin-2,3-butanediol by *budABC*, or nitrogen fixation by nitrogenase *nifHDK*.

Indirect plant growth-promoting mechanisms used by plant-growth-promoting rhizobacteria (PGPR) include induced systemic resistance, antibiotic protection against pathogens, reduction of iron availability in the rhizosphere by

Fig. 2.—Phytobeneficial effect of *Rhizobium* sp. NT-26 on *Phaseolus vulgaris*. Erlenmeyer flasks of *P. vulgaris* were inoculated with *Rhizobium* sp. NT-26 on the left, and with water on the right.

sequestration with siderophores, synthesis of fungal cell wall-lysing or lytic enzymes, and competition for nutrients and colonization sites with pathogens (Dobbelaere and Okon 2007). *Rhizobium* sp. NT-26 contains loci coding for polyketide synthases (NT26v4_3331, NT26v4_3332, and NT26v4_3333), which are involved in nonribosomal synthesis of antibiotics, or coding for proteins involved in siderophore transport (NT26v4_2008, NT26v4_4195, and NT26v4_4199). Taken together, these observations suggest that *Rhizobium* sp. NT-26 does not exert any direct interaction with plants but it may have an indirect role in plant growth and protection by its metabolic activities in the rhizosphere.

Unlike *H. arsenicoxydans* (Muller et al. 2007) and *T. arsenitoxydans* (Arsène-Ploetze et al. 2010), which metabolize and provide resistance to arsenic using proteins encoded by chromosomally borne genes, proteins involved in arsenic resistance in *Rhizobium* sp. NT-26 are encoded by *ars* genes present on both the chromosome and the megaplasmid. The *aio* genes involved in arsenite oxidation are present only on the megaplasmid (fig. 1*B*), as shown in *The. thermophilus* str. HB8. The *Rhizobium* sp. NT-26 *aio* cluster also contains genes coding for phosphate transport and molybdenum cofactor biosynthesis, as previously observed in other arsenite-oxidizing bacteria (Arsène-Ploetze et al. 2010; Bertin et al. 2011). In addition, like the metallo-resistant strain *Cupriavidus metallidurans* (Janssen et al. 2010), the *Rhizobium* sp. NT-26 megaplasmid contains numerous genes involved in resistance to heavy metals such as chromium, cadmium, and mercury. These observations suggest a loss of most plasmid-encoded functions known to be involved in bacteria–plant interactions and an acquisition of multiple genes allowing the organism to grow in its natural habitat, a goldmine known to contain toxic metals and metalloids.

The gene cluster coding for arsenite oxidase contains 5 *aio* genes in *Rhizobium* sp. NT-26. The survey of the nr database revealed the existence of numerous homologs of AioA, AioB, AioR, AioS, and AioX. Preliminary phylogenetic analyses of AioA homologs showed that the sequence from *Rhizobium* sp. NT-26 belongs to a well-supported clade of proteobacterial sequences corresponding to the groups I and II, which were recently described (Heinrich-Salmeron et al. 2011). Subsequent phylogenetic analyses revealed that the *Rhizobium* sp. NT-26 AioA branched among alphaproteobacterial sequences (group I), within a strongly supported clade composed of sequences from various *Agrobacterium* species and *Sinorhizobium* sp. M14 (Rhizobiaceae), from *Ochrobacterium tritici* (Brucellaceae), and uncultured organisms (bootstrap value [BV] = 98% and posterior probability [PP] = 1.00, supplementary fig. S2*A*, Supplementary Material online). Phylogenetic analyses of other Aio proteins showed similar branching patterns (supplementary fig. S2*B–E*, Supplementary Material online). This suggests that the five *aio* genes have co-evolved, which is not entirely surprising given that they are functionally related and clustered together when present in a genome.

A careful examination of the taxonomic distribution within the subgroups (supplementary fig. S2, Supplementary Material online) revealed that only 2 *Rhizobium/Agrobacterium* complete genomes contain the *aio* genes although nearly 30 genome sequences are available at NCBI. In addition, the relationships among the sequences were not always in agreement within the phylogeny of species, for example, the grouping of a member of *Brucellaceae* within *Rhizobiaceae*. This strongly suggests that horizontal gene transfers may have played a role in the spread of *aio* genes among these species. The alternative hypothesis, that is the presence of *aio* genes in the common ancestor of the *Rhizobium/Agrobacterium* group followed by multiple independent gene losses during the diversification of this lineage, appears less likely and does not explain the discrepancies among the Aio phylogenies and the taxonomy. On the contrary, these inconsistencies may be easily explained by the initial acquisition of the *aio* genes by one member of the *Rhizobium/Agrobacterium* group followed by a few horizontal gene transfers to related species or strains. Such transfers may have been favored by the colocation of *aio* genes on genomes and their location on plasmids in a few strains (e.g., pSinA in *Sinorhizobium* sp. M14 and pAt5A for *A. tumefaciens* 5A).

Taxonomic Relationship of *Rhizobium* sp. NT-26 with Other *Rhizobiaceae*

Rhizobium sp. NT-26 strain has been previously assigned to the *Rhizobium* genus on the basis of 16S RNA sequence (Santini et al. 2000), but its phylogenetic relationship with other *Rhizobiaceae* remains quite unclear. Therefore, the sequence of the *Rhizobium* sp. NT-26 chromosome has been compared with the sequences present in the "Prokaryotic Genome DataBase" (PkGDB) (Vallenet et al. 2006) and RefSeq (NCBI Reference Sequences) data banks. The highest synteny conservation was observed with the *A. tumefaciens* C58 circular chromosome, which is 67.85% of the CDS in this genome share synteny with the chromosome of *Rhizobium* sp. NT-26 and the average size of the syntons is 8.4 CDS. This gene order conservation is higher than that observed with *Rhizobium* spp. strains (6.9–7.4), suggesting a closer evolutionary relationship of strain *Rhizobium* sp. NT-26 with the *Agrobacterium* lineage. To determine more precisely the taxonomic position of *Rhizobium* sp. NT-26 among *Rhizobiaceae*, we compared its genome with a set of 23 other sequenced genomes of this family in a phylogenetic framework. We computed maximum-likelihood (ML) trees based on the concatenated alignments of sequences of either all-homologous genes that are common to and unique in all strains (822 "core" genes) or genes of ribosomal proteins (51 "ribosomal" genes). Intriguingly, both data sets yielded phylogenies that agree on all major splits in the taxon, but not on the position

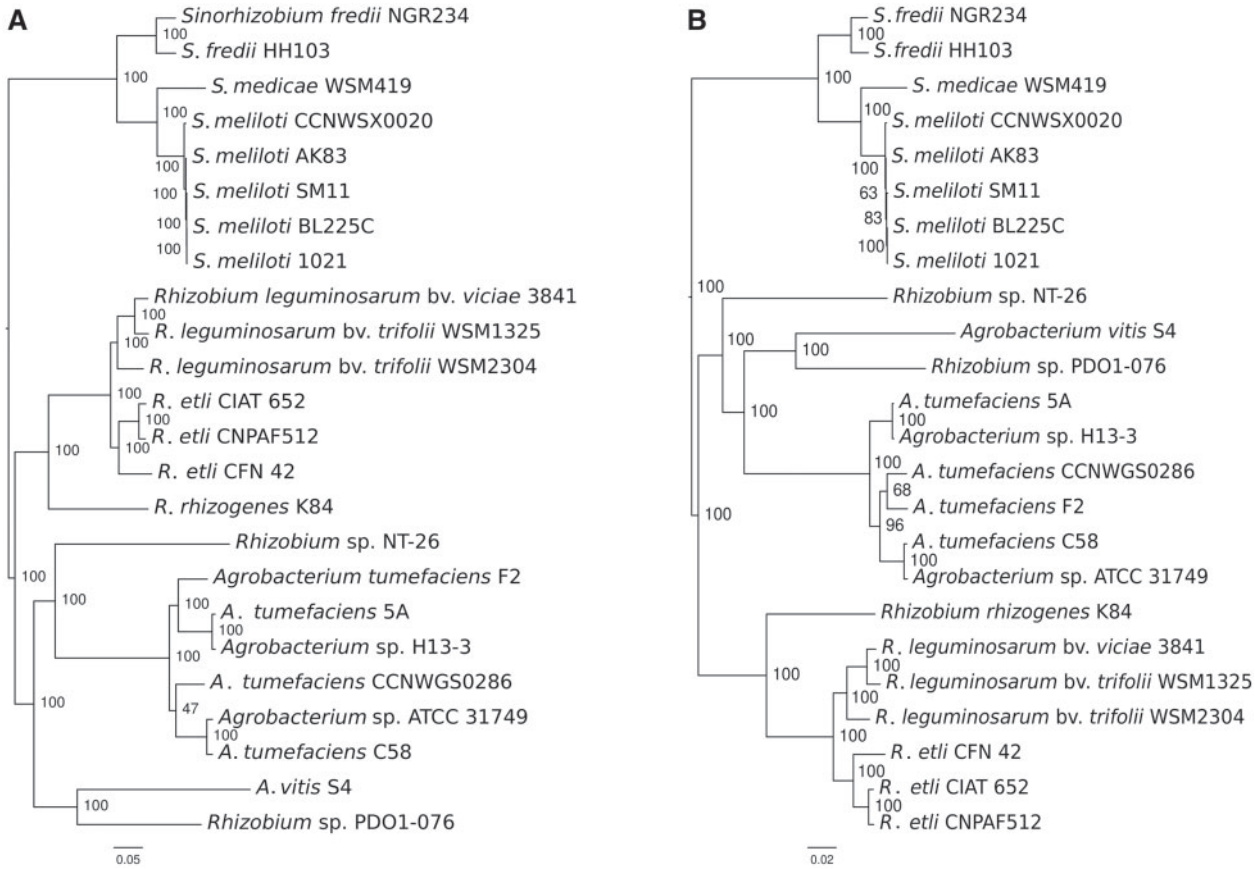


Fig. 3.—Phylogeny of *Rhizobium* sp. NT-26 among *Rhizobiaceae*. ML phylogenies of 24 *Rhizobiaceae* including strain NT-26 were built from concatenated alignments of (*A*) 822 core genes and (*B*) 51 ribosomal genes. Branch supports are percentage of the 200 and 1,000 bootstrap trees having the bipartition, respectively.

of *Rhizobium* sp. NT-26. According to core genes, this strain branched as an "in-group" of the *Agrobacterium* subgroup, being the brother clade of *A. tumefaciens* after the split with *A. vitis* (fig. 3*A*). Instead, according to ribosomal genes, strain NT-26 branched as an "out-group" of the *Agrobacterium* clade that encompasses *A. vitis* and *A. tumefaciens* (fig. 3*B*). In both cases, the conflicting bipartitions are well supported, and removing third codon positions in the alignment of core genes, because of the possible saturation of the substitution signal in non-housekeeping genes, did not change the observed pattern (supplementary methods S3, Supplementary Material online). This suggests that among the core gene set, which include the majority of the ribosomal gene set, different genes have different histories, causing the average history (core phylogeny) to be different from that of a subset (ribosomal phylogeny). This may have been caused by horizontal gene transfer to and from *A. tumefaciens*, *A. vitis*, *Rhizobium*, or other more phylogenetically distant taxon that would blur the signal for vertical inheritance.

We therefore computed individual phylogenies for all homologous families to determine what scenario each gene supported. On a set of 2,878 homologous gene family trees containing 3,537 *Rhizobium* sp. NT-26 genes, we searched for subtrees displaying the unambiguous patterns of either *Rhizobium* sp. NT-26 as a direct out-group of the *Agrobacterium* clade ("NT26outAgro") or as an in-group of the *Agrobacterium* clade and a brother group of *A. tumefaciens* ("NT26inAgro"). "NT26inAgro" was prevalent with 338 genes (268 considering only high branch support) versus 255 (146) for "NT26outAgro," and even though the amount of genes displaying such unambiguous patterns was relatively low, "NT26inAgro" was significantly more frequent (χ^2 test, P value $< 10^{-3}$). The location of those genes along the chromosome of *Rhizobium* sp. NT-26 showed no grouping with a particular pattern that could support a large-scale transfer event (Mann–Whitney–Wilcoxon test, P value >0.9; supplementary fig. S3, Supplementary Material online). The homogeneous dispersal of phylogenetic signatures rather suggests that numerous small-scale transfer events occurred, as observed in the case of frequent homologous recombination with partners of different taxa (Didelot et al. 2010). Alternatively, our observations may be the consequence of a

poor resolution of phylogenies. This may be due to the frequent artifacts in the phylogenetic reconstruction of the relationship of strain NT-26 to *Agrobacterium*, such as those caused by the long branch leading to *Rhizobium* sp. NT-26. In the future, more phylogenetic information might be provided by sampling strains branching at the base of the *Rhizobium/Agrobacterium* group.

Most of *Rhizobiaceae* contain a secondary chromosome or megaplasmids, generally referred to as chromids (Harrison et al. 2010), that are members of a same family of large replicons derived from a plasmid (Slater et al. 2009). Although *Rhizobium* sp. NT-26 genome contains a megaplasmid with chromid characteristics (Harrison et al. 2010), this megaplasmid show very limited homology with chromids of this family (supplementary table S5, Supplementary Material online). The absence of a typical *Rhizobiaceae* secondary replicon makes the genome structure of *Rhizobium* sp. NT-26 unusual when compared with other members of the family. Comparison of the homologous gene location in *Rhizobium* sp. NT-26 and related organisms may help with the understanding of its evolution history. With this aim, the closest homolog of each of its genes present in several strains of *Rhizobium* and *Agrobacterium* were mapped along the *Rhizobium* sp. NT-26 chromosome (supplementary fig. S4, Supplementary Material online). It appeared that the vast majority of strain NT-26 chromosomal genes map to the principal chromosome in *Rhizobium* and *A. vitis*, suggesting that the *Rhizobium* sp. NT-26 lineage has completely lost the secondary chromosome of the *Rhizobium/Agrobacterium* ancestor. The history of intragenomic translocations have been documented in this taxon (Slater et al. 2009) and locating *Rhizobium* sp. NT-26 genes whose homologs have migrated at a specific divergence time would help to date the age of the divergence of the *Rhizobium* sp. NT-26 lineage. In this respect, its chromosome possesses a large chromosomal fragment (spanning from 2.55 to 3.40 Mb), which is specifically present on the secondary (linear) chromosome in *A. tumefaciens* (supplementary fig. 3, Supplementary Material online), further supporting a divergence of the *Rhizobium* sp. NT-26 lineage predating the *A. tumefaciens* speciation and synapomorphic translocation events. Similarly, strain NT-26 appears to have conserved the majority of the genes that are specifically borne by the secondary chromosome of *R. rhizogenes* (119 homologs in *Rhizobium* sp. NT-26 over 129 specific translocated genes). If strain NT-26 belonged to the *Rhizobium* lineage, the majority of these genes would probably have been lost with this whole chromid, although we cannot rule out potential translocations of those genes back to the main chromosome along with the chromid loss. Taken together, and even though our current data do not allow a more accurate classification, our observations support the inclusion of *Rhizobium* sp. NT-26 in the *Agrobacterium* subgroup. Nevertheless, according to the current nomenclature (Young et al. 2001), *Rhizobium* is still a valid genus name for strain NT-26.

Finally, although chromosomes are mainly dedicated to housekeeping functions, chromids carry genes involved in specific ecological functions, that is, legume symbiosis enabled by symbiotic plasmids in *Rhizobium* and *Sinorhizobium* (Harrison et al. 2010) and plant-related functions in *A. tumefaciens* C58 (Lassalle et al. 2011). All these functions relate to the interactions inside the rhizosphere and soil that represent the canonical habitat of *Rhizobiaceae*. The loss by *Rhizobium* sp. NT-26 of this large replicon housing rhizosphere-associated functions may be related to the drastic shift in environment the lineage has experienced. Indeed (discussed earlier), this loss coincides with the gain of genes enabling resistance to heavy metals, and arsenic and sulfur metabolisms, both traits with a potentially great adaptive value in sustaining life on an arsenopyrite (FeAsS)-containing rock.

Functional Approaches to Investigate Arsenic Metabolism and Resistance

Proteomic and Transcriptomic Profiling

Genomic tools have been used to study the bacterial response to arsenic mainly in arsenite-oxidizing *Betaproteobacteria* such as *H. arsenicoxydans*, a chemoorganotroph (Carapito et al. 2006; Muller et al. 2007; Weiss et al. 2009), and *T. arsenitoxydans*, a chemolithoautotroph (Bryan et al. 2009; Arsène-Ploetze et al. 2010). Arsenic metabolism was investigated in the chemolithoautotrophic *Alphaprotebacterium Rhizobium* sp. NT-26 by two complementary approaches: protein and RNA profiling using 2D gel electrophoresis and DNA microarrays, respectively. The comparisons of expression were done on proteins and RNA isolated from strain NT-26 grown heterotrophically with and without arsenite. The main results are summarized in table 1 and a complete list of the data is presented in supplementary table S6, Supplementary Material online.

The 2D gel proteomic profile of *Rhizobium* sp. NT-26 was quite similar to those previously obtained for *H. arsenicoxydans* (Carapito et al. 2006) and *T. arsenitoxydans* (Bryan et al. 2009), which are also neutrophilic bacteria. Sixty-three spots showed a significant difference in their accumulation pattern in strain NT-26 grown with and without As(III). Their analysis by mass spectrometry led to the identification of 141 proteins (supplementary table S6a, Supplementary Material online), including arsenite oxidase, which was identified for the first time on 2D gels. Like membrane proteins, such periplasmic proteins are often eliminated with cell debris before their solubilization during sample preparation. Proteins up- or downregulated with a fold-change ranging from +39.6 to −5.8 when *Rhizobium* sp. NT-26 was grown in the presence of As(III) had a molecular mass ranging from 15 to 109 kDa and a pI value from 4.2 to 7.9. Among them, 24% were involved in cell envelope and cellular processes, 12% in

Table 1

Major Arsenic-Regulated Functional Categories Identified in Transcriptomics and Proteomics Experiments

Functional Category	MaGe ID	Gene	Function	FC	
				RNA[a]	Protein[a]
Oxidative stress	NT26v4_0103	*rpoN*	RNA polymerase σ^{54} factor	1.32	
	NT26v4_0389	*katA*	Catalase A	1.37	
	NT26v4_0773	*ohr*	Organic hyperoxide resistance	1.41	
	NT26v4_0799	*sodB*	Superoxide dismutase		3
Carbon metabolism	NT26v4_0674	*cbbF*	Fructose-1.6-bisphosphatase	1.3	
	NT26v4_0670	*cbbL*	RuBisCo large subunit	1.37	
	NT26v4_2684	*cbbT*	Transketolase		3.4
	NT26v4_0667	*cbbE*	Ribulose-phosphate 3-epimerase	1.35	
Nitrogen metabolism	NT26v4_3645	*norQ*	Putative NorD protein	1.52	
	NT26v4_3643	*norC*	Nitric oxide reductase subunit C	1.42	
	NT26v4_3641	*norE*	Involved in nitric oxide reduction	1.60	
	NT26v4_3654	*nirV*	Involved in nitrite reduction	1.51	
	NT26v4_3653	*nirK*	Cu-containing nitrite reductase		39.6
Arsenic metabolism	NT26v4_p10030	*aioA*	Arsenite oxidase large subunit	3.89	10
	NT26v4_p10029	*aioB*	Arsenite oxidase small subunit	4.27	
	NT26v4_p10118	*arsC1b*	Arsenate reductase ArsC		22
	NT26v4_p10122	*arsH1*	Arsenical resistance protein		2.9
Sulfur metabolism	NT26v4_2623	*soxG*	Sulfur oxidation protein	−1.37	
	NT26v4_2619	*soxV*	Sulfur oxidation protein	−1.39	
	NT26v4_2618	*soxW*	Thioredoxin	−1.52	
	NT26v4_2617	*soxX*	Sulfur oxidizing protein	−1.68	
	NT26v4_2616	*soxY*	Sulfur oxidation protein	−1.45	
	NT26v4_2615	*soxZ*	Sulfur oxidation protein	−1.51	
	NT26v4_2882	*cysT*	Sulfate/thiosulfate transport protein	−1.72	
Phosphate metabolism	NT26v4_1226	*phoE1*	Phosphonate ABC transporter	1.45	
	NT26v4_0079	*phoR*	Phosphate regulon kinase	1.32	
	NT26v4_p10016	*phoE2*	Phosphonate ABC transporter subunit	1.61	
	NT26v4_p10017	*phoT2*	Phosphonate ABC transporter subunit	1.42	
	NT26v4_p10024	*pstS2*	High-affinity phosphate transporter		3.7
Motility/biofilm	NT26v4_0204	*fliF*	Flagellar M-ring protein	1.44	
	NT26v4_0227	*flaA*	Flagellin A		7.2
	NT26v4_0228	*fla*	Flagellin		9.7
	NT26v4_0655	*qseB*	Quorum sensing regulator QseB	1.39	
	NT26v4_2748	*noeJ*	Mannose-1-P guanylyltransferase	1.41	
	NT26v4_1615	*kdsA*	KDO 8-P synthase	1.40	
Plant/bacteria interactions	NT26v4_1705	*cgmA*	Beta-1.2-glucan modification protein	1.7	
	NT26v4_p10302	*avhB10*	Type IV system transglycosylase	1.41	

Note.—Induced and repressed functions are shown in blue and black, respectively. No value is indicated in the FC column if the gene is not statistically differentially expressed in transcriptomics or if the protein has not been identified in proteomics. Complete data are presented in supplementary table S6, Supplementary Material online.

[a]Fold-change observed in transcriptomics and proteomics data, respectively.

transport and binding proteins, 11% in information and regulation pathways, 42% in metabolism, 1% in transcription, and 10% were of unknown function.

The second approach used whole genome microarrays to perform a differential expression profiling experiment. Under As(III) stress, the transcript level of 199 genes, that is 4.5% of the whole genome, showed an increase of up to more than four times with a *P* value ≤ 0.05. At the same time, the expression of 416 genes, that is 9.5% of the whole genome, decreased by up to more than three times (supplementary table S6c, Supplementary Material online).

General Response to Arsenic Stress

Several proteins involved in arsenic resistance were shown to be accumulated on 2D gels when the organism was grown in the presence of As(III), for example, an ArsH1 NADPH-dependent FMN reductase and an ArsC1 arsenate

reductase (table 1) with fold changes of 2.9 and 22, respectively. These two proteins are encoded by an *ars* operon located on the megaplasmid, which also contains genes coding for an ArsB efflux pump and an ArsR regulator. Moreover, a second operon located on the chromosome contains an *arsA* gene coding for an ATPase associated with an ArsB arsenite efflux pump. ArsA enables the strain to increase arsenic resistance by ATP-dependent extrusion of the metalloid out of the cell (Branco et al. 2008).

Microarray experiments showed that genes encoding the two arsenite oxidase subunits, that is, the small subunit, AioB, which contains the Rieske 2Fe-2S cluster and the catalytic subunit, AioA, which contains a molybdopterin guanine dinucleotide at the active site and a 3Fe–4S cluster (Santini and vanden Hoven 2004), were about 4-fold induced in the presence of arsenite (table 1). In contrast, the expression of two genes located downstream of the *aioBA* operon, that is, *cytC* encoding the periplasmic cytochrome c_{552}, which can serve as an electron acceptor to the arsenite oxidase (Santini et al. 2007) and *moeA1* encoding a molybdenum cofactor biosynthesis gene, was not significantly affected under arsenite stress (supplementary table S6c, Supplementary Material online). These observations are further supported by proteomic experiments showing that, among these proteins, AioA was found to be preferentially accumulated in the presence of As(III). All these results are consistent with previous data (Santini et al. 2007), which suggests that the expression of *aioBA* genes is induced by arsenite while genes located downstream are constitutively expressed even though they may have a role in arsenic metabolism.

Located upstream of *aioBA* are two regulatory genes, *aioS* and *aioR*, which encode a sensor histidine kinase and a response regulator, respectively (Sardiwal et al. 2010). Both proteins have been shown to be required for the transcriptional regulation of the *aioBA* genes (Koechler et al. 2010; Sardiwal et al. 2010). Moreover, it has been demonstrated that the expression of the *aioBA* genes requires the RpoN alternative sigma factor (σ^{54}) in *H. arsenicoxydans* (Koechler et al. 2010). Similarly, a role for RpoN in arsenite oxidation has been recently highlighted in *A. tumefaciens* 5A (Kang et al. 2012). In this respect, a putative σ^{54}-dependent promoter region has been detected upstream of the *aioB* gene in *Rhizobium* sp. NT-26 (Santini et al. 2007), suggesting that it is also involved in the expression of the *aioBA* operon in strain NT-26 (Sardiwal et al. 2010). This hypothesis is supported by our transcriptomic data, which revealed an induced expression of *rpoN* in *Rhizobium* sp. NT-26 when it was grown in the presence of As(III) (table 1), in contrast to the constitutive expression recently observed in *A. tumefaciens* 5A (Kang et al. 2012). Similarly, microarray and 2D-gel data showed a 2-fold increase in the expression of genes coding for general chaperones, that is, DnaK and GroEL, in the presence of arsenite (supplementary table S6a and c, Supplementary Material online), which is in agreement with the role played by proteins of the heat-shock family in As(III) oxidation in *H. arsenicoxydans* (Koechler et al. 2010).

Rhizobium sp. NT-26 also tolerates arsenate concentration greater than 0.5 M (Clarke A. and Santini J.M., unpublished data), suggesting the existence of an alternative mode of resistance. The first one is an Ars-type arsenic resistance system, components of which were found to be upregulated when the strain was grown with arsenite (discussed earlier) and the second is the presence of a specific phosphate transport system which is thought to limit arsenate entry into the cell (Weiss et al. 2009). Indeed, in *Rhizobium* sp. NT-26, the arsenic genomic island contains a *pst* operon in the vicinity of the *aio* operon. The *pst* operon encodes proteins implicated in the specific transport of phosphate into the cell to maintain a sufficient level of this ion despite the presence of arsenate, a structural analog of phosphate (Muller et al. 2007; Cleiss-Arnold et al. 2010). PstS2, a periplasmic protein involved in phosphate transport and encoded by this operon, had a 3.7-fold increase in expression when strain NT-26 was grown in the presence of As(III) (table 1). The *pst* operon is regulated by *phoR*, which encodes a membrane-associated protein kinase that phosphorylates PhoB in response to environmental signals. Indeed, microarray data showed that the expression of *phoR* was also upregulated in the presence of As(III) (table 1). Moreover, the PhoR protein may be involved in biofilm formation as *phoB* overexpression has been shown to increase biofilm formation in *A. tumefaciens* (Danhorn et al. 2004). This is supported by the presence in the vicinity of the *pho* chromosomal operon of a cluster of genes involved in EPS biosynthesis (NT26v4_1233–NT26v4_1263).

Arsenic is known to induce oxidative stress by generating free radicals (Bernstam and Nriagu 2000). An induction of genes involved in the resistance to such a stress has been previously observed in *Pseudomonas aeruginosa* and in *H. arsenicoxydans* under arsenite exposure (Parvatiyar et al. 2005; Weiss et al. 2009; Cleiss-Arnold et al. 2010). In *Rhizobium* sp. NT-26, an increase in *katA* mRNA, which encodes a catalase involved in the protection against oxidative stress by scavenging endogenously produced H_2O_2, was observed in microarray experiments (table 1). Similarly, the expression of *ohr*, which promotes bacterial resistance to hydroperoxide, was also up-regulated in *Rhizobium* NT-26 (table 1). Finally, results of the proteomic experiments showed a 3-fold increase in the SodB superoxide dismutase accumulation when strain NT-26 was grown in the presence of As(III). These observations further support the strong link, which exists in bacteria between arsenic response and protection against oxidative stress (Bertin et al. 2012).

Carbon, Nitrogen, and Energy Metabolism

Rhizobium sp. NT-26 is able to use various carbon or electron sources for growth. Indeed, multiple carbohydrates such as acetate, succinate, fumarate, lactate, glucose, fructose,

xylose, and galactose are potential carbon sources for this bacterium (Santini et al. 2000). Alternatively, *Rhizobium* sp. NT-26 is able to grow chemolithoautotrophically in the presence of bicarbonate as a carbon source, oxygen as an electron acceptor and arsenite as an electron donor (Santini et al. 2000). Transcriptomics and proteomics experiments revealed that several genes and proteins involved in the fixation of CO_2 via the Calvin cycle were upregulated with a fold change ranging from 1.3 to 3.4 when strain NT-26 was grown in the presence of As(III) (table 1). This is in agreement with the proteomics results obtained in *T. arsenivorans,* where an accumulation of the ribulose-1,5-biphosphate carboxylase/oxygenase large subunit and of the fructose-1,6-biphosphate has been observed when the organism was grown in the presence of As(III) (Bryan et al. 2009). Both strains may thus improve their capacity to fix CO_2 when arsenite is present.

In addition, our microarrays data showed that the expression of *nirV* encoding a protein involved in nitrite reduction was induced. Moreover, 2D-gel data showed that NirK, which also participates in the reduction of nitrite to nitric oxide, was 39.6 times more accumulated when strain NT-26 was grown in the presence of arsenite. These experiments also showed an induction of several genes of the *norEFCBQD* nitric oxide reductase gene cluster when *Rhizobium* sp. NT-26 was grown in the presence of As(III) (table 1). These genes encode proteins that catalyze the reduction of nitric oxide to nitrous oxide, that is, *norQ, norE* and *norC* coding for a protein involved in nitric oxide reduction, a nitric oxide reductase activating protein, and the small subunit of the nitric oxide reductase, respectively. This suggests that the chemolithoautotrophic bacterium *Rhizobium* sp. NT-26 may fix CO_2 and couple nitrite reduction with As(III) oxidation. However, as no growth was observed with arsenite on either nitrate or nitrite, arsenite oxidation using nitrite as electron acceptor in autotrophic conditions seems unable to support sufficient energy (ATP) generation to sustain growth.

Rhizobium sp. NT-26 has been shown to grow with hydrogen sulfide, elemental sulfur, and thiosulfate (Santini J.M., unpublished data). In agreement with these observations, a *sox* cluster implicated in the oxidation of thiosulfate is present in the *Rhizobium* sp. NT-26 genome. Nevertheless, many genes involved in sulfur metabolism, that is, *soxGVWXYZ* and *cysT* were downregulated by up to 2-fold when the strain was grown in the presence of arsenite (table 1). Our results therefore suggest that, even though *Rhizobium* sp. NT-26 may be able to grow by using sulfur as an electron donor, the strain represses sulfur oxidation when grown in the presence of As(III). One hypothesis may be that, in such a case, the strain expresses a repressor of the *sox* genes. Their products serve for the oxidation of thiosulfate to sulfate and the reaction intermediates, that is, sulfite, sulfide, and hydrogen sulfide, have been shown to inhibit arsenite oxidase activity (Lieutaud et al. 2010).

Physiological and Genetic Approaches: Flagellar Motility and Biofilm Formation

Flagellum Cascade Features

Rhizobium sp. NT-26 is motile by the means of two subterminal flagella (Santini et al. 2000). Genes involved in their biosynthesis are organized in a large chromosomal cluster of 55 genes showing a perfect synteny with those of *S. meliloti*. In this flagellar regulon, *visN* and *visR* form part of the master operon and encode the proteins forming the VisNR heterodimer that acts as a global transcriptional regulator. This master regulator activates the expression of genes located in the cascade that encode flagella, motor, and chemotaxis proteins (Sourjik et al. 2000). In *Rhizobium* sp. NT-26, microarray data showed that the expression of *fliF*, coding for the flagellum M-ring protein, was induced when the organism was grown in the presence of As(III) (table 1). Furthermore, proteomic data showed a 9.7- and 7.2-fold-increase in the accumulation of flagellin proteins Fla and FlaA, respectively (table 1). The expression of *qseB* was also induced when *Rhizobium* sp. NT-26 was grown in the presence of As(III). QseB has been shown to participate in the flagellum and motility bacterial regulatory network via a quorum-sensing mechanism. Indeed, in *E. coli*, *qseBC* expression enhances the transcription of flagellar genes in response to the autoinducer by a direct binding of QseB to the *flhDC* master operon promoter (Clarke and Sperandio 2005). Finally, microarray data showed the induction of genes possibly involved in the synthesis of an exopolysaccharide matrix in *Rhizobium* sp. NT-26, that is, *noeJ*, coding for a mannose-1-phosphate guanylyltransferase and *kdsA*, coding for a 2-dehydro-3-deoxyphosphooctonate aldolase. These observations suggest that As(III) has an impact on flagellum synthesis, that is to say motility, and biofilm formation in *Rhizobium* sp. NT-26, as observed in *H. arsenicoxydans* (Muller et al. 2007; Marchal et al. 2010). To test this hypothesis, swarming assays were performed on 0.3% agar plates. The presence of As(III) was shown to increase the swarming ring by up to 2-fold in the presence of 8 mM As(III) (fig. 4*A*). Remarkably, cell observation under a TEM revealed that flagellum biosynthesis occurred immediately in the presence of 8 mM As(III) while more than two days were needed to observe flagella in the absence of arsenite (fig. 4*B* and *C*), providing evidence that arsenite promotes motility in *Rhizobium* sp. NT-26. In addition, a two-fold reduction in biofilm formation was observed in the first 24 h of growth in the presence of As(III) (fig. 4*D*), which suggests a preferential development as motile planktonic cells rather than as unflagellated sessile cells as in *H. arsenicoxydans* (Marchal et al. 2010).

Random Mutagenesis

The *Rhizobium* sp. NT-26 genome organization suggests that motility and arsenite oxidation depend on genes located on its

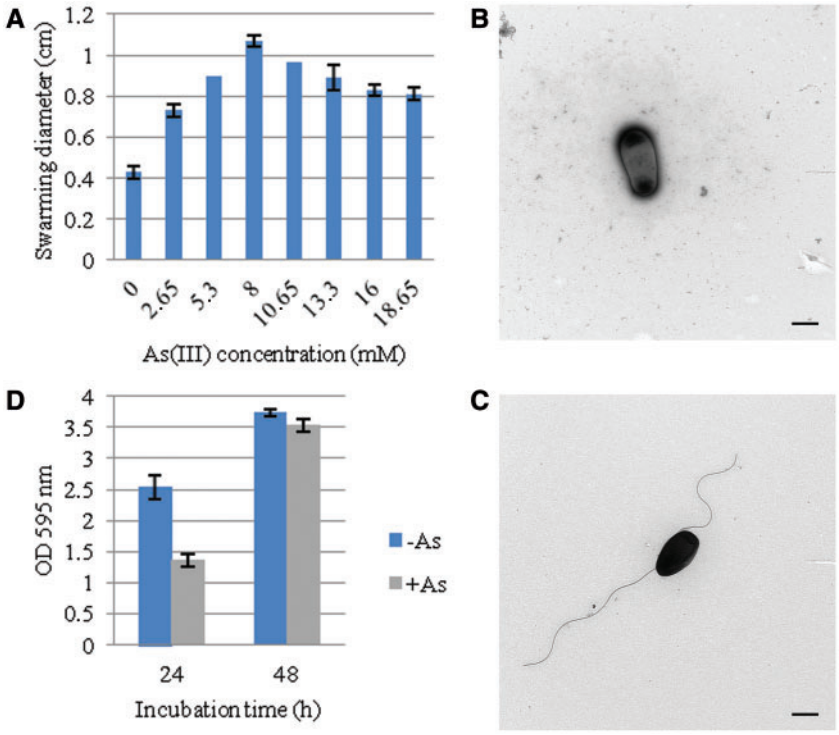


Fig. 4.—Motility phenotype of *Rhizobium* sp. NT-26 grown at different concentration of arsenite. (*A*) Swarming diameter measured after 48 h in MSM containing 0.04% yeast extract and supplemented by different concentrations of As(III). Results are the mean values of three independent experiments. (*B*) and (*C*) TEM observations of *Rhizobium* sp. NT-26 at 24 h of culture, without or with 8 mM As(III), respectively. The scale bar corresponds to 500 nm and the pictures are representative of 10 pictures. (*D*) Biofilm formation by strain NT-26, without or with 8 mM As(III) visualized by the crystal violet method. Results are the mean values of 24 replicates.

chromosome and on its megaplasmid, respectively. With the aim to analyze the possible link between these physiological processes, a mutant library was constructed by random transposon mutagenesis (Tang et al. 1999). The motility of 6,000 kanamycin-resistant transposition derivatives was tested on semisolid medium, which led to the isolation of 22 motility-deficient mutants. The mutations that resulted in a loss of motility were identified by sequencing the mini-Tn*5* transposon insertion sites. Fourteen mutations were shown to directly disrupt motility genes (table 2), and the proteins encoded by these genes are either structural or regulatory components of the flagellum cascade, that is, 5 Flg proteins (FlgE, FlgF, FlgG, FlgI, and FlgL), 5 Fli proteins (FliF, FliK, Flip, and FliR), 1 Flh protein (FlhA), and 1 Vis protein (VisR). No flagellin-defective mutant was obtained, which may be explained by the presence of four different flagellin-encoding genes on the chromosome.

Similarly, six mutations resulting in a lack of arsenite oxidation as compared with the wild-type and motility mutants were obtained after screening 6,000 kanamycin-resistant clones with the silver nitrate method (Muller et al. 2007) (table 2). First, two mutations were identified in the arsenite oxidase genes, that is, *aioA* and *aioB*. One mutation was also shown to affect *aioR,* which encodes the regulatory protein of the AioRS two-component system. No oxidation of As(III) to As(V) was detected by HPLC-ICP-AES (Muller et al. 2007) in this mutant, as compared with complete As(III) oxidation determined in the motility mutant deficient in the flagellum master regulator VisR. A fourth mutation was located in the *moeB* gene involved in the synthesis of the molybdopterin cofactor required for arsenite oxidase activity. Finally, the inactivation of the *aioX* gene, which is located upstream of *aioSR*, also resulted in a loss of arsenite oxidation in *Rhizobium* sp. NT-26. In *A. tumefaciens* 5A, the periplasmic AioX has been recently shown to be involved in the regulation of As(III) oxidation (Liu et al. 2012).

To determine the link between As(III) oxidation and colonization properties in *Rhizobium* sp. NT-26, the ability of various mutants to move and to form a biofilm was evaluated in the presence of arsenite (fig. 5). Mutations in flagellar genes resulted in a loss of motility and in a decrease in biofilm formation. Indeed, all the mutants we tested were nonmotile (fig. 5*A*) and lost between 12% and 45% of their ability to form a biofilm when compared with the wild-type strain (fig. 5*B*). This observation demonstrates that, although

Table 2

Rhizobium sp. NT-26 Mutants Isolated on the Basis of a Loss of Motility or Arsenite Oxidation

Mutant	MaGe ID[a]	Gene	Function	Gene Location[b]	Insertion[c]
Motility					
2B5	NT26v4_0222	*flgI*	Flagellar P-ring protein precursor	221267–222391	$258_{I\text{-}O}$
2D6	NT26v4_0245	*flhA*	Flagellar biosynthesis protein	243262–245349	$38_{I\text{-}O}$
4H7	NT26v4_0204	*fliF*	Flagellar M-ring protein	206351–208027	$477_{I\text{-}O}$
5B7	NT26v4_0226	*fliP*	Flagellar biosynthetic protein	224272–225009	$109_{O\text{-}I}$
6E4	NT26v4_0220	*flgG*	Flagellar basal-body rod protein	219981–220769	$182_{I\text{-}O}$
10A11	NT26v4_0248		Putative FlgJ-like protein	246658–247212	
10G2	NT26v4_2965		Conserved hypothetical protein	2880029–2881285	$44_{O\text{-}I}$
16B6	NT26v4_0245	*flhA*	Flagellar biosynthesis protein	243262–245349	$55_{I\text{-}O}$
16B7	NT26v4_0240	*flgL*	Flagellar hook-associated protein	240428–241549	$11_{I\text{-}O}$
18D11	NT26v4_0238	*flgE*	Flagellar hook protein	237398–238930	$363_{I\text{-}O}$
20E7	NT26v4_0246	*fliR*	Flagellar biosynthetic protein	245378–246130	$110_{O\text{-}I}$
23E5	NT26v4_0214	*flgF*	Flagellar basal-body rod protein	216054–216788	$44_{O\text{-}I}$
29G6	NT26v4_0247		Putative FliR/FliJ-like chaperone	246137–246553	$91_{I\text{-}O}$
35A8	NT26v4_2314		Putative two-component sensor histidine kinase	2263060–2264472	$113_{O\text{-}I}$
37C12	NT26v4_2314		Putative two-component sensor histidine kinase	2263060–2264472	$48_{O\text{-}I}$
37H1	NT26v4_2267		Putative ATP-dependent hydrolase protein	2217585–2219546	$337_{O\text{-}I}$
38B4	NT26v4_0206	*visR*	Master transcriptional regulator of flagellar regulon	209070–209798	$53_{O\text{-}I}$
38G3	NT26v4_0204	*fliF*	Flagellar M-ring protein	206351–208027	$235_{I\text{-}O}$
39G12	NT26v4_2671		Conserved protein of unknown function	2586119–2587117	$94_{O\text{-}I}$
40E5	NT26v4_0234	*fliK*	Flagellar hook-length regulator	234324–235835	$83_{I\text{-}O}$
50E11	NT26v4_0250		Conserved integral membrane protein of unknown function	247603–248142	$92_{O\text{-}I}$
61C2	NT26v4_3970		Conserved exported protein of unknown function	3918374–3918853	$196_{I\text{-}O}$
Arsenite oxidation					
8G1	NT26v4_p10026	*aioX*	Putative periplasmic phosphite-binding-like protein precursor; PtxB-like protein	23892–24806	$275_{I\text{-}O}$
11B3	NT26v4_4048	*moeB*	Putative molybdopterin biosynthesis protein MoeB	3998958–3999725	$245_{O\text{-}I}$
24B7	NT26v4_p10028	*aioR*	Two-component response regulator	26262–27584	$65_{I\text{-}O}$
37C3	NT26v4_p10026	*aioX*	Putative periplasmic phosphite-binding-like protein precursor; PtxB-like protein	23892–24806	$191_{I\text{-}O}$
55H7	NT26v4_p10029	*aioB*	Arsenite oxidase small subunit	27721–28248	
60E6	NT26v4_p10030	*aioA*	Arsenite oxidase large subunit	28261–30798	$818_{I\text{-}O}$

[a]Identification number of the gene in the MaGe interface.
[b]Position of the corresponding gene on the chromosome or the plasmid.
[c]Position of the codon immediately upstream of the transposon insertion site. Subscripts indicate the orientation of the insertion.

Rhizobium sp. NT-26 has a preferential motile life style in the presence of arsenite, flagella have a role as adhesive appendages in the first steps of biofilm formation, which has been shown previously in other studies (Kirov et al. 2004; Nejidat et al. 2008). This result is in agreement with those obtained with *H. arsenicoxydans*, where mutations resulting in nonfunctional flagella led to a more rapid adhesion as compared with the wild-type. Finally, *aioA* and *aioR* mutants were less motile and formed 30% and 45%, respectively, less biofilm than the wild-type (fig. 5), further supporting the role of motility and flagella in biofilm formation.

Regulation of Flagella Synthesis by AioR

The "omics" data showed that flagellar proteins and genes were upregulated when strain NT-26 was grown in the presence of arsenite (table 1). Remarkably, both *aioA* and *aioR* mutations resulted in a moderate reduction in motility (fig. 5*A*). This can be explained by the reduction in energy available to the cells as they are unable to metabolize arsenite (Santini et al. 2000). In addition, TEM observations of the *aioR* mutant, affected in the two-component signal transduction system, revealed the presence of flagella in the early log phase of growth even in the absence of As(III), which suggests that AioR may be involved in the repression of motility when no arsenite is present (fig. 6). AioR may thus interact, directly or indirectly, with components of the flagellar cascade.

To identify possible AioR-binding sites in the *Rhizobium* sp. NT-26 genome, multiple sequence alignments of all *aioBA* regulatory sequences available in databases were performed with fuzznuc (Rice et al. 2000). This enabled us to suggest the possible existence of two AioR putative binding sites upstream of the *aioBA* transcriptional start site, that is, GT[CT]CGN(6)CG[GA]AC in the *Rhizobiales* strains and GTTNCN(6)GNAAC in the *Burkholderiales*

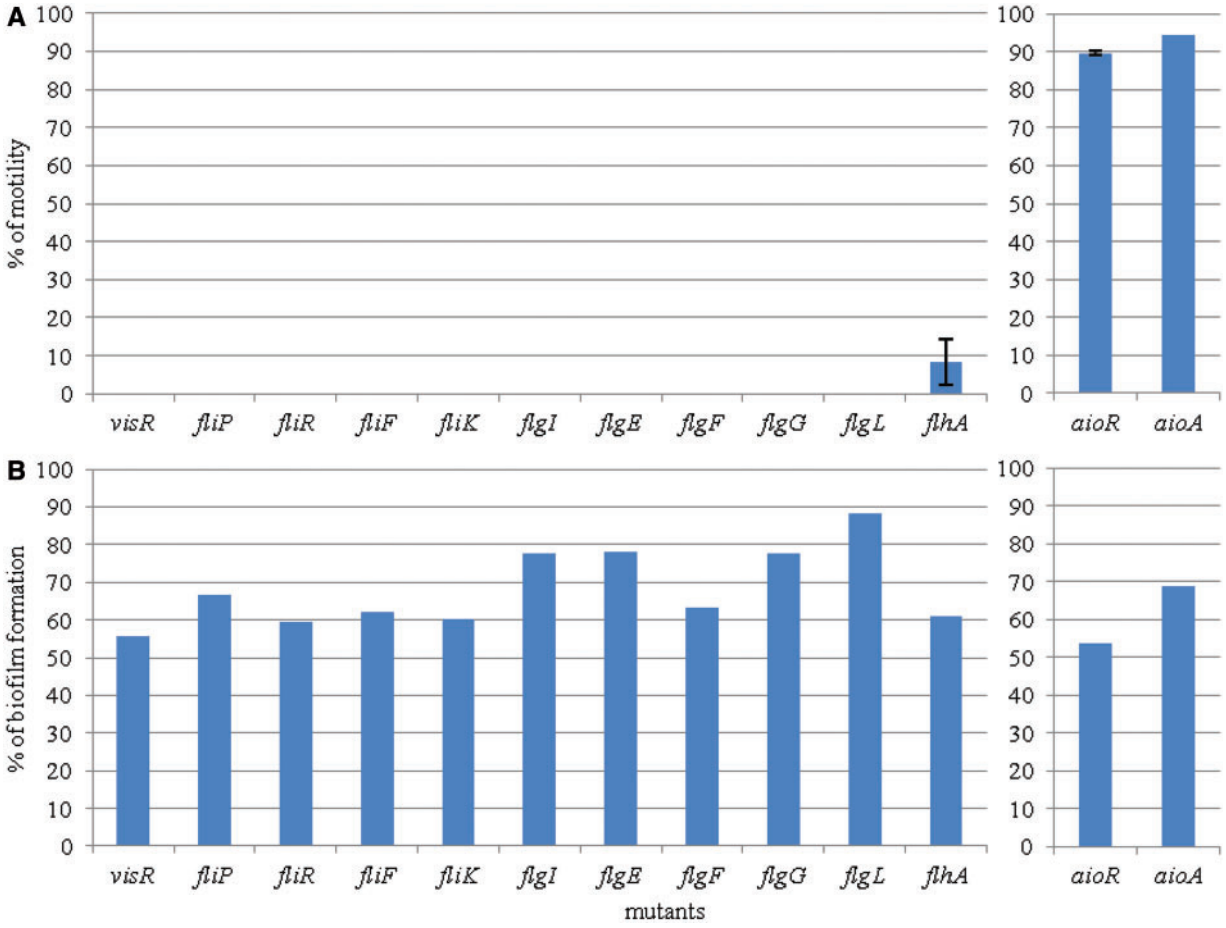


Fig. 5.—Percentage of motility and biofilm formation in various mutants as compared with *Rhizobium* sp. NT-26 wild-type strain. The left and the right panels show the results obtained in mutants affected in motility and As(III) oxidation, respectively. (*A*) % of swarming motility measured after 24 h. Results are the mean values calculated from the % of three independent experiments. In each experiment, mutant and wild-type strains were tested in triplicates. (*B*) % of biofilm formation visualized by crystal violet coloration. Results are the % calculated from the mean values of six replicates for each strain.

strains. In contrast, strains lacking the two-component system *aioSR* operon did not harbor any of these putative AioR-binding sites, which further supports a role for these motifs in the regulation of *aioBA* operon expression by AioR. Although the GT[CT]CGN(6)CG[GA]AC putative binding site was found at 49 locations on the *Rhizobium* sp. NT-26 chromosome, it is only in the upstream region of the *aioBA* operon that this motif was associated with the -12/-24 σ^{54}-dependent promoter sequence needed for the RpoN-dependent transcription initiation of arsenite oxidase genes (Koechler et al. 2010). Moreover, no clear connection could be observed between those putative binding sites and motility-related genes. Nevertheless, a search of the whole genome of strain NT-26 with a relaxed version of the pattern allowing any nucleotide at its degenerated positions, GTNCGN(6)CGNAC, yielded 39 more hits than with the canonical pattern. This low number of new hits suggests that the presence of this signature is not due to chance and this sequence may therefore have a potential regulatory role. None of the new hits was associated with a RpoN motif site although one hit was found within the coding sequence of the flagellar master regulatory gene *visN*. Therefore, although we cannot rule out an indirect effect of AioR by controlling another regulatory protein, we can hypothesize that the binding to this mildly degenerate motif of unphosphorylated AioR in the absence of arsenic would result in a *visN* repression and a delayed motility. Such a transcriptional repression via binding of the coding sequence of target genes has already been observed for other regulatory proteins and two-component system regulators, for example, OxyR (Zheng et al. 2001) and PrrA (Eraso et al. 2008). Taken together, our results demonstrate the importance of arsenite oxidation in the behavioral response of *Rhizobium* sp. NT-26, suggesting that arsenic metabolism enhances the ability of the organism to explore and colonize its environment.

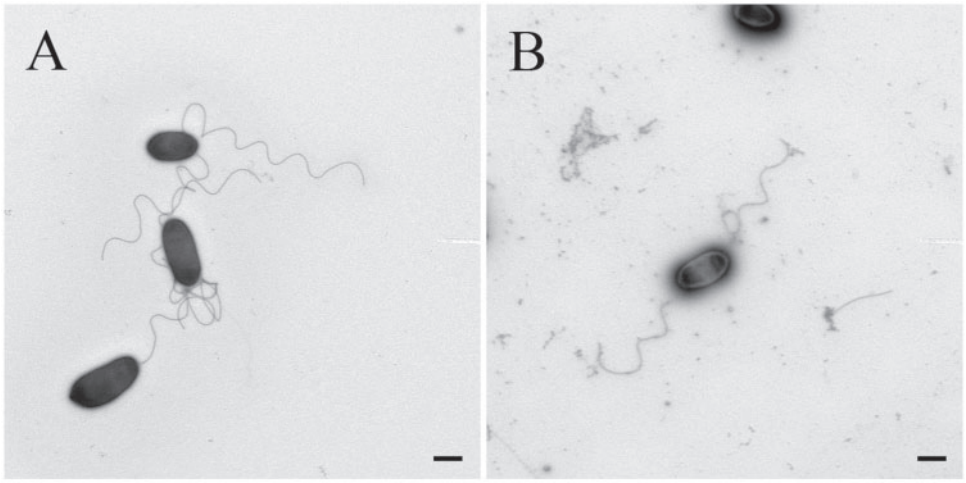


Fig. 6.—TEM observations of the *aioR* mutant. The *aioR* mutant was cultivated 24 h (*A*) in the absence of As(III) and (*B*) in the presence of 8 mM As(III). Pictures are representative of 10 photographs. The scale bar corresponds to 500 nm.

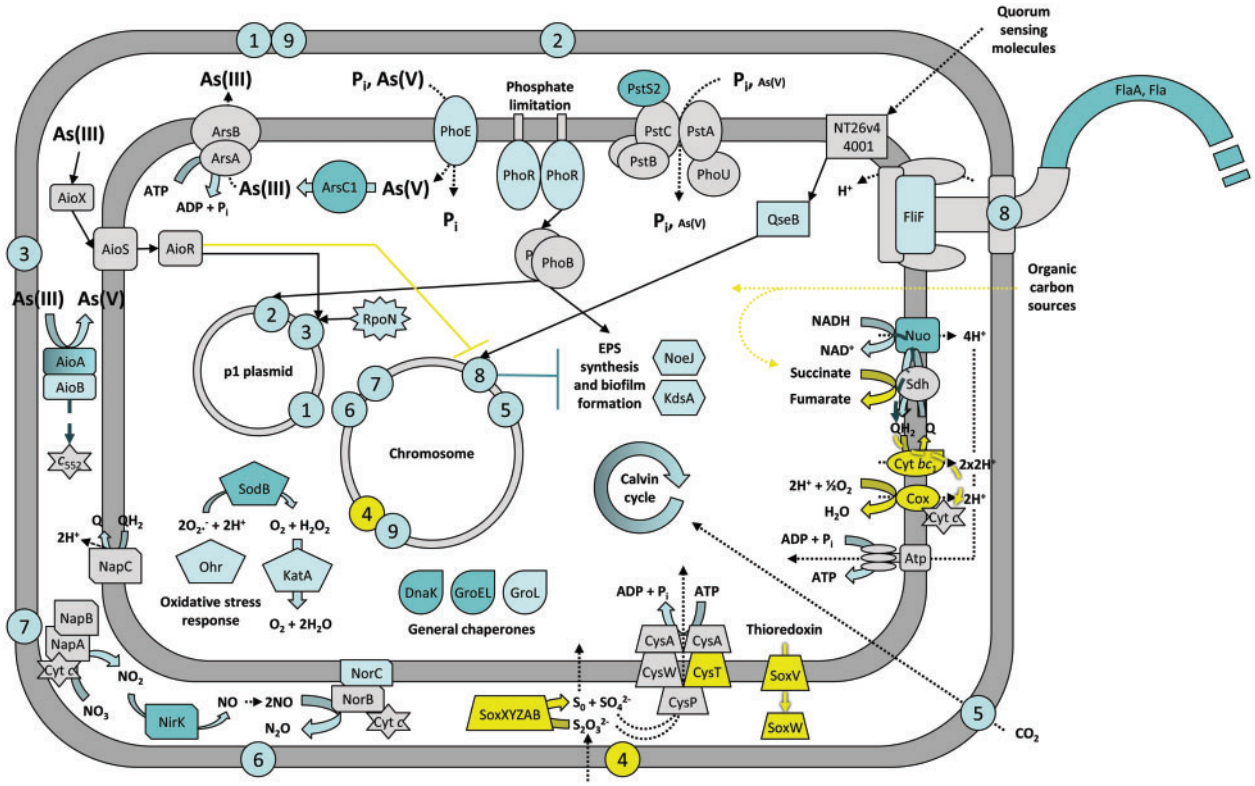


Fig. 7.—Conceptual representation of the *Rhizobium* sp. NT-26 response to arsenite exposure. This representation takes into account our genomic, transcriptomic, proteomic, and physiological results as well as data from the literature. Numbers 1 to 9 represent biological functions and the approximate genomic location of the gene clusters encoding their corresponding proteins. 1 and 9: arsenate reduction; 2: phosphate and arsenate transport; 3 and 4: arsenite and sulfur oxidation, respectively; 5: carbon fixation; 6 and 7: nitrate and nitrite reduction, respectively; 8: motility. Block, dashed, dotted, and standard arrows symbolize chemical reactions, electron flow, transport/utilization of molecules and signaling/regulatory pathways, respectively. When highlighted in dark blue, light blue, yellow or gray, elements have been identified as being induced by proteomic, induced by transcriptomic, repressed by transcriptomic, and present in the genome, respectively. For clarity reasons, proteins for which the exact function is still unknown but that are related to the different processes are not shown, the protein complexes of the respiratory chain, that is the NADH dehydrogenase, the fumarate reductase, the cytochrome bc_1 and the cytochrome *c* oxidase are designated by Nuo, Sdh, Cyt bc_1, and Cox, respectively, plasmid p2 is not shown and only one flagella is represented. Finally, Cyt *c* and c_{552} are for cytochrome *c* and cytochrome c_{552}, respectively, and NT26v4_4001 is a homolog of *qseC*.

Conclusion

This study extends our knowledge of the physiological response to arsenic in arsenite-oxidizing bacteria. Our results provide for the first time a reference set of genomic, transcriptomic, and proteomic data of an *Alphaproteobacterium* isolated from an arsenopyrite-containing goldmine, which allowed us to propose a model for the *Rhizobium* sp. NT-26 response to arsenite exposure (fig. 7). Although phylogenetically related to the plant-associated bacteria, strain sp. NT-26 has lost the major colonizing capabilities needed for symbiosis. Instead, this bacterium has acquired on a megaplasmid the various genes which allow it to metabolize arsenate. Remarkably, a link between flagellar motility/biofilm formation and arsenite oxidation was observed although the genes required for these physiological activities are carried by different genetic determinants, that is, the chromosome and the megaplasmid, respectively. This suggests the existence of a mechanism, probably indirect and which remains to be characterized at a molecular level, of a coordinate regulation of these two important biological processes. This underlines the importance of arsenite oxidation in the colonization of arsenic-rich ecosystems, a toxic element widespread on Earth. Importantly, our data also illustrate the major contribution of environmental pressure on the evolution of bacterial genomes, which results in a gain and loss of multiple functions, improving the fitness of the strains to extreme ecological niches.

Supplementary Material

Supplementary figures 1–4, tables S1–S6, and methods S1–S4 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the Université de Strasbourg (UdS), the Consortium National de Recherche en Génomique (CNRG), the Centre National de la Recherche Scientifique (CNRS), the Australian Research Council (DP034305), the French Ministère de l'Enseignement Supérieur et de la Recherche to J.A. and J.C.-A., the Direction Générale de l'Armement to L.G., a studentship from the Agence Nationale de la Recherche (ANR) COBIAS project (PRECODD 2007) to M.M., and the Ecole Normale Supérieure (ENS) of Lyon to F.L. The authors thank Mathieu Erhardt for the TEM observations and Ashleigh Clarke for her work on arsenic resistance in strain NT-26. This work was done in part in the frame of the Groupement de Recherche—Métabolisme de l'Arsenic chez les Microorganismes (GDR2909-CNRS) (http://gdr2909.alsace.cnrs.fr/, last accessed April 30, 2013).

Literature Cited

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Arsène-Ploetze F, et al. 2010. Structure, function, and evolution of the *Thiomonas* spp. genome. PLoS Genet. 6:e1000859.

Aury J-M, et al. 2008. High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. BMC Genomics 9:603.

Bernstam L, Nriagu J. 2000. Molecular aspects of arsenic stress. J Toxicol Environ Health B Crit Rev. 3:293–322.

Bertin PN, et al. 2011. Metabolic diversity among main microorganisms inside an arsenic-rich ecosystem revealed by meta- and proteogenomics. ISME J. 5:1735–1747.

Bertin PN, et al. 2012. Microbial arsenic response and metabolism in the genomics era. In: Santini JM, Ward SA, editors. The metabolism of arsenite. London: CRC Press. p. 99–114.

Bertin PN, Médigue C, Normand P. 2008. Advances in environmental genomics: towards an integrated view of micro-organisms and ecosystems. Microbiology 154:347–359.

Bigot T, Daubin V, Lassalle F, Perrière G. 2012. TPMS: a set of utilities for querying collections of gene trees. BMC Bioinformatics 14:109.

Borch T, et al. 2010. Biogeochemical redox processes and their impact on contaminant dynamics. Environ Sci Technol. 44:15–23.

Bradford MM. 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem. 72:248–254.

Branco R, Chung A-P, Morais PV. 2008. Sequencing and expression of two arsenic resistance operons with different functions in the highly arsenic-resistant strain *Ochrobactrum tritici* SCII24T. BMC Microbiol. 8:95.

Bryan CG, et al. 2009. Carbon and arsenic metabolism in *Thiomonas* strains: differences revealed diverse adaptation processes. BMC Microbiol. 9:127.

Carapito C, et al. 2006. Identification of genes and proteins involved in the pleiotropic response to arsenic stress in *Caenibacter arsenoxydans*, a metalloresistant beta-proteobacterium with an unsequenced genome. Biochimie 88:595–606.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17:540–552.

Clarke MB, Sperandio V. 2005. Transcriptional regulation of flhDC by QseBC and sigma (FliA) in enterohaemorrhagic *Escherichia coli*. Mol Microbiol. 57:1734–1749.

Cleiss-Arnold J, et al. 2010. Temporal transcriptomic response during arsenic stress in *Herminiimonas arsenicoxydans*. BMC Genomics 11:709.

Creus CM, et al. 2005. Nitric oxide is involved in the *Azospirillum brasilense*-induced lateral root formation in tomato. Planta 221:297–303.

Criscuolo A, Gribaldo S. 2010. BMGE (block mapping and gathering with entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol. 10:210.

Danhorn T, Hentzer M, Givskov M, Parsek MR, Fuqua C. 2004. Phosphorus limitation enhances biofilm formation of the plant pathogen *Agrobacterium tumefaciens* through the PhoR-PhoB regulatory system. J Bacteriol. 186:4492–4501.

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

Dénarié J, Debellé F, Promé JC. 1996. *Rhizobium* lipo-chitooligosaccharide nodulation factors: signaling molecules mediating recognition and morphogenesis. Annu Rev Biochem. 65:503–535.

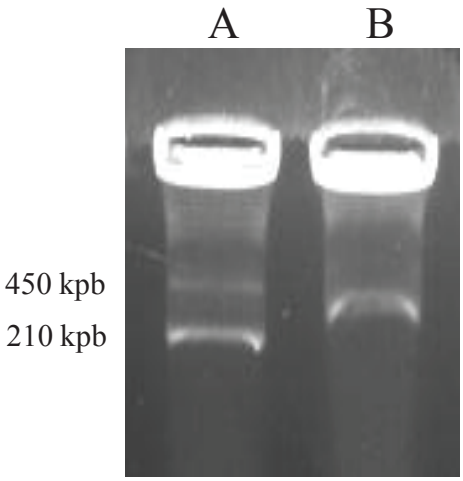
Didelot X, Lawson D, Darling A, Falush D. 2010. Inference of homologous recombination in bacteria using whole-genome sequences. Genetics 186:1435–1449.

Dobbelaere S, Okon Y. 2007. The plant growth-promoting effect and plant responses. In: Elmerich C, Newton WE, editors. Associative and endophytic nitrogen-fixing bacteria and cyanobacterial associations. Nitrogen fixation: origins, applications, and research progress V. Dordrecht (The Netherlands): Springer. p. 145–170.

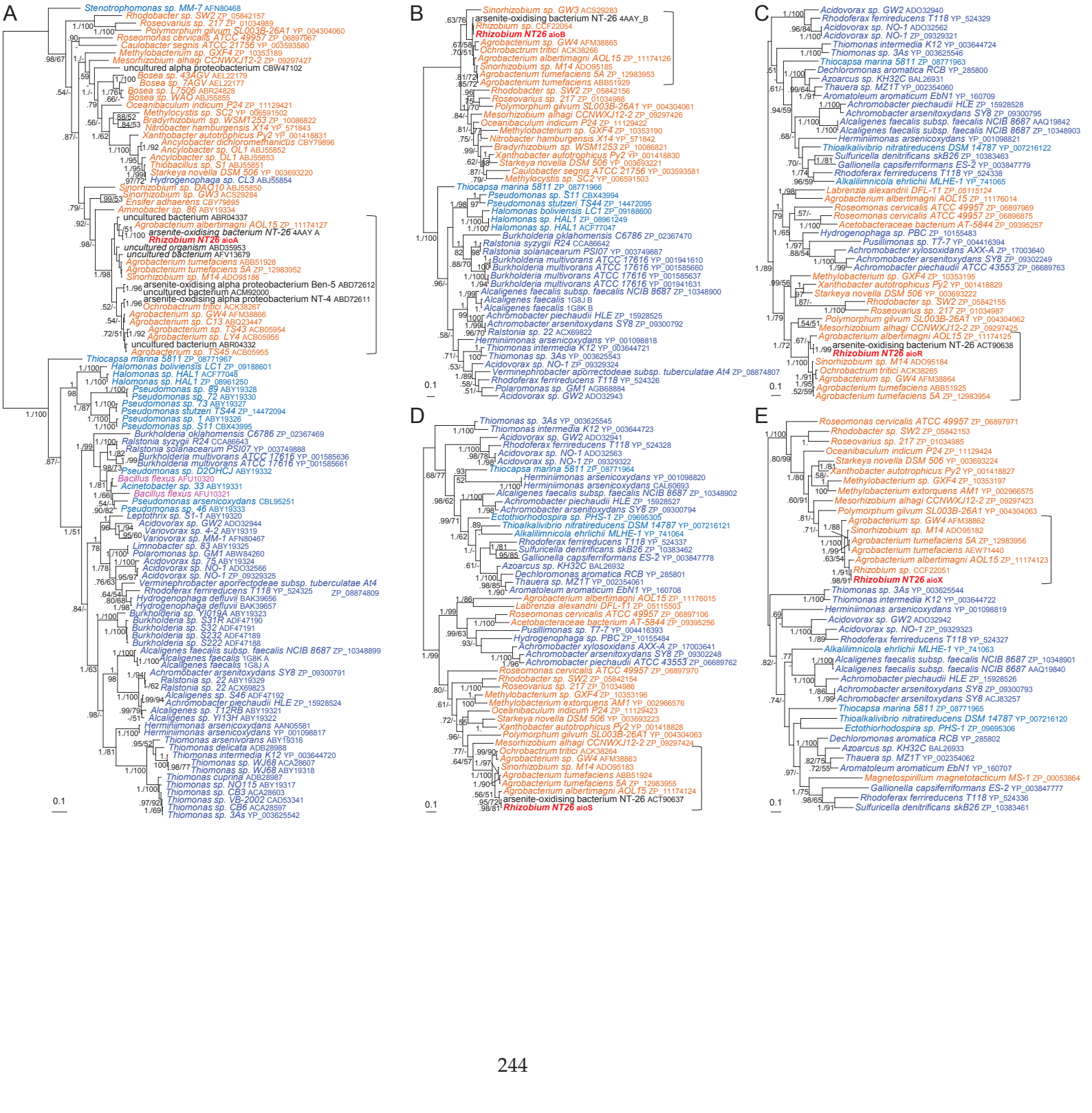
Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.
Eraso JM, et al. 2008. Role of the global transcriptional regulator PrrA in *Rhodobacter sphaeroides* 2.4.1: combined transcriptome and proteome analysis. J Bacteriol. 190:4831–4848.
Gadd GM. 2010. Metals, minerals and microbes: geomicrobiology and bioremediation. Microbiology 156:609–643.
Gallie DR, Kado CI. 1988. Minimal region necessary for autonomous replication of pTAR. J Bacteriol. 170:3170–3176.
Giraud E, et al. 2007. Legumes symbioses: absence of *Nod* genes in photosynthetic bradyrhizobia. Science 316:1307–1312.
Gonzalez V, et al. 2006. The partitioned *Rhizobium etli* genome: genetic and metabolic redundancy in seven interacting replicons. Proc Natl Acad Sci U S A. 103:3834–3839.
Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol. 27:221–224.
Gremaud MF, Harper JE. 1989. Selection and initial characterization of partially nitrate tolerant nodulation mutants of soybean. Plant Physiol. 89:169–173.
Guibaud G, van Hullebusch E, Bordas F. 2006. Lead and cadmium biosorption by extracellular polymeric substances (EPS) extracted from activated sludges: pH-sorption edge tests and mathematical equilibrium modelling. Chemosphere 64:1955–1962.
Guindon S, Delsuc F, Dufayard J-F, Gascuel O. 2009. Estimating maximum likelihood phylogenies with PhyML. Methods Mol Biol. 537:113–137.
Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52: 696–704.
Hall-Stoodley L, Costerton JW, Stoodley P. 2004. Bacterial biofilms: from the natural environment to infectious diseases. Nat Rev Microbiol. 2: 95–108.
Hao Y, Charles TC, Glick BR. 2011. ACC deaminase activity in avirulent *Agrobacterium tumefaciens* D3. Can J Microbiol. 57:278–286.
Harrison PW, Lower RPJ, Kim NKD, Young JPW. 2010. Introducing the bacterial 'chromid': not a chromosome, not a plasmid. Trends Microbiol. 18:141–148.
Heinrich-Salmeron A, et al. 2011. Unsuspected diversity of arsenite-oxidizing bacteria as revealed by widespread distribution of the aoxB gene in prokaryotes. Appl Environ Microbiol. 77:4685–4692.
Holmes A, et al. 2009. Comparison of two multimetal resistant bacterial strains: *Enterobacter* sp. YSU and *Stenotrophomonas maltophilia* ORO2. Curr Microbiol. 59:526–531.
Hommais F, et al. 2002. Effect of mild acid pH on the functioning of bacterial membranes in *Vibrio cholerae*. Proteomics 2:571–579.
Hynes MF, McGregor NF. 1990. Two plasmids other than the nodulation plasmid are necessary for formation of nitrogen-fixing nodules by *Rhizobium leguminosarum*. Mol Microbiol. 4:567–574.
Janssen PJ, et al. 2010. The complete genome sequence of *Cupriavidus metallidurans* strain CH34, a master survivalist in harsh and anthropogenic environments. PLoS One 5:e10433.
Juhas M, et al. 2009. Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol Rev. 33:376–393.
Kang Y-S, Bothner B, Rensing C, McDermott TR. 2012. Involvement of RpoN in regulating bacterial arsenite oxidation. Appl Environ Microbiol. 78:5638–5645.
Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform. 9:286–298.
Kirov SM, Castrisios M, Shaw JG. 2004. *Aeromonas flagella* (polar and lateral) are enterocyte adhesins that contribute to biofilm formation on surfaces. Infect Immun. 72:1939–1945.
Koechler S, et al. 2010. Multiple controls affect arsenite oxidase gene expression in *Herminiimonas arsenicoxydans*. BMC Microbiol. 10:53.
Lassalle F, et al. 2011. Genomic species are ecological species as revealed by comparative genomics in *Agrobacterium tumefaciens*. Genome Biol Evol. 3:762–781.
Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Mol Biol Evol. 25:1307–1320.
Lett M-C, Paknikar KM, Lièvremont D. 2001. A simple and rapid method for arsenite and arsenate speciation. In: Ciminelli VST, Garcia O Jr, editors. Biohydrometallurgy: fundamentals, technology and sustainable development, Part B. New York: Elsevier Science. p. 541–546.
Lieutaud A, et al. 2010. Arsenite oxidase from *Ralstonia* sp. 22. J Biol Chem. 285:20433–20441.
Liu G, et al. 2012. A periplasmic arsenite-binding protein involved in regulating arsenite oxidation. Environ Microbiol. 14:1624–1634.
López-Guerrero MG, et al. 2012. Rhizobial extrachromosomal replicon variability, stability and expression in natural niches. Plasmid 68: 149–158.
Marchal M, Briandet R, Koechler S, Kammerer B, Bertin PN. 2010. Effect of arsenite on swimming motility delays surface colonization in *Herminiimonas arsenicoxydans*. Microbiology 156:2336–2342.
Marchal M, et al. 2011. Subinhibitory arsenite concentrations lead to population dispersal in *Thiomonas* sp. PLoS One 6:e23181.
Marouga R, David S, Hawkins E. 2005. The development of the DIGE system: 2D fluorescence difference gel analysis technology. Anal Bioanal Chem. 382:669–678.
Martens M, et al. 2007. Multilocus sequence analysis of Ensifer and related taxa. Int J Syst Evol Microbiol. 57:489–503.
McDougald D, Rice SA, Barraud N, Steinberg PD, Kjelleberg S. 2012. Should we stay or should we go: mechanisms and ecological consequences for biofilm dispersal. Nat Rev Microbiol. 10:39–50.
Muller D, et al. 2007. A tale of two oxidation states: bacterial colonization of arsenic-rich environments. PLoS Genet. 3:e53.
Muller D, Lièvremont D, Simeonova DD, Hubert J-C, Lett M-C. 2003. Arsenite oxidase aox genes from a metal-resistant β-proteobacterium. J Bacteriol. 185:135–141.
Nejidat A, Saadi I, Ronen Z. 2008. Effect of flagella expression on adhesion of *Achromobacter piechaudii* to chalk surfaces. J Appl Microbiol. 105: 2009–2014.
Osborne TH, Santini JM. 2012. Prokaryotic aerobic oxidation of arsenite. In: Santini JM, Ward SA, editors. The metabolism of arsenite. London: CRC Press. p. 61–72.
Parvatiyar K, et al. 2005. Global analysis of cellular factors and responses involved in *Pseudomonas aeruginosa* resistance to arsenite. J Bacteriol. 187:4853–4864.
Penel S, et al. 2009. Databases of homologous gene families for comparative genomics. BMC Bioinformatics 10(6 Suppl):S3.
Ramírez-Bahena MH, Nesme X, Muller D. 2012. Rapid and simultaneous detection of linear chromosome and large plasmids in proteobacteria. J Basic Microbiol. 52:736–739.
Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16:276–277.
Richardson AE, Barea J-M, McNeill AM, Prigent-Combaret C. 2009. Acquisition of phosphorus and nitrogen in the rhizosphere and plant growth promotion by microorganisms. Plant Soil. 321:305–339.
Rigaud J, Puppo A. 1975. Indole-3-acetic acid catabolism by soybean bacteroids. J Gen Microbiol. 88:223–228.
Ronquist F, et al. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 61:539–542.
Santini JM, et al. 2007. The NT-26 cytochrome c552 and its role in arsenite oxidation. Biochim Biophys Acta. 1767:189–196.
Santini JM, Sly LI, Schnagl RD, Macy JM. 2000. A new chemolithoautotrophic arsenite-oxidizing bacterium isolated from a gold mine: phylogenetic, physiological, and preliminary biochemical studies. Appl Environ Microbiol. 66:92–97.

Santini JM, vanden Hoven RN. 2004. Molybdenum-containing arsenite oxidase of the chemolithoautotrophic arsenite oxidizer NT-26. J Bacteriol. 186:1614–1619.
Sardiwal S, Santini JM, Osborne TH, Djordjevic S. 2010. Characterization of a two-component signal transduction system that controls arsenite oxidation in the chemolithoautotroph NT-26. FEMS Microbiol Lett. 313:20–28.
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.
Sourjik V, Muschler P, Scharf B, Schmitt R. 2000. VisN and VisR are global regulators of chemotaxis, flagellar, and motility genes in *Sinorhizobium (Rhizobium) meliloti*. J Bacteriol. 182:782–788.
Spaepen S, Vanderleyden J, Okon Y. 2009. Chapter 7: Plant growth-promoting actions of rhizobacteria. In: Van Loon LC, editor. Advances in botanical research, Vol. 51. Burlington: Academic Press. p. 283–320.
Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.
Stolz JF. 2011. Microbial metal and metalloid metabolism: advances and applications, 1st ed. Washington DC: ASM Press.
Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34:W609–W612.
Tang X, Lu BF, Pan SQ. 1999. A bifunctional transposon mini-Tn5gfp-km which can be used to select for promoter fusions and report gene expression levels in *Agrobacterium tumefaciens*. FEMS Microbiol Lett. 179:37–42.
Vallenet D, et al. 2006. MaGe: a microbial genome annotation system supported by synteny results. Nucleic Acids Res. 34:53–65.
Vallenet D, et al. 2009. MicroScope: a platform for microbial genome annotation and comparative genomics. Database (Oxford) 2009: bap021.
van Lis R, Nitschke W, Duval S, Schoepp-Cothenet B. 2012. Evolution of arsenite oxidation. In: Santini JM, Ward SA, editors. The metabolism of arsenite. London: CRC Press. p. 125–144.
van Lis R, Nitschke W, Duval S, Schoepp-Cothenet B. 2013. Arsenics as bioenergetic substrates. Biochim Biophys Acta. 1827: 176–188.
Weiss S, et al. 2009. Enhanced structural and functional genome elucidation of the arsenite-oxidizing strain *Herminiimonas arsenicoxydans* by proteomics data. Biochimie 91:192–203.
Wilkins MJ, et al. 2009. Proteogenomic monitoring of Geobacter physiology during stimulated uranium bioremediation. Appl Environ Microbiol. 75:6591–6599.
Wood DW, et al. 2001. The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. Science 294:2317–2323.
Young JM, Kuykendall LD, Martínez-Romero E, Kerr A, Sawada H. 2001. A revision of *Rhizobium* Frank 1889, with an emended description of the genus, and the inclusion of all species of *Agrobacterium* Conn 1942 and *Allorhizobium* undicola de Lajudie et al. 1998 as new combinations: *Rhizobium radiobacter*, *R. rhizogenes*, *R. rubi*, *R. undicola* and *R. vitis*. Int J Syst Evol Microbiol. 51:89–103.
Young JPW, et al. 2006. The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. Genome Biol. 7: R34.
Zheng M, et al. 2001. Computation-directed identification of OxyR DNA binding sites in *Escherichia coli*. J Bacteriol. 183:4571–4579.
Zhong Z, et al. 2003. Nucleotide sequence based characterizations of two cryptic plasmids from the marine bacterium *Ruegeria* isolate PR1b. Plasmid 49:233–252.
Zumft WG. 1997. Cell biology and molecular basis of denitrification. Microbiol Mol Biol Rev. 61:533–616.

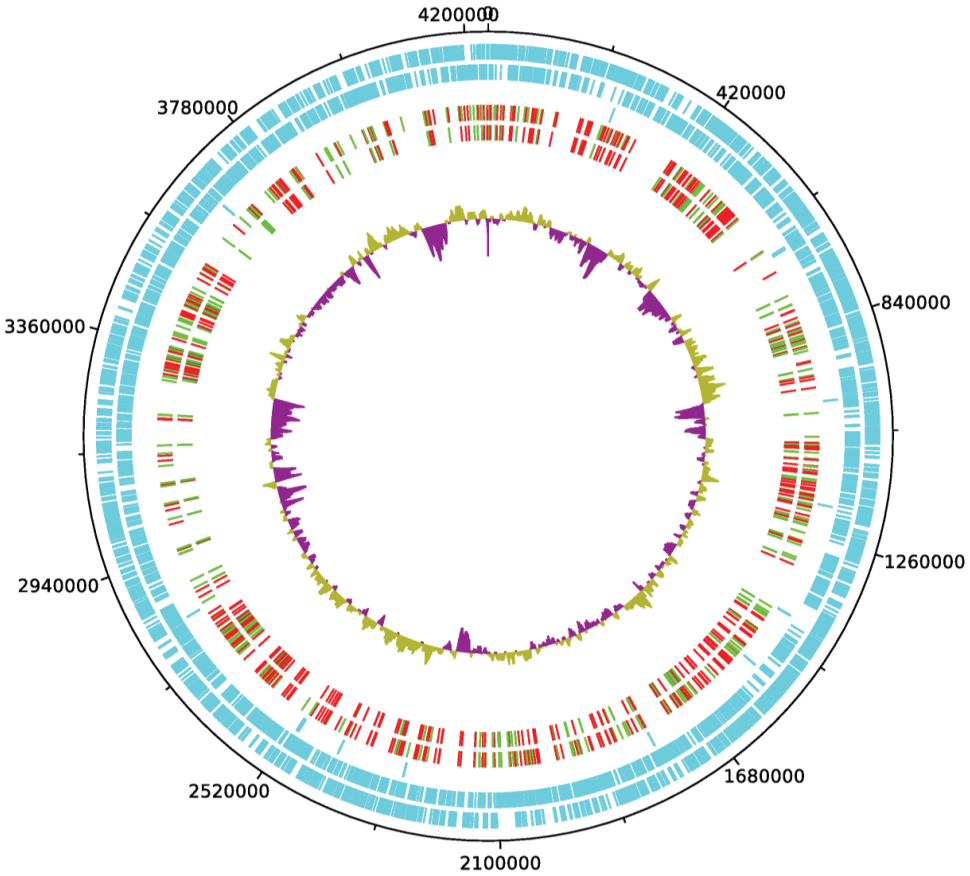
Associate editor: Purificación López-García



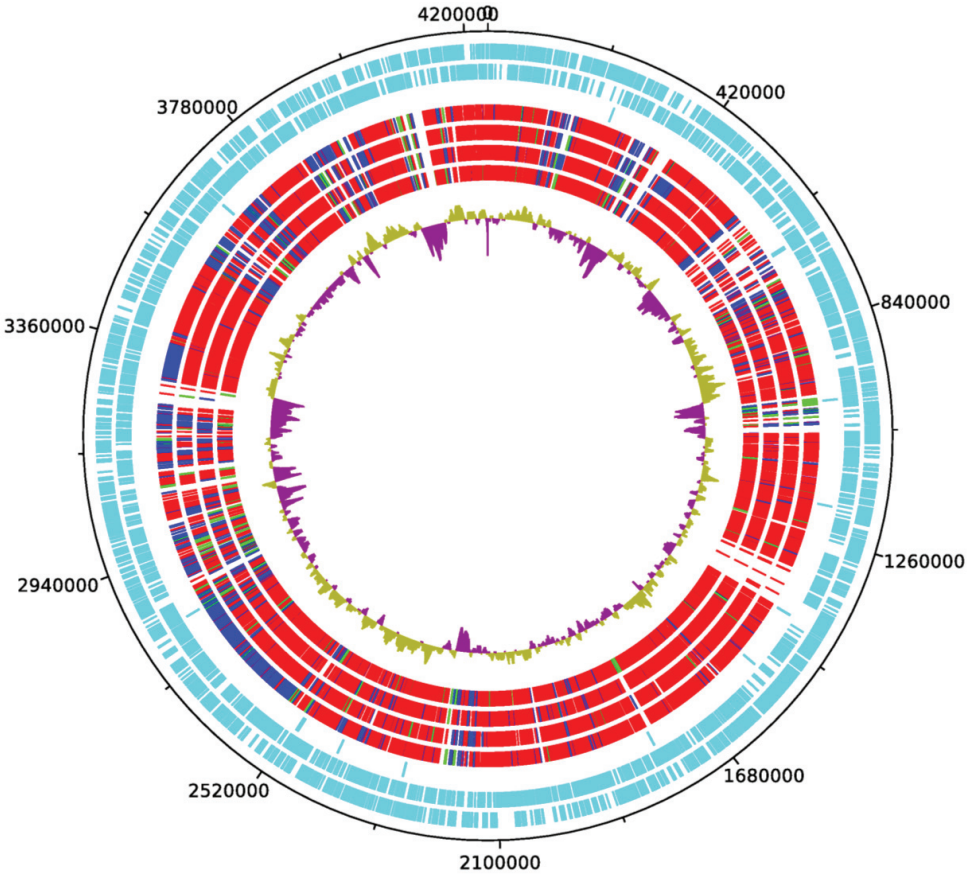
Supplementary Figure 1. Pulsed-Field Gel Electrophoresis analysis of *Rhizobium* sp. NT-26 genomic DNA. (*A*) *A. tumefaciens* C58 presenting 2 large plasmid: pAt (450 kbp) and pTi (210 kbp); (*B*) *Rhizobium* sp. NT-26 presenting the 322 kbp p1 plasmid. This electrophoretic profile was obtained by a modified Eckhardt procedure.


A
Stenotrophomonas sp. MM-7 AFN80468
Rhodobacter sp. SW2 ZP_05842157
Roseovarius sp. 217 ZP_01034989
Polymorphum gilvum SL003B-26A1 YP_004304060
Roseomonas cervicalis ATCC 49957 ZP_06897967
Caulobacter segnis ATCC 21756 YP_003593580
Methylobacterium sp. GXF4 ZP_10353189
Mesorhizobium alhagi CCNWXJ12-2 ZP_09297427
uncultured alpha proteobacterium CBW47102
Bosea sp. 43AGV AEL22179
Bosea sp. 7AGV AEL22177
Bosea sp. LT306 ABR24828
Bosea sp. WAO ABJ55855
Oceanibaculum indicum P24 ZP_11129421
Methylocystis sp. SC2 YP_006591502
Bradyrhizobium sp. WSM1253 ZP_10086822
Nitrobacter hamburgensis X14 YP_571843
Xanthobacter autotrophicus Py2 YP_001418831
Ancylobacter dichloromethanicus CBY79896
Ancylobacter sp. OL1 ABJ55852
Ancylobacter sp. OL1 ABJ55853
Thiobacillus sp. S1 ABJ55851
Starkeya novella DSM 506 YP_003693220
Hydrogenophaga sp. CL3 ABJ55854
Sinorhizobium sp. DAO10 ABJ55850
Sinorhizobium sp. GW3 ACS29284
Ensifer adhaerens CBY79895
Aminobacter sp. 86 ABY19334
uncultured bacterium ABR04337
Agrobacterium albertimagni AOL15 ZP_11174127
arsenite-oxidising bacterium NT-26 4AAY A
Rhizobium NT26 aioA
uncultured organism ABD35953
uncultured bacterium AFV13679
Agrobacterium tumefaciens ABB51928
Agrobacterium tumefaciens 5A ZP_12983952
Sinorhizobium sp. M14 ADO95186
arsenite-oxidising alpha proteobacterium Ben-5 ABD72612
uncultured bacterium ACM92000
arsenite-oxidising alpha proteobacterium NT-4 ABD72611
Ochrobactrum tritici ACK38267
Agrobacterium sp. GW4 AFM38866
Agrobacterium sp. C13 ABQ23447
Agrobacterium sp. TS43 ACB05954
Agrobacterium sp. LY4 ACB05956
uncultured bacterium ABR04332
Agrobacterium sp. TS45 ACB05955
Thiocapsa marina 5811 ZP_08771967
Halomonas boliviensis LC1 ZP_09188601
Halomonas sp. HAL1 ACF77048
Halomonas sp. HAL1 ZP_08961250
Pseudomonas sp. 89 ABY19328
Pseudomonas sp. 72 ABY19330
Pseudomonas sp. 73 ABY19327
Pseudomonas stutzeri TS44 ZP_14472094
Pseudomonas sp. 1 ABY19326
Pseudomonas sp. S11 CBX43995
Burkholderia oklahomensis C6786 ZP_02367469
Ralstonia syzygii R24 CCA86643
Ralstonia solanacearum PSI07 YP_003749888
Burkholderia multivorans ATCC 17616 YP_001585636
Burkholderia multivorans ATCC 17616 YP_001585661
Pseudomonas sp. D2OHCJ ABY19332
Bacillus flexus AFU10320
Acinetobacter sp. 33 ABY19331
Bacillus flexus AFU10321
Pseudomonas arsenicoxydans CBL95251
Pseudomonas sp. 46 ABY19333
Leptothrix sp. S1-1 ABY19320
Acidovorax sp. GW2 ADO32944
Variovorax sp. 4-2 ABY19319
Variovorax sp. MM-1 AFN80467
Limnobacter sp. 83 ABY19325
Polaromonas sp. GM1 ABW84260
Acidovorax sp. 75 ABY19324
Acidovorax sp. NO-1 ADO32566
Acidovorax sp. NO-1 ZP_09329325
Verminephrobacter aporrectodeae subsp. tuberculatae At4 ZP_08874809
Rhodoferax ferrireducens T118 YP_524325
Hydrogenophaga defluvii BAK39656
Hydrogenophaga defluvii BAK39657
Burkholderia sp. YI019A ABY19323
Burkholderia sp. S31R ADF47190
Burkholderia sp. S32 ADF47191
Burkholderia sp. S232 ADF47189
Burkholderia sp. S222 ADF47188
Alcaligenes faecalis subsp. faecalis NCIB 8687 ZP_10348899
Alcaligenes faecalis 1G8K A
Alcaligenes faecalis 1G8J A
Achromobacter arsenitoxydans SY8 ZP_09300791
Ralstonia sp. 22 ABY19329
Ralstonia sp. 22 ACX69823
Alcaligenes sp. S46 ADF47192
Achromobacter piechaudii HLE ZP_15928524
Alcaligenes sp. T12RB ABY19321
Alcaligenes sp. YI13H ABY19322
Herminiimonas arsenicoxydans AAN05581
Herminiimonas arsenicoxydans YP_001098817
Thiomonas arsenivorans ABY19316
Thiomonas delicata ADB28988
Thiomonas intermedia K12 YP_003644720
Thiomonas sp. WJ68 ACA28607
Thiomonas sp. WJ68 ABY19318
Thiomonas cuprina ADB28987
Thiomonas sp. NO115 ABY19317
Thiomonas sp. CB3 ACA28603
Thiomonas sp. VB-2002 CAD53341
Thiomonas sp. CB6 ACA28597
Thiomonas sp. 3As ZP_003625542
0.1
B
Sinorhizobium sp. GW3 ACS29283
arsenite-oxidising bacterium NT-26 4AAY_B
Rhizobium sp. CCF22054
Rhizobium NT26 aioB
Agrobacterium sp. GW4 AFM38865
Ochrobactrum tritici ACK38266
Agrobacterium albertimagni AOL15 ZP_11174126
Sinorhizobium sp. M14 ADO95185
Agrobacterium tumefaciens 5A ZP_12983953
Agrobacterium tumefaciens ABB51929
Rhodobacter sp. SW2 ZP_05842156
Roseovarius sp. 217 ZP_01034988
Polymorphum gilvum SL003B-26A1 YP_004304061
Mesorhizobium alhagi CCNWXJ12-2 ZP_09297426
Oceanibaculum indicum P24 ZP_11129422
Methylobacterium sp. GXF4 ZP_10353190
Nitrobacter hamburgensis X14 YP_571842
Bradyrhizobium sp. WSM1253 ZP_10088821
Xanthobacter autotrophicus Py2 YP_001418830
Starkeya novella DSM 506 YP_003693221
Caulobacter segnis ATCC 21756 YP_003593581
Methylocystis sp. SC2 YP_006591503
Thiocapsa marina 5811 ZP_08771966
Pseudomonas sp. S11 CBX43994
Pseudomonas stutzeri TS44 ZP_14472095
Halomonas boliviensis LC1 ZP_09188600
Halomonas sp. HAL1 ZP_08961249
Halomonas sp. HAL1 ACF77047
Burkholderia oklahomensis C6786 ZP_02367470
Ralstonia syzygii R24 CCA86642
Ralstonia solanacearum PSI07 YP_003749887
Burkholderia multivorans ATCC 17616 YP_001941610
Burkholderia multivorans ATCC 17616 YP_001585660
Burkholderia multivorans ATCC 17616 YP_001585637
Burkholderia multivorans ATCC 17616 YP_001941631
Alcaligenes faecalis subsp. faecalis NCIB 8687 ZP_10348900
Alcaligenes faecalis 1G8J B
Alcaligenes faecalis 1G8K B
Achromobacter piechaudii HLE ZP_15928525
Achromobacter arsenitoxydans SY8 ZP_09300792
Ralstonia sp. 22 ACX69822
Herminiimonas arsenicoxydans YP_001098818
Thiomonas intermedia K12 YP_003644721
Thiomonas sp. 3As YP_003625543
Acidovorax sp. NO-1 ZP_09329324
Verminephrobacter aporrectodeae subsp. tuberculatae At4 ZP_08874807
Rhodoferax ferrireducens T118 YP_524326
Polaromonas sp. GM1 AGB68884
Acidovorax sp. GW2 ADO32943
0.1
C
Acidovorax sp. GW2 ADO32940
Rhodoferax ferrireducens T118 YP_524329
Acidovorax sp. NO-1 ADO32562
Acidovorax sp. NO-1 ZP_09329321
Thiomonas intermedia K12 YP_003644724
Thiomonas sp. 3As YP_003625546
Thiocapsa marina 5811 ZP_08771963
Dechloromonas aromatica RCB YP_285800
Azoarcus sp. KH32C BAL26931
Thauera sp. MZ1T YP_002354060
Aromatoleum aromaticum EbN1 YP_160709
Achromobacter piechaudii HLE ZP_15928528
Achromobacter arsenitoxydans SY8 ZP_09300795
Alcaligenes faecalis subsp. faecalis NCIB 8687 AAQ19842
Alcaligenes faecalis subsp. faecalis NCIB 8687 ZP_10348903
Herminiimonas arsenicoxydans YP_001098821
Thioalkalivibrio nitratireducens DSM 14787 YP_007216122
Sulfuricella denitrificans skB26 ZP_10383463
Gallionella capsiferriformans ES-2 YP_003847779
Rhodoferax ferrireducens T118 YP_524338
Alkalilimnicola ehrlichii MLHE-1 YP_741065
Labrenzia alexandrii DFL-11 ZP_05115124
Agrobacterium albertimagni AOL15 ZP_11176014
Roseomonas cervicalis ATCC 49957 ZP_06897969
Roseomonas cervicalis ATCC 49957 ZP_06896875
Acetobacteraceae bacterium AT-5844 ZP_09395257
Hydrogenophaga sp. PBC ZP_10155483
Pusillimonas sp. T7-7 YP_004416394
Achromobacter xylosoxidans AXX-A ZP_17003640
Achromobacter arsenitoxydans SY8 ZP_09302249
Achromobacter piechaudii ATCC 43553 ZP_06689763
Methylobacterium sp. GXF4 ZP_10353195
Xanthobacter autotrophicus Py2 YP_001418829
Starkeya novella DSM 506 YP_003693222
Rhodobacter sp. SW2 ZP_05842155
Roseovarius sp. 217 ZP_01034987
Polymorphum gilvum SL003B-26A1 YP_004304062
Mesorhizobium alhagi CCNWXJ12-2 ZP_09297425
Agrobacterium albertimagni AOL15 ZP_11174125
arsenite-oxidising bacterium NT-26 ACT90638
Rhizobium NT26 aioR
Sinorhizobium sp. M14 ADO95184
Ochrobactrum tritici ACK38265
Agrobacterium sp. GW4 AFM38864
Agrobacterium tumefaciens ABB51925
Agrobacterium tumefaciens 5A ZP_12983954
0.1
D
Thiomonas sp. 3As YP_003625545
Thiomonas intermedia K12 YP_003644723
Acidovorax sp. GW2 ADO32941
Rhodoferax ferrireducens T118 YP_524328
Acidovorax sp. NO-1 ADO32563
Acidovorax sp. NO-1 ZP_09329322
Thiocapsa marina 5811 ZP_08771964
Herminiimonas arsenicoxydans YP_001098820
Herminiimonas arsenicoxydans CAL60693
Alcaligenes faecalis subsp. faecalis NCIB 8687 ZP_10348902
Achromobacter piechaudii HLE ZP_15928527
Achromobacter arsenitoxydans SY8 ZP_09300794
Ectothiorhodospira sp. PHS-1 ZP_09695305
Thioalkalivibrio nitratireducens DSM 14787 YP_007216121
Alkalilimnicola ehrlichii MLHE-1 YP_741064
Rhodoferax ferrireducens T118 YP_524337
Sulfuricella denitrificans skB26 ZP_10383462
Gallionella capsiferriformans ES-2 YP_003847778
Azoarcus sp. KH32C BAL26932
Dechloromonas aromatica RCB YP_285801
Thauera sp. MZ1T YP_002354061
Aromatoleum aromaticum EbN1 YP_160708
Agrobacterium albertimagni AOL15 ZP_11176015
Labrenzia alexandrii DFL-11 ZP_05115503
Roseomonas cervicalis ATCC 49957 ZP_06897106
Acetobacteraceae bacterium AT-5844 ZP_09395256
Pusillimonas sp. T7-7 YP_004416393
Hydrogenophaga sp. PBC ZP_10155484
Achromobacter xylosoxidans AXX-A ZP_17003641
Achromobacter arsenitoxydans SY8 ZP_09302248
Achromobacter piechaudii ATCC 43553 ZP_06689762
Roseomonas cervicalis ATCC 49957 ZP_06897970
Rhodobacter sp. SW2 ZP_05842154
Roseovarius sp. 217 ZP_01034986
Methylobacterium sp. GXF4 ZP_10353196
Methylobacterium extorquens AM1 YP_002966576
Oceanibaculum indicum P24 ZP_11129423
Starkeya novella DSM 506 YP_003693223
Xanthobacter autotrophicus Py2 YP_001418828
Polymorphum gilvum SL003B-26A1 YP_004304063
Mesorhizobium alhagi CCNWXJ12-2 ZP_09297424
Ochrobactrum tritici ACK38264
Agrobacterium sp. GW4 AFM38863
Sinorhizobium sp. M14 ADO95183
Agrobacterium tumefaciens ABB51924
Agrobacterium tumefaciens 5A ZP_12983955
Agrobacterium albertimagni AOL15 ZP_11174124
arsenite-oxidising bacterium NT-26 ACT90637
Rhizobium NT26 aioS
0.1
E
Roseomonas cervicalis ATCC 49957 ZP_06897971
Rhodobacter sp. SW2 ZP_05842153
Roseovarius sp. 217 ZP_01034985
Oceanibaculum indicum P24 ZP_11129424
Starkeya novella DSM 506 YP_003693224
Xanthobacter autotrophicus Py2 YP_001418827
Methylobacterium sp. GXF4 ZP_10353197
Methylobacterium extorquens AM1 YP_002966575
Mesorhizobium alhagi CCNWXJ12-2 ZP_09297423
Polymorphum gilvum SL003B-26A1 YP_004304063
Agrobacterium sp. GW4 AFM38862
Sinorhizobium sp. M14 ADO95182
Agrobacterium tumefaciens 5A ZP_12983956
Agrobacterium tumefaciens AEW71440
Agrobacterium albertimagni AOL15 ZP_11174123
Rhizobium sp. CCF22051
Rhizobium NT26 aioX
Thiomonas sp. 3As YP_003625544
Thiomonas intermedia K12 YP_003644722
Herminiimonas arsenicoxydans YP_001098819
Acidovorax sp. GW2 ADO32942
Acidovorax sp. NO-1 ZP_09329323
Rhodoferax ferrireducens T118 YP_524327
Alkalilimnicola ehrlichii MLHE-1 YP_741063
Alcaligenes faecalis subsp. faecalis NCIB 8687 ZP_10348901
Alcaligenes faecalis subsp. faecalis NCIB 8687 AAQ19840
Achromobacter piechaudii HLE ZP_15928526
Achromobacter arsenitoxydans SY8 ZP_09300793
Achromobacter arsenitoxydans SY8 ACJ83257
Thiocapsa marina 5811 ZP_08771965
Thioalkalivibrio nitratireducens DSM 14787 YP_007216120
Ectothiorhodospira sp. PHS-1 ZP_09695306
Dechloromonas aromatica RCB YP_285802
Azoarcus sp. KH32C BAL26933
Thauera sp. MZ1T YP_002354062
Aromatoleum aromaticum EbN1 YP_160707
Magnetospirillum magnetotacticum MS-1 ZP_00053864
Gallionella capsiferriformans ES-2 YP_003847777
Rhodoferax ferrireducens T118 YP_524336
Sulfuricella denitrificans skB26 ZP_10383461
0.1

Supplementary Figure 2. Bayesian phylogenetic trees of Aio proteins. AioA (*A*) (109 sequences, 352 amino acid positions), AioB (*B*) (49 sequences, 153 amino acid positions), AioR (*C*) (46 sequences, 291 amino acid positions), AioS (*D*) (49 sequences, 214 amino acid positions) and AioX (*E*) (40 sequences, 201 amino acid positions) inferred with MrBayes. The sequences from *Alphaproteobacteria*, *Betaproteobacteria*, *Gammaproteobacteria* and *Firmicutes* are shown in orange, dark blue, light blue and pink, respectively, whereas the *Rhizobium* sp. NT-26 sequences are in red. Numbers at nodes represent posterior probabilities and bootstrap values computed with MrBayes and PhyML, respectively. For clarity reasons, only values greater than 0.50 and 50% are shown. Scale bars indicate the estimated average number of substitutions per site. The subgroups containing the *Rhizobium* sp. NT-26 sequences discussed in the paper are indicated by brackets.



Supplementary Figure 3. Distribution of genes matching "NT26inAgro" or "NT26outAgro" patterns around the *Rhizobium* sp. NT-26 chromosome. From the outside, circles display: (1) coordinates in bp; (2), (3) and (4) genes on direct strand, reverse strand, and pseudogenes, respectively; (5) and (6) genes matching the simple patterns and patterns with support higher than 90%, respectively; in red: "NT26inAgro", in green: "NT26outAgro"; (7) GC % deviation from the genomic mean, represented in yellow and purple for high and low GC %, respectively.



Supplementary Figure 4. Replicon location of *Rhizobium* sp. NT-26 gene homologs in related *Rhizhobiaceae* genomes. From the outside, circles display: (1) coordinates in bp; (2), (3) and (4) genes on direct strand, reverse strand, and pseudogenes, respectively; (5), (6), (7) and (8) represent replicon location of closest homolog in *A. tumefaciens* C58, *A. vitis* S4, *R. rhizogenes* K84, and *R. etli* CFN42, respectively. Represented in red: genes that are located on the primary chromosome, in blue: on the secondary chromosome/chromid (include p42b, p42e and p42f), and in green: on the plasmids; (9) GC % deviation from the genomic mean, represented in yellow and purple for high and low GC %, respectively.

5.2 Acquisition of protelomerase and linearization of secondary chromosome led to the emmergence of a major clade within Rhizobiaceae

5.2.1 Introduction

The following manuscript was recently submitted to the journal *Molecular Phylogeny and Evolution*. It demonstrates of the exact coincidence of the acquisition of a pro-telomerase gene and the linearization of the secondary chromosome of a lineage of Rhizobiaceae that includes *A. tumefaciens* species complex, and discusses of the implication of this linearization event in the speciation of the clade.

My personal contribution in this work consists in the phylogenetic comparison of the agrobacterial TelA protein with distant protelomerase homologs, and the deduction of the evolutionary origin of this trait.

5.2.2 Manuscript

Single acquisition of protelomerase gave rise to speciation of a subclade within the *Agrobacterium*/*Rhizobium* supercluster characterized by the presence of a linear chromid.

Martha H. Ramírez-Bahena[1,2,3,4], Ludovic Vial[1,2,3], Florent Lassalle[1,2,3,4,5,6], Benjamin Diel[1,2,3], David Chapulliot[1,2,3,4], Vincent Daubin[1,2,5], Xavier Nesme[1,2,3,4], and Daniel Muller[1,2,3]

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

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

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

[4] *INRA, USC 1364, Ecologie Microbienne, 69622 Villeurbanne, France*

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

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

Corresponding author: X. Nesme
Fax: +33 4 72 44 82 89
E-mail: nesme@univ-lyon1.fr

Abstract

Linear chromosomes are atypical in bacteria and likely a secondary trait derived from ancestral circular molecules. Within the Rhizobiaceae family, whose genome contains at least two chromosomes, a particularity of *Agrobacterium fabrum* (formerly *A. tumefaciens*) secondary chromosome (chromid) is to be linear and hairpin-ended thanks to the TelA protelomerase. However, no data are available on linear chromid emergence and radiation through the Rhizobiaceae family. Linear topology and *telA* distribution within this bacterial family was thus screened by pulse field gel electrophoresis and PCR. In *A. rubi, A. larrymoorei*, *Rhizobium skierniewicense, A. viscosum, A. sp.* NCPPB 1650, and every genomospecies of the *A. tumefaciens* species complex (including *R. pusense*, *A. radiobacter, A. fabrum* and *R. nepotum*), linear chromid topologies were retrieved concomitantly with *telA* presence, whereas the remote species *A. vitis, Allorhizobium undicola*, *Rhizobium rhizogenes* and *Ensifer meliloti* harbored a circular chromid as well as no *telA* gene. Moreover, the *telA* phylogeny is congruent with that of *recA* used as a marker gene of the *Agrobacterium* phylogeny. Collectively, these findings strongly suggest that single acquisition of *telA* by an ancestor was the founding event in the speciation of a particular Rhizobiaceae subclade characterized by the presence of a linear chromid. This subclade, characterized by an unusual genome architecture, appears to be a relevant candidate to serve as a basis for redefining the controversial *Agrobacterium* genus.

Keywords: species complex, bacterial speciation, genome architecture, protelomerase, TelA.

1. Introduction

Possession of a single circular chromosome, sometimes accompanied by extrachromosal elements, is generally the characteristic of bacteria. However, the development of electrophoretic tools to visualize large DNA molecules and genomics has led to the discovery of linear chromosomes in bacterial cells (Volff and Altenbuchner, 2000). Linear replicons were first described in the spirochete *Borrelia burgdorferi* (Ferdows and Barbour, 1989), and then in genomes of other bacteria such as members of the *Actinobacteria* phylum (Lin et al., 1993), the alphaproteobacterial Rhizobiaceae *Agrobacterium tumefaciens* (Allardet-Servent et al., 1993), and more recently the Cyanobacteria *Cyanothece* 51142 (Bandyopadhyay et al., 2011).

The emergence of linear replicons from circular molecules requires specific systems to protect replicon ends (telomeres) against nuclease activity, and to fully replicate telomeres, since DNA polymerase requires both templates and the 3' end of a primer that is usually provided by the complementary strand. These problems have been solved in at least three independent ways throughout evolution. In eukaryotic cells, telomeres are regenerated by telomerases, i.e. specific DNA polymerases that synthesize DNA sequence repeats at the 3' end of DNA strands (Grandin and Charbonneau, 2008). In actinomycetes, telomere replication is protein primed, i.e. telomeres bind a specific protein to the 5'-ends, thus priming DNA synthesis. In *Borrelia* spp., *A. tumefaciens* C58 or several phages, telomeres form covalently closed hairpin loops presenting an uninterrupted DNA chain to the replication machinery (Casjens, 1999; Huang et al., 2012; Ravin, 2003). Replications through hairpin telomeres produces inverted repeat circular dimers. The replicated intermediate is a substrate for a DNA breakage and reunion enzyme that releases the linear strands (Chaconas et al., 2001). This enzyme activity is referred to as telomere resolvase (ResT) in *Borrelia* species or as protelomerase in the linear phage N15 (TelN) (Ravin, 2003) and *A. tumefaciens* C58 (TelA) (Huang et al., 2012).

Interestingly, among genomes of the Rhizobiaceae family, *Rhizobium*, *Agrobacterium, Allorhizobium* and *Ensifer* (formerly *Sinorhizobium*) genera have the unique feature of being composed of a standard circular chromosome and a second large replicon, called a chromid, and often plasmids. Chromids represent "second chromosomes" or "megaplasmids" carrying some core genes and plasmid-type replication machinery (Harrison et al., 2010). Two decades ago, the chromid of the *A. tumefaciens* strain C58 was shown to be a linear molecule (Allardet-Servent et al., 1993; Goodner et al., 2001; Wood et al., 2001). Although linear

chromids occur in *A. tumefaciens* and *A. rubi* strains, none have been detected in related Rhizobiaceae members such as *A. vitis* or *Rhizobium rhizogenes* K84 (formerly *Agrobacterium* biovar 2) (Galibert et al., 2001; Jumas-Bilak et al., 1998; Slater et al., 2013; Slater et al., 2009).

Historically, agrobacteria were characterized based on their disease phenotype, leading to a classification of strains that induced tumors, hairy roots or that were not pathogenic into *A. tumefaciens, A. rhizogenes* or *A. radiobacter*, respectively. This classification can, however, no longer be maintained since pathogenicity is determined by dispensable and exchangeable plasmids (i.e. tumor-inducing (Ti) and root-inducing (Ri) plasmids), while chromosomal traits conversely revealed strong strain similarities independently of their plasmid content (Kersters and De Ley, 1984; Sawada et al., 1993). These inadequate nomenclature concerned species epithet, but there is also a current controversy about the genus name *Agrobacterium*. Indeed, due to intermingled phylogenies of *Agrobacterium* and *Rhizobium*, it has been proposed to include all *Agrobacterium* members as well as *Allorhizobium* within the genus *Rhizobium* (Young et al., 2001). However, part of the scientific community rejected that proposal opening a still pending nomenclature controversy (Farrand et al., 2003; Young et al., 2003). This is the reason why the International Committee on Systematics of Prokaryotes subcommittee for the taxonomy of *Rhizobium* and *Agrobacterium* proposed to temporarily define *Agrobacterium* as a monophyletic clade in the *Agrobacterium/Rhizobium* supercluster until the discovery of decisive elements to provide a basis for a novel definition of this genus (Lindström and Young, 2011). Indeed, genomic-based classification has led to the delineation of a clade that pools all bacteria that were previously classified as belonging to *Agrobacterium*, except *A. rhizogenes* (Costechareyre et al., 2010). To rank *Agrobacterium* as a monophyletic group, the nomenclature was modified by transferring former *A. rhizogenes* into the *Rhizobium* genus, thus giving *R. rhizogenes*. The monophyletic *Agrobacterium* taxon defined as such is diverse, including several *bona fide* genomic species (*A. vitis*, *Allorhizobium undicola, A. rubi, A. larrymoorei, R. skierniewicense)*, an unnamed genomospecies (*A. sp*. strain NCPPB 1650), and a clade consisting of clearly different but closely related genomospecies encompassing the so-called *A. tumefaciens* species complex, which in turn contains *A. fabrum, A. radiobacter, R. nepotum, R. pusense* and seven other, yet unnamed, genomospecies (Costechareyre et al., 2010; Lindström and Young, 2011; Shams et al., 2013). However, the phyletic position of heterodox agrobacteria such as *A. viscosum* and *A. albertimagni* have not been considered in recent studies, and the presence of a linear chromid has not yet been extensively investigated

in all species/genomospecies of this taxon. Consequently, the *Agrobacterium/Rhizobium* supercluster appears to be a good model to study how a linear replicon emerged during evolution.

The aim of the current study was to investigate the evolution, radiation, and maintenance of the linear chromosome structure among members of the *Agrobacterium* clade and neighboring genera. To this end, in parallel to pulse field electrophoreses (PFGE) performed to visualize linear chromosomes, *telA* homologs in the *Agrobacterium* lineage were amplified and sequenced in a set of strains chosen to represent the whole species/genomospecies diversity of the taxon.

2. Materials and methods

2.1. Linear chromosome detection.

Bacteria used in this study are presented in Table S1. Linear replicons were detected by PFGE using a sample preparation procedure without plug preparation (Ramírez-Bahena et al., 2012). Bacterial strains grown overnight at 28°C in yeast extract mannitol (YEM) medium were harvested by centrifugation at 6000 g for 3 min at 4°C and resuspended in H_2O to achieve a concentration of 3 10^8 cells per ml. 400 µl of the cell suspension were harvested by centrifugation and washed with 500 µl of 0.3% sodium lauryl sarcosine solution. After centrifugation, the pellet was suspended in 25 µl lysis solution (lysozyme 500 U ml^{-1}, RNase A 3.15 U ml^{-1} and 13% sucrose in Tris-borate-EDTA (TBE) buffer (Euromedex)). Samples were immediately loaded into the wells of a 0.75% sodium dodecyl sulfate, 0.8% agarose gel (pulse field-certified agarose, Bio-Rad). PFGE was performed in the CHEF-DRIII Variable Angle System (Bio-Rad) in 0.5X TBE. The electrophoresis program consisted of 3.5 V cm^{-1} with a constant 106° field angle for a 50 s switch time for 0.5 h, followed by a switch time of 240 s for 20 h and finally 120 s for 4 h.

2.2. Amplification and sequencing procedures

The chromosomal *telA* gene (Atu2523) was amplified with F6682 CTAGCCATCTGCAACATGAAGA and F6683 AGCGACGTTTCGAGGTCGTT primers designed using the Primer3 program (http://frodo.wi.mit.edu/primer3/) in order to obtain a 964-bp amplification fragment in *A. tumefaciens* C58, under standard PCR conditions with the following cycles: 1 min at 94°C for DNA denaturation, 1 min at 55°C for annealing, and 1 min at 72°C for extension for 35 cycles. *recA* sequences were retrieved from GenBank

(http://www.ncbi.nlm.nih.gov) or amplified with F8198 TCTTTGCGKCTCGTAGAGGAYA and F8199 TGCAGGAAGCGGTCGGCRATSAG primers for all the strains, with the exception of *A. rubi*, *Al. undicola*, *R. skierniewicense* and *A. sp*. NCPPB 1650 for which F8925 AGGGMTCGATCATGAAGCTCG and F8928 CCATACATGATGTCCAATTC primers were used for *recA* amplification (Shams et al., 2013). *telA* and *recA* PCR products were sequenced by Genoscreen (Lille, France).

2.3. Sequence analysis

Analyses were performed using the SeaView multiplatform graphical user interface (available at http://pbil.univ-lyon1.fr/) (Gouy et al., 2010) using MUSCLE (default parameters) (Edgar, 2004) and phylogenetic tree building using PhyML (version 3.0) (Guindon et al., 2010), with a GTR or LG model of substitution for nucleotide or protein sequences, respectively, with four site categories of rate heterogeneity, an estimated proportion of invariant sites (LG/GTR+Γ4+I), a topology improved with the "best of NNI and SPR" moves. Branch supports are "SH-like" supports unless stated otherwise.

TelA proteins were retrieved using BLASTP (Altschul et al., 1990) at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov). To align highly divergent sequences, we took advantage of studies on telomerase structures carried out by Huang *et al.* (2012) to identify and align homologous core regions. Sequences were separated into three groups of homogeneous sequences: those belonging to *Agrobacterium* species, those belonging to *Borrelia* species, and others. Each group was aligned independently and alignments were subsequently reduced to the sites aligned with residues contained in the conserved region (Aihara et al., 2007; Huang et al., 2012; Shi et al., 2013) for the following reference sequences: NP_355469 (residues 218 to 442), YP_002776118 (residues 148 to 348) and 2V6E (residues 229 to 447) for groups of *Agrobacterium*, *Borrelia* and other sequences, respectively. Alignments were then put together and re-aligned.

3. Results and Discussion

3.1. Linear chromid presences in the Agrobacterium clade

The distribution of linear chromid was analyzed by PFGE among strains and species of *Agrobacterium* and neighboring genera that displayed the widest range of diversity revealed by phylogenetic markers such as the *recA* gene (Table S1). The plug-free PFGE approach revealed the presence of bands of ca. 2.0 Mb for tested strains of the *A. tumefaciens* species

176 complex as well as *A. viscosum, A. rubi, A. larrymoorei, R. skierniewicense* and the *A. sp.*
177 NCPPB1650 strain, while such bands were not retrieved for example in *A. vitis, R. rhizogenes*
178 or *E. meliloti* strains known to harbor only circular replicons (Fig. 1). Consequently, a linear
179 chromid was found exclusively in members of the monophyletic *Agrobacterium* taxon,
180 excluding the closely related genera *Ensifer* and *Rhizobium.*
181 Amongst members of the *A. tumefaciens* complex, linear chromosome sizes ranged from ca.
182 1.9 Mb in genomovar G9 to ca. 2.2 Mb in genomovars G1, G3, G4 and G7 (Fig. S1), while no
183 notable size difference in linear chromosomes was observed between members of the same
184 species (data not shown). Similarly, outside the *A. tumefaciens* complex, a ca. 2.0 Mb band
185 was generally observed except for *R. skierniewicense*, which yielded a 1.6 Mb chromid (Fig.
186 S2).
187
188 *3.2.Cooccurrence of protelomerase and linear chromosome*
189 Huang *et al.* (2012) recently demonstrated that replication of the *Agrobacterium* linear
190 chromid involved covalently closed hairpin loop telomeres and the protelomerase TelA. As
191 expected, we detected *telA* by PCR in all agrobacteria found to harbor a linear replicon and,
192 interestingly, never in other strains (Table S1). This confirmed the physiological relationship
193 between TelA and the presence of linear replicons. Nevertheless, the question remains as to
194 whether *telA* acquisition is a single founding event for a clade characterized by the presence
195 of a linear chromid in agrobacteria.
196
197 *3.3. telA and Rhizobiacae evolutionary history*
198 To distinguish between single and multiple event(s) for linear chromid acquisition, we
199 screened for congruence between the Rhizobiaceae phylogeny and the *telA* phylogeny. As
200 previously shown (Costechareyre et al., 2010; Shams et al., 2013), we used *recA* as a
201 convenient proxy for the Rhizobiaceae phylogeny. According to this phylogenetic marker, the
202 linear chromid-bearing and *telA*-bearing agrobacteria (i.e. the whole *A. tumefaciens* species
203 complex, *A. viscosum*, *A. larrymoorei*, *A. rubi*, *A.* sp NCPP 1650 and *R. skierniewicense*)
204 form a well-supported monophyletic group (Fig. 2). Moreover, within this clade, *telA* and
205 *recA* phylogenies were found to be congruent for almost all major supported splits, indicating
206 that *telA* has a history following that of the core genome. The *telA* gene thus appeared to have
207 been acquired horizontally once in a single ancestor and then have been vertically inherited
208 and stably maintained in the progeny. Remarkably, *A. vitis* and *Al. undicola* that were

209 included in *Agrobacterium* in the transitory proposal of Lindström and Young (Lindström and
210 Young, 2011) have a circular chromid and no *telA*. These two species are both outgroup of the
211 linear-chromid-bearing lineage.
212 To look for a candidate donor of the protelomerase gene to the lineage ancestor, TelA
213 homologs were sought in the nr protein database with BLASTP. An alignment was done
214 using the core catalytic domain of these distant proteins (Huang et al., 2012). The resulting
215 TelA tree presented four main clusters (Fig. S3). A first cluster grouped TelA proteins
216 retrieved from the different agrobacterial species. The second cluster contained different TelA
217 homologs retrieved in cyanobacterial genomes, i.e. *Synechococcus* sp. PCC 7335, two
218 different *Acaryochloris* species, *Crocosphaerea watsonii*, *Ktedonobacter racemifer* DSM
219 44963 and three different *Cyanothece* species. The third group clustered the ResT telomere
220 resolvase harbored on genomes of the *Borrelia* genus, which has been fully studied for its
221 linear replicons (Chaconas and Kobryn, 2010). The fourth group was clustering protelomerase
222 of gammaproteobacterial phage genomes such as N15, PY54 and phiKO2, which are known
223 to have linear genomes (Huang et al., 2012). Besides these four main clusters, two smaller
224 clusters showed putative protelomerase derived from genomes of eukaryotic virus members of
225 the Phycodnaviridae family, i.e. *Emiliania huxleyi* virus, *Feldmannia irregularis* virus and
226 *Ectocarpus siliculosus* virus, whose genome geometry is still unknown. Agrobacterial TelA
227 core domain sequences appear distantly related to the *Emiliania huxleyi* virus ones, but as
228 there is no support for this clustering and branches are very long, the relatedness to these viral
229 genes is not certain.
230 TelA is sparsely distributed amongst organisms and its phylogeny is not congruent with the
231 tree of life, indicating that protelomerase genes were acquired by lateral gene transfer at least
232 six independent times. The presence of characterized protelomerase proteins in distant parts of
233 the tree of TelA homologs suggests that the protelomerase activity is an ancient feature of this
234 protein family. *Borrelia* telomere resolvase ResT was proposed to have evolved from tyrosine
235 recombinases (Chaconas and Kobryn, 2010), raising the question of protelomerase function
236 arising from divergence of proteins pre-existing in organism genomes. However, the
237 similarity of tyrosine recombinases to protelomerases is almost undetectable and indeed lower
238 than the similarity existing amongst protelomerases themselves. This indicates that the
239 divergence between protelomerases and tyrosine recombinases is much older than the *telA*
240 acquisition events by the six different lineages.
241
242 *3.4. Evolutionary advantage of linear chromids*

As the linear geometry still persists in all of the tested progenies, it likely confers them selective advantages and is probably involved in the speciation process. However, although circular DNA seems easier to replicate, maintenance of linear dsDNA (which is ubiquitous in eukaryotes) questions the advantage it confers to cells. Indeed, neither circularized chromosomes of *Streptomyces lividans* (Volff et al., 1997) nor linearized chromosomes of *Escherichia coli* (Cui et al., 2007) have been shown to have a measurable effect on cell fitness. It is likely that the main advantage of a linear geometry is an enhanced recombination efficiency (Volff and Altenbuchner, 2000). It was shown that ends of linear chromosomes are recombinogenic in many biological systems, and that the plasticity of the sub-telomeric region may confer evolutionary advantages (Chaconas and Kobryn, 2010; Huang et al., 2012). By investigating the diversity of housekeeping genes, Marri *et al.* (2008) concluded that the evolution of the *Agrobacterium* linear chromid was not based on a facility to recombine, since recombination occurs at the same rate as that observed between genes mapping the distal portion of the circular chromosome. However, exhaustive genome comparisons showed that the linear chromosome is the site of large genetic exchanges that may involved up to 1 Mb within *A. fabrum* (Lassalle et al., 2011). In addition, we showed that species-specific genes whose acquisition led to the ecological speciation of *A. fabrum* were mostly located in its linear chromid (Lassalle et al., 2011). The adaptive gain conferred by a linear chromid would thus be the enhanced efficacy of sexuality, which facilitates the spread and maintenance of adaptive genomic innovations, especially those encoded by very large loci.

3.5. Presence of telA and linear chromids as key traits for redefining the Agrobacterium genus

Collectively, our findings strongly suggest that acquisition of *telA* is the founding event leading to the speciation of a major subclade within the *Agrobacterium*/*Rhizobium* supercluster. The classification of *Agrobacterium* as a valid genus has been controversial for decades leading to taxonomical confusion (Costechareyre et al., 2010). We showed here that the presence of linear chromids is a synapomorphic trait of a well-defined clade. Considering the key importance of this trait in the speciation process, we suggest that the *Agrobacterium* genus should now correspond to this clade. Consequently, the emended *Agrobacterium* genus should not include *A. vitis* and related species *Al. undicola* because they harbor a circular chromid. Similarly, *A. albertimagni*, which has no *telA* gene, is not a *bona fide Agrobacterium*. Conversely, all members of the emended *Agrobacterium* genus should be

given the genus name *Agrobacterium*. To date this concerns 16 species/genomospecies that either already received this genus name (i.e. *A. radiobacter*, *A. fabrum, A. viscosum, A. rubi, A. larrymoorei, A.* sp NCPP 1650) or taxa that were originally classified as *Rhizobium* due to taxonomic uncertainties (i.e. *R. skierniewicense, R. nepotum* and *R. pusense* that should now be named *A. skierniewicense, A. nepotum* and *A. pusense*, respectively), as well as genomospecies that should be transitorily named *A. sp.* genomovar G1, G3, G5, G6, G7, G9 or G13 until they receive a valid Latin epithet.

4. Conclusion

The acquisition of a secondary chromosome (i.e. chromid) is a feature of the Rhizobiaceae family. Within this family, the single acquisition event of a *telA* gene by an ancestor allowed chromid linearization and maintenance of the linear geometry in the ancestor progeny. The high species diversity of the emended *Agrobacterium* genus (defined on the basis of this feature) highlights the evolutionary success associated with this uncommon genome geometry and questions the actual nature of the advantages it confers.

Acknowledgements

M.H.R.B received a postdoctoral fellowship from the *Département de Santé des Plantes* (DSPE) of *Institut National de la Recherche Agronomique* (INRA). The authors would like to thank the *Collection Française de Bactéries associées aux Plantes* (CFBP) (http://www.angers.inra.fr/cfbp/) of the *Centre International de Ressources Microbiennes* (CIRM) network for access to the strain repository, and the translator Dr. D. Manley for reading the manuscript. This work was supported by the EcoGenome project of the *Agence Nationale de la Recherche* (ANR) (grant number 08-BLAN-0090).

References

Aihara, H., Huang, W.M., Ellenberger, T., 2007. An interlocked dimer of the protelomerase TelK distorts DNA structure for the formation of hairpin telomeres. Mol. Cell 27, 901-913.

Allardet-Servent, A., Michaux-Charachon, S., Jumas-Bilak, E., Karayan, L., Ramuz, M., 1993. Presence of one linear and one circular chromosome in the *Agrobacterium tumefaciens* C58 genome. J. Bacteriol. 175, 7869-7874.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403-410.

Bandyopadhyay, A., Elvitigala, T., Welsh, E., Stöckel, J., Liberton, M., Min, H., Sherman, L.A., Pakrasi, H.B., 2011. Novel Metabolic Attributes of the Genus *Cyanothece*, Comprising a Group of Unicellular Nitrogen-Fixing Cyanobacteria. mBio 2, e00214-00211.

312 Casjens, S., 1999. Evolution of the linear DNA replicons of the *Borrelia* spirochetes. Curr.
313 Opin. Microbiol. 2, 529-534.
314 Chaconas, G., Kobryn, K., 2010. Structure, function, and evolution of linear replicons in
315 *Borrelia*. Annu. Rev. Microbiol. 64, 185-202.
316 Chaconas, G., Stewart, P.E., Tilly, K., Bono, J.L., Rosa, P., 2001. Telomere resolution in the
317 Lyme disease spirochete. EMBO J 20, 3229-3237.
318 Costechareyre, D., Rhouma, A., Lavire, C., Portier, P., Chapulliot, D., Bertolla, F., Boubaker,
319 A., Dessaux, Y., Nesme, X., 2010. Rapid and efficient identification of *Agrobacterium*
320 species by *recA* allele analysis. Microb. Ecol. 60, 862-872.
321 Cui, T., Moro-oka, N., Ohsumi, K., Kodama, K., Ohshima, T., Ogasawara, N., Mori, H.,
322 Wanner, B., Niki, H., Horiuchi, T., 2007. *Escherichia coli* with a linear genome. EMBO Rep
323 8, 181-187.
324 Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high
325 throughput. Nucleic Acids Res. 32, 1792-1797.
326 Farrand, S.K., van Berkum, P.B., Oger, P., 2003. *Agrobacterium* is a definable genus of the
327 family *Rhizobiaceae*. Int J Syst Evol Microbiol 53, 1681-1687.
328 Ferdows, M.S., Barbour, A.G., 1989. Megabase-sized linear DNA in the bacterium *Borrelia*
329 *burgdorferi*, the Lyme disease agent. Proc. Natl Acac Sci. USA 86, 5969-5973.
330 Galibert, F., Finan, T.M., Long, S.R., Pühler, A., Abola, P., Ampe, F., Barloy-Hubler, F.,
331 Barnett, M.J., Becker, A., Boistard, P., Bothe, G., Boutry, M., Bowser, L., Buhrmester, J.,
332 Cadieu, E., Capela, D., Chain, P., Cowie, A., Davis, R.W., Dréano, S., Federspiel, N.A.,
333 Fisher, R.F., Gloux, S., Godrie, T., Goffeau, A., Golding, B., Gouzy, J., Gurjal, M.,
334 Hernandez-Lucas, I., Hong, A., Huizar, L., Hyman, R.W., Jones, T., Kahn, D., Kahn, M.L.,
335 Kalman, S., Keating, D.H., Kiss, E., Komp, C., Lelaure, V., Masuy, D., Palm, C., Peck, M.C.,
336 Pohl, T.M., Portetelle, D., Purnelle, B., Ramsperger, U., Surzycki, R., Thébault, P.,
337 Vandenbol, M., Vorhölter, F.-J., Weidner, S., Wells, D.H., Wong, K., Yeh, K.-C., Batut, J.,
338 2001. The composite genome of the legume symbiont *Sinorhizobium meliloti*. Science 293,
339 668-672.
340 Goodner, B., Hinkle, G., Gattung, S., Miller, N., Blanchard, M., Qurollo, B., Goldman, B.S.,
341 Cao, Y., Askenazi, M., Halling, C., Mullin, L., Houmiel, K., Gordon, J., Vaudin, M.,
342 Iartchouk, O., Epp, A., Liu, F., Wollam, C., Allinger, M., Doughty, D., Scott, C., Lappas, C.,
343 Markelz, B., Flanagan, C., Crowell, C., Gurson, J., Lomo, C., Sear, C., Strub, G., Cielo, C.,
344 Slater, S., 2001. Genome sequence of the plant pathogen and biotechnology agent
345 *Agrobacterium tumefaciens* C58. Science 294, 2323-2328.
346 Gouy, M., Guindon, S., Gascuel, O., 2010. SeaView version 4: a multiplatform graphical user
347 interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27, 221-224.
348 Grandin, N., Charbonneau, M., 2008. Protection against chromosome degradation at the
349 telomeres. Biochimie 90, 41-59.
350 Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New
351 algorithms and methods to estimate maximum-likelihood phylogenies: assessing the
352 performance of PhyML 3.0. Syst Biol 59, 307-321.
353 Harrison, P.W., Lower, R.P.J., Kim, N.K.D., Young, J.P.W., 2010. Introducing the bacterial
354 `chromid': not a chromosome, not a plasmid. Trends in Microbiology 18, 141-148.
355 Huang, W.M., DaGloria, J., Fox, H., Ruan, Q., Tillou, J., Shi, K., Aihara, H., Aron, J.,
356 Casjens, S., 2012. Linear chromosome generating system of *Agrobacterium tumefaciens* C58:
357 Protelomerase generates and protects hairpin ends. J. Biol. Chem.
358 Jumas-Bilak, E., Michaux-Charachon, S., Bourg, G., Ramuz, M., Allardet-Servent, A., 1998.
359 Unconventional genomic organization in the Alpha subgroup of the Proteobacteria. J.
360 Bacteriol. 180, 2749-2755.

Kersters, K., De Ley, J., 1984. Genus III. *Agrobacterium* Conn 1942, 359AL. In: Krieg, N.R., Holt, J.G. (Eds.), Bergey's Manual of Systematic Bacteriology. Williams & Wilkins, Baltimore, pp. 244-254.
Lassalle, F., Campillo, T., Vial, L., Baude, J., Costechareyre, D., Chapulliot, D., Shams, M., Abrouk, D., Lavire, C., Oger-Desfeux, C., Hommais, F., Gueguen, L., Daubin, V., Muller, D., Nesme, X., 2011. Genomic species are ecological species as revealed by comparative genomics in *Agrobacterium tumefaciens*. Genome Biol. Evol. 3, 762-781.
Lin, Y.S., Kieser, H.M., Hopwood, D.A., Chen, C.W., 1993. The chromosomal DNA of *Streptomyces lividans* 66 is linear. Mol Microbiol. 10, 923-933.
Lindström, K., Young, J.P.W., 2011. International Committee on Systematics of Prokaryotes; Subcommittee on the taxonomy of *Agrobacterium* and *Rhizobium:* Minutes of the meeting, 7 September 2010, Geneva, Switzerland International Journal of Systematic and Evolutionary Microbiology 61, 3089-3093.
Marri, P.R., Harris, L.K., Houmiel, K., Slater, S.C., Ochman, H., 2008. The effect of chromosome geometry on genetic diversity. Genetics 179, 511-516.
Ramírez-Bahena, M.H., Nesme, X., Muller, D., 2012. Rapid and simultaneous detection of linear chromosome and large plasmids in Proteobacteria. Journal of Basic Microbiology 52, 736-739.
Ravin, N.V., 2003. Mechanisms of replication and telomere resolution of the linear plasmid prophage N15. FEMS Microbiol. Ecol. 221, 1-6.
Sawada, H., Ieki, H., Oyaizu, H., Matsumoto, S., 1993. Proposal for rejection of *Agrobacterium tumefaciens* and revised descriptions for the genus *Agrobacterium* and for *Agrobacterium radiobacter* and *Agrobacterium rhizogenes*. Int J Syst Bacteriol 43, 694-702.
Shams, M., Vial, L., Chapulliot, D., Nesme, X., Lavire, C., 2013. Rapid and accurate species and genomic species identification and exhaustive population diversity assessment of *Agrobacterium* spp. using *recA*-based PCR. Syst. Appl. Microbiol. 36, 351-358.
Shi, K., Huang, W.M., Aihara, H., 2013. An enzyme-catalyzed multistep DNA refolding mechanism in hairpin telomere formation. PLoS Biol 11, e1001472.
Slater, S., Setubal, J.C., Goodner, B., Houmiel, K., Sun, J., Kaul, R., Goldman, B.S., Farrand, S.K., Almeida, N., Burr, T., Nester, E., Rhoads, D.M., Kadoi, R., Ostheimer, T., Pride, N., Sabo, A., Henry, E., Telepak, E., Cromes, L., Harkleroad, A., Oliphant, L., Pratt-Szegila, P., Welch, R., Wood, D., 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, S.C., Goldman, B.S., Goodner, B., Setubal, J.C., Farrand, S.K., Nester, E.W., Burr, T.J., Banta, L., Dickerman, A.W., Paulsen, I., Otten, L., Suen, G., Welch, R., Almeida, N.F., Arnold, F., Burton, O.T., Du, Z., Ewing, A., Godsy, E., Heisel, S., Houmiel, K.L., Jhaveri, J., Lu, J., Miller, N.M., Norton, S., Chen, Q., Phoolcharoen, W., Ohlin, V., Ondrusek, D., Pride, N., Stricklin, S.L., Sun, J., Wheeler, C., Wilson, L., Zhu, H., Wood, D.W., 2009. Genome sequences of three *Agrobacterium* biovars help elucidate the evolution of multichromosome genomes in bacteria. J. Bacteriol. 191, 2501-2511.
Volff, J.-N., Altenbuchner, J., 2000. A new beginning with new ends: linearisation of circular chromosomes during bacterial evolution. FEMS Microbiol. Letters 186, 143-150.
Volff, J.N., Viell, P., Altenbuchner, J., 1997. Artificial circularization of the chromosome with concomitant deletion of its terminal inverted repeats enhances genetic instability and genome rearrangement in *Streptomyces lividans*. Mol. Gen. Genet. 253, 753-760.
Wood, D.W., Setubal, J.C., Kaul, R., Monks, D.E., Kitajima, J.P., Okura, V.K., Zhou, Y., Chen, L., Wood, G.E., Almeida, N.F., Jr., Woo, L., Chen, Y., Paulsen, I.T., Eisen, J.A., Karp, P.D., Bovee, D., Sr., Chapman, P., Clendenning, J., Deatherage, G., Gillet, W., Grant, C., Kutyavin, T., Levy, R., Li, M.-J., McClelland, E., Palmieri, A., Raymond, C., Rouse, G.,

411 Saenphimmachak, C., Wu, Z., Romero, P., Gordon, D., Zhang, S., Yoo, H., Tao, Y., Biddle,
412 P., Jung, M., Krespan, W., Perry, M., Gordon-Kamm, B., Liao, L., Kim, S., Hendrick, C.,
413 Zhao, Z.-Y., Dolan, M., Chumley, F., Tingey, S.V., Tomb, J.-F., Gordon, M.P., Olson, M.V.,
414 Nester, E.W., 2001. The genome of the natural genetic engineer *Agrobacterium tumefaciens*
415 C58. Science 294, 2317-2323.
416 Young, J.M., Kuykendall, L.D., Martínez-Romero, E., Kerr, A., Sawada, H., 2001. A revision
417 of *Rhizobium* Frank 1889, with an emended description of the genus, and the inclusion of all
418 species of *Agrobacterium* Conn 1942 and *Allorhizobium undicola* de Lajudie *et al*. 1998 as
419 new combinations: *Rhizobium radiobacter*, *R. rhizogenes*, *R. rubi*, *R. undicola* and *R. vitis*. Int
420 J Syst Evol Microbiol 51, 89-103.
421 Young, J.M., Kuykendall, L.D., Martínez-Romero, E., Kerr, A., Sawada, H., 2003.
422 Classification and nomenclature of *Agrobacterium* and *Rhizobium* - a reply to Farrand *et al*.
423 (2003). Int J Syst Evol Microbiol 53, 1689-1695.
424

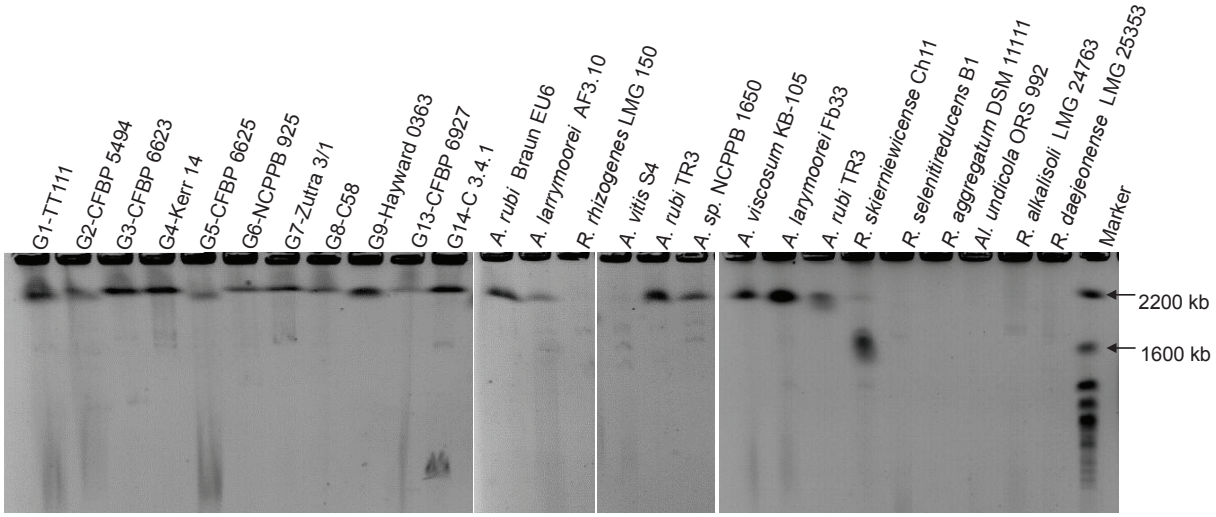


Fig. 1. Presence of linear chromids in *Agrobacterium* members and related taxa revealed by pulse field electrophoresis. *A. sp.* genomovars (indicated as G1, G2...) are indicated for members of the *A. tumefaciens* species complex. Marker, *Saccharomyces cerevisiae*. The pictures are representative of different experiments. All strains were tested at least twice with similar results.

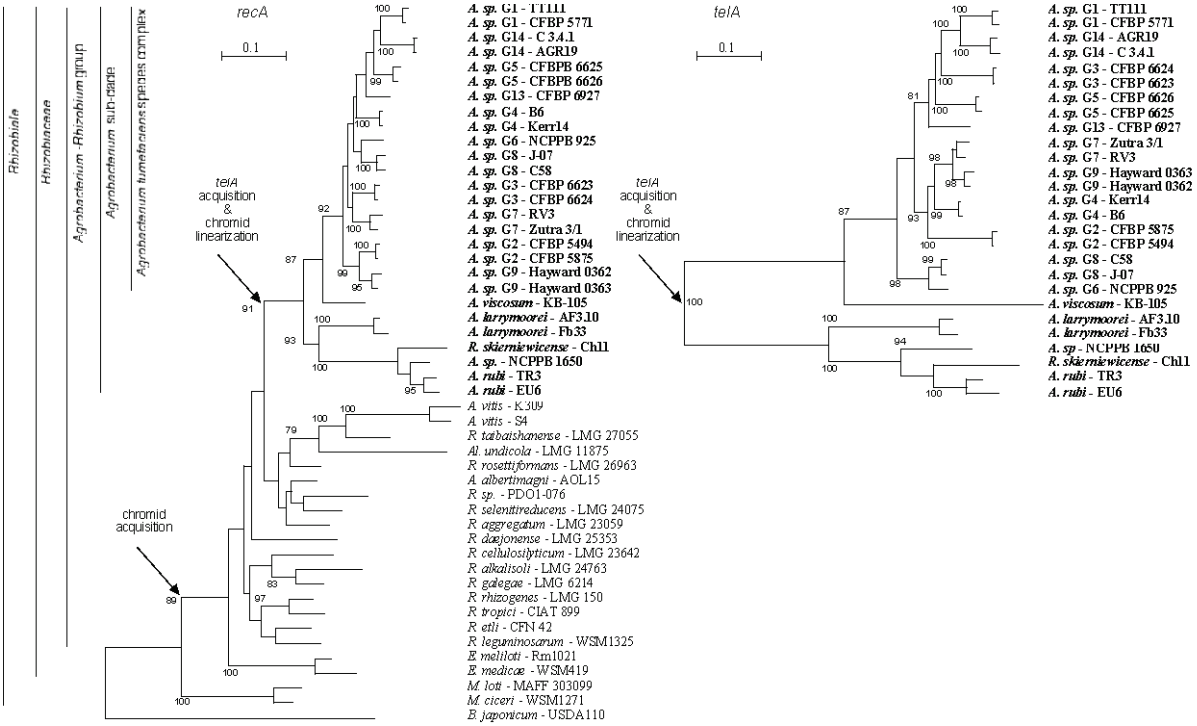


Fig. 2. Comparative phylogenies of *recA* and *telA* showing the emergence of a single homogeneous sub-clade within the *Agrobacterium*/*Rhizobium* supercluster characterized by concomitant presences of both *telA* and linear chromid structure. The presence of a large chromid is characteristic and coincides with speciation of the Rhizobiaceae family within Rhizobiales. The acquisition of *telA* is characteristic and coincides with the chromid linearization and the speciation of a sub-clade in the *Agrobacteriuum/Rhizobium* supercluster. Trees were constructed via maximum likelihood with values with 1098 and 1344 sites for *recA* and *telA*, respectively. Distances are in nucleotide substitution per site. Significant bootstrap values (100 bootstrap resamplings) over 80% are given.

List of Figures

List of Tables

Bibliography

Abby SS, Tannier E, Gouy M, and Daubin V. 2010. *Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests*. 11:324–324. PMID: 20550700 PMCID: 2905365.

Abby SS, Tannier E, Gouy M, and Daubin V. 2012. Lateral gene transfer as a support for the tree of life. *Proceedings of the National Academy of Sciences* 109(13):4962–4967.

Achtman M and Wagner M. 2008. Microbial diversity and the genetic nature of microbial species. *Nat Rev Micro* 6(6):431–440.

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. 1990. Basic local alignment search tool. *Journal of molecular biology* 215(3):403–410. PMID: 2231712.

Andres J, Arsène-Ploetze F, Barbe V, Brochier-Armanet C, Cleiss-Arnold J, Coppée JY, Dillies MA, Geist L, Joublin A, Koechler S, Lassalle F, Marchal M, Médigue C, Muller D, Nesme X, Plewniak F, Proux C, Ramirez-Bahena MH, Schenowitz C, Sismeiro O, Vallenet D, Santini JM, and Bertin PN. 2013. Life in an arsenic-containing gold mine: genome and physiology of the autotrophic arsenite-oxidizing bacterium rhizobium sp. NT-26. *Genome Biology and Evolution*. PMID: 23589360.

Angly FE, Willner D, Prieto-Davó A, Edwards RA, Schmieder R, Vega-Thurber R, Antonopoulos DA, Barott K, Cottrell MT, Desnues C, Dinsdale EA, Furlan M, Haynes M, Henn MR, Hu Y, Kirchman DL, McDole T, McPherson JD, Meyer F, Miller RM, Mundt E, Naviaux RK, Rodriguez-Mueller B, Stevens R, Wegley L, Zhang L, Zhu B, and Rohwer F. 2009. The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput Biol* 5(12):e1000593.

Ané C, Larget B, Baum DA, Smith SD, and Rokas A. 2007. Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution* 24(2):412–426. PMID: 17095535.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, and Sherlock G. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics* 25(1):25–29.

Aujoulat F, Jumas-Bilak E, Masnou A, Sallé F, Faure D, Segonds C, Marchandin H, and Teyssier C. 2011. Multilocus sequence-based analysis delineates a clonal population of agrobacterium

(rhizobium) radiobacter (agrobacterium tumefaciens) of human origin. *Journal of bacteriology* 193(10):2608–2618. PMID: 21398532.

Bansal MS, Alm EJ, and Kellis M. 2012. Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics (Oxford, England)* 28(12):i283–291. PMID: 22689773.

Barraud N, Schleheck D, Klebensberger J, Webb JS, Hassett DJ, Rice SA, and Kjelleberg S. 2009. Nitric oxide signaling in pseudomonas aeruginosa biofilms mediates phosphodiesterase activity, decreased cyclic di-GMP levels, and enhanced dispersal. *Journal of Bacteriology* 191(23):7333–7342. PMID: 19801410.

Bigot T, Daubin V, Lassalle F, and Perrière G. 2013. TPMS: a set of utilities for querying collections of gene trees. *BMC bioinformatics* 14:109. PMID: 23530580.

Boussau B, Karlberg EO, Frank AC, Legault BA, and Andersson SGE. 2004. Computational inference of scenarios for alpha-proteobacterial genome evolution. *Proceedings of the National Academy of Sciences of the United States of America* 101(26):9722–9727. PMID: 15210995.

Boussau B, Szöllősi GJ, Duret L, Gouy M, Tannier E, and Daubin V. 2013. Genome-scale coestimation of species and gene trees. *Genome Research* 23(2):323–330.

Bouzar H and Jones J. 2001. Agrobacterium larrymoorei sp. nov., a pathogen isolated from aerial tumours of ficus benjamina. *Int J Syst Evol Microbiol* 51(3):1023–1026.

Bouzar H, Ouadah D, Krimi Z, Jones JB, Trovato M, Petit A, and Dessaux Y. 1993. Correlative association between resident plasmids and the host chromosome in a diverse agrobacterium soil population. *Applied and environmental microbiology* 59(5):1310–1317. PMID: 16348927.

Brown JR and Doolittle WF. 1997. Archaea and the prokaryote-to-eukaryote transition. *Microbiology and molecular biology reviews: MMBR* 61(4):456–502. PMID: 9409149.

Bruen TC, Philippe H, and Bryant D. 2006. A simple and robust statistical test for detecting the presence of recombination. *Genetics* 172(4):2665–2681. PMID: 16489234.

Bérard S, Gallien C, Boussau B, Szöllősi GJ, Daubin V, and Tannier E. 2012. Evolution of gene neighborhoods within reconciled phylogenies. *Bioinformatics* 28(18):i382–i388.

Cadillo-Quiroz H, Didelot X, Held NL, Herrera A, Darling A, Reno ML, Krause DJ, and Whitaker RJ. 2012. Patterns of gene flow define species of thermophilic archaea. *PLoS Biol* 10(2):e1001265.

Caffrey BE, Williams TA, Jiang X, Toft C, Hokamp K, and Fares MA. 2012. Proteome-wide analysis of functional divergence in bacteria: exploring a host of ecological adaptations. *PloS one* 7(4):e35659. PMID: 22563391.

Cardoso PG, Macedo GC, Azevedo V, and Oliveira SC. 2006. Brucella spp noncanonical LPS: structure, biosynthesis, and interaction with host immune system. *Microbial cell factories* 5:13. PMID: 16556309.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution* 17(4):540–552. PMID: 10742046.

Chain PSG, Carniel E, Larimer FW, Lamerdin J, Stoutland PO, Regala WM, Georgescu AM, Vergez LM, Land ML, Motin VL, Brubaker RR, Fowler J, Hinnebusch J, Marceau M, Médigue C, Simonet M, Chenal-Francisque V, Souza B, Dacheux D, Elliott JM, Derbise A, Hauser LJ, and Garcia E. 2004. Insights into the evolution of yersinia pestis through whole-genome comparison with yersinia pseudotuberculosis. *Proceedings of the National Academy of Sciences of the United States of America* 101(38):13826–13831. PMID: 15358858.

Charlebois RL and Doolittle WF. 2004. Computing prokaryotic gene ubiquity: Rescuing the core from extinction. *Genome Research* 14(12):2469 –2477.

Choi SC, Rasmussen MD, Hubisz MJ, Gronau I, and Stanhope MJ. 2012. Replacing and additive horizontal gene transfer in streptococcus. *Molecular Biology and Evolution*.

Cohan FM. 2001. Bacterial species and speciation. *Systematic Biology* 50(4):513 –524.

Cohan FM. 2002a. Sexual isolation and speciation in bacteria. *Genetica* 116(2-3):359–370. PMID: 12555790.

Cohan FM. 2002b. What are bacterial species? *Annual Review of Microbiology* 56(1):457–487.

Cohan FM and Koeppel AF. 2008. The origins of ecological diversity in prokaryotes. *Current Biology: CB* 18(21):R1024–1034. PMID: 19000803.

Cohan FM and Perry EB. 2007. A systematics for discovering the fundamental units of bacterial diversity. *Current Biology: CB* 17(10):R373–386. PMID: 17502094.

Coleman ML and Chisholm SW. 2010. Ecosystem-specific selection pressures revealed through comparative population genomics. *Proceedings of the National Academy of Sciences* 107(43):18634 –18639.

Cooper VS, Vohr SH, Wrocklage SC, and Hatcher PJ. 2010. Why genes evolve faster on secondary chromosomes in bacteria. *PLoS Comput Biol* 6(4):e1000732.

Cordero OX, Ventouras LA, DeLong EF, and Polz MF. 2012. Public good dynamics drive evolution of iron acquisition strategies in natural bacterioplankton populations. *Proceedings of the National Academy of Sciences* 109(49):20059–20064. PMID: 23169633.

Cortez D, Forterre P, and Gribaldo S. 2009. A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and ORFans in archaeal and bacterial genomes. *Genome Biology* 10(6):R65.

Costechareyre D, Bertolla F, and Nesme X. 2009. Homologous recombination in agrobacterium: potential implications for the genomic species concept in bacteria. *Molecular Biology and Evolution* 26(1):167–176. PMID: 18936442.

Costechareyre D, Rhouma A, Lavire C, Portier P, Chapulliot D, Bertolla F, Boubaker A, Dessaux Y, and Nesme X. 2010. Rapid and efficient identification of agrobacterium species by recA allele analysis : Agrobacterium recA diversity. *Microbial Ecology* 60(4):862–72. PMID: 20521039.

Couturier E and Rocha EPC. 2006. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. *Molecular Microbiology* 59(5):1506–1518.

Csűrös M. 2008. Ancestral reconstruction by asymmetric wagner parsimony over continuous characters and squared parsimony over distributions. In Nelson CE and Vialette S, editors, *Comparative Genomics* number 5267 in Lecture Notes in Computer Science pages 72–86. Springer Berlin Heidelberg.

Csűrös M and Miklós I. 2009. Streamlining and large ancestral genomes in archaea inferred with a phylogenetic birth-and-death model. *Molecular Biology and Evolution* 26(9):2087–2095.

Dagan T, Artzy-Randrup Y, and Martin W. 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proceedings of the National Academy of Sciences* 105(29):10039–10044. PMID: 18632554.

Dagan T and Martin W. 2007. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proceedings of the National Academy of Sciences* 104(3):870 –875.

Daubin V, Lerat E, and Perrière G. 2003. The source of laterally transferred genes in bacterial genomes. *Genome Biology* 4(9):R57. PMID: 12952536.

Daubin V and Ochman H. 2004a. Bacterial genomes as new gene homes: The genealogy of ORFans in e. coli. *Genome Research* 14(6):1036 –1042.

Daubin V and Ochman H. 2004b. Start-up entities in the origin of new genes. *Current Opinion in Genetics & Development* 14(6):616–619.

Daubin V and Perrière G. 2003. G+C3 structuring along the genome: A common feature in prokaryotes. *Molecular Biology and Evolution* 20(4):471 –483.

David LA and Alm EJ. 2011. Rapid evolutionary innovation during an archaean genetic expansion. *Nature* 469(7328):93–96.

Davids W and Zhang Z. 2008. The impact of horizontal gene transfer in shaping operons and protein interaction networks - direct evidence of preferential attachment. *BMC Evolutionary Biology* 8(1):23.

de Queiroz K. 2005. Ernst mayr and the modern concept of species. *Proceedings of the National Academy of Sciences* 102(suppl_1):6600–6607.

Didelot X, Lawson D, Darling A, and Falush D. 2010. Inference of homologous recombination in bacteria using whole genome sequences. *Genetics* 186:1435–1449.

Dimmer EC, Huntley RP, Alam-Faruque Y, Sawford T, O'Donovan C, Martin MJ, Bely B, Browne P, Mun Chan W, Eberhardt R, Gardner M, Laiho K, Legge D, Magrane M, Pichler K, Poggioli D, Sehra H, Auchincloss A, Axelsen K, Blatter MC, Boutet E, Braconi-Quintaje S, Breuza L, Bridge A, Coudert E, Estreicher A, Famiglietti L, Ferro-Rojas S, Feuermann M, Gos A, Gruaz-Gumowski N, Hinz U, Hulo C, James J, Jimenez S, Jungo F, Keller G, Lemercier P, Lieberherr D, Masson P, Moinat M, Pedruzzi I, Poux S, Rivoire C, Roechert B, Schneider M, Stutz A, Sundaram S, Tognolli M, Bougueleret L, Argoud-Puy G, Cusin I, Duek-Roggli P, Xenarios I, and Apweiler R. 2011. The UniProt-GO annotation database in 2011. *Nucleic Acids Research* 40(D1):D565–D570.

Doolittle WF. 1999. Phylogenetic classification and the universal tree. *Science* 284(5423):2124–2128. PMID: 10381871.

Doolittle WF. 2013. Is junk DNA bunk? a critique of ENCODE. *Proceedings of the National Academy of Sciences* 110(14):5294–5300. PMID: 23479647.

Doolittle WF and Zhaxybayeva O. 2009. On the origin of prokaryotic species. *Genome research* 19(5):744–756. PMID: 19411599.

Doroghazi JR and Buckley DH. 2011. A model for the effect of homologous recombination on microbial diversification. *Genome Biology and Evolution* 3(0):1349–1356.

Doyon JP, Chauve C, and Hamel S. 2008. Algorithms for exploring the space of gene Tree/Species tree reconciliations. In Nelson CE and Vialette S, editors, *Comparative Genomics* number 5267 in Lecture Notes in Computer Science pages 1–13. Springer Berlin Heidelberg.

Doyon JP, Ranwez V, Daubin V, and Berry V. 2011. Models, algorithms and programs for phylogeny reconciliation. *Briefings in Bioinformatics* 12(5):392–400.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics* 5:113. PMID: 15318951.

Edwards RA and Rohwer F. 2005. Viral metagenomics. *Nat Rev Micro* 3(6):504–510.

Epstein B, Branca A, Mudge J, Bharti AK, Briskine R, Farmer AD, Sugawara M, Young ND, Sadowsky MJ, and Tiffin P. 2012. Population genomics of the facultatively mutualistic bacteria sinorhizobium meliloti and s. medicae. *PLoS Genet* 8(8):e1002868.

Esnault E, Valens M, Espéli O, and Boccard F. 2007. Chromosome structuring limits genome plasticity in escherichia coli. *PLoS Genet* 3(12):e226.

Farrand SK, Berkum PBv, and Oger P. 2003. Agrobacterium is a definable genus of the family rhizobiaceae. *International Journal of Systematic and Evolutionary Microbiology* 53(5):1681–1687. PMID: 13130068.

Felsenstein J. 1993. PHYLIP (phylogeny inference package) version 3.5c.

Francino MP. 2012. *Horizontal Gene Transfer in Microorganisms*. Horizon Scientific Press.

Fraser C, Alm EJ, Polz MF, Spratt BG, and Hanage WP. 2009. The bacterial species challenge: making sense of genetic and ecological diversity. *Science (New York, N.Y.)* 323(5915):741–746. PMID: 19197054.

Fraser C, Hanage WP, and Spratt BG. 2007. Recombination and the nature of bacterial speciation. *Science* 315(5811):476–480. PMID: 17255503.

Friedman J, Alm EJ, and Shapiro BJ. 2013. Sympatric speciation: When is it possible in bacteria? *PLoS ONE* 8(1):e53539.

Gerdes SY, Scholle MD, Campbell JW, Balazsi G, Ravasz E, Daugherty MD, Somera AL, Kyrpides NC, Anderson I, Gelfand MS, Bhattacharya A, Kapatral V, D'Souza M, Baev MV, Grechkin Y, Mseeh F, Fonstein MY, Overbeek R, Barabasi AL, Oltvai ZN, and Osterman AL. 2003. Experimental determination and system level analysis of essential genes in escherichia coli MG1655. *J. Bacteriol.* 185(19):5673–5684.

Gonzalez V, Acosta JL, Santamaria RI, Bustos P, Fernandez JL, Hernandez Gonzalez IL, Diaz R, Flores M, Palacios R, Mora J, and Davila G. 2010. Conserved symbiotic plasmid DNA sequences in the multireplicon pangenomic structure of rhizobium etli. *Appl. Environ. Microbiol.* 76(5):1604–1614.

González V, Bustos P, Ramírez-Romero MA, Medrano-Soto A, Salgado H, Hernández-González I, Hernández-Celis JC, Quintero V, Moreno-Hagelsieb G, Girard L, Rodríguez O, Flores M, Cevallos MA, Collado-Vides J, Romero D, and Dávila G. 2003. The mosaic structure of the symbiotic plasmid of rhizobium etli CFN42 and its relation to other symbiotic genome compartments. *Genome Biology* 4(6):R36–R36. PMID: 12801410 PMCID: 193615.

Goodner B, Hinkle G, Gattung S, Miller N, Blanchard M, Qurollo B, Goldman BS, Cao Y, Askenazi M, Halling C, Mullin L, Houmiel K, Gordon J, Vaudin M, Iartchouk O, Epp A, Liu F, Wollam C, Allinger M, Doughty D, Scott C, Lappas C, Markelz B, Flanagan C, Crowell C, Gurson J, Lomo C, Sear C, Strub G, Cielo C, and Slater S. 2001. Genome sequence of the plant pathogen and biotechnology agent agrobacterium tumefaciens c58. *Science (New York, N.Y.)* 294(5550):2323–2328. PMID: 11743194.

Gouy M, Gautier C, Attimonelli M, Lanave C, and di Paola G. 1985. ACNUC–a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. *Computer Applications in the Biosciences: CABIOS* 1(3):167–172. PMID: 3880341.

Guindon S and Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* 52(5):696–704.

Hao W and Golding GB. 2006. The fate of laterally transferred genes: Life in the fast lane to adaptation or death. *Genome Research* 16(5):636 –643.

Hao X, Lin Y, Johnstone L, Liu G, Wang G, Wei G, McDermott T, and Rensing C. 2012a. Genome sequence of the arsenite-oxidizing strain agrobacterium tumefaciens 5A. *Journal of Bacteriology* 194(4):903.

Hao X, Xie P, Johnstone L, Miller SJ, Rensing C, and Wei G. 2012b. Genome sequence and mutational analysis of plant-growth-promoting bacterium agrobacterium tumefaciens CC-NWGS0286 isolated from a zinc-lead mine tailing. *Applied and Environmental Microbiology* 78(15):5384–5394.

Hardin G. 1960. The competitive exclusion principle. *Science (New York, N.Y.)* 131:1292–1297. PMID: 14399717.

Harrison PW, Lower RP, Kim NK, and Young JPW. 2010. Introducing the bacterial 'chromid': not a chromosome, not a plasmid. *Trends in Microbiology* 18(4):141–148.

Hausdorf B. 2011. Progress toward a general species concept. *Evolution; international journal of organic evolution* 65(4):923–931. PMID: 21463293.

Hershberg R and Petrov DA. 2009. General rules for optimal codon choice. *PLoS Genet* 5(7):e1000556.

Hershberg R and Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet* 6(9):e1001115.

Hickman JW and Harwood CS. 2008. Identification of FleQ from pseudomonas aeruginosa as a c-di-GMP-responsive transcription factor. *Molecular Microbiology* 69(2):376–389.

Hildebrand F, Meyer A, and Eyre-Walker A. 2010. *Evidence of selection upon genomic GC-Content in bacteria*. 6(9). PMID: 20838593 PMCID: 2936529.

Hoffmann D, Hevel JM, Moore RE, and Moore BS. 2003. Sequence analysis and biochemical characterization of the nostopeptolide a biosynthetic gene cluster from nostoc sp. GSV224. *Gene* 311:171–180.

Holden MTG, Hsu LY, Kurt K, Weinert LA, Mather AE, Harris SR, Strommenger B, Layer F, Witte W, de Lencastre H, Skov R, Westh H, Zemlicková H, Coombs G, Kearns AM, Hill RLR, Edgeworth J, Gould I, Gant V, Cooke J, Edwards GF, McAdam PR, Templeton KE, McCann A, Zhou Z, Castillo-Ramírez S, Feil EJ, Hudson LO, Enright MC, Balloux F, Aanensen DM, Spratt BG, Fitzgerald JR, Parkhill J, Achtman M, Bentley SD, and Nübel U. 2013. A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant staphylococcus aureus pandemic. *Genome research* 23(4):653–664. PMID: 23299977.

Homma K, Fukuchi S, Nakamura Y, Gojobori T, and 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. *Molecular Biology and Evolution* 24(3):805 –813.

Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, and Polz MF. 2008. Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* 320(5879):1081–1085. PMID: 18497299.

Jones BR, Rajaraman A, Tannier E, and Chauve C. 2012. ANGES: reconstructing ANcestral GEnomeS maps. *Bioinformatics (Oxford, England)* 28(18):2388–2390. PMID: 22820205.

Juhas M, van der Meer JR, Gaillard M, Harding RM, Hood DW, and Crook DW. 2009. Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS microbiology reviews* 33(2):376–393. PMID: 19178566.

Keane P, Kerr A, and New P. 1970. Crown gall of stone fruit II. identification and nomenclature of agrobacterium isolates. *Australian Journal of Biological Sciences* 23(3):585–596.

Kersters K, Ley JD, Sneath PHA, and Sackin M. 1973. Numerical taxonomic analysis of agrobacterium. *Journal of General Microbiology* 78(2):227–239.

Kolter R and Greenberg EP. 2006. Microbial sciences: The superficial life of microbes. *Nature* 441(7091):300–302.

Konstantinidis KT, Ramette A, and Tiedje JM. 2006. The bacterial species definition in the genomic era. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 361(1475):1929–1940. PMID: 17062412.

Konstantinidis KT and Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America* 102(7):2567–2572. PMID: 15701695.

Koonin EV and Wolf YI. 2008. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucl. Acids Res.* 36(21):6688–6719.

Krasteva PV, Fong JCN, Shikuma NJ, Beyhan S, Navarro MVAS, Yildiz FH, and Sondermann H. 2010. Vibrio cholerae VpsT regulates matrix production and motility by directly sensing cyclic di-GMP. *Science* 327(5967):866–868.

Krimi Z, Petit A, Mougel C, Dessaux Y, and Nesme X. 2002. Seasonal fluctuations and long-term persistence of pathogenic populations of agrobacterium spp. in soils. *Applied and Environmental Microbiology* 68(7):3358–3365. PMID: 12089015.

Kristensen DM, Wolf YI, Mushegian AR, and Koonin EV. 2011. Computational methods for gene orthology inference. *Briefings in Bioinformatics* 12(5):379–391.

Kubler-Kielb J and Vinogradov E. 2013. The study of the core part and non-repeating elements of the o-antigen of brucella lipopolysaccharide. *Carbohydrate Research* 366:33–37.

Kunin V, Goldovsky L, Darzentas N, and Ouzounis CA. 2005. The net of life: Reconstructing the microbial phylogenetic network. *Genome Research* 15(7):954 –959.

Kuo CH and Ochman H. 2009. *Deletional bias across the three domains of life*. 2009:145–152. PMID: 20333185 PMCID: 2817411.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, and Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biology* 5(2):R12. PMID: 14759262.

Lan R and Reeves PR. 2000. Intraspecies variation in bacterial genomes: the need for a species genome concept. *Trends in Microbiology* 8(9):396–401. PMID: 10989306.

Lan R and Reeves PR. 2001. When does a clone deserve a name? a perspective on bacterial species based on population genetics. *Trends in Microbiology* 9(9):419–424. PMID: 11553453.

Lapierre P and Gogarten JP. 2009. Estimating the size of the bacterial pan-genome. *Trends in Genetics* 25(3):107–110.

Lassalle F, Campillo T, Vial L, Baude J, Costechareyre D, Chapulliot D, Shams M, Abrouk D, Lavire C, Oger-Desfeux C, Hommais F, Guéguen L, Daubin V, Muller D, and Nesme X. 2011. Genomic species are ecological species as revealed by comparative genomics in agrobacterium tumefaciens. *Genome Biology and Evolution* 3:762 –781.

Lathe WC, Snel B, and Bork P. 2000. Gene context conservation of a higher order than operons. *Trends in Biochemical Sciences* 25(10):474–479.

Lawrence J. 1999. Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. *Current opinion in genetics & development* 9(6):642–648. PMID: 10607610.

Lawrence JG and Ochman H. 1997. Amelioration of bacterial genomes: Rates of change and exchange. *Journal of Molecular Evolution* 44(4):383–397.

Lawrence JG and Retchless AC. 2009. The interplay of homologous recombination and horizontal gene transfer in bacterial speciation. *Methods in Molecular Biology (Clifton, N.J.)* 532:29–53. PMID: 19271178.

Lawrence JG and Roth JR. 1996. Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* 143(4):1843–1860. PMID: 8844169.

Lefébure T, Pavinski Bitar PD, Suzuki H, and Stanhope MJ. 2010. Evolutionary dynamics of complete campylobacter pan-genomes and the bacterial species concept. *Genome Biology and Evolution* 2:646 –655.

Lerat E, Daubin V, Ochman H, and Moran NA. 2005. Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol* 3(5):e130.

Lercher MJ and Pal C. 2008. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Mol Biol Evol* 25(3):559–567.

Ley JD, Tijtgat R, Smedt JD, and Michiels M. 1973. Thermal stability of DNA: DNA hybrids within the genus agrobacterium. *Journal of General Microbiology* 78(2):241–252.

Li A, Geng J, Cui D, Shu C, Zhang S, Yang J, Xing J, Wang J, Ma F, and Hu S. 2011. Genome sequence of agrobacterium tumefaciens strain f2, a bioflocculant-producing bacterium. *Journal of Bacteriology* 193(19):5531–5531. PMID: 21914861.

Lindstrom K and Martinez-Romero ME. 2002. International committee on systematics of prokaryotes: Subcommittee on the taxonomy of agrobacterium and rhizobium. *International Journal of Systematic and Evolutionary Microbiology* 52(6):2337–2337.

Lindström K and Young JPW. 2011. International committee on systematics of Prokaryotes-Subcommittee on the taxonomy of agrobacterium and rhizobium minutes of the meeting, 7 september 2010, geneva, switzerland. *International Journal of Systematic and Evolutionary Microbiology* 61(12):3089–3093. PMID: 22156799.

Lobry JR and Sueoka N. 2002. Asymmetric directional mutation pressures in bacteria. *Genome Biology* 3(10):RESEARCH0058. PMID: 12372146.

Loehfelm TW, Luke NR, and Campagnari AA. 2008. Identification and characterization of an acinetobacter baumannii biofilm-associated protein. *Journal of bacteriology* 190(3):1036–1044. PMID: 18024522.

Luo C, Walk ST, Gordon DM, Feldgarden M, Tiedje JM, and Konstantinidis KT. 2011. Genome sequencing of environmental escherichia coli expands understanding of the ecology and speciation of the model bacterial species. *Proceedings of the National Academy of Sciences* 108(17):7200–7205.

Majewski J, Zawadzki P, Pickerill P, Cohan FM, and Dowson CG. 2000. Barriers to genetic exchange between bacterial species: Streptococcus pneumoniae transformation. *Journal of bacteriology* 182(4):1016–1023. PMID: 10648528.

Makarova K, Slesarev A, Wolf Y, Sorokin A, Mirkin B, Koonin E, Pavlov A, Pavlova N, Karamychev V, Polouchine N, Shakhova V, Grigoriev I, Lou Y, Rohksar D, Lucas S, Huang K, Goodstein DM, Hawkins T, Plengvidhya V, Welker D, Hughes J, Goh Y, Benson A, Baldwin K, Lee JH, Díaz-Muñiz I, Dosti B, Smeianov V, Wechter W, Barabote R, Lorca G, Altermann E, Barrangou R, Ganesan B, Xie Y, Rawsthorne H, Tamir D, Parker C, Breidt F, Broadbent J, Hutkins R, O'Sullivan D, Steele J, Unlu G, Saier M, Klaenhammer T, Richardson P, Kozyavkin S, Weimer B, and Mills D. 2006. Comparative genomics of the lactic acid bacteria. *Proceedings of the National Academy of Sciences* 103(42):15611 –15616.

Makarova KS and Koonin EV. 2005. Evolutionary and functional genomics of the archaea. *Current Opinion in Microbiology* 8(5):586–594.

Marchal M, Briandet R, Koechler S, Kammerer B, and Bertin PN. 2010. Effect of arsenite on swimming motility delays surface colonization in herminiimonas arsenicoxydans. *Microbiology* 156(8):2336–2342. PMID: 20447996.

Markowitz VM, Chen IMA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Anderson I, Lykidis A, Mavromatis K, Ivanova NN, and Kyrpides NC. 2009. The integrated microbial genomes system: an expanding comparative analysis resource. *Nucleic Acids Research* 38(Database):D382–D390.

Marri PR, Hao W, and Golding GB. 2007. The role of laterally transferred genes in adaptive evolution. *BMC Evolutionary Biology* 7(Suppl 1):S8–S8. PMID: 17288581 PMCID: 1796617.

Marri PR, Harris LK, Houmiel K, Slater SC, and Ochman H. 2008. The effect of chromosome geometry on genetic diversity. *Genetics* 179(1):511–516. PMID: 18493068 PMCID: 2390628.

Mauchline TH, Fowler JE, East AK, Sartor AL, Zaheer R, Hosie AHF, Poole PS, and Finan TM. 2006. Mapping the sinorhizobium meliloti 1021 solute-binding protein-dependent transportome. *Proceedings of the National Academy of Sciences of the United States of America* 103(47):17933–17938. PMID: 17101990.

Mayr E. 1942. *Systematics and the origin of species, from the viewpoint of a zoologist*. Harvard University Press.

Meikle PJ, Perry MB, Cherwonogrodzky JW, and Bundle DR. 1989. Fine structure of a and m antigens from brucella biovars. *Infection and immunity* 57(9):2820–2828. PMID: 2474504.

Miele V, Penel S, Daubin V, Picard F, Kahn D, and Duret L. 2012. High-quality sequence clustering guided by network topology and multiple alignment likelihood. *Bioinformatics (Oxford, England)*. PMID: 22368255.

Miele V, Penel S, and Duret L. 2011. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics* 12:116. PMID: 21513511.

Mira A, Ochman H, and Moran NA. 2001. Deletional bias and the evolution of bacterial genomes. *Trends in Genetics* 17(10):589–596.

Mirkin B, Fenner T, Galperin M, and Koonin E. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evolutionary Biology* 3(1):2.

Mizoguchi H, Mori H, and Fujio T. 2007. Escherichia coli minimum genome factory. *Biotechnology and Applied Biochemistry* 46(3):157.

Mizoguchi H, Sawano Y, Kato Ji, and Mori H. 2008. Superpositioning of deletions promotes growth of escherichia coli with a reduced genome. *DNA Research* 15(5):277 –284.

Moore LW, Chilton WS, and Canfield ML. 1997. Diversity of opines and opine-catabolizing bacteria isolated from naturally occurring crown gall tumors. *Applied and Environmental Microbiology* 63(1):201–207. PMID: 16535484.

Moran NA and Plague GR. 2004. Genomic changes following host restriction in bacteria. *Current Opinion in Genetics & Development* 14(6):627–633.

Morrow JD and Cooper VS. 2012. Evolutionary effects of translocations in bacterial genomes. *Genome Biology and Evolution* 4(12):1256–1262.

Morton ER, Merritt PM, Bever JD, and Fuqua C. 2013. Large deletions in the pAtC58 megaplasmid of agrobacterium tumefaciens can confer reduced carriage cost and increased expression of virulence genes. *Genome Biology and Evolution* 5(7):1353–1364. PMID: 23783172.

Mougel C. 2000. *Structure génétique des populations d'Agrobacterium spp.: effet sélectif de la plante et implication dans la diffusion conjugative du plasmide Ti*. Number 00-90 in Thèses Doctorat Lyon 1/Sciences. Lyon.

Mougel C, Cournoyer B, and Nesme X. 2001. Novel tellurite-amended media and specific chromosomal and ti plasmid probes for direct analysis of soil populations of agrobacterium biovars 1 and 2. *Applied and environmental microbiology* 67(1):65–74. PMID: 11133429.

Muller J, Szklarczyk D, Julien P, Letunic I, Roth A, Kuhn M, Powell S, von Mering C, Doerks T, Jensen LJ, and Bork P. 2010. eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Research* 38(Database issue):D190–195. PMID: 19900971.

Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, Croucher NJ, Choi SY, Harris SR, Lebens M, Niyogi SK, Kim EJ, Ramamurthy T, Chun J, Wood JLN, Clemens JD, Czerkinsky C, Nair GB, Holmgren J, Parkhill J, and Dougan G. 2011. Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* 477(7365):462–465.

Myllykallio H, Lopez P, Lopez-Garcia P, Heilig R, Saurin W, Zivanovic Y, Philippe H, and Forterre P. 2000. Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. *Science* 288(5474):2212–2215.

Médigue C, Rouxel T, Vigier P, Hénaut A, and Danchin A. 1991. Evidence for horizontal gene transfer in escherichia coli speciation. *Journal of Molecular Biology* 222(4):851–856. PMID: 1762151.

Nguyen TH, Doyon JP, Pointet S, Chifolleau AMA, Ranwez V, and Berry V. 2012. Accounting for gene tree uncertainties improves gene trees and reconciliation inference. In Raphael B

and Tang J, editors, *Algorithms in Bioinformatics* number 7534 in Lecture Notes in Computer Science pages 123–134. Springer Berlin Heidelberg.

Nguyen TH, Ranwez V, Pointet S, Chifolleau AMA, Doyon JP, and Berry V. 2013. Reconciliation and local gene tree rearrangement can be of mutual profit. *Algorithms for molecular biology: AMB* 8(1):12. PMID: 23566548.

Normand P, Lapierre P, Tisa LS, Gogarten JP, Alloisio N, Bagnarol E, Bassi CA, Berry AM, Bickhart DM, Choisne N, Couloux A, Cournoyer B, Cruveiller S, Daubin V, Demange N, Francino MP, Goltsman E, Huang Y, Kopp OR, Labarre L, Lapidus A, Lavire C, Marechal J, Martinez M, Mastronunzio JE, Mullin BC, Niemann J, Pujic P, Rawnsley T, Rouy Z, Schenowitz C, Sellstedt A, Tavares F, Tomkins JP, Vallenet D, Valverde C, Wall LG, Wang Y, Medigue C, and Benson DR. 2007. Genome characteristics of facultatively symbiotic frankia sp. strains reflect host range and host plant biogeography. *Genome Research* 17(1):7 –15.

Ochman H and Davalos LM. 2006. The nature and dynamics of bacterial genomes. *Science (New York, N.Y.)* 311(5768):1730–1733. PMID: 16556833.

Ochman H and Moran NA. 2001. Genes lost and genes found: Evolution of bacterial pathogenesis and symbiosis. *Science* 292(5519):1096–1099.

Ophel K and Kerr A. 1990. Agrobacterium vitis sp. nov. for strains of agrobacterium biovar 3 from grapevines. *International Journal of Systematic Bacteriology* 40(3):236–241.

Otten L, Burr T, and Szegedi E. 2008. Agrobacterium: A disease-causing bacterium. In Tzfira T and Citovsky V, editors, *Agrobacterium: From Biology to Biotechnology* pages 1–46. Springer New York.

Pagel M. 1997. Inferring evolutionary processes from phylogenies. *Zoologica Scripta* 26(4):331–348.

Pal C, Papp B, and Lercher MJ. 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet* 37(12):1372–1375.

Penel S, Arigon AM, Dufayard JF, Sertier AS, Daubin V, Duret L, Gouy M, and Perriere G. 2009. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics* 10(Suppl 6):S3.

Pesquita C, Faria D, Falcão AO, Lord P, and Couto FM. 2009. Semantic similarity in biomedical ontologies. *PLoS Comput Biol* 5(7):e1000443.

Popa O, Hazkani-Covo E, Landan G, Martin W, and Dagan T. 2011. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Research* 21(4):599–609. PMID: 21270172.

Popoff MY, Kersters K, Kiredjian M, Miras I, and Coynault C. 1984. [taxonomic position of agrobacterium strains of hospital origin]. *Annales De Microbiologie* 135A(3):427–442. PMID: 6087709.

Portier P, Fischer-Le Saux M, Mougel C, Lerondelle C, Chapulliot D, Thioulouse J, and Nesme X. 2006. Identification of genomic species in agrobacterium biovar 1 by AFLP genomic markers. *Applied and Environmental Microbiology* 72(11):7123–7131. PMID: 16936063.

Porwollik S, Wong RMY, and McClelland M. 2002. Evolutionary genomics of salmonella: Gene acquisitions revealed by microarray analysis. *Proceedings of the National Academy of Sciences of the United States of America* 99(13):8956–8961. PMID: 12072558 PMCID: 124405.

Puigbò P, Wolf YI, and Koonin EV. 2009. Search for a 'Tree of life' in the thicket of the phylogenetic forest. *Journal of biology* 8(6):59. PMID: 19594957.

Puławska J, Willems A, De Meyer SE, and Süle S. 2012. Rhizobium nepotum sp. nov. isolated from tumors on different plant species. *Systematic and Applied Microbiology* 35(4):215–220.

Pál C, Papp B, and Lercher MJ. 2005. Horizontal gene transfer depends on gene content of the host. *Bioinformatics (Oxford, England)* 21 Suppl 2:ii222–223. PMID: 16204108.

Pósfai G, Plunkett r Guy, Fehér T, Frisch D, Keil GM, Umenhoffer K, Kolisnychenko V, Stahl B, Sharma SS, de Arruda M, Burland V, Harcum SW, and Blattner FR. 2006. Emergent properties of reduced-genome escherichia coli. *Science (New York, N.Y.)* 312(5776):1044–1046. PMID: 16645050.

Ragan MA. 2001. On surrogate methods for detecting lateral gene transfer. *FEMS Microbiology Letters* 201(2):187–191.

Rahman MH, Biswas K, Hossain MA, Sack RB, Mekalanos JJ, and Faruque SM. 2008. Distribution of genes for virulence and ecological fitness among diverse vibrio cholerae population in a cholera endemic area: Tracking the evolution of pathogenic strains. *DNA and Cell Biology* 27(7):347–355.

Rasmussen MD and Kellis M. 2012. Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome research* 22(4):755–765. PMID: 22271778.

Retchless AC and Lawrence JG. 2007. Temporal fragmentation of speciation in bacteria. *Science (New York, N.Y.)* 317(5841):1093–1096. PMID: 17717188.

Retchless AC and Lawrence JG. 2010. Phylogenetic incongruence arising from fragmented speciation in enteric bacteria. *Proceedings of the National Academy of Sciences of the United States of America* 107(25):11453–11458. PMID: 20534528.

Roberts MS and Cohan FM. 1993. The effect of DNA sequence divergence on sexual isolation in bacillus. *Genetics* 134(2):401–408. PMID: 8325477.

Rocha EP. 2008. The organization of the bacterial genome. *Annual Review of Genetics* 42(1):211–233.

Rocha EPC. 2006. Inference and analysis of the relative stability of bacterial chromosomes. *Molecular Biology and Evolution* 23(3):513 –522.

Rocha EPC and Danchin A. 2003. Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Research* 31(22):6570–6577. PMID: 14602916.

Rocha EPC and Danchin A. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Molecular Biology and Evolution* 21(1):108–116. PMID: 14595100.

Romiguier J, Figuet E, Galtier N, Douzery EJP, Boussau B, Dutheil JY, and Ranwez V. 2012. Fast and robust characterization of time-heterogeneous sequence evolutionary processes using substitution mapping. *PLoS ONE* 7(3):e33852.

Ruffing AM, Castro-Melchor M, Hu WS, and Chen RR. 2011. Genome sequence of the curdlan-producing agrobacterium sp. strain ATCC 31749. *Journal of bacteriology* 193(16):4294–4295. PMID: 21685288.

Schirmer T and Jenal U. 2009. Structural and mechanistic determinants of c-di-GMP signalling. *Nature Reviews Microbiology* 7(10):724–735.

Schlicker A, Domingues FS, Rahnenführer J, and Lengauer T. 2006. A new measure for functional similarity of gene products based on gene ontology. *BMC bioinformatics* 7:302. PMID: 16776819.

Shams M, Campillo T, Lavire C, Muller D, Nesme X, and Vial L. 2012. Rapid and efficient methods to isolate, type strains and determine species of agrobacterium spp. in pure culture and complex environments. In Jimenez-Lopez JC, editor, *Biochemical Testing*. InTech.

Shams M, Vial L, Chapulliot D, Nesme X, and Lavire C. 2013. Rapid and accurate species and genomic species identification and exhaustive population diversity assessment of agrobacterium spp. using recA-based PCR. *Systematic and applied microbiology* 36(5):351–358. PMID: 23578959.

Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabó G, Polz MF, and Alm EJ. 2012. Population genomics of early events in the ecological differentiation of bacteria. *Science* 336(6077):48–51. PMID: 22491847.

Sheppard SK, Didelot X, Jolley KA, Darling AE, Pascoe B, Meric G, Kelly DJ, Cody A, Colles FM, Strachan NJC, Ogden ID, Forbes K, French NP, Carter P, Miller WG, McCarthy ND, Owen R, Litrup E, Egholm M, Affourtit JP, Bentley SD, Parkhill J, Maiden MCJ, and Falush D. 2013. Progressive genome-wide introgression in agricultural campylobacter coli. *Molecular ecology* 22(4):1051–1064. PMID: 23279096.

Sikorski J and Nevo E. 2005. Adaptation and incipient sympatric speciation of bacillus simplex under microclimatic contrast at "Evolution canyons" i and II, israel. *Proceedings of the National Academy of Sciences of the United States of America* 102(44):15924 –15929.

Slater S, Setubal JC, Goodner B, Houmiel K, Sun J, Kaul R, Goldman BS, Farrand SK, Almeida N, Burr T, Nester E, Rhoads DM, Kadoi R, Ostheimer T, Pride N, Sabo A, Henry E, Telepak E, Cromes L, Harkleroad A, Oliphant L, Pratt-Szegila P, Welch R, and Wood D. 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. *Applied and Environmental Microbiology* 79(4):1414–1417.

Slater SC, Goldman BS, Goodner B, Setubal JC, Farrand SK, Nester EW, Burr TJ, Banta L, Dickerman AW, Paulsen I, Otten L, Suen G, Welch R, Almeida NF, Arnold F, Burton OT, Du Z, Ewing A, Godsy E, Heisel S, Houmiel KL, Jhaveri J, Lu J, Miller NM, Norton S, Chen Q, Phoolcharoen W, Ohlin V, Ondrusek D, Pride N, Stricklin SL, Sun J, Wheeler C, Wilson L, Zhu H, and Wood DW. 2009. Genome sequences of three agrobacterium biovars help elucidate the evolution of multichromosome genomes in bacteria. *Journal of Bacteriology* 191(8):2501–2511. PMID: 19251847.

Smith NH, Gordon SV, de la Rua-Domenech R, Clifton-Hadley RS, and Hewinson RG. 2006. Bottlenecks and broomsticks: the molecular evolution of mycobacterium bovis. *Nature Reviews Microbiology* 4(9):670–681.

Snel B, Bork P, and Huynen MA. 2002. Genomes in flux: The evolution of archaeal and proteobacterial gene content. *Genome Research* 12(1):17 –25.

Song JK, Ahn HJ, Kim HS, and Song BK. 2006. Molecular cloning and expression of perhydrolase genes from pseudomonas aeruginosa and burkholderia cepacia in escherichia coli. *Biotechnology Letters* 28(12):849–856.

Stackebrandt E, Frederiksen W, Garrity GM, Grimont PAD, Kämpfer P, Maiden MCJ, Nesme X, Rosselló-Mora R, Swings J, Trüper HG, Vauterin L, Ward AC, and Whitman WB. 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *International Journal of Systematic and Evolutionary Microbiology* 52(Pt 3):1043–1047. PMID: 12054223.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21):2688–2690.

Sueoka N. 1988. Directional mutation pressure and neutral molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America* 85(8):2653.

Suyama M, Torrents D, and Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research* 34(Web Server):W609–W612.

Szöllosi GJ and Daubin V. 2012. Modeling gene family evolution and reconciling phylogenetic discord. *Methods in molecular biology (Clifton, N.J.)* 856:29–51. PMID: 22399454.

Szöllosi GJ, Tannier E, Lartillot N, and Daubin V. 2013. Lateral gene transfer from the dead. *Systematic biology*. PMID: 23355531.

Szöllősi GJ, Boussau B, Abby SS, Tannier E, and Daubin V. 2012. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proceedings of the National Academy of Sciences* 109(43):17513–17518.

Szöllősi GJ, Rosikiewicz W, Boussau B, Tannier E, and Daubin V. 2013. Efficient exploration of the space of reconciled gene trees. arXiv e-print 1306.2167.

Team RDC. 2009. *R: A language and environment for statistical computing*. 1(09/18/2009):ISBN 3–900051–07–0.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, DeBoy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, and Fraser CM. 2005. Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: Implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America* 102(39):13950 –13955.

Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, Calteau A, Chiapello H, Clermont O, Cruveiller S, Danchin A, Diard M, Dossat C, Karoui ME, Frapy E, Garry L, Ghigo JM, Gilles AM, Johnson J, Le Bouguénec C, Lescat M, Mangenot S, Martinez-Jéhanne V, Matic I, Nassif X, Oztas S, Petit MA, Pichon C, Rouy Z, Ruf CS, Schneider D, Tourret J, Vacherie B, Vallenet D, Médigue C, Rocha EPC, and Denamur E. 2009. Organised genome dynamics in the escherichia coli species results in highly diverse adaptive paths. *PLoS Genetics* 5(1):e1000344. PMID: 19165319.

Treangen TJ and Rocha EPC. 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet* 7(1):e1001284.

Ussery DW, Kiil K, Lagesen K, Sicheritz-Pontén T, Bohlin J, and Wassenaar TM. 2009. The genus burkholderia: analysis of 56 genomic sequences. *Genome Dynamics* 6:140–157. PMID: 19696499.

Vallenet D, Engelen S, Mornico D, Cruveiller S, Fleury L, Lajus A, Rouy Z, Roche D, Salvignol G, Scarpelli C, and Médigue C. 2009. MicroScope: a platform for microbial genome annotation and comparative genomics. *Database: The Journal of Biological Databases and Curation* 2009:bap021. PMID: 20157493.

van Passel MWJ, Marri PR, and Ochman H. 2008. The emergence and fate of horizontally acquired genes in escherichia coli. *PLoS Computational Biology* 4(4). PMID: 18404206 PMCID: 2275313.

Vesth T, Wassenaar TM, Hallin PF, Snipen L, Lagesen K, and Ussery DW. 2010. On the origins of a vibrio species. *Microbial Ecology* 59(1):1–13. PMID: 19830476.

Vetsigian K and Goldenfeld N. 2005. Global divergence of microbial genome sequences mediated by propagating fronts. *Proceedings of the National Academy of Sciences of the United States of America* 102(20):7332–7337. PMID: 15878987.

Viklund J, Ettema TJG, and Andersson SGE. 2012. Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. *Molecular Biology and Evolution* 29(2):599–615.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, and Birney E. 2009. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research* 19(2):327–335.

Vizcaíno N, Cloeckaert A, Zygmunt MS, and 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. *Infection and immunity* 69(11):6738–6748. PMID: 11598046.

Vogel J, Normand P, Thioulouse J, Nesme X, and Grundmann GL. 2003. Relationship between spatial and genetic distance in agrobacterium spp. in 1 cubic centimeter of soil. *Applied and Environmental Microbiology* 69(3):1482–1487. PMID: 12620832.

Vos M. 2011. A species concept for bacteria based on adaptive divergence. *Trends in microbiology* 19(1):1–7. PMID: 21071229.

Watson B, Currier TC, Gordon MP, Chilton MD, and Nester EW. 1975. Plasmid required for virulence of agrobacterium tumefaciens. *Journal of Bacteriology* 123(1):255–264. PMID: 1141196.

Wibberg D, Blom J, Jaenicke S, Kollin F, Rupp O, Scharf B, Schneiker-Bekel S, Sczcepanowski R, Goesmann A, Setubal JC, Schmitt R, Pühler A, and Schlüter A. 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. *Journal of Biotechnology* In Press, Uncorrected Proof.

Williams D, Gogarten JP, and Papke RT. 2012. Quantifying homologous replacement of loci between haloarchaeal species. *Genome Biology and Evolution* 4(12):1223–1244.

Williams KP, Sobral BW, and Dickerman AW. 2007. A robust species tree for the alphaproteobacteria. *J. Bacteriol.* 189(13):4578–4586.

Williamson SJ, Rusch DB, Yooseph S, Halpern AL, Heidelberg KB, Glass JI, Andrews-Pfannkoch C, Fadrosh D, Miller CS, Sutton G, Frazier M, and Venter JC. 2008. The sorcerer II global ocean sampling expedition: Metagenomic characterization of viruses within aquatic microbial samples. *PLoS ONE* 3(1):e1456.

Woese CR and Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences* 74(11):5088 –5090.

Woese CR, Stackebrandt E, Macke TJ, and Fox GE. 1985. A phylogenetic definition of the major eubacterial taxa. *Systematic and applied microbiology* 6:143–151. PMID: 11542017.

Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, and Koonin EV. 2001. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC evolutionary biology* 1:8. PMID: 11734060.

Wood DW, Setubal JC, Kaul R, Monks DE, Kitajima JP, Okura VK, Zhou Y, Chen L, Wood GE, Almeida NF, Woo L, Chen Y, Paulsen IT, Eisen JA, Karp PD, Bovee D, Chapman P, Clendenning J, Deatherage G, Gillet W, Grant C, Kutyavin T, Levy R, Li MJ, McClelland E, Palmieri A, Raymond C, Rouse G, Saenphimmachak C, Wu Z, Romero P, Gordon D, Zhang S, Yoo H, Tao Y, Biddle P, Jung M, Krespan W, Perry M, Gordon-Kamm B, Liao L, Kim S, Hendrick C, Zhao ZY, Dolan M, Chumley F, Tingey SV, Tomb JF, Gordon MP, Olson MV, and Nester EW. 2001. The genome of the natural genetic engineer agrobacterium tumefaciens c58. *Science (New York, N.Y.)* 294(5550):2317–2323. PMID: 11743193.

Xu J, Kim J, Koestler BJ, Choi JH, Waters CM, and Fuqua C. 2013. Genetic analysis of agrobacterium tumefaciens unipolar polysaccharide production reveals complex integrated control of the motile-to-sessile switch. *Molecular Microbiology* page n/a–n/a.

Yang K, Heath LS, and Setubal JC. 2012. REGEN: ancestral genome reconstruction for bacteria. *Genes* 3(4):423–443.

Yap WH, Zhang Z, and Wang Y. 1999. Distinct types of rRNA operons exist in the genome of the actinomycete thermomonospora chromogena and evidence for horizontal transfer of an entire rRNA operon. *Journal of bacteriology* 181(17):5201–5209. PMID: 10464188.

Yasuhara A, Akiba-Goto M, and Aisaka K. 2005. Cloning and sequencing of the aldehyde oxidase gene from methylobacillus sp. KY4400. *Bioscience, biotechnology, and biochemistry* 69(12):2435–2438. PMID: 16377905.

Yin Y and Fischer D. 2008. Identification and investigation of ORFans in the viral world. *BMC Genomics* 9(1):24.

Yomtovian I, Teerakulkittipong N, Lee B, Moult J, and Unger R. 2010. Composition bias and the origin of ORFan genes. *Bioinformatics* 26(8):996–999.

Young JM, Kuykendall LD, Martínez-Romero E, Kerr A, and Sawada H. 2001. A revision of rhizobium frank 1889, with an emended description of the genus, and the inclusion of all species of agrobacterium conn 1942 and allorhizobium undicola de lajudie et al. 1998 as new combinations: Rhizobium radiobacter, r. rhizogenes, r. rubi, r. undicola and r. vitis. *International journal of systematic and evolutionary microbiology* 51(Pt 1):89–103. PMID: 11211278.

Young JM, Kuykendall LD, Martínez-Romero E, Kerr A, and Sawada H. 2003. Classification and nomenclature of agrobacterium and rhizobium. *International journal of systematic and evolutionary microbiology* 53(Pt 5):1689–1695. PMID: 13130069.

Young JPW, Crossman LC, Johnston AW, Thomson NR, Ghazoui ZF, Hull KH, Wexler M, Curson AR, Todd JD, Poole PS, Mauchline TH, East AK, Quail MA, Churcher C, Arrowsmith C, Cherevach I, Chillingworth T, Clarke K, Cronin A, Davis P, Fraser A, Hance Z, Hauser H, Jagels K, Moule S, Mungall K, Norbertczak H, Rabbinowitsch E, Sanders M, Simmonds M, Whitehead S, and Parkhill J. 2006. The genome of rhizobium leguminosarum has recognizable core and accessory components. *Genome Biology* 7(4):R34. PMID: 16640791.

Zuckerkandl E and Pauling L. 1965. Molecules as documents of evolutionary history. *Journal of theoretical biology* 8(2):357–366. PMID: 5876245.