Divergence et flux de gènes au sein du complexe d'espèces Xanthomonas axonopodis : histoires neutres et adaptatives

Karine Durand

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE
ÉCOLE PRATIQUE DES HAUTES ÉTUDES
Sciences de la Vie et de la Terre
Mémoire présenté par
DURAND Karine

pour l'obtention du Diplôme de l'École Pratique des Hautes Études

---

Divergence et flux de gènes au sein du complexe d'espèces *Xanthomonas axonopodis* : histoires neutre et adaptative

---

soutenu le 7 Décembre 2017 devant le jury composé de

Président	Bruno CANQUE (DE1-HDR)
Tuteur pédagogique	Thierry WIRTH (DE1-HDR)
Tuteurs scientifiques	Christophe LEMAIRE (MCF-HDR)
Tuteurs scientifiques	Marion LE SAUX (CR1)
Rapporteur	Caroline MONTEIL (Post-doc)
Examinateur	Céline LAVIRE (MCF-HDR)

Mémoire préparé sous la direction de :

Marion LE SAUX et Christophe LEMAIRE à l'IRHS, INRA, AGROCAMPUS-Ouest, Université d'Angers, SFR 4207 QUASAV. Directeur : Jean-Pierre RENOU

et de Thierry WIRTH au Laboratoire EPHE de Biologie intégrative des populations, UMR-CNRS 7205. Directeur : Thierry WIRTH

REMERCIEMENTS

Tout d'abord, je remercie chaleureusement les membres du jury d'avoir accepté d'évaluer mon travail dans le cadre de mon Diplôme de l'École Pratique des Hautes Études en Sciences de la Vie et de la Terre.

J'adresse mes remerciements au directeur de l'IRHS, Mr Jean pierre Renou qui m'a soutenue et encouragée dès le début de mon projet de formation. Il a été disponible pour formaliser mon stage entre mon équipe et l'équipe Ecofun.

Marie-Agnès Jacques, je souhaite te remercier d'avoir accepté mon projet. Je te remercie aussi de tous les échanges constructifs que nous avons eus. Du temps que tu as passé à m'expliquer tout l'intérêt de formuler correctement les questions de recherche et les hypothèses. Et pour m'offrir la possibilité de valoriser mes nouvelles compétences au travers un nouveau projet de recherche.

Je remercie Marion Le-Saux et Christophe Lemaire d'avoir accepté d'être mes tuteurs scientifiques. Christophe je te remercie particulièrement pour ton encadrement, ta patience, ainsi que pour la confiance que tu m'accordes. Tu m'as toujours encouragée et soutenue dans les périodes difficiles surtout quand les résultats du chapitre II n'étaient pas au rendez-vous. J'espère que ça n'a pas été trop dur pour toi de supporter mes fréquentes interrogations, mes "je n'ai pas bien compris...", et mes "est-ce que tu aurais 5 minutes ?", et cela malgré un emploi du temps très chargé. Grâce à toi, les mots coalescence, chaîne de Markov, D de Tajima .... ne sont plus des termes abstraits. Marion Le-Saux, je te remercie pour le temps que tu as passé, notamment lors de la rédaction de mon rapport. Merci à vous deux vous m'avez appris à réfléchir, à me poser des questions, à analyser, avec l'exigence et la rigueur qu'exige le travail du scientifique.

Merci à Thierry Wirth d'avoir accepté d'être mon tuteur pédagogique. J'ai beaucoup apprécié vos cours à l'EPHE, votre disponibilité, ainsi que vos conseils et vos encouragements notamment lors des comités de suivi. Merci aussi pour les relectures du rapport.

Merci à Martial Briand, bio-informaticien de l'équipe. Heureusement que tu étais là pour m'aider à faire mes premiers scripts en Bash et me faire bénéficier de tes scripts Perls. Merci de ta patience, tu as toujours su régler mes problèmes d'installation, ou de scripts qui ne fonctionnaient pas.

Merci aux personnes qui étaient présentes lors de mon comité de suivi et qui m'ont aiguillée lors de ces réunions.

Merci à Deborah Merda pour ses conseils, les multiples discussions et le partage de nos expériences sur l'utilisation de certains outils de génétique des populations.

Merci à Sophie qui a bien voulu jouer le rôle de candide (pas si candide que ça !) en relisant ce rapport et qui m'a souvent aidé à reformuler plus clairement mes idées.

Merci à tous mes collègues des équipes Emersys et Ecofun pour leur aide.

Merci à mes collègues de la pause déjeuner, Sophie, et Céline qui m'ont beaucoup aidé à évacuer le stress !

Je remercie ma famille pour m'avoir soutenue et encouragée dans cette voie.

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR
ET DE LA RECHERCHE

Divergence et flux de gènes au sein du complexe d'espèces *Xanthomonas axonopodis* : histoires neutre et adaptative

---

DURAND Karine

le 2017

Résumé

RÉSUMÉ

La compréhension des mécanismes d'évolution permettant la différenciation et la spéciation des bactéries est cruciale pour estimer le potentiel évolutif des agents pathogènes et permettre la prévention des nouvelles maladies des plantes. L'espèce bactérienne est à l'heure actuelle définie sur la base de similarité/dissemblance génomique. Cependant, cette approche ne nous renseigne en rien sur les mécanismes évolutifs conduisant à la divergence génétique et génomique sur laquelle est basée la notion d'espèce bactérienne. Dans ce travail, nous avons étudié les processus gouvernant la divergence génétique au sein du complexe d'espèces *Xanthomonas axonopodis* qui regroupe des souches responsables de maladies sur des cultures d'importance socio-économique majeure. L'étude de 73 génomes représentatifs de ce complexe a révélé une structuration en cinq groupes. La structuration de ce complexe n'étant pas liée à la géographie ou à l'hôte, nous avons étudié le rôle que peuvent avoir les flux de gènes dans la divergence et montré que l'impact de la recombinaison sur le polymorphisme était équivalent à celui de la mutation. Nous avons de plus noté une prépondérance de la recombinaison sur les branches précédant les événements de divergence en groupes. Une analyse de génomique des populations et des scénarios de divergence ont montré que les flux de gènes étaient plus faibles entre des groupes ayant divergé récemment qu'entre des groupes plus distants. L'absence de corrélation entre la distance génétique et le flux de gènes serait indicatrice de la présence de barrières génétiques ou écologiques entre ces populations. Le flux de gènes entre groupes distants serait imputable à du contact secondaire. Dans un deuxième temps il a été montré que la structuration du complexe basée sur la matrice de présence-absence des gènes du génome accessoire était différente de celle basée sur le core génome, traduisant l'impact du transfert horizontal de gènes (HGT). Nous avons pu identifier que le partage d'une même niche écologique était une condition nécessaire au HGT mais non suffisante. L'inférence des gains et des pertes de gènes pendant l'histoire évolutive suggère une intensification récente en accord avec les scénarios de divergence suivie de contacts secondaires inférés entre certains groupes. L'évolution des pratiques agricoles et la mondialisation pourraient être responsables de ces contacts secondaires. Le gain d'un cluster de gènes codant pour la biosynthèse des lipopolysaccharides avant la diversification du groupe 9.5 pourrait être impliqué dans la divergence de ce groupe. Les résultats présentés montrent que la divergence de ces groupes peut être due à l'accumulation de mutations, à la diminution de la recombinaison homologue suite à la présence de barrières plutôt qu'à la distance génétique, ou à l'acquisition de gènes qui pourraient favoriser l'isolement écologique de certains groupes.

MOTS-CLÉS : *Xanthomonas*, flux de gènes, recombinaison, spéciation, contacts secondaires, inférences démographiques, adaptation, Transfert Horizontal de Gènes.

Sommaire

Liste des figures

Liste des tableaux

Abbréviations

F_{ST} Indice de fixation

ACP Analyse en composantes principales

AIC Critère d'information d'Akaike

ANI Identités nucléotidiques moyennes

bp Base pair

BSC Biological Species Concept

HGT Transfert horizontal de gènes

IS Séquences d'insertions

$LG1$ Lignée génétique 1

LPS Lipopolysaccharide

$MCMC$ Markov Chain Monte Carlo

$PAMP$ Motif moléculaire associé aux agents pathogènes

RM Système de Restriction-Modification

SNP Single Nucleotide Polymorphism

SRM Système de réparation des mésappariements

$T3SS$ Système de Sécrétion de Type III

TA Toxine-Antitoxine

Première partie

Contexte

Les maladies des plantes sont un problème important entraînant de graves conséquences économiques et environnementales dans le Monde. Ce problème est d'autant plus grave que l'on note un important accroissement de ces maladies depuis le début du XXI$^{\text{ème}}$ siècle [Bartoli et al., 2016]. Parmi ces maladies certaines sont causées par des bactéries phytopathogènes pour lesquelles il existe peu de traitements efficaces, si l'on exclut les traitements antibiotiques interdits dans l'Union Européenne. La prévention et le contrôle sont alors les seuls moyens de limiter l'impact des maladies bactériennes sur plantes. En outre de nouvelles maladies apparaissent régulièrement soit en raison de facteurs favorisant leur émergence (conditions climatiques, introduction de nouvelles souches dans une zone géographique *via* les échanges internationaux...), soit par acquisition de nouveaux traits liés à la pathogénie [Giraud et al., 2010]. Parmi les nombreux exemples récents, on peut noter la détection, en 2013, en Italie, de *Xylella fastidiosa* agent pathogène sur un hôte d'intérêt économique, l'olivier [Saponari et al., 2014], ou encore le développement mondial du chancre bactérien du kiwi causé par *Pseudomonas syringae* pv. *actinidiae* originaire de Chine [Kim et al., 2016]. La compréhension des mécanismes d'évolution permettant la différenciation et la spéciation [1] des bactéries est cruciale pour estimer le potentiel évolutif des agents pathogènes et permettre la prévention et le contrôle des nouvelles maladies des plantes. L'interaction hôte-pathogène a été la plus étudiée en tant que facteur-clef de l'évolution des organismes pathogènes. Cependant, la focalisation sur cet unique aspect a souvent conduit à n'entrevoir la diversification de ces organismes que sous l'angle de la divergence écologique.

La divergence entre populations [2] bactériennes résulte de l'accumulation au cours du temps de nouvelles mutations, neutres ou adaptatives, dans chaque lignée génétique. L'analyse des polymorphismes ainsi générés, permet non seulement d'étudier les éventuels processus de divergence écologique dont la spécialisation sur différents hôtes mais aussi, bien que de manière moins évidente, les proces-

1. Le terme spéciation ici n'est pas équivalent *sensu stricto* à celui utilisé chez les eucaryotes. L'espèce bactérienne est elle-même difficile à définir (voir II.2), au-delà du simple isolement reproductif, les échanges génétiques étant possible entre différents genres bactériens.

2. Une population est définie comme une communauté d'individus qui vivent dans le même habitat (i.e. avec une cohésion démographique) et interagissent les uns avec les autres (i.e. avec une cohésion reproductive) [Waples and Gaggiotti, 2006]

sus démographiques impliquant, par exemple, la divergence neutre et le flux de gènes. Lorsque cette divergence est suffisamment importante elle permet l'identification de l'espèce bactérienne qui est basée sur des critères pragmatiques et empiriques utilisant le pourcentage d'hybridation ADN-ADN ou l'identité nucléotidique moyenne (ANI). On notera ici que l'identification d'espèces bactériennes repose avant tout sur la divergence génétique entre souches, et non, comme c'est le cas chez la plupart des organismes eucaryotes, sur l'existence d'un isolement reproductif. Toutefois, l'accumulation de différences génétiques neutres ou adaptatives peut aussi conduire, chez les bactéries, à un isolement reproductif partiel ou presque complet [Wielgoss et al., 2016].

La petite taille des génomes des bactéries (de 0,6 MB pour *Buchnera*, à 8,6 MB pour *Streptomyces coelicolor*), les rend faciles à séquencer. Ceci explique certainement la présence de nombreux génomes bactériens dans les bases de données accessibles en ligne (ex : NCBI, JGI DOE, etc...). Contrairement aux eucaryotes, dans lesquels il existe une large variation de la densité de gènes, et peu d'association entre la taille du génome et le nombre de gènes ou la complexité des organismes, la taille des génomes bactériens est fortement corrélée avec le nombre de gènes [Kuo and Ochman, 2009]. Les bactéries peuvent alors apparaître comme des organismes au génome plus compact, rare en éléments non-codant, et donc *a priori* plus susceptibles d'être affectés par des processus sélectifs que les eucaryotes. En effet, une plus faible proportion de génome non-codant, rend plus forte la probabilité d'apparition d'une mutation dans un gène, avec effet potentiel sur la fitness (valeur adaptative) de la bactérie.

Dans ce projet, nous proposons d'analyser par des approches génomiques, les processus adaptatifs et démographiques à l'œuvre dans le complexe d'espèces bactériennes *Xanthomonas axonopodis*. Les bactéries du genre *Xanthomonas* provoquent d'importantes maladies sur des plantes hôtes d'intérêt économique comme le haricot, le chou, le manioc, des agrumes, le chanvre, le poivron, le riz, la canne à sucre, la tomate, ou le blé [Denancé et al., 2016]. Le pathovar *Xanthomonas phaseoli* pv. *manihotis* par exemple, se place en $6^{ème}$ place des 10 bactéries phytopathogènes perçues comme les plus importantes scientifiquement ou économiquement [Mansfield et al., 2012]. Cette maladie affecte sévèrement la production de manioc dans le monde entier avec des pertes de rendement et de plants allant de

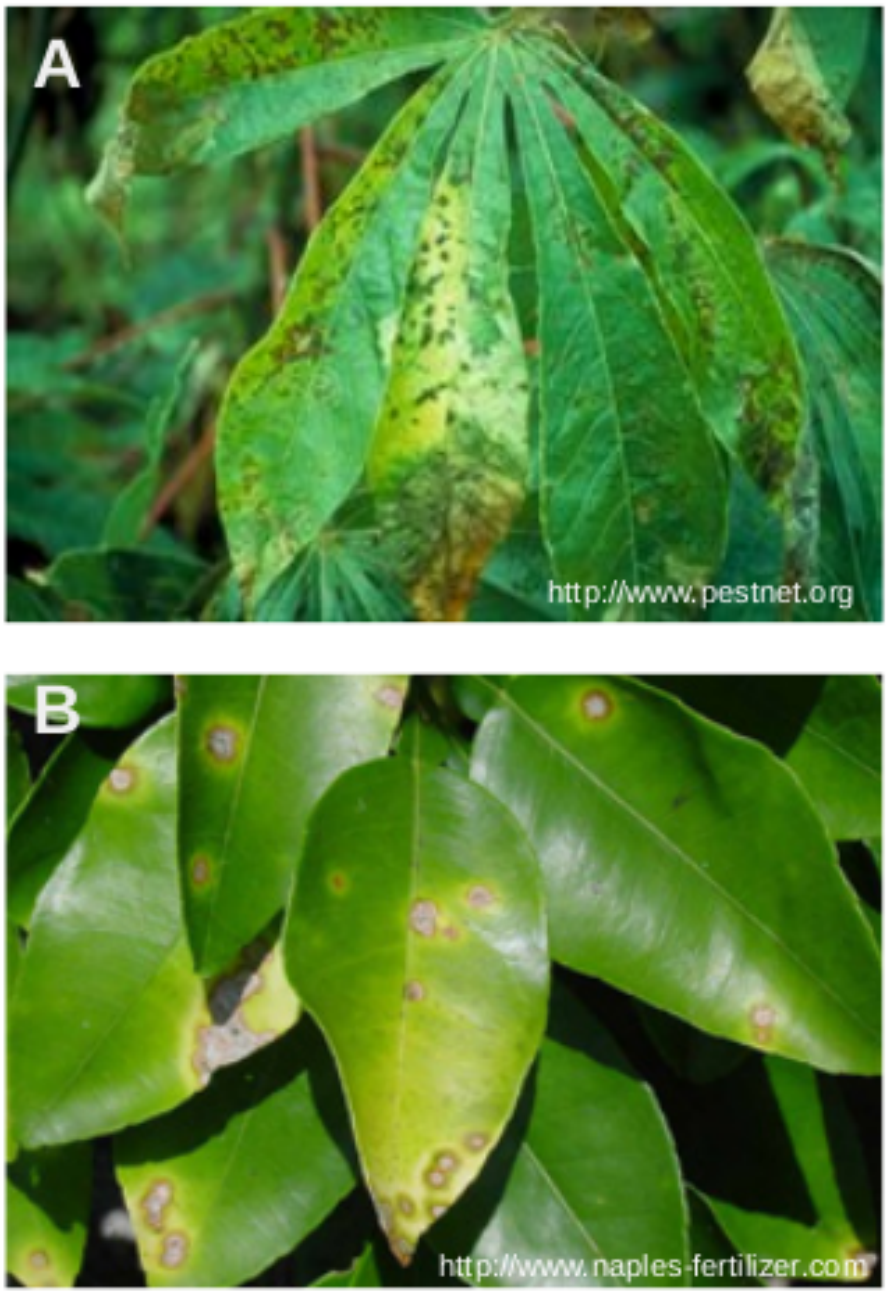


FIGURE 1 – Symptômes de bactérioses dues à *Xanthomonas*.
a) Brûlures foliaires provoquées par la bactérie *Xanthomonas phaseoli* pv. *manihotis* sur manioc b) Lésions sur feuilles de Citrus atteint par *Xanthomonas citri* pv. *citri*

12% à 100% [Verdier et al., 2004]**(fig. 1a)**. Un autre pathovar *Xanthomonas citri* pv. *citri* est un organisme de quarantaine A1 de l' Organisation européenne et méditerranéenne pour la protection des plantes [Smith et al., 1997]. Cette bactérie provoque en Floride, sur des grandes surfaces, des pertes de rendement moyennes estimées entre 2 et 5% sur oranger et 5 à 10% sur pomelo. En zone tropicale où la bactérie est établie, des pertes de rendement, par chutes de fruits avant récolte, de 30 à 50% sont observées couramment sur cultivars sensibles [Pruvost, 2004]**(fig. 1b)**.

Le complexe d'espèces *Xanthomonas axonopodis* présenterait plusieurs niveaux hiérarchisés de différenciation [Mhedbi-Hajri et al., 2013]. Un premier niveau de structuration, permet de distinguer cinq groupes de souches selon une logique non liée de manière évidente à l'écologie ou à l'origine géographique de ces souches. En revanche, le deuxième niveau, plus récent du point de vu évolutif, indique une forte structure de population en fonction des hôtes d'isolement. Notre équipe possède la plus importante collection de génomes de *Xanthomonas axonopodis* échantillonnés à des dates variant de 1942 à 2012, dans le monde entier sur une très grande gamme d'hôtes. Cette collection génomique nous a permis d'analyser les processus évolutifs responsables de la diversification et de la spécialisation d'hôte *via* les outils de la génomique des populations.

Deuxième partie

Synthèse bibliographique

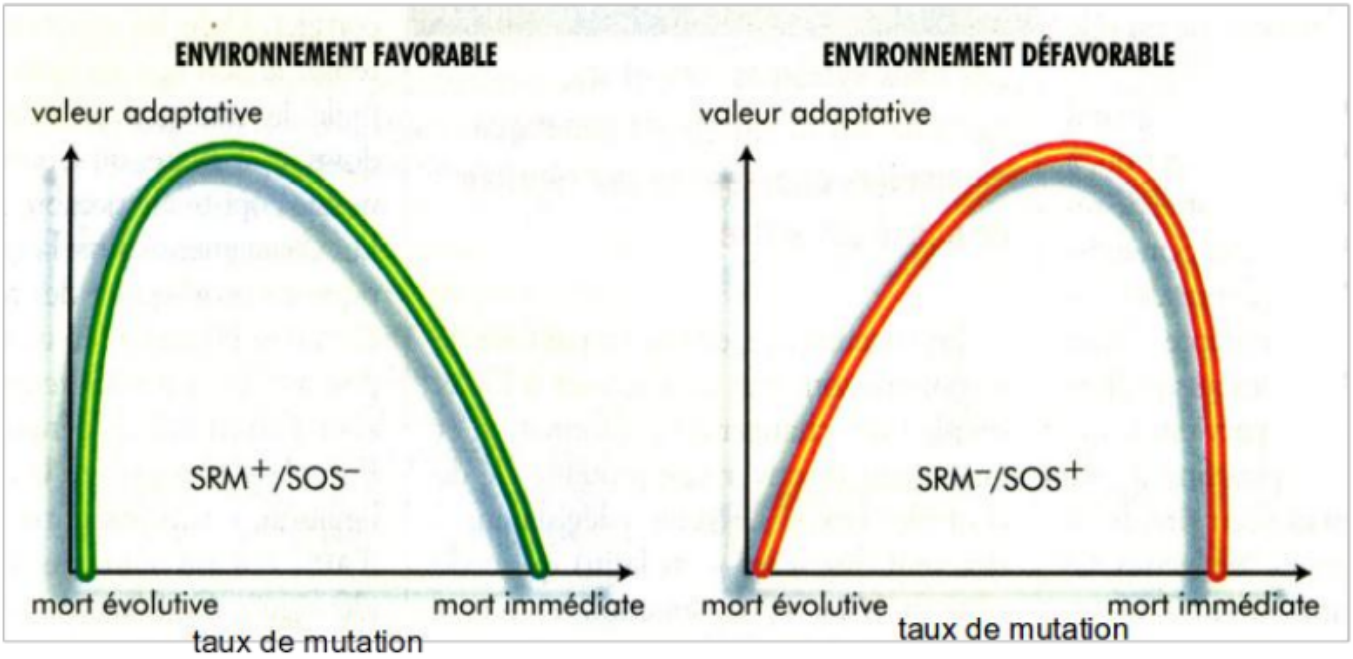


FIGURE 2 – La modulation des systèmes SRM et SOS.
Lorsque l'environnement est favorable à la bactérie, le SRM assure que lors de la recombinaison entre des séquences d'ADN partiellement différentes, la différence n'excède pas un certain seuil risquant de provoquer un réarrangement chromosomique ou d'autoriser la recombinaison avec une séquence provenant d'une espèce différente. En cas de stress, lorsque la bactérie doit s'adapter à un nouvel environnement pour survivre, la réponse SOS est déclenchée, cela va permettre l'insertion et le réarrangement des cassettes de gènes, ce qui constitue un avantage adaptatif. Si le SRM est inactivé, l'ADN lésé peut se réarranger avec de l'ADN venu d'une autre espèce. D'après Taddei F., Matic I., et Radman M. (1996)[Taddei F. et al., 1996].

1 Les mécanismes moléculaires conditionnant la divergence chez les bactéries

Le niveau de divergence des populations bactériennes dépend de l'intensité de la mutation, de la dérive, de la sélection et du flux de gènes. En effet, si la mutation et la dérive ont pour effet d'augmenter en moyenne les différences génétiques entre lignées, le flux génique diminuera ces différences. La reproduction sexuée n'existant pas chez les bactéries, le flux génique est réalisé *via* d'autres modalités comme la recombinaison homologue ou le transfert horizontal de gènes.

1.1 La mutation

La mutation est l'un des mécanismes conditionnant la divergence. Elle se produit le plus souvent suite à des erreurs dans le processus de réplication de l'ADN. Les mutations ponctuelles correspondent à la substitution d'une base par une autre (SNP pour *Single Nucleotide Polymorphism*) ou à l'insertion/délétion d'une base (indel). Chez la bactérie *Escherichia coli*, le taux de mutation par génération est estimé entre $4{,}1.10^{-3}$ et 6.10^{-2} [Wielgoss et al., 2013]. Cependant toutes les substitutions n'ont pas la même probabilité d'apparition. Chez les bactéries, il a été montré que les transitions G $\rightarrow$ A et C $\rightarrow$ T sont les mutations ponctuelles les plus fréquentes [Hershberg and Petrov, 2010].

De nombreux gènes sont indispensables à la survie, trop de polymorphisme dans ces gènes conduirait à la mort de la cellule. D'un autre côté, trop peu de variabilité empêche l'adaptation à un nouvel environnement, pouvant aboutir à l'extinction de la population. Les altérations génétiques apparaissent donc comme le prix à payer pour l'évolution. La diversité génétique est finement modulée par un couple antagoniste, le système SOS et le système de réparation des mésappariements (SRM) [Schofield and Hsieh, 2003]. Le SRM empêche l'apparition de mutations. Le système SOS, créateur de variabilité, peut être induit par un stress. Le stress provoqué par des agents chimiques (Méthanesulfonate d'éthyle, agents alkylants...) ou physiques (UV, radiations ionisantes...) entraîne une moins bonne fidélité de la machinerie de réplication et une introduction accrue de mutations ponctuelles. Ceci va entraîner un relâchement du système de réparation des mésappariements

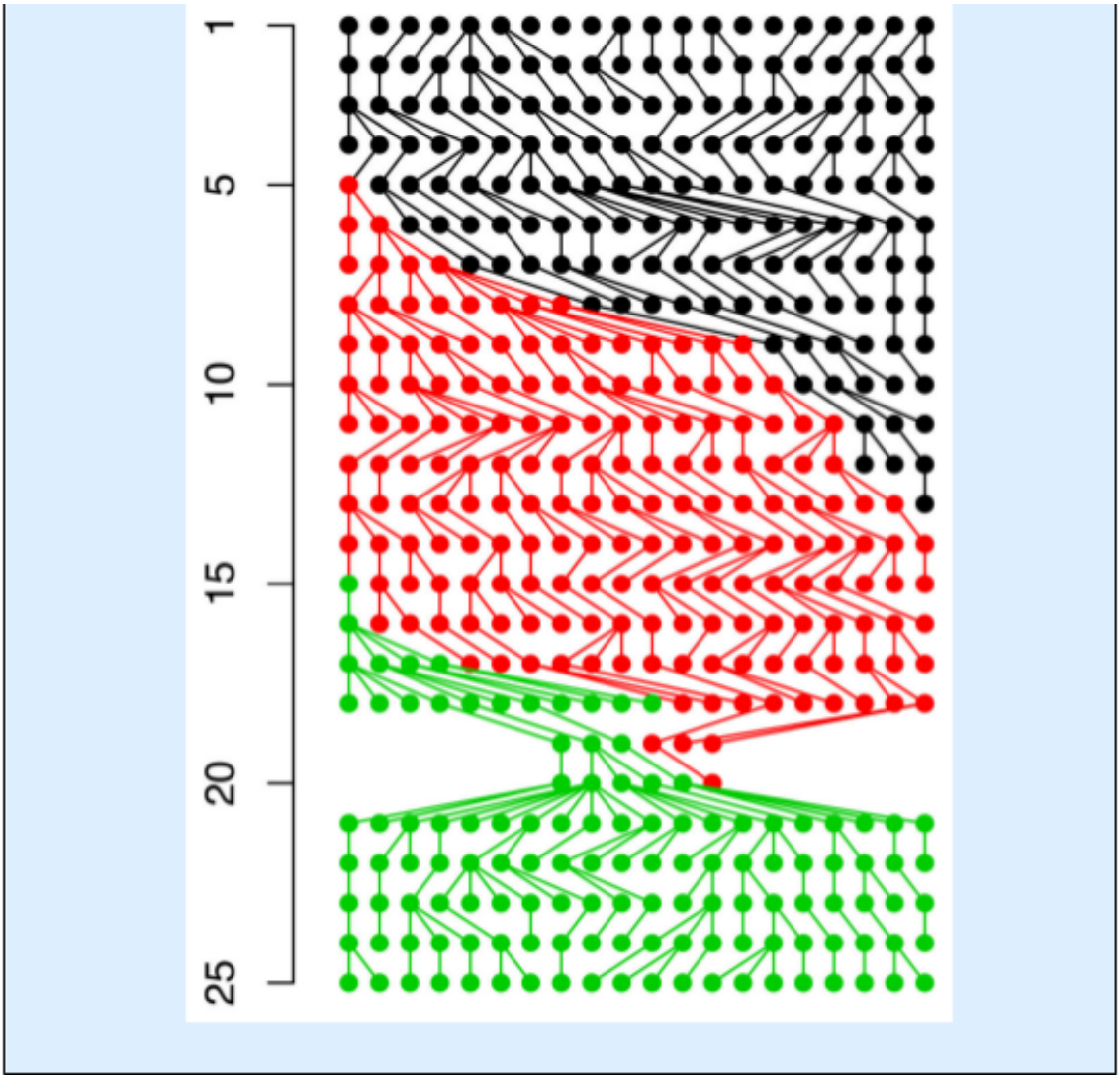


FIGURE 3 – La dérive génétique.
La dérive génétique est le processus par lequel les fréquences alléliques fluctuent au fil du temps. Pour illustrer ce concept, l'évolution d'une population bactérienne a été simulée conformément aux suppositions du modèle de Wright-Fisher. La population a une taille de population efficace initiale de 20 bactéries et chaque génération (indiquée en ordonnée) est formée en choisissant aléatoirement des parents dans la génération précédente. À la génération 5, et 15 l'individu de gauche subit une mutation (en rouge ou vert) et la nouvelle mutation est héritée par tous ses descendants. Cette mutation avantage légèrement les individus ce qui fait qu'ils ont plus de descendants que les autres bactéries et sa fréquence augmente dans la population. Aux générations 19 et 20, les bactéries subissent un goulet d'étranglement, la taille efficace de la population est réduite à seulement 6 individus. Les goulets d'étranglement augmentent l'effet de la dérive génétique. Dans la génération 21, l'allèle vert augmente rapidement dans la population et l'allèle rouge s'est éteint. Cette perte de variabilité est connue comme l'effet fondateur, qui se produit quand une population est initiée d'un petit nombre d'individus. D'après Didelot et al. (2016).

et donc, un état “mutateur” transitoire [McKenzie et al., 2001]. Il existe donc un taux de variabilité génétique qui optimise la valeur sélective (*fitness*) **(fig. 2)**.

Cette valeur optimale dépend de l'adaptation de l'individu à son environnement et varie donc avec les changements d'environnement, d'où la nécessité d'une modulation du taux de variabilité génétique dans la population [Radman et al., 1993]. Une fois ces mutations apparues, l'évolution de leur fréquence dépendra de trois forces évolutives : la dérive, la sélection et le flux génique.

1.2 La dérive génétique

La dérive génétique est une force évolutive dont les effets dépendent de l'effectif d'une population parce qu'elle est caractérisée par une fluctuation aléatoire, d'une génération à l'autre, des fréquences alléliques. Lors de l'échantillonnage aléatoire des allèles transmis à chaque génération, soit un nouveau variant se fixe dans la population, (cet allèle augmente en fréquence dans la population jusqu'à atteindre 1), soit il disparaît (cet allèle diminue en fréquence dans la population jusqu'à atteindre 0). Les effets de la dérive génétique sont particulièrement importants sur les populations d'effectif limité [Didelot et al., 2016]. Par exemple, quand une lignée bactérienne s'adapte à un mode de vie qui réduit sa taille efficace [3], ou à un habitat restreint (comment c'est le cas des *Dehalococcoides*) les nouvelles mutations ont plus de chances d'être fixées dans la population à cause de l'élévation de la dérive génétique [Kuo et al., 2009]. Dans le modèle de dérive génétique de Wright-Fisher, on modélise la transmission des gènes d'une génération à l'autre de façon très schématique, il s'agit d'un tirage aléatoire de 2N gamètes dans un ensemble infini de gamètes. Dans ce cas, N individus diploïdes (N/2 de chaque sexe) génèrent un pool infini de gamètes pour former N nouveaux individus diploïdes à la prochaine génération. Ce modèle est une simplification considérable du cycle de reproduction des populations naturelles. L'effet de la dérive génétique se produit lors du tirage des 2N gamètes pour former la nouvelle génération. Les hypothèses principales de ce modèle incluent le non chevauchement des générations, la même fitness de tous les individus, et une taille constante de la population au cours du temps. Ce modèle

3. On définit la taille efficace de la population comme l'effectif d'une population idéale (de type Wright-Fisher) pour laquelle on aurait une fluctuation du polymorphisme équivalente à celle de la population naturelle.

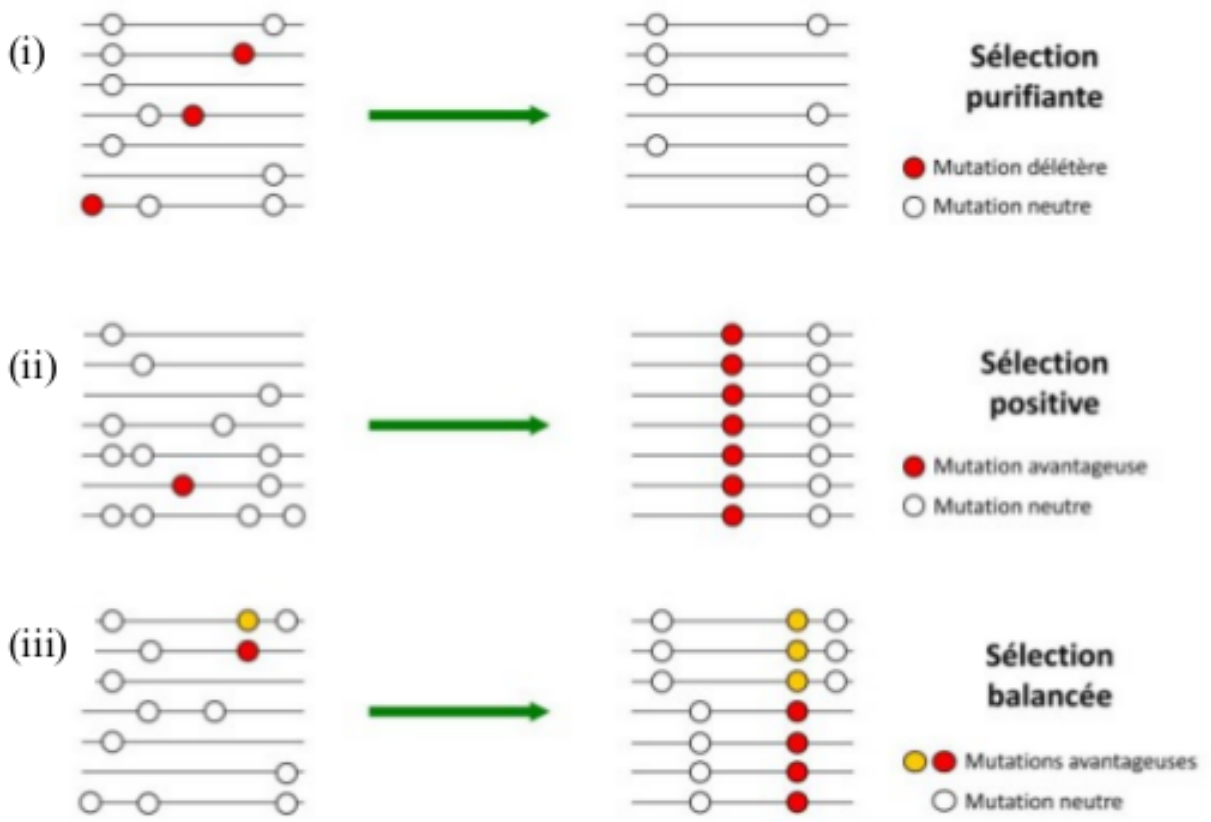


FIGURE 4 – La sélection.
Illustration des différents types de sélection au niveau moléculaire au sein d'une population. D'après [Montaigne, 2011]. (i) la sélection purifiante élimine rapidement les mutations défavorables, (ii) la sélection positive fixe rapidement les mutations favorables et (iii) la sélection balancée maintient plusieurs mutations bénéfiques en fréquence intermédiaire

peut s'adapter aux individus haploïdes. Sur le long terme, la dérive génétique conduit à un appauvrissement de la diversité génétique des petites populations, avec surreprésentation de l'un des allèles et sous-représentation des autres **(fig. 3)**

1.3 La sélection

Charles Darwin fut le premier des naturalistes à concevoir la sélection naturelle comme une force active. Les individus les mieux adaptés à des contraintes seront plus à même de survivre et de se reproduire. Pour que la sélection agisse, il est nécessaire que les individus d'une population portent des traits différents, que cette variation affecte la capacité à survivre ou à se reproduire. Les individus portant des mutations avantageuses auront plus de descendants, la fréquence de cette mutation dans la population va augmenter. La sélection n'agit que sur certains locus, et sur leur voisinage direct [Fay and Wu, 2000]. Il existe différents types de sélection qui engendrent différentes signatures moléculaires **(fig. 4)**. La sélection dite purificatrice ou négative diminue la fréquence de mutations défavorables dans un environnement donné. La sélection équilibrante permet à plusieurs allèles de coexister à un locus donné, s'ils sont avantageux individuellement ou ensemble, ce type de sélection favorise le maintien de la diversité dans une population. Le maintien du polymorphisme peut être nécessaire à la survie dans un environnement présentant une hétérogénéité spatiale ou temporelle [Nielsen, 2005]. La sélection positive va augmenter en fréquence la mutation avantageuse dans la population, et accroître l'adaptation des individus dans un environnement. Une forte sélection positive peut conduire à un balayage sélectif qui entraîne une réduction de la diversité autour de l'allèle sélectionné, ce phénomène s'appelle l'auto-stop génétique (*hitchhiking*) des variants neutres avoisinants qui sont en déséquilibre de liaison avec l'allèle sélectionné [Charlesworth, 2007].

1.4 Le flux génique

Le flux génique est l'échange de portions de génome ou d'allèles entre les populations, ou au sein d'une même population. La migration est parfois utilisée comme synonyme de flux de gènes. A long terme, lorsque les flux de gènes sont réciproques d'une population à une autre, les fréquences alléliques entre populations

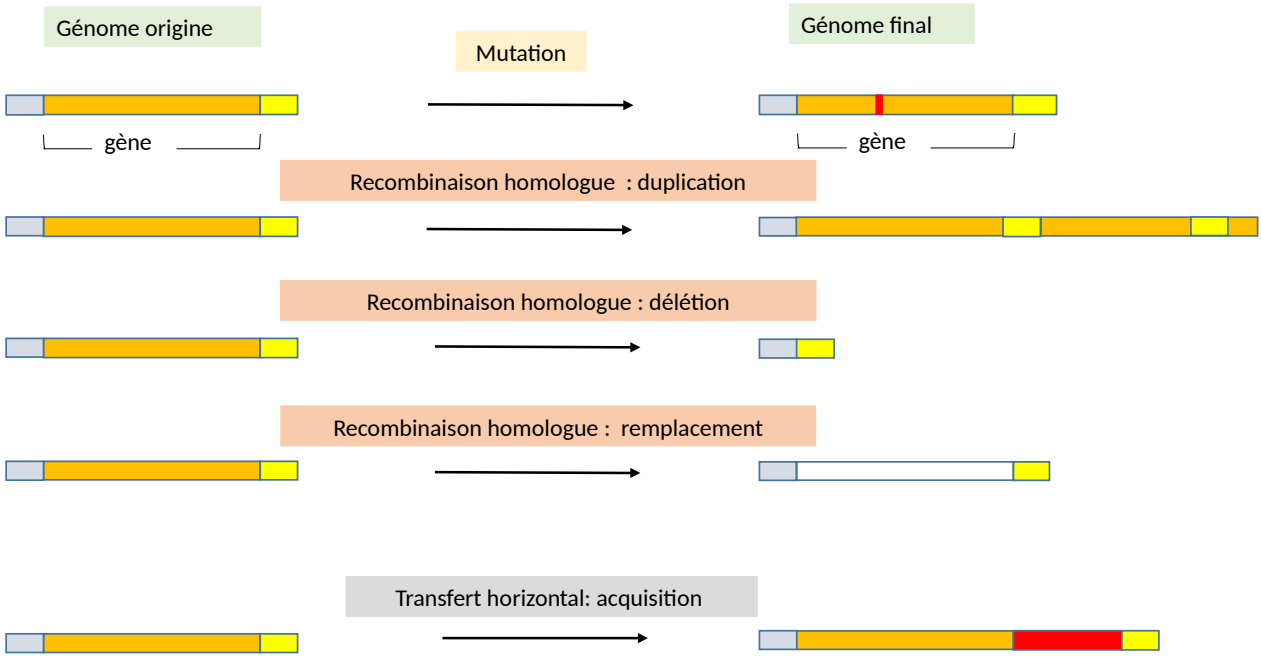


FIGURE 5 – Les mécanismes moléculaires.
Les mécanismes moléculaires apportant du polymorphisme génétique chez les bactéries [Bartoli et al., 2016]

s'équilibrent [Slarkin, 1985] induisant une réduction de la différenciation génétique. L'homogénéisation est d'autant plus forte que les flux de gènes sont importants.

En l'absence de reproduction sexuée de type eucaryote, trois mécanismes permettent à la bactérie d'intégrer de l'ADN étranger : la transformation (l'acquisition de fragments d'ADN libre dans le milieu, par des bactéries compétentes), la conjugaison (elle consiste en une transmission de plasmides conjugatifs d'une bactérie donneuse à une bactérie receveuse), et la transduction (transfert grâce à un bactériophage) [Whittam, TS. and Ake, SE., 1993][Ochman et al., 2000][Errington et al., 2001]. Comme expliqué dans les paragraphes suivants, ces mécanismes aboutissent au remplacement d'une séquence homologue (recombinaison homologue) ou à l'ajout d'un fragment d'ADN exogène (HGT) **(fig. 5)**.

1.4.1 La recombinaison homologue

La recombinaison homologue est le remplacement d'une séquence par sa séquence correspondante provenant d'un autre gènome. La recombinaison homologue est à l'origine de réarrangements génomiques (délétions, duplications, inversions). Ces réarrangements peuvent profondément affecter l'expression des gènes, pouvant conduire à la perte complète d'une fonction quand le réarrangement se produit dans le cadre de lecture. La recombinaison homologue implique un appariement entre séquences homologues, par exemple des transposons, des séquences d'insertions (IS), des prophages [Brussow et al., 2004]. Elle est donc dépendante d'une similarité de séquences.

La prévention de la recombinaison entre des séquences divergentes est assurée par certains gènes comme *mutS* ou *mutL* (impliquées dans le système de réparation des mésappariements (SRM)). Leur inactivation augmente la fréquence de recombinaison interspécifique d'un facteur 1000, elle autorise la recombinaison entre espèces voisines, *Salmonella typhimurium* et *Escherichia coli* dont les séquences nucléiques divergent de 15% [Rayssiguier et al., 1989]. L'étude de l'influence de la divergence sur la recombinaison entre des séquences longues de 400 pb chez *E. coli*, a montré que la fréquence de recombinaison est diminuée de 240 fois lorsque la similitude entre les séquences décroît de 10%, alors que cette fréquence n'est affectée que d'un facteur 9 dans une souche *mutS* déficiente [Shen and Huang, 1989]. Plus

les séquences sont proches plus la probabilité qu'un événement de recombinaison se produise est élevée. Ainsi, chez *E. coli* la recombinaison homologue se produit plus fréquemment à l'intérieur d'un clade qu'entres clades, ce qui appuie l'hypothèse que la recombinaison homologue serait une force cohésive [Didelot et al., 2012]. Ce mécanisme induit des effets significatifs sur le phénotype des bactéries, comme des acquisitions de traits associés à la pathogénie et à la virulence [García-Solache et al., 2016].

1.4.2 La recombinaison non-homologue ou transfert horizontal de gènes (HGT)

La recombinaison non-homologue introduit du matériel génétique nouveau, elle est aussi appelée transfert horizontal de gènes [Didelot et al., 2012]. Les bactéries qui vont partager un même environnement, réservoir, ou hôte seront plus à même d'échanger ou d'acquérir du matériel génétique. C'est l'un des processus d'évolution les plus puissants car il peut immédiatement changer le phénotype et introduire la possibilité de coloniser un nouvel habitat. Chez les *Xanthomonas*, les régions acquises par HGT sont principalement des îlots de pathogénie contenant des gènes de virulence ou des gènes de régulation [Lima et al., 2008]. Le HGT se produit plus fréquemment lorsqu'il y a une similitude de séquence ou de protéome entre le donneur et le receveur, ce qui implique que la divergence entre les séquences est une barrière au HGT chez les procaryotes [Popa and Dagan, 2011]. Ces auteurs précisent cependant que cette barrière n'est pas insurmontable, puisqu'il peut se produire des HGT inter génériques, par exemple entre les *Gammaproteobacteria* et les *BetaproteobacteriA*. Un système de réparation de l'ADN (les protéines de jonction d'extrémité non-homologues, NHEJ) jouerait un rôle dans le contournement de cette barrière. En effet, les génomes ayant reçu des gènes très divergents contiennent plus fréquemment des NHEJ, que les génomes qui n'en contiennent pas.

Les polymorphismes génétiques chez les bactéries proviennent donc de trois mécanismes moléculaires : la mutation, la recombinaison homologue, et le transfert horizontal de gènes (HGT) [Bartoli et al., 2016]. La persistance de ces polymorphismes dépendra, elle, de la dérive et de la sélection. Ces trois mécanismes peuvent

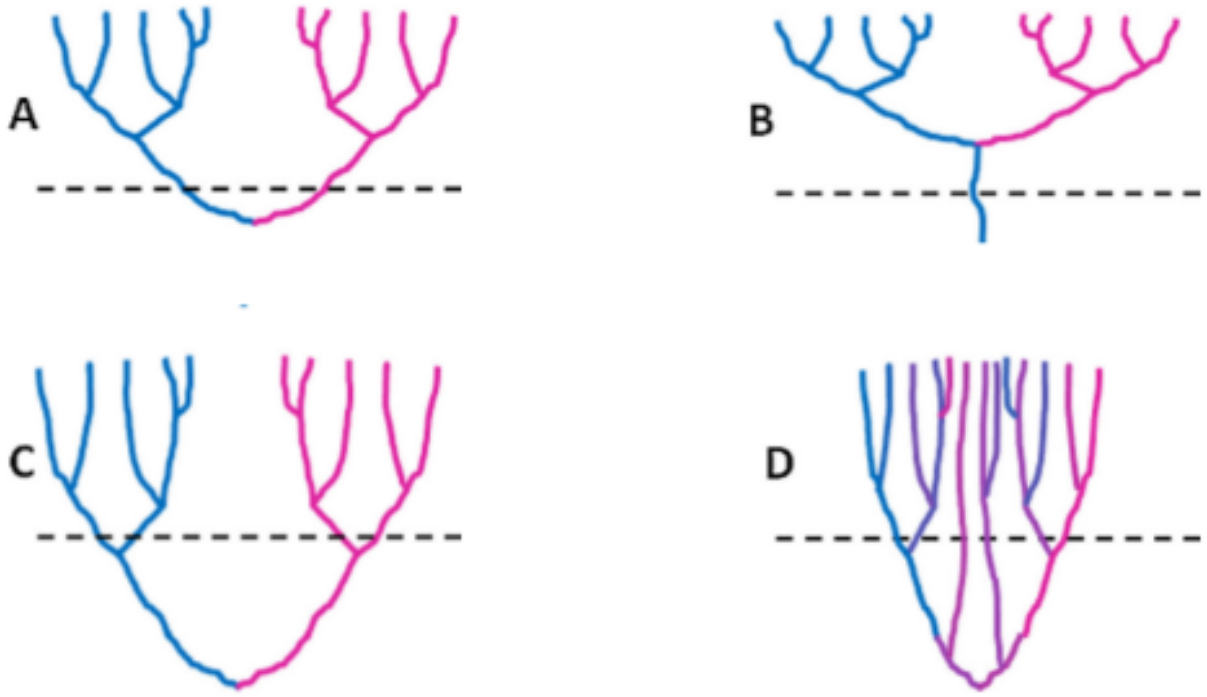


FIGURE 6 – Les seuils de délimitation d'espèces.
Les seuils sont nécessaires mais pas suffisants pour délimiter les espèces. Les seuils basés sur des valeurs d'ANI par exemple sont indiqués en pointillé pour quatre situations hypothétiques. Les lignées avec des propriétés biologiques différentes sont indiquées par deux couleurs. Dans la situation A, le seuil permet de discriminer correctement les deux lignées. Le seuil dans la situation B ne permet pas de discriminer les 2 lignées, il les regroupe de façon inappropriée. Dans C, le seuil identifie 4 lignées ayant des propriétés biologiques identiques deux à deux. Dans D, les propriétés biologiques des souches ne sont pas distribuées selon la phylogénie et il serait inapproprié de différencier des lignées en appliquant un seuil. [Whitman, 2015]

provoquer une altération du phénotype bactérien soit bénéfique, par exemple en permettant l'adaptation à un nouvel environnement, soit délétère, en diminuant la fitness de la bactérie, soit neutre, c'est à dire sans effet sur la fitness bactérienne.

Comme mentionné plus haut, la divergence génétique via l'accumulation de différences génétiques entre populations, peut aboutir à un isolement génétique au moins partiel, c'est à dire à une réduction du flux génique voire sa suppression. L'évolution de cet isolement génétique peut déboucher sur l'apparition de nouvelles espèces bactériennes.

2 La spéciation chez les bactéries

Le concept d'espèce est très discuté chez les procaryotes [Krause and Whitaker, 2015]. L'espèce bactérienne est définie grâce à des similitudes génotypiques qui peuvent être complétées par des critères phénotypiques [Stackebrandt et al., 2002]. Différentes méthodes ont été utilisées pour mesurer cette similitude génétique :

— le pourcentage d'hybridation ADN-ADN pour lequel on assigne deux isolats à la même espèce lorsque le pourcentage est supérieur ou égal à 70%. Cette mesure peut être complétée par la mesure de la stabilité thermique des hybrides (ΔTm inférieur à $5°C$) [Wayne et al., 1987].

— La mesure de l'identité nucléotidique moyenne (ANI) possible aujourd'hui grâce au nombre croissant de génomes disponibles dans les bases de données connaît un fort développement. Une valeur d'ANI supérieure à 95% correspond au seuil de 70% d'hybridation ADN-ADN [Konstantinidis and Tiedje, 2005].

Bien qu'il soit nécessaire d'appliquer des seuils en taxonomie, une certaine flexibilité est nécessaire selon les situations **(fig. 6)** [Whitman, 2015]. D'autres facteurs, comme la physiologie, l'écologie, les flux de gènes doivent être considérés [Stackebrandt et al., 2002].

Les connaissances théoriques sur les effets des processus biologiques permettant la cohésion dans l'espèce, et la divergence entre espèces sont encore limitées [Gevers et al., 2006]. Le concept de sélection périodique des écotypes a été proposé par Cohan en 2001, [Achtman and Wagner, 2008] **(fig. 7a)**. L'intérêt de ce concept est de prendre en compte l'écologie des bactéries pour expliquer le processus de

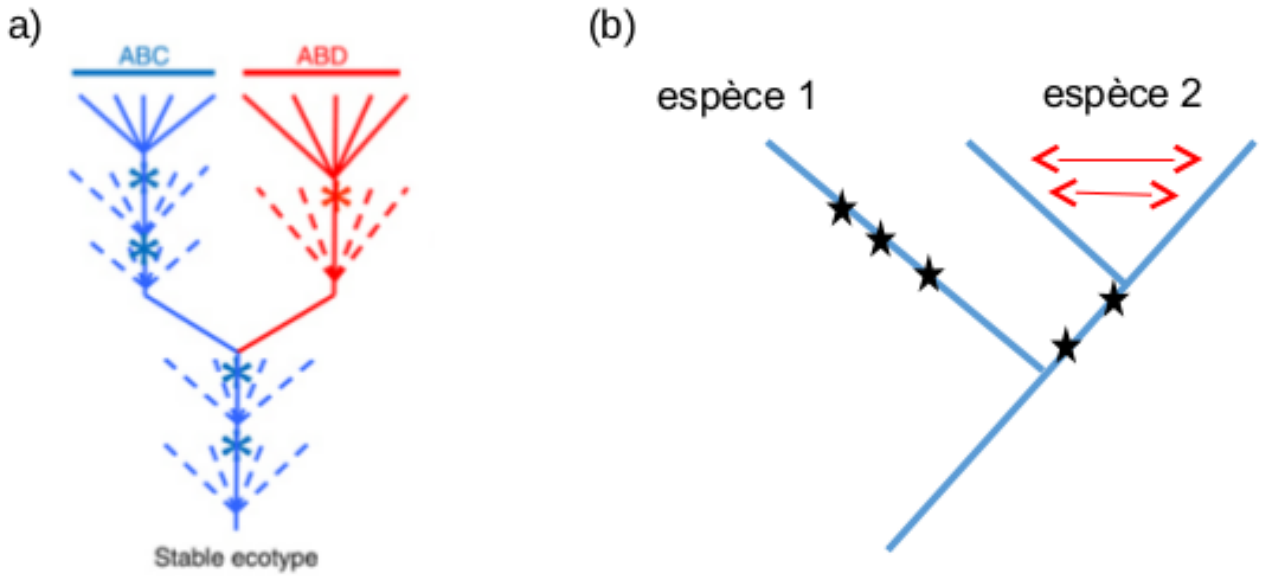


FIGURE 7 – Concepts d'espèce.
a) La sélection périodique des écotypes. Dans le modèle de l'écotype stable, les écotypes sont créés et éteints à faible fréquence. Chaque écotype E1, E2, subit une série d'événements de sélection périodiques (indiqués par des astérisques) pendant l'histoire de la divergence. La sélection est à l'origine de la cohésion et de l'homogénéité à l'intérieur d'un écotype. A chaque événement de sélection périodique, un mutant acquiert une mutation adaptative et supplante tous les autres individus du clade, la sélection purge toute la diversité présente. Les descendants du mutant survivant divergent (indiqués par des lignes pointillés), mais à la sélection périodique suivante à nouveau un seul un mutant survit [Cohan, 2006]. b) Le Biological Species Concept (BSC). L'accumulation de polymorphisme par exemple (étoiles noires), ou un isolement écologique entraînent des barrières et une réduction du flux de gènes entre l'espèce 1 et 2 [Seehausen et al., 2014]

spéciation. On appelle alors écotype une population asexuée adaptée à une niche écologique et dont la diversité est régulièrement purgée par la sélection naturelle. L'isolement génétique et écologique entre les écotypes ainsi que la cohésion phylogénétique leur confèrent alors le statut d'espèces. Le principal écueil de ce concept est qu'il néglige le flux de gènes qui peut introduire des allèles n'étant pas sous sélection et supprimer la cohésion d'un écotype [Doolittle and Zhaxybayeva, 2009].

Beaucoup d'autres concepts existent **(tab. I)**, certains s'appliquant plus aux eucaryotes [De Queiroz, 2007]. Un des concepts d'espèces le plus communément accepté est le *Biological Species Concept* (BSC) **(fig. 7b)** proposé par Mayr [Mayr E, 1942]. Mayr propose qu'une espèce biologique soit effectivement ou potentiellement interféconde, et génétiquement isolée d'autres groupes similaires. À cette définition, il a ensuite été ajouté que cette espèce doit pouvoir engendrer une progéniture viable et féconde. Ainsi, l'espèce est la plus grande unité de population au sein de laquelle le flux de gènes est possible dans des conditions naturelles, les individus d'une même espèce étant génétiquement isolés d'autres ensembles équivalents du point de vue reproductif. Bien que cette définition ait été développée pour les eucaryotes, il semble intéressant de pouvoir la modifier suffisamment pour l'appliquer aux procaryotes en considérant la recombinaison, même rare, comme de la reproduction entre les souches. L'écueil principal de l'adaptation du concept biologique d'espèce aux procaryotes est sans aucun doute une prédominance forte de la clonalité et la rareté des évènements de recombinaison.

L'application des concepts de génétique des populations aux bactéries peut nous permettre de mieux appréhender l'évolution de ces organismes. Notamment, il est important de comprendre comment des bactéries pathogènes de plantes cultivées ont pu s'adapter et se diversifier aussi rapidement sur des hôtes dont l'apparition remonte à quelques siècles (si on considère les variations intra-spécifiques liées à la sélection), voire à quelques milliers d'années (pour les variations liées à la domestication).

TABLE I – Les choix de concepts d'espèces
Principaux concepts d'espèces et les propriétés qui les distinguent. D'après [De Queiroz, 2007]

Concepts d'espèce	Propriétés	Sources
Biologique	Croisement possible et isolement reproductif intrinsèque	[Dobzhansky, 1950] [Mayr, 1942]
Écologique	Même niche ou zone adaptative(Partage les mêmes composants de l'environnement)	[Valen, 1976] [Andersson, 1990]
Phylogénétique	Groupe monophylétique partageant des caractères moléculaires ou morphologiques hérités d'un ancêtre commun direct	[Donoghue, 1985]
Généalogique	Coalescence exclusive des allèles(Tous les allèles d'un gène descendent d'un allèle ancêtre commun qui n'est pas partagé par d'autres espèces)	[Baum and Shaw, 1995] [Avise and Ball, 1990]
Morphologique	Population morphologiquement distincte d'une autre par des critères diagnostiques	[Cronquist, 1978]

3 Les mécanismes de pathogénie chez les bactéries phythopathogènes

Le cycle infectieux d'une bactérie à Gram-négatif épiphyte colonisant les parties aériennes des plantes, comme dans le cas des *Xanthomonas*, répond à des étapes successives : la phase d'attraction, la phase d'installation et la phase d'invasion. Lors de la phase d'attraction, le chimiotactisme permet aux bactéries phytopathogènes d'identifier l'hôte en répondant aux signaux chimiques émis par la plante, et d'établir une relation étroite avec l'hôte. L'adhésion à la surface est un prérequis à la formation d'un biofilm, qui est composé de microorganismes agrégés dans une matrice d'exopolysaccharides et attachés à une surface [Costerton et al., 1995]. Le mécanisme d'adhérence fonctionne grâce à des protéines de surface, les adhésines, qui reconnaissent des récepteurs spécifiques au niveau des tissus colonisés [Mhedbi-Hajri et al., 2011]. Certaines bactéries pénètrent dans la plante par des blessures ou des ouvertures naturelles, comme les stomates. Par la suite, lors de la phase d'invasion, ces bactéries se multiplient en causant des dégâts aux cellules de l'hôte. C'est dans cette phase que la survie de la bactérie dépend de sa capacité de contournement aux mécanismes de défenses de l'hôte, condition indispensable à l'établissement de la phase d'infection. Le contournement des défenses de la plante s'effectue grâce à un répertoire de fonctions qui sont régulées en réponse aux signaux environnementaux rencontrés sur l'hôte [Hajri et al., 2009][Wilson et al., 2002]. Les facteurs de virulence sont exprimés au cours de la phase de multiplication de la bactérie. Chez les bactéries à Gram négatif, plusieurs systèmes hautement spécialisés sécrètent des protéines impliquées dans l'acquisition de nutriments, la compétition et la pathogénèse. Parmi ceux-ci, le Système de Sécrétion de Type III (T3SS), une structure de surface très répandue chez les bactéries pathogènes, permet d'injecter des protéines à l'intérieur des cellules hôtes. Ce système de sécrétion est nécessaire au développement de la maladie sur les plantes hôtes. Les protéines effectrices ou effecteurs de type III (T3SE) sont des protéines transférées dans la cellule de l'hôte grâce au T3SS et interférant avec certaines fonctions de la cellule eucaryote. D'autres facteurs de virulence tels que les enzymes pectinolytiques, les toxines ou encore les effecteurs des autres systèmes de sécrétion jouent aussi un

rôle déterminant dans l'infection. Ainsi, la spécificité d'hôte devrait être déterminée par un répertoire de gènes codant pour des facteurs de virulence spécifique aux bactéries infectant un certain type d'hôte [Sarkar et al., 2006]. Toutefois, un tel répertoire se définit non seulement par la présence de certaines protéines effectrices, mais aussi par l'absence d'autres, pouvant par exemple être reconnues par l'hôte [Hajri et al., 2009].

4 Impact des facteurs environnementaux sur l'émergence des maladies

L'émergence de nouvelles maladies peut être la conséquencede l'augmentation en incidence d'un agent pathogène, de l'élargissement de sa gamme d'hôte, ou de sa répartition géographique [Anderson et al., 2004]. Cette émergence peut être liée à l'acquisition d'un nouveau trait comme un gène de résistance par exemple, ou liée à des facteurs environnementaux comme détaillé ci-dessous [Engering et al., 2013].

4.1 La domestication

La domestication des plantes, a commencé dans les zones tropicales et subtropicales entre 10000 à 7000 ans de Cal BP [4] [Gupta, 2004]. La domestication a provoqué des changements radicaux tant dans la diversité génétique des plantes cultivées, que dans leur expansion et leur densité dans les agroécosystèmes [Stukenbrock and McDonald, 2008]. Dans la plupart des pays en voie de développement la Révolution Verte qui a commencé vers 1960, a entraîné la diminution de la diversité génétique avec dans le même temps un accroissement de la taille des zones cultivées rendant l'agriculture plus vulnérable à la maladie [Pingali, 2012]. L'évolution des pratiques agricoles, comme la rotation des cultures, les produits phytosanitaires, l'irrigation, et toute modification de l'hôte vont avoir un impact direct sur les populations d'agents pathogènes.

4. La mention « avant le présent calibrées» ou « Cal BP » s'applique généralement aux dates exprimées en nombre d'années comptées vers le passé à partir de l'année 1950, calibrées par datation à partir du carbone 14.

4.2 Les échanges commerciaux

Depuis 1492, suite à la découverte de l'Amérique par Christophe Colomb, l'échange Colombien a marqué le début des transferts de plantes cultivées, et des micro-organismes associés, entre l'ancien et le nouveau Monde [Nunn and Qian, 2010]. L'intensification de ces échanges a permis l'introduction de nouvelles espèces et l'émergence de maladies par colonisation d'une nouvelle aire géographique, ou d'un nouvel hôte.

4.3 Les changements climatiques mondiaux

Les changements climatiques mondiaux observés actuellement pourraientt apporter des conditions favorables au développement et à la dispersion des maladies (survie, nombre de cycles par an, humidité, variations de température) [Garrett et al., 2006]. Par conséquent, le nombre d'épidémies risque de s'accroître dans les prochaines décennies [Evans et al., 2008]. Par exemple, le stress induit par la sécheresse et celui de l'infection par *Xylella fastidiosa* ont des effets additifs sur les plantes, en aggravant les symptômes et leurs progressions le long de la tige [McElrone et al., 2003]. Le réchauffement climatique peut avoir un effet sur l'aire de distribution des espèces, comme cela a été étudié chez un oomycète pathogène forestier *Phytophthora cinnamomi* où selon des prédictions théoriques l'aire de répartition s'élargirait de 1 à 100 km en un peu plus d'un siècle suite à un réchauffement d'environ $1,8°C$ pour la température moyenne pendant la période d'hiver [Bergot et al., 2004].

5 Les bactéries du genre *Xanthomonas*

Le genre *Xanthomonas* appartient au phylum des *Proteobacteria*, classe des *Gammaproteobacteria* et contient environ 27 espèces [Ryan et al., 2011]. Ce sont des bactéries Gram négatives provoquant d'importantes maladies sur plus de 400 plantes hôtes, dont certaines présentent un intérêt économique majeur (ex : le riz, les agrumes, la canne à sucre, le haricot...) [Hayward, 1993]. Les bactéries du genre *Xanthomonas* peuvent infecter beaucoup d'espèces cultivées dont par

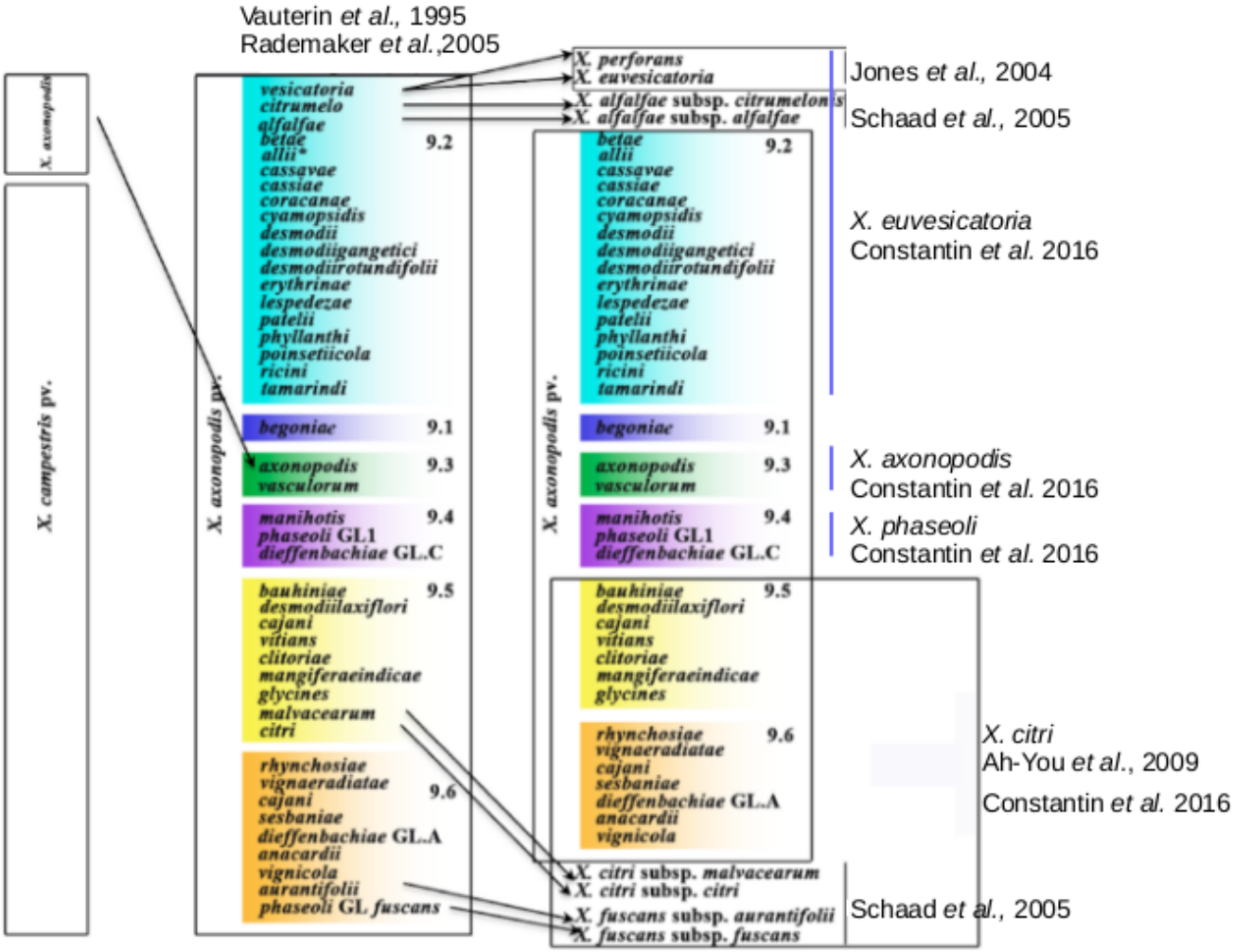


FIGURE 8 – Remaniements taxonomiques au sein de l'espèce *X. axonopodis*.
Remaniements taxonomiques au sein de l'espèce *X. axonopodis*. D'après [Mhedbi-Hajri, 2010]

exemple, des *Solanaceae*, des *Brassicaceae*, des céréales en provoquant des taches, chancres, chloroses et nécroses sur feuilles et fruits. Chaque souche a généralement une gamme d'hôte étroite. Au sein d'une espèce, les souches provoquant le même type de symptômes sur une même gamme d'hôte sont regroupées en pathovars [Dye et al., 1980a]. Les *Xanthomonas* spp. ont été traditionnellement décrits comme des bactéries associées aux plantes qui ne sont pas retrouvées dans d'autres environnements [Hayward, 1993]. Ces bactéries infectent aussi bien des plantes pérennes (ex : Agrumes) que des plantes annuelles (ex : Haricot). La plupart des *Xanthomonas* sont initialement épiphytes (vie à la surface des feuilles), et peuvent entrer dans leur hôte par les ouvertures naturelles (hydathodes [5] ou par des blessures). Ils peuvent ensuite infecter toute la plante grâce au système vasculaire (*Xanthomonas phaseoli* pv. *manihotis*), ou coloniser le parenchyme (*Xanthomonas citri* pv. *citri*).

5.1 Taxonomie et phylogénie du complexe d'espèces *Xanthomonas axonopodis*

La diversité pathologique a été à l'origine du concept «new host-new species» pour lequel chaque variant montrant une gamme d'hôte différente ou produisant des symptômes particuliers était classé dans une nouvelle espèce [Starr, 1981]. En 1974, seules cinq espèces du genre *Xanthomonas* ont été retenues sur la base des caractéristiques phénotypiques, *X. campestris* (espèce type), *X. fragariae*, *X. axonopodis*, *X. albilineans* et *X. ampelina* [Dye et al., 1980b]. L'espèce *X. axonopodis* a été redéfinie sur la base d'une taxonomie polyphasique (taxonomie qui prend en compte un maximum de données, génétiques, phénotypiques, écologiques..) et d'hybridations ADN-ADN par Vauterin et al. (1995)[VAUTERIN et al., 1995]. Par la suite, six groupes au sein de *X. axonopodis* (nommés de 9.1 à 9.6) **(fig. 8)** ont été décrits sur la base de rep-PCR (repetitive extragenic palindromic-PCR) [Rademaker et al., 2005]. Plus récemment, certains pathovars de *X. axonopodis*, pv. *alfalfae* et pvs. *citrumelonis* du groupe 9.2, *citri* et *malvacearum* du groupe 9.5, et *fuscans* et *aurantifolii* du groupe 9.6, ont été reclassés respectivement dans les nouvelles

5. Les hydathodes sont des tissus sécréteurs qui rejettent l'eau issue d'un parenchyme aquifère ou d'un vaisseau de xylème par des orifices aménagés entre des cellules épidermiques foliaires. Ces pores opèrent notamment lors du phénomène de guttation par exsudation de gouttelettes.

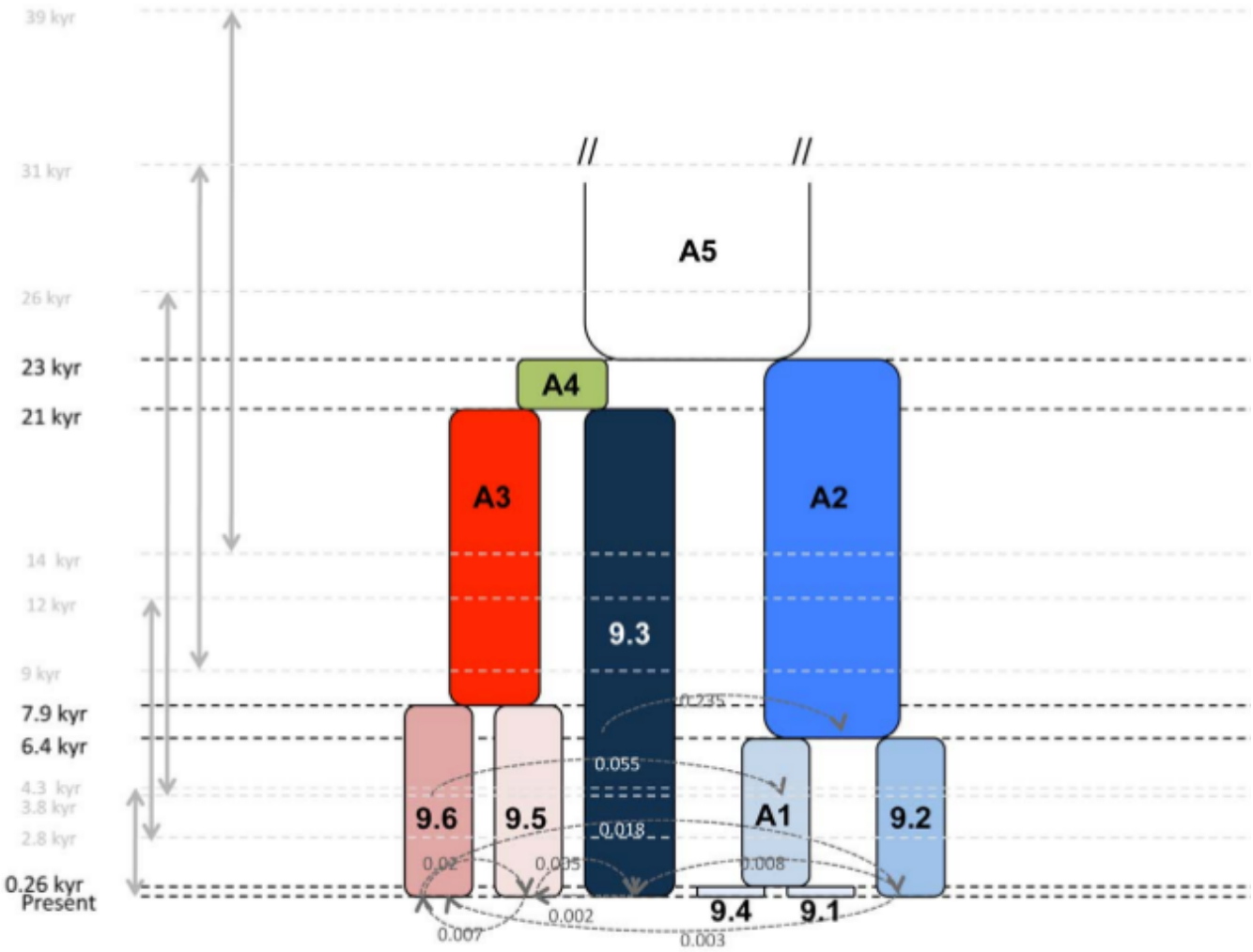


FIGURE 9 – Histoire évolutive des *Xanthomonas axonopodis*.
Histoire évolutive en IMa2 de 131 souches de *X. axonopodis* basée sur 7 gènes de ménage. Les temps de divergence sont indiqués à droite en Kyr (milliers d'années). Les flèches grises pointillées indiquent le flux de gènes et les chiffres le nombre effectif de migrants. Les groupes A1 à A5 représentent les populations ancestrales [Mhedbi-Hajri et al., 2013]

espèces, *X. alfalfae*, *X. citri*, et *X. fuscans* respectivements [Schaad et al., 2005a] [Schaad et al., 2006]. Jones et al. [Jones et al., 2004] ont proposé de diviser *X. axonopodis* pv. *vesicatoria* en deux espèces *X. euvesicatoria* et *X. perforans* sur la base d'hybridation ADN-ADN. Sur la base des valeurs d'ANI, Barak et al. (2016) proposent de regrouper les espèces *X. euvesicatoria* et *X. perforans* en une seule espèce. Ah-You et al. (2009) ont suggéré que *X. citri* et *X. fuscans* sont des espèces synonymes sur la base d'une taxonomie polyphasique et d'une MLSA (Multilocus Sequence Analysis). Récemment, une proposition visant à remanier et élever la plupart de ces groupes définis au sein de l'espèce *X. axonopodis* au rang d'espèce a été publiée [Constantin et al., 2016]. Ces nouvelles propositions ont été récemment validées du point de vue du Code international de la nomenclature. Nous les avons utilisées pour la taxonomie de nos souches dans l' **Annexe B**. Ainsi, le groupe 9.4 forme l'espèce *X. phaseoli*, le groupe 9.2, l'espèce *X. euvesicatoria*, le groupe 9.3, l'espèce *X. axonopodis* et les groupes 9.5 et 9.6 forment l'espèce *X. citri* **(fig. 8)**.

6 Histoire évolutive du complexe *X. axonopodis*

Les travaux de Mhedbi-Hajri et al. (2013) basés sur l'analyse de séquences partielles de sept gènes de ménage et des approches de coalescence et génétique des populations ont montré que les bactéries du complexe d'espèces *Xanthomonas axonopodis* auraient connu une première phase de diversification en 5 groupes (9.5, 9.6, 9.3, 9.2, et A1) indépendamment de l'hôte et de la géographie **(fig. 9)**. Des souches pathogènes sur *Citrus* spp. sont retrouvées dans les groupes 9.5 et 9.6 qui auraient divergé il y a environ 8000 ans, et aussi dans le groupe 9.2 qui aurait divergé de son ancêtre commun avec les groupes 9.6 et 9.5 il y a 25000 ans **(fig. 8)**. A l'intérieur des 5 groupes, il existe des clusters monophylétiques regroupant des souches pathogènes d'un même hôte. Cependant certains agents pathogènes attaquant le même hôte ne sont pas monophylétiques, c'est le cas des pathovars pathogènes sur haricot. Trois lignées du pathovar *fuscans* font partie du groupe 9.6 (*X. citri* pv. *fuscans* lignée 2 (LG2), *X. citri* pv. *fuscans* lignée 3 (LG3), *X. c.* pv. *fuscans* lignée *fuscans*) et le pathovar *phaseoli* fait partie du groupe A1 (*X. p.* pv. *phaseoli* lignée 1, LG1). La divergence en pathovars se serait produite durant les deux derniers siècles avec le développement de la monoculture et l'expansion de

l'agriculture. Une troisième étape impliquant des contacts secondaires et un fort flux de gènes entre les groupes se serait produite à la faveur du développement de l'agriculture et de la mondialisation des échanges. Il semblerait que ce complexe d'espèces évoluerait 3 fois plus par recombinaison que par mutation (r/m=3,18, r/m étant la mesure l'impact de la recombinaison r par rapport à la mutation m).

7 Les questions de recherche, objectifs

L'objectif principal de ce stage était d'inférer l'histoire évolutive du complexe d'espèces *X. axonopodis* à partir de données génomiques. Les bactéries du genre *Xanthomonas* sont des agents pathogènes majeurs sur une grande variété de plantes d'intérêt agronomique et économique. Étonnamment certains pathovars appartenant à des groupes différents sont pathogènes sur les mêmes hôtes, comme des souches de *X. citri* pv. *citri* (groupe 9.5) et des souches de *X. citri* pv. *aurantifollii* (groupe 9.6) sur agrumes, ou *X. citri* pv. *fuscans* (groupe 9.6) et *X. phaseoli* pv. *phaseoli* (groupe 9.4) sur haricot. Les questions de recherche posées étaient les suivantes :

Retrouve-t-on avec des données génomiques la structuration en cinq groupes décrite par Mhedbi-Hajri et al. (2013) sur la base de sept gènes de ménages ? Quelles forces évolutives ont conduit à cette différenciation en groupes ? Quels sont les composantes adaptatives, les fonctions biologiques qui auraient pu induire la divergence dans le complexe d'espèces *Xanthomonas axonopodis* ? Quelles barrières conduiraient de nouvelles souches à former une nouvelles espèce ? Y-a-t-il un mécanisme de spéciation pouvant s'appliquer à toutes les bactéries, indépendamment de l'isolement par diminution du taux de recombinaison connu pour être extrêmement variable ? La faculté à infecter le même hôte en appartenant à des groupes différents serait-elle due à une adaptation indépendante ou correspondrait-elle à un transfert de caractères entres les groupes via le flux de gènes ?

Pour répondre à ces questions, j'ai utilisé 73 génomes représentant la diversité de ce complexe. Dans une première partie, mon objectif était d'identifier les forces évolutives qui ont participé à la différentiation de ce complexe d'espèces. Pour cela, la structure des populations a été caractérisée, et les forces évolutives ont été analysées (l'intensité de la recombinaison homologue à été comparée à celle

de la mutation). Ensuite, l'impact des flux de gènes tout au long de l'histoire évolutive (principalement au moment de la divergence en groupe, et lors de la spécialisation en pathovars) a été analysé. En me focalisant sur deux groupes de populations, j'ai tenté d'inférer l'histoire démographique associée à l'origine du premier niveau de différenciation en groupes en testant une grande diversité de scénarios démographiques. Dans une deuxième partie, le génome accessoire a été étudié avec comme objectif d'identifier son potentiel adaptatif. J'ai caractérisé le transfert horizontal de gènes (HGT) le long de l'histoire évolutive en inférant les événements de gains et de pertes aux nœuds d'un arbre phylogénétique. Est-ce que le développement des échanges et de l'agriculture a provoqué une intensification récente des flux de gènes qui serait responsable de la convergence pathologique de certains pathovars ? L'étude des fonctions biologiques plus particulièrement acquises ou échangées, permettrait d'essayer d'y répondre.

Enfin, ces résultats ont été discutés dans leur globalité et en essayant de montrer en quoi ils pourraient contribuer à une meilleure compréhension des émergences d'agents pathogènes chez les *Xanthomonas*.

Troisième partie

CHAPITRE I

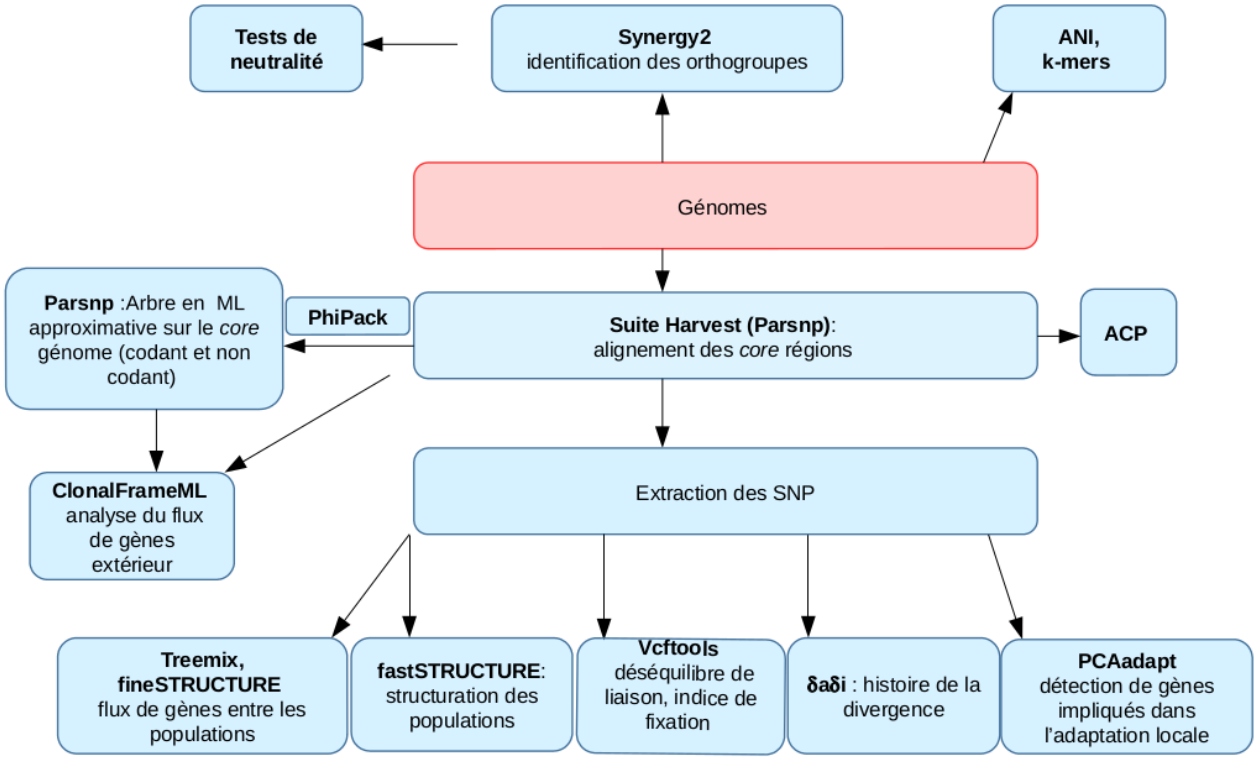


FIGURE 11 – Schéma synoptique du traitement des données.
Les flèches indiquent les données d'entrées des différents outils utilisés dans le Chapitre I

1 Introduction

La taxonomie du complexe d'espèces *Xanthomonas axonopodis* est en constante évolution. Six groupes (nommés 9.1 à 9.6) ont été précédemment décrits sur la base de rep-PCR [Rademaker et al., 2005]. La génétique des populations basée sur sept gènes de ménage a montré que le complexe d'espèces *Xanthomonas axonopodis* serait composé de cinq groupes génétiques [Mhedbi-Hajri et al., 2013]. L'objectif principal de ce chapitre 1 est de caractériser les forces évolutives. Pour cela, la structure de population de ce complexe a été analysée à partir d'une collection de 73 génomes. Nous avons identifié cinq groupes (que nous appellerons 9.3, 9.2, 9.41, 9.5 et 9.6) parmi lesquels les groupes 9.1 et 9.4 correspondent à l'espèce *X. phaseoli*, le groupe 9.2 à l'espèce *X. euvesicatoria*, le groupe 9.3 à l'espèce *X. axonopodis* et les groupes 9.5 et 9.6 à l'espèce *X. citri*. Afin de déterminer le rôle que peuvent avoir les flux de gènes dans la divergence et l'apparition de nouvelles maladies, nous avons étudié la recombinaison homologue d'origine extérieure et la recombinaison homologue entre les souches de notre collection. Ensuite, une analyse des scénarios de divergence a permis de valider le sens des flux de gènes et a également montré que de façon surprenante ce flux de gènes était plus faible entre des groupes ayant divergé récemment. Au final, afin de détecter d'éventuelles zones génomiques sous sélection positive (*i.e.* adaptation locale) les régions fortement différenciées entre les génomes des souches des groupes 9.5 et 9.6 ainsi que celles ayant de fortes valeurs négatives au D de Tajima [Tajima, 1983] ont été recherchées. Les signatures de sélection positive délimitent les régions du génome qui sont, ou ont été, fonctionnellement importantes dans l'adaptation d'une population à une contrainte environnementale. Ainsi, une zone génomique couplant à la fois un F_{ST} excessivement fort et un D de Tajima excessivement bas (indicateur d'une baisse de diversité) peut être suspectée comme étant sous sélection positive. Les génomes ont aussi été scannés afin de rechercher des SNP du *core genome* montrant une structure excessive (en terme variation de fréquences alléliques, par exemple) entre populations. Ces gènes sont supposés être des marqueurs d'une adaptation locale (voir la revue de [Beaumont, 2005]), et pourraient nous apporter des indications sur les fonctions impliquées dans la spécificité d'hôte.

TABLE II – Nombre de génomes et de pathovars de *Xanthomonas axonopodis*
Nombre de génomes et de pathovars de *Xanthomonas axonopodis* représentés dans notre collection de séquences génomiques.

Groupes	Nbre de séquencesgénomiques	Nbre de pathovars représentés dans le groupe
9.1	2	2
9.2	13	7
9.3	2	2
9.4	13	3
9.5	21	6
9.6	22	5

2 Matériels et méthodes

Le matériel et méthodes est résumé dans un schéma synoptique figure 11.

2.1 Les génomes du complexe d'espèces *Xanthomonas axonopodis*

Un jeu de 73 génomes a été constitué, il est composé de 56 génomes issus de la collection de génomes séquencés du laboratoire et de 17 génomes des banques de données publiques, en choisissant le maximum de diversité géographique, chronologique, et de pathovars par groupe **(Annexe B et tab. II)**. La taille des génomes est en moyenne de 5,09 Mb, et le pourcentage de GC est compris entre 64 et 65,23%. Les génomes de la collection du laboratoire ont été séquencés en Illumina HiSeq avec une couverture de 100X et après assemblage sont composés de 1 à 328 contigs. Quatre souches de *Xanthomonas* isolées de semences de haricot mais non pathogènes sur cet hôte ont été ajoutées à la collection, deux dans le groupe 9.2 (CFBP 7916, CFBP 7920) et deux dans le 9.6 (CFBP 7765, CFBP 7923). Pour les souches pathogènes des agrumes du groupe 9.5 nous avons choisi des souches de *Xanthomonas citri* pv. *citri* du pathotype A (à large gamme d'hôtes parmi les Rutacées) et du pathotype A* (souches à gamme d'hôtes restreinte au limettier Mexicain et espèces proches). La souche de *Xanthomonas vasicola* pv. *holcicola* (CFBP 2543), espèce phylogénétiquement proche du complexe d'espèces *X. axonopodis* a été utilisée comme groupe externe.

2.2 Identification du *core genome*, et extraction des SNP

Déterminer le *core genome*, c'est à dire toutes les régions codantes et non codantes communes aux 73 génomes, est indispensable pour analyser la structure et les flux géniques homologues du complexe d'espèces *Xanthomonas axonopodis*. Nous avons utilisé le programme Parsnp de la suite de logiciels Harvest [Treangen et al., 2014] pour l'identification du *core genome*. Parsnp est recommandé pour analyser des génomes de bonne qualité, il a été conçu pour aligner le *core genome* de 100 jusqu'à 1000 génomes bactériens en quelques minutes. Parsnp est

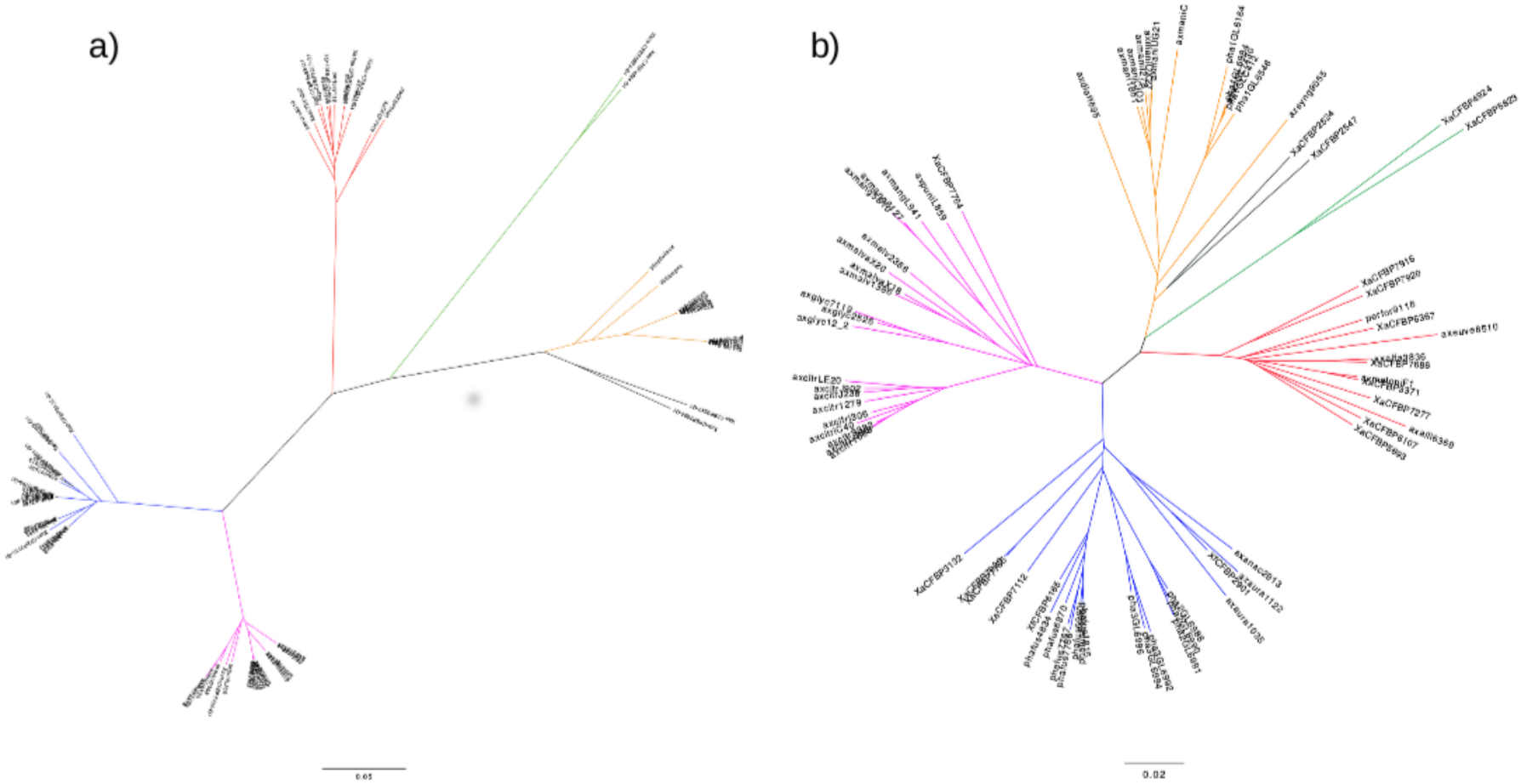


FIGURE 10 – Comparaison d'arbres phylogénétiques non enracinés des 73 génomes.
a) Arbre phylogénétique généré par la suite Harvest en maximum de vraisemblance approximative à partir du *core genome* b) Arbre phylogénétique généré par la méthode CVtree, qui prend en compte les protéomes entiers. Les couleurs indiquent les groupes, 9.1 en noir, 9.2 en rouge, 9.3 en vert, 9.4 en marron, 9.5 en violet et 9.6 en bleu. Les échelles des branches sont en nombre de substitutions par site.

recommandé pour des alignements intra-spécifiques et nécessite des génomes très similaires (≥97% d'ANI), toutefois si tous nos génomes ne valident pas ces conditions, l'alignement produit semble de bonne qualité (les locus semblent bien alignés et le signal phylogénétique est cohérent avec les précédentes études). La validité de cet alignement a été vérifiée en comparant l'arbre obtenu avec Parsnp et une autre phylogénie générée par CVtree (Composition Vector Tree) [Xu and Hao, 2009] (cf. **fig. 10**). CVtree n'utilise pas d'alignement mais génère une matrice de distance à partir du protéome des génomes.

La suite Harvest permet de produire une extraction des SNP et une phylogénie du *core genome*. L'alignement est filtré afin d'enlever les SNP localisés dans des régions identifiées comme recombinantes avec PhiPack [Bruen and Bruen, 2005], afin de permettre la construction d'un l'arbre phylogénétique en utilisant la méthode de vraisemblance approximative implémentée dans FastTree2 [Price et al., 2010]. L'arbre inféré à partir des 73 génomes a été importé dans FIGTREE http ://tree.bio.ed.ac.uk/software/figtree/. L'extraction des SNP (du *core genome* obtenu sans avoir filtré pour la recombinaison afin de permettre l'étude des flux de gènes), est réalisée avec Gingr de la suite Harvest. Grâce à la profondeur du séquençage (>100X), nous avons confiance en l'identification des singletons et aucune autre méthode n'a été employée [Chimenti Michael S., 2016].

Nous avons procédé de la même manière sur différents sous-échantillonnages du jeu de données total, afin d'obtenir des phylogénies et SNP sur les différents groupes, avec ou sans outgroup, en vue d'analyses détaillées ci-dessous, comme celles réalisées avec ClonalFrame et $\delta a \delta$i.

2.3 Annotation des génomes

Afin d'identifier les gènes du *core genome* et d'avoir une annotation homogène, une ré-annotation de tous les génomes a été réalisé avec le pipeline Galaxy [https ://bbric-pipelines.toulouse.inra.fr/galaxy/]. Ce pipeline utilise EUGENE-PP [Sallet et al., 2014] avec une librairie de RNA-Seq (Les protéines bactériennes de Swiss-Prot) et la base de données des protéines de *Xanthomonas. citri* pv. *fuscans* CFBP 4885 (NC_022541.1). Comme les génomes de plus de 100 contigs génèrent des erreurs avec ce pipeline, ils ont été ré-annotés avec Rapid Annotation

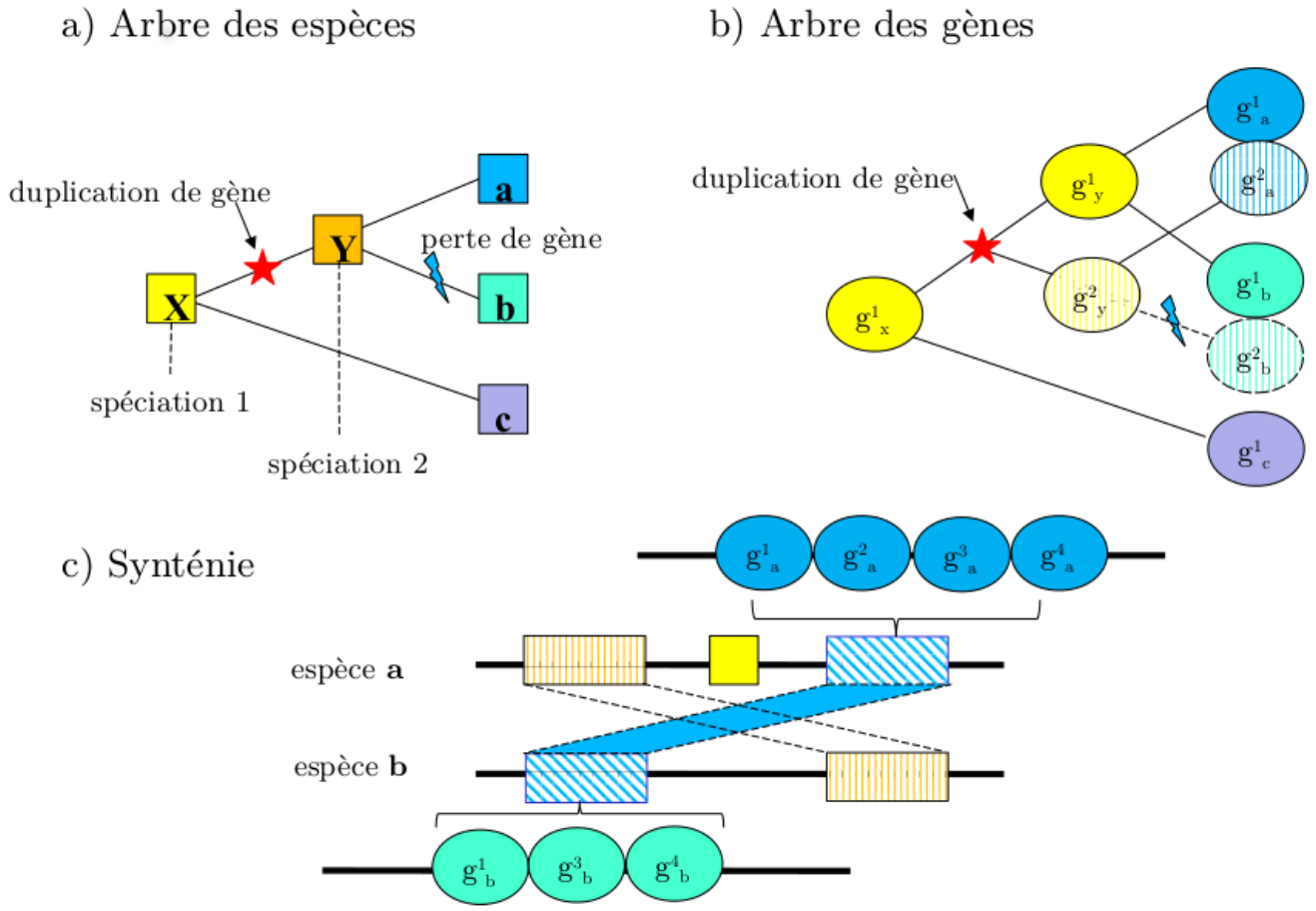


FIGURE 12 – Principe de Synergy2.
L'algorithme divise les gènes en ensembles (orthogroupes) contenant tous les gènes qui descendent d'un unique gène du dernier ancêtre commun de l'espèce. Synergy2 reconstruit l'arbre des gènes pour chaque orthogroupe, en remontant nœud après nœud dans l'arbre des espèces. a) Arbre des espèces, chaque nœud (carré) dans l'arbre représente une espèce. Les événements de spéciation 1 et 2 ont donné les espèces actuelles a, b et c. b) Arbre des gènes décrivant les événements évolutifs pour les gènes g^1, g^2. Chaque nœud (cercle) est un gène. L'arbre montre la descendance d'un gène ancestral g1 en paralogue (cercle hachuré) et orthologue (cercle plein) après la duplication de gène de l'espèce Y. Le gène g^2 a été perdu (éclair bleu) dans l'espèce b. c) Chaque chromosome contient plusieurs blocs synténiques (hachures) constitués de plusieurs gènes. Une région dans l'un des génomes (rectangle jaune) n'a pas de bloc synténique dans l'autre séquence. Le score de similarité synténique pour cette paire de gènes est la fraction des voisins qui sont orthologues aux autres (par exemple le score pour g^3a et g^3b est 2/3)d'après [Wapinski et al., 2007]

with Subsystems Technology (RAST) [Overbeek et al., 2014]. La distribution du nombre et des tailles moyennes des gènes n'a pas montré de différences significatives.

2.4 Identification des orthogroupes : les gènes orthologues, les gènes présents-absents, les gènes en multicopies

Le logiciel Synergy2 [Wapinski et al., 2007] a été utilisé afin de définir les groupes d'orthologues ou "orthogroupes". Trois orthogroupes sont constitués, *(i)* les gènes orthologues, l'ensemble des gènes communs à toutes les 73 souches, *(ii)* les gènes présents-absents, les gènes présents uniquement chez une ou plusieurs souches des 73 génomes, *(iii)* les gènes en multicopies. L'algorithme Synergy2 regroupe les gènes par similarité en s'appuyant sur un arbre phylogénétique des individus et en reconstruisant simultanément un arbre phylogénétique des gènes afin de différencier les orthologues des paralogues **(fig. 12)**. Cette méthode nécessite une estimation *a priori* de différents paramètres. Nous avons réalisé plusieurs analyses en modifiant certains paramètres, concernant notamment la synténie car Synergy2 a été conçu pour des eucaryotes or les génomes bactériens sont beaucoup moins synténiques. Nous avons utilisé tous les paramètres par défaut excepté pour la fenêtre de synténie que nous avons réduite de 5000bp à 1000bp et la contribution de la synténie dans le poids des branches pour les arbres (réduite de $Synscale = 1$ à 0,5). Ces ajustements ont permis de réduire le coefficient de variation des orthogroupes. L'arbre phylogénétique précédemment généré par Parsnp a servi d'arbre guide pour Synergy2. Les résultats de cette analyse se présentent sous forme d' "orthogroupes" pour différentes catégories : les gènes orthologues ubiquistes et unicopies au sein des 73 génomes, les gènes présents-absents (PS) présents dans au moins deux génomes en simple copie et les gènes en multicopies (MCC). Une recherche par Blastn [Altschul et al., 1990] des orthologues chez *Xanthomonas vasicola* pv. *holcicola* (CFBP 2543), a été réalisée et cette séquence a été ajoutée à chaque orthogroupe avant de les aligner avec l'outil MACSE [Ranwez et al., 2011].

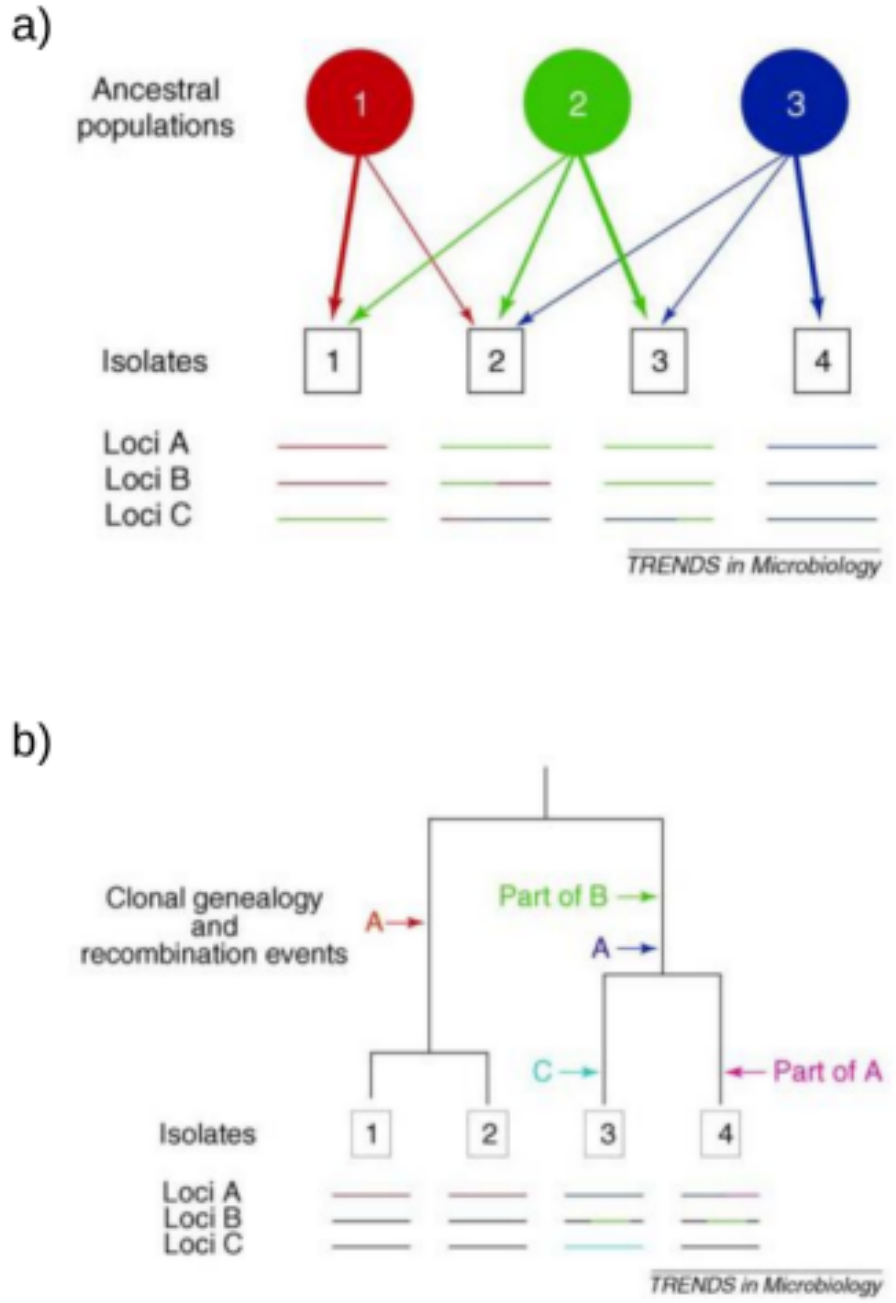


FIGURE 13 – Principe de STRUCTURE et de ClonalFrame.
a) STRUCTURE essaye de regrouper les individus en populations ayant les mêmes fréquences alléliques distinctes selon l'équilibre de Hardy-Weinberg. Le génotype de chaque individu est fait de blocs provenant des populations ancestrales. Le profil d'admixture est indicatif du flux de gènes qui s'est produit entre et parmi les populations. STRUCTURE ne modélise pas la généalogie clonale sous-jacente de la population bactérienne. Cela implique de lorsque deux isolats partagent un fragment de la même population ancestrale, ce n'est pas possible de dire si c'est le résultat d'un héritage clonal ou d'une récente recombinaison affectant l'un ou l'autre. Par conséquence STRUCTURE ne produit pas une estimation du flux de gènes et est plus adapté à des d'espèces très recombinantes car le signal clonal a probablement été effacé par la recombinaison. b) Illustration du modèle de ClonalFrame. Les événements de recombinaison sont indiqués en couleur et la généalogie clonale en noir. La partie basse montre les génotypes des isolats, avec les fragments affectés par la recombinaison colorés selon les événements dont ils sont originaires. ClonalFrame construit un arbre qui représente la généalogie clonale, et il localise les événements de recombinaison qui se sont produits sur chaque branche de la généalogie. Il n'infère pas l'origine des imports, et ne peut pas être utilisé pour étudier le flux de gènes entre les lignées. D'après [Didelot and Maiden, 2010].

2.5 Structure du complexe d'espèces *Xanthomonas axonopodis*

Le concept de population chez les bactéries est assez mal défini puisqu'il est basé sur la panmixie [6] et que les bactéries n'ont pas de reproduction sexuée. Nous utiliserons donc ici différentes méthodes permettant de délimiter des groupes de souches assimilables à des populations.

2.5.1 Structure génétique du complexe d'espèces *Xanthomonas axonopodis*

Dans le but d'identifier la structure des populations, nous avons comparé les résultats d'une approche multivariée : analyse en composante principale (ACP), avec ceux d'une analyse bayésienne. L'ACP a été réalisée avec le paquet R {adegenet} [Jombart and Ahmed, 2011] à partir de l'alignement produit par la suite Harvest. L'analyse bayésienne a été réalisée avec fastSTRUCTURE, qui est un algorithme dérivant du logiciel STRUCTURE et optimisé pour les gros jeux de données (voir principe **fig. 13a**) [Raj et al., 2014]. Adapté à notre jeu de données haploïde fastSTRUCTURE tente d'identifier des groupes d'individus minimisant le déséquilibre de liaison entre les SNP liés à une structure génétique. Pour une partition donnée de k populations, fastSTRUCTURE calcule pour chaque individu sa probabilité postérieure d'appartenir à l'une de ces k populations. Nous avons testé le nombre de population entre $k = 1$ et $k = 6$ avec 10 répétitions en utilisant les SNP extraits du *core genome* avec la suite Harvest.

Les 73 génomes ont été comparés deux à deux avec un script ANI.pl [Chen et al., 2015] utilisant l'algorithme de JSpecies [Richter and Rosselló-Móra, 2009] qui permet le calcul par BLAST des identités nucléotidiques moyennes (ANI). La matrice de distance basée sur $dist = -ln(ANI/100)$ a été calculée sur tous les génomes. D'autre part, une matrice de distance a été réalisée avec Simka [Benoit et al., 2015] sur la base des k-mers de 22 bp partagés. La visualisation des relations entre souches basées sur ces deux matrices de distances a été réalisée avec Declic (pour Delimitation of species with cliques) [Rimet et al., 2016]. En effet, selon la

6. La panmixie, en génétique des populations, est le principe qui considère que les individus sont répartis de manière homogène au sein de la population et se reproduisent tous aléatoirement

définition de l'espèce bactérienne, deux souches ayant une valeur ANI supérieure au seuil approximatif de 95% appartiennent à la même espèce [Richter and Rosselló-Móra, 2009]. Selon cette définition, l'espèce serait une clique, c'est-à-dire un sous-ensemble dont tous les éléments sont connectés les uns avec les autres par une distance inférieure à un seuil fixé.

2.6 Tests de neutralité

Une fois les "populations" définies, des statistiques descriptives ont pu être calculées notamment pour tester la neutralité et permettre l'utilisation d'outils de génétiques de population conditionnés par cette hypothèse. Les spectres de fréquences alléliques ont été calculés avec VCFtools [Danecek et al., 2011] sur le *core genome*, c'est-à-dire les régions codantes et non codantes, obtenu avec la suite Harvest. Trois tests de neutralité incluant le D de Tajima [Tajima, 1983], une version normalisée du H de Fay et Wu's [Fay and Wu, 2000], et le EW de Ewens-Watterson [Watterson, 1978] ont été calculés pour détecter les écarts au modèle neutre avec le programme DH [Zeng et al., 2006] sur chaque alignement de gène orthologue (c'est-à-dire les gènes du *core genome* identifiés avec Synergy 2), au sein de chaque groupe. Chacun de ces tests se base sur les attendus neutres du spectre de fréquences alléliques. Les tests de Tajima et de Fay et Wu reposent sur la comparaison de deux estimateurs de $\theta = 2Ne\mu$ (chez les haploïdes), le paramètre de diversité nucléotidique. Sous l'hypothèse de neutralité sélective ou d'équilibre démographique, la différence entre chacun de ces estimateurs de thêta, pris deux à deux, est censée être nulle. Le test du D de Tajima compare la moyenne du nombre de différences entre individus pris deux à deux (θ_π) avec le nombre de sites polymorphes (θ_W), il est défini par $D = \frac{\theta_\pi - \theta_W}{\sqrt{Var(\theta_\pi - \theta_W)}}$.

Un D de Tajima positif indique un excès d'allèles à fréquence intermédiaire et donc un goulet d'étranglement, une sélection balancée récente ou une sous-structuration. Un D de Tajima négatif indique au contraire un excès de polymorphisme (un excès de variants rares) et donc une population en expansion ou un balayage sélectif.

Le H normalisé utilise un estimateur de diversité nucléotidique basé sur la fréquence des allèles dérivés (θ_L), il est défini par $H = \frac{\theta_\pi - \theta_L}{\sqrt{Var(\theta_\pi - \theta_L)}}$. Les indices

de diversité sont calculés sur des parties différentes du spectre de fréquence c'est pour cela que θ_W est sensible aux changements de variants à basse fréquence, θ_π aux changements des variants à fréquence intermédiaire, et θ_L aux variants à forte fréquence.

EW est basé sur les fréquences des haplotypes, le test consiste à comparer la valeur de la statistique $F = \sum_{i=1}^{k} (pi^2)$ (k étant le nombre d'haplotypes dans un échantillon de taille n et pi est la fréquence du $i^{ème}$ haplotype) à celles obtenues à partir d'échantillons simulés sous l'hypothèse de neutralité de la population. Il détecte un excès d'haplotypes, qui sont des déviations par rapport au modèle neutre attendu lors de sélection positive ou balayage sélectif.

Les trois tests statistiques composés (DH, HEW, DHEW) qui combinent les probabilités des tests de neutralité précédents ont aussi été calculés. Ces tests statistiques composés prennent les avantages des différents tests et sont plus robustes aux *biais* comme la démographie et la sélection d'arrière-plan (*background selection*) [Zeng et al., 2007]. Le test composé DH, est un composé du D de Tajima et du H de Fay and Wu. Les HEW et DHEW sont des composés de H ou DH avec EW. Le DHEW est le plus robuste puisqu'il combine les probabilités de trois tests, il est relativement insensible à sélection d'arrière-plan et à la démographie. La significativité des tests de neutralité et des statistiques composées a été évalué avec 10000 simulations coalescentes neutres sans recombinaison, conditionné par la taille de l'échantillon et θ_W estimé à partir des données. L'hypothèse d'absence de recombinaison est conservative pour les tests basés sur le spectre de fréquence et a peu d'impact sur les tests statistiques composés [Zeng et al., 2007].

2.7 Analyse des flux de gènes et de la recombinaison

Après avoir vérifié la neutralité de nos populations, il est possible d'estimer selon différentes méthodes le flux de gènes. Il s'agit alors de mesurer indirectement l'isolement génétique entre groupes de souches.

2.7.1 Flux de gènes d'origine extérieure au complexe d'espèces *Xanthomonas axonopodis*

Le *core genome* non-filtré pour la recombinaison, issu de Harvest a été analysé avec ClonalFrameML [Didelot and Wilson, 2015] **(fig. 13b)**. La phylogénie obtenue avec la suite logicielle Harvest est utilisée comme topologie de départ. ClonalFrameML est un programme d'inférence bayésienne permettant de détecter les relations phylogénétiques discordantes le long d'un alignement de séquences. Cette méthode est particulièrement adaptée pour retracer les généalogies des organismes clonaux. Il détecte les imports de séquences exogènes, introduisant un nombre important de substitutions dans notre jeu de données. Pour ce faire, l'algorithme modélise l'import de ces fragments et tend à sous-estimer le nombre d'événements de recombinaison si le donneur est génétiquement proche ou dans la population étudiée. Contrairement aux modèles phylogénétiques, cette méthode prend en compte à la fois les événements de mutation et de recombinaison homologue pour reconstruire une généalogie clonale de l'échantillon. Le modèle de coalescence utilisé correspond au modèle de Kingman (1982), dans lequel la taille de l'échantillon est supposée constante au cours du temps. Après avoir validé l'hypothèse de taille constante des populations étudiées avec les différents tests de neutralité, une première analyse a été réalisée dans les conditions standards. Les valeurs des paramètres R/θ, $1/\delta$ et nu ont été estimées. Le paramètre R représente le taux de recombinaison, θ la diversité nucléotidique, δ la longueur moyenne des fragments importés, et nu la divergence moyenne de l'ADN importé. Une seconde analyse a ensuite été réalisée en indiquant comme *a priori* les valeurs de paramètres estimés lors de la première analyse, et en calculant cette fois les paramètres par branche. Ce paramètre *em_branch* spécifie les contraintes sur la variabilité des paramètres de recombinaison sur les branches de l'arbre. Dans cette deuxième analyse nous avons utilisé $em\_branch = 1$ le modèle le moins contraint (le plus dispersé).

2.7.2 Flux de gènes au sein du complexe d'espèces *Xanthomonas axonopodis*

Afin d'estimer l'importance du flux génique contemporain entre les groupes précédemment identifiés, nous avons utilisé deux méthodes reposant sur des hypothèses complémentaires comme cela sera détaillé plus bas. L'indépendance des sites doit être vérifiée avant d'utiliser certaines méthodes. Le déséquilibre de liaison représente l'association statistique d'allèles présents à des locus différents. Le degré d'association a été utilisé pour inférer le niveau de clonalité des organismes, une forte association étant corrélée avec un fort niveau de reproduction clonale [Guttman, 1997]. La mesure du déséquilibre de liaison a été réalisée sur les SNP extraits du *core genome* avec Harvest pour chacun des groupes comportant plus de deux génomes : 9.2, 9.41, 9.5 et 9.6. VCFtools [Danecek et al., 2011] calcule ce déséquilibre pour chaque paire de sites dans une fenêtre 50 kb.

a) Analyse avec TreeMix Afin d'inférer les relations entre les populations à partir du polymorphisme, TreeMix [Decker et al., 2014] construit tout d'abord un arbre des populations en maximum de vraisemblance. Les populations sont représentées par les feuilles de l'arbre, et les branches sont les relations inférées entre ces populations. Il arrive que plusieurs branches conduisent aux mêmes populations, ce qui suggère alors de la migration entre ces populations. Nous avons utilisé les SNP extraits du *core genome* avec la suite Harvest. Par défaut, TreeMix utilise les sites bialléliques ce qui correspond à un modèle de mutation en site infinis [7], les SNP ont donc été filtrés avec VCFtools (option *–max-alleles* 2) pour obtenir au final 107759 SNP. Nous avons enraciné le dendrogramme avec le groupe 9.3, le plus divergent (voir III.3.1). Pour tenir compte du fait que les SNP voisins pourraient ne pas être indépendants, nous avons regroupé les SNP en fenêtre de 10kb dans le génome (la valeur choisie doit excéder la mesure du déséquilibre de liaison, voir plus bas figure 23). Plusieurs analyses ont été exécutées en considérant un nombre d'événements de migrations allant de 1 à 4 (-m 1-4). Les événements sont ajoutés jusqu'à ce que 99,9% de la variance entre les populations soit expliquée par le modèle. Les résultats sont tracés sur un arbre des populations en utilisant le script

7. Selon ce modèle, toute nouvelle mutation se produit à un nouveau site qui n'a encore jamais été affecté par une mutation

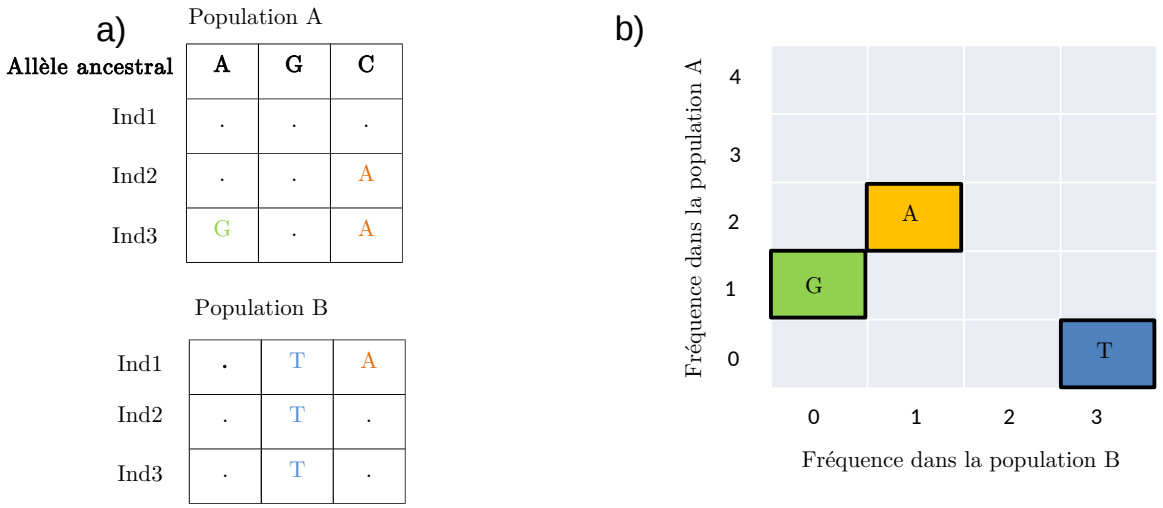


FIGURE 14 – Principe du spectre de fréquence joint.
a) Distribution des allèles dans chacun des individus (Ind) des populations A et B. Par exemple, l'allèle dérivé « G »est en fréquence 1 dans la population A et 0 dans la population B. b)Spectre de fréquence joint des deux populations A et B. Chaque axe représente la fréquence d'un allèle dérivé dans les deux populations étudiées. Ainsi, les allèles partagés entre les deux populations se retrouvent proches de la diagonale sur la figure, tandis que les allèles propres à chaque population vont être représentés proches des axes[Gutenkunst et al., 2010].

R de l'outil Treemix.

b) Analyse avec fineSTRUCTURE Afin de préciser les flux de gènes entre les populations, le pipeline fineSTRUCTURE v.2 [Yahara et al., 2013] a également été utilisé afin de réaliser une peinture chromosomique (*chromosome painting*). Cet algorithme fait l'hypothèse que tous les échanges génétiques se produisent entre les souches du jeu de données contrairement à ClonalFrameML. Cette méthode recherche l'haplotype voisin le plus proche appelé le "co-ancêtre" parmi les individus échantillonnés. L'algorithme ChromoPainter génère une matrice de similarité des co-ancêtres (un résumé des relations entre les haplotypes du jeu de données) qui est utilisée par fineSTRUCTURE pour réaliser un regroupement des souches avec une approche bayésienne. FineSTRUCTURE utilise un algorithme de Markov Chain Monte Carlo (MCMC) pour lequel nous avons imposé 1000000 itérations, avec une période de préchauffage (burn-in) de 50000 itérations et un échantillonnage toutes les 100 itérations. Il se peut que parfois, des clones, ou des génomes tellement similaires qu'un seul donneur est considéré par le modèle perturbent l'algorithme (Lawson d., comm. pers.). Afin de contourner ce problème, nous avons enlevé les génomes très similaires jusqu'à obtenir une matrice de co-ancêtre satisfaisante. Le jeu de données ainsi nettoyé comprend 33 souches. Une *heatmap* du flux de gènes est obtenue avec l'interface fineSTRUCTURE.

2.8 Inférence de l'histoire démographique

Pour inférer le scénario de divergence le plus probable entre les groupes que nous avons identifiés nous avons utilisé δaδi [Gutenkunst et al., 2009] qui est une méthode d'inférence de scénarios évolutifs basée sur un modèle de diffusion. Cette méthode permet d'obtenir des spectres de fréquence joints dont les profils diffèrent en fonction des différents scénarios évolutifs. En effet, les événements démographiques affectent le spectre de fréquence d'une population [Achaz, 2009][Tellier and Lemaire, 2014]. Les spectres de fréquence joints correspondent à une matrice représentant la fréquence de chaque allèle dérivé au sein des populations testées **(fig. 14)**. Les SNP sont filtrés avec VCFtools afin de ne conserver que les bi-alléliques . La vraisemblance d'un modèle dans le cas de liaisons entre sites est

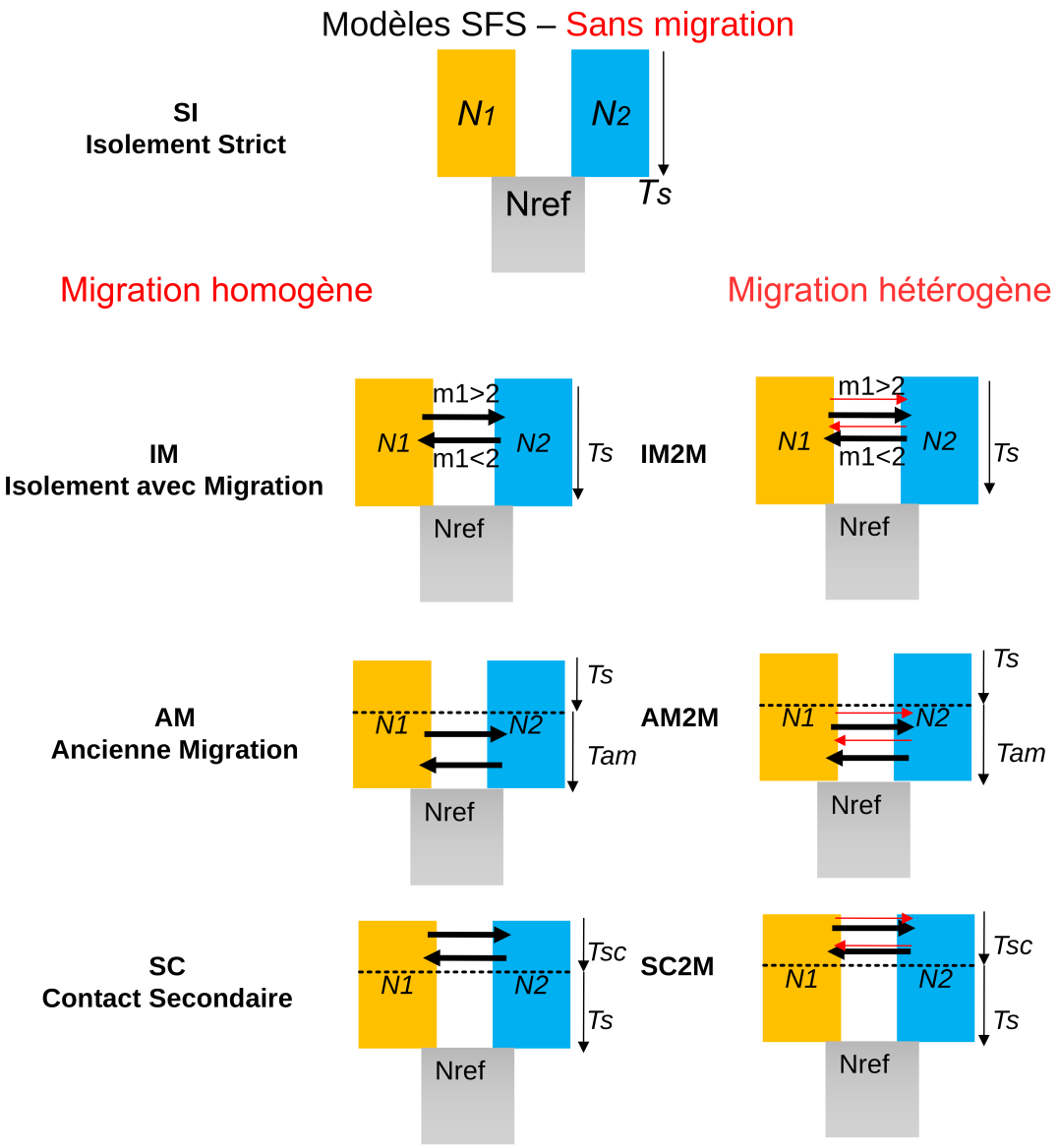


FIGURE 15 – Représentation des modèles démographiques testés dans cette étude. Chaque modèle consiste en une population ancestrale de taille Nref qui diverge en deux populations de taille N1 et N2 pendant le temps Ts (dans les modèles SI, IM), Tam+Ts(AM), ou Ts+Tsc (SC) générations. Il y a échange de migrants pendant Ts (IM), Tam (AM), ou Tsc (SC) générations à un taux de m1<2 de la population 2 vers la population 1, et m2<1 dans l'autre sens. Les modèles en migration hétérogène infèrent deux taux de migration (2M), Un taux pour la migration dans le génome évoluant de façon neutre et un taux dans les îlots génomiques. A ces modèles on peut ajouter les paramètres P1 et P2 qui donnent la proportion de génome évoluant de façon neutre dans les population 1 et 2 respectivement. Ces modèles sont IM2M2P, AM2M2P, et SC2M2P. A tous ces modèles on intègre une variation exponentielle de la taille effective de population qui modélise un bottleneck ou une expansion, les tailles effectives finales sont N'1 et N'2 pendant une durée de Te génération. Ces modèles sont SIex2, IMex, IMex2, AMex2, SCex2, IM2M2Pex2, AM2M2Pex2, SC2M2Pex2, IM2Mex2, AM2Mex2, SC2Mex2.

une vraisemblance composite que l'on ne peut pas utiliser pour choisir le meilleur modèle. C'est pourquoi, pour minimiser l'impact du déséquilibre de liaison, nous avons utilisé des SNP filtrés tous les 3000bp, sur la base des résultats d'analyse de décroissance du déséquilibre de liaison mentionnée plus haut. Ce jeu de données contient 1223 SNP pour les paires de populations 9.5 et 9.6, et 1235 SNP pour les paires de populations 9.5 et 9.2. Les SNP ont été polarisés en utilisant la séquence du génome *Xanthomonas vasicola* pv. *holcicola* (CFBP 2543), qui représente l'état ancestral. Les paramètres démographiques des différents modèles ont été estimés à partir du spectre de fréquence joint de la distribution des SNP.

Les différents modèles démographiques sont décrits dans la **figure 15**. Nous avons utilisé les scripts développés par C. Fraïsse et al. (*comm. Pers.*). Vingt analyses indépendantes sur les 20 modèles testés ont été réalisées pour vérifier la convergence. Le meilleur modèle a été choisi sur la base du critère d'information d'Akaike (AIC), et la probabilité d'avoir choisi le meilleur modèle a été calculée avec l'Akaike weight (W_{AIC}) [Rougeux et al., 2016]. Lorsque la vraisemblance d'un modèle complexe est meilleure que celle d'un modèle directement plus contraint (moins de paramètres), un test LRT a été réalisé.

Une fois le meilleur modèle choisi, on peut utiliser le jeu de SNP complet (non filtré tous les 3000bp) pour calculer des paramètres, car cette fois le déséquilibre de liaison n'a pas d'impact. Pour estimer l'incertitude des paramètres associés aux meilleurs modèles nous avons utilisé la méthode de bootstrap implémentée dans δaδi. Cette fonction calcule l'incertitude à partir de 100 ré-échantillonnages (bootstraps non paramétriques : échantillonnage avec remplacement) des 257587 SNP non filtrés pour les groupes 9.5 et 9.6, et des 121881 SNP non filtrés pour les groupes 9.5 et 9.2. Les paramètres sont calculés en unité de taille efficace (Nref). La taille efficace est égale à $Nref = \frac{\theta}{2L\mu}$ (μ étant le taux de mutation) et $L = \frac{xL}{y}$(x nombre de sites retenu par δaδi, L longueur du core-génome en bp et y le nombre total de SNP du *core genome*) [Rougeux et al., 2016]. Le taux de mutation μ a été précédemment estimé à 2.10^{-8} par site et par an [Mhedbi-Hajri et al., 2013].

Afin de pouvoir comparer sur le même jeu de données les flux de gènes entre les trois populations 9.6, 9.5 et 9.2, une analyse à trois populations a été réalisée. L'inférence d'un modèle de divergence à trois populations avec changement de taille, n'est pas possible à cause du trop grand nombre de paramètres et du temps

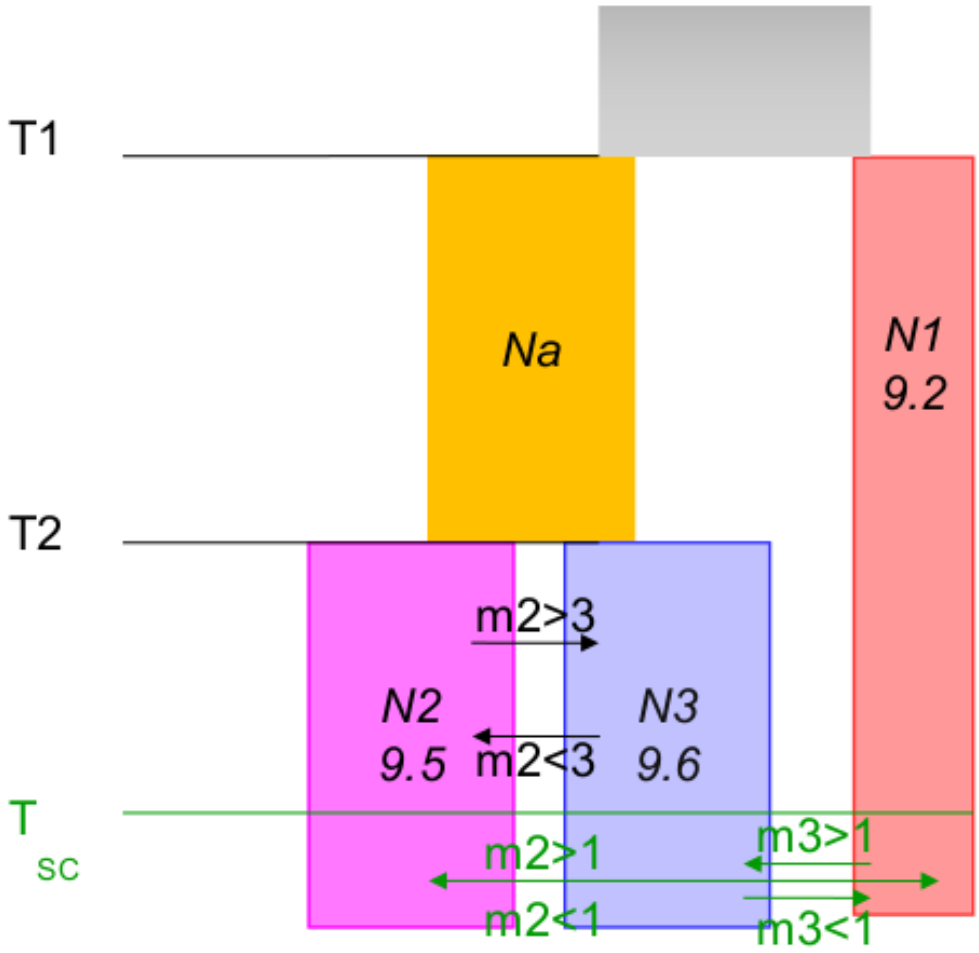


FIGURE 16 – Modèle à trois populations IMSC.
Le modèle consiste en un premier événement de divergence se produisant au temps T1 entre la population 9.2 de taille efficace N1 et une population ancestrale Na. A T2 la population ancestrale Na diverge en deux populations, 9.5 de taille N2 et 9.6 de taille N3. Le flux de gènes m2>3 de 9.5 vers 9.6, et m2<3 dans l'autre sens, est constant depuis la divergence. Alors que le flux de gènes se produit depuis TSC entre les populations 9.6 et 9.2 (m3>2 et m3<2) et 9.5 et 9.2 (m2>1 et m2<1).

de calcul nécessaire. C'est pour cela que j'ai testé un modèle simple prenant en compte les résultats des deux analyses par paires, et mis au point le script pour le modèle IMSC : modèle avec migration constante entre 9.5 et 9.6 et contact secondaire entre 9.2 et 9.6 et 9.5 **(fig. 16)**. Dix-sept analyses ont été réalisées sous ce modèle, avec un jeu de données de 106 142 SNP pour les trois populations 9.6, 9.5 et 9.2.

2.9 Recherche des gènes liés à l'adaptation

La divergence des groupes 9.5 et 9.6 est assez récente et probablement accentuée par des barrières génomiques. Afin de rechercher les gènes pouvant être liés à la divergence adaptative entre groupes, deux approches ont été utilisées. Tout d'abord le *core genome* des groupes 9.5 et 9.6 a été scanné afin de localiser les régions fortement différenciées (pics de F_{ST}) entre ces deux groupes révélant des signatures de sélection. Toute mutation porteuse d'une adaptation locale verra sa fréquence augmenter dans la population cible, augmentant localement la différenciation génétique (*i.e.* F_{ST}) avec les autres populations soumises à d'autres contraintes environnementales. Cette augmentation en fréquence de la mutation avantageuse aura aussi pour conséquence de diminuer localement la diversité nucléotidique, et ce dans une région génomique dont la largeur sera inversement proportionnelle au taux local de recombinaison. C'est ce qu'on appelle le phénomène d'auto-stop génétique (*hitchhiking*)[Smith and Haigh, 1974]. Par conséquent, les différences de fréquence des allèles entre les populations peuvent être plus extrêmes dans les régions du génome sous sélection [Akey, 2002]. Une corrélation entre ces régions fortement différenciées et des valeurs négatives de D de Tajima confirmerait la sélection. Une deuxième approche a consisté en l'identification par Analyse en Composantes Principales (ACP) sur le core genome des SNP montrant une différenciation excessive particulière entre groupes et qui seraient donc supposés être des marqueurs d'une adaptation locale. Cette méthode est implémentée pas le paquet R {PCAdapt} [Luu et al., 2016].

2.9.1 Par la recherche des SNP sous sélection dans des régions différenciées

Un indice de la structure de la population (F_{ST}) a été utilisé pour rechercher les régions différenciées entre les populations des groupes 9.5 et 9.6. Les valeurs de F_{ST} varient de 0 (populations identiques) à 1 (populations complètement différenciées). Les valeurs de F_{ST} et des D de Tajima ont été calculés à partir des SNP extraits de l'alignement du *core genome* issu de la suite Harvest des groupes 9.5 et 9.6 en utilisant VCFtools [Danecek et al., 2011] sur des fenêtres de 1000bp. Ces valeurs ont été placées sur l'alignement du *core genome*.

2.9.2 Par *genome scan* avec PCAdapt

Le *genome scan* a été réalisé avec le paquet R {PCAdapt} [Luu et al., 2016] afin de détecter les SNP potentiellement sous sélection pour l'adaptation locale à partir d'un alignement du *core genome* des groupes 9.5 et 9.6. Les SNP dits *outliers* sont considérés comme atypiques pour la différenciation génétique, puisqu'ils sont censés différencier plus les populations écologiquement différentes que ne le font les marqueurs neutres. Quand l'adaptation se produit dans deux lignées différentes un génome scan basé sur une ACP peut permettre d'identifier des locus indiquant un type de structure qualitativement ou quantitativement différent de celle observée sur marqueurs neutres. Dans le cas de deux populations, l'axe 1 de l'ACP reflète la structure neutre des populations, c'est à dire celle obtenue aux marqueurs uniquement soumis à la migration et la dérive. Les axes supplémentaires (PC2 ou plus) rendront compte de la variance génétique aux marqueurs potentiellement sous sélection. Les gènes qui seraient alors identifiés pourraient apporter des éléments sur les déterminants biologiques ayant conduit ou du moins liés à la différenciation génétique.

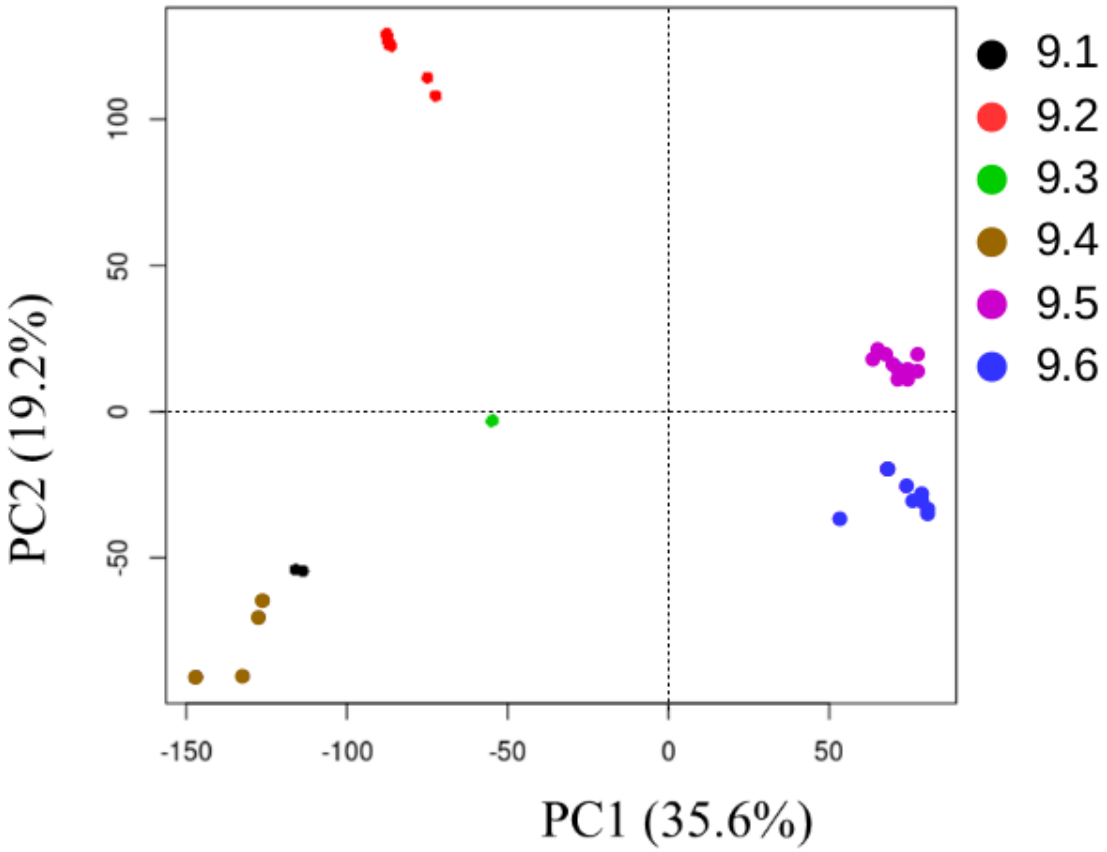


FIGURE 17 – Analyse en composantes principales (ACP).
Analyse en composantes principales (ACP) basée sur 246 585 SNP du *core genome* des 73 souches du complexe d'espèces *Xanthomonas axonopodis*. L'axe PC1 explique 35,6% de la variance et l'axe PC2 en explique 19,2%. Les souches sont indiquées par des points colorés selon leur appartenance aux groupes

3 Résultats

3.1 Le complexe d'espèces *Xanthomonas axonopodis* possède différents niveaux de structuration

Le complexe d'espèces *Xanthomonas axonopodis* présente une structure hiérarchique et peut se structurer en 2, 4, 5 ou 6 groupes. Selon les seuils utilisés les groupes 9.4 - 9.1 et 9.5 - 9.6 sont soit regroupés soit séparés par les différentes méthodes.

L'ACP **(fig. 17)** a permis de définir deux groupes selon l'axe PC1, confirmés par fastSTRUCTURE $k = 2$ **(fig. 18b)**. Le premier groupe est composé de l'espèce *Xanthomonas citri* (avec les groupes communément appelés 9.5 et 9.6), le deuxième contient les espèces *X. euvesicatoria*, *X. phaseoli* et *X. axonopodis* (communément appelées 9.2, 9.4-9.1, et 9.3). L'arbre phylogénétique **(fig. 18a)** produit par Harvest a été établi à partir des *core* régions des 73 génomes qui représentent 36% du génome de référence CFBP 4885 (1,9 Mb) et 246 585 SNP. L'arbre a confirmé que *X. citri*, composé des groupes 9.5 et 9.6, forme un groupe monophylétique.

La visualisation des relations entre souches basées sur les valeurs d'ANI avec un seuil de 95% **(fig. 19a)** et des k-mers avec un seuil de 45% d'identité **(fig. 19c)** permet de dénombrer quatre cliques, correspondant aux quatre espèces *Xanthomonas citri*, *X. euvesicatoria*, *X. phaseoli* et *X. axonopodis* qui s'individualisent sous forme de groupes monophylétiques sur l'arbre phylogénétique **(fig. 18a)** avec des valeurs de bootstrap de 100%.

L'analyse fastSTRUCTURE avec $k = 4$ n'a pas discriminé les mêmes 4 groupes que ceux décrits ci-dessus. En effet, avec $k = 4$, les espèces *X. phaseoli* (9.4 et 91) et *X. axonopodis* (9.3) ne sont pas discriminées, et c'est l'espèce *X. citri* composée des groupes 9.5 et 9.6 qui est divisée en deux. Cela peut être dû au faible effectif du groupe 9.3 de l'espèce *X. axonopodis* qui rend difficile l'inférence d'une population avec 2 génomes.

L'ACP **(fig. 17)**, fastSTRUCTURE $k = 5$ **(fig. 18b)**, l'ANI lorsque l'on ajuste le seuil à 96,5%, et les k-mers au seuil de 55% d'identité **(fig. 19b et d)** structurent la population en cinq groupes, correspondant à 9.6, 9.5, 9.3, 9.2 et 9.4/9.1.

Finalement seul fastSTRUCTURE avec $k = 6$ permet de distinguer les groupes

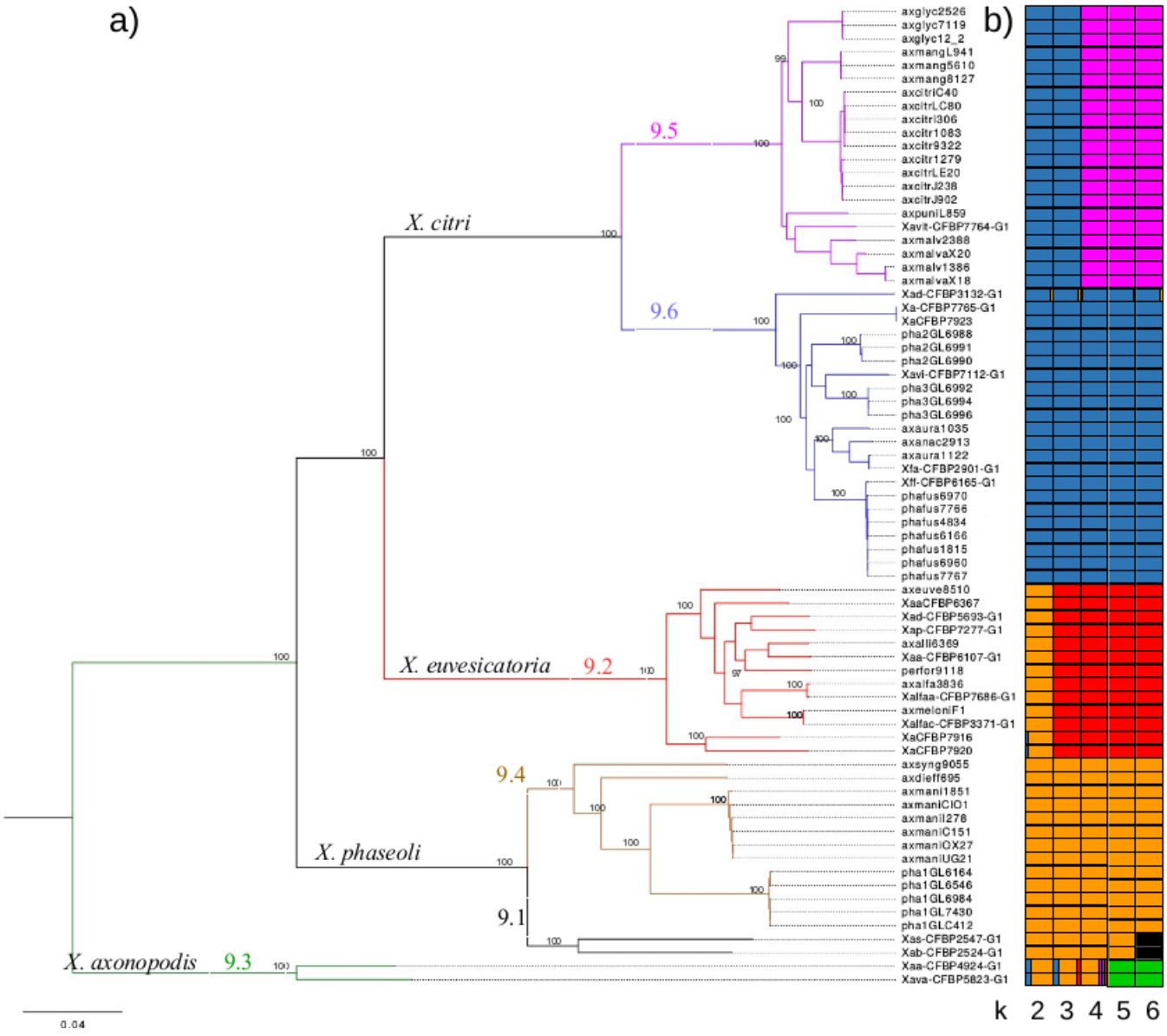


FIGURE 18 – Structure du complexe d'espèces *X. axonopodis*.
a)Phylogénie en maximum de vraisemblance approximative de 73 souches du complexe d'espèces *Xanthomonas axonopodis* basé les SNP non recombinants du *core genome* (1,87 Mb). Les branches violettes correspondent aux souches du groupe de Rademaker 9.5, les bleues au groupe 9.6, les rouges au groupe 9.2, les marrons au groupe 9.4, les noirs au groupe 9.1, les vertes au groupe 9.3. Les valeurs de bootstrap sont calculées en pourcentage pour 1000 répétitions. b) Structure de la population basée sur une inférence Bayésienne de 73 souches. Les couleurs indiquent les différents groupes inférés selon la valeur de *k*.

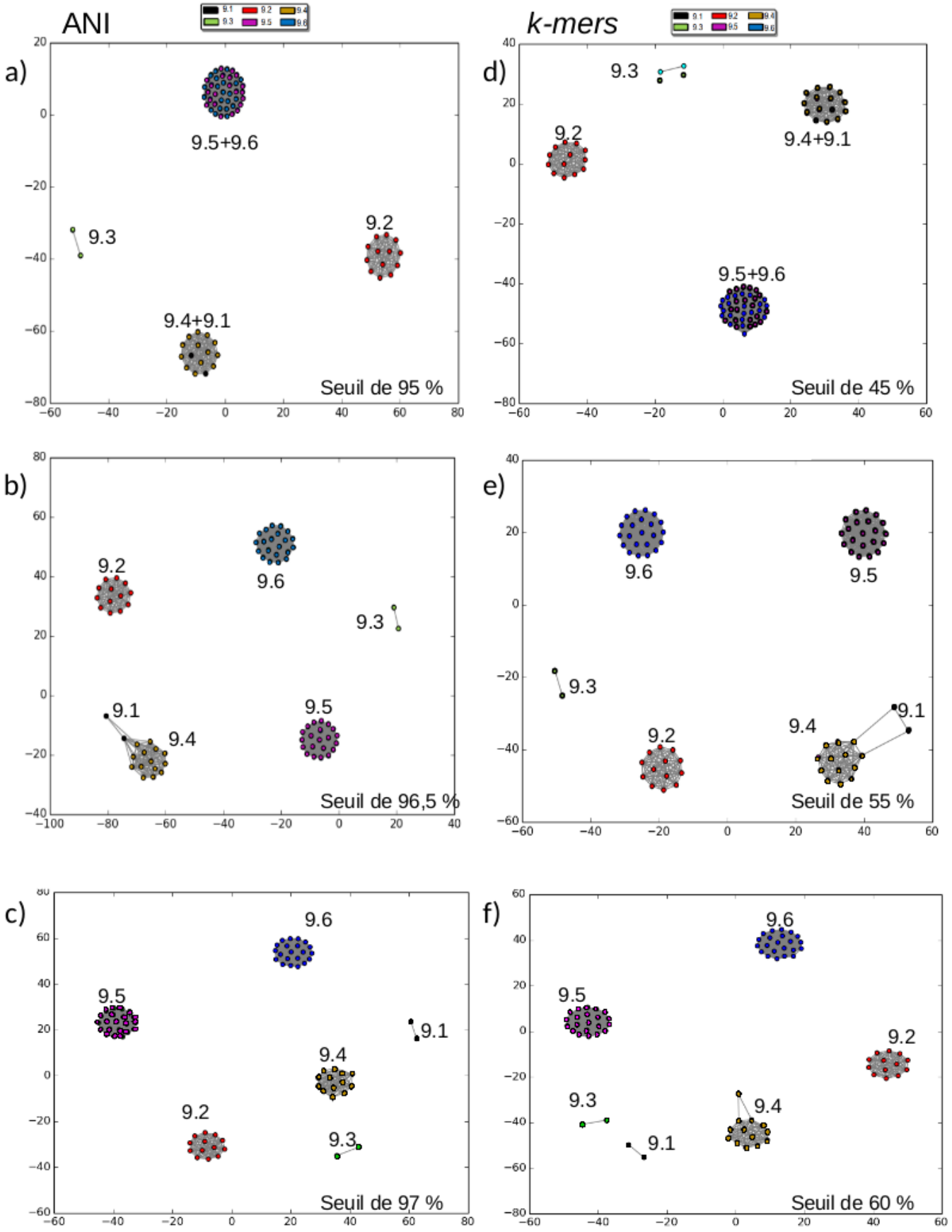


FIGURE 19 – Représentation des cliques.
Cliques pour les matrices de distances des 73 souches du complexe d'espèces *Xanthomonas axonopodis*. Les souches sont indiquées par des points colorés selon leur appartenance aux groupes. a,b,c) ANI aux seuils de 95%, 96,5%, et 97% d, e, f) K-mers aux seuils de 45%, 55% , et 60% d'identité des kmers.

TABLE III – Valeurs moyennes des différents tests de neutralité par groupes
Les tests réalisés sont le D de Tajima, Le H de Fay et Wu's et le EW de Ewens-Watterson

Groupes	D de Tajima	H de Fay et Wu	EW Ewens-Watterson
9.2	$-0,32 \pm 0,78$	$-0,84 \pm 1,16$	$0,21 \pm 0,15$
9.41	$-0,75 \pm 0,76$	$-1,26 \pm 1,04$	$0,29 \pm 0,11$
9.5	$-0,10 \pm 0,82$	$-0,77 \pm 1,24$	$0,32 \pm 0,15$
9.6	$-0,41 \pm 0,93$	$-1,01 \pm 1,49$	$0,29 \pm 0,15$

9.4 et 9.1 ce qui correspond exactement aux 6 groupes de Rademaker.

Si on se base sur les valeurs d'ANI les 4 espèces, *X. phaseoli*, *X.euvesicatoria*, *X. citri* et *X. axonopodis* sont différentiées. *X. phaseoli* comprend les groupes 9.1 (représenté par 2 génomes) et le groupe 9.4 ; nous appellerons ce groupe 9.41 pour la suite des analyses. Pour l'espèce *X. citri*, il ne nous semblait pas pertinent de ne considérer qu'une population au vu de la forte sous structuration identifiée par fastSTUCTURE, l'ACP, l'ANI dès que l'on ajuste le seuil à 96,5%, et la phylogénie. C'est pour cette raison que 5 groupes ont été retenus, appelés 9.2, 9.3, 9.41, 9.5, 9.6 qui ne correspondent donc pas tous à une espèce. La pertinence de ce regroupement sera vérifiée par l'analyse des flux de gènes entre les souches de ces groupes.

3.2 Identification des orthogroupes

Un orthogroupe est défini comme un set de gènes descendant d'un unique gène dans le dernier ancêtre commun de l'espèce considérée. Synergy2 identifie 2071 orthogroupes de gènes communs aux 73 génomes. Seul 2058 orthologues de *Xanthomonas vasicola* pv. *holcicola* (CFBP 2543) ont pu être alignés aux orthogroupes. La matrice des orthogroupes présents dans au moins deux génomes en simple copie (PS) contient 7288 gènes, les génes multicopies (MCC) représentent 318 gènes.

3.3 Tests de la neutralité du polymorphisme au sein de chaque population

La majorité des valeurs du D de Tajima est plutôt négatives. Les valeurs du D de Tajima sont en moyenne de $-0,32 \pm 0,78$ dans le groupe 9.2, $-0,75 \pm 0,76$ dans le groupe 9.41, $-0,10 \pm 0,82$ dans le groupe 9.5, $-0,41 \pm 0,93$ dans le groupe 9.6 **(tab. III)**. Les probabilités sont majoritairement non significatives ($pval > 0,05$) indiquant que ces populations ne sont pas différentes du modèle neutre **(fig. 20)**. Les valeurs du EW qui se base sur les fréquences alléliques sont toutes positives, ce qui indique une sous-structuration des populations qui semble plus importante dans le groupe 9.41 **(fig. 20c)**. Les trois tests composés montrent une dispersion plutôt positive des probabilités ($pval > 0,05$) indiquant que la majorité des gènes ne rejettent pas l'hypothèse du modèle neutre **(fig. 21)**. Pour le test DHEW le plus

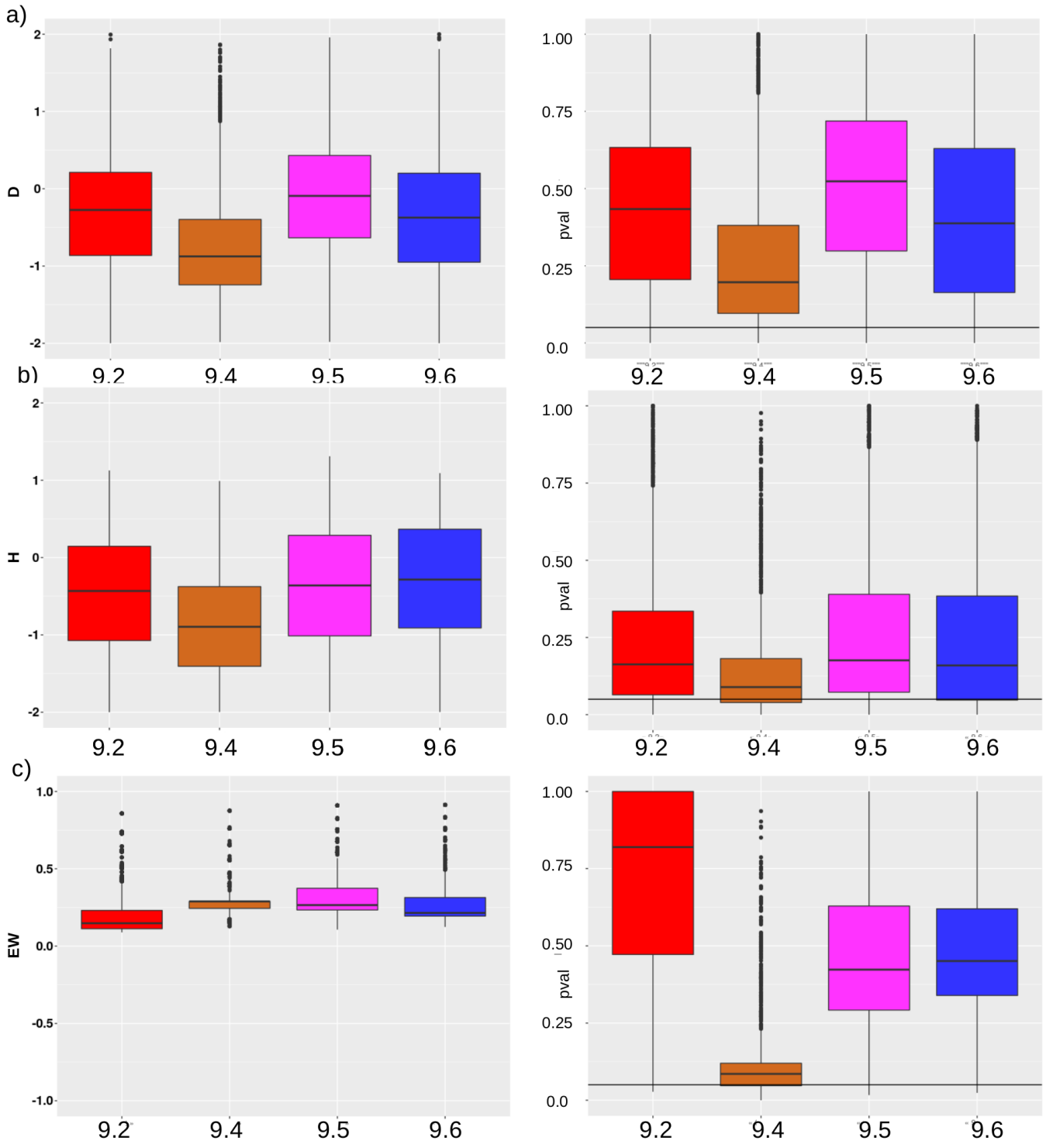


FIGURE 20 – Distribution des valeurs des tests de neutralité par groupe.
a) Boxplots des valeurs du D de Tajima à droite, avec à gauche la distribution des probabilités b) Boxplots des valeurs du H de Fay et Wu, avec la distribution des probabilités à gauche c) Boxplots des valeurs du EW de Ewens-Watterson avec à gauche les probabilités. La ligne noire indique une probabilité de 0,05. Les valeurs n'ont pas été calculées pour les groupes 9.3 et 9.1 car les effectifs étaient trop faibles.

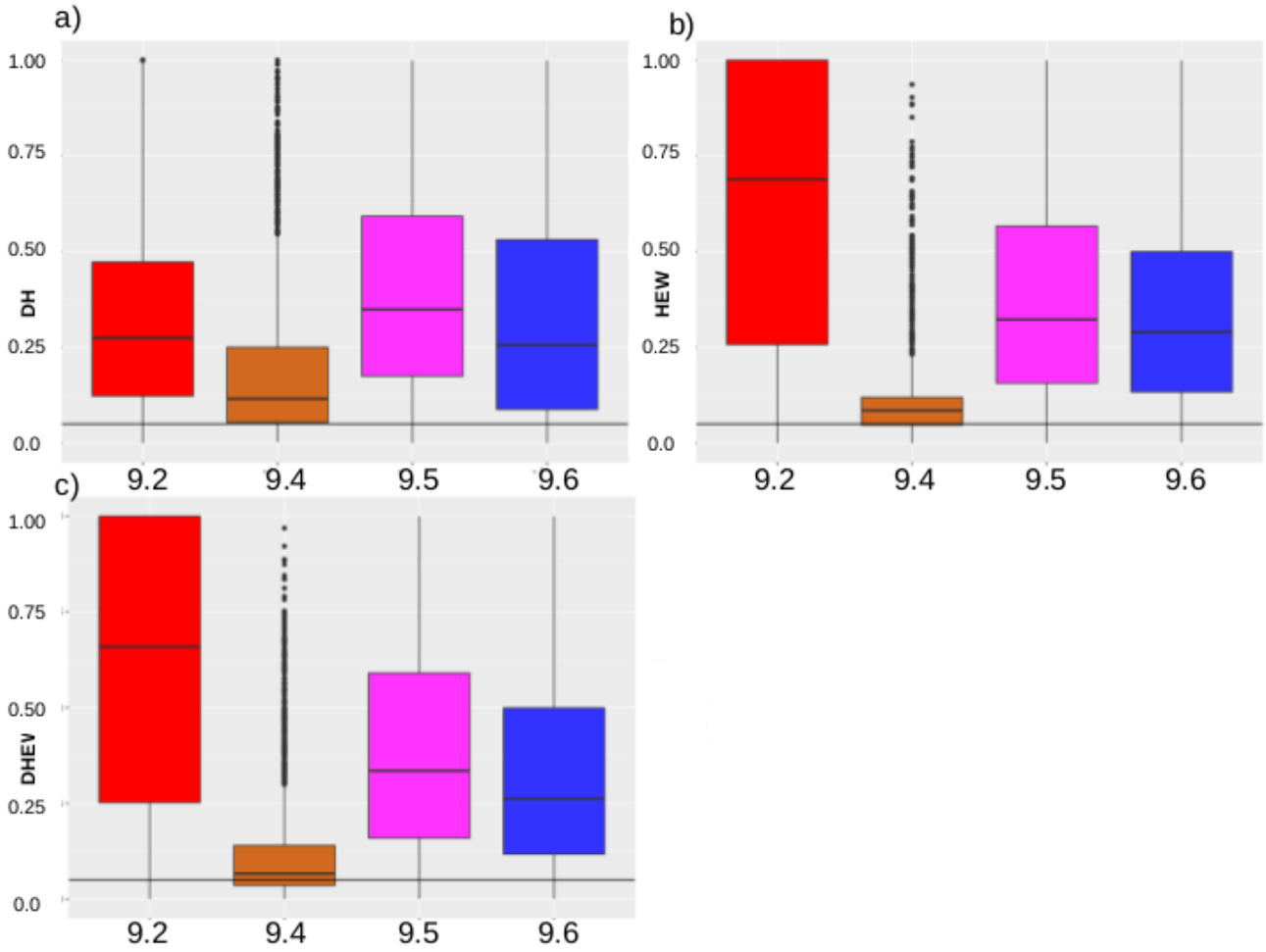


FIGURE 21 – Distributions des probabilités des tests de neutralité composés.
Probabilités des tests de neutralité composés a) DH b) HEW et c) DHEW par groupe. Les valeurs n'ont pas été calculées pour les groupe 9.3 et 9.1 car les effectifs étaient trop faibles. La ligne noire indique une probabilité de 0,05.

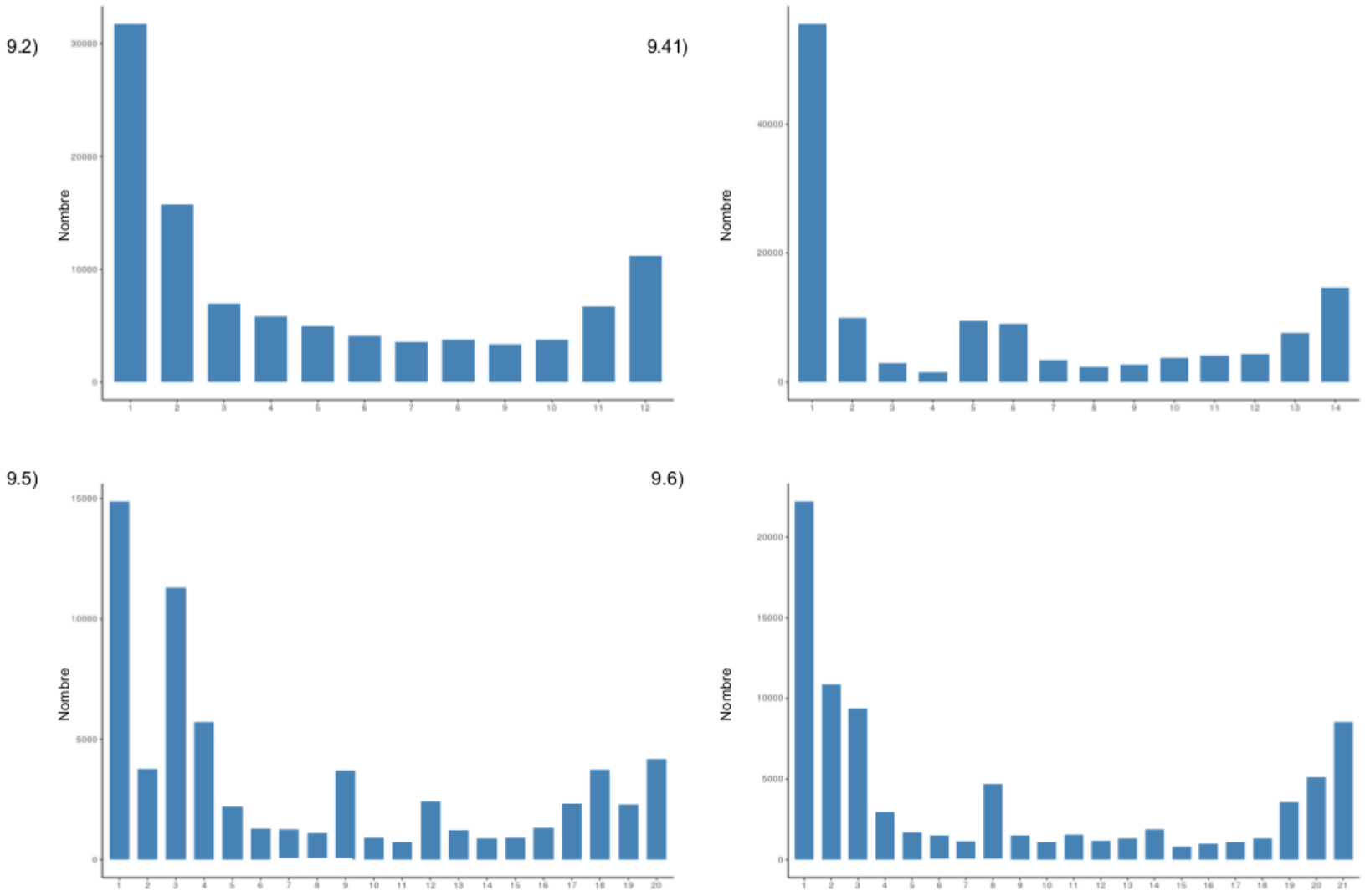


FIGURE 22 – Spectres de fréquence de l'allèle dérivé dans chaque groupe.
La fréquence est indiquée en ordonnée et le nombre d'allèles est indiqué en abscisse.

robuste, 7,8% des gènes dans le groupe 9.2, 7% dans le groupe 9.5, 9,6% dans le groupe 9.6, 35% dans la population 9.41 rejettent l'hypothèse neutre. Les spectres de fréquence des allèles dérivés **(fig. 22)** nous indiquent que dans les groupes 9.41, 9.5 et 9.6, il semble exister un excès de mutations en fréquence intermédiaire. Une telle distribution pourrait indiquer une sous structuration, ou de la sélection balancée. Dans le groupe 9.41, les singletons représentent 42% des SNP, ce qui pourrait indiquer une population en expansion, ce qui est confirmé par le test DHEW avec 35% des gènes dans ce groupe qui rejettent le modèle neutre, et une valeur moyenne du D de Tajima pour ces gènes de $-1,77$. Le spectre de fréquence des allèles dérivés du groupe 9.2 semble le plus proche d'un modèle neutre.

3.4 Rôle de la recombinaison dans la diversité du complexe d'espèces *Xanthomonas axonopodis*

Sur l'ensemble des souches étudiées, l'étude de la généalogie a permis d'estimer le ratio r/m=0,97 signifiant que l'impact de la recombinaison sur le polymorphisme est presque égal à celui de la mutation **(tab. IV)**. Ces valeurs sont variables en fonction des groupes et comprises entre 0,61 dans le groupe 9.41 et 1,13 dans le groupe 9.2. La recombinaison se produit 3,5 fois moins souvent que la mutation au sein du complexe. Ces résultats sont confirmés par la décroissance du déséquilibre de liaison jusqu'au seuil $r^2 \simeq 0,2$ pour tous les groupes, ce qui indique que les groupes seraient plutôt clonaux **(fig. 23)**. Le rapport du taux de recombinaison sur le taux de mutation observé dans le jeu de données complet est de $R/\theta = 0,28$, avec des valeurs comprises entre 0,06 pour le groupe 9.6 et 0,26 pour le groupe 9.2. Les événements de recombinaisons ont été inférés le long de la généalogie clonale, et varient fortement **(fig. 24)**. ClonalFrameML a détecté 5631 imports de séquences homologues au sein des 73 isolats. Le groupe le plus recombinant est le groupe 9.2. Les imports les plus nombreux se localisent sur les branches précédent la divergence en groupes : sur la branche menant à l'espèce *X. phaseoli* (9.4), au groupe 9.1, sur la branche menant à l'espèce *X. citri* (groupes 9.5 et 9.6), aux branches menant aux groupes 9.5, 9.6 et à l'espèce *X. euvesicatoria* 9.2. On remarque aussi beaucoup d'imports dans l'espèce *X. phaseoli* sur les branches conduisant aux pathovars *X. phaseoli* pv. *phaseoli* et *X. phaseoli* pv. *manihotis*, et à l'intérieur de

TABLE IV – Estimation de la recombinaison et de la mutation
Les paramètres ont été estimés avec ClonalFrameML sur l'alignement du *core genome* Harvest.

Groupes	N	δ	nu	r/m	R/θ
9.2	13	101,8	0.045	1,13	0,25
9.3	2	na	na	na	na
9.41	15	86,2	0,063	0,61	0,11
9.5	21	130,4	0,061	0,66	0,082
9.6	22	224,9	0,064	0,91	0,063
Tous	73	82,12	0,042	0,97	0,28

N : nombre de séquences

δ: Taille moyenne des imports en bp

nu : divergence moyenne des imports

r/m : ratio de l'impact de la recombinaison par rapport à la mutation

R/θ : ratio des événements de recombinaison par rapport aux événements de mutation

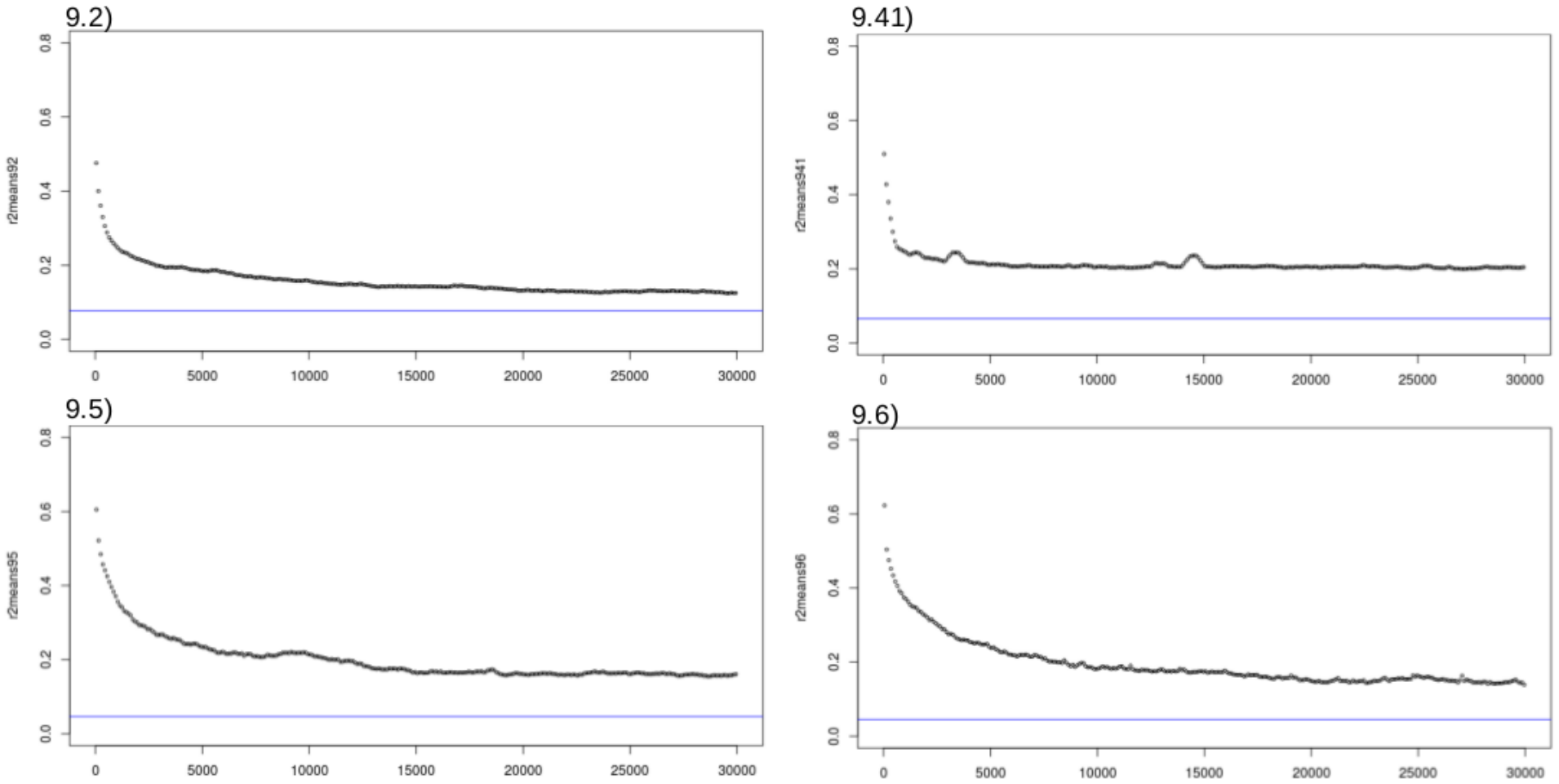


FIGURE 23 – Décroissance du déséquilibre de liaison.
Décroissance du déséquilibre de liaison par groupe en fonction de la distance en bp. Pour le groupe 9.2, le déséquilibre de liaison décroît jusqu'à 0,15 à partir de 15000 bp. Pour le groupe 9.41, le déséquilibre de liaison décroît rapidement jusqu'à 0,2 à partir de 5000 bp. Pour le groupe 9.5, le déséquilibre de liaison décroît jusqu'à 0,2 à partir de 15000 bp. Pour le groupe 9.6, le déséquilibre de liaison décroît jusqu'à 0,2 à partir de 10000 bp . La ligne de base bleue indique l'équilibre de liaison attendu (1/N) dans une population panmictique, ou N est la taille de l'échantillon.

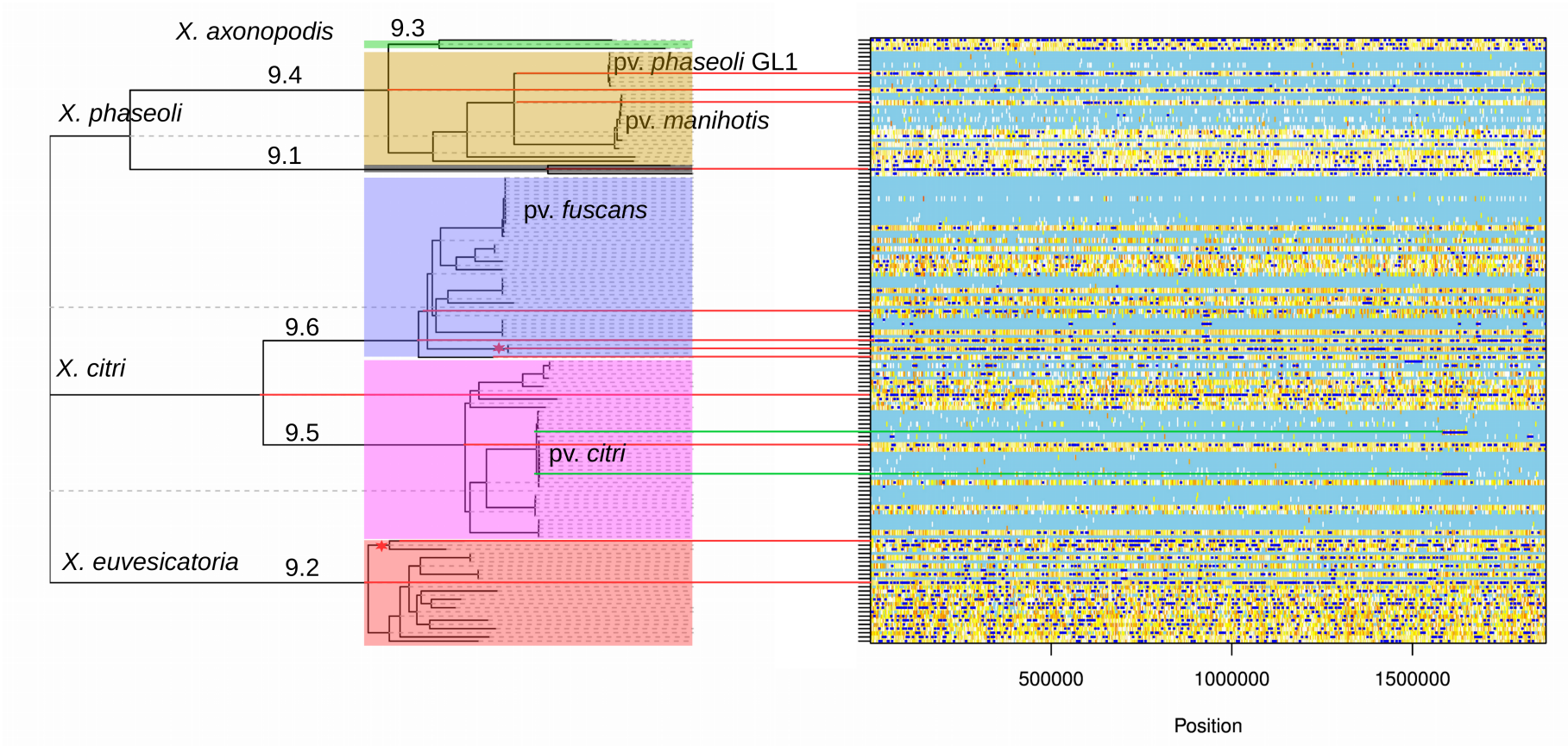


FIGURE 24 – Relation phylogénétique et sites de recombinaison, ou de substitution inférés par ClonalFrameML.
Les événements de recombinaison sont indiqués en bleu foncé, les substitutions en bleu clair. Les sites avec des substitutions non homoplasiques sont en blanc, avec une augmentation vers le rouge en fonction de l'homoplasie. Les branches portant le plus d'événements de recombinaison sont indiquées en rouge, les souches non pathogènes avec une étoile. Les deux branches vertes montrent les deux événements en miroir à l'intérieur du pathovar *X. c. pv. citri*.

ce clade sur la branche conduisant à *X. phaseoli* pv. *phaseoli*. La branche portant les deux souches non pathogènes du groupe 9.6, la branche portant la souche non pathogène du groupe 9.2 (CFBP 7920), et la branche portant la souche de *X. citri* CFBP 3132 montrent aussi beaucoup d'événements de recombinaison. Dans *X. c.* pv. *citri*, un large événement (d'environ 63 kb) marqué par une ligne horizontale bleue foncée, est en miroir sur les branches conduisant aux clades frères (branches vertes de la figure **24**). Cette symétrie peut être expliquée par des événements de substitutions qui se produisent sur la branche immédiatement ancestrale aux deux clades frères voir [Didelot and Wilson, 2015]. Cette région recombinante (gènes *atpAHFEB*) code pour des protéines liées à la production d'énergie. Les tailles des imports varient de 2 à 66 517 pb. La longueur moyenne des fragments recombinants (δ) est de 82 pb. La distribution des tailles des imports sur les branches de l'arbre correspondant aux groupes, et aux pathovars majeurs **(fig. 25)**, montre que le groupe 9.5 a reçu de plus gros imports que les autres groupes, ainsi que les pathovars *X. citri* pv. *mangiferaeindicae*, *X. citri* pv. *citri*, et les deux souches non pathogènes du groupe 9.6. La plus grosse séquence recombinante (66kb) a été importée au niveau de la branche portant deux souches du pathovar *citri* (306 et FDC1083), et le deuxième est de 62 kb au niveau du groupe 9.3.

3.5 Le flux de gènes entre les groupes au sein du complexe d'espèces *Xanthomonas axonopodis* permet de définir cinq groupes

Nous avons analysé le flux de gènes entre les groupes du complexe d'espèces *Xanthomonas axonopodis* en utilisant une approche dite de *chromosome painting* **(fig. 26)**. Nous observons cinq groupes de souches avec une forte intensité de flux de gènes le long de la diagonale. Une certaine sous structuration existe au sein de chacun des groupes ce qui est confirmé par les valeurs positives du test EW (voir paragraphe III.3.3). Dans le cas du groupe 9.41, cette sous structuration ne permet pas d'effacer le signal d'expansion indiqué par un D de Tajima négatif. Nous avons montré du déséquilibre de liaison et fineSTRUCTURE le prend en compte, contrairement aux autres méthodes comme l'ACP ; les deux analyses assignent pourtant le même nombre de populations. Les relations entre les groupes sont

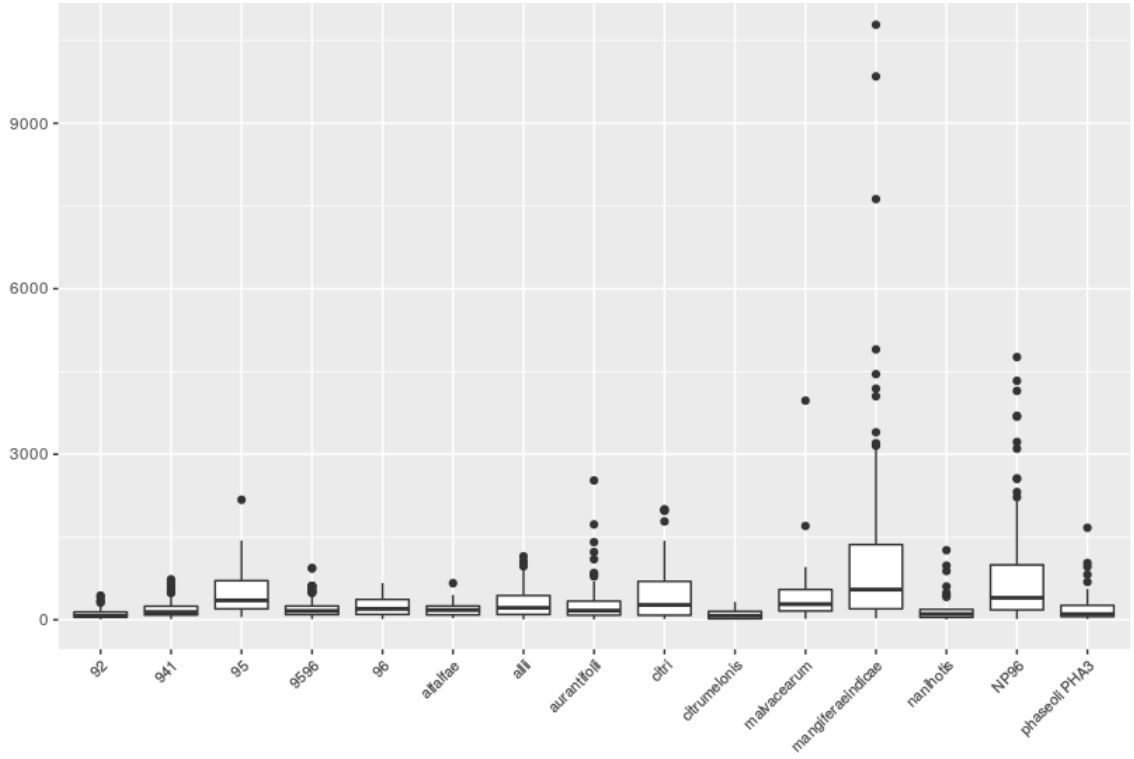


FIGURE 25 – Distribution des tailles des imports par branche le long de la phylogénie.
Les imports ne sont indiqués qu'aux branches supportant les pathovars majeurs et les groupes. La branche supportant le groupe 9.3 n'est pas représentée car elle comporte un seul import d'environ 62kb.

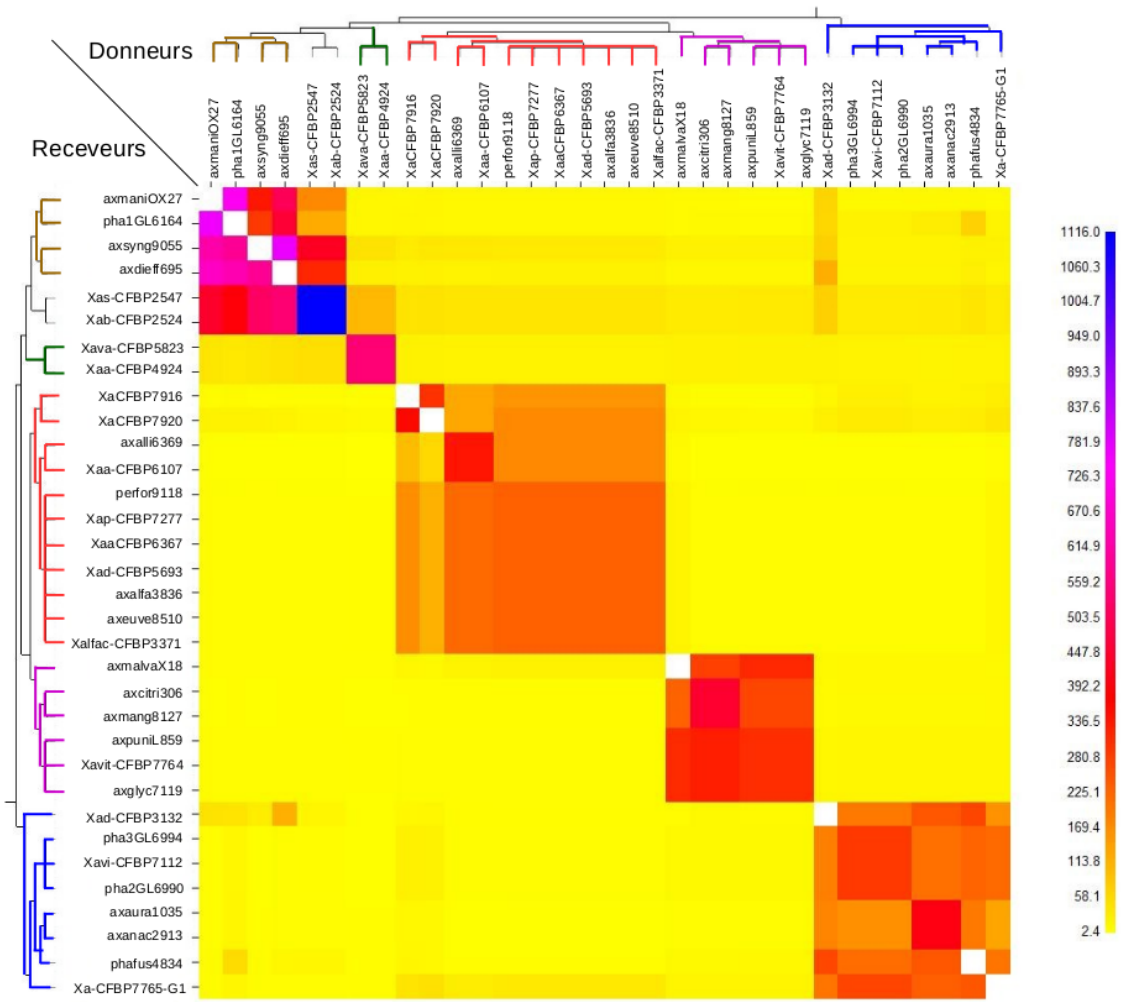


FIGURE 26 – Heatmap fineSTRUCTURE.

Heatmap de la matrice de "co-ancêtre" avec la structure des populations inférée par fineSTRUCTURE sur 33 souches représentatives. La couleur de chaque cellule indique le nombre de fragments importés d'un génome donneur (colonne), vers un génome receveur (en ligne). Les branches de l'arbre (montrant l'assignation par clustering à une population) sont colorées selon l'appartenance des souches au groupes (9.1 en noir, 9.2 en rouge, 9.3 en vert, 9.4 en marron, 9.5 en violet, 9.6 en bleu). L'intensité du flux est indiquée sur l'échelle à droite.

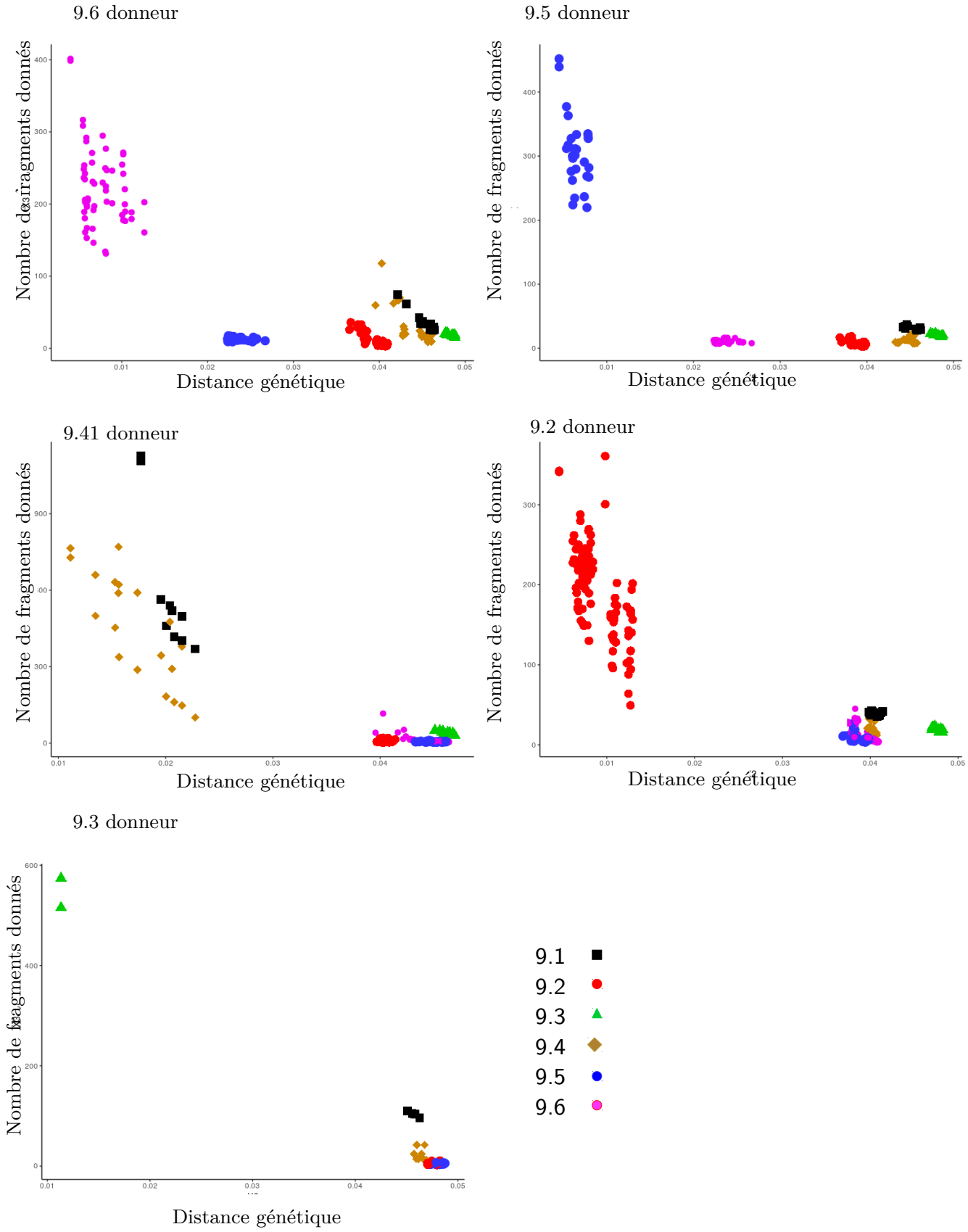


FIGURE 27 – Graphiques représentants le nombre de fragments donnés en fonction de la distance génétique.
Nombre de fragments donnés en fonction de la distance génétique entre les génomes du donneur et du receveur, calculé d'après la matrice fineSTRUCTURE.

différentes de l'arbre phylogénétique, le groupe 9.6 est isolé des autres, il n'est pas proche du groupe 9.5.

3.5.1 Le flux de gènes entre les groupes ne dépend pas que de la distance génétique entre génomes

Le nombre absolu d'événements montre qu'il existe de rares transferts entre les souches de *X. citri* du groupe 9.5 et 9.6, (environ 11 événements) contre 224 événements au sein des souches du groupe 9.6. La recombinaison est faible entre les groupes, elle est plus importante à l'intérieur de chaque groupe. Un flux de gènes important se produit entre les souches de *X. phaseoli* des groupes 9.4 et 9.1, ce qui nous conforte dans notre choix de les regrouper. Il existe quelques exemples de flux inter-groupes entre des souches qui partagent le même hôte, notamment entre (i) la souche CFBP 4885 (groupe 9.6) et la souche CFBP 6164 (groupe 9.41) pathogène sur les *Phaseolus* sp., (ii) entre les souches CFBP 3132 (groupe 9.6) et axdieff695 (CFBP 3133) (groupe 9.41) pathogène sur le *Dieffenbachia*. La matrice des co-ancêtres révèle une asymétrie de flux entre les populations. Les groupes 9.41 (et plus particulièrement les deux génomes du groupe 9.1), et 9.3, reçoivent de tous les autres groupes, et au sein de chacun de ces groupes le flux est aussi plus intense que chez les autres groupes. Ces résultats sont en accord avec le déséquilibre de liaison, qui est beaucoup plus faible au sein groupe 9.41. Ces deux groupes sont plus perméables aux flux de gènes. On peut aussi noter que les 3 souches non pathogènes sur leur hôte d'isolement (CFBP 7765, CFBP 7916, CFBP 7920) reçoivent plus des autres groupes que les souches de leur groupe. La souche CFBP 7923 n'a pas pu être testée dans l'analyse à cause de sa similarité avec CFBP 7765.

Il n'existe pas de relation directe entre le nombre de fragments donnés par un groupe et la distance génétique entre les génomes. Le flux de gènes n'est pas toujours fonction de la distance génétique **(fig. 27)**. On retrouve que le groupe 9.6 donne plus à certains génomes du groupe 9.4, 9.1, et 9.2 qu'au groupe 9.5 qui est plus proche génétiquement. Le nombre moyen de fragments donnés par 9.6 à 9.41 est de 26,4 ce qui est significativement supérieur ($pval = 3,915.10^{-6}$) à 13 pour le 9.5. Le groupe 9.3 donne aussi plus de fragments aux génomes du groupe 9.1 qu'aux

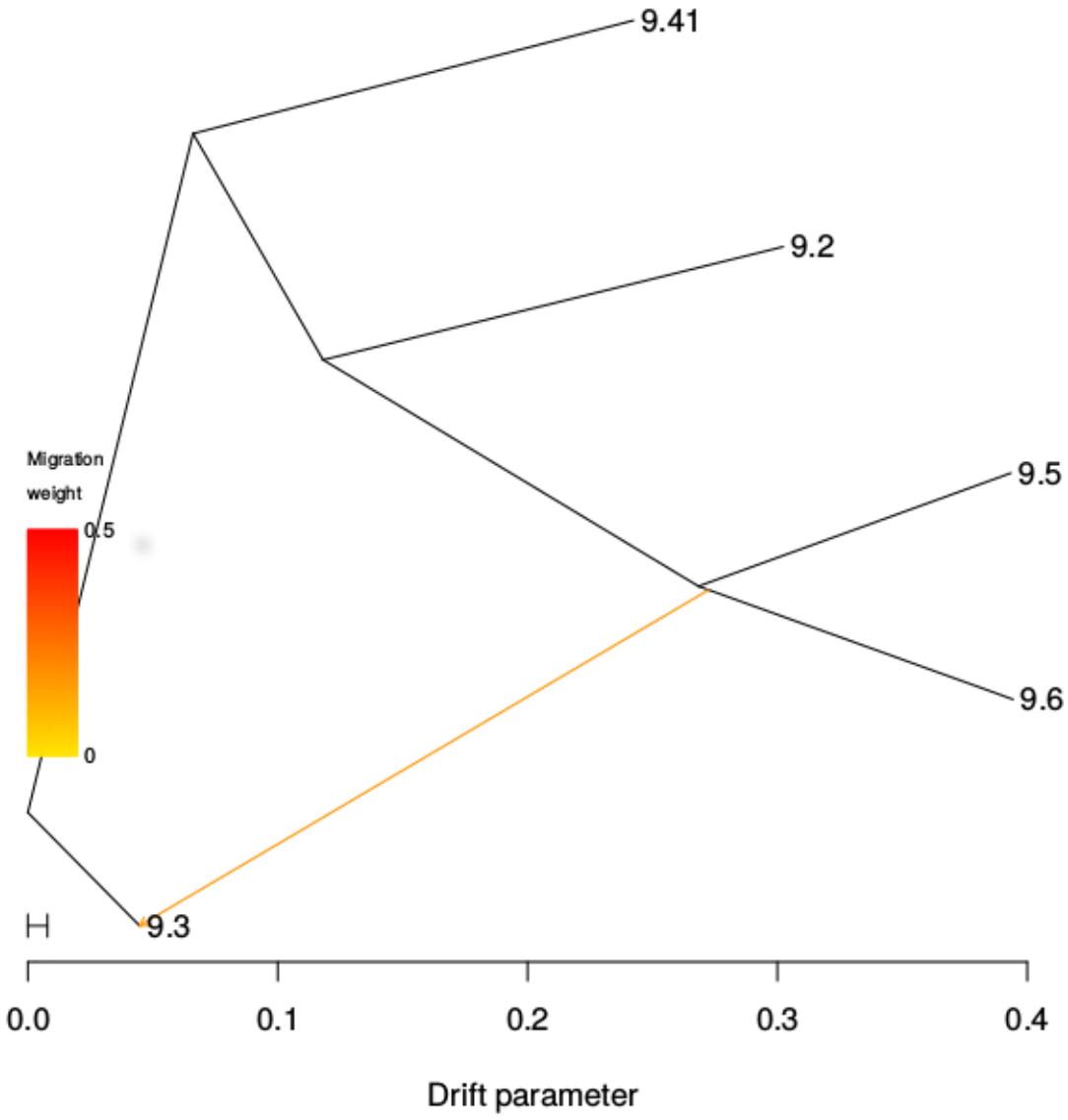


FIGURE 28 – Réseau phylogénétique des relations inférées entre les groupes par TreeMix. L'événement de migration montre une introgression du groupe 9.6 vers le groupe 9.3.

génomes du groupe 9.4 pour une distance génétique équivalente. Nous avons utilisé le graphe implémenté dans TreeMix pour confirmer le faible flux inter-groupes. Le modèle avec un seul événement de migration est le meilleur et explique 99,9% de la variance **(fig. 28)**. L'événement de migration est statistiquement significatif ($pval < 0,05$), le groupe 9.3 a reçu 17% du groupe 9.6.

3.6 Inférence de l'histoire de la divergence avec δaδi

Nous avons voulu analyser l'histoire évolutive de deux paires de groupes présentant des flux de gènes contrastés. Tout d'abord, il a été choisi d'inférer un scénario de divergence pour les groupes 9.5 et 9.6 qui appartiennent à la même espèce *X. citri*, mais qui forment deux populations bien distinctes avec un flux de gènes très faible entre elles. Quant aux groupes 9.5 et 9.2, ils n'appartiennent pas à la même espèce, mais présentent certaines souches perméables au flux de gènes inter-groupes (axmalvaX18, CFBP 7920), et des pathovars pathogènes sur les mêmes hôtes (*X. euvesicatoria* pv. *citrumelonis*, et *X. citri* pv. *citri* sur agrumes).

3.6.1 Situation 1 : les groupes 9.5 et 9.6 de la même espèce avec peu de flux de gènes

Choix du meilleur modèle Les meilleurs scénarios sur la base du critère du ΔAIC<10 pour les groupes 9.5 et 9.6 sont indiqués dans le **tableau V**. Six modèles ont été retenus dérivant des modèles de divergence avec migration (IM) et de contact secondaire (SC). Le meilleur modèle est un modèle de divergence avec migration et changement de taille de population (IMex2)**(fig. 30)** avec une probabilité de 0,79 avec l'Akaike pondéré (W_{AIC}). Les modèles suivant IM2Mex2 (dérivé du modèle précédent mais avec un flux de gènes hétérogène), SCex2 (contact secondaire avec changement démographique), et SC2M2Pex2 (SCex2 avec un flux de gènes hétérogène) ont une probabilité équivalente de 0,06. La dispersion des valeurs d'AIC **(fig. 29)** montre que le modèle hétérogène SC2M2Pex2 a une meilleure convergence que le même modèle homogène SCex2.

Estimation des paramètres Les paramètres estimés pour les 4 meilleurs modèles sont indiqués dans le **tableau VI**. Les valeurs sont toutes du même ordre

TABLE V – Meilleurs modéles pour 9.5 et 9.6

Probabilités logarithmiques (Log-Likelihood) et critère d'information d'Akaike (AIC) des 6 meilleurs modèles (ΔAIC <10) testés avec $\delta a \delta i$ pour les groupes 9.5 et 9.6.

Modèles	k	**Log-Likelihood**	LRT	AIC	ΔAIC
IMex2	9	-254,85		527,71	0
IM2Mex2	10	-256,42		532,84	5,13
SCex2	10	-256,43		532,84	5,13
SC2M2Pex2	12	-256,42	(SC2M2Pex2/SC2Mex2)=0,02 dll 1	532,84	5,13
SC2Mex2	11	-256,43		534,87	7 ,1
IM2M2Pex2	11	-257,60		537,21	9,5

k Nombre de paramètres du modèle

Log-Likelihood estimée parmi 23 analyses indépendantes

AIC Critère d'information d'Akaike

ΔAIC Différence d'AIC entre le modèle i et le meilleur modèle (IMex2)

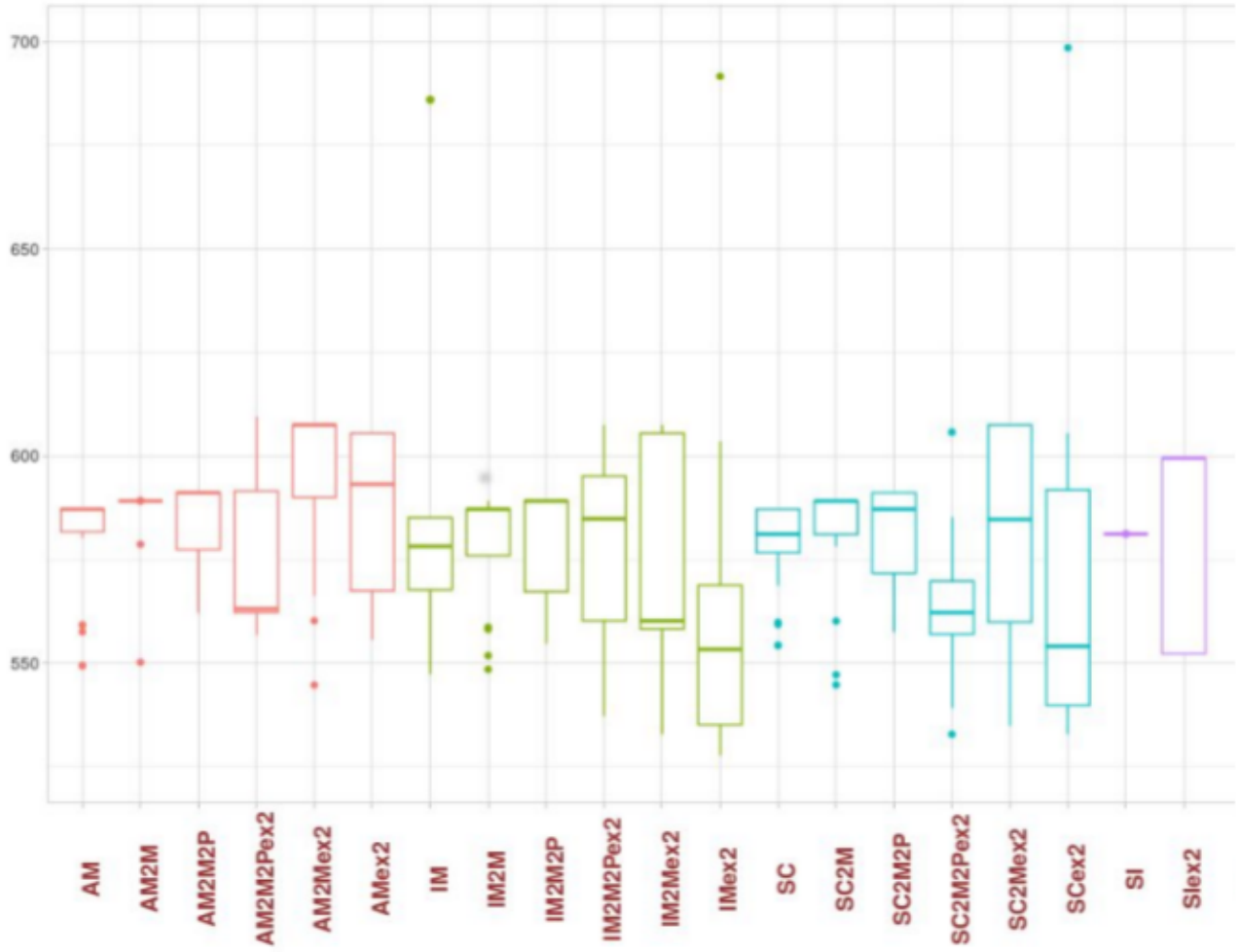


FIGURE 29 – Dispersion des valeurs d'AIC pour les groupes 9.5 et 9.6.
Dispersion des valeurs d'AIC pour 23 analyses indépendantes de $\delta a \delta i$ des groupes 9.5 et 9.6. En abscisses les 20 modèles testés

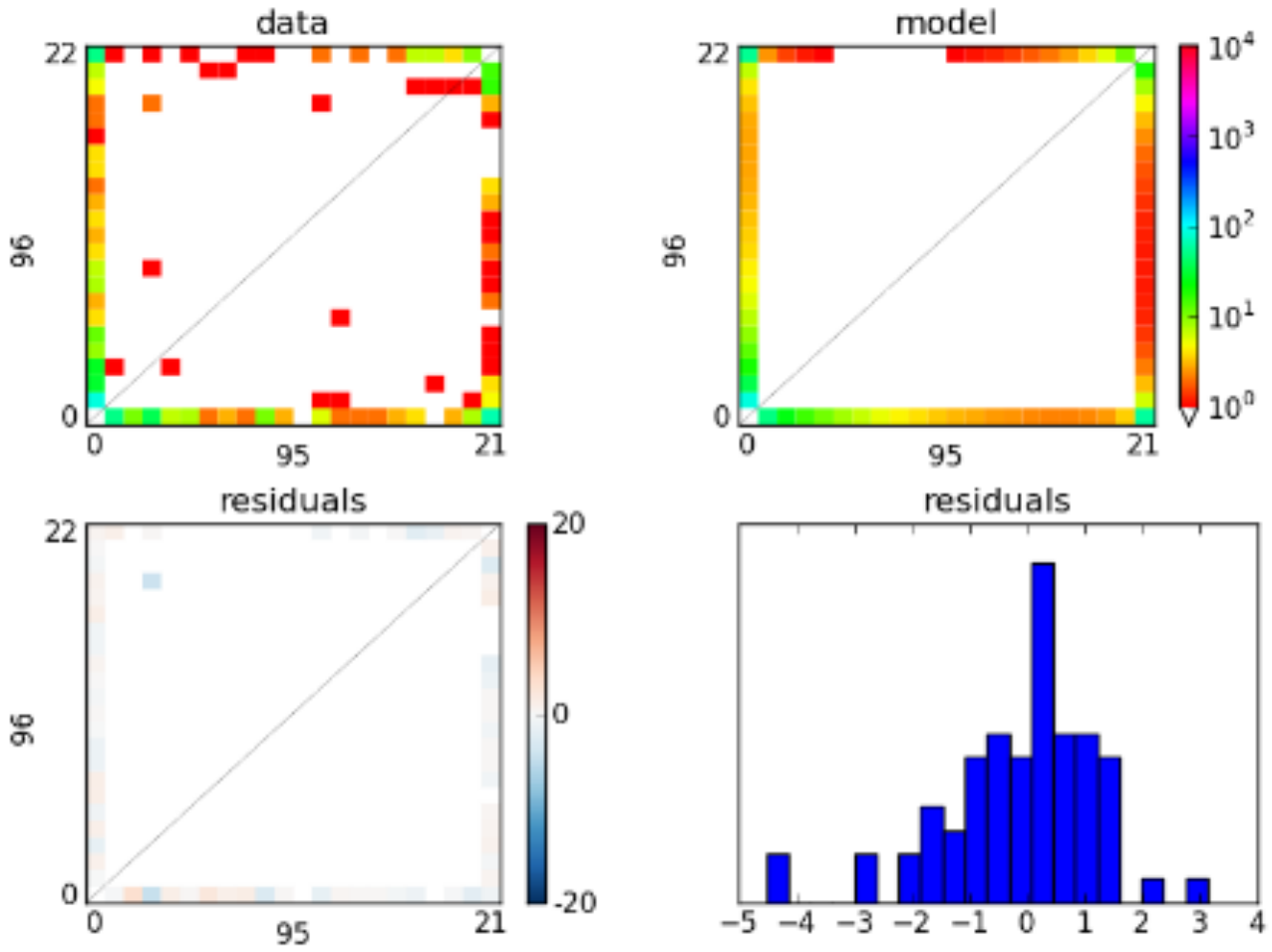


FIGURE 30 – Spectres de fréquence de l'allèle dérivé pour 9.5 et 9.6.
La heatmap en haut à gauche, correspond au spectre de fréquence du groupe 9.6 sur l'axe vertical, et du groupe 9.5 en abscisse. La heatmap de droite correspond au spectre de fréquence de l'allèle dérivé inféré par δaδi selon le modèle IMex2. Les graphiques en bas correspondent à la distribution des résidus ($résidus = modèle\frac{data}{\sqrt{modèle}}$) ; en bleu, le modèle ne prévoit pas assez d'allèles, en rouge le modèle infère trop d'allèles

TABLE VI – Paramètres inférés pour la démographie des groupes 9.5 et 9.6 selon les différents modèles.

Modèle	**θ**	**Nref**	**n1**	**n2**	**n'1**	**n'2**	**m1<2**	**m2<1**	**Ts**	**Tsc**	**TD**	**Te**	**P1**	**P2**	**P**	**O**
IMex2	25564,94	278181,00	0,293 ∓ 0,0246	0,292 ∓ 0,044	0,776 ∓ 0,0779	0,337 ∓ 0,0251	0,052 ∓ 0,0036	0,056 ∓ 0,0044	0,944 ∓ 0,1061		0,94	0,20 ∓ 0,0015				0,83
IM2Mex2	25047,33	272548,67	0,301 ∓ 0,0204	0,300 ∓ 0,0203	0,795 ∓ 0,0688	0,345 ∓ 0,0114	0,055 ∓ 0,0079	0,059 ∓ 0,0080	0,984 ∓ 0,0798		0,98	0,21 ∓ 0,0138			0,95	0,83
SCex2	28803,22	313417,81	0,256 ∓ 0,0061	0,254 ∓ 0,0063	0,659 ∓ 0,0412	0,286 ∓ 0,0131	0,069 ∓ 0,0033	0,075 ∓ 0,0051	0,654 ∓ 8,5E-5	1,97E-22 ∓ 3,7E-14	0,65	0,17 ∓ 0,0124				0,83
SC2M2Pex2	24688,92	268648,74	0,336 ∓ 0,0233	0,312 ∓ 0,0207	0,816 ∓ 0,0748	0,350 ∓ 0,0248	0,158 ∓ 0,0415	0,724 ∓ 0,138	0,959 ∓ 0,0856	8,26E-7 ∓ 6,53E-5	0,96	0,19 ∓ 0,0146	0,43 ∓ 0,085	0,17 ∓ 0,017		0,82

L = 2297510 (Nombre de SNP retenu par dadi x longueur du core génome)/Nombre de SNP total

θ Thêta pour la population ancestrale avant la divergence

Nref Taille efficace de la population ancestrale **Nref = θ/2Lμ**

n1 Taille de la population 9.6 après la divergence

n2 Taille de la population 9.5 après la divergence

n'1 Taille actuelle de la population 9.6

n'2 Taille actuelle de la population 9.5

m1<2 Migration neutre de la population 2 (9.5) vers la population 1 (9.6)

m2<1 Migration neutre de la population 1 (9.6) vers de la population 2 (9.5)

P1 La proportion de génome évoluant de façon neutre dans la population 1 (9.6)

P2 La proportion de génome évoluant de façon neutre dans la population 2 (9.5)

O Proportion d'allèles correctement orientés

Ts Temps entre la divergence et Tsc en unités de Nref générations

TD Temps depuis la divergence en unités de Nref générations

Tsc Temps entre le début de la migration et le présent en unités Nref

Te Temps depuis le changement démographique

P La proportion de génome évoluant de façon neutre

TABLE VII – Meilleurs modèles pour 9.5 et 9.2

Probabilités logarithmiques (Log-Likelihood) et critère d'information d'Akaike (AIC) des 6 meilleurs modèles (ΔAIC<10) testés avec $\delta a \delta i$ pour les groupes 9.5 et 9.2. k =Nombre de paramètres du modèle , Log-Likelihood estimée parmi 20 analyses indépendantes, ΔAIC=Différence d'AIC entre le modèle i et IMex2 le meilleur modèle, LRT=Likelihood-ratio test, * si le test est significatif à p<0,05.

Modèles	k	Log-Likelihood	LRT	AIC	ΔAIC
SCex2	10	-234,16	(SCex2/SC)=8,22* dll3	488,32	0
IM2Mex2	10	-234,25	(IM2Mex2/IM)=10,8* dll4	488,50	0,17
SC2M2Pex2	12	-232,47	(SCex2/SC)=8,22* dll5	488,95	0,62
IM2M2P	8	-236,84		489,68	1,35
SC2Mex2	11	-234,18		490,36	2,04
SC	7	-238,27		490,54	2,22
IM	6	-239,65		491,31	2,99
SC2M	8	-237,92		491,86	3,54
AM	7	-238,95		491,90	3,58
IM2M	7	-239,27		492,54	4,22
IM2M2Pex2	11	-235,31		492,62	4,3
IMex2	9	-237,66		493,32	5

de grandeur quel que soit le modèle. La taille efficace ancestrale de chacune des populations 9.6 et 9.5 est semblable. Elle est de l'ordre de 0,3 soit 81507 individus dans la population 9.6 et 81229 dans la population 9.5. Seule la taille actuelle de la population 9.6 a augmenté d'un facteur 2,6 soit 215868 individus. Le taux de migration efficace est symétrique entre les deux populations, il est de l'ordre de 0,05. La fraction d'individus à chaque génération de la population 9.6 qui est constituée de nouveaux migrants de la population 9.5 par génération est égale à $f = 1,87.10^{-7}$ ($f = \frac{m1>2}{Nref}$), et 2.10^{-7} de la population 9.6 vers 9.5 par génération. Le temps de divergence est d'environ 0,94 ce qui correspond si on utilise un temps de génération de 0,003, à une divergence il y a 951 ans ($TD \times Nref \times 0,003$). Le changement démographique se serait produit il y a 167 années. Le temps depuis le contact secondaire pour le modèle SC2M2Pex2 est estimé à moins d'une demi-journée ce qui explique la difficulté pour le modèle d'être significatif.

3.6.2 Situation 2, les groupes 9.5 et 9.2 deux espèces avec du flux inter-groupes

Choix du meilleur modèle Les meilleurs scénarios sur la base du critère du ΔAIC<5 pour les groupes 9.5 et 9.2 sont indiqués dans le **tableau VII**. Douze modèles ont été retenus dérivant des modèles de contact secondaire (SC) et de divergence avec migration (IM). Trois modèles ont des probabilités presque équivalentes, SCex2, IM2Mex2 et SC2M2Pex2. Le meilleur modèle est un modèle de contact secondaire et changement de taille de population (SCex2) avec une probabilité de 0,21 avec le Akaike pondéré (W_{AIC}). Le modèle suivant est un modèle hétérogène de divergence avec flux de gènes et changement de taille de population (IM2Mex2) avec une probabilité de 0,19, puis on retrouve un modèle de contact secondaire hétérogène avec changement démographique (SC2M2Pex2) avec un probabilité de 0,15. La dispersion des valeurs d'AIC pour les 20 répétitions des 20 modèles **(fig. 31)**, les deux médianes les plus basses correspondent aux modèles (SC et SCex2)

Estimation des paramètres Les paramètres des 3 meilleurs modèles sont indiqués dans le **tableau VIII**. La taille ancestrale de chacune des populations 9.5 et 9.2 est de 3581745 et 29402 individus respectivement. La taille actuelle de chacune

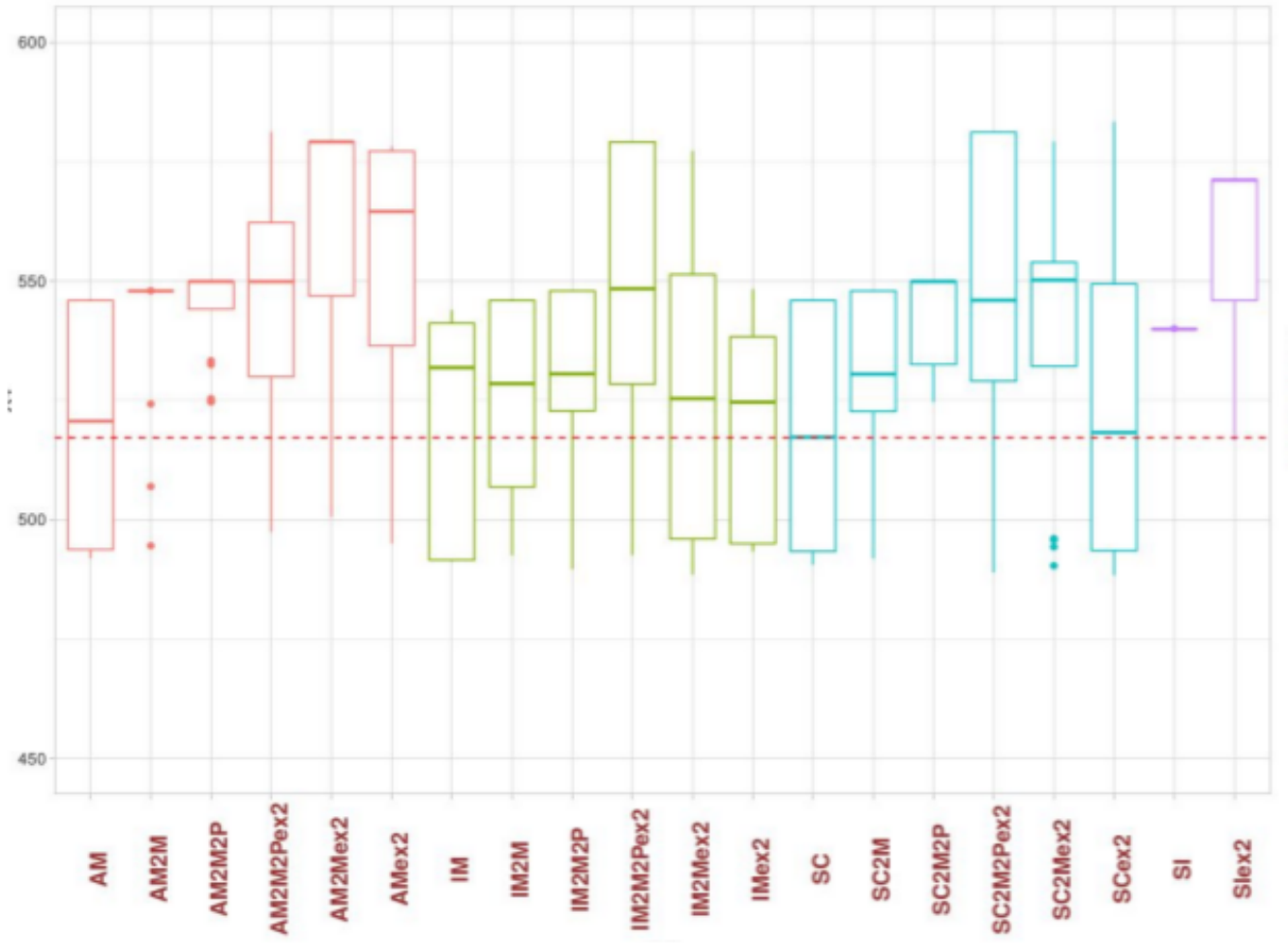


FIGURE 31 – Dispersion des valeurs d'AIC pour les groupes 9.5 et 9.2.
Dispersion des valeurs d'AIC pour 20 analyses indépendantes de $\delta a \delta i$ des groupes 9.5 et 9.2. En abscisses les 20 modèles testés

TABLE VIII – Paramètres inférés pour la démographie des groupes 9.5 et 9.2 selon les différents modèles

Modèle	θ	Nref	N1	N2	N'1	N'2	m1<2	m2<1	Ts	Tsc	TD	Te	P1	P2	P	O
SCex2	6302.64	53458,89	6,70	0,550	0,019	0,041	0,679	0,915	17,12	$3{,}8.10^{-13}$	17,27	0,149				0,998
			358175	**29402**	**1016**	**2192**	**$1{,}2.10^{-5}$**	**$1{,}7.10^{-5}$**	**2746**	**$6{,}1.10^{-11}$**	**2770**	**24**				
IM2Mex2	19009.93	161241,9	16,69	6,40	0,0184	0,0158	0,772	2,41	15,95		16,17	0,216			0,998	0,997
			2691127	**1031948**	**2967**	**2548**			**7715**		**7820**	**105**				
SC2M2Pex2	7217.98	60374,59	18,75	6,69	0,0410	0,0432	0,517	0,845	15,62	$3{,}3.10^{-7}$	15,98	0,363	0,76	0,98		0,989
			1132024	**403906**	**2475**	**2608**			**2829**	**$6{,}1.10^{-5}$**	**2894**	**65**				

En gras, les paramètres sont indiqués en nombre d'individus pour les tailles, en fraction de migrant par génération pour la migration et en années pour les temps. L = 2947423,92

θ Thêta pour la population ancestrale avant la divergence

Nref Taille efficace de la population ancestrale

N1 Taille de la population 9.5 après la divergence

N2 Taille de la population 9.2 après la divergence

N'1 Taille actuelle de la population 9.5

N'2 Taille actuelle de la population 9.2

m1<2 Migration neutre de la population 2 (9.2) vers la population 1 (9.5)

P1 La proportion de génome évoluant de façon neutre dans la population 1 (9.5)

Ts Temps entre la divergence

TD Temps depuis la divergence en unités de Nref générations

Tsc Temps entre le début de la migration et le présent en unités Nref

Te Temps depuis le changement démographique

P La proportion de génome évoluant de façon neutre

O Proportion d'allèles correctement orientés

m2<1 Migration neutre de la population 1 (9.5) vers de la population 2 (9.2)

P2 La proportion de génome évoluant de façon neutre dans la population 2 (9.2)

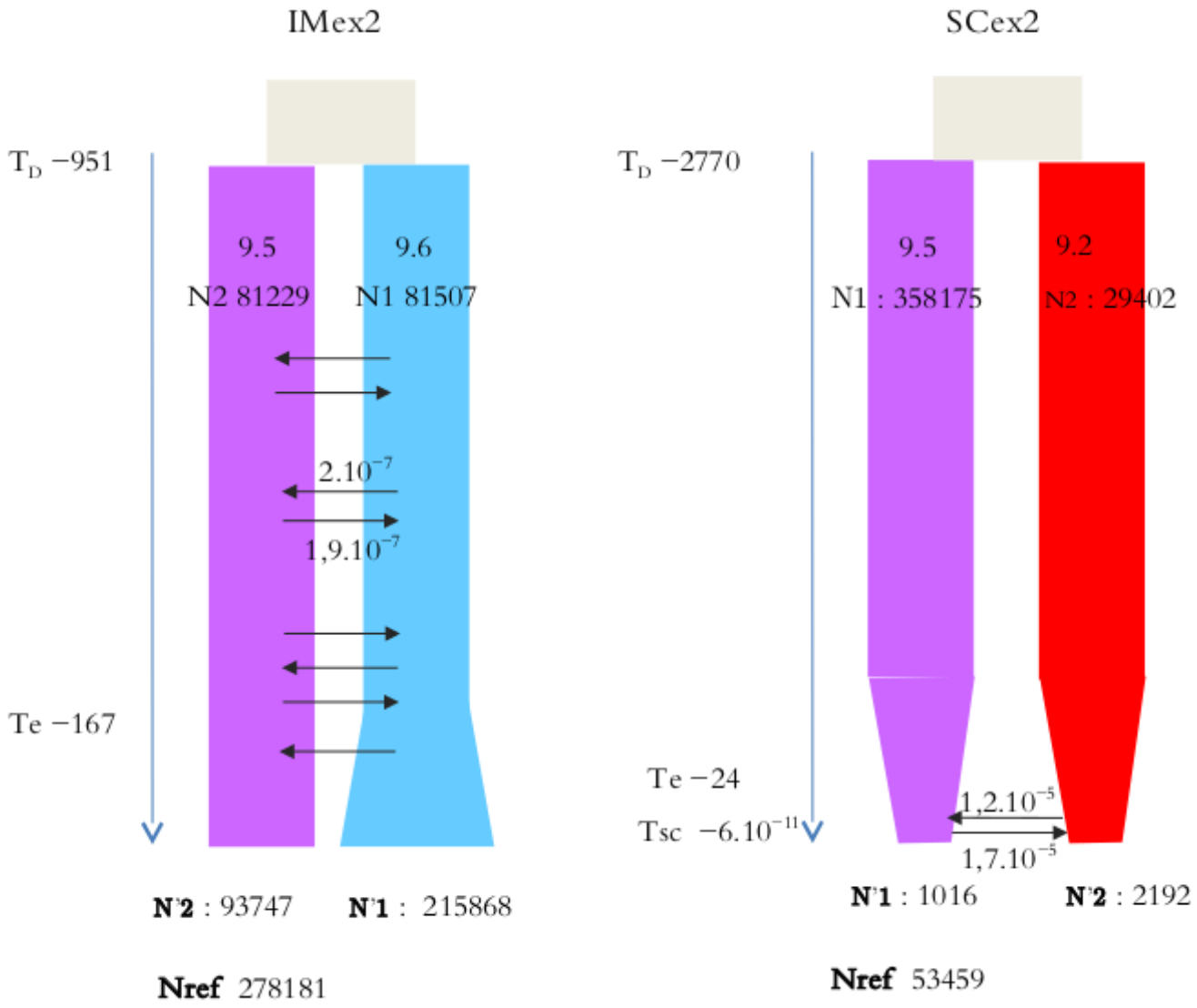


FIGURE 32 – Représentation des paramètres des meilleurs modèles.
A gauche, les paramètres inférés pour les groupes 9.5 et 9.6 selon le modèle IMex2. A droite, les paramètres inférés pour les groupes 9.5 et 9.2 selon le modèle SCex2. Les temps (T) sont indiqués en années, la migration est la fraction d'individus qui sont des migrants venant de l'autre groupe par génération.

des populations est réduite d'un facteur 350 (10156 individus) pour la population 9.5 et d'un facteur 13 (2192 individus) pour la population 9.2. La migration représente une fraction $f = 1,2.10^{-5}$ individus migrants de la population 9.2 vers la population 9.5 par génération, et $f = 1,7.10^{-5}$ individus qui sont des migrants de la population 9.5 vers la population 9.2 par génération. Le temps de divergence est estimé à 2770 années. Le temps de contact secondaire est estimé à 2.10^{-8} jour. Le temps depuis le changement démographique est de 24 années.

3.6.3 Le modèle à trois populations confirme un flux de gènes plus important entre 9.5 et 9.2 qu'entre 9.6 et 9.5

Les deux analyses précédentes nous donnent des résultats qui ne sont pas comparables entre eux car ils n'ont pas été établis sur les mêmes jeux de données, ce qui fait que la population 9.5 dans l'analyse 9.5/9.6 a une taille stable, alors que dans l'analyse 9.5/9.2 cette population montre une forte réduction de sa taille **(fig. 32)**. Le modèle IMSC avec le meilleur AIC confirme que la fraction de migrants entre les groupes 9.5 et 9.6 (valeurs comprises entre 10^{-13} et 10^{-17}) est beaucoup moins importante qu'entre les groupes 9.5 et 9.2, ou 9.6 et 9.2 (valeurs comprises entre 10^{-6} et 10^{-7}) **(tab. IX)**.

3.7 Recherche de gènes liés à l'adaptation

3.7.1 Les régions à fort F_{ST} ne semblent pas sous sélection positive

La moyenne des valeurs de F_{ST} calculée sur des fenêtres de 1000bp est de $0,41 \pm -0,15$. Les distributions des valeurs de F_{ST} et D de Tajima sont indiquées sur l'alignement du *core genome* des groupes 9.5 et 9.6. On ne détecte pas de régions fortement différenciées qui seraient colocalisées avec des valeurs de D de Tajima négatives indiquant de la sélection positive **(fig. 33)**.

3.7.2 Les groupes 9.5 et 9.6 sont très différenciés

L'ACP distcrimine les deux groupes 9.5 et 9.6 **(fig. 34a)**. La distribution des probabilités associées aux SNP montre énormément de SNP avec de faibles valeurs **(fig. 34b)** donc les SNP significativement différents. Il y a un grand nombre de

TABLE IX – Paramètres inférés pour la démographie des groupes 9.2, 9.5 et 9.6 selon IMSC

	Na	N1	N2	N3	m1<2	m2<1	m3<1	m1<3	m2<3	m3<2	T1	T2	TSC	O
Paramètres	10,07	0,375	0,203	0,353	1,68	0,180	0,169	0,129	$7{,}57.10^{-12}$	$1{,}62.10^{-7}$	10,80	5,98	0,72	0,84
					$3{,}9.10^{-6}$	**$4{,}3.10^{-7}$**	**4.10^{-7}**	**3.10^{-7}**	**$1{,}8.10^{-17}$**	**$3{,}8.10^{-13}$**	**13698**	**7585**	**913**	

En gras les valeurs sont en nombre d'individus qui sont des migrants par génération pour la migration, et en années pour les temps.

L=764341, Nref=422778 individus.

Na : Taille de la population ancestrale

N1 : Taille de la population 9.2 en Nref

N2: Taille de la population 9.5 en Nref

N3: Taille de la population 9.6 en Nref

m1<2 Migration neutre de la population 2 (9.5) vers la population 1 (9.2)

m2<1 Migration neutre de la population 1 (9.2) vers la population 2 (9.5)

m1<3 Migration neutre de la population 3 (9.6) vers la population 1 (9.2)

m3<1 Migration neutre de la population 1 (9.2) vers la population 3 (9.6)

m3<2 Migration neutre de la population 2 (9.5) vers la population 3 (9.6)

m2<3 Migration neutre de la population 3 (9.6) vers la population 2 (9.5)

T1 Temps entre la divergence et T2 en Nref générations

T2 Temps entre la divergence et TSC en Nref générations

TSC Temps depuis le début de la migration par contact secondaire en Nref générations

O Proportion d'allèles correctement orientés

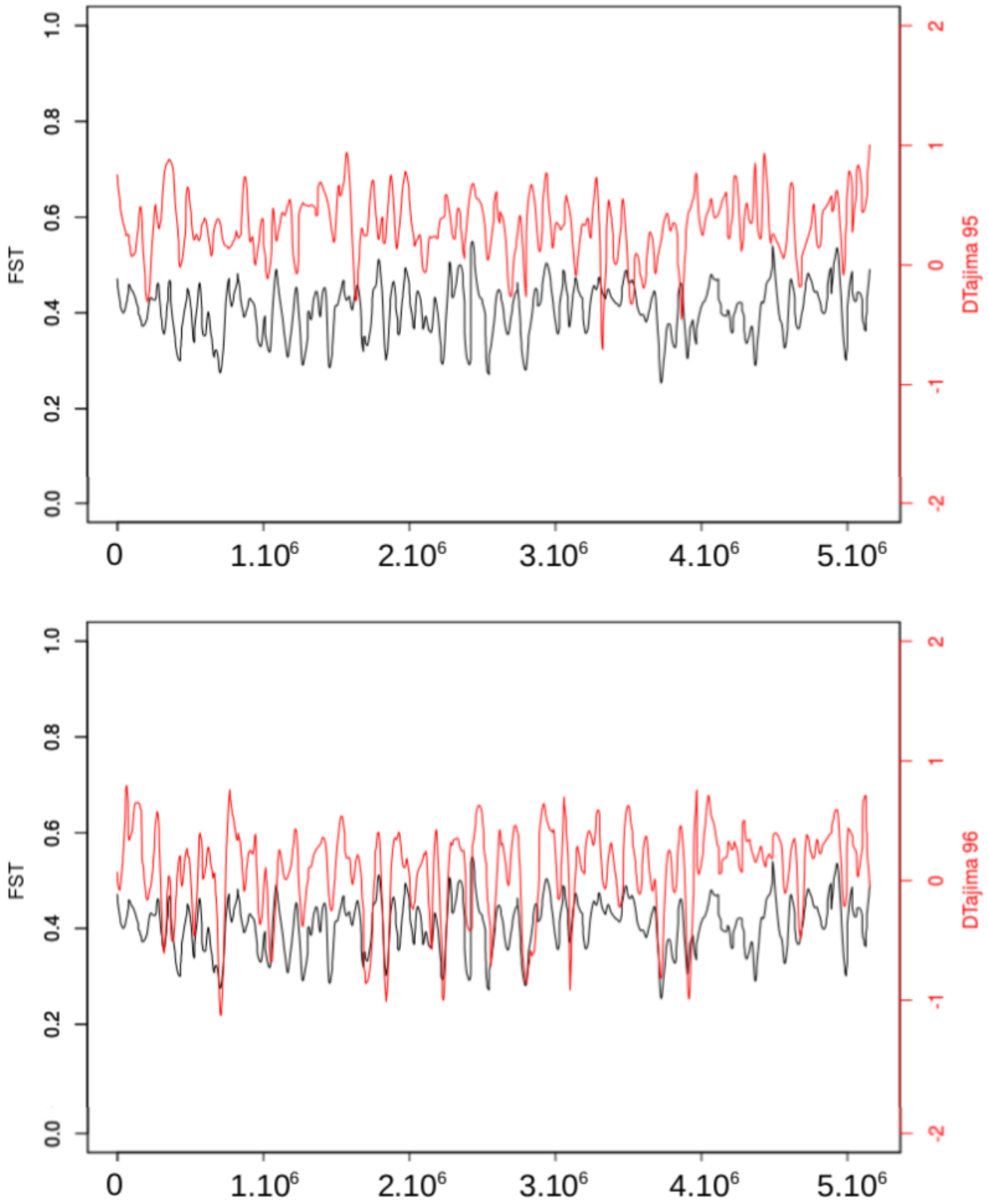


FIGURE 33 – La sélection sur le génome.
La distribution de la F_{ST} calculée dans une fenêtre glissante de 1000bp le long du *core genome* des groupes 9.5 et 9.6, avec les valeurs du D de Tajima pour les mêmes fenêtres. Les valeurs F_{ST} sont indiquées en noires. Les valeurs du D de Tajima sont indiquées en rouge, en haut les valeurs pour le groupe 9.5, en bas les valeurs pour le groupe 9.6.

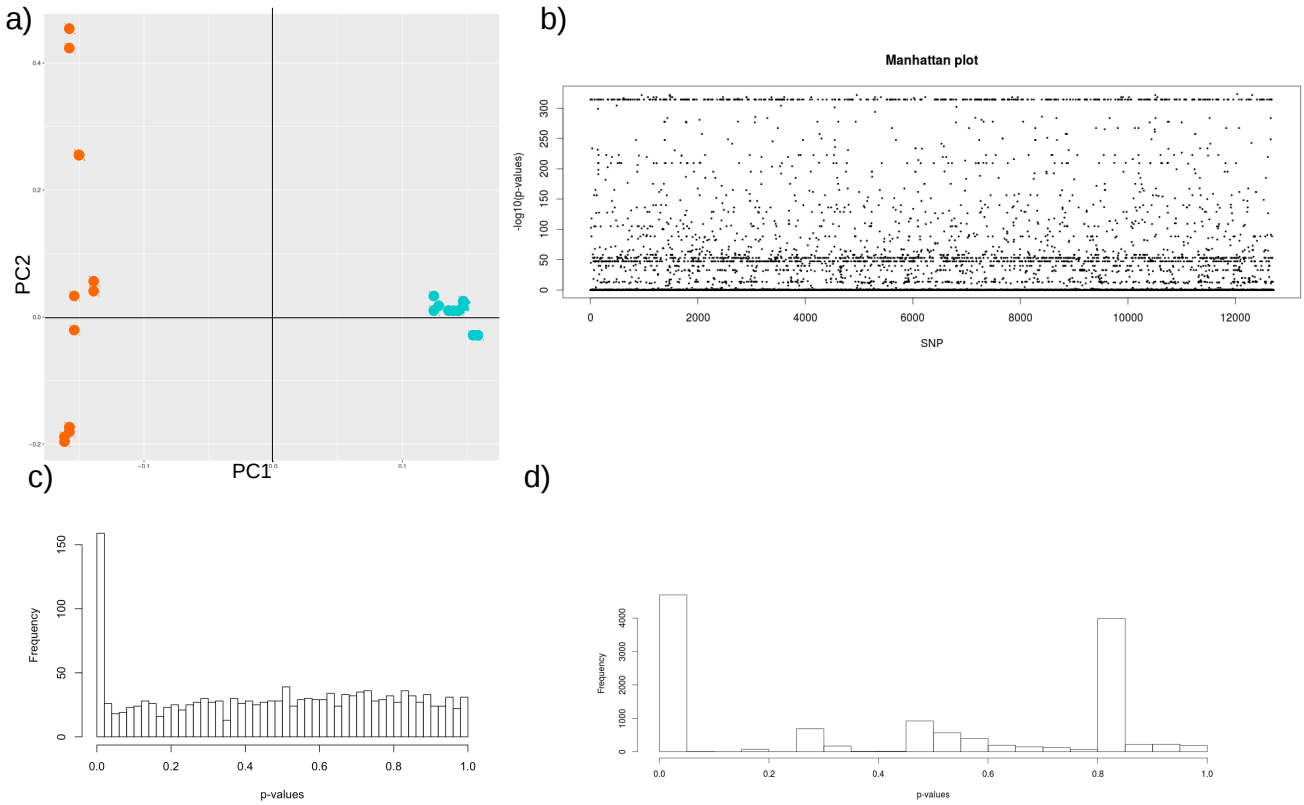


FIGURE 34 – Résultats de l'analyse PCAdapt
a) Structure des populations sur les deux premiers axes de l'ACP b) Distribution des valeurs de P-values des SNP c) Histogramme des fréquences de P-values attendues d) Histogramme des fréquences de P-values de notre jeu de données

marqueurs, 16 419 sur les 230 850 SNP, qui sont fixés entre les deux populations. D'autre part, 77% des SNP sont en faibles fréquences (<0,05). L'histogramme des probabilités devrait indiquer que la plupart suivent une distribution uniforme avec un excès de petites probabilités qui indiquent les *outliers* comme sur cet exemple **(fig. 34c)**. La distribution des probabilités **(fig. 34d)** de nos groupes 9.5 et 9.6 est bimodale, il y a bien un excès de petites valeurs de probabilités, mais aussi une forte fréquence de probabilités autour de 0,8. Cette méthodologie ne peut pas donner de résultats dans notre cas, car un grand nombre de SNP est fixé entre les deux populations. On ne peut donc pas identifier des *outliers* pour la différentiation génétique.

4 Discussion

Notre étude a consisté en l'analyse génomique de 73 souches du complexe d'espèces bactériennes *Xanthomonas axonopodis* afin de déterminer quelles étaient les forces évolutives déterminant sa structure génétique. Nous avons confirmé une structure en 5 groupes correspondant pour la majorité aux groupes de Rademaker, mais cependant sans pouvoir différencier les groupes 9.4 et 9.1. Nos résultats indiquent que la capacité d'échange de gènes par recombinaison homologue entre groupes ne semble pas suivre la proximité génétique de ces groupes. D'autres facteurs, sans doutes sélectifs, semblent être à l'œuvre. Dans les paragraphes suivants, nous tenterons tout d'abord d'identifier les éventuels *biais* d'analyse. En effet, la majeure partie des méthodes utilisées sont conçues pour des organismes eucaryotes à reproduction sexuée. Ensuite nous discuterons du flux de gènes selon les groupes, et des barrières à la recombinaison, et du statut des groupes 9.5 et 9.6 en tenant compte à la fois de leur divergence récente et de leur isolement reproductif important.

4.1 *Biais* des méthodes

Nous avons fait l'hypothèse que nos populations sont en équilibre de liaison, pour pouvoir utiliser fastSTUCTURE. Un tel équilibre de liaison est attendu dans les grandes populations panmictiques à l'équilibre démographique, deux condi-

tions insuffisamment documentées chez les bactéries. En effet les populations bactériennes n'ont pas toujours une taille constante et infinie. *A fortiori* les bactéries phytopathogènes subissent des réductions drastiques et régulières de taille de populations, notamment pendant l'hiver où les hôtes se raréfient et/ou perdent leurs feuilles [Tellier and Lemaire, 2014]. Ainsi les populations d'organismes pathogènes subissent régulièrement des goulets d'étranglement (*bottlenecks*) suivis d'expansions démographiques [Martiny et al., 2006]. De manière générale, les inférences de changements démographiques sont sensibles à la recombinaison, la sélection, et aux *biais* d'échantillonnage et doivent donc être effectuées avec précaution [Lapierre et al., 2016]. Les statistiques descriptives et certains indices comme le D de Tajima sont utilisés pour détecter d'éventuelles déviations au modèle neutre pouvant être causées par ces *biais*. L'utilisation de tests composés tels que DH, HEW et DHEW indiquent toutefois que les populations sont plutôt à l'équilibre (excepté pour le groupe 9.41, voir plus bas).

Cependant l'inférence démographique effectuée avec δaδi semble privilégier les modèles avec changement démographique. Ces changements sont inférés en tenant compte de la divergence et du flux de gènes avec d'autres populations, contrairement au test de Tajima effectué sur les échantillons de population seuls. Par exemple, pour la paire 9.5-9.2, une situation de contact secondaire suivi de flux de gènes récents aura pour effet d'apporter de nouvelles mutations en basses fréquences dans chacune des populations, ce qui a pour effet de diminuer le D de Tajima. En revanche, une réduction de taille de population aura pour effet d'augmenter D. Ainsi la distribution du D de Tajima apparemment centrée sur 0 reflèterait plus une compensation d'effets antagonistes qu'un réel équilibre démographique des populations.

Compte tenu de ce *biais*, nous avons comparé les résultats issus de différentes méthodes reposant sur des hypothèses totalement différentes, comme des hypothèses probabilistes pour la phylogénie, des statistiques descriptives avec l'ACP, des analyses bayésiennes avec fastSTRUCTURE et fineSTRUCTURE. Le modèle de fineSTRUCTURE tient compte du déséquilibre de liaison entre SNP contrairement à fastSTRUCTURE, cependant dans les deux cas cinq groupes sont assignés. Une récente étude sur basées sur des génomes humains a démontré que la méthode du "chromosome painting" avec fineSTRUCTURE est capable de détecter des

structures de populations plus subtiles, plus récentes par rapport aux structures détectées avec STRUCTURE ou l'ACP [Lawson et al., 2012]. C'est peut-être pour cette raison que l'organisation des groupes les uns par rapport aux autres est différente entre STRUCTURE et fineSTRUCTURE. Ce dernier isole le groupe 9.6 des autres groupes.

Cas du groupe 9.4.1 Dans ce groupe on trouve beaucoup de singletons qui conduisent à un D de Tajima négatif, et une importante sous-structuration indiquée par fineSTRUCTURE et EW. Ces deux phénomènes produisant des effets contraires, on peut supposer que l'expansion est nettement plus marquée que ce qu'indique le D de Tajima. Le groupe 9.41 comprend des agents pathogènes de cultures majeures, comme le manioc et le haricot, son expansion pourrait s'expliquer par le développement intensif de ces cultures. Le manioc est la quatrième matière première la plus importante après le riz, le blé et le maïs, et la croissance annuelle de la production mondiale de manioc dans la période 1961 à 1997 était 2.35% par an [El-Sharkawy et al., 2012].

Dans l'ensemble, les résultats apportés par fineSTRUCTURE et δaδi sont cohérents. Ces deux méthodes sont complémentaires dans l'étude des processus démographiques au sein du complexe d'espèce *X. axonopodis*. FineSTRUCTURE a ceci d'intéressant qu'il ne considère aucune population *a priori* et peut traiter l'ensemble des souches échantillonnées. Cette méthode est donc très utile en première instance pour permettre l'identification des "populations" ainsi qu'une première estimation des échanges génétiques entre les différentes souches composant ces populations. Les résultats de δaδi apportent cependant plus de précision dans l'estimation des paramètres démographiques car cette méthode s'appuie sur un modèle démographique explicite. Ainsi contrairement à fineSTRUCTURE, les apparentes similarités génétiques entre souches sont interprétées dans un contexte évolutif complexe, impliquant à la fois de la divergence, du flux de gènes et des changements de tailles de population. Toutefois, δaδi, n'offre la possibilité d'analyser qu'un nombre limité de populations. Dans notre étude, si l'analyse de paires de populations était possible, nous n'avons pas réussi à analyser 3 populations pour des modèles complexes (par exemple en introduisant des changements de

TABLE X – Probabilités logarithmiques (Log-Likelihood) et critère d'information d'Akaike (AIC) des meilleurs modèles (ΔAIC <10) testés avec $\delta a \delta i$ avec 20 analyses pour chacune des paires testées en prenant l'ensemble des SNP non-filtré. Les modèles hétérogènes sont les modèles 2M.

Groupes	Modèles	Log-likelihood	AIC
9.5-9.6	SC2M2Pex2	-10450,78	20925,56
9.5-9.6	IM2M2Pex2	-10562,47	21146,93
9.5-9.6	SCex2	-10595,46	21210,92
9.5-9.6	IM2Mex2	-10649,04	21318,09
9.5-9.2	SCex2	-20100,42	40220,85
9.5-9.2	SC2M2Pex2	-20205,48	40434,96
9.5-9.2	SC2Mex2	-20208,91	40439,82
9.6-9.2	SC2M	-21260,34	42536,69
9.6-9.2	SC2Mex2	-21778,17	43578,34
9.6-9.2	SCex2	-21979,66	43979,32
9.6-9.41	SC2M2Pex2	-26086,46	52196,93
9.6-9.41	SCex2	-26360,39	52740,79
9.6-9.41	IM2Mex2	-26447,84	52915,68

taille de populations). Ceci tient sûrement au fait que pour identifier le meilleur scénario avec précision sous $\delta a \delta i$, les SNP utilisés doivent être indépendants ce qui contraint souvent à une réduction importante du jeu de données initial, compte tenu de l'étendue du déséquilibre de liaison entre SNP voisins. Les SNP indépendants permettent de choisir le meilleur modèle, mais cependant la réduction du jeu de donnéE diminue la puissance statistique. Les modèles hétérogènes sont presque systématiquement les meilleurs lorsque les analyses sont conduites sur les SNP totaux (cf. **tab. X**). Lorsque qu'on retire du jeu de données les SNP liés, on supprime alors aussi les régions liées aux sites sous sélection positive par autostop génétique [Smith and Haigh, 1974]. Ces régions sous sélection positive peuvent constituer des barrières génétiques au flux de gènes [Krause and Whitaker, 2015]. Ainsi supprimer les SNP en déséquilibre de liaison peut rendre plus difficile la détection d'un modèle incorporant l'hétérogénéité de flux de gènes le long du génome.

4.2 Le flux de gènes est variable entre les groupes

L'impact sur le polymorphisme de la recombinaison par rapport à la mutation $r/m = 0{,}97$ est qualifié de faible (< 1) et similaire aux valeurs reportés pour *Enterococcus faecium* et *Ralstonia solanacearum* [Vos and Didelot, 2009]. Ce *ratio* est nettement inférieur à celui précédemment estimé par Mhedbi-Hajri et al. (2013) sur sept portions de gènes de ménage $r/m = 3{,}18$. Une telle différence dans les estimations de r/m est certainement imputable au nombre différent de locus échantillonnés dans les deux études. En effet, le taux de recombinaison, loin d'être homogène sur l'ensemble du génome, varie selon les régions génomiques [MARAIS, 2002]. Nous avons estimé r/m à partir de données d'alignement du *core genome* produit par Harvest et contenant des régions codantes et non codantes. Une telle différence dans les estimations de notre étude et celle de Mhedbi-Hajri (2013) pourrait aussi refléter une hétérogénéité de la recombinaison entre les gènes et les régions non-codantes. Une autre hypothèse serait que cette différence serait induite par des différences entre les phylogénies, notamment pour le groupe 9.3. Cette hétérogénéité du taux de recombinaison rendrait donc l'estimation du paramètre r/m très sensible à l'échantillonnage des locus. C'est pour cela que l'utilisation de nombreux locus tamponne les effets possibles, des variations du taux de recombinaison

breux locus tamponne les effets possibles, des variations du taux de recombinaison le long du génome, et des variations stochastiques [Vos and Didelot, 2009]. Notre estimation serait donc être plus robuste, car elle est basée sur un *core genome* de 1,9 Mb. Les différences d'estimation du paramètre r/m sont susceptibles d'être causées par des différences dans les méthodes analytiques, mais aussi par le fait que les méthodes analytiques sont sensibles à la stratégie d'échantillonnage utilisée pour collecter les isolats bactériens [Didelot and Maiden, 2010]. Dans le cas du clade séro-résistant de *Moraxella catarrhalis*, qui contient la plupart des souches virulentes de cet organisme a un taux de recombinaison six fois plus important que la population séro-sensible [Wirth et al., 2007].

De même, malgré le faible nombre de souches non pathogènes de notre collection (4), celles-ci semblent plus recombinantes que les souches pathogènes **(fig. 26)**. Les souches non pathogènes possèdent un système de sécrétion de type III, mais ne provoquent pas de symptômes sur leur hôte d'isolement. Ces résultats sont en accord avec les travaux de Merda et al. (2016)[Merda et al., 2016b], où le réseau recombinant est constitué par des souches «non pathogènes» au sein duquel émergent des clones épidémiques beaucoup moins recombinants. Concernant le rapport entre le taux de recombinaison et le taux de mutation R/θ, les groupes présentant les valeurs les plus faibles sont les groupes 9.5 et 9.6 (respectivement $R/\theta = 0,08$ et $R/\theta = 0,06$). Ces deux groupes sont les moins divergents au sein du complexe d'espèces *X. axonopodis*. Le scénario évolutif inféré par $\delta a \delta i$ correspond au modèle IMex2, c'est à dire un modèle de divergence avec flux de gènes et changement de taille efficace. Néanmoins, les taux de migration inférés sont très faibles (c.a. 2.10_{-7}), indiquant la présence possible de barrières génétiques au flux de gènes. Bien que plus divergents que la paire précédente, les groupes 9.5 et 9.2 présentent un taux de migration plus important entre eux, de l'ordre de $1,5.10^{-5}$ migrants par génération. Ceci pourrait impliquer une plus grande compatibilité génétique entre les groupes 9.5 et 9.2, qu'entre les groupes 9.5 et 9.6, où le flux de gènes est bien plus limité. Ce résultat pose la question de l'existence et la nature des barrières au flux de gènes entre ces groupes.

4.3 Existence d'une barrière à la recombinaison homologue entre les groupes étroitement liés 9.5 et 9.6

Le faible flux de gènes entre les groupes 9.5 et 9.6 pourrait donc refléter l'existence de facteurs limitant les échanges entre ces groupes.

Premièrement, les limitations d'échanges de gènes ne semblent pas être dues à des facteurs liés aux hôtes. En effet, la phylogénie des hôtes ne suit pas la phylogénie des groupes. Certaines souches au sein des trois groupes 9.2, 9.5 et 9.6 partagent même certains hôtes, c'est le cas des souches pathogènes des *Citrus*.

Deuxièmement, la divergence ne semble pas non plus conduite par la géographie car dans chacun des groupes des souches ont été isolées sur plusieurs continents. Il n'existe donc pas de relation évidente entre le site géographique de prélèvement des souches et leur appartenance à un groupe particulier. En revanche plusieurs souches appartenant à des groupes génétiques différents (9.5 ou 9.6) ont été isolées dans des régions proches.

Enfin, on peut faire l'hypothèse de barrières génétiques lors de la conjugaison, résultant d'incompatibilités de *pili* et de protéines de surface, ou des systèmes RM, ou SRM (voir paragraphe II.1.1), ou encore d'incompatibilité génétique entre locus (épistasie négative) provenant de fond génétiques différents seraient autant de causes qui ne permettrait plus à la recombinaison de jouer son rôle de force cohésive. Un tel type d'isolement reproductif a été détecté entre deux types de bactéries du sol *Myxococcus xanthus*, vivant en sympatrie [Wielgoss et al., 2016]. Les auteurs ont identifié une région avec un contenu génomique variable qui pourrait expliquer cette différence de phénotype. Chez les groupes 9.5 et 9.6, la faible divergence (environ 7585 ans estimé avec IMSC **(tab. IX)**) contraste fortement avec le faible flux de gènes. Ce résultat suggère fortement l'existence de barrières génétiques écologiques ou non ayant permis la mise en place rapide d'un tel isolement reproductif. De telles barrières sont difficiles à identifier. Par exemple, dans le cas d'espèces d'*Archaea* thermophiles en sympatrie, seule une différence de croissance suggère que ces espèces sont maintenues par une différenciation écologique [Cadillo-Quiroz et al., 2012]. L'étude de la variation des valeurs de F_{ST} et du D de Tajima sur le *core genome* des groupes 9.5 et 9.6 n'a pas permis d'identifier de régions montrant des traces de balayage sélectif.

4.4 Il ne semble pas y avoir de barrières à la recombinaison entre les groupes plus divergents

L'inférence démographique a permis d'identifier un scénario de divergence avec contact secondaire suivi de flux de gènes pour les groupes 9.5 et 9.2. Ces groupes auraient divergé il y a environ 7700 ans (estimation du modèle IMSC **(tab. IX)**) sans flux de gènes avant d'entrer à nouveau en contact très récemment. Ce contact, récent (c.a. 316 ans d'apès le modèle IMSC et 10^{-8} jour pour le modèle SCex2), est certainement dû à la mise en sympatrie de souches ayant divergé en allopatrie, et ce grâce aux échanges internationaux et à la mondialisation de l'agriculture. L'introduction de nouvelles espèces dans de nouvelles aires géographiques, la production de semences regroupées dans certains pays, l'échange de plants, permettraient la mise en contact des populations d'agents pathogènes jusqu'alors isolées [Leroy et al., 2016]. On peut supposer qu'il n'existe pas de barrières entre ces groupes empêchant la mise en place de flux de gènes. Le fort flux de gènes inter-groupes inféré pour des souches partageant les mêmes hôtes des groupes 9.6 et 9.41, confirme l'absence de barrières entre des groupes plus divergents que 9.6 et 9.5. C'est le cas des souches CFBP 4885 (groupe 9.6) et CFBP 6164 (groupe 9.41) pathogènes sur des phaseolées, et des souches CFBP 3132 (groupe 9.6) et CFBP 3133 (groupe 9.41) pathogènes du dieffenbachia. Les travaux de Aritua et al. 2015[Aritua et al., 2015] montrent aussi l'existence d'un flux de gènes entre 9.6 et 9.41 en identifiant plus de 100 gènes probablement acquis de *X. citri* pv. *fuscans* (9.6) par *X. phaseoli* pv. *phaseoli* LG1 (9.41). Des analyses au laboratoire dans des conditions optimisées pourraient permettre de calculer la fréquence de recombinaison lors de la conjugaison entre des souches des groupes 9.5 et 9.6. La recombinaison lors de la conjugaison entre *S. typhimurium* et *E. coli* est par exemple 105 fois plus basse que la recombinaison intra spécifique. Cela confirme l'existence de barrière à l'échange génétique entre ces deux espèces [Rayssiguier et al., 1989]. Sans ce genre d'expériences, on ne peut savoir si le peu de recombinaison identifié dans cette étude entre les groupes est le résultat de barrières écologiques ou génétiques.

4.5 Les groupes 9.5 et 9.6 forment-ils deux espèces ?

Notre objectif était de comprendre la structure des populations du complexe d'espèces *Xanthomonas axonopodis*, dans l'objectif de le comparer aux espèces déjà décrites. Nous avons montré que ce complexe avait plusieurs niveaux de structuration, et que la structuration en cinq groupes nous semblait la plus pertinente. Les groupes identifiés par les différentes méthodes (ACP, fineSTRUCTURE, ANI...) sans ambiguïté sont les groupes 9.2, 9.3, et 9.41. Ils correspondent exactement aux espèces *X. euvesicatoria*, *X. axonopodis*, et *X. phaseoli* respectivement. FineSTRUCTURE est la méthode qui confirme notre choix de regrouper 9.4 avec 9.1 car ces deux groupes présentent du flux de gènes entres les génomes et constituent un groupe homogène **(fig. 26)**. Un autre jeu de données comportant plus de souches du groupe 9.1 permettrait d'avoir plus de robustesse dans les analyses notamment pour fastSTRUCTURE.

Enfin, les groupes 9.5 et 9.6 sont clairement différenciés par toutes les méthodes : ACP, arbre phylogénétique, fastSTRUCTURE et calcul d'ANI. Ces deux groupes sont déjà tellement divergents que l'étude des SNP *outliers* par PCAdapt, n'a pas permis d'identifier des SNP liés à l'adaptation.

Les groupes 9.5 et 9.6 sont regroupés au sein l'espèce *X. citri* au seuil utilisé pour la définition de l'espèce bactérienne de 95% [Constantin et al., 2016], mais à un seuil plus élevé de 96,5% deux cliques bien indépendantes sont formées. Dans notre étude, nous avons montré que les groupes 9.5 et 9.6 présentent un faible flux de gènes, environ 20 fois plus faible qu'entre les souches d'un même groupe, ce qui est également confirmé par une faible migration (méthode $\delta a \delta i$) de l'ordre de 10^{-3} individus par génération malgré une divergence assez récente ($Td = 7585$ ans estimé par IMSC **(tab. IX)**). Est-ce judicieux de garder au sein de l'espèce *Xanthomonas citri* des souches pour lesquelles on met en évidence un isolement reproductif ? Le statut d'espèce de ces deux groupes 9.5 et 9.6 a déjà fait l'objet de débats. Ces deux groupes ont en effet déjà été décrits comme deux espèces bactériennes *X. citri* et *X. fuscans* mais uniquement sur la base de certains pathovars (pv. *citri* et *malvacearum* pour 9.5 ; et pv. *fuscans* et *aurantifolii* pour 9.6)[Schaad et al., 2005b]. Notre étude semble donc confirmer ces observations mais cependant enrichit chacune des espèces bactériennes *X. citri* et *X. fuscans* d'autres pathovars

tels que pv. *glycines*, pv. *mangiferaeindicae* du 9.5 et pv. *anacardii*, pv. *vignicola du 9.6*.

Cependant, d'autres études contestent ce statut d'espèce aux groupes 9.5 et 9.6. *X. fuscans* et *X. citri* seraient des synonymes pouvant être reclassés comme sous espèces de *X. axonopodis* [Young et al., 2008]. Ah -you et al. (2009) propose aussi la synonymie entre *X. fuscans* et *X. citri* mais suggère de les regrouper en une seule espèce *X. citri*.

En comparaison avec les eucaryotes, les groupes 9.5 et 9.6 sont dans une "zone grise" qui engloberait des souches dont la spéciation n'est pas achevée, avec un flux de gènes extrêmement faible. Une telle zone existe chez les animaux et a été mise en évidence par Roux et al. (2016)[Roux et al., 2016], sur un continuum de spéciation de 61 paires de populations (espèces) animales avec des degrés de divergence variables. Une "zone grise" située entre 0.5% et 2% de divergence synonyme correspond aux situations où le flux de gènes est hétérogène le long du génome à cause de l'existence de barrières génétiques au flux de gène. L'existence de telles barrières est un prérequis à la spéciation. Ce genre de corrélation n'a pas encore été testée chez les bactéries. Cependant on peut facilement imaginer que chez de tels organismes, la spéciation obéit certainement à d'autres règles, considérant notamment l'asexualité et les mécanismes d'échanges de gènes. Toutefois, l'existence d'un flux de gène hétérogène entre les groupes 9.5 et 9.6 (cf. **tab. X**) semble indiquer que chez ces deux groupes l'isolement reproductif apparaît de manière progressive, comme cela a été observé chez les animaux [Roux et al., 2016]. Nos résultats nous permettent tout de même d'affirmer que les groupes 9.5 et 9.6 présentent un degré substantiel d'isolement reproductif, ce qui exclut d'emblée leur regroupement au sein de la même espèce. Il devient alors intéressant de connaître les causes d'un tel isolement.

Quatrième partie

CHAPITRE II

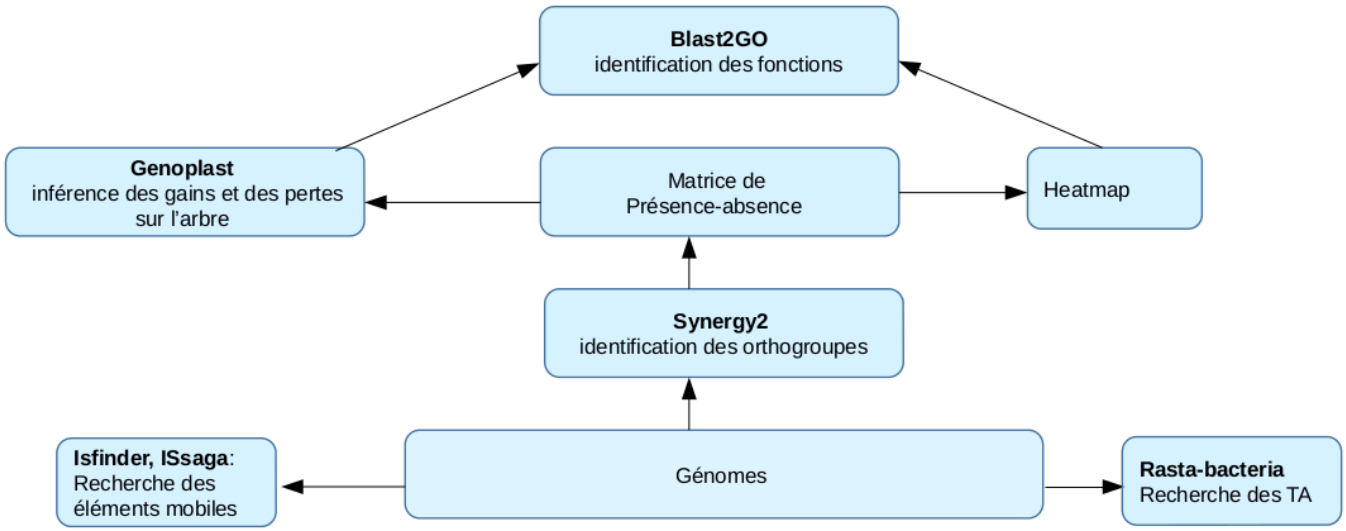


FIGURE 35 – Schéma synoptique du traitement des données
Les flèches indiquent les données d'entrées des différents outils utilisés dans le Chapitre II

1 Introduction

L'objectif de ce deuxième chapitre est de comparer les dynamiques évolutives du *core genome* et du génome accessoire, afin d'identifier des gènes et des fonctions qui pourraient avoir un rôle dans l'isolement reproductif et la divergence adaptative des groupes précédemment identifiés. Par cette approche nous essayerons de comprendre l'impact du transfert horizontal de gènes lors de l'histoire de la divergence au sein du complexe d'espèces *Xanthomonas axonopodis*. En effet, ce mécanisme moléculaire peut être impliqué dans l'émergence de nouveaux agents pathogènes [Bartoli et al., 2016]. L'acquisition ou la perte de gènes peut alors entraîner d'importantes modifications phénotypiques, notamment concernant le pouvoir pathogène. Par exemple, dans le cadre d'un modèle gène pour gène, on prédit qu'une protéine de résistance chez la plante va reconnaître une protéine effectrice provenant de la bactérie, et déclencher une réaction d'hypersensibilité (HR) [Flor, 1971]. La perte du gène codant pour cette protéine effectrice chez la bactérie permet alors de contourner efficacement la résistance de la plante [Kim et al., 2009], [Li et al., 2015].

La distribution du génome accessoire a été étudiée sur les 73 génomes de notre collection. Les fonctions des génomes *core* et accessoires ont été analysées dans le but de savoir si certaines catégories de fonctions sont différentiellement représentées dans les groupes définis dans le chapitre précédent, et si les gènes impliqués dans certaines fonctions s'échangent plus *via* la recombinaison homologue ou le HGT. La présence de gènes spécifiques d'un groupe génétique potentiellement responsables d'un isolement écologique entre groupes de souches a été recherchée dans la matrice de présence-absence des gènes accessoires sur l'ensemble des 73 souches de notre jeu de données.

Nous avons ensuite estimé les gains et les pertes de gènes pendant l'histoire évolutive du complexe d'espèces *Xanthomonas axonopodis* dans l'objectif d'identifier les moments-clefs de l'apport du polymorphisme par ce mécanisme. L'étude des fonctions codées par les gènes acquis ou perdus n'a pas permis de démontrer si certaines fonctions étaient sur ou sous-représentées, au moment de la divergence en groupes, ou au moment de la spécialisation en pathovars par exemple.

Nous avons en outre étudié plus particulièrement la distribution des séquences

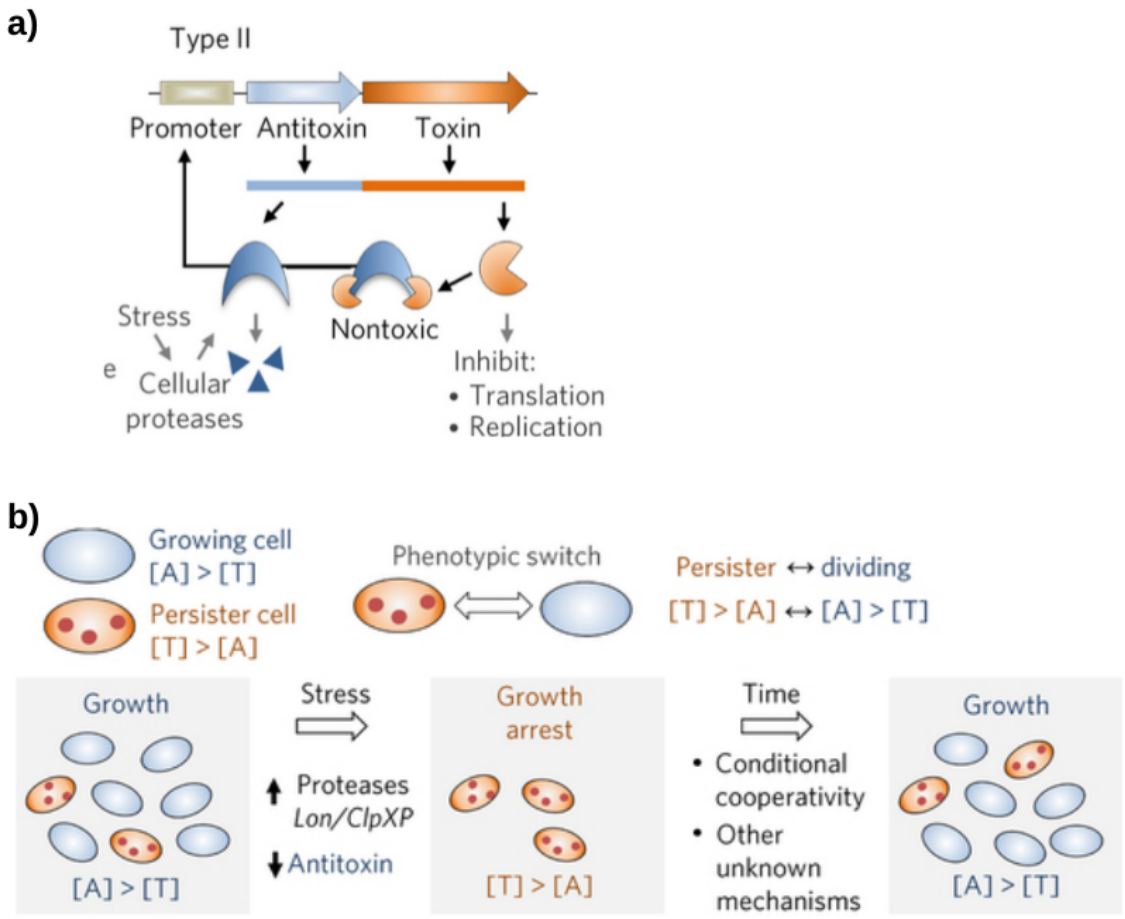


FIGURE 36 – Système Toxine Antitoxine

a) Les toxines sont indiquées en orange et les antitoxines en bleues ; les activitées qui ne sont pas toxiques sont en noires, et les toxiques en grises. Toxine Antitoxine de type II : L'antitoxine et la toxine sont des protéines ; en phase de croissance, la toxine est liée à l'antitoxine ce qui inhibe son activité. L'antitoxine et le complexe TA, se lient au promoteur pour réprimer la transcription. En condition de stress, les protéases cellulaires comme Lon et ClpXP sont activées et clive préférentiellement l'antitoxine, libérant ainsi la toxine qui inhibe la croissance en inhibant la traduction et la réplication b) Dans des cellules en croissance (bleues), la concentration en antitoxine exède celle de la toxine. C'est l'inverse dans les cellules persistantes (orange), Le basculement d'un état à l'autre nécessite un changement de ce ratio toxine-antitoxine. L'exposition à un stress, comme une carence en nutriment ou des antibiotiques, tend a activer les protéases qui clivent les antitoxines, il en résulte un excès de toxines. Cela va rapidement éteindre les activités de croissance cellulaire et aussi permuter une petite fraction de cellule à l'état persistant. D'après [Page and Peti, 2016]

d'insertions (IS) sur l'ensemble des génomes disponibles. Il est en effet connu que ces IS jouent un rôle important d'agent mutagène chez les bactéries permettant, l'adaptation aux nouveaux environnements et la colonisation de nouvelles niches [Touchon and Rocha, 2007]. Ainsi, l'analyse de la répartition des IS entre les groupes peut nous informer sur leur rôle dans l'apport de diversité génétique et la différenciation. Les IS sont impliquées dans l'inactivation de gènes affectant la virulence, la résistance et le métabolisme [Vandecraen et al., 2017]. Par exemple, l'effet des IS sur la virulence, *via* la modulation de la formation de biofilm et la production de polymères extra-cellulaires a été décrit pour différentes bactéries, comme chez *Staphylococcus epidermidis* et *Staphylococcus aureus* [Ziebuhr et al., 1999].

Enfin, nous avons voulu étudier la répartition des systèmes Toxine-Antitoxine (TA) dans les génomes du complexe d'espèces *Xanthomonas axonopodis*. Les systèmes Toxine-Antitoxine (TA) sont de petits éléments génétiques composés d'un gène de toxine et de celui de son antitoxine associée **(fig. 36)**. Les toxines sont toutes connues pour être des protéines, alors que les antitoxines peuvent être des protéines ou des ARN non codant [Unterholzner et al., 2013]. Les toxines sont stables pendant que les antitoxines sont rapidement dégradées par des protéases. Ce système permet de maintenir les plasmides lors de leur transmission de générations en générations [Sengupta and Austin, 2011]. Cependant, beaucoup de TA ont été trouvées dans les chromosomes bactériens, dans ce cas leur fonction biologique est autre. Elles permettent aux bactéries de passer d'un état de croissance à un état de latence, en réponse aux stress environnementaux induit par une carence en nutriment ou par des antibiotiques [Martins et al., 2016]. Les TA pourraient aussi participer à la diminution des HGT dans leur environnement génomique proche, de manière similaire au système de restriction-modification (RM) empêchant l'insertion d'ADN étranger dans les génomes bactériens [Pandey, 2005]. Dans ce cas, les groupes contenant le plus grand nombre de TA pourraient être indicateurs d'une diminution du HGT dans ces groupes.

2 Matériels et méthodes

Le matériel et méthode est résumé dans un schéma synoptique figure 35.

2.1 Annotation fonctionnelle

Nous avons tout d'abord cherché à associer chacun des gènes de notre jeu de données à une fonction biologique. Le but est de déterminer si certaines fonctions, ou classes de fonctions sont particulièrement acquises par HGT. Ainsi ces acquisitions auraient-elles pu provoquer la divergence en groupes, ou la spécialisation en pathovars ? L'annotation fonctionnelle a été déterminée à l'aide de Blast2GO [Conesa and Götz, 2008]. Blast2GO recherche par BlastX (recherche d'une séquence nucléotidique traduite dans une base de données protéique) des séquences similaires déjà caractérisées, et compare les résultats obtenus à la base de données *Gene Ontology* afin de *in fine* transférer les données fonctionnelles aux séquences consensus de chaque orthogroupe. Ici, la séquence référence de chaque gène était la séquence consensus de chacun des orthogroupes précédemment défini (voir III.3.3.2).

2.2 Distribution du génome accessoire entre les souches

Afin d'analyser la répartition des gènes au sein des souches, et la structuration de la collection basée sur le génome accessoire, une matrice ordonnée de présence-absence a été générée avec la fonction *heatmap* du paquet R {stat}[Team, 2013]. Les lignes (individus) et colonnes (gènes) de cette matrice ont été ordonnées sur la base de la similitude calculée à partir de la matrice binaire de présence-absence obtenue avec Synergy2. Ainsi, les gènes spécifiques de chaque groupe génétique ont été identifiés grâce à cette matrice ordonnée. Afin de déterminer s'il existait un lien entre la distribution des gènes au sein des 73 souches et les fonctions codées par ces mêmes gènes, un test de corrélation a été réalisé entre la matrice de similitude calculée sur le profil de distribution des gènes dans la collection et une matrice de similitude sémantique des termes GO associés à chaque gène. Cette dernière a été obtenue à l'aide du paquet R {GOSemSim} [Yu et al., 2010]. La corrélation a été vérifiée avec un test de Mantel réalisé avec le paquet R {ade4} et 999999 permutations. Cette analyse permet de tester si la distribution des fonctions dans la matrice de présence-absence est aléatoire, ou corrélé à la distance entre les souches.

2.3 Divergence et adaptation des groupes 9.5 et 9.6

La recherche des gènes spécifiques de chaque groupe génétique d'après la matrice ordonnée est assez restrictive, puisque nous avons recherché les gènes présents dans tous les génomes d'un groupe et strictement absents dans tous les autres génomes. Dans cette partie nous nous focaliserons sur les groupes 9.5 et 9.6. Comme cela a été montré dans le chapitre précédent, ces deux groupes sont génétiquement les plus proches et leur histoire évolutive est maintenant bien caractérisée. La connaissance de cette histoire évolutive permet de fixer un véritable contexte démographique et évolutif dans lequel les gains et pertes de gènes par transferts horizontaux ont pu se produire. Les régions différentiellement présentes entre les groupes 9.5 et 9.6 ont été recherchées en effectuant des *scans* des génomes pour identifier des régions (*k-mers*) spécifiques avec SkIf [Martial Briand et al., 2016]. SkIf est un outil de comparaison génomique qui détecte des régions, ici la taille du *k-mer* a été fixée à 22 bp, spécifiquement associées à des groupes d'organismes.

2.4 Analyse des gains et des pertes lors de l'histoire évolutive

Dans le chapitre précédent, nous avons montré que les moments clés de la divergence (en groupes ou en pathovars) sont souvent associés à un nombre plus important d'événements de recombinaison **(fig. 24)**. Dans le but de déterminer à quel moment de l'histoire évolutive le HGT intervient, et de préciser son rôle dans la différenciation en groupes, ou en pathovars, les gains et les pertes de gènes ont été inférés sur la phylogénie du complexe d'espèces *Xanthomonas axonopodis* à l'aide de Genoplast [Didelot et al., 2008]. Ces évènements de gains et de pertes de gènes sont placés sur un arbre ultramétrique transformé à partir de la phylogénie obtenue avec FastTree2 dans la suite Harvest (voir paragraphe III.3.3.1) grâce au paquet R {ape} [Paradis et al., 2004]. Genoplast implémente un modèle autorisant une variation des probabilités de pertes et de gains le long des branches de l'arbre, ce qui est censé mieux correspondre à des évènements de HGT non constants au cours de l'histoire évolutive des populations bactériennes. La recherche des paramètres du modèle se fait par Monte-Carlo Markov Chain (MCMC). Après

plusieurs essais, le nombre d'itérations a été finalement fixé à 1 million plus 500000 itérations de préchauffage (*burn in*) et une collecte des paramètres toutes les 2000 itérations, ce qui nous permet d'estimer les distributions de paramètres à partir de 500 valeurs. Dans notre étude, deux exécutions ont été réalisées afin de tester la convergence. Les gènes gagnés et perdus sur certaines branches d'intérêt, précédent par exemple la divergence en groupes ou en pathovars ont étés localisés sur un génome de référence : CFBP 4885 pour le groupe 9.6, axcitri306 pour 9.5, CFBP 6984 pour 9.41, CFBP 5610 pour 9.2 afin de détecter si ces gènes appartiennent à des clusters. En effet, les transferts d'un grand nombre de gènes sur une branche pourraient correspondre au transfert d'un opéron, ou d'un plasmide. Les fonctions de ces gènes ont ensuite été recherchées avec Blast2GO.

2.5 Analyse des IS et du système TA

Les éléments génétiques mobiles jouent un rôle important dans la plasticité des génomes. Dans cette étude il a été choisi de rechercher les séquences d'insertions (IS). Nous nous sommes aussi intéressés aux gènes codant les systèmes Toxine-Antitoxine.

2.5.1 Recherche des IS

La détection des séquences d'insertion a été réalisée avec le *pipeline* ISsaga [Varani et al., 2011], qui utilise la base de données ISfinder [Siguier, 2006]. Les paramètres pour la recherche de ces séquences sont fixés par ce *pipeline*. Ces paramètres sont une identité supérieure à 97%, une e-value de 10^{-5}, un *word-size* de 3. ISsaga détecte uniquement les IS entières. Ceci est particulièrement important car comme nos séquences sont assemblées en contigs, et que les IS sont connues pour provoquer des ruptures d'assemblages, le nombre d'IS sera sous-estimé.

2.5.2 Recherche des TA

L'étude de la distribution des Toxine-Antitoxine (TA) a été réalisée afin déterminer si une corrélation existe entre le nombre de gènes codant pour ces TA et une diminution du HGT, sachant que ces gènes joueraient un rôle répresseur de

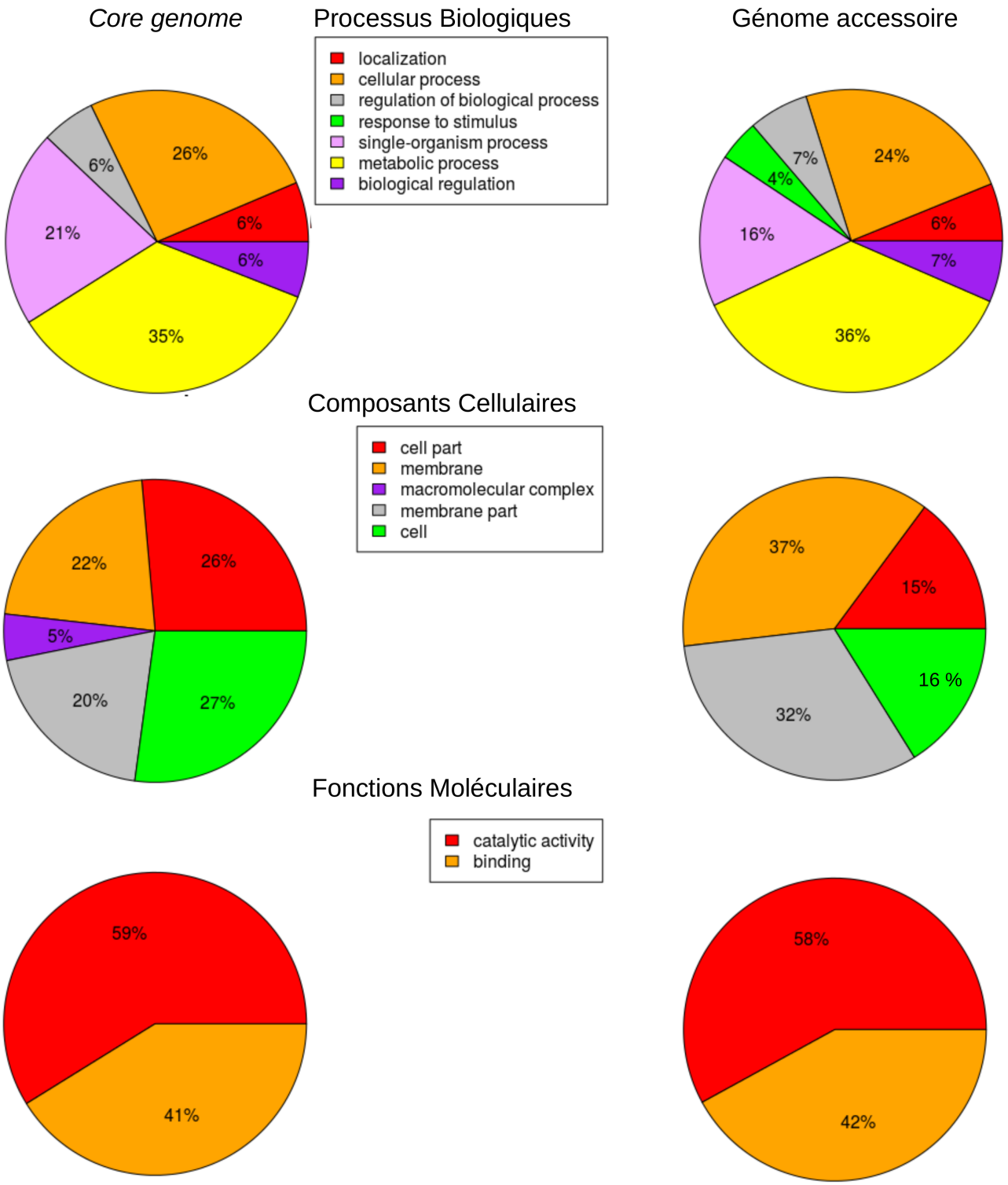


FIGURE 37 – Comparaison des catégories de fonction entre le *core genome* et le génome accessoire Proportion des fonctions des gènes pour les trois catégories, Processus Biologiques, Composants cellulaires, et Fonctions Moléculaires, chez les gènes orthologues et les gènes du génome accessoire

l'insertion d'ADN étranger [Pandey, 2005]. La recherche des TA a été réalisée avec le programme RASTA-Bacteria (Rapid Automated Scan for Toxins and Antitoxins in Bacteria) [Sevin and Barloy-Hubler, 2007] en ne gardant que les protéines pour lesquelles le score est supérieur à 70% (ce score représente une forte probabilité d'appartenir à la famille des TA). Cette recherche a été complétée avec les TA identifiés par [Martins et al., 2016] sur les génomes de *X. citri* pv. *citri* 306 et *X. euvesicatoria* 8510 qui ont été recherchés par Blastn, et notés présents si le seuil d'identité était supérieur ou égal à 80% sur 80% de la longueur.

3 Résultats

3.1 Comparaison des fonctions des gènes orthologues avec celles du génome accessoire

Globalement les fonctions des gènes sont réparties de la même façon et dans les mêmes proportions entre le génome accessoire et le *core genome* **(fig. 37)**. On remarque ainsi que 69% des gènes du génome accessoire sont associées à la membrane contre seulement 42% dans le *core genome*. On note aussi que la catégorie des gènes impliqués dans les réponses aux stimulus est uniquement représentée dans le génome accessoire où ils représentent 4% des fonctions présentes. Dans cette catégorie, on trouve les processus qui résultent d'un changement d'état ou d'activité (mouvement, sécrétion, production d'enzymes, expression de gènes) de la bactérie en réponse à un stimulus. Ces résultats sont en accord avec l'idée que le génome accessoire code pour les fonctions liées à l'adaptation rapide à un environnement changeant. Ainsi, les gènes accessoires pourraient être partagés par des populations pathogènes ayant des histoires adaptatives similaires, et en particulier à une pathogénie commune [Wiedenbeck and Cohan, 2011].

La catégorie des fonctions de gènes liés aux complexes macromoléculaires est quant à elle présente uniquement dans le *core genome* (avec une proportion de 5%). Dans cette catégorie on retrouve des fonctions liées à l'adhésion, des macromolécules transmembranaires, ou des bicouches lipidiques attachées à la membrane.

Afin de tester l'hypothèse d'une répartition aléatoire des fonctions des gènes dans la matrice de présence-absence, nous avons réalisé un test de Mantel. Pour

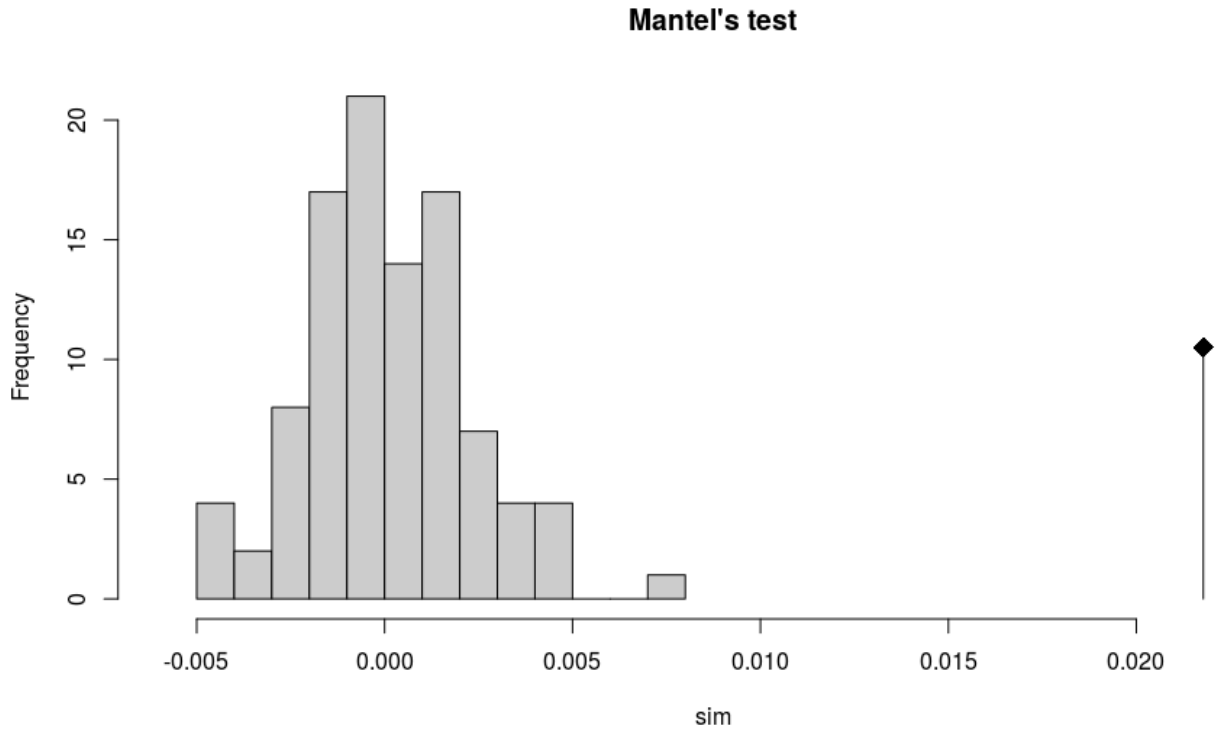
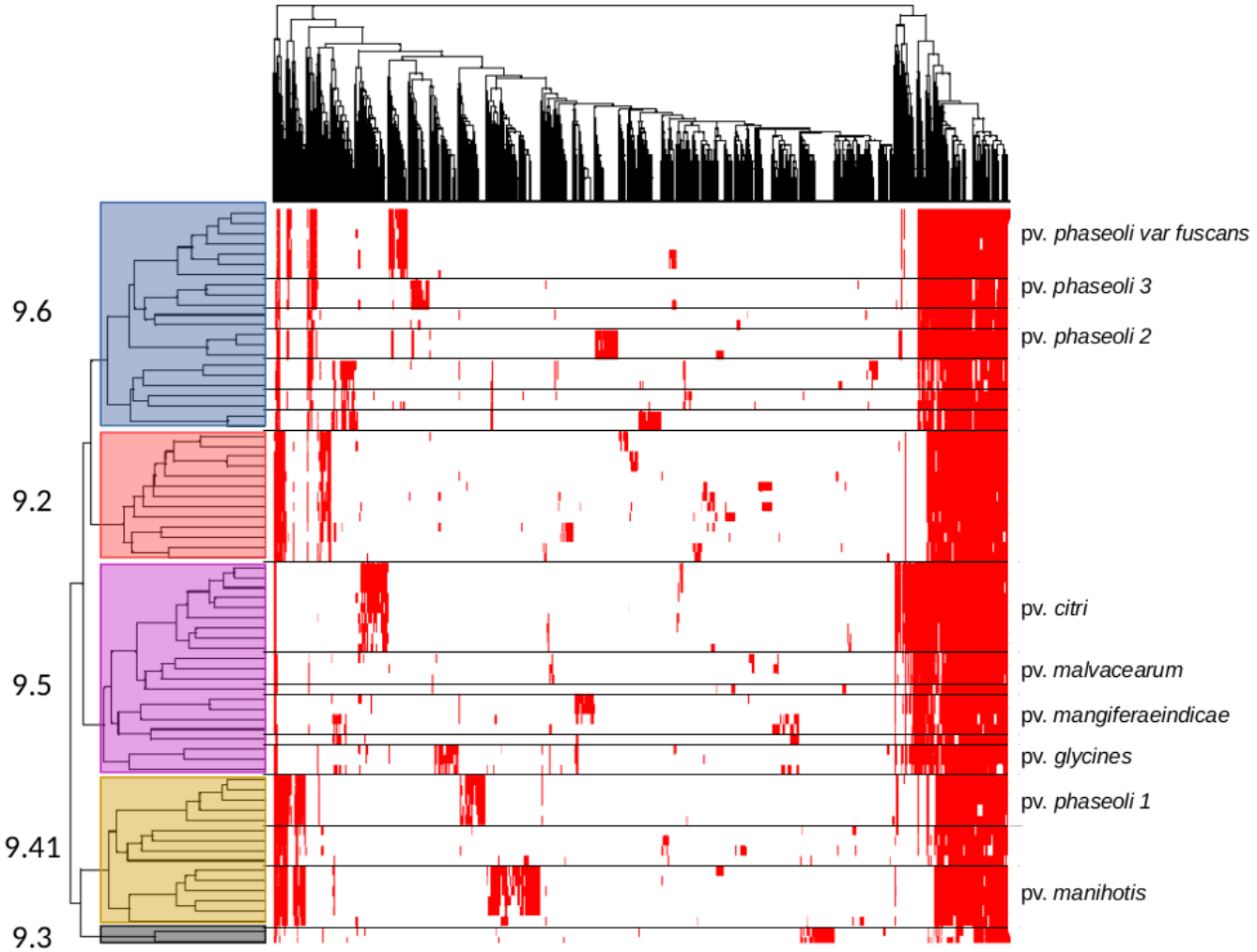


FIGURE 38 – Test de Mantel
Distribution des valeurs de r pour 999999 permutations, le losange indique la valeur de r de nos données

cela nous avons comparé la matrice de distance des génomes basée sur les présence-absence de gènes, avec la matrice de distance sémantique de GO (*Gene Ontology*). L'hypothèse nulle est rejetée : les fonctions des gènes et leur distribution chez les souches sont corrélées (r =0,02178266, $p = 10^{-6}$) **(fig. 38)**. Le regroupement des génomes en fonction de la matrice binaire des gènes du génome accessoire permet d'identifier cinq groupes, de la même façon qu'avec le *core genome* **(fig. 39)**. Cependant, bien que ces 5 groupes soient clairement identifiés sur la base des présence-absence de gènes, leur place sur le dendrogramme diffère de celle observée sur l'arbre établi sur la base du *core genome*. En effet, contrairement à la *core* phylogénie, le groupe 9.6 est plus proche du groupe 9.2 que du groupe 9.5.

3.2 Identification des gènes spécifiques des groupes

Des gènes spécifiques de chacun des groupes sont identifiés, 3 sont spécifiques du groupe 9.2, 195 gènes sont spécifiques du groupe 9.3, 6 gènes sont spécifiques du groupe 9.41, 26 pour le groupe 9.5, et seulement 2 gènes sont spécifiques du groupe 9.6. Les résultats de BlastX et les ontologies des gènes (GO) sont indiqués dans l'**Annexe C**. Les gènes spécifiques de chacun des groupes de plus de 2 génomes ont été placés sur un génome de référence (CFBP 4885 pour le groupe 9.6, axcitri306 pour 9.5, CFBP 6984 pour 9.41, CFBP 5610 pour 9.2) ce qui a permis de montrer que 23 des 26 gènes spécifiques du groupe 9.5 font partie d'un cluster. Les 23 gènes du groupe 9.5 organisés en opéron représentent une partie de la région XACSR1 de *X. citri* pv. *citri* 306 [Moreira et al., 2010] **(fig. 40)** qui contient des gènes impliqués dans la biosynthèse des lipopolysaccharides (LPS). Cette région semble similaire, selon les résultats de Blast, avec une région d'un génome d'*Acidovorax avenae* subsp. *aspargine* avec 73% identité sur 64% de la longueur. Cette bactérie est responsable de maladies sur le maïs et l'avoine, et elle distribuée dans toutes les zones de production de maîs et d'avoine. Seul les trois premiers gènes de ce cluster (de XAC0037 à XAC0039), ne sont pas spécifiques au groupe 9.5, on les retrouve chez des souches du groupe 9.2 (pour les gènes XAC0037, XAC0038, XAC0039) et 9.41(XAC0038, XAC0039). Les LPS sont des composants clés de la membrane externe des bactéries à Gram négatif. Il sera intéressant de replacer les gènes de cet opéron dans l'histoire évolutive du complexe d'espèce *Xanthomonas axonopodis*,



center)

FIGURE 39 – Matrice ordonnée des 7288 gènes présent-absent
Les groupes sont indiqués à droite, les principaux pathovars à gauche. Le dendrogramme en haut représente le regroupement des 7288 gènes. La présence et l'absence de gène est indiquée respectivement en rouge et blanc

afin de déterminer si ces gènes ont pu être gagnés par ce groupe, ou perdu chez les autres groupes. Pour les autres groupes, 9.6, 9.41 et 9.2 aucun opéron n'a été identifié. Les gènes présents sont dispersés dans les génomes. L'insertion d'un opéron comme dans le cas du groupe 9.5 semble être un événement rare.

3.3 Absence de gènes communs pouvant expliquer la convergence pathologique de pathovars de groupes différents

Dans la matrice de présence-absence quelques gènes semblent spécifiques à des pathovars. Cependant malgré la convergence pathologique de pathovars de différents groupes comme *Xanthomonas citri* pv. *citri* (9.5), *Xanthomonas citri* pv. *aurantifolii* (9.6), et *Xanthomonas euvesicatoria* pv. *citrumelonis* (9.2) sur agrumes, aucun gène n'est présent spécifiquement chez toutes les souches pathogènes des agrumes. C'est aussi le cas des pathovars *Xanthomonas citri* pv. *fuscans*, pv. *fuscans* LG2, pv. *fuscans* LG3 (9.6), et *Xanthomonas phaseoli* pv. *phaseoli* LG1 (9.41) dont les souches sont pathogènes des Phaseolus (haricots). Entre les pathovars sur haricot, un gène codant pour une *hypothetical protein* a été gagné sur la branche conduisant au pv. *fuscans* LG2 et au pv. *fuscans* lignée *fuscans*, un gène codant pour une *hypothetical protein* entre pv. *fuscans* LG2 et LG3, et un gène codant pour un *TonB* entre pv. *fuscans* LG3 et *fuscans*. Les transporteurs TonB peuvent être impliqués dans la transport actif d'éléments nutritifs ou dans la transduction de signaux [Ryan et al., 2011].

3.4 Détection de gènes impliqués dans l'adaptation locale ayant pu conduire à la divergence entre 9.5 et 9.6

L'algorithme SkIf identifie 29% (1 535 814 bp) du génome de FDC1083 (utilisé comme référence du groupe 9.5) comme spécifique aux génomes du groupe 9.5 par rapport aux génomes du groupe 9.6. On retrouve par cette méthode l'opéron XACSR1 identifié plus haut dans la matrice présence-absence. Pour le groupe 9.6 c'est 25% (1 311 641 bp) du génome de CFBP 2913 (utilisé comme référence du groupe 9.6) qui lui est spécifique par rapport au groupe 9.5. Dans ce groupe 9.6, l'approche basée sur les *k-mers* identifient une région de 6 gènes spécifiques

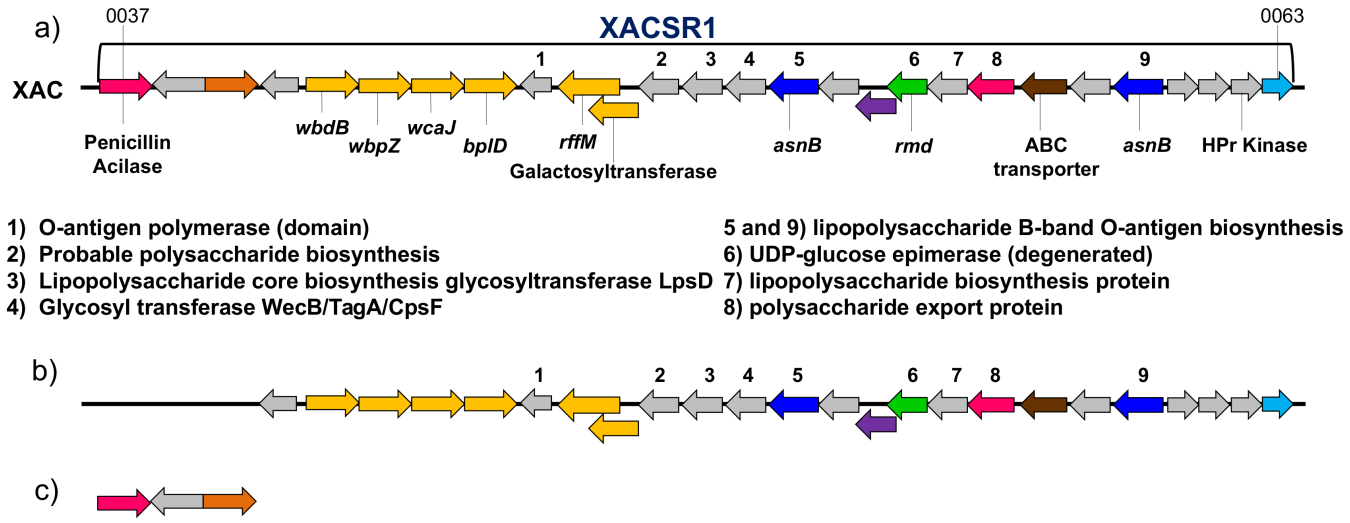


FIGURE 40 – Région XACSR1

a) Région XACSR1 du génome de *Xanthomonas citri* pv. *citri 306* [Moreira et al., 2010]. Ce cluster contient plusieurs gènes impliqués dans la synthèse de lipopolysaccharide (LPS). Il contient deux copies de *asnB* codant pour une aspargine synthase. Un homologue de ce gène chez *Pseudomonas aeruginosa* est impliqué dans la biosynthèse d'antigène O. b) Les gènes de l'opéron XACRS1 spécifiques du groupe 9.5 c) Les trois gènes présents qui ne sont pas spécifiques du groupe 9.5

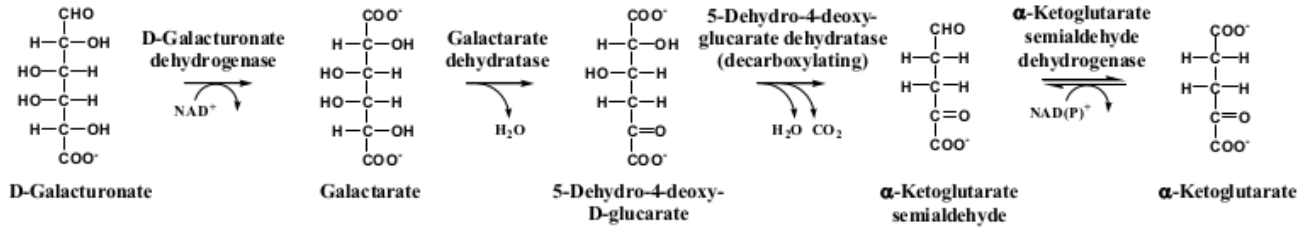


FIGURE 41 – Voie métabolique du D-galacturonate chez *Agrobacterium tumefaciens* [Hilditch and Valtion teknillinen tutkimuskeskus, 2010]

(les gènes $XFF4834R\_chr.1255$, $XFF4834R\_chr.1256$, $XFF4834R\_chr.1257$, $XFF4834R\_chr.1259$, $XFF4834R\_chr.1260$ et $XFF4834R\_chr.1261$ de notre annotation) dont trois sont semblables à ceux de la voie catabolique du D-galacturonate chez *Agrobacterium tumefaciens*. Deux voies cataboliques différentes du D-galacturonate ont été identifiées chez les bactéries. L'une avec une première étape d'isomérisation, l'autre avec une étape d'oxydation. Cette deuxième voie a été identifiée chez *Pseudomonas syringae* et *Agrobacterium tumefaciens* **(fig. 41)** [Hilditch and Valtion teknillinen tutkimuskeskus, 2010].

3.5 L'apport du polymorphisme par transfert horizontal n'est pas un processus continu le long de la phylogénie

Les deux analyses indépendantes réalisées avec Genoplast sont convergentes (Gelman-Rubin $Rc = 0,999$). Les événements ancestraux avec une probabilité supérieure ou égale à 0,50 ont été positionnés à chaque nœud de l'arbre **(fig. 42)**. Une hypothèse sous-jacente de cette méthode est l'indépendance d'occurrence des événements de gains et de pertes. Dans la réalité, ce n'est pas toujours le cas, parce que beaucoup de gènes peuvent être gagnés (ou perdus) dans un événement de transfert unique (par exemple, l'acquisition d'un plasmide). Les résultats présentés ici pourraient donc surestimer le nombre d'événements de transfert impliqués. La variation du nombre de gains ou de pertes de gènes est considérable le long des branches. Généralement, un nombre très important d'événements de gains et de pertes se sont produits le long des branches terminales et ce nombre diminue le long des branches plus profondes de la phylogénie. Une corrélation positive existe entre le nombre de gènes gagnés et la distance à la racine **(fig. 43)**. Les gains terminaux les plus importants ont été inférés sur les branches conduisant aux pathovars *Xanthomonas phaseoli* pv. *manihotis* (420 gains), *Xanthomonas citri* pv. *glycines* (374 gains) ainsi qu'aux deux souches non pathogènes de *Xanthomonas citri* CFBP 7765 et CFBP 7923 (556 gains). En moyenne, 8,3% du génome total moyen dans le pathovar *X. p.* pv. *manihotis* a été gagné récemment, mais cette valeur augmente à 15,6% lorsque l'on considère seulement la proportion relative à la taille du génome accessoire. Les pathovars avec le nombre le plus important de pertes terminales inférées sont *Xanthomonas citri* pv. *mangiferaeindicae* (46 pertes), les

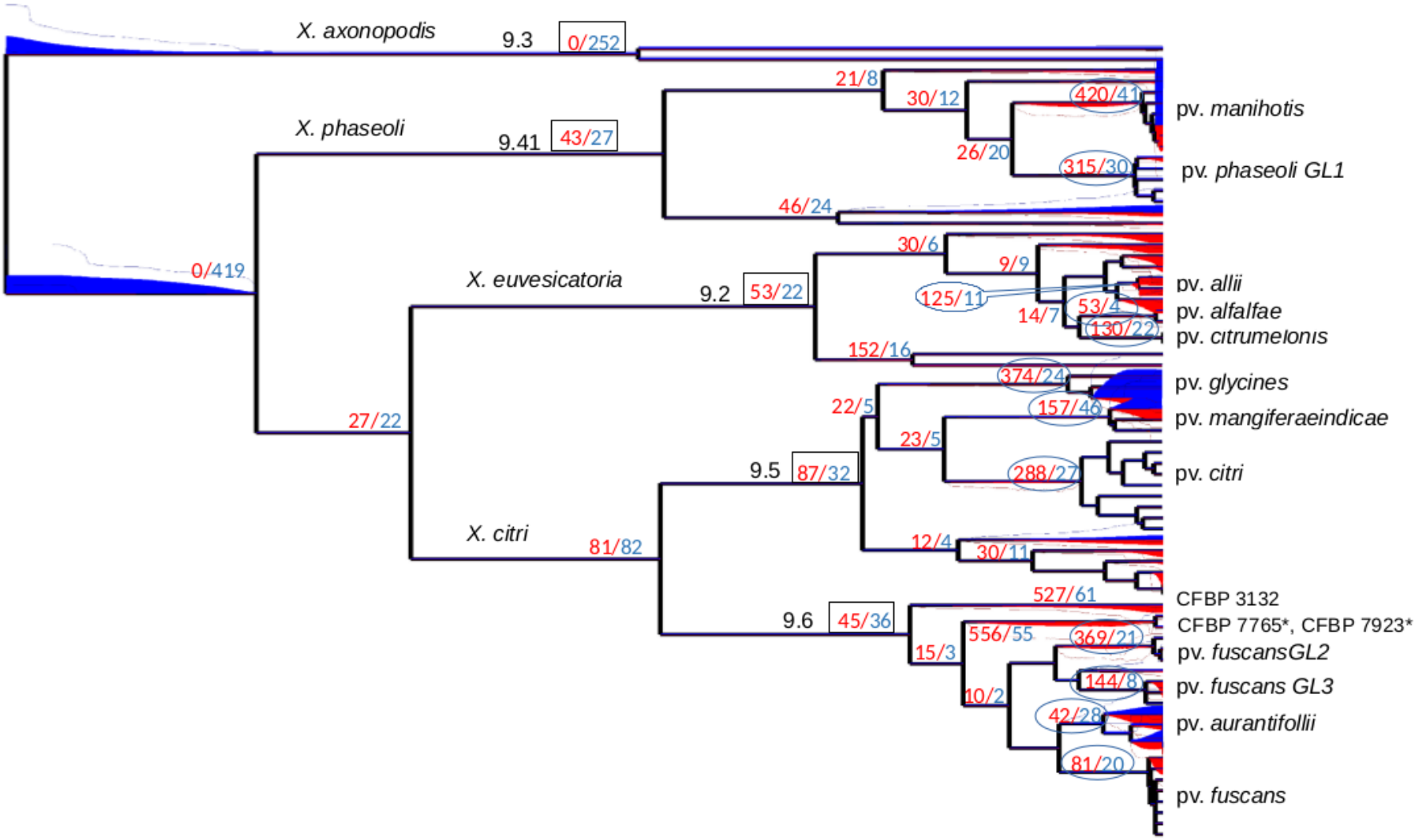


FIGURE 42 – Inférence des gènes gagnés et perdus

Genoplast infère les événements de gain et de perte à chaque branche de l'arbre phylogénétique. Les gains sont indiqués en rouge, les pertes en bleus. Plus la couleur de la branche est forte plus la probabilité est élevée.

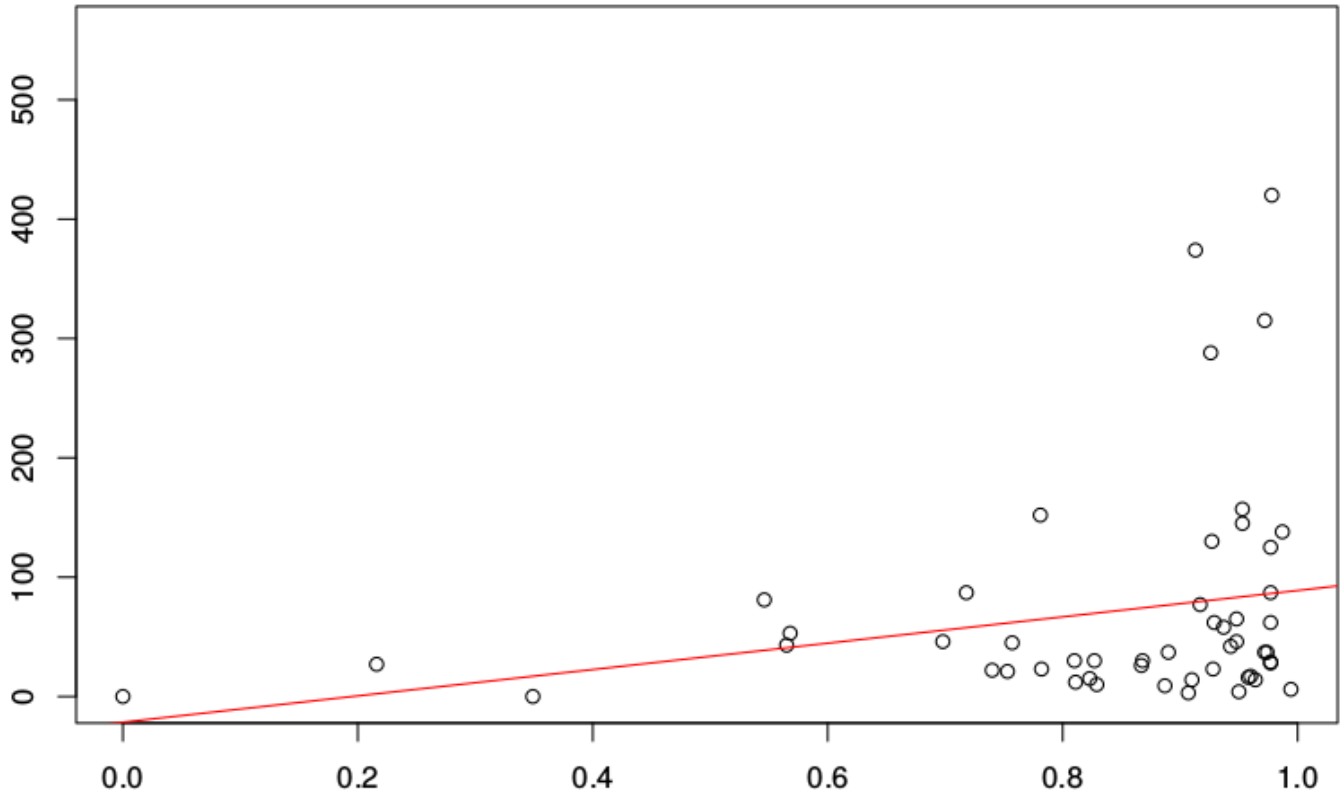


FIGURE 43 – Plot du nombre de gains en fonction de la distance à la racine
En rouge la droite de régression.

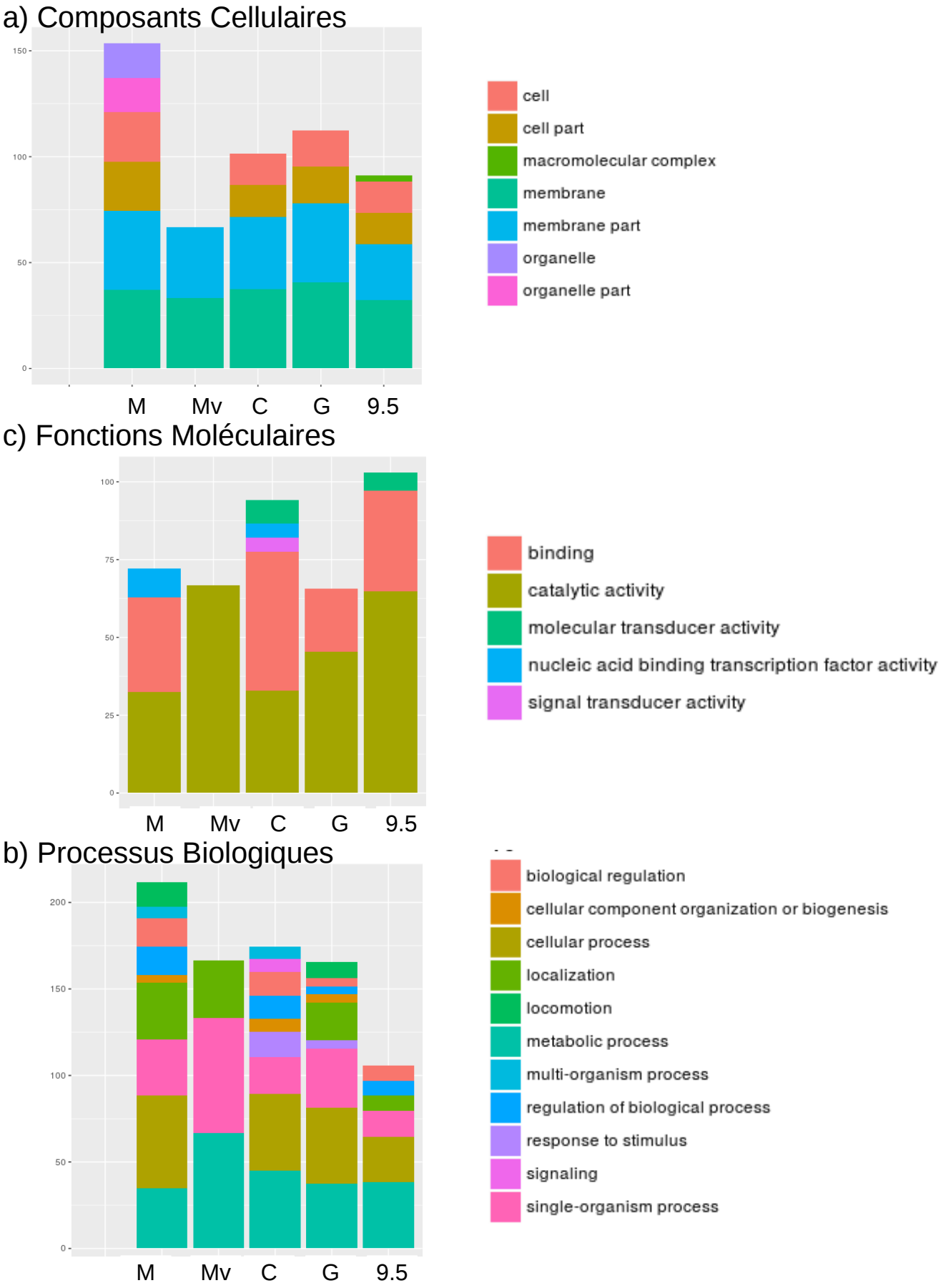


FIGURE 44 – Histogrammes des proportions de fonctions des gènes gagnés
Les fonctions des gènes gagnés au niveau des branches précédents les pathovars *mangiferaeindicae* (M), *malvacearum* (Mv), *citri* (C) et *glycines* (G) et du groupe 9.5 sont indiqués pour les trois classes de GO (a,b et c).

deux souches non pathogènes de *Xanthomonas citri* CFBP 7765 et CFBP 7923 (55 pertes) et la souche *Xanthomonas citri* CFBP 3132 (61 pertes) **(fig. 42)**. Cela représente 1% du génome total moyen dans le pathovar *Xanthomonas citri* pv. *mangiferaeindicae* et 1,9% lorsque l'on considère seulement la proportion relative à la taille du génome accessoire.

3.6 Fonctions gagnées ou perdues aux noeuds stratégiques de l'arbre phylogénétique

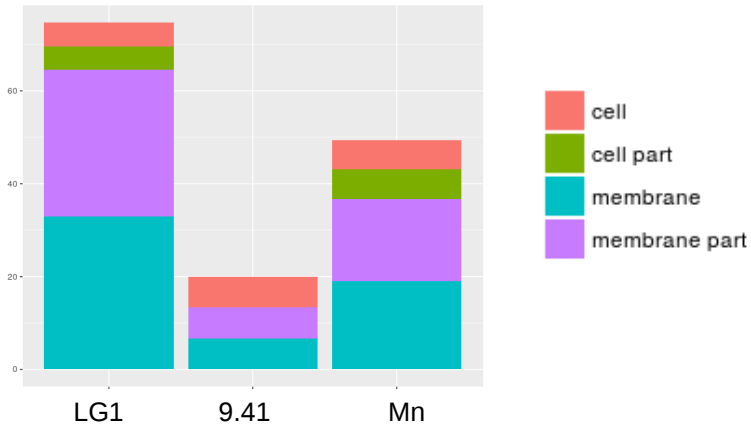
Les fonctions des gènes (gagnés ou perdus) au niveau des branches précédents la divergence en groupes sont similaires aux fonctions des gènes inférées aux branches conduisant aux pathovars (des exemples sont représentés **(figure 44** et **45)**). Certains gènes gagnés sur la branche conduisant au groupe 9.5 sont regroupés en deux clusters, le cluster spécifique de ce groupe XACSR1, et un autre groupe de quatre gènes constitués d'une histidine kinase, d'une protéine hypothétique, d'une protéine de stress et d'une endonucléase de réparation d'excision de nucléotide. Le cluster de 6 gènes précédemment identifié par SkIf comme étant spécifique du groupe 9.6, a été gagné sur la branche conduisant à ce groupe. D'autres gènes ont été localisés sur le génome de référence pour les autres groupes mais ils n'appartiennent pas à un opéron identifié.

3.7 Recherche du contenu génomique en éléments mobiles IS et du système TA

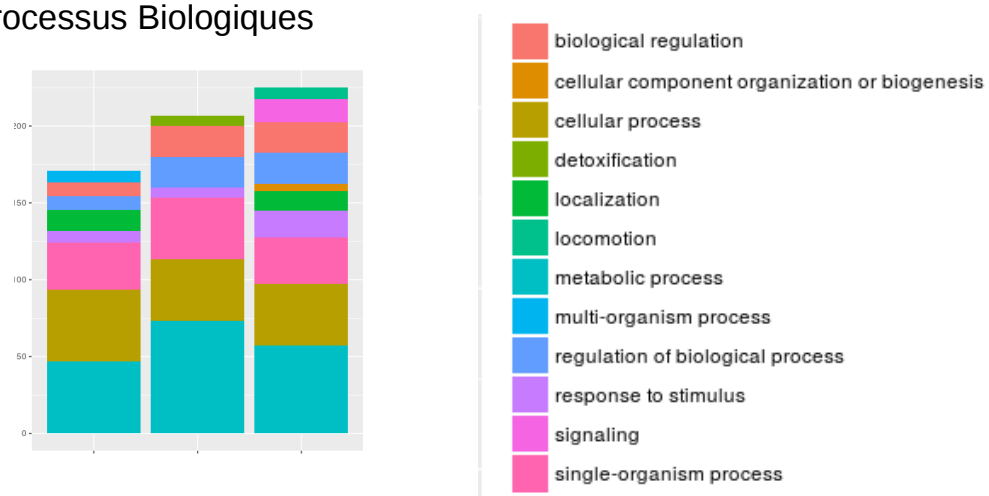
Le nombre moyen des IS par génome est de 35 dans un génome du groupe 9.6, 37 dans un génome du groupe 9.5, 52 dans un génome du groupe 9.41 et 39 dans le groupe 9.2. Les groupes 9.5 et 9.6 contiennent tous les deux la plus grande diversité d'IS avec 17 familles différentes, contre 15 pour 9.41 et 11 pour le groupe 9.2. Les familles d'IS spécifiques des groupes 9.5 et 9.6 sont les IS4 (0,62%), IS4_is10 (0,15%), IS110 (0.77%), IS1634 (1,08%). L'IS701 (1,14%) est spécifique du groupe 9.41 **(fig. 46)**.

Concernant les gènes codant pour les toxines ou antitoxines (TA), le pathovar *X. citri* pv. *citri* est celui qui contient le plus de TA avec une moyenne de 40 ± 3.

a) Composants Cellulaires



b) Processus Biologiques



c) Fonctions Moléculaires

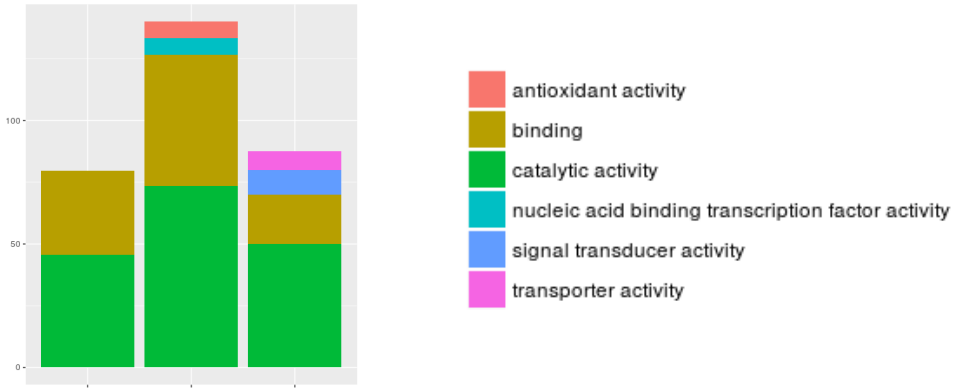
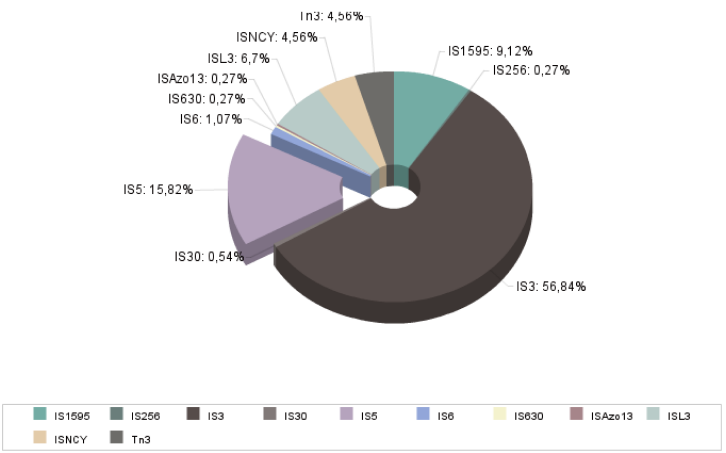
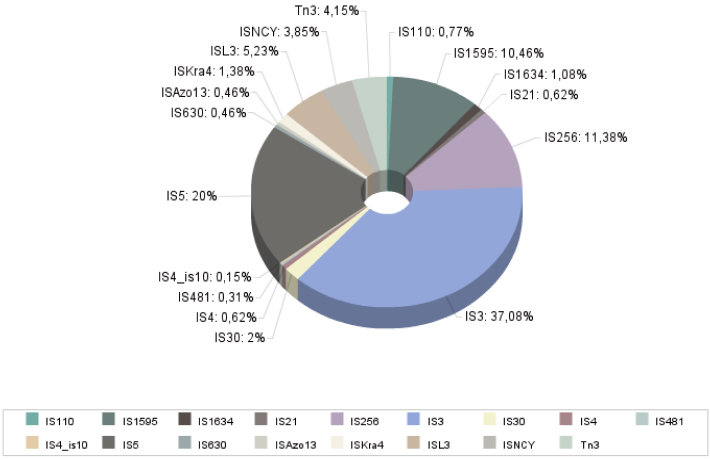


FIGURE 45 – Histogrammes des proportions de fonctions des gènes gagnés
Les fonctions des gènes gagnés au niveau des branches précédents les pathovars *X. axonopodis* pv. *phaseoli* (LG1), *X. axonopodis* pv. *manihotis* (Mn) et le groupe 9.41 sont indiqués pour les trois classes de GO (a,b et c).

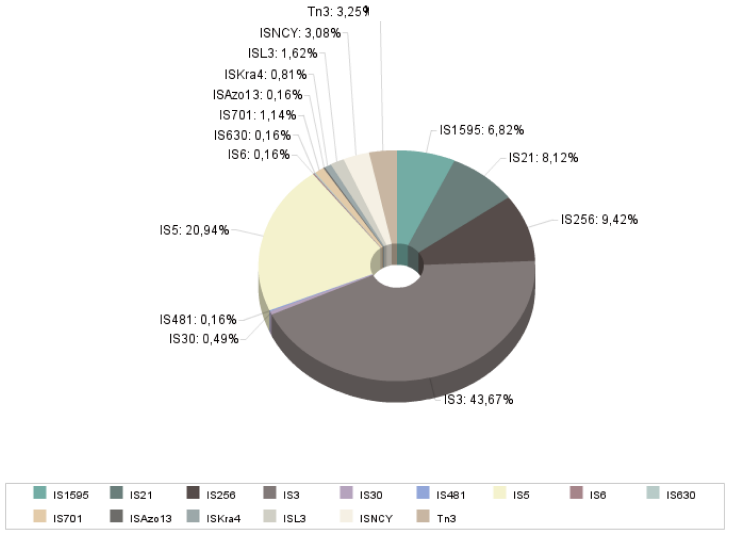
a) 9.2



b) 9.5



c) 9.41



d) 9.6

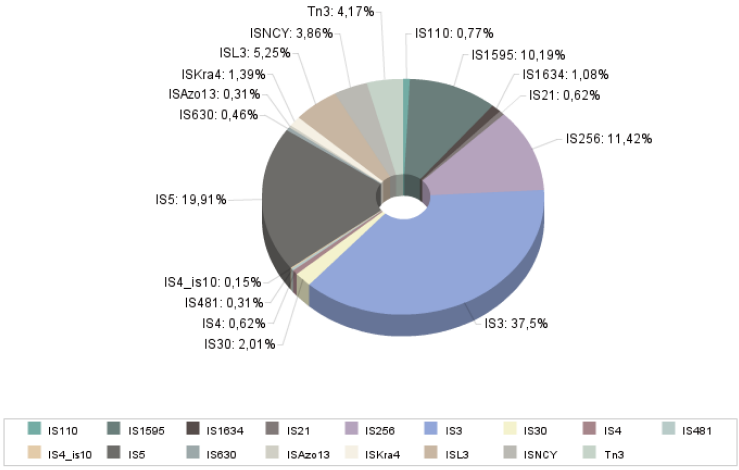


FIGURE 46 – Répartition des familles d'IS au sein des groupes.

TABLE XI – Nombre de Toxines et Antitoxines moyen par groupes et par pathovars

Groupes	Nombre de T et A	*pathovars*	Nombre de T et A
9.3	21±2		
9.2	21±4	*citrumelonis*	18
		alfalfae	21±3
9.41	21±3	*phaseoli LG1*	24±2
		manihotis	20±2
9.5	34±6	*citri*	40±3
		glycines	29±3
		malvacearum	32±6
		mangiferaeindicae	21±1
9.6	22±3	*fuscans LG2*	24±1
		fuscans LG3	23±3
		fuscans	23±3
		aurantifollii	20±1

De manière plus générale, le groupe 9.5 est aussi celui qui en contient le plus par rapport au autres groupes avec une moyenne de 36 ± 6 **(tab. XI)**. La topologie basée sur la présence-absence des TA regroupent les souches par groupes dans le cas du groupe 9.2 et du groupe 9.5. Ils regroupent aussi les groupes 9.41 et 9.6 **(fig. 47)**, la composition en TA des lignées de *Xanthomonas phaseoli* pv. *phaseoli* LG1, et de *Xanthomonas citri* pv. *fuscans* semblent rapprocher ces souches. La plupart des *Xanthomonas citri* pv. *fuscans* sont dans un clade, excepté CFBP 6165 et CFBP 6970 qui semblent plus proches des lignées LG2 et LG3 que des autres *fuscans*. Cela s'expliquerait peut être par leur origine, ces souches sont les seules originaires du continent américain.

Des délétions de toxines sont retrouvées dans des génomes. Le gène XAC1194 codant pour une toxine avec un domaine COG3654-doc est absent des génomes des souches du pathovar *manhihotis* et de tous les génomes des groupes 9.6 et 9.3. Le gène XCV1189 codant pour une toxine avec un domaine de stabilisation de plasmide RelE/ParE est délété chez la plupart des souches du pathovar *citri*, excepté chez axcitr1279 et axcitr9322, chez les souches du pathovar *mangiferaeindicae* et chez des souches du groupe 9.6 et 9.2. Le gène XCV0819 codant pour une toxine avec un domaine pfam01845-CcdB est absent de certain génome des souches du groupe 9.2 et du pv. *citri*. Le gène XACb0033 qui code pour une toxine avec un domaine pfam11455-DUF3018 est absent des souches du pv. *phaseoli* (9.41) et de certaines souches du groupe 9.6. On trouve aussi uniquement une antoxine relB sans la toxine associée chez certaines souches du pathovar *X. c.* pv. *fuscans* (excepté LG2 et LG3), du pv. *aurantifolii*, ainsi que des souches du groupe 9.41 appartenant au pathovar *manihotis* **Annexe E**. On trouve des délétions d'antitoxines telles que le gène XAC0188 (domaine COG1396HipBcd00093HTH_XRE). Ce dernier est absent des génomes du pathovar *mangifereaindicae*.

Les gains et les pertes des TA ont été localisés sur l'arbre grâce à Genoplast **(fig. 48)**. De nombreux gains ont été inférés au niveau des branches supportant les pathovars. Certaines TA ont été acquises par des pathovars de groupes différents, mais qui partagent les même hôtes. C'est le cas de XFF4834R_RS06680, XFF4834R_RS21050 et XFF4834R_RS20795 qui ont été gagnés au niveau du pathovar *X. phaseoli* pv. *phaseoli* LG1 du groupe 9.41 et du pathovar *X. c.* pv. *fuscans* qui sont des pathovars qui regroupent des souches pathogènes sur ha-

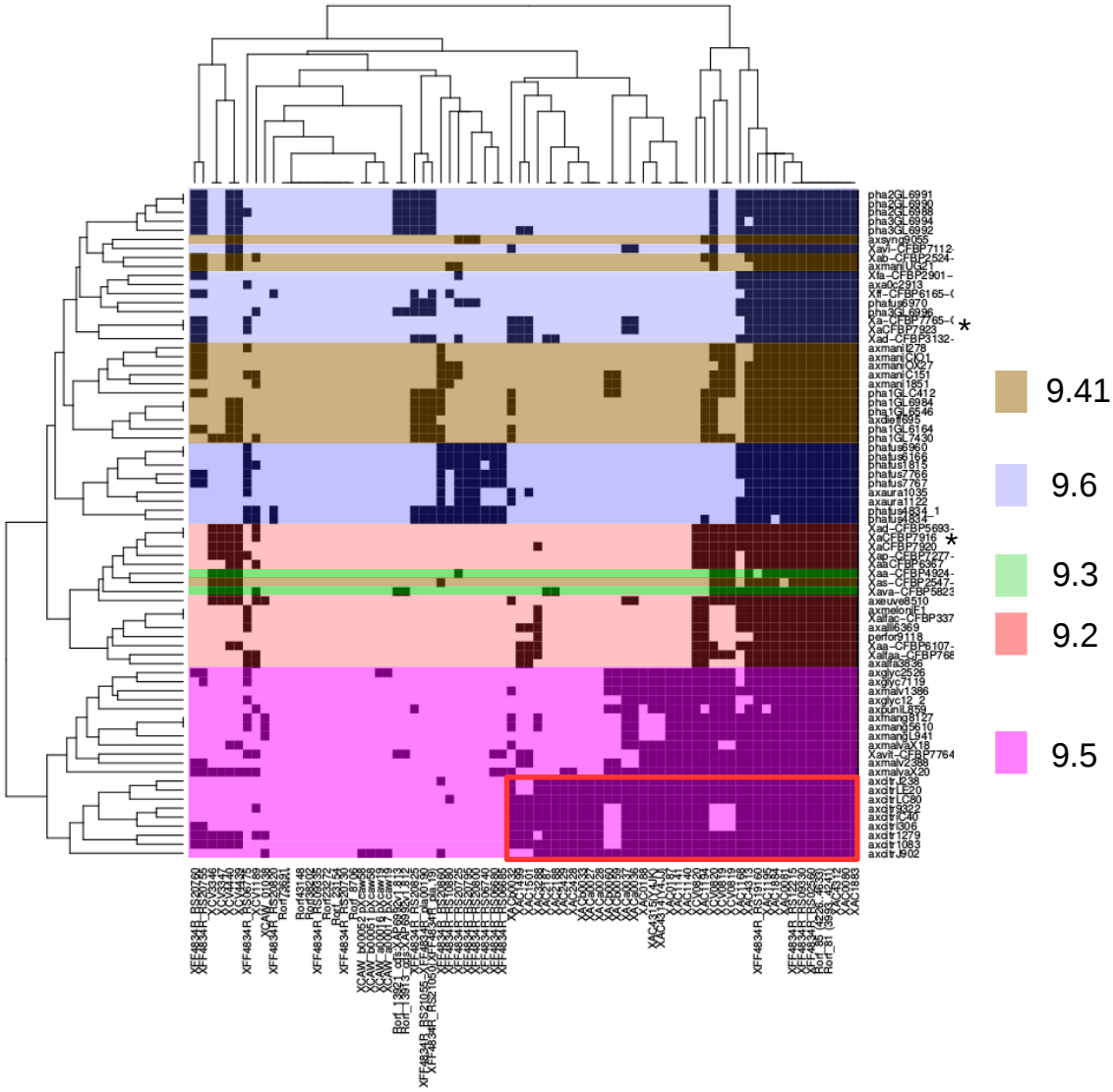


FIGURE 47 – Matrice ordonnée en fonction des TA.
Les groupes sont indiqués par des couleurs. Les TA sont indiqués en colonnes, les souches en lignes. Les souches du pathovars *citri* sont encadrées en rouge.

ricot. D'autres TA (Rorf_13913, XFF4834R_RS20820, XCAW_a00017) ont été acquises au niveau des pathovars *X. c.* pv. *fuscans* LG2 et LG3 du groupe 9.6 ainsi que du pathovar *X. phaseoli* pv. *phaseoli* LG1 du groupe 9.41. Le pathovar *X. c.* pv. *fuscans* a gagné 6 toxines ou antitoxines.

4 Discussion

Dans ce chapitre II nous nous sommes intéressés à l'évolution du contenu en gènes du génome accessoire ainsi qu'aux échanges de ces gènes entre populations. La topologie des souches basée sur la matrice de présence-absence confirme les groupes que nous avons définis au chapitre précédent. Les différences entre la topologie basée sur le *core genome* et celle basée sur le génome accessoire seront discutées ci-dessous. L'inférence des gains et des pertes de ces gènes lors de l'histoire évolutive montre que le HGT s'est intensifié récemment lors de l'apparition des pathovars. Dans les paragraphes suivants, nous mettrons en parallèle le polymorphisme apporté par HGT avec celui apporté par recombinaison homologue lors de l'histoire évolutive de ce complexe. Afin d'identifier des fonctions qui pourraient expliquer la divergence des groupes, certains gènes spécifiques de groupes ont été recherchés. Nous discuterons pour finir de leur rôle dans la divergence écologique et l'adaptation à l'hôte.

4.1 Comparaison des topologies des arbres basés sur les génomes *core* et accessoire

Nous avons inféré l'histoire évolutive du complexe d'espèces *X. axonopodis* en explorant les différences d'échanges génétiques entre le *core genome* d'une part et le génome accessoire d'autre part. Les évènements de gains et de pertes de gènes, par transfert horizontal, ne sont pas soumis aux mêmes contraintes que la recombinaison homologue. En effet contrairement à la recombinaison homologue, le HGT peut se produire entre des souches très éloignées phylogénétiquement [Popa and Dagan, 2011]. Cela implique que les phylogénies du *core genome* et celle réalisée à partir de la matrice des gènes présent-absent puissent ne pas être identiques. De manière générale, les topologies de souches bactériennes basées sur le génome

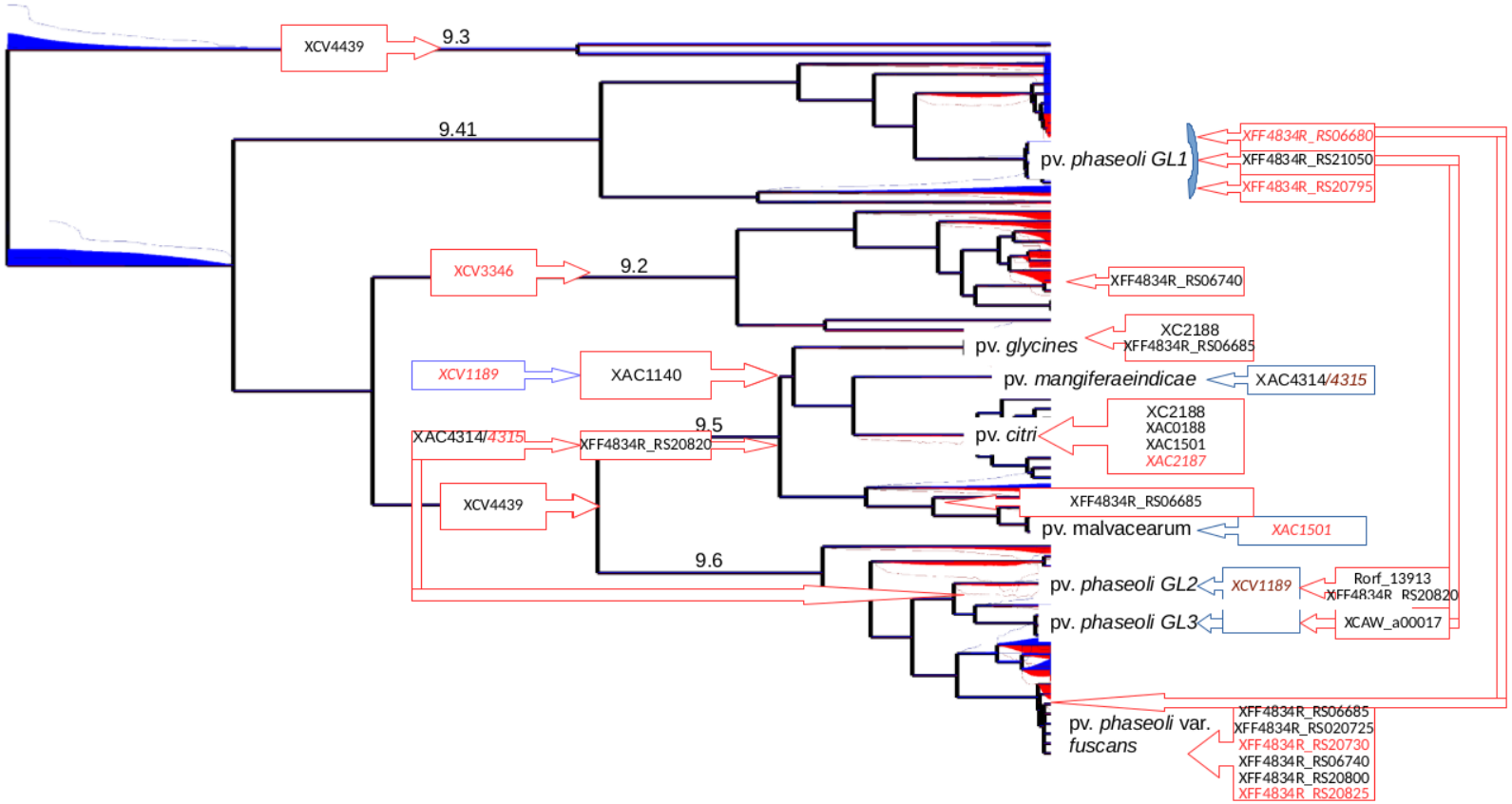


FIGURE 48 – Localisation des gains et pertes de TA.
Les gains sont indiqués dans des rectangles rouges, les pertes dans des rectangles bleus. Les gènes de toxine sont en rouge, et les gènes d'antitoxine en noir

accessoire montrent une évolution plus fortement influencée par l'écologie que celles issues du *core genome* [Wiedenbeck and Cohan, 2011]. Par exemple, une étude sur *E. coli* [Didelot et al., 2012] montre une discordance pour certains clades entre la topologie des arbres des génomes core et accessoire. La topologie basée sur le génome accessoire regroupe des souches du même pathotype, c'est à dire des souches qui partagent les mêmes traits d'histoire de vie, ce que ne montre pas la topologie issue du *core genome*.

Concernant notre étude, en dépit du fort taux de gains estimé, un signal phylogénétique précis et robuste est présent dans la matrice de présence-absence des gènes du génome accessoire qui permet d'identifier les mêmes groupes et pathovars qu'avec la phylogénie basée sur le *core genome*. Cependant, la topologie générale de l'arbre basé sur le génome accessoire, diffère de celle obtenue précédemment sur le *core genome* **(fig. 18)**. On note, par exemple, que l'arbre basé sur la matrice de présence-absence rassemble les groupes 9.6 et 9.2 puis le groupe 9.5, ce qui ne correspond pas à la phylogénie issue du *core genome*. La recombinaison non-homologue par HGT se produirait donc plus fréquemment entre les groupes 9.2 et 9.6 génétiquement plus éloignés qu'entre les groupe 9.5 et 9.6, pourtant décrits précédemment comme plus proches. Toutefois, ce regroupement ne peut pas s'expliquer totalement par la spécialisation à l'hôte. En effet, ces trois groupes contiennent chacun des souches pathogènes sur agrumes. De plus, aucun hôte n'est spécifique aux souches des groupes 9.6 et 9.2. Les génomes du groupe 9.5 sont ceux qui contiennent en moyenne le plus grand nombre de TA. Leur rôle dans ce groupe pourrait être de stabiliser le génome (en contribuant au maintien de plasmides, ou d'îlot génomique) en empêchant l'introduction d'ADN étranger [Van Melderen and Saavedra De Bast, 2009]. Le plus faible flux de gène par HGT entre 9.5 et 9.6, qu'entre 9.6 et 9.2 confirmerait la présence de barrières entre les groupes 9.5 et 9.6.

Un autre exemple de non congruence existe entre la phylogénie basée sur le *core genome*, et celle basée sur la matrice des Toxine-Antitoxine. Cette matrice regroupe les souches des groupes 9.41 et 9.6. Ce résultat illustrerait la présence de HGT entre des groupes génétiquement distants concernant ces gènes. De manière intéressante, les TA sont majoritairement portées par des plasmides [Unterholzner et al., 2013]. Il est alors probable que les transferts horizontaux entre les groupes

9.41 et 9.6 aient pu être facilités par la proximité écologique des certaines souches de ces deux groupes, notamment celles pathogènes sur haricot (genre *Phaseolus*).

4.2 La recombinaison homologue et le HGT ne se produisent pas au même moment dans l'histoire évolutive

Un très grand nombre d'événements de gains et de pertes se sont produits le long des branches terminales et suggère un rôle important des transferts horizontaux dans l'histoire évolutive récente du complexe d'espèces *Xanthomonas axonopodis*. Des fluctuations considérables du nombre de gains ou de pertes sur des périodes évolutives relativement courtes sont observées sur les branches conduisant aux pathovars. Une précédente étude sur des *Pseudomonas syringae* [Nowell et al., 2014] a montré que les gains et les pertes se produisent plus fréquemment sur les branches terminales, et que leur nombre diminue le long des branches à mesure que l'on se rapproche de la racine de la phylogénie. Une des raisons de cette diminution du HGT le long des branches plus profondes de la phylogénie, réside dans la difficile estimation des évènements de gains anciens. En effet, une majorité des gains de gènes sont transitoires et rapidement éliminés du génome [Kuo and Ochman, 2009][Touchon et al., 2009]. Nous observons aussi un excès de gains par rapport aux pertes sur les branches terminales. Cette observation impliquerait donc une augmentation récente de la taille des génomes des souches de *Xanthomonas axonopodis*. Cependant les souches du complexe *X. axonopodis* ont une taille de génome comparable aux autres *Xanthomonas*. Ainsi il faudrait conclure à une augmentation générale de la taille du génome au sein du genre *Xanthomonas*. Il semblerait plus probable qu'il s'agisse d'un *biais* méthodologique conduisant à une surestimation des taux de gains ou de pertes [Zhaxybayeva et al., 2007][Hao and Golding, 2008]. A ce titre, des problèmes d'assemblage de génome peuvent conduire à une mauvaise estimation du nombre de gènes. Certains gènes peuvent manquer à cause d'une couverture incomplète du génome entier. Ainsi des parties de gènes peuvent se trouver dans les parties non ou mal assemblées du génome. D'autres gènes peuvent être coupés à cause d'erreur de cadre de lecture [Han et al., 2013]. Ainsi, ces absences artefactuelles de gènes, induisent un *biais* dans l'estimation du génome accessoire, et par conséquent dans les gains inférés par Genoplast à

certaines souches. Les gènes codant pour les TA fonctionnent en couple : un gène codant pour la toxine et son gène d'antitoxine associé. Dans certains génomes, des gènes de toxines seuls ont été observés, impliquant *de facto* l'inférence d'une perte d'antitoxine. Or, la présence d'une toxine seule serait responsable de la mort de la cellule. On peut alors supposer que ces pertes d'antitoxines seules seraient probablement dues à des problèmes d'annotations, ou d'assemblage. Cependant, un tel *biais* d'annotation semble rare, seul 4 gènes d'antitoxines sur les 35 recherchés, seraient délétés dans 13 génomes au total. La profondeur de séquençage de la majorité de nos génomes est de 100X au minimum, ce qui nous conforte dans la fiabilité de nos résultats.

Tout comme les flux de gènes décrits au chapitre I, l'augmentation récente du HGT sur les branches terminales pourrait aussi être expliquée par des contacts secondaires liés la modification des pratiques culturales et à la mondialisation des échanges commerciaux. Bien que les *Xanthomonas* soient des bactéries phytopathogènes avec différentes gammes d'hôtes, dans certains contextes agricoles (cultures intercalaires ou rotation) la proximité physique des plantes contaminées par des agents pathogènes distincts peut faciliter le contact et le HGT. Par exemple, des espèces de *Xanthomonas* proches, *X. gardneri* et *X. citri* pv. *citri* portent le même plasmide [Richard et al., 2017] ce qui suggère que ces bactéries ont été en contact malgré des hôtes différents, la tomate et les agrumes respectivement. Le pathovar *manihotis* qui appartient au groupe 9.41, est celui qui a gagné le plus de gènes (420) d'après l'analyse Genoplast. Le groupe 9.41 est aussi le groupe qui contient le plus d'IS par génome (52) ce qui indique que ce groupe est fortement exposé aux HGT. L'augmentation récente du HGT vers ce pathovar pourrait avoir un lien avec la démographie en expansion que nous avons identifié précédemment. Une augmentation de l'aire de répartition des bactéries suite à l'expansion pourrait permettre une plus grande exposition à de nouveaux pools de gènes. Les souches du pathovar *manihotis* hébergent trois plasmides [Lin et al., 1979][Verdier, 1988], possiblement acquis récemment, ce qui pourrait expliquer le grand nombre de gènes gagnés récemment. Cependant, bien que ces plasmides aient été décrits, aucune séquence n'est pour l'instant disponible dans les bases de données. Il nous est donc impossible de vérifier l'hypothèse d'une acquisition récente de ces plasmides par le pathovar *manihotis*. Les deux souches non pathogènes du groupe 9.6 CFBP 7765

et CFBP 7923 ont aussi gagné énormément de gènes (556). Nous avons montré que la souche CFBP 7765 est plus perméable au flux de gènes par la recombinaison homologue que les autres souches de ce groupe (paragraphe III.3.3.5.1 et **figure 26**). Elle pourrait donc aussi être plus perméable au HGT. L'importance des gains inférés sur la branche précédent ces deux souches pourrait aussi s'expliquer par la longueur de cette branche qui les sépare des autres pathovars. Ces souches auraient donc eu plus le temps pour accumuler des différences de contenu de génome.

Une autre force évolutive, la recombinaison homologue ne se produit pas au même moment que le HGT dans l'histoire évolutive de ce complexe d'espèces *Xanthomonas axonopodis*. Elle se produit plus fréquemment sur les branches précédents les événements majeur de divergence, sur les branches conduisant au groupes par exemple (voir **fig. 24)** que sur les branches terminales. Ce patron de recombinaison pendant l'adaptation suivi de faible recombinaison une fois adapté semble être commun à beaucoup de bactéries pathogènes [Didelot and Maiden, 2010]. Une fois un maximum adaptatif atteint, la recombinaison devient délétère puisqu'elle réinjecte des allèles mal adaptés dans la population. L'isolement reproductif permet au contraire de garder les combinaisons favorables de gènes dans la population. Plusieurs agents pathogènes très spécialisés montrent peu de signe de recombinaison par rapport à leurs parents non spécialisés [Achtman and Wagner, 2008]. C'est le cas par exemple chez *Mycobacterium tuberculosis* [Liu et al., 2006], et chez *Xanthomonas arboricola* [Merda et al., 2016a]. Quel que soit le mécanisme provoquant la diminution de la recombinaison homologue, cela aboutira à une augmentation de la divergence des séquences, et à la spéciation. Ce processus de spécialisation conduit par une réduction de la recombinaison homologue est particulièrement attrayant par le parallèle qu'il offre avec le BSC de Mayr [Mayr, 1942]. Des mécanismes d'isolements reproductifs autres que ceux du à la divergence des séquences existent. Un premier mécanisme pourrait être l'isolement par la distance, qui empêche les interactions entre les souches qui favoriseraient la recombinaison [Wielgoss et al., 2016]. Ce ne semble pas être le cas pour nos groupes 9.5 et 9.6, car les souches proviennent des mêmes aires géographiques (continents), certains hôtes sont communs entre ces groupes, et les cas d'isolement allopatrique chez les bactéries sont rares [Fenchel, 2003]. Un deuxième mécanisme serait que la divergence des séquences diminuerait l'efficacité de la recombinaison, mais le fait que

nos 2 groupes 9.5 et 9.6 échangent moins que les groupes plus divergent 9.6 et 9.41 ne s'accorde pas avec cette hypothèse. Un troisième serait, la présence d'éléments génétiques, comme des systèmes de Restriction-Modification (RM), qui empêcheraient la réussite de la recombinaison entre des lignées. Il existe un parallèle entre ces systèmes et certains systèmes TA [Mruk and Kobayashi, 2014]. L'abondance des systèmes TA dans les génomes du groupe 9.5 pourrait être un élément en faveur de cette hypothèse. Un dernier mécanisme serait, qu'une différenciation de niche provoquée par l'acquisition de différentes fonctions ait induit la divergence de ces groupes, comme l'acquisition de XACSR1 dans le cas du groupe 9.5 que nous développons ci-dessous. Nos résultats ne nous permettent pas de choisir entre ces deux dernières hypothèses.

4.3 Divergence induite par l'adaptation

Les mécanismes de transferts horizontaux de gènes (HGT) sont à la base de l'évolution des génomes bactériens et concernent souvent des gènes impliqués dans les interactions hôtes-bactéries (incluant le pouvoir pathogène) et l'adaptation aux niches écologiques. Par exemple, chez *X. oryzae* un événement de HGT a été détecté impliquant des gènes de la biosynthèse des lipopolysaccharides [Patil et al., 2007]. Le HGT et les pertes de gènes peuvent directement faciliter l'accumulation de divergence entre des populations bactériennes *via* l'adaptation différentielle des lignées à des niches écologiques distinctes [Polz et al., 2013]. Ce mécanisme entre en jeu dans l'évolution et pourrait expliquer la divergence des groupes, et la large gamme d'hôte du complexe d'espèces *X. axonopodis*. Le cluster de biosynthèse de LPS précédemment décrit comme étant un élément de la spécificité d'hôte de *X. c.* pv. *citri* est partagé par tous les membres du groupe 9.5 et pourrait être un des facteurs ayant contribué à la divergence de ce groupe. Les LPS sont composés de trois entités synthétisées séparément : le lipide A, le noyau et l'antigène O. Le lipide A, nommé aussi l'endotoxine, est la partie responsable de l'induction de la réponse immunitaire non spécifique. L'antigène O influence la relation hôte-bactérie à différents niveaux, il représente la partie immunogénique spécifique du LPS. Chez les bactéries phytopathogènes, le LPS est un important facteur de virulence, il est aussi de plus en plus reconnu comme un motif moléculaire associé

aux agents pathogènes (PAMP) majeur reconnu par les plantes et induisant l'expression de gènes liés aux défenses des plantes, comme la production de composés phénoliques, la suppression de réaction d'hypersensibilité (HR) [Patil et al., 2007]. Les modifications de l'antigène O semblent jouer un rôle important dans plusieurs stades de l'infection, dans la phase de colonisation (adhérence) et dans la capacité de contourner ou surmonter les systèmes de défense de l'hôte [Lerouge and Vanderleyden, 2002]. Les 3 gènes impliqués dans la voie catabolique du D-galacturonate spécifique du groupe 9.6, pourraient aussi avoir contribué à la divergence de ce groupe. L'acide D-galacturonique est un des composants majeur de la pectine, un polysaccharide constituant les parois cellulaires des plantes. C'est une importante source de carbone pour les microorganismes vivant sur des végétaux [Zhang et al., 2011]. L'acquisition de ces fonctions a pu apporter une adaptation différentielle à une nouvelle niche écologique et favoriser la divergence de ce groupe.

La convergence pathologique ne semble pas due à des caractères acquis par tous les pathovars pathogènes sur un même hôte, mais plutôt par plusieurs acquisitions de gènes différents entre les pathovars. Il n'y a pas de gènes partagés par chacune des lignées du groupe 9.6 (*X. c.* pv. *fuscans* lignée *fuscans*, LG2 et LG3) et la lignée 1 (*X. p.* pv. *phaseoli*) du groupe 9.41. Il existe peut-être un *biais* géographique car les lignées LG2 et LG3 contiennent uniquement des souches isolées à la Réunion, ce qui pourrait expliquer que peu de gènes soient partagés avec la lignée *X. c.* pv. *fuscans*. Malgré cela les deux lignées de la Réunion LG2 et LG3 ne partagent pas plus de gènes accessoires entre elles (1 gène) que chacune de ces lignées avec *fuscans* (1 gène). L'acquisition de plusieurs gènes différents entre les pathovars d'un même hôte rejoint l'hypothèse que la spécificité de l'hôte résulterait de l'interaction entre les répertoires de gènes de virulence bactériens et les répertoires de gènes impliqués dans les défenses de l'hôte [Hajri et al., 2009].

Cinquième partie

Conclusion et Perspectives

L'objectif de cette étude était d'identifier les principales forces évolutives impliquées dans la divergence des bactéries du complexe d'espèces *Xanthomonas axonopodis*. Nous avons tout d'abord étudié la structuration, et confirmé cinq groupes sur des données génomiques. Les différents tests de neutralité indiquent des populations à l'équilibre démographique excepté pour le groupe 9.41 qui semble en expansion. Cependant, les différents scénarios d'inférence démographique testés avec $\delta a \delta i$ montrent que lorsqu'on se place au niveau de paires de populations, et qu'on prend en compte la migration, les meilleurs modèles sont des modèles avec changement de taille de populations (ex : IMex2, SCex2).

Nous avons montré que l'impact de la recombinaison est équivalent à celui de la mutation dans l'histoire évolutive de ce complexe d'espèces. La recombinaison homologue semble avoir joué un rôle important juste avant la divergence des cinq groupes. Ensuite nous avons observé une diminution des évènements de recombinaisons, sans doute liée à un isolement reproductif ou géographique ayant conduit à la divergence entre ces 5 cinq groupes. Cette divergence a pu être induite par l'accumulation de mutations et par l'effet de la dérive, ou provoquée par des gains (ou des pertes) de gènes favorisant l'adaptation d'une population à un nouvel environnement. Nous avons pu identifier, par exemple, le gain d'un opéron codant pour des LPS au niveau du groupe 9.5. Ce cluster n'est pas retrouvé par Blastn au sein des autres génomes du genre *Xanthomonas*. Des études fonctionnelles de délétion et de complémentation des gènes de ce cluster pourraient apporter des réponses sur son rôle dans la divergence ou l'adaptation des souches du groupe 9.5 à leur environnement.

Après la divergence des 5 groupes, la recombinaison homologues entre souches de groupes différents ne semble être pas avoir été contrainte par la proximité génétique des souches, mais plutôt favorisée par la proximité écologique (*i.e.* l'hôte). Ainsi un flux de gènes plus intense a été identifié entre des souches de groupes différents, pathogènes sur un même hôte. Ce résultat est confirmé par le transfert horizontal de gènes qui se produit par exemple plus fréquemment entre des groupes éloignés génétiquement comme 9.6 et 9.2 dont certaines souches ont des hôtes communs (les agrumes) qu'entre des groupes moins divergents comme 9.5 et 9.6. Ces groupes 9.5 et 9.6 ont aussi des hôtes communs, mais il semble exister entre les souches de ces deux groupes des barrières génétiques ou écologiques aux

flux de gènes. Dans le cas des deux groupes 9.5 et 9.6 le scénario démographique le plus probable est celui impliquant une divergence avec faible flux de gènes continu. Cependant, lorsque tous les SNP sont utilisés le meilleur modèle est un modèle divergence avec contact secondaire et flux de gènes hétérogène (SC2M2Pex2) ce qui confirme bien l'hypothèse de barrières. Cette différence dans le choix du modèle est révélatrice de la difficulté chez les bactéries de choisir suffisamment de SNP non-liés. Cette difficulté constitue ici une importante limite dans la puissance des méthodes d'inférence démographique quand elles sont appliquées aux procaryotes.

Il serait intéressant de constituer d'autres jeux de données qui ne seraient pas sous structurés (comme c'est le cas des groupes qui sont constitués de pathovars). Ainsi un échantillonnage plus important d'individus d'une même populations d'agents pathogènes de mêmes hôtes et des mêmes régions pourrait permettre de tester l'inférence de scénarios démographiques, entre par exemple, les pathovars *X. citri* pv. *fuscans* et *X. phaseoli* pv. *phaseoli*, ou entre *X. citri* pv. *citri*, *X. citri* pv. *aurantifollii*, et *X. euvesicatoria* pv. *citrumelonis*. La recherche de barrières seraient probablement facilitée sur ces jeux de données plus homogènes, les traces de sélection n'étant pas brouillées par des effets contradictoires entre différents pathovars. Ainsi des genome scans réalisés avec différentes statistiques, comme des F_{ST}, du déséquilibre de liaison, des indices de diversité, ou des test de McDonald–Kreitman [McDonald and Kreitman, 1991] sur ces différents jeux de données pourraient révéler avec plus de précisions les régions sous sélection.

Contrairement à la recombinaison homologue, le HGT s'est lui intensifié récemment. En effet, Genoplast a inféré beaucoup de gains et de pertes sur les branches terminales de l'arbre qui pourraient refléter des échanges entre des souches remises en contact par la mondialisation ou lES modifications des pratiques agricoles. Une des perspectives de ce travail pourrait être d'inclure à notre collection *i)* de véritables populations (souches prélevées ensemble sur un site géographique) et *ii)* une plus forte proportion de génome de souches isolées très récemment pour confirmer le scénario de contact secondaire.

Pour finir, notre étude décrit un modèle de divergence qui ne pourrait peut-être pas être généralisé à toutes les bactéries mais qui apporte des éléments pour comprendre le processus de spéciation chez les bactéries phytopathogènes. Au vu de ces résultats l'étude des flux de gènes aurait besoin d'être étendue à d'autres

bactéries afin de généraliser un concept d'espèce bactérienne basé sur une réalité biologique.

Annexe A

Liste des gènes spécifiques et des annotations fonctionnelles pour chaque groupe

Groupe		GO	Family	Integration	Description
9.41	WP_017157620|multidrug efflux RND transporter permease subunit [Xanthomonas axonopodis]	GO0006810 GO0005215 GO0016020	PF00873	Acriflavin resistance protein (IPR001036)	The Escherichia coli acrA and acrB genes encode a multi-drug efflux system that is believed to protect the bacterium against hydrophobic inhibitors
	gi|746539912|ref|WP_039572264.1|chitinase [Xanthomonas axonopodis]	GO0004553 GO0005975	PF00704 SM00636	Glycoside hydrolase family 18, catalytic domain (IPR001223) Chitinase II (IPR011583)	O-Glycosyl hydrolases (EC:3.2.1.) are a widespread group of enzymes that hydrolyse the glycosidic bond between two or more carbohydrates, or between a carbohydrate and a non-carbohydrate moiety.
	NA	GO0016301 GO0006468 GO0005524 GO0004672	PS50290	Phosphatidylinositol 3-/4-kinase, catalytic domain (IPR000403) Protein kinase domain (IPR000719) Ricin B, lectin domain (IPR000772)	This domain is present in a wide range of protein kinases, involved in diverse cellular functions, such as control of cell growth, regulation of cell cycle progression, a DNA damage checkpoint, recombination, and maintenance of telomere length. Despite significant homology to lipid kinases, no lipid kinase activity has been demonstrated for any of the PIK-related kinases [PMID: 12456783].
	NA	NA	PS50231	Ricin B, lectin domain (IPR000772)	Primary structure analysis has shown the presence of a similar domain in many carbohydrate-recognition proteins like plant and bacterial AB-toxins, glycosidases or proteases [PMID: 9603958, PMID: 7664090, PMID: 8844840]. This domain, known as the ricin B lectin domain, can be present in one or more copies and has been shown in some instance to bind simple sugars, such as galactose or lactose.
	WP_057683835|hypothetical protein [Xanthomonas axonopodis]				
	gi|515723723|ref|WP_017156323.1|DNA polymerase I [Xanthomonas axonopodis]	GO0006260 GO0003887 GO0003677 GO0003676 GO0008408 GO0006139 GO0003824 GO0004553	PF00476 PF01612 PF01367 PF02739	DNA-directed DNA polymerase, family A, palm domain (IPR001098)	DNA-directed DNA polymerases(EC:2.7.7.7) are the key enzymes catalysing the accurate replication of DNA This domain is responsible for the 3'-5' exonuclease proofreading activity of Escherichia coli DNA polymerase I (polI) and other enzymes, it catalyses the hydrolysis of unpaired or mismatched nucleotides
9.2	EGD14785|transcriptional regulator [Xanthomonas perforans 91-118]	GO0006355 GO0003700	PF00126	Transcription regulator HTH, LysR (IPR000847)	Numerous bacterial transcription regulatory proteins bind DNA via a helix-turn-helix (HTH) motif. The majority of these proteins appear to be transcription activators and most are known to negatively regulate their own expression
	KLB55430|cinnamyl-alcohol dehydrogenase [Xanthomonas euvesicatoria].	GO0045551 GO0052747 GO0050662	PF01370	NAD-dependent epimerase/dehydratase (IPR001509)	This domain is found in proteins that utilise NAD as a cofactor and use nucleotide-sugar substrates for a variety of chemical reactions [PMID: 9174344]
	WP_031425249|LysR family transcriptional regulator [Xanthomonas axonopodis]	GO0003677 GO0006351	PF00126	Transcription regulator HTH, LysR (IPR000847)	LysR-type transcriptional regulators (LTTRs) regulate a diverse set of genes, including those involved in

		GO0003700			virulence, metabolism, quorum sensing and motility
9.5	gi|489577626|ref|WP_003482072.1|MULTISPECIES: aryl sulfotransferase [Xanthomonas]	GO0008146	PF00685	Sulfotransferase domain (IPR000863)	These enzymes are responsible for the transfer of sulphate groups to specific compounds
	gi|515318003|ref|WP_016850972.1|ligase [Xanthomonas axonopodis]	GO0016021 GO0016874 GO0008152	PF04932	O-antigen ligase-related (IPR007016)	This group of bacterial proteins is involved in the synthesis of O-antigen, a lipopolysaccharide found in the outer membrane in Gram-negative bacteria
	gi|489577652|ref|WP_003482098.1|hypothetical protein [Xanthomonas citri]	GO0016021	PS51257		NA
	CEF38679 |conserved hypothetical protein [Xanthomonas citri pv. citri]	NA			
	CEE15480|conserved hypothetical protein [Xanthomonas citri pv. citri]	GO0016021	NA		
	AJZ64715|hypothetical protein J168_00528 [Xanthomonas citri subsp. citri]	NA	PF14907		Uncharacterised nucleotidyltransferase
	gi|21106091|gb|AAM34940.1|hypothetical protein XAC0048 [Xanthomonas axonopodis pv. citri str. 306]	GO0016021	NA		
	gi|749509152|ref|WP_040149208.1|undecaprenyl-phosphate glucose phosphotransferase [Xanthomonas citri]	GO0016021 GO0016740 GO0008152	PF02397 PF13727	Bacterial sugar transferase (IPR003362)	This entry represents a conserved region from a number of different bacterial sugar transferases, involved in diverse biosynthesis pathways
	gi|469765176|gb|AGH75593.1|polysaccharide export protein [Xanthomonas axonopodis Xac29-1]	GO0016020	PF01943	Polysaccharide biosynthesis protein (IPR002797)	Members of this family are integral membrane proteins [PMID: 8118055], and many are implicated in the production of polysaccharide. The family includes RfbX part of the O antigen biosynthesis operon [PMID: 7517390], and SpoVB from Bacillus subtilis (Q00758), which is involved in spore cortex biosynthesis [PMID: 1744050].
	gi|746548904|ref|WP_039580964.1|ligand-gated channel [Xanthomonas axonopodis]	GO0009279 GO0004872 GO0005506 GO0015891	PF00593	TonB-dependent receptor, beta-barrel (IPR000531)	In Escherichia coli the TonB protein interacts with outer membrane receptor proteins that carry out high-affinity binding and energy-dependent uptake of specific substrates into the periplasmic space [PMID: 14499604].
	gi|499361751|ref|WP_011050060.1|MULTISPECIES: hypothetical protein [Xanthomonas]	GO0005622 GO0000155 GO0005524 GO0000160 GO0006109 GO0023014	G3DSA:3.40.50.300	P-loop containing nucleoside triphosphate hydrolase (IPR027417)	The P-loop NTPase fold is the most prevalent domain of the several distinct nucleotide-binding protein folds.
	gi|492677552|ref|WP_005919999.1|MFS transporter [Xanthomonas axonopodis]	GO0016021 GO0055085	PF07690	Major facilitator superfamily (IPR011701)	Among the different families of transporter, only two occur ubiquitously in all classifications of organisms. These are the ATP-Binding Cassette (ABC) superfamily and the Major Facilitator Superfamily (MFS). The MFS transporters are single-polypeptide secondary carriers capable only of transporting small

					solutes in response to chemiosmotic ion gradients [PMID: 9529885, PMID: 9868370].
	gi|21106101|gb|AAM34949.1|transport protein [Xanthomonas axonopodis pv. citri str. 306]	GO0016887 GO0005524 GO0055085 GO0016021 GO0006810 GO0042626	PF00005 PF00664	ABC transporter-like (IPR003439) ABC transporter type 1, transmembrane domain (IPR011527)	ABC transporters belong to the ATP-Binding Cassette (ABC) superfamily, which uses the hydrolysis of ATP to energise diverse biological systems.ABC transporters are involved in the export or import of a wide variety of substrates ranging from small ions to macromolecules. The major function of ABC import systems is to provide essential nutrients to bacteria
	gi|492706839|ref|WP_005929055.1|MULTISPECIES: glycosyl transferase family 1 [Xanthomonas]	GO0008152 GO0016757	PF13439 PF13692	Glycosyltransferase subfamily 4-like, N-terminal domain (IPR028098)	
	CEE30917|conserved hypothetical protein [Xanthomonas citri pv. citri]	NA	PF05402	Coenzyme PQQ synthesis D (IPR008792)	This protein is required for coenzyme pyrrolo-quinoline-quinone (PQQ) biosynthesis
	gi|21106982|gb|AAM35740.1|TonB-dependent receptor [Xanthomonas axonopodis pv. citri str. 306]	GO0009279 GO0006810 GO0004872	PF00593 PF07715	TonB-dependent receptor, beta-barrel (IPR000531)	In Escherichia coli the TonB protein interacts with outer membrane receptor proteins that carry out high-affinity binding and energy-dependent uptake of specific substrates into the periplasmic space [PMID: 14499604].
	gi|489577667|ref|WP_003482113.1|MULTISPECIES: UDP-N-acetyl glucosamine 2-epimerase [Xanthomonas]	GO0008761 GO0016740	PF02350	UDP-N-acetylglucosamine 2-epimerase domain (IPR003331)	Proteins containing this domain also include UDP-N,N'-diacetylbacillosamine 2-epimerase, which is involved in biosynthesis of legionaminic acid (5,7-diamino-3,5,7,9-tetradeoxy-D-glycero-D-galacto-non-2-ulosonic acid)(Leg), a sialic acid-like derivative that is incorporated into virulence-associated cell surface glycoconjugates such as lipopolysaccharide (LPS) which could be a key determinant in the ability of L. pneumophila to inhibit the fusion of phagosomes with lysosomes
	gi|489577672|ref|WP_003482118.1|mannosyltransferase [Xanthomonas citri]	GO0016757	PF00534 PF13439	Glycosyl transferase, family 1 (IPR001296)	The biosynthesis of disaccharides, oligosaccharides and polysaccharides involves the action of hundreds of different glycosyltransferases. These enzymes catalyse the transfer of sugar moieties from activated donor molecules to specific acceptor molecules, forming glycosidic bonds. A classification of glycosyltransferases using nucleotide diphospho-sugar, nucleotide monophospho-sugar and sugar phosphates (EC:2.4.1.-)
	gi|492706807|ref|WP_005929035.1|MULTISPECIES: transposase [Xanthomonas]	GO0003978 GO0050662	PF01370	NAD-dependent epimerase/dehydratase (IPR001509)	This domain is found in proteins that utilise NAD as a cofactor and use nucleotide-sugar substrates for a variety of chemical reactions [PMID: 9174344].
	gi|21106084|gb|AAM34934.1|glycosyltransferase [Xanthomonas axonopodis pv. citri str. 306]	GO0016757	PF00534	Glycosyl transferase, family 1 (IPR001296)	The biosynthesis of disaccharides, oligosaccharides and polysaccharides involves the action of hundreds of

					different glycosyltransferases
	AAM34952|hypothetical protein XAC0060 [Xanthomonas axonopodis pv. citri str.306]	GO0008158 GO0032259	PF13847	Methyltransferase domain (IPR025714)	This domain appears to be a methyltransferase domain.
	gi|21106089|gb|AAM34938.1|UDP-N-acetyl-D-mannosamine transferase [Xanthomonas axonopodis pv. citri str. 306]	GO0016757 GO0009058	PF03808	Glycosyl transferase WecB/TagA/CpsF (IPR004629)	Sugar and amino sugar residues are converted to sugar nucleotides prior to their incorporation into structural polysaccharides via UDP-sugar transferases. In this pathway, UDP-N-acetyl-D-glucosamine (UDP-GlcNAc) is the amino sugar nucleotide donor of N-acetyl-D-glucosamine (GlcNAc) residues for the biosynthesis of glycosylated proteins and cell surface structures.. Also included in this group is Xanthomonas campestris pv. campestris GumM, a glycosyltransferase participating in the biosynthesis of the exopolysaccharide xanthan [PMID: 8830246, PMID: 11673418, PMID: 12618464, PMID: 18156271, PMID: 16953575, PMID: 3275612, PMID: 9537354].
	gi|492653097|ref|WP_005912530.1|MULTISPECIES: asparagine synthetase B [Xanthomonas]	GO0004066 GO0006529	PF00733 PF13537	Asparagine synthase (IPR001962) Glutamine amidotransferase type 2 domain (IPR017932)	Family members that contain this domain catalyse the conversion of aspartate to asparagine. Asparagine synthetase B (EC:6.3.5.4) A large group of biosynthetic enzymes are able to catalyse the removal of the ammonia group from glutamine and then to transfer this group to a substrate to form a new carbon-nitrogen group
	gi|372554053|emb|CCF68489.1|methyltransferase [Xanthomonas axonopodis pv. punicae str. LMG 859]	GO0008168 GO0032259	PF05711	Macrocin-O-methyltransferase (IPR008884)	This family consists of bacterial macrocin O-methyltransferase (TylF) proteins. TylF is responsible for the methylation of macrocin to produce tylosin. Tylosin is a macrolide antibiotic used in veterinary medicine to treat infections caused by Gram-positive bacteria and as an animal growth promoter in the Sus scrofa (Pig) industry. It is produced by several Streptomyces species. As with other macrolides, the antibiotic activity of tylosin is due to the inhibition of protein biosynthesis by a mechanism that involves the binding of tylosin to the ribosome, preventing the formation of the mRNA-aminoacyl-tRNA-ribosome complex [PMID: 10220165].
	gi|489577655|ref|WP_003482101.1|MULTISPECIES: methyltransferase type 12 [Xanthomonas]	GO0032259 GO0008168	PF13489 G3DSA:3.40.50.150	S-adenosyl-L-methionine-dependent methyltransferase (IPR029063)	This entry represents a domain found in S-adenosyl-L-methionine-dependent methyltransferases (SAM MTases
	gi|515315363|ref|WP_016849966.1|TonB-dependent receptor [Xanthomonas axonopodis]		PF00593 PF07715	TonB-dependent receptor, beta-barrel (IPR000531)	In Escherichia coli the TonB protein interacts with outer membrane receptor proteins that carry out high-affinity binding and energy-dependent uptake of

					specific substrates into the periplasmic space [PMID: 14499604].
9.6	WP_007966013|LacI family transcriptional regulator [Xanthomonas]	GO0003677 GO0003700 GO0006351 GO0006355	PF00356 PF13377	LacI-type HTH domain (IPR000843)	The lacI-type HTH domain is a DNA-binding, helix-turn-helix (HTH) domain of about 50-60 residues present in the lacI/galR family of transcriptional regulators involved in metabolic regulation in prokaryotes. Most of these bacterial regulators recognize sugar-inducers
	KLB09953|integration host factor subunit alpha [Xanthomonas gardneri]	GO0005829 GO0003677 GO0006310 GO0006351 GO0006355 GO0006417	PF00216	Histone-like DNA-binding protein (IPR000119)	The integration host factor (IHF), a dimer of closely related chains which seem to function in genetic recombination as well as in translational and transcriptional control [PMID: 2972385] is found in enterobacteria and viral proteins include the African Swine fever virus protein A104R (or LMW5-AR) [PMID: 8464748].

Annexe B

Liste des souches bactériennes utilisées pour les analyses génomiques.

Code génome	*Nom dans les collections*	*Origine*	*Année d'isolement*	*Hôte d'isolement*	*Nomenclature*	*Groupes selon* Rademaker *et al.* (2005)	*%GC*	*Nombre de contigs*	*Taille du génome (Mb)*	*Nombre de CDS*
Xab-CFBP 2524-G1	CFBP 2524^{PT}	Nouvelle-Zélande	1962	*Begonia sp.*	*"X. phaseoli* pv. *Begoniae"$*	9.1	64.75	108	5,07	4265
Xas-CFBP 2547-G1	CFBP 2547^{PT}	Martinique	1985	*Spondias cytherea*	*"X. phaseoli* pv. *Spondiae"$*	9.1 et 9.4*	65.23	167	4,55	4289^{R}
Xalfaa-CFBP 7686-G1	CFBP 7686^{T}	Indes	1954	*Medicago sativa*	*X. euvesicatoria* pv. *alfalfae*	9.2	64.80	46	5,02	4181
axalfa3836	CFBP 3836^{PT}	Soudan	1996	*Medicago sativa ()*	*X. euvesicatoria* pv. *alfalfae*	9.2	64.74	6	5,04	4246
Xalfac-CFBP 3371-G1	CFBP 3371	USA	1989	*Poncirus trifoliata X Citrus paradisi*	*X. euvesicatoria* pv. *citrumelonis*	9.2	64.94	46	4,93	4092
axmeloniF1	F1	USA	1984	*Citrus*	*X. euvesicatoria* pv. *citrumelonis*	9.2	64.92	1	4,96	4102
Xaa-CFBP 6107-G1	CFBP 6107^{PT}	Japon	1998	*Allium fistulosum*	*X. euvesicatoria* pv. *allii*	9.2	64.90	54	5,10	4372
XaaCFBP 6367	CFBP 6367	Barbades	2002	*Allium cepa*	*X. euvesicatoria.* pv.. *allii*	9.2	64.98	182	5,05	4430^{R}
axalli6369	CFBP 6369	Réunion	2003	*Allium cepa*	*X. euvesicatoria* pv. *allii*	9.2	64.36	3	5,42	4641
Xad-CFBP 5693-G1	CFBP 5693	USA	2001	*Philodendron scandens subsp. oxycardium ()*	*X. euvesicatoria*pv. *dieffenbachiae*	9.2	64.84	38	5,10	4305
Xap-CFBP 7277-G1	CFBP 7277^{PT}	Indes	1950	*Euphorbia pulcherrima*	*"X. euvesicatoria* pv. *Poinsettiicola"$*	9.2	64.94	77	5,03	4214
axeuve8510	85-10, CFBP 5618	USA		*Capsicum annuum*	*X. Euvesicatoria* pv. *euvesicatoria*	9.2	64.75	5	5,17	4691
perfor9118	91-118	USA	1991	*Solanum lycopersicum*	*X. euvesicatoria* pv. *perforans*	9.2	65.04	291	5,26	4632^{R}
XaCFBP 7916	CFBP 7916*, SNES 29		2010	*Phaseolus vulgaris*	*X. euvesicatoria*	9.2	64.85	106	5,05	4361
XaCFBP 7920	CFBP 7920*, SNES 5		2010	*Phaseolus vulgaris*	*X. euvesicatoria*	9.2	64.89	104	5,05	4334
Xaa-CFBP 4924-G1	CFBP 4924^{T}	Colombie	1949	*Axonopus scoparius*	*X. axonopodis* pv. *axonopodis*	9.3	64.46	288	4,51	4383^{R}
Xava-CFBP 5823-G1	CFBP 5823^{PT}	Mauritanie	2001	*Saccharum officinarum*	*X. axonopodis* pv. *vasculorum*	9.3	64.18	328	4,69	4529^{R}
axsyng9055	LMG 9055, CFBP 3451^{PT}	USA	1994	*Syngonium podophyllum*	*X. phaseoli* pv. *syngonii*	9.4	64.82	6	5,00	4455

Code génome	*Nom dans les collections*	*Origine*	*Année d'isolement*	*Hôte d'isolement*	*Nomenclature*	*Groupes selon* Rademaker *et al.* (2005)	*%GC*	*Nombre de contigs*	*Taille du génome (Mb)*	*Nombre de CDS*
pha1GLC412	CFBP 412	USA	1963	*Phaseolus vulgaris*	*X. phaseoli* pv. *phaseoli*	9.4	64.88	1	5,03	4268
pha1GL6164	CFBP 6164	Roumanie	1966	*Phaseolus vulgaris*	*X. phaseoli* pv. *phaseoli*	9.4	64.26	1	5,31	4678
pha1GL6984	CFBP 6984	Réunion	2000	*Phaseolus vulgaris*	*X. phaseoli* pv. *phaseoli*	9.4	64.74	1	5,10	4346
pha1GL7430	CFBP 7430	Iran	2006	*Phaseolus*	*X. phaseoli* pv. *phaseoli*	9.4	64.80	1	5,08	4333
pha1GL6546	CFBP 6546	USA		*Phaseolus vulgaris*	*X. phaseoli* . pv. *phaseoli*	9.4	64.91	4	4,99	4274
axmaniC151	CFBP 7661	Colombie	1995	*Cassavae*	*X. phaseoli* pv. *manihotis*	9.4	64.55	36	5,15	4637
axmani1851	CFBP 1851	USA	1978	*Manihot esculenta*	*X. phaseoli* pv. *manihotis*	9.4	65.18	123	4,79	4316^{R}
axmaniCIO1	CIO 1	Colombie	1995		*X. phaseoli* pv. *manihotis*	9.4	65.1	128	4,89	4424^{R}
axmaniI278	IBSBF278	Brésil	1965	*Manihot esculenta*	*X. phaseoli* pv. *manihotis*	9.4	64.89	138	5,02	4532^{R}
axmaniOX27	ORSTX27	Togo	1989	*Manihot esculenta*	*X. phaseoli* pv. *manihotis*	9.4	65.07	133	4,93	4427^{R}
axmaniUG21	CFBP 8188	Ouganda	2011		*X. phaseoli* pv. *manihotis*	9.4	65.06	148	5,01	4528^{R}
axdieff695	CFBP 3133^{PT}	Brésil	1965	*Anthurium sp.*	*X. phaseoli* pv. *dieffenbachiae*	9.4	64.88	1	5,03	4234
axmangL941	CFBP 2531^{PT}	Indes	1957	*Mangifera indica*	*X. citri* pv. *mangiferaeindicae*	9.5	64.85	195	5,11	4547^{R}
axmang5610	LG56-10				*X. citri* pv. *mangiferaeindicae*	9.5	64.56	4	5,28	4521
axmang8127	LG81-27				*X. citri* pv. *mangiferaeindicae*	9.5	64.68	6	5,20	4461
Xavit-CFBP 7764-G1	CFBP 7764	Brésil	2012	*Vitios vinifera*	*"X. citri* pv. *Viticola"$*	9.5	64.32	76	5,31	4504
axglyc12_2	12_2	Thaïlande		*Soybean, Glycine max*	*X. citri* pv. *glycines*	9.5	64.36	465	5,2	4714^{R}
axglyc2526	CFBP 2526^{PT}	Soudan	1956	*Glycine hispida*	*X. citri* pv. *glycines*	9.5	64.63	4	5,26	4480
axglyc7119	CFBP 7119	Brésil	1981	*Glycine max*	*X. citri* pv. *glycines*	9.5	64.63	4	5,25	4741
axmalv2388	GSPB2388	Soudan	1994	*Gossypium sp.*	*X. citri.* pv. *malvacearum*	9.5	64.39	61	5,11	4347
axmalv1386	GSPB1386	Nicaragua	1986	*Gossypium sp.*	*X. citri.* pv.	9.5	64.53	127	4,97	4391

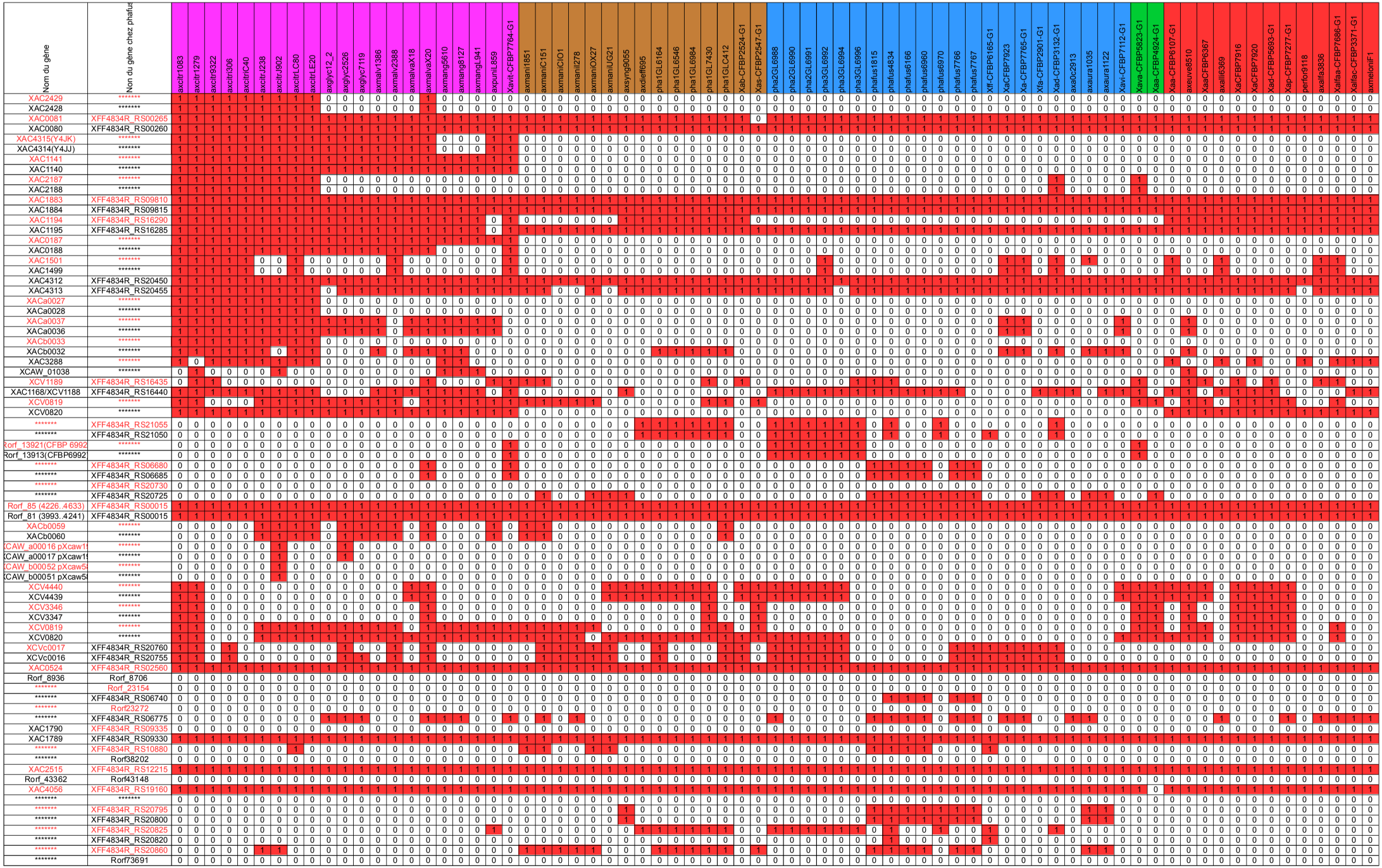
Code génome	*Nom dans les collections*	*Origine*	*Année d'isolement*	*Hôte d'isolement*	*Nomenclature*	*Groupes selon* Rademaker *et al.* (2005)	*%GC*	*Nombre de contigs*	*Taille du génome (Mb)*	*Nombre de CDS*
					malvacearum					
axmalvaX18	X18	Burkina Faso		*Gossypium sp.*	*X. citri* pv. *malvacearum*	9.5	64.73	4	4,99	4235
axmalvaX20	X20	Burkina Faso		*Gossypium sp.*	*X. citri* pv. *malvacearum*	9.5	64.46	4	5,16	4440
axpuniL859	LMG859^{PT}	Indes	1957	*Punica granatum*	*X. citri* pv. *punicae*	9.5	64.87	217	4,95	4413^{R}
axcitr9322	CFBP 3369^{T}	USA	1989	*Citrus aurantifolia*	*X. citri* pv. *citri*	9.5	64,73	1	5,19	4397
axcitr1279	A^{w}12879 - CP003778	USA		*Citrus aurantifolia*	*X. citri* pv. *citri*	9.5	64.68	3	5,32	4625
axcitri306	306				*X. citri* pv. *citri*	9.5	64.68	3	5,17	4485
axcitr1083	FDC1083	Brésil	1980	*C. reticulata*	*X. citri* pv. *citri*	9.5	64.71	3	5,27	4403
axcitrLC80	LC80	Mali	2006	*C. reticulata x C. sinensis*	*X. citri* pv. *citri*	9.5	64.69	1	5,22	4421
axcitrJ902	JF90-2	Oman	1986	*C.aurantifolia*	*X. citri* pv. *citri*	9.5	64.67	1	5,23	4515
axcitrJ238	JJ238-24	Thailande	1989	*C.aurantifolia*	*X. citri* pv. *citri*	9.5	64.63	1	5,25	4513
axcitriC40	C40	Réunion	1988	*Citrus sinensis*	*X. citri* pv. *citri*	9.5	64.75	1	5,28	4482
axcitrLE20	LE20-1	Ethiopie	2008	*C.aurantifolia*	*X. citri* pv. *citri*	9.5	64.67	3	5,27	4533
Xad-CFBP 3132-G1	CFBP 3132	USA	1950	*Dieffenbachia sp*	*X. citri*	9.6	64.49	185	5,15	4512^{R}
Xavi-CFBP 7112-G1	CFBP 7112^{PT}		1942	*Vigna sinensis*	*X. citri* pv. *vignicola*	9.6	64.69	82	5,08	4483
axanac2913	CFBP 2913^{PT}	Brésil		*Mangifera indica ()*	*X. citri* pv. *anacardii*	9.6	64,57	1	5,20	4296
phafus4834	4834R, CFBP 4885	France	1998	*Phaseolus vulgaris*	*X. citri* pv. *fuscans*	9.6	64.73	4	5,09	4290
phafus1815	CFBP 1815	Grèce	1978	*Phaseolus sp*	*X. citri* pv. *fuscans*	9.6	64.79	1	4,95	4200
phafus6166	CFBP 6166	Afrique du sud	1963	*Phaseolus vulgaris*	*X. citri* pv. *fuscans*	9.6	64.78	1	4,98	4238
phafus6960	CFBP 6960	Réunion	2000	*Phaseolus vulgaris*	*X. citri* pv. *fuscans*	9.6	64.77	1	4,99	4233
phafus6970	CFBP 6970	USA	1990	*Phaseolus sp*	*X. citri* pv. *fuscans*	9.6	64.88	1	5,00	4250
phafus7766	CFBP 7766	Cameroun	2009		*X. citri* pv. *fuscans*	9.6	64.60	1	5,18	4444
Xff-CFBP 6165-G1	CFBP 6165^{PT}	Canada	1957	*Phaseolus vulgaris*	*X. citri* pv. *fuscans*	9.6	64.91	133	4,94	4354
phafus7767	CFBP 7767	Cameroun	2009		*X. citri* pv. *fuscans*	9.6	64.67	1	5,10	4356
axaura1035	ICPB10535				*X. citri* pv. *aurantifolii*	9.6	64.81	237	5,01	4503^{R}
Xfa-CFBP	CFBP 2901^{PT}	Argentine	1982	*Citrus limon*	*X. citri* pv.	9.6	64.91	163	4,82	4310^{R}

Code génome	***Nom dans les collections***	***Origine***	***Année d'isolement***	***Hôte d'isolement***	***Nomenclature***	***Groupes selon*** Rademaker *et al.* (2005)	***%GC***	***Nombre de contigs***	***Taille du génome (Mb)***	***Nombre de CDS***
2901-G1					*aurantifolii*					
axaura1122	ICPB11122				*X. citri* pv. *aurantifolii*	9.6	64,88	237	4,88	4406[R]
pha2GL6990	CFBP 6990	Réunion	2000	*Phaseolus vulgaris*	*X. citri* pv. *fuscans*	9.6	64.62	1	5,12	4317
pha2GL6991	CFBP 6991	Réunion	2000	*Phaseolus vulgaris*	*X. citri* pv. *fuscans*	9.6	64.28	1	5,34	4526
pha3GL6992	CFBP 6992	Réunion	2000	*Phaseolus vulgaris*	*X. citri* pv. *fuscans*	9.6	64.45	1	5,51	4521
pha3GL6994	CFBP 6994	Tanzanie	1990	*Phaseolus vulgaris*	*X. citri* pv. *fuscans*	9.6	64.26	1	5,25	4452
pha2GL6988	CFBP 6988	Réunion	2000	*Phaseolus vulgaris*	*X. citri* pv. *fuscans*	9.6	64.64	1	5,13	4338
pha3GL6996	CFBP 6996	Réunion	2000	*Phaseolus vulgaris*	*X. citri* pv. *fuscans*	9.6	64.73	4	5,09	4186
Xa-CFBP 7765-G1	CFBP 7765*, SNES 27			*Phaseolus vulgaris*	*X. citri*	9.6	64.57	122	5,06	4441[R]
XaCFBP 7923	CFBP 7923*,SNES 9		2011	*Phaseolus vulgaris*	*X. citri*	9.6	64.57	239	5,11	4484[R]
Xvaho-CFBP 2543-G1	CFBP 2543	Nouvelle-Zélande	1969	*Sorghum vulgare*	*X. vasicola* pv. *holcicola*		63.3	135	5,4	4533[R]

CFBP (Collection Française de Bactéries associées aux Plantes), Groupes (groupe génétique) d'après Rademaker *et al.* (2005) et * Ah-You *et al.*2009, [R] Annotation génomique avec RAST. Les génomes soulignés proviennent des bases de données publiques, T : indique les souches types, PT : indique les souches pathotypes, * souches non pathogènes sur leur hôte d'isolement, $ nouveaux noms non formellement proposés (pathovar non inclus dans Constantin et al. (2016)).

Annexe C

Matrice de présence absence des gènes codant pour des Toxines Antitoxines. Le nom des gènes chez la souche *X. c.* pv. *citri* 306 est indiqué par XAC, le nom des gènes chez la souche *X. euvesicatoria* 8510 est indiqué par XCV, et pour les Orfs le nom de la souche est indiqué entre parenthèses. Les couleurs dans la $1^{ère}$ ligne indiquent l'appartenance des souches à un groupe : rouge pour 9.2, vert pour 9.3, marron pour 9.41, violet pour 9.5 et bleu pour 9.6. Les lignes rouges indiquent les Toxines, les noires les Antitoxines.


Nom du gène
Non du gène chez phafus
axcitr1083
axcitr1279
axcitr9322
axcitr306
axcitrC40
axcitrJ238
axcitrJ902
axcitrLC80
axcitrLE20
axglyc12_2
axglyc2526
axglyc7119
axmalv1386
axmalv2388
axmalvaX18
axmalvaX20
axmang5610
axmang8127
axmangL941
axpunL859
Xavit-CFBP7764-G1
axmani1851
axmaniC151
axmaniClO1
axmanil278
axmaniOX27
axmaniUG21
axsyng9055
axdieff695
pha1GL6164
pha1GL6546
pha1GL6984
pha1GL7430
pha1GLC412
Xab-CFBP2524-G1
Xas-CFBP2547-G1
pha2GL6988
pha2GL6990
pha2GL6991
pha3GL6992
pha3GL6994
pha3GL6996
phafus1815
phafus4834
phafus6166
phafus6960
phafus6970
phafus7766
phafus7767
Xfl-CFBP6165-G1
XaCFBP7923
Xa-CFBP7765-G1
Xfa-CFBP2901-G1
Xad-CFBP3132-G1
axa0c2913
axaura1035
axaura1122
Xavi-CFBP7112-G1
Xavra-CFBP5823-G1
Xca-CFBP4924-G1
Xaa-CFBP6107-G1
axeuve8510
XaaCFBP6367
axalil6369
XaCFBP7916
XaCFBP7920
Xad-CFBP5693-G1
Xap-CFBP7277-G1
perfor9118
axalfa3836
Xalfaa-CFBP7686-G1
Xalfac-CFBP3371-G1
axmelonF1

Annexe D

Domaines et fonctions des TA. La première colonne indique si le gène code pour une toxine (T) ou une antitoxine (A). Les autres colonnes indiquent le nom des gènes chez les diférentes souches.

Identification	Domaines	Nom du gène	Non du gène chez phafusc4834
T	COG3905 - Transcriptio0l regulator	XAC2429	*******
A	COG3668 - ParE	XAC2428	*******
T	pfam05016 - plasm. stab. prot. RelE/ParE	XAC0081	XFF4834R_RS00265
A	TIGR01552/pfam02604 - phd / COG2161 - StbD	XAC0080	XFF4834R_RS00260
T	cd09881 - PIN_VapC-FitB	XAC4315(Y4JK)	*******
A	COG4691 - StbC	XAC4314(Y4JJ)	*******
T	COG3668 - ParE	XAC1141	*******
A	TIGR02606 - antidote_CC2985	XAC1140	*******
T	cd09855 - PIN_VapC-Smg6-like	XAC2187	*******
A	COG2336 - MazE	XAC2188	*******
T	smart00966 - SpoVT/AbrB	XAC1883	XFF4834R_RS09810
A	cd09875 - PIN_VapC-FitB-like	XAC1884	XFF4834R_RS09815
T	COG3654 - doc	XAC1194	XFF4834R_RS16290
A	TIGR02609 – hypothetical_protein	XAC1195	XFF4834R_RS16285
T	TIGR03071 - Couple_hipA / pfam07805 - HipA_N	XAC0187	*******
A	COG1396 - HipB / cd00093 - HTH_XRE	XAC0188	*******
T	pfam02661 - Fic/DOC	XAC1501	*******
A	cd00093 - HTH_XRE	XAC1499	*******
	(no known domain)	XAC4312	XFF4834R_RS20450
	pfam05016 - plasm. stab. prot. RelE/ParE	XAC4313	XFF4834R_RS20455
T	cl00995 - superf. PemK / PRK09812 - ChpB	XACa0027	*******
A	cl00877 - superf. MazE / PRK11347 - ChpS	XACa0028	*******
T	RHH (recently described by Gallo, 2010)	XACa0037	*******
A	COG1569 - with PIN domain	XACa0036	*******
T	pfam11455 - DUF3018	XACb0033	*******
A	pfam02452 - PemK -like proteins	XACb0032	*******
A	cd00093 - HTH_XRE	XAC3288	*******
	plasmid. stabil - TIGR02683	XCAW_01038	*******
T	pfam05016 - plasm. stab. prot. RelE/ParE	XCV1189	XFF4834R_RS16435
A	(no known domain)	XAC1168/XCV1188	XFF4834R_RS16440
T	cl03380/pfam01845 - CcdB	XCV0819	*******
A	cl02188/pfam07362 - CcdA	XCV0820	*******
T	VapC/PIN/COG1848	*******	XFF4834R_RS21055
A	StbC	*******	XFF4834R_RS21050
T	VapC/PIN/COG1848	Rorf_13921_cds(CFBP 6992)	*******
A	StbD	Rorf_13913_(CFBP6992)	*******
T	HipA	*******	XFF4834R_RS06680
A	HTH_XRE/HTH_XRE	*******	XFF4834R_RS06685
A		*******	XFF4834R_RS20730
T	RelB=antitoxine	*******	XFF4834R_RS20725
T	cd09861 PIN-VapC-like	Rorf_85 (4226..4633)	XFF4834R_RS00015
A	spoVT_AbrB/MazE	Rorf_81 (3993..4241)	XFF4834R_RS00015
		XACb0059	*******
		XACb0060	*******
T	cd09881 - PIN_VapC-FitB	XCAW_a00016 pXcaw19	*******
A	pfam02604 - PhdYeFM_antitox	XCAW_a00017 pXcaw19	*******
T	COG3668 - ParE	XCAW_b00052 pXcaw58	*******
A	(no known domain)	XCAW_b00051 pXcaw58	*******
	cl21503 - Plasmid_stabil super family / COG3549 - HigB	XCV4440	*******
	cd00093 - HTH_XRE	XCV4439	*******
T	pfam05016 - Plasmid_stabil RelE/ParE	XCV3346	*******
A	cl22461 - RHH_1 COG3905	XCV3347	*******
T	cl03380/pfam01845 - CcdB	XCV0819	*******
A	cl02188/pfam07362 - CcdA	XCV0820	*******
T	pfam05016 - Plasmid_stabil RelE/ParE	XCVc0017	XFF4834R_RS20760
A	(no known domain)	XCVc0016	XFF4834R_RS20755
T	HTH_XRE/HTH_3/HipB/HTH_XRE	XAC0524	XFF4834R_RS02560
A		Rorf_8936	Rorf_8706
T		*******	Rorf_23154
A	HTH_XRE/HipB/HTH_XRE/HTH_3	*******	XFF4834R_RS06740
T		*******	Rorf23272
A	HTH_XRE/HTH_XRE/HTH_3/HipB	*******	XFF4834R_RS06775
T	hypothetical protein	XAC1790	XFF4834R_RS09335
A	StbD	XAC1789	XFF4834R_RS09330
T	HTH_XRE/HipB/HTH_XRE/HTH_3/VapI	*******	XFF4834R_RS10880
A		*******	Rorf38202
T	HTH_XRE	XAC2515	XFF4834R_RS12215
A		Rorf_43362	Rorf43148
T	HTH_XRE/HTH_XRE/HTH_3/HipB/VapI	XAC4056	XFF4834R_RS19160
A		*******	*******
T	ParE/Plasmid_stabil	*******	XFF4834R_RS20795
A		*******	XFF4834R_RS20800
T	VapC/PIN/COG1848	*******	XFF4834R_RS20825
A	StbC	*******	XFF4834R_RS20820
T	HTH_4	*******	XFF4834R_RS20860
A		*******	Rorf73691

Références bibliographiques

[Achaz, 2009] Achaz, G. (2009). Frequency Spectrum Neutrality Tests : One for All and All for One. *Genetics*, 183(1) :249–258.

[Achtman and Wagner, 2008] Achtman, M. and Wagner, M. (2008). Microbial diversity and the genetic nature of microbial species. *Nature Reviews Microbiology*, 6(6) :431–440.

[Akey, 2002] Akey, J. M. (2002). Interrogating a High-Density SNP Map for Signatures of Natural Selection. *Genome Research*, 12(12) :1805–1814.

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3) :403–410.

[Anderson et al., 2004] Anderson, P. K., Cunningham, A. A., Patel, N. G., Morales, F. J., Epstein, P. R., and Daszak, P. (2004). Emerging infectious diseases of plants : pathogen pollution, climate change and agrotechnology drivers. *Trends in Ecology & Evolution*, 19(10) :535–544.

[Aritua et al., 2015] Aritua, V., Harrison, J., Sapp, M., Buruchara, R., Smith, J., and Studholme, D. J. (2015). Genome sequencing reveals a new lineage associated with lablab bean and genetic exchange between Xanthomonas axonopodis pv. phaseoli and Xanthomonas fuscans subsp. fuscans. *Frontiers in Microbiology*, 6.

[Bartoli et al., 2016] Bartoli, C., Roux, F., and Lamichhane, J. R. (2016). Molecular mechanisms underlying the emergence of bacterial pathogens : an ecological perspective. *Molecular plant pathology*, 17(2) :303–310.

[Beaumont, 2005] Beaumont, M. A. (2005). Adaptation and speciation : what can Fst tell us ? *Trends in Ecology & Evolution*, 20(8) :435–440.

[Benoit et al., 2015] Benoit, G., Peterlongo, P., Lavenier, D., and Lemaitre, C. (2015). Simka : fast kmer-based method for estimating the similarity between numerous metagenomic datasets.

[Bergot et al., 2004] Bergot, M., Cloppet, E., Perarnaud, V., Deque, M., Marcais, B., and Desprez-Loustau, M.-L. (2004). Simulation of potential range expansion of oak disease caused by Phytophthora cinnamomi under climate change. *Global Change Biology*, 10(9) :1539–1552.

[Bruen and Bruen, 2005] Bruen, T. and Bruen, T. (2005). PhiPack PHI test and other tests of recombination. *McGill University, Montreal, Quebec*.

[Brussow et al., 2004] Brussow, H., Canchaya, C., and Hardt, W.-D. (2004). Phages and the Evolution of Bacterial Pathogens : from Genomic Rearrangements to Lysogenic Conversion. *Microbiology and Molecular Biology Reviews*, 68(3) :560–602.

[Cadillo-Quiroz et al., 2012] Cadillo-Quiroz, H., Didelot, X., Held, N. L., Herrera, A., Darling, A., Reno, M. L., Krause, D. J., and Whitaker, R. J. (2012). Patterns of Gene Flow Define Species of Thermophilic Archaea. *PLoS Biology*, 10(2) :e1001265.

[Charlesworth, 2007] Charlesworth, B. (2007). A hitch-hiking guide to the genome : a commentary on 'The hitch-hiking effect of a favourable gene' by John Maynard Smith and John Haigh. *Genetical Research*, 89(5-6) :389.

[Chen et al., 2015] Chen, J., Yang, X., Chen, J., Cen, Z., Guo, C., Jin, T., and Cui, Y. (2015). SISP : a Fast Species Identification System for Prokaryotes Based on Total Nucleotide Identity of Whole Genome Sequences. *Infectious Diseases and Translational Medicine*, 1(1) :30–55.

[Chimenti Michael S., 2016] Chimenti Michael S. (2016). Sequencing depth for accurate SNP calling : bcbio case study. `http://www.michaelchimenti.com/2016/07/sequencing-depth-genome-variant-calling/`.

[Cohan, 2006] Cohan, F. M. (2006). Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. *Philosophical Transactions of the Royal Society of London B : Biological Sciences*, 361(1475) :1985–1996.

[Conesa and Götz, 2008] Conesa, A. and Götz, S. (2008). Blast2go : A Comprehensive Suite for Functional Analysis in Plant Genomics. *International Journal of Plant Genomics*, 2008.

[Constantin et al., 2016] Constantin, E. C., Cleenwerck, I., Maes, M., Baeyen, S., Van Malderghem, C., De Vos, P., and Cottyn, B. (2016). Genetic characterization of strains named as *Xanthomonas axonopodis* pv. *dieffenbachiae* leads to a taxonomic revision of the *X. axonopodis* species complex. *Plant Pathology*, 65(5) :792–806.

[Costerton et al., 1995] Costerton, J. W., Lewandowski, Z., Caldwell, D. E., Korber, D. R., and Lappin-Scott, H. M. (1995). Microbial Biofilms. *Annual Review of Microbiology*, 49(1) :711–745.

[Danecek et al., 2011] Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., and 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15) :2156–2158.

[De Queiroz, 2007] De Queiroz, K. (2007). Species Concepts and Species Delimitation. *Systematic Biology*, 56(6) :879–886.

[Decker et al., 2014] Decker, J. E., McKay, S. D., Rolf, M. M., Kim, J., Molina Alcalá, A., Sonstegard, T. S., Hanotte, O., Götherström, A., Seabury, C. M., Praharani, L., Babar, M. E., Correia de Almeida Regitano, L., Yildiz, M. A., Heaton, M. P., Liu, W.-S., Lei, C.-Z., Reecy, J. M., Saif-Ur-Rehman, M., Schnabel, R. D., and Taylor, J. F. (2014). Worldwide Patterns of Ancestry, Divergence, and Admixture in Domesticated Cattle. *PLoS Genetics*, 10(3) :e1004254.

[Denancé et al., 2016] Denancé, N., Lahaye, T., and Noël, L. D. (2016). Editorial Genomics and Effectomics of the Crop Killer Xanthomonas. *Frontiers in Plant Science*, 7.

[Didelot et al., 2008] Didelot, X., Darling, A., and Falush, D. (2008). Inferring genomic flux in bacteria. *Genome Research*, page gr.082263.108.

[Didelot and Maiden, 2010] Didelot, X. and Maiden, M. C. J. (2010). Impact of recombination on bacterial evolution. *Trends in microbiology*, 18(7) :315.

[Didelot et al., 2012] Didelot, X., Méric, G., Falush, D., and Darling, A. E. (2012). Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli. *BMC Genomics*, 13(1) :256.

[Didelot et al., 2016] Didelot, X., Walker, A. S., Peto, T. E., Crook, D. W., and Wilson, D. J. (2016). Within-host evolution of bacterial pathogens. *Nature Reviews Microbiology*, 14(3) :150–162.

[Didelot and Wilson, 2015] Didelot, X. and Wilson, D. J. (2015). ClonalFrameML : Efficient Inference of Recombination in Whole Bacterial Genomes. *PLoS Comput Biol*, 11(2) :e1004041.

[Doolittle and Zhaxybayeva, 2009] Doolittle, W. F. and Zhaxybayeva, O. (2009). On the origin of prokaryotic species. *Genome Research*, 19(5) :744–756.

[Dye et al., 1980a] Dye, D. W., Bradbury, J. F., Goto, M., Hayward, A. C., Lelliott, R. A., and Schroth, M. N. (1980a). International standards for naming pathovars of phytopathogenic bacteria and a list of pathovar names and pathotype strains. *Review of Plant Pathology*, 59(4) :153–168.

[Dye et al., 1980b] Dye, D. W., Bradbury, J. F., Goto, M., Hayward, A. C., Lelliott, R. A., and Schroth, M. N. (1980b). International standards for naming pathovars of phytopathogenic bacteria and a list of pathovar names and pathotype strains. *Review of Plant Pathology*, 59(4) :153–168.

[El-Sharkawy et al., 2012] El-Sharkawy, M. A., de Tafur, S. M., and López, Y. (2012). *A Multidisciplinary Approach to Crop Improvement and Sustainable Production*. CIAT, Calyuca, Cali-Palmira, Colombia, eds. b. ospina and h. ceballos edition.

[Engering et al., 2013] Engering, A., Hogerwerf, L., and Slingenbergh, J. (2013). Pathogen-host-environment interplay and disease emergence. *Emerging Microbes & Infections*, 2(2) :e5.

[Errington et al., 2001] Errington, J., Bath, J., and Wu, L. J. (2001). DNA transport in bacteria. *Nature Reviews. Molecular Cell Biology*, 2(7) :538–545.

[Evans et al., 2008] Evans, N., Baierl, A., Semenov, M. A., Gladders, P., and Fitt, B. D. (2008). Range and severity of a plant disease increased by global warming. *Journal of The Royal Society Interface*, 5(22) :525–531.

[Fay and Wu, 2000] Fay, J. C. and Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, 155(3) :1405–1413.

[Fenchel, 2003] Fenchel, T. (2003). MICROBIOLOGY : Biogeography for Bacteria. *Science*, 301(5635) :925–926.

[Flor, 1971] Flor, H. H. (1971). Current Status of the Gene-For-Gene Concept. *Annual Review of Phytopathology*, 9(1) :275–296.

[García-Solache et al., 2016] García-Solache, M., Lebreton, F., McLaughlin, R. E., Whiteaker, J. D., Gilmore, M. S., and Rice, L. B. (2016). Homologous Recombination within Large Chromosomal Regions Facilitates Acquisition of -Lactam and Vancomycin Resistance in Enterococcus faecium. *Antimicrobial Agents and Chemotherapy*, 60(10) :5777–5786.

[Garrett et al., 2006] Garrett, K. A., Dendy, S. P., Frank, E. E., Rouse, M. N., and Travers, S. E. (2006). Climate change effects on plant disease : genomes to ecosystems. *Annu. Rev. Phytopathol.*, 44 :489–509.

[Gevers et al., 2006] Gevers, D., Dawyndt, P., Vandamme, P., Willems, A., Vancanneyt, M., Swings, J., and Vos, P. D. (2006). Stepping stones towards a new prokaryotic taxonomy. *Philosophical Transactions of the Royal Society of London B : Biological Sciences*, 361(1475) :1911–1916.

[Giraud et al., 2010] Giraud, T., Gladieux, P., and Gavrilets, S. (2010). Linking emergence of fungal plant diseases and ecological speciation. *Trends in ecology & evolution*, 25(7) :387–395.

[Gupta, 2004] Gupta, A. (2004). Origin of agriculture and domestication of plants and animals linked to early Holocene climate amelioration. *Current Science*, 87(1) :54–59.

[Gutenkunst et al., 2009] Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genetics*, 5(10) :e1000695.

[Gutenkunst et al., 2010] Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2010). Diffusion Approximations for Demographic Inference : DaDi. *Nature Precedings*, (713).

[Guttman, 1997] Guttman, D. S. (1997). Recombination and clonality in natural populations of Escherichia coli. *Trends in Ecology & Evolution*, 12(1) :16–22.

[Hajri et al., 2009] Hajri, A., Brin, C., Hunault, G., Lardeux, F., Lemaire, C., Manceau, C., Boureau, T., and Poussier, S. (2009). A Repertoire for Repertoire ? Hypothesis : Repertoires of Type Three Effectors are Candidate Determinants of Host Specificity in Xanthomonas. *PLoS ONE*, 4(8) :e6632.

[Han et al., 2013] Han, M. V., Thomas, G. W. C., Lugo-Martinez, J., and Hahn, M. W. (2013). Estimating Gene Gain and Loss Rates in the Presence of Error in Genome Assembly and Annotation Using CAFE 3. *Molecular Biology and Evolution*, 30(8) :1987–1997.

[Hao and Golding, 2008] Hao, W. and Golding, G. B. (2008). High rates of lateral gene transfer are not due to false diagnosis of gene absence. *Gene*, 421(1-2) :27–31.

[Hayward, 1993] Hayward, A. C. (1993). The hosts of Xanthomonas. In Swings, J. G. and Civerolo, E. L., editors, *Xanthomonas*, pages 1–119. Springer Netherlands. DOI : 10.1007/978-94-011-1526-1_1.

[Hershberg and Petrov, 2010] Hershberg, R. and Petrov, D. A. (2010). Evidence That Mutation Is Universally Biased towards AT in Bacteria. *PLOS Genet*, 6(9) :e1001115.

[Hilditch and Valtion teknillinen tutkimuskeskus, 2010] Hilditch, S. and Valtion teknillinen tutkimuskeskus (2010). *Identification of the fungal catabolic D-galacturonate pathway*. PhD thesis, VTT, Espoo, Finland. OCLC : 697545881.

[Jombart and Ahmed, 2011] Jombart, T. and Ahmed, I. (2011). adegenet 1.3-1 : new tools for the analysis of genome-wide SNP data. *Bioinformatics*, page btr521.

[Jones et al., 2004] Jones, J. B., Lacy, G. H., Bouzar, H., Stall, R. E., and Schaad, N. W. (2004). Reclassification of the Xanthomonads Associated with Bacterial Spot Disease of Tomato and Pepper. *Systematic and Applied Microbiology*, 27(6) :755–762.

[Kim et al., 2016] Kim, G. H., Kim, K.-H., Son, K. I., Choi, E. D., Lee, Y. S., Jung, J. S., and Koh, Y. J. (2016). Outbreak and Spread of Bacterial Canker

of Kiwifruit Caused by Pseudomonas syringae pv. actinidiae Biovar 3 in Korea. *The Plant Pathology Journal*, 32(6) :545–551.

[Kim et al., 2009] Kim, J.-G., Li, X., Roden, J. A., Taylor, K. W., Aakre, C. D., Su, B., Lalonde, S., Kirik, A., Chen, Y., Baranage, G., McLane, H., Martin, G. B., and Mudgett, M. B. (2009). Xanthomonas T3s Effector XopN Suppresses PAMP-Triggered Immunity and Interacts with a Tomato Atypical Receptor-Like Kinase and TFT1. *THE PLANT CELL ONLINE*, 21(4) :1305–1323.

[Konstantinidis and Tiedje, 2005] Konstantinidis, K. T. and Tiedje, J. M. (2005). Towards a Genome-Based Taxonomy for Prokaryotes. *Journal of Bacteriology*, 187(18) :6258–6264.

[Krause and Whitaker, 2015] Krause, D. J. and Whitaker, R. J. (2015). Inferring speciation processes from patterns of natural variation in microbial genomes. *Systematic Biology*, page syv050.

[Kuo et al., 2009] Kuo, C.-H., Moran, N. A., and Ochman, H. (2009). The consequences of genetic drift for bacterial genome complexity. *Genome Research*, 19(8) :1450–1454.

[Kuo and Ochman, 2009] Kuo, C.-H. and Ochman, H. (2009). The fate of new bacterial genes. *FEMS Microbiology Reviews*, 33(1) :38–43.

[Lapierre et al., 2016] Lapierre, M., Blin, C., Lambert, A., Achaz, G., and Rocha, E. P. C. (2016). The Impact of Selection, Gene Conversion, and Biased Sampling on the Assessment of Microbial Demography. *Molecular Biology and Evolution*, 33(7) :1711–1725.

[Lawson et al., 2012] Lawson, D. J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of Population Structure using Dense Haplotype Data. *PLoS Genetics*, 8(1) :e1002453.

[Lerouge and Vanderleyden, 2002] Lerouge, I. and Vanderleyden, J. (2002). O-antigen structural variation : mechanisms and possible roles in animal/plant–microbe interactions. *FEMS microbiology reviews*, 26(1) :17–47.

[Leroy et al., 2016] Leroy, T., Caffier, V., Celton, J.-M., Anger, N., Durel, C.-E., Lemaire, C., and Le Cam, B. (2016). When virulence originates from nonagricultural hosts : evolutionary and epidemiological consequences of introgressions fol-

lowing secondary contacts in *Venturia inaequalis*. *New Phytologist*, 210(4) :1443–1452.

[Li et al., 2015] Li, S., Wang, Y., Wang, S., Fang, A., Wang, J., Liu, L., Zhang, K., Mao, Y., and Sun, W. (2015). The Type III Effector AvrBs2 in Xanthomonas oryzae pv. oryzicola Suppresses Rice Immunity and Promotes Disease Development. *Molecular Plant-Microbe Interactions*, 28(8) :869–880.

[Lima et al., 2008] Lima, W. C., Paquola, A. C. M., Varani, A. M., Sluys, M.-A. V., and Menck, C. F. M. (2008). Laterally transferred genomic islands in Xanthomonadales related to pathogenicity and primary metabolism. *FEMS Microbiology Letters*, 281(1) :87–97.

[Lin et al., 1979] Lin, P. C., Tai, H. C., Chen, S. C., and Chien, M. C. (1979). Isolation and characterization of plasmids in Xanthomonas manihotis. *Botanical bulletin of Academia Sinica. New series.*

[Liu et al., 2006] Liu, X., Gutacker, M. M., Musser, J. M., and Fu, Y.-X. (2006). Evidence for Recombination in Mycobacterium tuberculosis. *Journal of Bacteriology*, 188(23) :8169–8177.

[Luu et al., 2016] Luu, K., Bazin, E., and Blum, M. G. (2016). pcadapt : an R package to perform genome scans for selection based on principal component analysis.

[Mansfield et al., 2012] Mansfield, J., Genin, S., Magori, S., Citovsky, V., Sriariyanum, M., Ronald, P., Dow, M., Verdier, V., Beer, S. V., Machado, M. A., Toth, I., Salmond, G., and Foster, G. D. (2012). Top 10 plant pathogenic bacteria in molecular plant pathology : Top 10 plant pathogenic bacteria. *Molecular Plant Pathology*, 13(6) :614–629.

[MARAIS, 2002] MARAIS, G. (2002). L'estimation des variations du taux de recombinaison dans un génome eucaryote : limites méthodologiques. `http://www.sfbi.fr/sites/default/files/jobim/jobim2002/papiers/P-p053_144.pdf`.

[Martial Briand et al., 2016] Martial Briand, Romain Gaborieau, Marie-Agnes Jacques, Matthieu Barret, Tristan Boureau, Sylvain Gaillard, and Nicolas Chen WG (2016). SkIf : a tool for rapid identification of genes or regulators of interest.

[Martins et al., 2016] Martins, P. M. M., Machado, M. A., Silva, N. V., Takita, M. A., and de Souza, A. A. (2016). Type II Toxin-Antitoxin Distribution and Adaptive Aspects on Xanthomonas Genomes : Focus on Xanthomonas citri. *Frontiers in Microbiology*, 7.

[Martiny et al., 2006] Martiny, J. B. H., Bohannan, B. J. M., Brown, J. H., Colwell, R. K., Fuhrman, J. A., Green, J. L., Horner-Devine, M. C., Kane, M., Krumins, J. A., Kuske, C. R., Morin, P. J., Naeem, S., Øvreås, L., Reysenbach, A.-L., Smith, V. H., and Staley, J. T. (2006). Microbial biogeography : putting microorganisms on the map. *Nature Reviews Microbiology*, 4(2) :102–112.

[Mayr, 1942] Mayr, E. (1942). Systematics and the Origin of Species(Comlumbia Univ. Press, New York).

[Mayr E, 1942] Mayr E (1942). *Systematics and the Origin of Species*. Columbia University Press.

[McDonald and Kreitman, 1991] McDonald, J. H. and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *, Published online : 20 June 1991 ; | doi :10.1038/351652a0*, 351(6328) :652–654.

[McElrone et al., 2003] McElrone, A. J., Sherald, J. L., and Forseth, I. N. (2003). Interactive effects of water stress and xylem-limited bacterial infection on the water relations of a host vine. *Journal of Experimental Botany*, 54(381) :419–430.

[McKenzie et al., 2001] McKenzie, G. J., Lee, P. L., Lombardo, M.-J., Hastings, P., and Rosenberg, S. M. (2001). SOS Mutator DNA Polymerase IV Functions in Adaptive Mutation and Not Adaptive Amplification. *Molecular Cell*, 7(3) :571–579.

[Merda et al., 2016a] Merda, D., Bonneau, S., Guimbaud, J.-F., Durand, K., Brin, C., Boureau, T., Lemaire, C., Jacques, M.-A., and Fischer-Le Saux, M. (2016a). Recombination-prone bacterial strains form a reservoir from which epidemic clones emerge in agroecosystems : Recombinant strains as a reservoir for epidemics. *Environmental Microbiology Reports*, 8(5) :572–581.

[Merda et al., 2016b] Merda, D., Bonneau, S., Guimbaud, J.-F., Durand, K., Brin, C., Boureau, T., Lemaire, C., Jacques, M.-A., and FischerLe Saux, M.

(2016b). Recombinationprone bacterial strains form a reservoir from which epidemic clones emerge in agroecosystems. *Environmental Microbiology Reports*, 8(5) :572–581.

[Mhedbi-Hajri, 2010] Mhedbi-Hajri, N. (2010). *theses.fr – Nadia Mhedbi-Hajri , Approches cumulées de phylogénie et d'écologie pour déterminer les bases génétiques de la spécificité d'hôte des bactéries phytopathogènes, cas des Xanthomonas spp.* PhD thesis, Angers.

[Mhedbi-Hajri et al., 2011] Mhedbi-Hajri, N., Darrasse, A., Pigné, S., Durand, K., Fouteau, S., Barbe, V., Manceau, C., Lemaire, C., and Jacques, M.-A. (2011). Sensing and adhesion are adaptive functions in the plant pathogenic xanthomonads. *BMC Evolutionary Biology*, 11(1) :67.

[Mhedbi-Hajri et al., 2013] Mhedbi-Hajri, N., Hajri, A., Boureau, T., Darrasse, A., Durand, K., Brin, C., Saux, M. F.-L., Manceau, C., Poussier, S., Pruvost, O., Lemaire, C., and Jacques, M.-A. (2013). Evolutionary History of the Plant Pathogenic Bacterium Xanthomonas axonopodis. *PLoS ONE*, 8(3) :e58474.

[Montaigne, 2011] Montaigne, W. (2011). *Diversité génétique et adaptation au milieu chez les arbres forestiers tropicaux : étude chez le genre Virola (Myristicaceae).* Antilles-Guyane.

[Moreira et al., 2010] Moreira, L. M., Almeida, N. F., Potnis, N., Digiampietri, L. A., Adi, S. S., Bortolossi, J. C., da Silva, A. C., da Silva, A. M., de Moraes, F. E., de Oliveira, J. C., de Souza, R. F., Fancincani, A. P., Ferraz, A. L., Ferro, M. I., Furlan, L. R., Gimenez, D. F., Jones, J. B., Kitajima, E. W., Laia, M. L., Leite, R. P., Nishiyama, M. Y., Rodrigues Neto, J., Nociti, L. A., Norman, D. J., Ostroski, E. H., Pereira, H. A., Staskawicz, B. J., Tezza, R. I., Ferro, J. A., Vinatzer, B. A., and Setubal, J. C. (2010). Novel insights into the genomic basis of citrus canker based on the genome sequences of two strains of Xanthomonas fuscans subsp. aurantifolii. *BMC Genomics*, 11(1) :238.

[Mruk and Kobayashi, 2014] Mruk, I. and Kobayashi, I. (2014). To be or not to be : regulation of restriction–modification systems and other toxin–antitoxin systems. *Nucleic Acids Research*, 42(1) :70.

[Nielsen, 2005] Nielsen, R. (2005). Molecular Signatures of Natural Selection. *Annual Review of Genetics*, 39(1) :197–218.

[Nowell et al., 2014] Nowell, R. W., Green, S., Laue, B. E., and Sharp, P. M. (2014). The Extent of Genome Flux and Its Role in the Differentiation of Bacterial Lineages. *Genome Biology and Evolution*, 6(6) :1514–1529.

[Nunn and Qian, 2010] Nunn, N. and Qian, N. (2010). The Columbian Exchange : A History of Disease, Food, and Ideas. *Journal of Economic Perspectives*, 24(2) :163–188.

[Ochman et al., 2000] Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). [No Title]. *Nature*, 405(6784) :299–304.

[Overbeek et al., 2014] Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A. R., Xia, F., and Stevens, R. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research*, 42(D1) :D206–D214.

[Page and Peti, 2016] Page, R. and Peti, W. (2016). Toxin-antitoxin systems in bacterial growth arrest and persistence. *Nature Chemical Biology*, 12(4) :208–214.

[Pandey, 2005] Pandey, D. P. (2005). Toxin-antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes. *Nucleic Acids Research*, 33(3) :966–976.

[Paradis et al., 2004] Paradis, E., Claude, J., and Strimmer, K. (2004). APE : Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, 20(2) :289–290.

[Patil et al., 2007] Patil, P. B., Bogdanove, A. J., and Sonti, R. V. (2007). The role of horizontal transfer in the evolution of a highly variable lipopolysaccharide biosynthesis locus in xanthomonads that infect rice, citrus and crucifers. *BMC Evolutionary Biology*, 7(1) :243.

[Pingali, 2012] Pingali, P. L. (2012). Green Revolution : Impacts, limits, and the path ahead. *Proceedings of the National Academy of Sciences*, 109(31) :12302–12308.

[Polz et al., 2013] Polz, M. F., Alm, E. J., and Hanage, W. P. (2013). Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends in Genetics*, 29(3) :170–175.

[Popa and Dagan, 2011] Popa, O. and Dagan, T. (2011). Trends and barriers to lateral gene transfer in prokaryotes. *Current Opinion in Microbiology*, 14(5) :615–623.

[Price et al., 2010] Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, 5(3) :e9490.

[Pruvost, 2004] Pruvost, O. (2004). Analyse du Risque Phytosanitaire AGR- b1 : Xanthomonas axonopodis pv. citri.

[Rademaker et al., 2005] Rademaker, J. L. W., Louws, F. J., Schultz, M. H., Rossbach, U., Vauterin, L., Swings, J., and de Bruijn, F. J. (2005). A Comprehensive Species to Strain Taxonomic Framework for *Xanthomonas*. *Phytopathology*, 95(9) :1098–1111.

[Radman et al., 1993] Radman, M., Taddei, F., and Halliday, J. (1993). Correction des erreurs dans l'ADN : de la génétique bactérienne aux mécanismes de prédisposition héréditaire aux cancers chez l'homme. *Nature*, page 363 : 13.

[Raj et al., 2014] Raj, A., Stephens, M., and Pritchard, J. K. (2014). fastSTRUCTURE Variational Inference of Population Structure in Large SNP Datasets. *Genetics*, page genetics.114.164350.

[Ranwez et al., 2011] Ranwez, V., Harispe, S., Delsuc, F., and Douzery, E. J. P. (2011). MACSE Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. *PLOS ONE*, 6(9) :e22594.

[Rayssiguier et al., 1989] Rayssiguier, C., Thaler, D. S., and Radman, M. (1989). The barrier to recombination between Escherichia coli and Salmonella typhimurium is disrupted in mismatch-repair mutants. *, Published online : 23 November 1989 ; | doi :10.1038/342396a0*, 342(6248) :396–401.

[Richard et al., 2017] Richard, D., Ravigné, V., Rieux, A., Facon, B., Boyer, C., Boyer, K., Grygiel, P., Javegny, S., Terville, M., Canteros, B. I., Robène, I., Vernière, C., Chabirand, A., Pruvost, O., and Lefeuvre, P. (2017). Adaptation of genetically monomorphic bacteria : evolution of copper resistance through multiple horizontal gene transfers of complex and versatile mobile genetic elements. *Molecular Ecology*, 26(7) :2131–2149.

[Richter and Rosselló-Móra, 2009] Richter, M. and Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences of the United States of America*, 106(45) :19126–19131.

[Rimet et al., 2016] Rimet, F., Chaumeil, P., Keck, F., Kermarrec, L., Vasselon, V., Kahlert, M., Franc, A., and Bouchez, A. (2016). R-Syst : :diatom : an open-access and curated barcode database for diatoms and freshwater monitoring. *Database : The Journal of Biological Databases and Curation*, 2016.

[Rougeux et al., 2016] Rougeux, C., Bernatchez, L., and Gagnaire, P.-A. (2016). Modeling the multiple facets of speciation-with-gene-flow towards improving divergence history inference of a recent fish adaptive radiation. *bioRxiv*, page 068932.

[Roux et al., 2016] Roux, C., Fraïsse, C., Romiguier, J., Anciaux, Y., Galtier, N., and Bierne, N. (2016). Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic Divergence. *PLOS Biology*, 14(12) :e2000234.

[Ryan et al., 2011] Ryan, R. P., Vorhölter, F.-J., Potnis, N., Jones, J. B., Van Sluys, M.-A., Bogdanove, A. J., and Dow, J. M. (2011). Pathogenomics of Xanthomonas : understanding bacterium–plant interactions. *Nature Reviews Microbiology*, 9(5) :344–355.

[Sallet et al., 2014] Sallet, E., Gouzy, J., and Schiex, T. (2014). EuGene-PP A Next Generation Automated Annotation Pipeline for Prokaryotic Genomes. *Bioinformatics*, page btu366.

[Saponari et al., 2014] Saponari, M., Loconsole, G., Cornara, D., Yokomi, R. K., De Stradis, A., Boscia, D., Bosco, D., Martelli, G. P., Krugner, R., and Porcelli, F. (2014). Infectivity and transmission of Xylellua fastidiosa by Philaenus spumarius (Hemiptera : Aphrophoridae) in Apulia, Italy. *Journal of Economic Entomology*, 107(4) :1316–1319.

[Sarkar et al., 2006] Sarkar, S. F., Gordon, J. S., Martin, G. B., and Guttman, D. S. (2006). Comparative Genomics of Host-Specific Virulence in Pseudomonas syringae. *Genetics*, 174(2) :1041–1056.

[Schaad et al., 2006] Schaad, N. W., Postnikova, E., Lacy, G., Sechler, A., Agarkova, I., Stromberg, P. E., Stromberg, V. K., and Vidaver, A. K. (2006). Emen-

ded classification of xanthomonad pathogens on citrus. *Systematic and Applied Microbiology*, 29(8) :690–695.

[Schaad et al., 2005a] Schaad, N. W., Postnikova, E., Lacy, G. H., Sechler, A., Agarkova, I., Stromberg, P. E., Stromberg, V. K., and Vidaver, A. K. (2005a). Reclassification of Xanthomonas campestris pv. citri (ex Hasse 1915) Dye 1978 forms A, B/C/D, and E as X. smithii subsp. citri (ex Hasse) sp. nov. nom. rev. comb. nov., X. fuscans subsp. aurantifolii (ex Gabriel 1989) sp. nov. nom. rev. comb. nov., and X. alfalfae subsp. citrumelo (ex Riker and Jones) Gabriel et al., 1989 sp. nov. nom. rev. comb. nov. ; X. campestris pv malvacearum (ex Smith 1901) Dye 1978 as X. smithii subsp. smithii nov. comb. nov. nom. nov. ; X. campestris pv. alfalfae (ex Riker and Jones, 1935) Dye 1978 as X. alfalfae subsp. alfalfae (ex Riker et al., 1935) sp. nov. nom. rev. ; and “var. fuscans” of X. campestris pv. phaseoli (ex Smith, 1987) Dye 1978 as X. fuscans subsp. fuscans sp. nov. *Systematic and Applied Microbiology*, 28(6) :494–518.

[Schaad et al., 2005b] Schaad, N. W., Postnikova, E., Lacy, G. H., Sechler, A., Agarkova, I., Stromberg, P. E., Stromberg, V. K., and Vidaver, A. K. (2005b). Reclassification of Xanthomonas campestris pv. citri (ex Hasse 1915) Dye 1978 forms A, B/C/D, and E as X. smithii subsp. citri (ex Hasse) sp. nov. nom. rev. comb. nov., X. fuscans subsp. aurantifolii (ex Gabriel 1989) sp. nov. nom. rev. comb. nov., and X. alfalfae subsp. citrumelo (ex Riker and Jones) Gabriel et al., 1989 sp. nov. nom. rev. comb. nov. ; X. campestris pv malvacearum (ex smith 1901) Dye 1978 as X. smithii subsp. smithii nov. comb. nov. nom. nov. ; X. campestris pv. alfalfae (ex Riker and Jones, 1935) dye 1978 as X. alfalfae subsp. alfalfae (ex Riker et al., 1935) sp. nov. nom. rev. ; and "var. fuscans" of X. campestris pv. phaseoli (ex Smith, 1987) Dye 1978 as X. fuscans subsp. fuscans sp. nov. *Systematic and Applied Microbiology*, 28(6) :494–518.

[Schofield and Hsieh, 2003] Schofield, M. J. and Hsieh, P. (2003). DNA Mismatch Repair : Molecular Mechanisms and Biological Function. *Annual Review of Microbiology*, 57(1) :579–608.

[Seehausen et al., 2014] Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., Peichel, C. L., Saetre, G.-P., Bank, C., Brännström, , Brelsford, A., Clarkson, C. S., Eroukhmanoff, F., Feder, J. L., Fi-

scher, M. C., Foote, A. D., Franchini, P., Jiggins, C. D., Jones, F. C., Lindholm, A. K., Lucek, K., Maan, M. E., Marques, D. A., Martin, S. H., Matthews, B., Meier, J. I., Möst, M., Nachman, M. W., Nonaka, E., Rennison, D. J., Schwarzer, J., Watson, E. T., Westram, A. M., and Widmer, A. (2014). Genomics and the origin of species. *Nature Reviews Genetics*, 15(3) :176–192.

[Sengupta and Austin, 2011] Sengupta, M. and Austin, S. (2011). Prevalence and Significance of Plasmid Maintenance Functions in the Virulence Plasmids of Pathogenic Bacteria. *Infection and Immunity*, 79(7) :2502–2509.

[Sevin and Barloy-Hubler, 2007] Sevin, E. W. and Barloy-Hubler, F. (2007). RASTA-Bacteria : a web-based tool for identifying toxin-antitoxin loci in prokaryotes. *Genome biology*, 8(8) :R155.

[Shen and Huang, 1989] Shen, P. and Huang, H. V. (1989). Effect of base pair mismatches on recombination via the RecBCD pathway. *Molecular and General Genetics MGG*, 218(2) :358–360.

[Siguier, 2006] Siguier, P. (2006). ISfinder : the reference centre for bacterial insertion sequences. *Nucleic Acids Research*, 34(90001) :D32–D36.

[Slarkin, 1985] Slarkin, M. (1985). Gene Flow in Natural Populations. *Annual Review of Ecology and Systematics*, 16(1) :393–430.

[Smith et al., 1997] Smith, I., DG, M., PR, S., and CABI, H. M. (1997). *Quarantine Pests for Europe*. 2nd edition. EPPO/CABI, UK, wallingford edition.

[Smith and Haigh, 1974] Smith, J. M. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research*, 23(1) :23–35.

[Stackebrandt et al., 2002] Stackebrandt, E., Frederiksen, W., Garrity, G. M., Grimont, P. A. D., Kämpfer, P., Maiden, M. C. J., Nesme, X., Rosselló-Mora, R., Swings, J., Trüper, H. G., Vauterin, L., Ward, A. C., and Whitman, W. B. (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(3) :1043–1047.

[Starr, 1981] Starr, M. P. (1981). The Genus Xanthomonas. In Starr, M. P., Stolp, H., Trüper, H. G., Balows, A., and Schlegel, H. G., editors, *The Prokaryotes*, pages 742–763. Springer Berlin Heidelberg. DOI : 10.1007/978-3-662-13187-9_62.

[Stukenbrock and McDonald, 2008] Stukenbrock, E. H. and McDonald, B. A. (2008). The Origins of Plant Pathogens in Agro-Ecosystems. *Annual Review of Phytopathology*, 46(1) :75–100.

[Taddei F. et al., 1996] Taddei F., Matic I., and Radman M. (1996). Du nouveau sur l'origine des espèces. *La Recherche*, (291) :52 – 72.

[Tajima, 1983] Tajima, F. (1983). Evolutionary Relationship of Dna Sequences in Finite Populations. *Genetics*, 105(2) :437–460.

[Team, 2013] Team, R. C. (2013). R : A Language and Environment for Statistical Computing.

[Tellier and Lemaire, 2014] Tellier, A. and Lemaire, C. (2014). Coalescence 2.0 : a multiple branching of recent theoretical developments and their applications. *Molecular Ecology*, 23(11) :2637–2652.

[Touchon et al., 2009] 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, M. E., Frapy, E., Garry, L., Ghigo, J. M., Gilles, A. M., Johnson, J., Le Bougu ?nec, C., Lescat, M., Mangenot, S., Martinez-J ?hanne, V., Matic, I., Nassif, X., Oztas, S., Petit, M. A., Pichon, C., Rouy, Z., Ruf, C. S., Schneider, D., Tourret, J., Vacherie, B., Vallenet, D., M ?digue, C., Rocha, E. P. C., and Denamur, E. (2009). Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths. *PLoS Genetics*, 5(1) :e1000344.

[Touchon and Rocha, 2007] Touchon, M. and Rocha, E. P. C. (2007). Causes of Insertion Sequences Abundance in Prokaryotic Genomes. *Molecular Biology and Evolution*, 24(4) :969–981.

[Treangen et al., 2014] Treangen, T. J., Ondov, B. D., Koren, S., and Phillippy, A. M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol*, 15(524) :2.

[Unterholzner et al., 2013] Unterholzner, S. J., Poppenberger, B., and Rozhon, W. (2013). Toxin–antitoxin systems : Biology, identification, and application. *Mobile Genetic Elements*, 3(5) :e26219.

[Van Melderen and Saavedra De Bast, 2009] Van Melderen, L. and Saavedra De Bast, M. (2009). Bacterial Toxin ?Antitoxin Systems : More Than Selfish Entities ? *PLoS Genetics*, 5(3) :e1000437.

[Vandecraen et al., 2017] Vandecraen, J., Chandler, M., Aertsen, A., and Van Houdt, R. (2017). The impact of insertion sequences on bacterial genome plasticity and adaptability. *Critical Reviews in Microbiology*, pages 1–22.

[Varani et al., 2011] Varani, A. M., Siguier, P., Gourbeyre, E., Charneau, V., and Chandler, M. (2011). ISsaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. *Genome Biology*, 12(3) :R30.

[VAUTERIN et al., 1995] VAUTERIN, L., HOSTE, B., KERSTERS, K., and SWINGS, J. (1995). Reclassification of Xanthomonas. *International Journal of Systematic and Evolutionary Microbiology*, 45(3) :472–489.

[Verdier, 1988] Verdier, V. (1988). *Contribution à l'étude de la variabilité de Xanthomonas campestris pv. manihotis (Arthaud Berthet et Bondar) Starr. agent causal de la bactériose vasculaire du manioc (Manihot esculenta Crantz)*. THE : Thèses, ORSTOM, Paris.

[Verdier et al., 2004] Verdier, V., Restrepo, S., Mosquera, G., Jorge, V., and Lopez, C. (2004). Recent progress in the characterization of molecular determinants in the Xanthomonas axonopodis pv. manihotis ?cassava interaction. *Plant Molecular Biology*, 56(4) :573–584.

[Vos and Didelot, 2009] Vos, M. and Didelot, X. (2009). A comparison of homologous recombination rates in bacteria and archaea. *The ISME Journal*, 3(2) :199–208.

[Wapinski et al., 2007] Wapinski, I., Pfeffer, A., Friedman, N., and Regev, A. (2007). Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics*, 23(13) :i549–i558.

[Waples and Gaggiotti, 2006] Waples, R. S. and Gaggiotti, O. (2006). What is a population ? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity : WHAT IS A POPULATION ? *Molecular Ecology*, 15(6) :1419–1439.

[Watterson, 1978] Watterson, G. A. (1978). The homozygosity test of neutrality. *Genetics*, 88(2) :405–417.

[Wayne et al., 1987] Wayne, L. G., Brenner, D. J., Colwell, R. R., Grimont, P. A. D., Kandler, O., Krichevsky, M. I., Moore, L. H., Moore, W. E. C., Murray, R. G. E., Stackebrandt, E., Starr, M. P., and Truper, H. G. (1987). Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *International Journal of Systematic and Evolutionary Microbiology*, 37(4) :463–464.

[Whitman, 2015] Whitman, W. B. (2015). Genome sequences as the type material for taxonomic descriptions of prokaryotes. *Systematic and Applied Microbiology*, 38(4) :217–222.

[Whittam, TS. and Ake, SE., 1993] Whittam, TS. and Ake, SE. (1993). *Genetic polymorphisms and recombination in natural populations of Escherichia coli*. In : Takahata, N. ; Clark, AG., editors. Mechanisms of Molecular Evolution. Sinauer Associates, Sunderland.

[Wiedenbeck and Cohan, 2011] Wiedenbeck, J. and Cohan, F. M. (2011). Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiology Reviews*, 35(5) :957–976.

[Wielgoss et al., 2013] Wielgoss, S., Barrick, J. E., Tenaillon, O., Wiser, M. J., Dittmar, W. J., Cruveiller, S., Chane-Woon-Ming, B., Médigue, C., Lenski, R. E., and Schneider, D. (2013). Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proceedings of the National Academy of Sciences*, 110(1) :222–227.

[Wielgoss et al., 2016] Wielgoss, S., Didelot, X., Chaudhuri, R. R., Liu, X., Weedall, G. D., Velicer, G. J., and Vos, M. (2016). A barrier to homologous recombination between sympatric strains of the cooperative soil bacterium Myxococcus xanthus. *The ISME Journal*.

[Wilson et al., 2002] Wilson, J., Schurr, M., LeBlanc, C., Ramamurthy, R., Buchanan, K., and Nickerson, C. (2002). Mechanisms of bacterial pathogenicity. *Postgraduate Medical Journal*, 78(918) :216–224.

[Wirth et al., 2007] Wirth, T., Morelli, G., Kusecek, B., van Belkum, A., van der Schee, C., Meyer, A., and Achtman, M. (2007). The rise and spread of a new

pathogen : Seroresistant Moraxella catarrhalis. *Genome Research*, 17(11) :1647–1656.

[Xu and Hao, 2009] Xu, Z. and Hao, B. (2009). CVTree update : a newly designed phylogenetic study platform using composition vectors and whole genomes. *Nucleic Acids Research*, 37(Web Server) :W174–W178.

[Yahara et al., 2013] Yahara, K., Furuta, Y., Oshima, K., Yoshida, M., Azuma, T., Hattori, M., Uchiyama, I., and Kobayashi, I. (2013). Chromosome Painting In Silico in a Bacterial Species Reveals Fine Population Structure. *Molecular Biology and Evolution*, 30(6) :1454–1464.

[Young et al., 2008] Young, J., Park, D.-C., Shearman, H., and Fargier, E. (2008). A multilocus sequence analysis of the genus Xanthomonas. *Systematic and Applied Microbiology*, 31(5) :366–377.

[Yu et al., 2010] Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). GOSemSim : an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26(7) :976–978.

[Zeng et al., 2006] Zeng, K., Fu, Y.-X., Shi, S., and Wu, C.-I. (2006). Statistical Tests for Detecting Positive Selection by Utilizing High-Frequency Variants. *Genetics*, 174(3) :1431–1439.

[Zeng et al., 2007] Zeng, K., Shi, S., and Wu, C.-I. (2007). Compound Tests for the Detection of Hitchhiking Under Positive Selection. *Molecular Biology and Evolution*, 24(8) :1898–1908.

[Zhang et al., 2011] Zhang, L., Thiewes, H., and van Kan, J. A. (2011). The d-galacturonic acid catabolic pathway in Botrytis cinerea. *Fungal Genetics and Biology*, 48(10) :990–997.

[Zhaxybayeva et al., 2007] Zhaxybayeva, O., Nesbø, C. L., and Doolittle, W. F. (2007). Systematic overestimation of gene gain through false diagnosis of gene absence. *Genome Biology*, 8(2) :402.

[Ziebuhr et al., 1999] Ziebuhr, W., Krimmer, V., Rachid, S., Lossner, I., Gotz, F., and Hacker, J. (1999). A novel mechanism of phase variation of virulence in Staphylococcus epidermidis : evidence for control of the polysaccharide intercellular adhesin synthesis by alternating insertion and excision of the insertion sequence element IS256. *Molecular Microbiology*, 32(2) :345–356.