



HAL
open science

Apport de l'ADN ancien dans l'étude de l'émergence et de l'évolution d'agents pathogènes de cultures

Paola E Campos

► **To cite this version:**

Paola E Campos. Apport de l'ADN ancien dans l'étude de l'émergence et de l'évolution d'agents pathogènes de cultures. Sciences du Vivant [q-bio]. Museum National d'Histoire Naturelle, 2021. Français. NNT: . tel-03936457

HAL Id: tel-03936457

<https://hal.science/tel-03936457v1>

Submitted on 12 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License



Muséum national d'Histoire naturelle

Sciences de la nature et de l'Homme : écologie et évolution – ED227

UMR Peuplements végétaux et bio-agresseurs en milieu tropical, CIRAD, Université de La Réunion

UMR Institut de systématique, évolution, biodiversité, Muséum national d'Histoire naturelle, CNRS,
SU, EPHE, UA

Apport de l'ADN ancien dans l'étude de l'émergence et de l'évolution d'agents pathogènes de cultures

Par Paola Campos

Thèse de doctorat de Biologie des organismes

Spécialité : Phylogénétique, épidémiologie moléculaire, phytopathologie

Dirigée par Nathalie Becker

Présentée et soutenue publiquement le 8 décembre 2021

Devant un jury composé de :

Dr Denis Fargette, Directeur de recherche, IRD PHIM, Rapporteur

Dr Marie-Agnès Jacques, Directrice de recherche, INRAE IRHS, Rapportrice

Pr Serge Muller, Professeur émérite, MNHN ISyEB, Examineur et Président du jury

Dr Clio Der Sarkissian, Chercheuse, CNRS CAGT, Examinatrice

Dr Nathalie Becker, Maître de conférences, MNHN ISyEB, Directrice de thèse

Dr Adrien Rieux, Chercheur, CIRAD PVBMT, Encadrant de thèse

Dr Lionel Gagnevin, Chercheur, CIRAD PHIM, Co-encadrant de thèse (invité)

*Pour peu qu'on ait l'esprit sensé,
Et que du Monde on sache le grimoire,
On voit bientôt que cette histoire
Est un conte du temps passé [.]*

Charles Perrault, La Barbe Bleue, 1697

A ma famille.
A ma mère et à mon frère.
A Tortue et Crocodile.

Remerciements

Je souhaiterais, pour commencer, remercier mes trois encadrants de thèse qui m'ont fait confiance durant ce projet et qui m'ont tant et patiemment enseignée, chacun dans votre domaine. Merci de m'avoir accordée autant de votre temps pour que j'intègre ne serait-ce qu'un peu vos connaissances, pour votre rigueur scientifique, merci pour tous vos conseils et pour votre soutien ces trois ans qui ont fait de cette thèse une expérience grandiose pour moi. Nathalie, merci pour ta direction et ta gentillesse, pour ces *heures* que tu as sans doute dues passées à m'aiguiller ou à me rassurer, merci pour ton enseignement au labo, ces rires durant les manip, les moments de pause aussi en dehors du boulot et au lagon. Adrien, merci de m'avoir autant expliquée et permis d'explorer en bioinformatique, merci de m'avoir recadrée quand j'en avais besoin, merci de ton aide si précieuse quand j'étais débordée, merci de ton calme olympien. Adrien, tu es le roc auquel j'aspire quand je serai grande :) Lionel, merci pour toutes les pépites de savoir que tu divulgues sans peine, véritable puits de connaissances sur Xantho et la science en général, merci de ton dynamisme et de ta réactivité dans les analyses et mes questionnements malgré les 10000 km qui nous séparaient.

Je remercie chaleureusement les membres de mon jury d'avoir accepté d'examiner et d'évaluer mon travail de thèse, Marie-Agnès Jacques (INRAE), Denis Fargette (IRD), Clio Der Sarkissian (CNRS) et Serge Muller (MNHN). Je remercie également les membres de mes comités de thèse pour les discussions expertes et leurs nombreux conseils qui ont servi à orienter cette thèse, Violaine Llaurens (MNHN), François Balloux (University College of London), Régis Debruyne (MNHN) et Álvaro Pérez-Quintero (Colorado State University). Merci aussi à Olivier Pruvost, Frédéric Chiroleu et Thuy Trang Cao pour leur relecture de ce manuscrit.

J'adresse mes remerciements à tous ceux qui m'ont accompagnée dans cette formidable aventure. Tout d'abord, merci à Karine Boyer avec qui le labo fut comme la découverte de la caverne d'Ali Baba. Merci à Claudine Boyer, Véronique Maillot-Lebon, Christophe Simiand (Lulu ! et tes goûts musicaux plus qu'éclectiques :D pour s'enjailler en Bio Mol), Murielle Hoareau, Stéphanie Javegny pour vos conseils et votre bonne humeur au laboratoire. Merci à Olivier Pruvost, Jean-Michel Lett, Pierre Leveuvre, Damien Richard, Frédéric Chiroleu, Anna Doizy et Isabelle Robène pour vos discussions enrichissantes et vos recommandations tout au long de ce travail.

Merci également à tous ces messieurs et ces dames qui ont fait du 3P un environnement convivial de travail et de rigolade, Cyril, Carine, Nico, Gilles, Virginie, Nathalie, Delphine, Hélène, Séverine, Willy, Océane, Micheline, Jean-Michel, Clara, Cédric, Jean-Philippe, Olivier, Moutou. Merci à Sohini et Sarah et Alizée pour votre accueil chaleureux à mon arrivée ici et vos conseils de survie. Merci aux potes ingé, VSC, doctorants (courage à vous !) et docteurs pour leur entrain et leur compagnie revigorante et sans

qui cette thèse aurait eu une autre saveur, Ange Ca' (cœur sur toi, gros, gros cœur sur toi), Anna (sauveuse), Maëva, Franky, Mathilde, Emma, Margot, Corentin, Corentin, Rachel, Justine, Marine, Pauline (dernière ligne droite, meuf ! après inondations et chutes de plafond), Sélim (merci de ton accueil au nouveau bureau :P merci pour les bonbecs et ta playlist 60s), Maëva, Félicien (Félicien !), Hasina, Benoît, Marine, Jocelin, Gaëlle, et Nirry, Karim, Damien, Nico, Cathucia, Anziz, Seb' qui sont déjà docteurs !

Spéciale dédicace à Ismaël, monsieur l'oiseau, binôme de bureau merveilleux et déjanté. Comme tu as pu être communicatif dans ta joie et ta bonne humeur, dans les hauts comme les bas de cette aventure, mon admiration pour toi est sans limite pour ta capacité à te dépasser encore et toujours ! On se retrouvera de l'autre côté, alors en attendant, un énorme merci à toi pour ton amitié, tes soupapes de décompression, tes jeux de mots, blagues et autres doigts dans le nez :D

Mes sincères remerciements aux amis que j'ai laissé à Paris, qui ont excusé mon absence et mon manque de réactivité et réponses à vos messages et qui me retrouvaient à bras ouverts lors de passages éclairs. Merci Diane (Djigr, you beautiful Dino Lady, I'd probably not have done the same without you :^) tout mon courage pour toi, ma belle), Julie & Julien (hé hé), Thomas, Mathieu, Gwen, Zach, Maria (besitos guapa !), Alfred & Tracy, et l'équipe de choc du 4^{ème} Mathilde, Lucille, Lucy, Felipe et Mathieu.

Merci aux princesses du ciel, Judith & Madeleine, cette triade géographique à quatre a pour sûr été une sacrée épreuve ! Alors merci pour tout et le reste, parce que, où qu'on ait été, on a tenu, n'est-ce pas ? Tenons la barre, Mesdames, et n'oubliez pas, mardi ou mercredi, dernier délai !

Pour finir, je remercie ma famille qui a toujours su avoir confiance en moi et me soutenir, pour votre chaleureuse présence malgré la distance, pour vos réserves de câlins, vos appels à pas d'heure et vos histoires de la vie. Merci d'avoir été là pour moi dans cette aventure, vous savez tout ce que vous avez fait pour moi.

Table des matières

Remerciements	3
Table des illustrations.....	3
Table des tableaux.....	3
Liste des publications	4
Chapitre 1 – Revue bibliographique	5
1.1. Les maladies de plantes cultivées	5
1.1.1. Pathogènes et maladies infectieuses	5
1.1.2. Forces évolutives menant à l'émergence, la diffusion et l'évolution de bactéries pathogènes	7
1.2. Epidémiologie	10
1.2.1. Épidémiologies classique et moléculaire.....	10
1.2.2. Etude de l'histoire évolutive des agents pathogènes	12
1.3. Ressources historiques	20
1.3.1. L'avènement de l'ADN ancien	20
1.3.2. Application aux cas des maladies infectieuses.....	21
1.3.3. Les échantillons d'herbier: une ressource pour l'étude des maladies des plantes	28
1.3.4. Caractéristiques et modalités d'acquisition de matériel génétique ancien depuis échantillons d'herbier	32
1.4. Le pathosystème: <i>Citrus / Xanthomonas citri pathovar citri</i>	35
1.4.1. Chancre asiatique des agrumes.....	35
1.4.2. <i>Xanthomonas citri</i> pv. <i>citri</i>	38
1.4.3. Cycle biologique de <i>Xanthomonas citri</i> pv. <i>citri</i> et interactions avec la plante	38
1.5. Ressources et état des connaissances avant le début de la thèse.....	40
1.6. Annexes	42
Objectifs de la thèse	67
Chapitre 2 – Optimisation et application de protocoles innovants au laboratoire et en bioinformatique	68
2.1. Optimisation des protocoles au laboratoire	69
2.2. Optimisation d'un pipeline bioinformatique.....	71
2.3. Caractérisation des échantillons et génération des génomes historiques	75
2.4. Analyses des patrons de dégradation de l'ADN	82
2.5. Annexes	88
Chapitre 3 – Reconstruction de l'histoire d'une émergence à échelle locale.....	111
Chapitre 4 – Reconstruction de l'histoire évolutive d'un pathogène à échelle mondiale	138
Chapitre 5 – Collaborations et travaux annexes	182
Contribution des ARNs interférents de petites tailles historiques issus d'herbier dans l'étude d'une maladie virale de cultures	182
Discussion générale.....	196
Apport de l'ADN ancien dans l'étude de l'émergence et de l'évolution de pathogènes de cultures	196

Caractérisation de la dégradation <i>post-mortem</i> de l'ADN d'une bactérie de plante	197
Limites de la reconstruction des génomes par alignement sur séquences de référence.....	198
Avantages et contraintes de l'approche moléculaire non ciblée.....	200
Quel(s) apport(s) des 250+ autres échantillons d'agrumes symptomatiques collectés mais non utilisés dans l'étude?	202
Conclusions et perspectives	203
Bibliographie.....	205

Table des illustrations

Figure 1.1.1.A. Comparaison d'un écosystème naturel et d'un agroécosystème (Stukenbrock and McDonald, 2008, Figure 1).	7
Figure 1.2.1.A. Contribution des analyses génomiques aux études épidémiologiques des maladies infectieuses émergentes (Li, Grassly and Fraser, 2014, Figure 1).	11
Figure 1.2.2.A. Calibration temporelle d'un arbre phylogénétique par « <i>tip-dating</i> » et estimation temporelle de l'émergence d'un pathogène.	15
Figure 1.2.2.B. Méthodes permettant de tester la présence d'un signal temporel au sein d'un jeu de données de séquences nucléotidiques hétérochrones (Rieux and Balloux, 2016, modification de la Figure 2).....	17
Figure 1.2.2.C. Phylogéographie discrète (Rasmussen and Grünwald, 2020, Figure 1A&B).	19
Figure 1.3.2.A. Méthodes d'obtention et d'analyse des données génomiques issues de pathogènes (Spyrou et al., 2019, Figures 2 modifiée et 3). Aperçu des analyses typiquement réalisables sur données génomiques de pathogènes comprenant de l'ADN ancien. Abréviations : NGS, <i>Next Generation Sequencing</i> ; SNP, <i>Single-Nucleotide Polymorphism</i> ; PCA, <i>Principal Component Analysis</i> ; PC, <i>Principal Component</i> ; MRCA, <i>Most Recent Common Ancestor</i> ; tMRCA, <i>time since Most Recent Common Ancestor</i>	24
Figure 1.3.3.A. Précision de la reconstruction des routes invasives de <i>Phytophthora infestans</i> du XIX ^{ème} siècle grâce à la paléogénomique (Yoshida et al., 2014, Figure 2).....	30
Figure 1.3.4.A. Salles de collection de l'Herbier national d'Histoire naturelle (à gauche) et du Naturalis Biodiversity Center (à droite).	33
Figure 1.3.4.B. Fragmentation et désamination de l'ADN (Dabney, Meyer and Pääbo, 2013, Figure 1).	34
Figure 1.4.1.A. Symptômes du chancre asiatique des agrumes sur fruit (gauche) et feuille (droite)...	36
Figure 1.4.1.B. Coupes histologiques d'une feuille au niveau d'une lésion de chancre asiatique des agrumes (gauche) et d'une feuille saine (droite) (Lin, Hsu and Tzeng, 2009, Figure 2).	36
Figure 1.4.1.C. Distribution géographique et émergences du chancre asiatique des agrumes.....	37
Figure 1.4.3.A. Cycle du chancre asiatique des agrumes (Schubert <i>et al.</i> , 2001, Figure 3).	40
Figure 1.5.A. Symptômes du chancre asiatique des agrumes sur spécimens d'herbiers.	41
Figure 2.2.A. Pipeline bioinformatique utilisé incluant les bibliothèques issues d'échantillons d'herbier...	72
Figure 2.2.B. Distribution géographique des génomes historiques (issus d'herbiers) et modernes (issus de cultures bactériennes) de <i>Xci</i> utilisés durant la thèse.....	75
Figure 2.3.A. Patrons de dégradation <i>post-mortem</i> des 13 génomes <i>Xci</i> issus d'échantillons historiques d'herbier.	82

Figure 2.4.A. Analyse factorielle des correspondances des catégories de tailles de fragments des génomes <i>Xci</i> issus de 13 échantillons historiques.....	85
Figure 2.4.B. Pourcentage de désamination à la dernière position de l'extrémité 3' des génomes <i>Xci</i> issus de 13 échantillons historiques.....	86
Figure 2.4.C. Pourcentage de désamination à la dernière position de l'extrémité 3' des ADN des trois séquences (chromosome, plasmides pXAC33 et pXAC64) des génomes <i>Xci</i> issus de 13 échantillons historiques.....	87
Figure 6.A. Origine généalogique des principaux groupes variétaux des <i>Citrus</i> (Curk <i>et al.</i> , 2016, Figure 8).....	201

Table des tableaux

Tableau 1.3.2.A. Données génomiques d'agents pathogènes anciens issus de spécimens historiques ou anciens (Spyrou <i>et al.</i> , 2019, Tableau 1).	25
Tableau 1.3.3.A. Données génomiques d'agents pathogènes anciens issus de spécimens d'herbiers.	31
Tableau 2.3.A. Caractéristiques générales de reconstruction et dégradation des génomes <i>Xci</i> issus de 13 échantillons historiques d'herbier.....	76

Liste des publications

Gagnevin, L., Rieux, A., Lett, J.-M., Roumagnac, P., Szurek, B., Campos, P. , Baider, C., Gaudeul, M. and Becker, N. (2021) 'Les herbiers, une fenêtre ouverte sur l'histoire évolutive des agents pathogènes des cultures', in <i>Les collections naturalistes dans la science du XXI^e siècle</i> . ISTE Editions, pp. 197–219.	43
Robène, I., Maillot-Lebon, V., Chabirand, A., Moreau, A., Becker, N., Moumène, A., Rieux, A., Campos, P. , Gagnevin, L., Gaudeul, M., Baider, C., Chiroleu, F. and Pruvost, O. (2020) 'Development and comparative validation of genomic-driven PCR-based assays to detect <i>Xanthomonas citri</i> pv. <i>citri</i> in citrus plants', <i>BMC Microbiology</i> , 20(296), pp. 1–13. doi:10.1186/s12866-020-01972-8.	98
Campos, P.E. , Groot Crego, C., Boyer, K., Gaudeul, M., Baider, C., Richard, D., Pruvost, O., Roumagnac, P., Szurek, B., Becker, N., Gagnevin, L. and Rieux, A. (2021) 'First historical genome of a crop bacterial pathogen from herbarium specimen: Insights into citrus canker emergence', <i>PLOS Pathogens</i> , 17(7), pp. 1–25. doi:10.1371/journal.ppat.1009714.	113
Campos, P.E. , Pruvost, O., Boyer, K., Gaudeul, M., Baider, C., Utteridge, T., Shannon, D., Toner, M., Becker, N., Rieux, A. and Gagnevin, L. (in prep) 'Improved reconstruction of the crop pathogenic bacterium <i>Xanthomonas citri</i> pv. <i>citri</i> diversification history using historical herbarium genomes.'	140
Rieux, A., Campos, P. , Duvermy, A., Scussel, S., Martin, D., Gaudeul, M., Lefeuvre, P., Becker, N. and Lett, J.-M. (2021) 'Contribution of historical herbarium small RNAs to the reconstruction of a cassava mosaic geminivirus evolutionary history', <i>Scientific Reports</i>	184

Chapitre 1 – Revue bibliographique

1.1. Les maladies de plantes cultivées

1.1.1. Pathogènes et maladies infectieuses

Les maladies infectieuses se définissent comme une altération de la santé et des fonctions d'un organisme causée par des agents pathogènes (prions, virus, bactéries, champignons ou parasites eucaryotes), capables d'être transmis, directement ou non, à un autre individu (définition de l'Organisation Mondiale de la Santé, 2021). Chez les plantes, les pathogènes de cultures menacent la production des ressources alimentaires avec des baisses de rendement et des pertes cumulées pouvant dépasser les 40% de récoltes, mettant ainsi en péril la sécurité alimentaire et engendrant de forts déficits économiques (Savary *et al.*, 2019). De plus, afin de limiter la dispersion de ces pathogènes, les plantes et leurs produits font l'objet de restrictions de mouvements et de commercialisation dans les zones de quarantaine. Dans les zones exemptes du pathogène, des systèmes de surveillance doivent être mis en place afin d'éviter l'arrivée du pathogène. Enfin, l'utilisation de pesticides souvent mis en place dans les stratégies de lutte contre les agents pathogènes est néfaste pour l'environnement, la santé publique et la biodiversité (Bernades *et al.*, 2015).

Les maladies infectieuses sont de différents types :

- endémique lorsque l'agent pathogène et la maladie qu'il cause sont trouvés de manière habituelle et récurrente dans une zone géographique, induisant une incidence (nombre de nouveaux cas par période donnée) stable sur la population ou région déterminée ;
- épidémique lorsque la maladie se développe ou se propage par une croissance rapide de son incidence dans une population à une région donnée (zone originaire de la maladie ou périphérique) ;
- pandémique lorsque l'épidémie atteint une très grande envergure et est présente sur de vastes territoires (au moins deux continents ou 100 pays) ;
- émergence lorsque la maladie est détectée dans une nouvelle population ou dans un nouvel espace-temps où sa prévalence (proportion de cas dans la population) augmente de manière significative ; la maladie peut également réémerger là où elle avait auparavant disparu.

L'émergence d'une maladie peut être conditionnée par différents facteurs, intrinsèques ou extrinsèques à l'agent pathogène. Les premiers correspondent aux traits de vie de l'agent pathogène et son potentiel évolutif ; ils se traduisent par la capacité de l'organisme à devenir pathogène et être plus ou moins virulent, à développer une tolérance, voire une résistance aux pesticides, à contourner la résistance de son hôte ou encore à s'adapter à un nouvel hôte ou à de nouvelles conditions environnementales (McDonald and Linde, 2002; Stukenbrock and McDonald, 2008). Les facteurs

ε

extrinsèques représentent les opportunités conférées par une modification de l'environnement où se trouve habituellement l'agent pathogène. A petite échelle, l'environnement correspond à l'hôte du pathogène avec lequel il interagit. Cette interaction hôte-pathogène conduit à une coévolution où l'hôte recherchera l'immunité face au pathogène *via* l'acquisition de mécanismes de défense (gènes de résistance, inactivation des gènes de sensibilité ciblés par le pathogène...) que le pathogène cherchera à contourner *via* son propre contenu en gènes. Les facteurs extrinsèques à large échelle sont de l'ordre de la dispersion de l'agent pathogène par événement météorologique ou liée à l'activité humaine par l'introduction de matériel infecté dans une nouvelle aire. Le dérèglement climatique, dont les projections prévoient des modifications de la fréquence des événements météorologiques inhabituels et des stress (hydriques ou de température) qui leur sont liés, peut également contribuer à l'expansion de l'aire de distribution de l'agent pathogène ou à son émergence dans une zone jusqu'alors peu propice à son développement (Anderson et al., 2004).

De nombreuses maladies bactériennes animales et humaines datent de la révolution néolithique (Mira, Pushker and Rodríguez-Valera, 2006), une période qui se caractérise par le développement de l'élevage et de l'agriculture il y a environ 13 000 à 10 000 ans (Balter, 2007). La transition du comportement chasseur-cueilleur des populations humaines vers des communautés agricoles a induit une modification du paysage par la création d'agroécosystèmes (Figure 1.1.1.A), des écosystèmes cultivés par l'homme où sont entretenus des plantes mais également leurs organismes associés (Larson *et al.*, 2014; Zeder, 2017), notamment leurs agents pathogènes (Stukenbrock and McDonald, 2008). Les conditions de culture des agroécosystèmes, comme l'homogénéisation du milieu, la densité importante et l'uniformité génétique des populations de plantes-hôtes ou encore la dispersion des propagules liée aux activités humaines, sont propices à l'émergence d'agents pathogènes spécialistes et à leur propagation, ainsi qu'au développement de lignées plus virulentes (Mira, Pushker and Rodríguez-Valera, 2006; Stukenbrock and McDonald, 2008).



Figure 1.1.1.A. Comparaison d'un écosystème naturel et d'un agroécosystème (Stukenbrock and McDonald, 2008, Figure 1). Les agroécosystèmes (droite) présentent une plus grande homogénéité environnementale, une plus faible diversité spécifique et une plus grande densité d'hôtes qui facilitent la transmission de pathogènes aux plants sains avoisinants. Ces propriétés permettent le développement d'épidémies caractérisées par plusieurs cycles de reproduction du pathogène chaque année et conduisent à l'émergence de pathogènes de plantes hautement virulents et spécialisés à l'hôte. Les écosystèmes naturels (gauche) ont une plus grande hétérogénéité environnementale, une plus grande diversité spécifique et une plus faible densité d'hôtes qui favorisent des pathogènes moins virulents et généralistes.

Ces conditions sont devenues d'autant plus favorables dans les périodes récentes marquées par le développement de la sélection variétale consciente, des techniques de greffage et de l'agriculture coloniale de rentes ainsi que l'intensification des échanges intercontinentaux de matériels aux XVIII et XIX^{ème} siècles. Au XX^{ème} siècle, la sélection des plantes-hôtes dans le but d'obtenir des variétés résistantes aux maladies devient un outil majeur en agronomie et s'accompagne d'une augmentation des aires monocoltivées et de l'utilisation de pesticides mais aussi du trafic de marchandises et de matériel sur de plus grandes distances par la mondialisation (Anderson *et al.*, 2004; Stukenbrock and McDonald, 2008). L'ensemble de ces conditions se traduit par l'expansion des zones où l'agent pathogène est présent et l'augmentation des opportunités de combinaisons entre pathogène et plante hôte (d'une nouvelle population naïve de la même espèce ou d'espèces proches de la plante hôte) ou de rencontres entre différents pathogènes; mais également par l'application de pressions de sélection directionnelle sur les populations de pathogènes qui vont impacter les risques d'émergence (Anderson *et al.*, 2004; Stukenbrock and McDonald, 2008).

1.1.2. Forces évolutives menant à l'émergence, la diffusion et l'évolution de bactéries pathogènes

L'émergence d'une maladie est déterminée par les interactions multiples, complexes et changeantes des facteurs liés à l'agent pathogène, à l'hôte et à l'environnement. Au niveau du pathogène, elle est le résultat de l'interaction entre plusieurs forces évolutives par lesquelles un trait dérivé génétiquement héritable lié à la pathogénie apparaît au sein d'une population de microorganismes et

confère un avantage sélectif à ses porteurs dans le milieu dans lequel ils évoluent par rapport au reste de la population. Ces forces affectent la structure génétique des populations, définie comme la quantité et la distribution de la variation génétique au sein et entre populations. La structure génétique est donc la conséquence de l'histoire évolutive de ces populations et son étude permet de reconstruire les processus qui l'ont façonnée mais également d'anticiper leur futur potentiel évolutif (McDonald and Linde, 2002). On distingue cinq forces évolutives, définies comme des mécanismes ayant la capacité de faire varier au sein d'une population, la fréquence des allèles d'une génération à la suivante et nous nous intéresserons ici à leurs effets chez les bactéries, modèle biologique d'étude principal de cette thèse :

- la mutation : elle peut être ponctuelle *via* la substitution d'un nucléotide par un autre, donnant lieu à un polymorphisme d'un seul nucléotide (SNP, *single nucleotide polymorphism*) ou par l'insertion/délétion d'un nucléotide (InDel). Elle peut aussi affecter un plus grand nombre de sites par des mécanismes d'élongation/raccourcissement de motifs répétés, des duplications/pertes de gènes ou de larges régions génomiques, des transpositions et des inversions et il a pu être estimé que leurs taux de substitutions ponctuelles étaient compris entre 10^{-8} et 10^{-5} substitution par site et par an (Duchêne *et al.*, 2016).
- la recombinaison : chez les organismes à reproduction asexuée, comme les bactéries qui se reproduisent par mitose, l'acquisition de nouveaux génotypes peut être obtenue par recombinaison homologue ou non-homologue. Elle est reconnue aujourd'hui comme une force évolutive majeure chez les bactéries (pour revues, Pallen and Wren, 2007; Didelot and Maiden, 2010). Le processus de recombinaison homologue, associé à des régions nucléotidiques répétées et impliqué dans la réparation de l'ADN, peut conduire à des réarrangements génomiques (accompagnés d'éventuelles réductions génomiques (Jackson *et al.*, 2011; Murray *et al.*, 2021)) et ainsi à de la diversité génotypique. La recombinaison hétérologue, transfert de gènes horizontal inter-individus, est médiée par trois mécanismes : conjugaison, transduction et transformation. Elle peut amener à l'acquisition de nouveaux gènes provenant d'individus parfois phylogénétiquement éloignés (Stukenbrock and McDonald, 2008; Bartoli, Roux and Lamichhane, 2016; McCann, 2020). Les types d'éléments mobiles acquis, plasmides, bactériophages et îlots de pathogénicité, peuvent porter des facteurs de virulence (Ziebuhr *et al.*, 1999; Spratt, 2001; Pallen and Wren, 2007; Stukenbrock and McDonald, 2008).
- la sélection naturelle : elle est le processus évolutif dirigé par lequel un allèle avantageux pour la survie dans l'environnement où évolue son porteur est sélectionné, conduisant à l'augmentation de sa fréquence dans la population tandis que la fréquence d'allèles

désavantageux diminuera. La sélection est à l'origine du changement du phénotype des individus d'une population vers celui plus adapté à son environnement (concept d'adaptation locale) donnant lieu, par exemple, à la fixation d'allèles de résistance chez l'hôte puis à celle d'allèles obtenus par mutation ou recombinaison permettant son contournement par le pathogène (McDonald and Linde, 2002). Certaines pratiques agricoles comme l'utilisation de pesticides ou le déploiement de variétés végétales résistantes représentent donc une forte pression de sélection sur les populations bactériennes.

- la dérive génétique : contrairement à la sélection, la dérive génétique est un processus aléatoire entraînant une variation de la fréquence des allèles dans une population suite à des phénomènes aboutissant à un échantillonnage des individus. Son effet est d'autant plus marqué lors de forte réduction de taille de la population, comme par goulot d'étranglement ou par effet de fondation (lorsque la dérive génétique est accompagnée de migration), ce qui peut amener à perte ou à la fixation d'allèles rares, respectivement (McDonald and Linde, 2002; McDonald, 2004). Cependant, dans le cas du pathogène bactérien multi-hôte (animal) *Staphylococcus aureus*, des études d'évolution expérimentale *in vivo* ont montré que des dizaines de mutations bénéfiques à l'infection animale avaient pu être conservées malgré une succession de goulots d'étranglement, soulignant les capacités d'adaptation de certains pathogènes bactériens aux changements d'hôtes (Bacigalupe *et al.*, 2019).
- la migration : cette force évolutive permet le déplacement de matériel génétique d'une population-source vers une population-puits au-delà des barrières isolant ces populations (McDonald and Linde, 2002; McDonald, 2004). Elle peut conduire à l'introduction de nouveaux allèles dans la population-puits et s'opposer à la sélection naturelle en limitant l'adaptation locale (Lenormand, 2002). Dans le cas particulier des émergences de bactéries pathogènes, la notion de réservoir-source et réservoir-puits est également utilisée pour décrire les migrations. Les espèces animales ou l'environnement (espèces non cultivées), d'une part, et l'espèce humaine ou les plantes cultivées, d'autre part, peuvent être considérés respectivement comme les sources et les puits (Gandon *et al.*, 2013).

L'étude de ces différentes forces *via* leurs effets sur la structure génétique des populations permet d'établir des profils de risques posés par les agents pathogènes ainsi que des stratégies de lutttes adaptées (McDonald and Linde, 2002) et constitue la base de l'épidémiologie moléculaire.

1.2. Épidémiologie

1.2.1. Épidémiologies classique et moléculaire

L'épidémiologie (au sens classique du terme) se définit comme l'étude de la fréquence et de la distribution spatio-temporelle des maladies dans le but de mieux comprendre les facteurs de risques impliqués dans leur propagation ainsi que de proposer des méthodes de gestion adaptées (définition de l'Organisation Mondiale de la Santé, 2021). Elle se base sur l'analyse de données mesurant la prévalence et l'incidence des maladies dans les populations afin de comprendre et prédire leur dynamique. La construction de modèles, incluant les caractéristiques de l'agent pathogène mais aussi la biologie de son hôte et les paramètres définissant son environnement, permet de tester nombre d'hypothèses, notamment sur la dynamique des populations, la dispersion et la transmission de la maladie et la réponse des hôtes au pathogène. L'épidémiologie s'intéresse ainsi à l'évolution et l'issue des maladies au sein des populations.

L'épidémiologie moléculaire est une amélioration récente de l'épidémiologie classique qui a vu le jour grâce aux outils de biologie moléculaire. L'utilisation de données génétiques permet entre autres d'identifier l'agent pathogène causatif d'une maladie, de détecter sa présence et suivre sa dynamique spatio-temporelle au sein d'une population, d'établir les relations de parentés entre individus afin de mesurer un réseau de transmission ou encore d'étudier les déterminants génétiques de sa virulence. De plus, comme illustré en Figure 1.2.1.A, lorsque l'échantillonnage est associé à des métadonnées telles que la lignée évolutive de l'hôte, la localité et date de prélèvement ou encore, les caractéristiques environnementales, l'épidémiologie moléculaire permet d'inférer une histoire évolutive plus précise du pathogène, en révélant son hôte, lieu et date d'origine ainsi qu'en précisant le rythme de son expansion spatio-temporelle. Ces connaissances peuvent alors contribuer à l'édification de moyens de surveillance et de luttés efficaces de la maladie en identifiant les populations-sources ou les voies de dissémination du pathogène et en estimant les facteurs biotiques et abiotiques favorables à son expansion (Li, Grassly and Fraser, 2014).

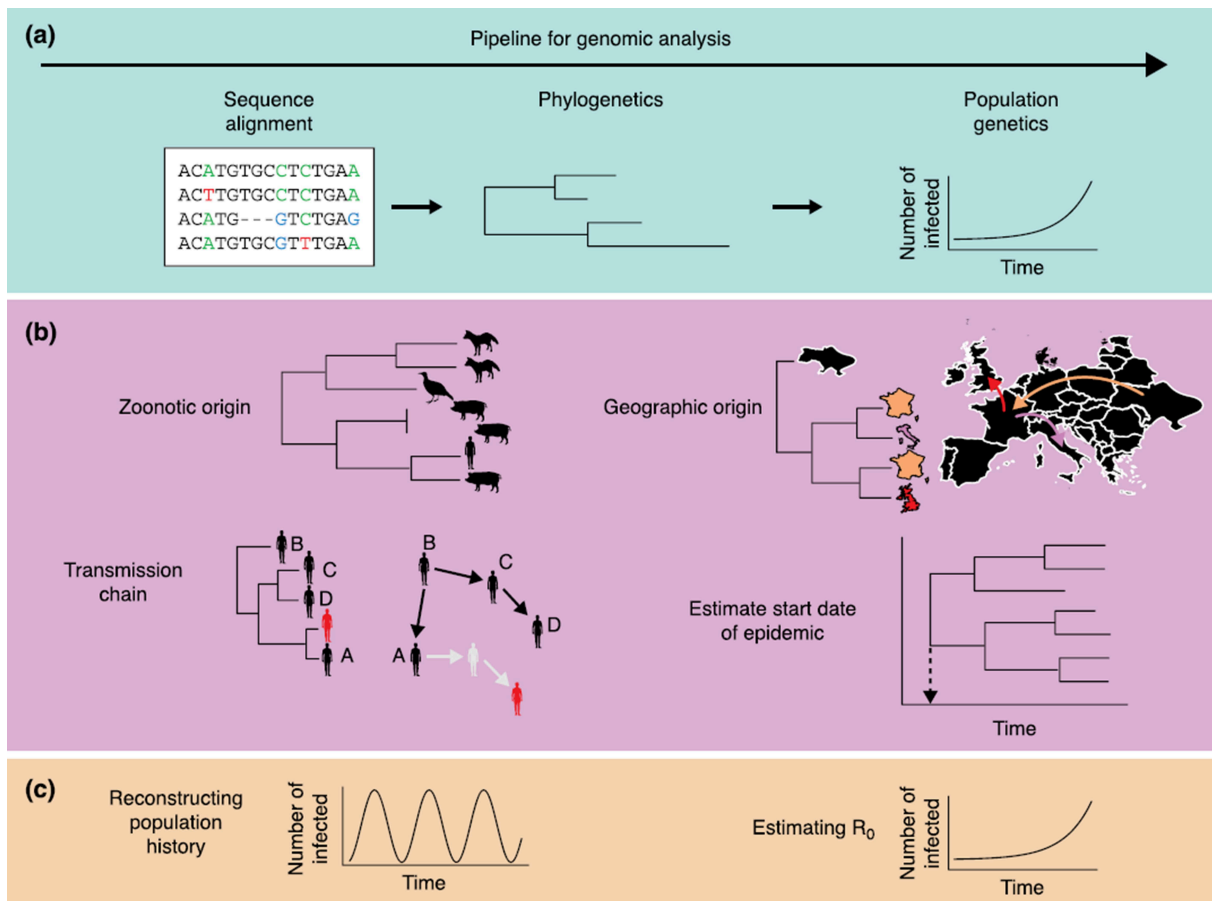


Figure 1.2.1.A. Contribution des analyses génomiques aux études épidémiologiques des maladies infectieuses émergentes (Li, Grassly and Fraser, 2014, Figure 1). a. L'analyse génomique débute par le séquençage et l'alignement de plusieurs séquences nucléotidiques du pathogène à partir desquelles une phylogénie peut être construite pour représenter les relations de parentés entre échantillons. A partir de cette phylogénie, des modèles simples de génétique des populations basés sur la théorie de la coalescence peuvent renseigner l'histoire démographique du pathogène. b. Associées aux métadonnées, les analyses phylogénétiques peuvent par la suite révéler les hôtes d'origine, les patrons spatio-temporels de la dispersion du pathogène et permettre de reconstruire les chaînes de transmission. c. L'utilisation de méthodes modélisant la démographie permet la reconstruction des dynamiques épidémiologiques passées et l'estimation de paramètres épidémiologiques tels que le nombre de reproduction de base.

Avec l'avènement des méthodes de séquençage haut débit (*High-Throughput Sequencing*, aussi appelé séquençage parallèle massif), il est aujourd'hui possible de suivre une épidémie en temps réel en capturant une fraction non négligeable de la diversité des pathogènes en circulation, comme illustré dans le cadre du contexte actuel de la pandémie de COVID-19 (van Dorp *et al.*, 2021). De plus, ces dernières décennies ont été marquées par l'amélioration des méthodologies en génomique, modélisation et phylogénie, permettant d'analyser des données génétiques de plus en plus longues et volumineuses, affinant ainsi notre capacité à étudier l'histoire de l'émergence et de l'évolution des pathogènes (Croucher and Didelot, 2015).

1.2.2. Etude de l'histoire évolutive des agents pathogènes

La reconstruction de l'histoire évolutive d'agents pathogènes depuis des données génétiques peut s'effectuer grâce à de nombreuses méthodes d'analyses, la plupart étant basées sur la théorie de la coalescence (pour revues, Stukenbrock and McDonald, 2008; Grünwald and Goss, 2011; Hartfield, Murall and Alizon, 2014; Li, Grassly and Fraser, 2014; Croucher and Didelot, 2015). La théorie de la coalescence est un modèle classique de génétique des populations théorisé au début des années 1980 par Kingman et Tajima, reposant sur un principe simple : la simulation des généalogies possibles d'un échantillonnage de gènes en remontant dans le passé jusqu'à l'ancêtre commun le plus récent (MRCA, *most recent common ancestor*) de l'échantillonnage (Kingman, 1982; Tajima, 1983). Dans ce contexte, deux grandes classes de méthodes se distinguent selon la nature des échantillons requis pour leur utilisation : celles nécessitant des données échantillonnées à l'échelle de la population ou de l'individu.

Les méthodes dites « populations centrées » nécessitent des données échantillonnées au sein de plusieurs (au moins deux) populations et se basent sur l'étude de la répartition de la diversité génétique au sein et entre ces populations pour mesurer et quantifier les forces évolutives qui s'y appliquent et ainsi reconstruire des *scenarii* évolutifs. Ces derniers peuvent par exemple permettre d'identifier les populations d'origine par rapport aux populations introduites, d'estimer les échanges (flux de gènes) entre les populations ainsi que les changements démographiques au cours du temps ou encore d'identifier des traces d'admixture suggérant l'échange horizontal de matériel génétique entre populations (Estoup and Guillemaud, 2010). De telles analyses peuvent être réalisées *via* des modèles relativement simples d'analyse du F_{ST} (un indice mesurant de la différenciation génétique inter-populations à partir de loci neutres) ou encore, avec des outils de modélisation plus raffinés comme l'ABC (*approximate Bayesian computation*), une approche Bayésienne qui prend en compte la stochasticité de l'histoire démographique et génétique des populations considérées et permet de reconstruire les routes invasives les plus probables (Beaumont, Zhang and Balding, 2002). Bien qu'en théorie très puissantes et informatives, l'inconvénient majeur de ces méthodes réside dans les hypothèses sous-jacentes aux modèles théoriques qui les constituent, incluant généralement la panmixie (reproduction aléatoire des gamètes) ou l'absence de sélection, deux conditions rarement applicables dans l'analyse de données génomiques de pathogènes (Grünwald and Goss, 2011).

Les méthodes « individus centrées » s'intéressent à l'analyse des relations de parentés directes entre individus, indépendamment de leur appartenance à une ou plusieurs populations. Elles renferment notamment les outils de reconstruction phylogénétique visant à inférer les relations de parenté entre individus sur la base de la distribution de caractères génétiques partagés. Un arbre phylogénétique est un graphe connexe non cyclique, suivant un cours historique, dans lequel les individus se trouvent aux

sommets des branches (feuilles ou nœuds externes) et sont reliés entre eux par des nœuds internes représentant leurs ancêtres communs hypothétiques. Il est indicatif de la question « qui est le plus proche parent de qui ? » (Darlu and Tassy, 1993). Les outils de reconstruction phylogénétique font l'hypothèse forte que les différences génétiques permettant de reconstruire les relations de parenté entre individus apparaissent par mutation seulement et sont transmises verticalement à la descendance. Dans le cas d'individus ou d'espèces pratiquant fréquemment la recombinaison, cela implique de devoir détecter et s'affranchir des portions génomiques issues de recombinaison pour ne pas biaiser la reconstruction phylogénétique (Awadalla, 2003; Didelot and Maiden, 2010).

Trois grandes approches existent pour reconstruire une phylogénie (Darlu and Tassy, 1993) :

- l'approche cladistique repose sur la mise en évidence des séries de transformation des caractères de l'état plésiomorphe (ancestral) à l'état apomorphe (dérivé) et est basée sur l'héritage de caractères provenant d'une ascendance commune (homologie) avec modification de caractères indépendants entre eux sans modèle explicite de leur évolution. C'est une approche hypothético-déductive qui nécessite l'application du principe de parcimonie mais peut être limitée dans son utilisation à cause de la saturation des données (probabilité que la transformation des caractères entraîne un retour à l'état ancestral) et l'impossibilité de reconstruire avec robustesse l'évolution des séquences par son manque de modèle d'évolution explicite.
- l'approche phénétique, basée sur la ressemblance globale observée, repose sur le postulat que le degré de ressemblance entre individus est corrélé à leur degré de parenté. La ressemblance globale est alors calculée selon une méthode de distance, ou dissimilitude, entre paires de séquences alignées dont la longueur des branches représente la distance génétique entre séquences (Felsenstein, 1996). Cependant, la ressemblance observée entre individus peut aussi bien être due au partage de caractères hérités d'un ancêtre commun proche (apomorphies) mais aussi plus éloigné (plésiomorphies) ou à une identité non due à un ancêtre commun (homoplasies par convergences, réversions...). En cherchant à répondre à « qui se ressemble le plus? », la phénétique s'éloigne de la question « qui est le plus proche parent de qui? ».
- l'approche probabiliste repose sur la conception que la transformation des caractères obéit à des lois de probabilité définies *a priori*. C'est une méthode basée sur la vraisemblance (probabilité conditionnelle d'observer des données selon un modèle d'évolution donné), utilisant un modèle d'évolution explicite permettant de compenser la saturation des données. Le modèle d'évolution comprend un arbre phylogénétique et une description de comment les

caractères se transforment le long des branches de l'arbre (ces longueurs sont ainsi proportionnelles au nombre de transformations). Pour les données génétiques, le modèle d'évolution est défini avec un modèle de substitution des séquences nucléiques. Il en existe plusieurs de complexité (et représentativité de la réalité) variable (Arenas, 2015). La méthode de reconstruction de la phylogénie par maximum de vraisemblance (*maximum likelihood*, ML) recherche les valeurs des paramètres du modèle d'évolution (dont l'arbre) dont la vraisemblance est maximisée. La méthode en inférences Bayésiennes (*Bayesian inferences*, IB) définit des valeurs de paramètres du modèle puis leur assigne une distribution des probabilités *a priori* sur laquelle elle réalise des inférences des valeurs des paramètres. Des probabilités postérieures (probabilités que le résultat inféré soit vrai sachant les données de départ) sont calculées à partir des combinaisons de distributions et permettent le choix du résultat le plus probable.

Appliquées à des données génétiques provenant d'agents pathogènes, les inférences phylogénétiques permettent de reconstruire leur histoire évolutive (Hartfield, Murall and Alizon, 2014), comme précédemment illustré en Figure 1.2.1.A. Dans ce contexte, l'inclusion de la temporalité dans la phylogénie permet de convertir une divergence mesurée en taux de transformations par caractère (substitution par sites pour les séquences nucléiques) en unité de temps, et ainsi de dater les événements de divergence entre lignées. Dans un contexte épidémiologique, dater (ou calibrer) un arbre phylogénétique permet de borner temporellement l'émergence d'un pathogène ou d'une lignée évolutive. Comme schématisé sur la Figure 1.2.2.A, l'émergence d'un pathogène se situe entre deux événements de divergence : celui liant la lignée du pathogène et celle de son plus proche parent (nœud de l'origine de la lignée du pathogène) et celui représentant l'ancêtre commun de l'ensemble des souches du pathogène (nœud de diversification du pathogène) (Ho and Duchêne, 2020). De façon similaire, la datation d'un arbre peut également servir à borner temporellement d'autres événements, comme par exemple l'acquisition/perte de matériel génétique (*e.g.* éléments insertionnels, gènes, plasmides...) modifiant le phénotype d'un pathogène (*e.g.* motilité, virulence, résistance à des pesticides) (Spagnoletti *et al.*, 2014; Wang *et al.*, 2018). La calibration temporelle d'un arbre phylogénétique peut être réalisée selon trois méthodes distinctes, applicables de façon indépendantes ou conjointes (Ho and Duchêne, 2014) :

- en *rate-dating*, calibration par un taux de mutation connu exprimé en nombre de mutations par site et par unité de temps. Ce taux peut avoir été estimé directement au laboratoire sur l'espèce d'intérêt, provenir d'une datation indépendante réalisée avec une autre méthode ou bien avoir été extrapolé à partir d'un taux connu sur une espèce proche de celle étudiée.

La calibration en *rate-dating* fait donc l'hypothèse forte que le taux de mutation utilisé est représentatif de l'ensemble des individus sur le pas de temps analysé.

- en *node-dating*, en associant des dates à un (ou plusieurs) nœud(s) interne(s) d'un arbre sur la base d'informations issues du registre fossile, d'événements géologiques ou biogéographiques datés ou encore d'éléments historiques. Cependant, ces événements sont datés indirectement et rarement avec une grande précision. Cette approche est couramment utilisée pour des phylogénies au-dessus du niveau populationnel sur des temps géologiques mais est plus rarement applicable sur pathogènes qui ne fossilisent pas (Ho *et al.*, 2011).
- en *tip-dating*, la calibration est effectuée en utilisant l'information temporelle contenue par les individus aux sommets (feuilles ou nœuds terminaux) de l'arbre pour lesquels les dates d'échantillonnages sont connues et variables (Figure 1.2.2.A). Cette méthode est plus directe que les deux précédentes, réalisant moins d'hypothèse quant à la composante temporelle du jeu de données ; elle ne peut cependant être appliquée que sur populations dites en « évolution mesurable » (*measurably evolving populations, MEP*) présentant une structure temporelle (Drummond *et al.*, 2002; Rieux and Balloux, 2016). Historiquement, les MEP étaient plutôt des populations de microorganismes caractérisés par des taux de substitution rapides et échantillonnés intensément durant une épidémie mais les innovations des techniques de séquençage à haut débit combinés à l'avènement de la paléogénomique (section 1.3.1.) ont récemment permis la généralisation des MEP à une large gamme d'organismes (Biek *et al.*, 2015).

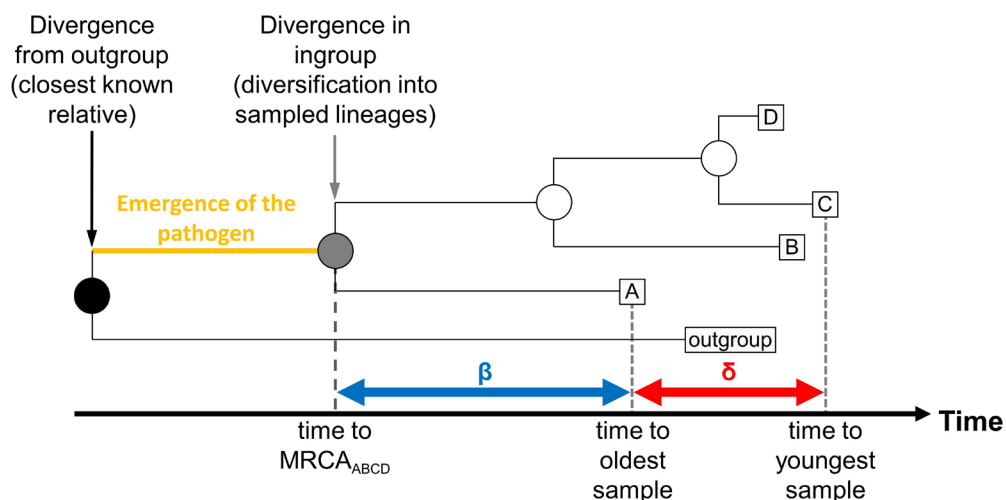


Figure 1.2.2.A. Calibration temporelle d'un arbre phylogénétique par *tip-dating* et estimation temporelle de l'émergence d'un pathogène. Schématisation d'un arbre phylogénétique composé d'un échantillonnage hétérochrone des individus A, B, C et D d'un pathogène d'intérêt (ingroup) et un outgroup utilisé pour orienter et enraciner la phylogénie. Les individus (carrés) sont reliés entre eux par leurs ancêtres communs hypothétiques (cercles). La population de l'ingroup présente une structure temporelle si la divergence moléculaire δ accumulée entre les

dates de collectes des échantillons A et C est statistiquement mesurable et détectable lorsque comparée à la divergence moléculaire β depuis son ancêtre commun hypothétique le plus récent (*Most Recent Common Ancestor*, MRCA). Dans une telle situation, une calibration de l'arbre par *tip-dating* est possible et permet de co-estimer taux de substitution et âge de tous les nœuds internes de l'arbre (cercle). La période d'émergence du pathogène étudié peut ainsi être estimée et bornée par les dates des nœuds de la branche reliant le MRCA des individus ABCD à celui incluant l'outgroup (branche jaune).

La présence de structure temporelle dans un jeu de données hétérochronique peut être détectée selon trois tests (Figure 1.2.2.B) :

- un test de régression linéaire réalisé à partir d'un arbre phylogénétique entre la date d'échantillonnage des individus du groupe d'étude (ou ingroup) et la distance les séparant de la racine (*i.e.*, le nombre de mutations accumulées chez les échantillons depuis leur ancêtre commun hypothétique le plus récent, Buonagurio *et al.*, 1986). Un signal temporel est considéré présent si une corrélation linéaire positive est détectée. Dans ce cas, le test fournit également une estimation du taux de substitution sous l'hypothèse d'horloge moléculaire stricte (la valeur de la pente), le degré d'association des séquences à une horloge moléculaire stricte (la valeur du coefficient de régression R^2) ainsi que l'âge de la racine de l'arbre (la valeur de l'ordonnée à l'origine). Ce test a été tout d'abord implémenté dans l'outil TempEst (anciennement Path-O-Gen) permettant d'explorer la structure temporelle à l'échelle globale d'un arbre (Rambaut *et al.*, 2016) et plus récemment dans l'outil PhyloSTemS, développé afin de tester l'existence du signal temporel à chacun des nœuds de l'arbre plutôt que seulement à sa racine (Doizy *et al.*, 2020). La rapidité et la facilité d'exécution de ce test ont participé à sa popularité et à son utilisation massive. Cependant, le test dans sa version classique repose sur l'hypothèse d'indépendance des données qui n'est pas respectée puisque les individus dont sont issues les séquences sont affiliés par la structure de l'arbre phylogénétique. Afin de corriger ce biais, une approche non-paramétrique du test, basée sur la permutation des dates d'échantillonnage a été proposée (Navascués, Depaulis and Emerson, 2010).
- un test de randomisation des dates dont le principe est de générer un ensemble de jeux de données au sein desquels les dates d'échantillonnages ont été aléatoirement permutées entre les différentes séquences avant d'être analysés *via* la méthode Bayésienne d'inférence (Duchêne *et al.*, 2015). Un signal temporel est considéré présent si l'intervalle de confiance à 95% d'un paramètre estimé (souvent l'âge de la racine ou le taux de substitution) sur le jeu de données réel ne chevauche avec aucun des mêmes intervalles obtenus à partir des jeux de données randomisés. Une implémentation astucieuse de ce test permet de gagner en rapidité en réalisant les permutations durant une seule et même analyse Bayésienne (Trovão *et al.*,

2015). Le test de randomisation des dates montre ses limites lorsque le nombre de dates d'échantillonnages est faible puisque le nombre possible de permutations devient limité.

- un test de comparaison de l'ajustement statistique entre deux modèles Bayésiens dans lesquels les dates d'échantillonnage sont fixées ou ignorées, respectivement (Rambaut, 2000; Murray *et al.*, 2016). Un signal temporel est considéré présent lorsque l'inclusion des dates d'échantillonnage dans l'analyse améliore l'ajustement du modèle. Ce test a récemment été implémenté dans le logiciel BETS (Duchêne *et al.*, 2020).

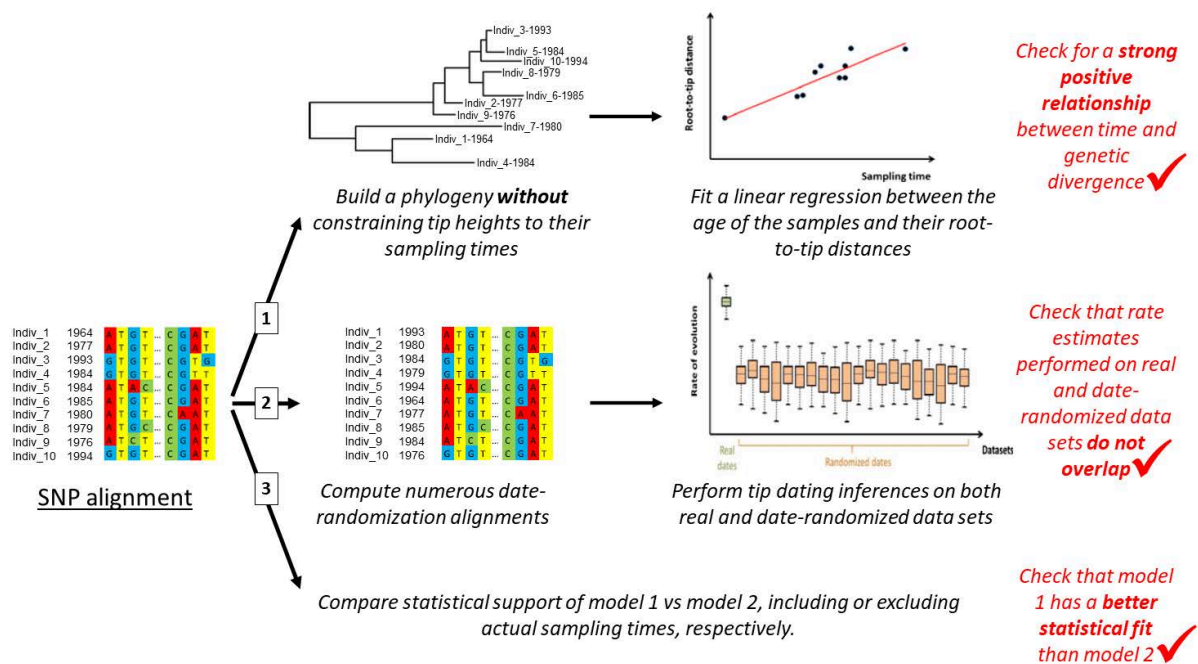


Figure 1.2.2.B. Méthodes permettant de tester la présence d'un signal temporel au sein d'un jeu de données de séquences nucléotidiques hétérochrones (Rieux and Balloux, 2016, modification de la Figure 2). 1, Ajustement d'une régression linéaire entre l'âge des échantillons et leurs distances à la racine; 2, Test de randomisation des dates et 3, Test de comparaison de l'ajustement statistique entre deux modèles Bayésiens incluant ou excluant les dates d'échantillonnages, respectivement.

Les tests de corrélation linéaire ainsi que de randomisation des dates peuvent cependant mener à une détection erronée de signal temporel lorsque la structure temporelle de l'arbre phylogénétique est « confondue » avec sa structure génétique, c'est-à-dire, lorsque des séquences proches phylogénétiquement ont tendance à provenir d'individus échantillonnés à des périodes similaires (Murray *et al.*, 2016). La présence de confusion entre ces structures peut se tester avec un test de Mantel entre une matrice de distances génétiques calculées par paire de séquences et une matrice de différences absolues des dates d'échantillonnage. Lorsqu'une confusion est identifiée, le test de randomisation doit être refait en prenant en définissant des clusters de séquences sur la base des dates d'échantillonnage et en procédant aux randomisations uniquement entre clusters (Murray *et al.*, 2016).

De façon analogue à la calibration temporelle d'une phylogénie, il est également possible d'intégrer une couche supplémentaire d'informations associée à la géographie dans les inférences phylogénétiques. La phylogéographie est l'étude phylogénétique de la distribution des géotypes dans l'espace permettant d'inférer l'histoire évolutive de populations dans un contexte géographique et de reconstruire l'évolution du caractère géographique des individus. Elle permet d'estimer les aires ancestrales et localiser l'origine géographique des lignées, ainsi que d'inférer le taux de migration entre populations grâce à la phylogéographie discrète, ou le taux de diffusion par phylogéographie continue (Figure 1.2.2.C) (pour revue, Rasmussen and Grünwald, 2020). Une analyse de phylogéographie discrète retrace les mouvements des lignées entre états discrets (les localisations) et construit une matrice de transition entre ces états correspondant à la matrice de migration permettant d'estimer à quel taux les différentes lignées se déplacent. La contribution de différentes variables environnementales à la transition entre localisations peut être estimée grâce à une paramétrisation de type GLM (*generalised linear model*, modèle linéaire généralisé) et permettre ainsi de prédire la dynamique de dispersion du pathogène (Lemey *et al.*, 2014). La phylogéographie continue retrace explicitement les mouvements des lignées dans l'espace en assumant que les lignées se diffusent le long d'un espace continu au lieu de transiter d'un état à un autre. Lorsqu'appliqué à une phylogénie calibrée temporellement, le taux auquel l'épidémie peut se diffuser dans l'espace (vitesse du front d'onde) ainsi que la distance que peuvent parcourir les lignées selon une période donnée (coefficient de diffusion) peuvent être estimés (Rasmussen and Grünwald, 2020). Le choix de traiter l'espace comme une variable discrète ou continue dépend de l'épidémiologie du pathogène, c'est-à-dire, sa capacité à transiter d'un état à un autre sur de grandes distances par dispersion de propagules ou sous l'effet de l'activité humaine, ainsi que l'échantillonnage dans l'espace. Ainsi, une analyse de phylogéographie discrète est plus appropriée lorsque le pathogène peut « sauter » de longues distance ou lorsque les états ancestraux correspondent à des populations différentes de leurs descendants, et inversement, la phylogéographie continue se prête mieux à des échantillonnages distribués continuellement dans l'espace ou lorsque les pathogènes se diffusent de proche en proche (Rasmussen and Grünwald, 2020).

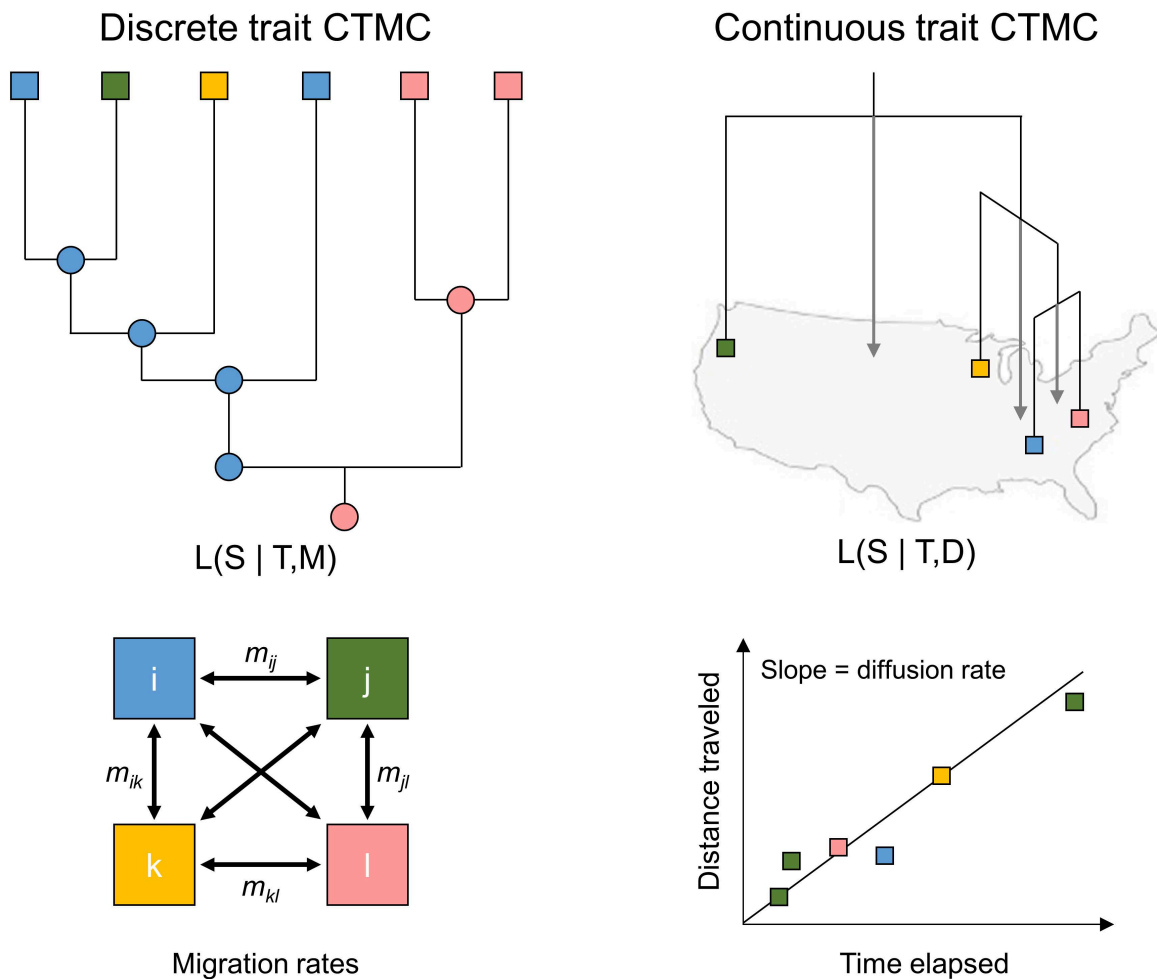


Figure 1.2.2.C. Phylogéographie discrète (Rasmussen and Grünwald, 2020, Figure 1A&B). Dans les modèles de phylogéographie de traits discrets en chaîne de Markov en temps continu (*continuous-time Markov chain*, CTMC), les lignées migrent entre les différentes localisations à des taux déterminés dans une matrice de taux. Sous les modèles CTMC, la vraisemblance (*likelihood*, L) des états S des individus échantillonnés est calculée depuis la phylogénie T et la matrice de migration M. Les lignées sont supposées migrer indépendamment les unes des autres. Pour les modèles CTMC de traits continus, les lignées sont assumées comme diffusant de manière continue dans l'espace. Si le temps et la localisation aux nœuds ancestraux (flèches grisées) peuvent être inférés, il est possible d'estimer la distance parcourue par chaque lignée sur l'intervalle de temps allant du début à la fin de sa branche. La régression de la distance parcourue par rapport à la longueur des branches fournit une estimation du taux de diffusion D d'une épidémie dans l'espace.

Ainsi, la reconstruction des états ancestraux aux nœuds internes d'une phylogénie peut s'effectuer pour un trait lié à la géographie, mais cela est également réalisable pour d'autres traits d'intérêt chez un pathogène comme l'hôte (Weinert *et al.*, 2012; Patané *et al.*, 2019) ou d'autres caractères associés à la virulence (*e.g.* taux de sporulation, niche écologique... (Ismail *et al.*, 2016)). De la même façon qu'une datation par *tip-dating* nécessite l'existence d'un signal temporel, une analyse de reconstruction d'état ancestraux requiert un signal phylogénétique qui se définit comme l'existence d'une corrélation entre la structure phylogénétique et la distribution du trait d'intérêt (géographie,

hôte ou autre). Sa présence peut être testée grâce à un test d'association (Parker, Rambaut and Pybus, 2008). Cependant, l'analyse phylogéographique est sensible à l'échantillonnage et la surreprésentation d'une population par rapport aux autres ou l'absence de représentation de la population fondatrice peuvent aboutir à une reconstruction erronée de l'état géographique ancestral du pathogène (Rasmussen and Grünwald, 2020).

Dans une étude ayant combiné calibration temporelle et phylogéographie, *Trovão et al.* (2015) ont pu inférer une origine tanzanienne datant de 1852 du virus de la panachure jaune du riz (*rice yellow mottle virus*, RYMV), une maladie circulant dans la plupart des pays producteurs de riz en Afrique et reportée pour la première fois au Kenya en 1966. L'arrivée de la maladie en Afrique de l'ouest aurait eu lieu en Côte d'Ivoire en 1887. En combinant une reconstruction phylogéographique discrète de la diffusion du virus avec l'analyse statistique d'un modèle linéaire généralisé (GLM, *generalised linear model*), les auteurs ont démontré l'importante contribution de la connectivité et de l'intensité de la culture du riz dans la dispersion des populations virales. La réalisation d'une analyse de reconstruction phylogéographique en espace continu a permis de révéler une dynamique d'expansion virale plus rapide en Afrique de l'ouest, une différence que les auteurs associent d'une part à l'existence de barrières à la dispersion à l'est et d'autre part à celle d'un système fluvial ayant pu favoriser la propagation virale à l'ouest. Ainsi, la dispersion spatio-temporelle de RYMV aurait été globalement guidée par l'intensification et l'expansion de la culture du riz en Afrique (adoption de nouveaux modes de production et utilisation de variétés asiatique sensibles).

A ce jour, la majorité des études d'épidémiologie moléculaire chez les pathogènes des cultures ont été réalisées à partir d'échantillons contemporains datant des quelques dernières décennies. Comme illustré par l'étude de l'histoire et de la dispersion de la panachure du riz en Afrique, ces études épidémiologiques peuvent être très informatives mais elles ne sont pas toujours réalisables par manque de signal temporel. Cette limitation peut cependant être dépassée en intégrant des échantillons historiques augmentant l'âge de l'échantillon le plus ancien et la probabilité de mettre en évidence un signal temporel (Spyrou *et al.*, 2019; Ho and Duchêne, 2020; Duchêne *et al.*, 2020).

1.3. Ressources historiques

1.3.1. L'avènement de l'ADN ancien

L'étude de l'ADN ancien ou paléogénétique a débuté dans les années 1980 avec le clonage de quelques centaines de paires de bases d'un gène chez un échantillon de couagga (*Higuchi et al.*, 1984), une sous-espèce de zèbre d'Afrique du Sud disparue près d'un siècle plus tôt, ainsi que chez une momie égyptienne (Pääbo, 1985). Ces études ont démontré la persistance des acides nucléiques dans les échantillons historiques (issus de collections naturalistes) ou anciens (issus de restes archéologiques

ou paléontologiques) (Yoshida, Sasaki and Kamoun, 2015) mais également la possibilité de les extraire et de les séquencer. L'avènement de l'amplification de l'ADN par réaction en chaîne (PCR, *Polymerase Chain Reaction*), peu après, a permis une meilleure accessibilité et utilisation de leur ADN endogène, souvent en très faible quantité et caractérisé par des dégradations moléculaires nécessitant un traitement spécifique au laboratoire (voir partie dédiée en chapitre 1.3.4.). L'introduction des NGS (*Next generation sequencing*, séquençage de seconde génération, aussi appelé séquençage à haut débit) a substantiellement augmenté la quantité d'ADN qui pouvait être traitée et de récentes avancées méthodologiques ont ensuite résulté en des protocoles d'extraction d'acides nucléiques et de préparation de bibliothèques de séquençage de plus en plus efficaces (Shapiro and Hofreiter, 2014). La paléogénétique reste cependant toujours limitée par la quantité et la qualité des acides nucléiques survivant dans un échantillon.

Grâce aux innovations des techniques de séquençage à haut débit, l'étude de l'ADN ancien, extrait sous la forme de nombreux fragments de petite taille, a permis l'essor de la paléogénomique avec l'obtention du premier génome ancien issu d'un mammouth (Miller *et al.*, 2008), d'hominidés avec un Néandertal (Green *et al.*, 2010), puis de la bactérie pathogène responsable de la peste (Bos *et al.*, 2011). Le génome ancien le plus âgé obtenu à ce jour provient d'un mammouth de plus d'un million d'années (van der Valk *et al.*, 2021). Une multitude de tissus a pu fournir de l'ADN ancien : peau, poils et cheveux, plumes, os et dents pour les animaux ; feuilles, fruits et graines pour les plantes, ces différents tissus pour leurs microorganismes associés ; calculs, coprolithes et sédiments pour les microorganismes du microbiote également... (pour revues, Pääbo *et al.*, 2004; Wandeler, Hoeck and Keller, 2007; Hagelberg, Hofreiter and Keyser, 2015; Arning and Wilson, 2020).

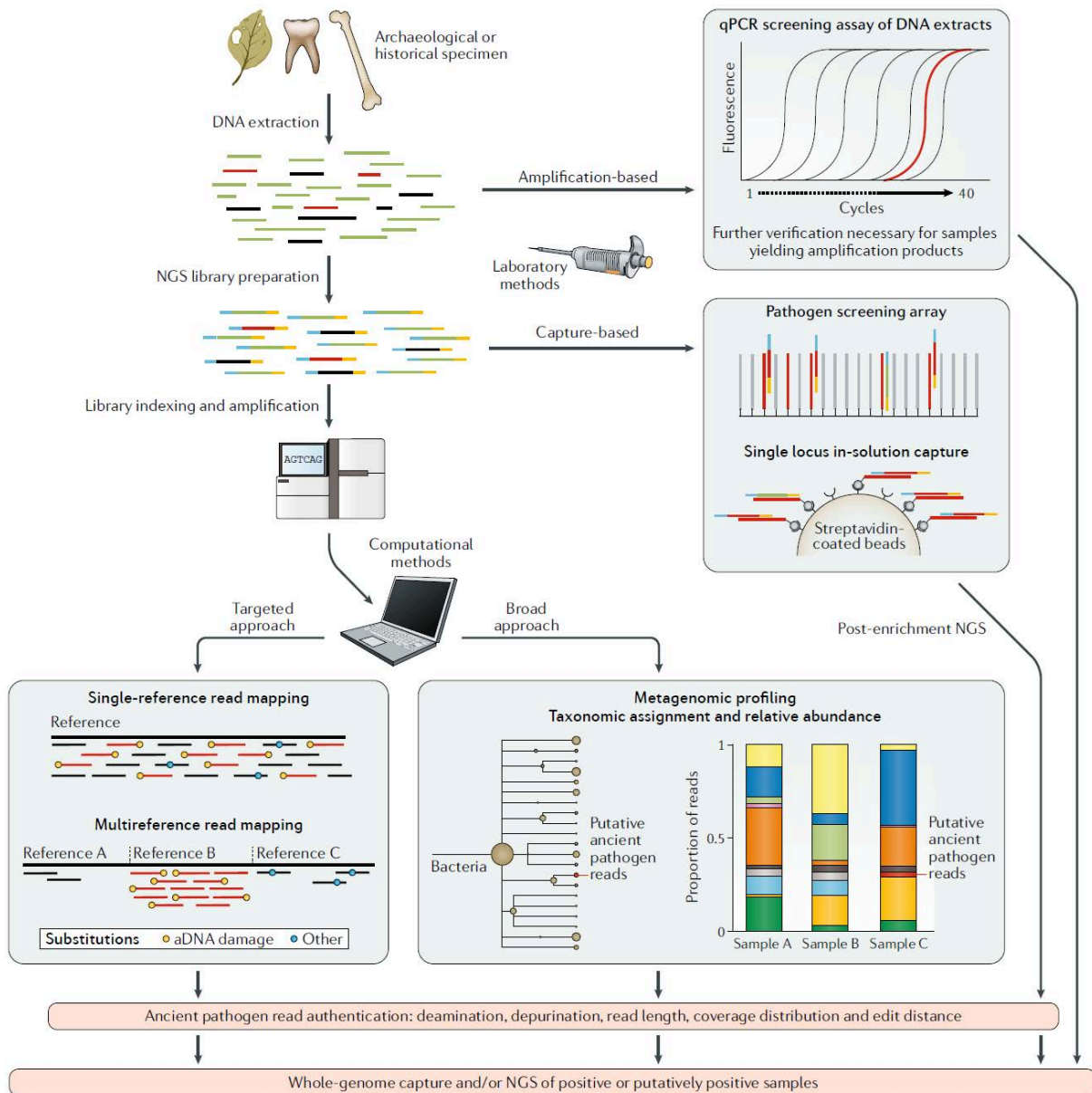
Les ADNs anciens ont ainsi pu être utilisés pour caractériser génétiquement les individus du passé et construire des phylogénies moléculaires des espèces, pouvant désormais inclure des espèces éteintes. Lors d'échantillonnages abondants au cours du temps, l'histoire et la dynamique de populations ont également pu être étudiées, notamment par rapport à leur environnement (*e.g.* Thomas *et al.*, 1990; Shapiro *et al.*, 2004; Lorenzen *et al.*, 2011). Sur plantes et animaux dont hominidés, les ADNs anciens ont pu permettre d'estimer leur origine et les processus de leur domestication (pour revues, Pääbo *et al.*, 2004; Orlando and Cooper, 2014; Hofreiter *et al.*, 2015). Sur individus malades, les agents pathogènes causatifs ont pu être identifiés et caractérisés et l'histoire des maladies reconstruites (pour revues, Spyrou *et al.*, 2019; Arning and Wilson, 2020; Arriola, Cooper and Weyrich, 2020).

1.3.2. Application aux cas des maladies infectieuses

La paléogénomique appliquée à des agents pathogènes rassemble diverses disciplines comme la microbiologie, l'épidémiologie, la biologie évolutive, la génomique comparative et l'histoire. Elle

s'intéresse à améliorer la compréhension des interactions entre les pathogènes et leurs hôtes, à découvrir les origines des pathogènes et distinguer les processus (génétiques, démographiques, sociaux-environnementaux...) impliqués dans leur(s) émergence(s) au sein de populations afin de reconstruire leur histoire évolutive (Spyrou *et al.*, 2019).

L'étude des pathogènes anciens, le plus souvent inclus dans leurs hôtes et non cultivables, nécessite de ségréger les données moléculaires de l'agent pathogène, compris dans la masse des données majoritaires de l'hôte. Pour cela, deux approches sont possibles : *i)* en séquençant par une approche dite de « *shotgun* » l'ensemble de l'ADN extrait d'un échantillon, au sein duquel la présence du pathogène a été éventuellement préalablement détectée, avant d'identifier et de sélectionner la part des séquences du pathogène avec des outils bioinformatiques ; *ii)* alternativement, en ne séquençant que l'ADN du pathogène d'intérêt préalablement « capturé » grâce à des sondes spécifiques (Bahcall, 2013) (Figure 1.3.2.A). Dans le cas *i)*, l'approche globale permet d'intégrer la bactérie d'intérêt dans un contexte métagénomique, en termes quantitatifs et taxonomiques. Dans les deux cas *i)* et *ii)*, la séquence ADN du pathogène peut être reconstruite (par alignement sur séquence de référence ou assemblage *de novo*) et, par comparaison à des séquences modernes, ses sites variables identifiés. L'évaluation du contenu en gènes est d'intérêt particulier pour ce qui est des facteurs de virulence impliqués dans la pathogénie. Grâce à des analyses phylogénétiques, les relations de parenté entre pathogènes anciens et modernes peuvent être reconstruites et la dynamique de leurs populations inférée, permettant éventuellement l'identification de lignées éteintes ou le remplacement de populations au fil des années (après s'être affranchis des régions recombinantes). Les spécimens anciens peuvent alors être utilisés directement pour la calibration de l'horloge moléculaire lorsqu'un signal temporel est identifiable. Celle-ci permet d'estimer le taux de mutation du pathogène, les dates de divergence des différentes lignées, ou encore la taille effective des populations (pour revue, Spyrou *et al.*, 2019).



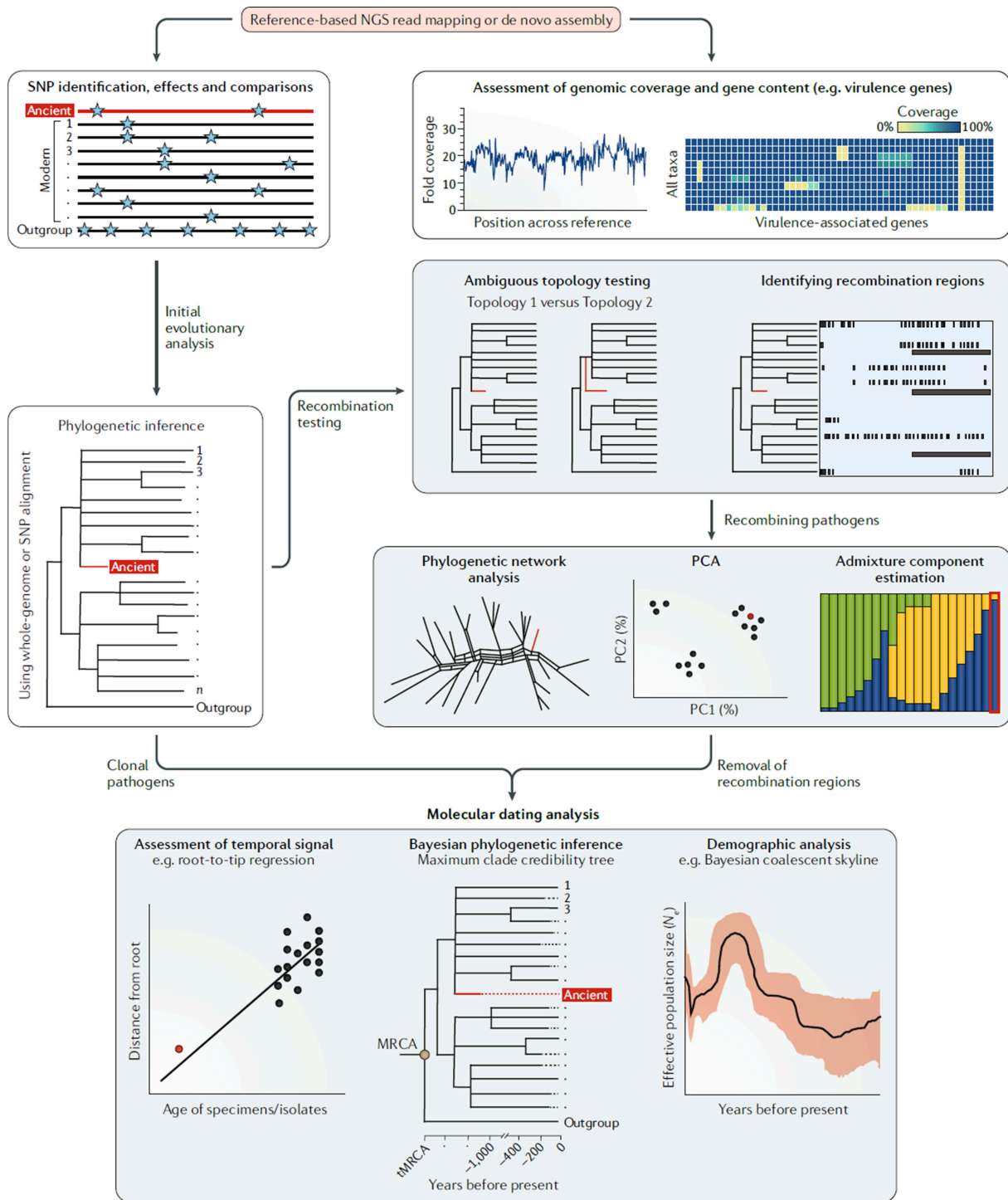


Figure 1.3.2.A. Méthodes d'obtention et d'analyse des données génomiques issues de pathogènes (Spyrou et al., 2019, Figures 2 modifiée et 3). Aperçu des analyses typiquement réalisables sur données génomiques de pathogènes comprenant de l'ADN ancien. Abréviations : NGS, *Next Generation Sequencing* ; SNP, *Single-Nucleotide Polymorphism* ; PCA, *Principal Component Analysis* ; PC, *Principal Component* ; MRCA, *Most Recent Common Ancestor* ; tMRCA, *time since Most Recent Common Ancestor*.

Les premières données d'ADN ancien issues de microorganismes ont été obtenues à partir de restes osseux anciens symptomatiques, porteurs de *Mycobacterium tuberculosis* (Spigelman and Lemma,

1993), la bactérie pathogène responsable de la tuberculose. Suite à cette preuve de concept, les travaux sur ADN ancien de microorganismes se sont orientés vers les agents pathogènes pouvant être présents dans les os chez l'Homme, faisant de *Mycobacterium tuberculosis*, *Mycobacterium leprae*, responsable de la lèpre, et *Yersinia pestis*, bactérie responsable de la peste, les microorganismes pathogènes les mieux étudiés en paléomicrobiologie (pour revue, Arriola, Cooper and Weyrich, 2020). La liste des génomes anciens s'est ensuite étendue à d'autres hôtes et microorganismes : dans leur revue de 2019, Spyrou *et al.* citent une quinzaine de pathogènes bactériens ou viraux anciens affectant l'Homme ou des animaux d'élevage, un premier virus de plante, ainsi que les premiers génomes eucaryotes anciens du parasite responsable du paludisme, et de l'oomycète du mildiou de la pomme de terre (Tableau 1.3.2.A)

Tableau 1.3.2.A. Données génomiques d'agents pathogènes anciens issus de spécimens historiques ou anciens (Spyrou *et al.*, 2019, Tableau 1).

	Pathogen	Infectious disease	Method of retrieval	Number of genomes	Biological insights	References
Bacterial pathogens	<i>Borrelia recurrentis</i>	Relapsing fever	Shotgun sequencing	1	<ul style="list-style-type: none"> Isolation from 15th-century CE human remains from Norway Genome signatures of reductive evolution, associated with typical virulence profile, and recent ecological adaptation 	(Guellil <i>et al.</i> , 2018)
	<i>Brucella melitensis</i>	Brucellosis	Shotgun sequencing	1	<ul style="list-style-type: none"> Isolation from a calcified nodule identified in an individual's pelvic girdle Presence of <i>B. melitensis</i> in Sardinia during the 14th century CE 	(Kay <i>et al.</i> , 2014)
	<i>Gardnerella vaginalis</i>	Bacterial vaginosis	Shotgun sequencing	1	<ul style="list-style-type: none"> Identified in human remains from Troy dating to 13th century CE Association with women's mortality during childbirth in the past The identified strain clusters among modern <i>G. vaginalis</i> diversity 	(Devault <i>et al.</i> , 2017)
	<i>Helicobacter pylori</i>	<ul style="list-style-type: none"> Ulcers of the upper gastrointestinal tract Increased risk of gastric carcinoma 	In-solution capture followed by NGS	1	<ul style="list-style-type: none"> Isolation from European Copper Age, 5,300-year-old mummy (Ötzi) Unadmixed strain, contrary to modern European strains, which are hybrids of two ancestral populations 	(Maixner <i>et al.</i> , 2016)
	<i>Mycobacterium leprae</i>	Lepromatous leprosy	<ul style="list-style-type: none"> Shotgun sequencing Microarray-based capture followed by 	27	<ul style="list-style-type: none"> First de novo assembled ancient pathogen genome Estimated emergence >5,000 years ago European origin of leprosy in the Americas High <i>M. leprae</i> diversity in medieval Europe 	(Schuenemann <i>et al.</i> , 2013; Mendum <i>et al.</i> , 2014; Schuenemann, Avanzi, <i>et al.</i> , 2018; Krause-Kyora, Nutsua, <i>et</i>

		NGS		<i>al.</i> , 2018)	
<i>Mycobacterium tuberculosis</i>	Tuberculosis	<ul style="list-style-type: none"> • Shotgun sequencing • Microarray-based capture followed by NGS 	19	<ul style="list-style-type: none"> • Genomes from pre-Columbian human infections show phylogenetic clustering within animal-adapted lineage present today in seals • Molecular dating analysis suggests emergence of MTBC <6,000 years ago • Analysis of European genomes shows past occurrence of multiple infections and suggests origin of lineage 4 during the 4th to 5th century CE 	(Chan <i>et al.</i> , 2013; Bos <i>et al.</i> , 2014; Kay <i>et al.</i> , 2015)
<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar Paratyphi C	Enteric (paratyphoid) fever	<ul style="list-style-type: none"> • Shotgun sequencing • Microarray-based capture followed by NGS • In-solution capture followed by NGS 	11	<ul style="list-style-type: none"> • <i>S. enterica</i> subsp. <i>enterica</i> serovar Paratyphi C presence in 12th-century CE Norway • Paratyphi C serovar was also identified among 16th-century individuals from Mexico that were associated with the major post-contact 'cocoliztli' Epidemic 	(Vågene <i>et al.</i> , 2018; Zhou <i>et al.</i> , 2018)
<i>Staphylococcus saprophyticus</i>	<ul style="list-style-type: none"> • Urinary tract infections • Puerperal fever 	Shotgun sequencing	1	<ul style="list-style-type: none"> • Identified in ~800-year-old human remains from Troy • Association with women's mortality during childbirth in the past • The identified lineage is not commonly associated with human disease today 	(Devault <i>et al.</i> , 2017)
<i>Tannerella forsythia</i>	Periodontal disease	Shotgun sequencing	1	<ul style="list-style-type: none"> • Isolation from medieval human remains (circa 950–1200 CE) • First pathogen genome reconstructed from ancient dental calculus 	(Warinner <i>et al.</i> , 2014)
<i>Treponema pallidum</i>	<ul style="list-style-type: none"> • Syphilis (<i>Treponema pallidum</i> subsp. <i>pallidum</i>) • Yaws (<i>Treponema pallidum</i> subsp. <i>pertenue</i>) • Bejel (<i>Treponema pallidum</i> subsp. <i>endemicum</i>) 	Microarray-based capture followed by NGS	3	<ul style="list-style-type: none"> • Isolated from individuals who lived in Mexico City between the 17th and 19th centuries CE • Different <i>Treponema</i> subspecies (<i>T. pallidum</i> subsp. <i>pallidum</i> and subsp. <i>pertenue</i>) caused similar skeletal lesions usually identifiable as skeletal syphilis in infants 	(Schuenemann, Kumar Lankapalli, <i>et al.</i> , 2018)

Viral pathogens	<i>Vibrio cholerae</i>	Cholera	Microarray-based capture followed by NGS	1	<ul style="list-style-type: none"> Isolation from 19th-century alcohol-preserved intestinal specimen from an individual affected during the second cholera pandemic The identified strain shows highest similarity with the classic pathogenic biotype O1 	(Devault <i>et al.</i> , 2014)
	<i>Yersinia pestis</i>	Bubonic, pneumonic and septicemic plague	<ul style="list-style-type: none"> Shotgun sequencing Microarray-based capture followed by NGS In-solution capture followed by NGS 	38	<ul style="list-style-type: none"> Bacterium affected humans as early as 5,000 years ago Both flea-adapted and non-adapted variants were present in Eurasia during the Bronze Age Causative agent of the Plague of Justinian (6th century CE) Causative agent of Black Death and persistence in Europe during the second plague pandemic (14th to 18th century CE) Possible European origin of third plague pandemic lineage 	(Bos <i>et al.</i> , 2011, 2016; Wagner <i>et al.</i> , 2014; Rasmussen <i>et al.</i> , 2015; Feldman <i>et al.</i> , 2016; Spyrou <i>et al.</i> , 2016, 2018; Andrades Valtueña <i>et al.</i> , 2017; de Barros Damgaard <i>et al.</i> , 2018; Namouchi <i>et al.</i> , 2018; Rascovan <i>et al.</i> , 2019)
	HBV	Viral hepatitis	<ul style="list-style-type: none"> Shotgun sequencing In-solution capture followed by NGS Whole-genome PCRb 	17	<ul style="list-style-type: none"> Identified in ancient human specimens as early as 7,000 years ago Neolithic genome lineage related to contemporary strains identified in African non-human primates Complex evolutionary history of HBV and identification of ancient recombination event giving rise to genotype A strains 	(Kahila Bar-Gal <i>et al.</i> , 2012; Krause-Kyora, Susat, <i>et al.</i> , 2018; Mühlemann, Jones, <i>et al.</i> , 2018; Patterson Ross <i>et al.</i> , 2018)
	HIV	AIDS	Whole-genome PCRb	8	<ul style="list-style-type: none"> Analysis of HIV RNA from archival specimens of seropositive individuals enrolled in HBV studies during the late 1970s HIV was introduced into the Americas from the Caribbean in the early 1970s 	(Worobey <i>et al.</i> , 2016)
	B19V	<ul style="list-style-type: none"> Erythema infectiosum (fifth disease) in children Arthropathies in adults Hydrops fetalis or fetal death in pregnant women Pure red-cell aplasia 	In-solution capture followed by NGS	10	<ul style="list-style-type: none"> Genomic signatures of B19V identified in human remains dating as early as ~7,000 years ago Contrary to previous estimates of a most recent common ancestor younger than 200 years, phylogenetic and molecular dating analysis of ancient genomes showed a much lengthier association of B19V with human populations 	(Mühlemann, Margaryan, <i>et al.</i> , 2018)

	Influenza virus	Influenza	Whole-genome PCRb	1	<ul style="list-style-type: none"> • First reconstructed genome from historical RNA virus • Avian source of 1918 influenza pandemic (Spanish flu, 1918–1920) • Reconstructed virus particle displayed increased virulence under laboratory conditions 	(Taubenberger <i>et al.</i> , 2005; Tumpey, 2005)
	VARV	Smallpox	In-solution capture followed by NGS	1	<ul style="list-style-type: none"> • Genome reconstruction from a 17th-century mummy from Lithuania • Recent emergence of 20th century VARV lineages (divergence during the 18th century CE) 	(Duggan <i>et al.</i> , 2016)
Eukaryotic pathogens	<i>Phytophthora infestans</i>	Late blight (also known as potato blight)	Shotgun sequencing	18	<ul style="list-style-type: none"> • First sequenced ancient eukaryotic (plant) pathogen genomes • Isolated from historical herbarium specimens • A unique <i>Phytophthora infestans</i> genotype caused the Irish potato famine and during the 1900s became replaced by the US-1 lineage that dominated worldwide until the 1970s 	(Martin <i>et al.</i> , 2013; Yoshida <i>et al.</i> , 2013)
	<i>Plasmodium falciparum</i> and <i>Plasmodium vivax</i>	Malaria	In-solution capture followed by NGS	5	<ul style="list-style-type: none"> • Oldest <i>Plasmodium falciparum</i> detection from southern Italy (1st to 2nd century CE) • <i>Plasmodium falciparum</i> and <i>Plasmodium vivax</i> mitochondrial genome isolation from 20th century microscopy slides • Possible introduction of <i>Plasmodium vivax</i> in the Americas through European contact 	(Gelabert <i>et al.</i> , 2016; Marciniak <i>et al.</i> , 2016)

a The indicated numbers include whole pathogen genomes and specimens yielding genome- wide data.

b Whole-genome PCR amplicons from the studies of influenza virus58, HIV57 and HBV54 that were sequenced using capillary sequencing (Sanger method).

Abbreviations: B19V, human parvovirus B19; CE, current era; HBV, hepatitis B virus; MTBC, *Mycobacterium tuberculosis* complex; NGS, next- generation sequencing; VARV, variola virus.

1.3.3. Les échantillons d'herbier: une ressource pour l'étude des maladies des plantes

Parmi les spécimens de plantes anciens et historiques, les herbiers offrent de vastes collections naturalistes sur les 400 dernières années renfermant plus de 350 millions d'échantillons d'une grande diversité taxonomique (Bieker and Martin, 2018). Dans nombre de cas, les espèces présentent un échantillonnage conséquent composé de réplicats génétiques, spatiaux, temporels qui permet aujourd'hui l'étude de leur évolution au cours du temps et de l'espace. Ces collections couvrent notamment des événements globaux (Lang *et al.*, 2019; Kistler *et al.*, 2020) comme la révolution industrielle dès le XVIII^{ème} siècle ou le dérèglement climatique depuis la fin du XX^{ème} siècle. D'autres touchent plus spécifiquement des espèces cultivées, comme le développement de la sélection variétale consciente aux XVIII et XIX^{ème} siècles ainsi que la révolution verte (politique de modernisation des techniques et d'augmentation des productions agricoles *via* l'utilisation de variétés sélectionnées et de produits phytosanitaires de synthèse) au XX^{ème} siècle. Géographiquement, les échantillons historiques peuvent provenir d'aires aujourd'hui difficiles à prélever (zones de conflit, réserves de la

biodiversité...) ou de populations protégées ou désormais disparues. Ils peuvent ainsi venir compléter un échantillonnage qui ne pourrait être effectué de nos jours.

Les échantillons infectés d'herbier, bien que moins fréquents que les spécimens sains, possèdent ces mêmes caractéristiques. Ils présentent alors plusieurs intérêts dans l'amélioration de la compréhension des mécanismes de l'émergence et évolution des pathogènes des cultures :

- si un échantillon d'herbier présente les symptômes d'une maladie, il constitue une trace directe de la présence du pathogène sur la plante hôte à la date et localité renseignées lors de la mise en collection. Cela a pu être démontré pour un basidiomycète pathogène de *Silene*, responsable de la maladie des anthères de fleurs, redéfinissant sa distribution spatio temporelle et ses hôtes aux Etats-Unis d'Amérique (Antonovics *et al.*, 2003). Une authentification moléculaire est néanmoins nécessaire lorsque les symptômes ne sont pas spécifiques ;
- si l'ADN de l'échantillon est exploitable, les séquences, voire le génome, du pathogène peuvent être reconstruits et, dans le cas d'approche de séquençage non ciblée, celles d'autres organismes associés (microbiote des tissus, plante hôte, communautés microbiennes de l'herbier) (Bieker *et al.*, 2020). Ces séquences peuvent alors être utilisées dans différentes analyses visant à étudier l'évolution des génomes, des facteurs de virulence (ou des gènes de résistance) et à reconstruire les relations phylogénétiques entre le ou les génome(s) ancien(s) du pathogène et ses homologues modernes (Martin *et al.*, 2013; Yoshida *et al.*, 2013; Yoshida, Sasaki and Kamoun, 2015; Duchêne *et al.*, 2020).
- si la séquence du pathogène historique apporte un signal temporel au sein de l'échantillonnage, une calibration temporelle de la phylogénie par *tip-dating* est applicable et des inférences phylogénétiques plus poussées des paramètres démographiques et évolutifs peuvent être réalisées (Drummond *et al.*, 2002). Ces estimations peuvent alors être corrélées avec les données d'événements historiques et socioéconomiques afin de proposer des *scenarii* de l'histoire évolutive du pathogène.

A ce jour, l'agent pathogène de plantes le plus étudié en faisant usage de spécimens historiques d'herbier est *Phytophthora infestans*, l'oomycète pathogène responsable du mildiou de la pomme de terre (Martin *et al.*, 2013; Yoshida *et al.*, 2013, 2014; Saville, Martin and Ristaino, 2016; Ristaino, 2020). Grâce à l'utilisation d'échantillons infectés d'herbier, le génome de la souche ayant déclenché l'épidémie européenne en 1845 a été identifié et caractérisé, ce qui a permis d'estimer l'évolution du génome et d'identifier une variation de la ploïdie du pathogène au cours du temps (Martin *et al.*, 2013; Yoshida *et al.*, 2013; Yoshida, Sasaki and Kamoun, 2015). Le génome de la souche historique a été placé

phylogénétiquement afin de reconstruire ses relations par rapport aux souches modernes et d'identifier sa lignée évolutive comme désormais disparue et sans homologues modernes (Martin *et al.*, 2013; Yoshida *et al.*, 2013). Cette souche ne serait pas issue de l'aire d'origine du pathogène (Vallée de Toluca au Mexique) mais plutôt d'une zone secondaire de diversification située entre cette aire d'origine et les Etats-Unis d'Amérique d'où un ou quelques événements de dispersion auraient permis l'émergence de *Phytophthora infestans* en Europe (Yoshida *et al.*, 2013; Saville, Martin and Ristaino, 2016). L'utilisation de données génomiques d'échantillons historiques a alors redéfini l'histoire évolutive du pathogène et ses routes d'invasion par rapport aux études menées sur souches modernes uniquement (Goodwin, Cohen and Fry, 1994; Ristaino, Groves and Parra, 2001; May and Ristaino, 2004; Martin *et al.*, 2013; Yoshida *et al.*, 2013; Saville, Martin and Ristaino, 2016) (Figure 1.3.3.A).

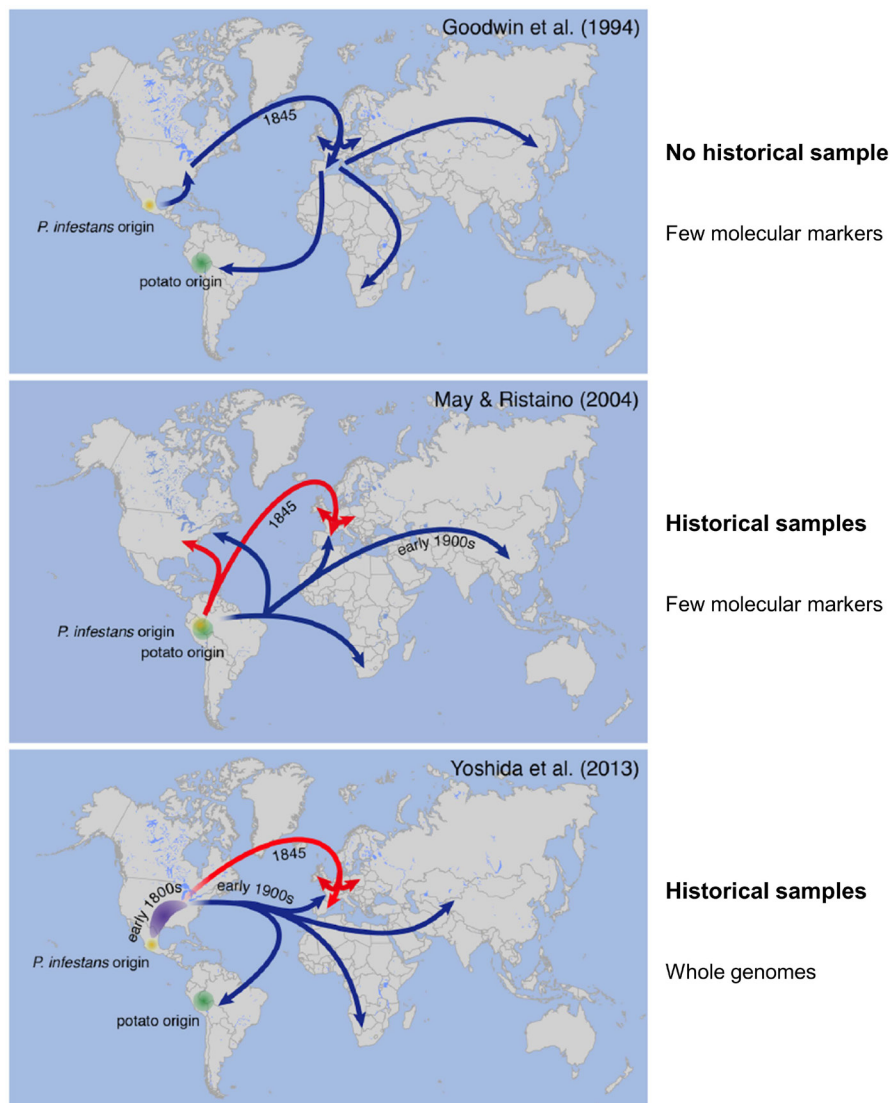


Figure 1.3.3.A. Précision de la reconstruction des routes invasives de *Phytophthora infestans* du XIX^{ème} siècle grâce à la paléogénomique (Yoshida et al., 2014, Figure 2). Modèles de migration successivement proposés (Goodwin, Cohen and Fry, 1994; May and Ristaino, 2004; Yoshida *et al.*, 2013) avec, en rouge, la lignée évolutive responsable du déclenchement de l'épidémie de 1845,

identifiée depuis échantillons d'herbier et, en bleu, celle du XX^{ème} siècle. Le *scenario* évolutif le plus vraisemblable au vu des données disponibles a été proposé suite à l'analyse de génomes historiques complets : l'épidémie européenne du XIX^{ème} siècle aurait été causée par un événement unique de dispersion d'une lignée provenant de la zone secondaire de diversification du pathogène, remplacée au XX^{ème} siècle par une autre lignée issue d'une zone d'origine similaire.

En complément à l'avancée majeure apportée par échantillon d'herbier dans la compréhension des routes d'invasion de *Phytophthora infestans*, d'autres pathogènes non eucaryotes ont également été étudiés depuis échantillons d'herbier, bien que plus anecdotiquement : la bactérie *Xanthomonas citri* pathovar *citri* (Li, Brlansky and Hartung, 2006; Li *et al.*, 2007), ou des virus de l'orge (Malmstrom *et al.*, 2007) ou d'agrumes (Hartung *et al.*, 2015) (Tableau 1.3.3.A).

Tableau 1.3.3.A. Données génomiques d'agents pathogènes anciens issus de spécimens d'herbiers.

	Pathogen	Infectious disease	Method of retrieval	Biological insights	Refs
Eukaryotic pathogens	<i>Phytophthora infestans</i>	Late blight (also known as potato blight)	Shotgun sequencing	<ul style="list-style-type: none"> • First sequenced ancient eukaryotic (plant) pathogen genomes • A unique <i>Phytophthora infestans</i> genotype caused the Irish potato famine and during the 1900s became replaced by the US-1 lineage that dominated worldwide until the 1970s 	(Martin <i>et al.</i> , 2013; Yoshida <i>et al.</i> , 2013)
Bacterial pathogens	<i>Xanthomonas citri</i> pathovar <i>citri</i>	Asiatic citrus canker	PCR-based amplification	<ul style="list-style-type: none"> • Identification of unprecedented genetic diversity in the pathogen with distinct genotype groups in the herbarium specimens • Determination of the origin of the disease emergence in the USA 	(Li, Brlansky and Hartung, 2006; Li <i>et al.</i> , 2007)
Viral pathogens	Barley yellow dwarf virus	Barley yellow dwarf	RT-PCR-based sequencing	<ul style="list-style-type: none"> • Determination of the role played by Barley yellow dwarf viruses in the facilitation of the 18th and 19th century invasion of Eurasian grasses in California • Evidence of virus spread from California to Australia in the late 19th century • Potential correspondence of virus diversification events with the beginning of extensive human exchange between the Old and New worlds 	(Malmstrom <i>et al.</i> , 2007)
	Citrus leprosis virus	Citrus leprosis	Small RNA-based sequencing	<ul style="list-style-type: none"> • Identification of the pathogen responsible for citrus leprosis emergence in Florida between 1911 and 1960s • Revision of citrus leprosis status as either endemic or reemerging disease in Mexico 	(Hartung <i>et al.</i> , 2015)

Depuis l'avènement de l'ADN ancien, les collections naturalistes connaissent un regain d'intérêt scientifique au-delà de la fonctionnalité des herbiers à stocker et cataloguer le vivant. L'utilisation de

ces collections, loin de rester statique, a ainsi évolué grâce à l'accessibilité de nouvelles technologies et leur application originale sur ces spécimens.

Le Muséum national d'Histoire naturelle (MNHN) a pour mission première la gestion et la conservation de ses collections débutées il y a quatre siècles et toujours alimentées par les chercheurs. La valorisation des collections ainsi que la diffusion des connaissances qu'elles apportent font partie intégrante du rôle que doit jouer le MNHN dans le monde de la recherche et auprès du public. Cela a été le projet de Roseli Pellens (UMR Institut de Systématique Evolution Biodiversité (ISyEB), MNHN) dans l'édition d'un ouvrage de vulgarisation des utilisations innovantes des collections naturalistes intitulé « **Les collections naturalistes dans la science du XXI^{ème} siècle** » pour lequel j'ai participé à la rédaction du chapitre « **Les herbiers, une fenêtre ouverte sur l'histoire évolutive des agents pathogènes des cultures** » (Annexe 1.3.3.A.). Nous y faisons état des modalités d'acquisition des données historiques et leur traitement dans un objectif d'étude et de reconstruction de l'histoire évolutive de différents pathogènes de cultures.

1.3.4. Caractéristiques et modalités d'acquisition de matériel génétique ancien depuis échantillons d'herbier

Les herbiers sont des locaux ou bâtiments recueillant des collections naturalistes de spécimens de plantes, algues et champignons, montés sur planche, séchés, en pots ou conservés en liquides (Figure 1.3.4.A)... Ces collections peuvent être organisées de diverses manières mais souvent suivant la classification phylogénique (famille, genre) et la région géographique (continents) puis enfin l'espèce. Cette organisation simplifie la prospection par plantes-hôtes dans la région où la maladie a été décrite puis l'éventuel élargissement à d'autres régions où sa présence est suspectée ou à d'autres espèces du même genre potentiellement hôtes du pathogène. De plus, les récents efforts de numérisation des planches d'herbier, bien que ne représentant qu'une minorité d'échantillons pour le moment (Bakker, 2017), permettent également d'avoir un certain aperçu des collections à distance, facilitant l'accès aux collections.



Figure 1.3.4.A. Salles de collection de l'Herbier national d'Histoire naturelle (à gauche) et du Naturalis Biodiversity Center (à droite). Photographies par L. Gagnevin.

Durant la mise en collection, les spécimens d'herbier subissent un traitement dans le but de leur conservation. Cependant, différents traitements existent : dessiccation à l'air libre, à l'alcool, au gel de silice, par presse, en four... empoisonnement par fumigation, par Kew mixture (solution de mercure, phénol et éthanol), par températures extrêmes... Pour la plupart, ils ont été conçus dans un souci de fixation de la morphologie à des époques où les études moléculaires n'existaient pas encore et certains peuvent accélérer le processus de dégradation *post-mortem* de l'ADN, rendant la majorité de l'ADN contenu en échantillons, même récents, inaccessible à l'amplification (Staats *et al.*, 2011).

Ainsi, suite à la mort de l'individu, les mécanismes cellulaires de réparation de l'ADN sont stoppés, laissant libre cours aux processus de digestion enzymatique, d'oxydation et d'hydrolyse qui induisent la modification et la perte du matériel génétique mais qui peuvent être inhibés dans des conditions de dessiccation rapide des tissus (pour revues, Pääbo *et al.*, 2004; Dabney, Meyer and Pääbo, 2013). Les conséquences de ces dégradations *post-mortem* peuvent être observées sur l'ADN ancien par la faible quantité d'ADN extractible, la réduction de la taille des fragments ADN par fragmentation, la présence de lésions induisant l'incorporation d'un mauvais nucléotide lors de la réplication de l'ADN (Figure 1.3.4.B) ainsi que la présence de lésions empêchant l'amplification de l'ADN par les enzymes polymérase (pour revues, Pääbo *et al.*, 2004; Dabney, Meyer and Pääbo, 2013; Kistler *et al.*, 2020).

La fragmentation de l'ADN est en partie expliquée par la dépurination (spontanée ou due à la chaleur), c'est-à-dire la perte des bases puriques adénine (A) et guanine (G), suivie de l'hydrolyse du squelette en sucre-phosphate, induisant la cassure des brins et l'apparition d'extrémités simple brin (Lindahl and Andersson, 1972). L'effet de la dépurination peut être mesuré par l'excès de bases puriques en amont des cassures (Briggs *et al.*, 2007). Aux extrémités générées par les cassures, la désamination des bases

pyrimidiques cytosines (C) en uraciles (U) est facilitée (Figure 1.3.4.B) (Pääbo *et al.*, 2004; Briggs *et al.*, 2007; Weiß *et al.*, 2016). Lors de l'amplification de l'ADN, les uraciles, lues comme des bases pyrimidiques thymines (T) par les enzymes polymérase tolérantes aux U, induisent l'incorporation d'adénines, ce qui se traduit par des apparentes substitutions C vers T en 5' (G vers A sur le brin complémentaire 3'). Elles représentent les mauvaises incorporations majoritaires chez l'ADN ancien (Briggs *et al.*, 2007). Ces différents patrons de dégradation *post-mortem* de l'ADN peuvent alors être utilisés pour authentifier les ADNs anciens (Rohland *et al.*, 2015; Kistler *et al.*, 2017). D'autres lésions (oxydations, hydrolyse) provoquent des mésincorporations additionnelles ou, de concert avec la formation de liaisons (dites « *cross-links* ») de protéines-ADN ou ADN-ADN, inhibent l'amplification de l'ADN par PCR (Pääbo, 1989; Fulton, 2012).

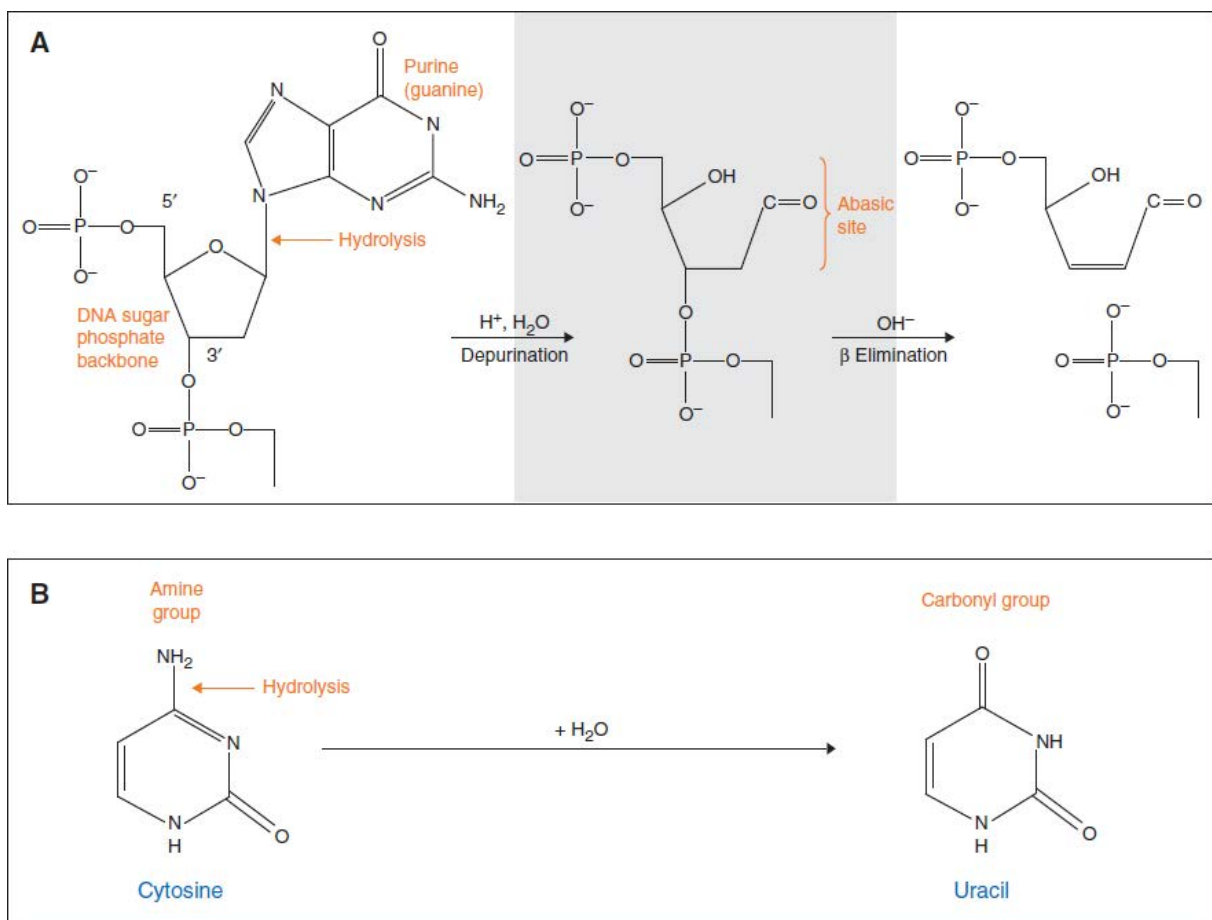


Figure 1.3.4.B. Fragmentation et désamination de l'ADN (Dabney, Meyer and Pääbo, 2013, Figure 1).

A. Une cause probable de fragmentation chez l'ADN ancien est la dépurination par laquelle le lien N-glycosyl entre un sucre et une base adénine (A) ou guanine (G) est rompu, résultant en un site abasique. Le brin d'ADN est alors fragmenté par β élimination, laissant apparaître des extrémités 3'-aldehydique et 5'-phosphate. B. La désamination de cytosine (C) en uracile (U) est le mécanisme principal menant à des lésions de codage erroné des bases de l'ADN ancien. Les ADN polymérase incorporent un A en face d'un U, et un T (thymine) en face d'un A, donnant lieu à d'apparentes substitutions G vers A et C vers T.

L'ensemble de ces caractéristiques des ADNs induit l'application de modalités d'acquisition du matériel génétique adaptées à l'ADN ancien afin de maximiser la quantité d'ADN endogène de petites tailles extraite et son utilisation par des enzymes tolérantes aux lésions tout en conservant des moyens d'authentifier l'ADN face à des contaminations exogènes ou modernes. Celles-ci peuvent avoir lieu durant le stockage, lors de transfert de matériel entre spécimens d'herbier, par exemple, lors de manipulation par le collecteur ou conservateur au cours du temps. Elles peuvent également avoir lieu au laboratoire *via* l'utilisation de réactifs ou de kits contaminés (Glassing *et al.*, 2016). Ces dernières peuvent être mesurées grâce à l'utilisation de témoins lors des manipulations moléculaires.

1.4. Le pathosystème: *Citrus* / *Xanthomonas citri* pathovar *citri*

Les agrumes rassemblent sous cette appellation six genres: *Citrus*, *Clymenia*, *Eremocitrus*, *Fortunella*, *Microcitrus* et *Poncirus*, tous de la famille des *Rutaceae*. Le genre *Citrus* comprend la plupart des agrumes commercialisés, issus d'événements d'hybridation primaire, secondaire, voire plus complexe, de quatre espèces d'origine : *Citrus maxima* (pamplemoussier), *C. medica* (cédratier), *C. micrantha* (papada) et *C. reticulata* (mandarinier) (Curk *et al.*, 2016). Le genre serait originaire d'une zone à l'intersection de l'Asie du sud (nord-est de l'Inde), de l'est (nord-ouest du Yunnan, province du sud de la Chine) et du sud-est (nord du Myanmar) (Wu *et al.*, 2018). Les *Citrus* ont connu une histoire de domestication, de culture et de gestion longue de plus de 4000 ans en Chine et des restes ont été trouvés dès -500 en méditerranée. Connus pour leur propriété anti-oxydante, notamment grâce à la présence d'acide ascorbique, ils sont utilisés par les explorateurs chinois et arabes dès le XII^{ème} siècle mais c'est au XV^{ème} siècle qu'ils sont dispersés plus largement, aux Amériques et en Afrique pour cultivation (pour revue, Talon, Caruso and Gmitter Jr., 2020). L'agrumiculture à grande échelle se déploie au XIX^{ème} et est aujourd'hui la première production fruitière au monde avec près de 150 millions de tonnes produites en 2019 sur plus de 140 pays (données FAO <http://www.fao.org/faostat>). Elle est cependant menacée par de nombreuses maladies, comme le chancre asiatique des agrumes, l'une des pathologies majeures des agrumes avec le huanglongbing (maladie bactérienne causée par *Candidatus Liberibacter* spp.) et la tristeza des agrumes (maladie virale causée par *Citrus tristeza virus*).

1.4.1. Chancre asiatique des agrumes

Le chancre asiatique des agrumes (CAA) touche les espèces du genre *Citrus*. La maladie consiste en l'apparition de lésions éruptives chancreuses sur les parties aériennes de la plante, d'aspect liégeux avec bordures imbibées d'eau ou huileuses, entourées d'un halo chlorotique (Figure 1.4.1.A) qui peuvent causer une défoliation ainsi qu'une chute précoce des fruits. Au niveau cellulaire, les symptômes se présentent sous forme d'une hypertrophie (augmentation du volume) ainsi qu'une hyperplasie (augmentation du nombre) cellulaire (Figure 1.4.1.B).



Figure 1.4.1.A. Symptômes du chancre asiatique des agrumes sur fruit (gauche) et feuille (droite). Photographies par C. Vernière (gauche) et D. Richard (droite).



Figure 1.4.1.B. Coupes histologiques d'une feuille au niveau d'une lésion de chancre asiatique des agrumes (gauche) et d'une feuille saine (droite) (Lin, Hsu and Tzeng, 2009, Figure 2). La feuille saine a été inoculée avec de l'eau distillée pour contrôle. La barre d'échelle représente 125 μm .

Les plus vieilles descriptions au champ du CAA datent de 1899 au Japon (Kuhara, 1978) mais l'identification de symptômes typiques de la maladie sur des échantillons d'herbier prélevés en 1842-1844 à Java et en 1827-1831 au nord-ouest de l'Inde (Fawcett and Jenkins, 1933) constitue son signalement le plus ancien. Pour ces raisons, la maladie a été supposée être d'origine d'asiatique et elle ne sera signalée hors de sa zone d'origine qu'à partir du début du XX^{ème} siècle : en 1911 aux Etats-Unis d'Amérique, en 1912 en Australie et en 1916 en Afrique du Sud (Skaria and da Graça, 2012). Elle est aujourd'hui largement distribuée dans l'ensemble des pays producteurs d'agrumes des zones tropicales et sub-tropicales (Graham *et al.*, 2004) (Figure 1.4.1.C).

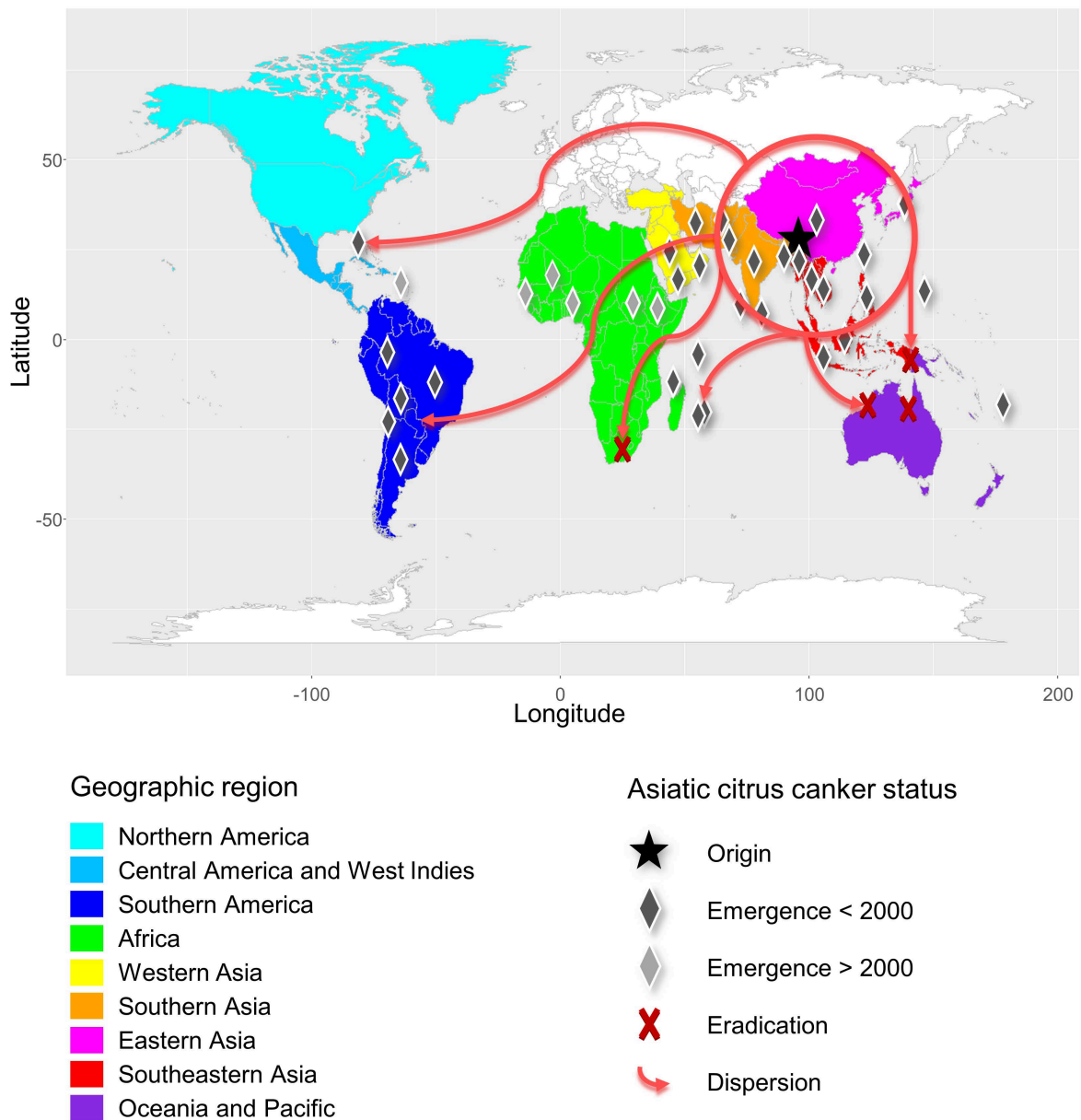


Figure 1.4.1.C. Distribution géographique et émergences du chancre asiatique des agrumes.

La maladie entraîne une diminution de la qualité des fruits et de l'état global des arbres et généralement une baisse de rendement de l'agrumiculture. De plus, par le statut d'organisme de quarantaine de l'agent causal, l'exportation est soumise à restriction et, dans les pays où la maladie est encore absente comme ceux du bassin méditerranéen, d'Afrique australe et les Etats-Unis d'Amérique et le Mexique, la gestion avec la mise en place de moyens de surveillance et d'éradication est coûteuse. Enfin, l'utilisation massive d'intrants et de pesticides souvent mise en place dans les stratégies de lutte est néfaste pour l'environnement, la santé publique et la biodiversité.

1.4.2. *Xanthomonas citri* pv. *citri*

Xanthomonas citri pathovar *citri* (Hasse, 1915) Gabriel et al., 1989 (ci-après *Xci*), identifiée comme l'agent responsable du chancre asiatique des agrumes en 1915 par Hasse (1915) est une γ -protéobactérie monomorphe à multiplication clonale (Li et al., 2007) prenant la forme d'un bâtonnet de 1,5-2,0 sur 0,5-0,8 μm . La bactérie fait partie de la famille des *Xanthomonadaceae* au sein de l'ordre des *Xanthomonadales*. Le pathovar (pv.) est un niveau taxonomique infrasubspécifique regroupant les bactéries pathogènes d'une même espèce présentant la même symptomatologie et la même gamme d'hôtes (Dye et al., 1980). À l'intérieur du pathovar, trois pathotypes (groupement d'organismes causant une même maladie partageant des caractéristiques communes de pouvoir pathogène (marqueurs moléculaires, symptômes, hôtes), n'est pas un niveau taxonomique) ont été décrits : le pathotype A regroupe des souches dont la gamme d'hôtes est la plus large attaque la plupart des *Citrus* ; le pathotype A* et le pathotype A^W qui touchent tous les deux seulement deux espèces de *Citrus* (*Citrus aurantiifolia* et *Citrus macrophylla*) mais diffèrent par la capacité du pathotype A^W à induire une réaction d'hypersensibilité chez certains agrumes non-hôtes (Vernière et al., 1998; Rybak et al., 2009). De manière intéressante, les trois pathotypes forment chacun un groupe monophylétique ou clade (Gordon et al., 2015; Zhang et al., 2015; Patané et al., 2019). Le pathotype A est trouvé à l'échelle mondiale alors que le pathotype A* n'est trouvé que dans quelques pays d'Asie de l'ouest, du sud et du sud-est plus les Fidji (Vernière et al., 1998) et en Afrique de l'est ; le pathotype A^W, quant à lui, est restreint aux Etats-Unis d'Amérique (Schubert et al., 2001), à l'Inde (Bui Thi Ngoc et al., 2009) et à Oman. *Xci* présente un chromosome circulaire de 5,18 Mb (séquence de la souche de référence IAPAR 306 *Xanthomonas citri* pv. *citri* (pathotype A), NC_003919.1 (da Silva et al., 2002)) et renferme quasi-systématiquement un ou plusieurs plasmides. De récents travaux ont révélé une grande plasticité plasmidique entre les différentes souches (Gochez et al., 2018; Richard et al., 2020).

1.4.3. Cycle biologique de *Xanthomonas citri* pv. *citri* et interactions avec la plante

Xanthomonas citri pv. *citri* (*Xci*), arrivé sur la plante, mobilise des facteurs de la virulence et des gènes impliqués dans la pathogénie. La bactérie doit d'abord assurer sa fixation à la surface de la plante puis son déplacement, régulés par le système chimiosensitif (Mhedbi-Hajri et al., 2011). La bactérie sécrète des protéines et polysaccharides, ce qui lui permet de constituer un biofilm de protection qui l'aide également à se déplacer (Rigano et al., 2007). Elle se déplace suite à détection des signaux environnementaux par chimiotaxie (vers zones riches en molécules bénéfiques et pauvres en molécules néfastes) à l'aide de son flagelle. *Xci* pénètre alors la plante par des ouvertures naturelles comme les stomates ou des blessures causées par les épines de l'arbre, le vent, les pratiques agricoles ou encore les insectes tels la mineuse asiatique des agrumes. Elle active alors des gènes du pouvoir pathogène afin de manipuler les cellules de la plante. Si la plante est sensible, c'est-à-dire, s'il y a

interaction moléculaire entre elle et la bactérie, elle est alors considérée comme hôte. Cette interaction se réalise entre le système immunitaire de la plante et principalement grâce au système de sécrétion de type III permettant à la bactérie d'injecter directement dans les cellules végétales des effecteurs (Ryan *et al.*, 2011). Vingt-six effecteurs de type III (T3E) sont trouvés systématiquement chez *Xci* parmi les 66 identifiés chez les *Xanthomonadales*, la présence de trois autres (*xopAF*, *xopC1* et *xopAG*, ce dernier étant responsable du déclenchement de la réaction hypersensible sur *Citrus paradisi* par les souches du pathotype A^w (Sun *et al.*, 2004)) est variable au sein des clades des pathotypes A* et A^w (Escalon *et al.*, 2013; Jalan *et al.*, 2013). Parmi les T3E, les effecteurs TAL (*Transcription Activator-Like Effectors*, TALE) sont une famille particulière et jouent le rôle d'activateurs transcriptionnels eucaryotes capables d'interaction directe avec l'ADN de la plante hôte suite à la reconnaissance de motifs nucléotidiques au niveau de promoteurs de gènes de sensibilité et/ou de résistance (Boch and Bonas, 2010). Chez *Xci*, *pthA4* est le gène de TALE indispensable à l'induction de symptômes (Duan *et al.*, 1999; Domingues *et al.*, 2010). Il compte trois homologues, *pthA1*, *pthA2* et *pthA3*, les quatre gènes étant situés sur deux plasmides (chez la souche de référence IAPAR 306, pXAC33 NC_003921.3 et pXAC64 NC_003922.1 (da Silva *et al.*, 2002)). Les TALE activant des gènes de sensibilité en se liant à leur région promotrice (chez les *Citrus*, *LOB1* (Hu *et al.*, 2014), impliqué dans l'expansion cellulaire) induisent alors des conditions favorables à la colonisation. La bactérie colonise ainsi l'apoplaste (espace intercellulaire des tissus impliqué dans le transport de l'eau) jusqu'à rupture de l'épiderme sous l'effet de l'hyperplasie et l'hypertrophie des cellules de la plante que le pathogène induit (Pruvost *et al.*, 2002). Cela contribue au relargage de bactéries à la surface de la plante qui se propagent par contact de l'eau de pluie sur les lésions de chancre, à petite échelle par éclaboussures et à grande échelle par événements météorologiques importants ou par déplacement de fruits, plants et greffons et matériels contaminés par l'Homme (Schubert *et al.*, 2001; Gottwald, Graham and Schubert, 2002; Pruvost *et al.*, 2002; Graham *et al.*, 2004) (Figure 1.4.3.A).

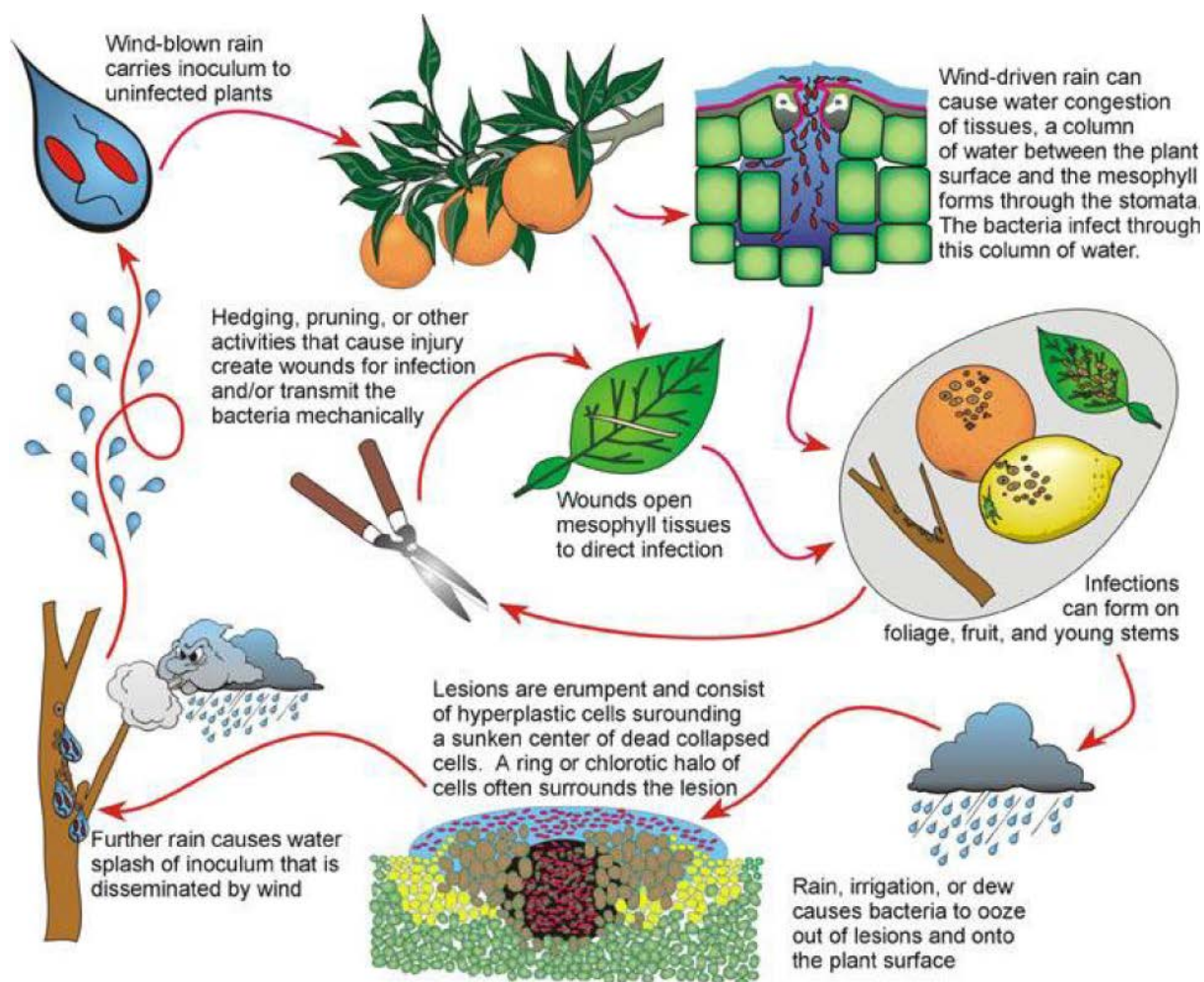


Figure 1.4.3.A. Cycle du chancre asiatique des agrumes (Schubert *et al.*, 2001, Figure 3).

1.5. Ressources et état des connaissances avant le début de la thèse

Les premières analyses génétiques cherchant à reconstruire les relations de parentés entre souches de *Xci* ont été effectuées sur des marqueurs moléculaires de type éléments insertionnels (Li *et al.*, 2007) et minisatellites (unité de répétition de motif nucléique de plus de 8 nucléotides sur une longueur de séquence de 0,1 à 20 kilobases) (Pruvost *et al.*, 2014). Ces travaux ont notamment permis de déterminer l'appartenance des trois pathotypes A, A^W et A* à trois groupes génétiquement distincts, avec une diversité génétique supérieure des populations asiatiques en accord avec l'hypothèse d'origine du pathogène basée sur les premières descriptions de la maladie. Cependant, les marqueurs utilisés n'étant pas assez résolutifs dans la reconstruction des relations de parenté, plusieurs études ont ensuite cherché à améliorer le signal phylogénétique *via* le séquençage du génome complet de nombreuses souches provenant d'origines géographiques diverses. L'analyse comparative de ces génomes a notamment permis de mieux décrire l'histoire évolutive de *Xci* et comprendre l'importance relative des différentes forces évolutives (Gordon *et al.*, 2015; Zhang *et al.*,

2015; Patané *et al.*, 2019; Richard *et al.*, 2020). Néanmoins, ces différents jeux de données composés de génomes séquencés à partir de souches bactériennes échantillonnées dans les années 1970/1980 pour les plus anciennes ne présentent pas de signal temporel, ce qui ne permet pas une datation précise des événements évolutifs de *Xci*. Afin de pallier cette lacune, et inspiré par les travaux menés sur *Phytophthora infestans* (pages 30-32), les collections naturalistes de plusieurs herbiers ont été prospectées entre 2016 et 2018 afin d'identifier et collecter des spécimens historiques de *Citrus* potentiellement infectés par le chancre asiatique des agrumes. Une approche d'échantillonnage systématique en présence de symptômes comparables aux lésions typiques causées par cette maladie a été adoptée. Il est cependant à noter que la dessiccation des tissus peut induire une perte de couleurs ou d'aspect, notamment pour le halo chlorotique ou les bordures imbibées d'eau ou huileuses qui deviennent alors plus difficilement observables autour des lésions chancreuses (Figure 1.5.A).



Figure 1.5.A. Symptômes du chancre asiatique des agrumes sur spécimens d'herbiers. Photographies par L. Gagnevin (gauche, milieu) et A. Rieux (droite).

Ainsi, depuis 2016, plus de 300 échantillons de feuilles, fruits et branches potentiellement infectés de *Citrus* et de *Rutaceae* proches ont été collectés depuis une dizaine d'herbiers dans le monde:

- quatre en Europe, l'herbier du Muséum national d'Histoire naturelle (Paris, France), du *Natural History Museum* (Londres, Royaume-Uni), celui des *Royal Botanic Gardens* (Kew, Royaume-Uni) et du *Naturalis Biodiversity Center* (Leiden, Pays-Bas);
- deux en Amérique du nord, l'*U.S. National Herbarium* (Washington, Etats-Unis d'Amérique) et le *Steere Herbarium, New York Botanical Garden* (New York, Etats-Unis d'Amérique);
- un en Amérique du sud, le *National Colombian herbarium* (Bogotá, Colombie);
- un en Afrique, le *Bolus herbarium library* (Cape Town, Afrique du sud);

- trois dans l'océan indien, l'herbier de La Réunion (Saint-Denis, France), le *Mauritius herbarium* (Réduit, Maurice) et l'*Antananarivo herbarium* (Antananarivo, Madagascar)
- deux en Asie, l'herbier de Bangkok (Bangkok, Thaïlande) et le *Gedung herbarium Bogoriense* (Bogor, Indonésie).

A ces herbiers s'ajoute l'*U.S. National Fungus Collection* du département de l'agriculture des Etats-Unis d'Amérique (USDA), qui comprend des spécimens de végétaux infectés interceptés et mis en collection par cet organisme.

Le prélèvement de matériel pour chacun des spécimens a été variable et décidé par les curateurs (quelques centimètres carrés à quelques feuilles entières). Les spécimens historiques échantillonnés, bien que majoritairement originaires du continent asiatique couvrent également d'autres zones géographiques et datent de 1820 à 2010.

À cette collection d'échantillons historiques s'ajoute une vaste collection au laboratoire d'accueil de plus de 2000 souches de *Xci* lyophilisées datant de 1948 à 2017 et couvrant l'ensemble des zones où la maladie est présente ou a été éradiquée. Au total, 415 génomes de souches contemporaines de *Xci* ont été générés et étaient disponibles avant le début de la thèse. Ces génomes (incluant ceux issus des souches datant de la deuxième moitié du XX^{ème} siècle) seront qualifiés de « modernes » dans la suite de ce document, l'extraction ADN ayant été réalisée sur des cultures fraîches de la bactérie réalisées à partir des souches lyophilisées.

1.6. Annexes

Annexe 1.3.3.A. Les herbiers, une fenêtre ouverte sur l'histoire évolutive des agents pathogènes des cultures

SCIENCES

BIOLOGIE

Systematique, phylogénomique et taxonomie

Les collections naturalistes dans la science du XXI^e siècle

*une ressource durable
pour la science ouverte*

sous la direction de
Roseli Pellens

ISTE
editions

13

Les herbiers, une fenêtre ouverte sur l'histoire évolutive des agents pathogènes des cultures

**Lionel GAGNEVIN¹, Adrien RIEUX², Jean-Michel LETT²,
Philippe ROUMAGNAC¹, Boris SZUREK¹, Paola CAMPOS²,
Claudia BAIDER³, Myriam GAUDEUL⁴ et Nathalie BECKER⁴**

¹ PHIM Plant Health Institute, CIRAD, IRD, INRAE,
Institut Agro, Université de Montpellier, Montpellier, France

² UMR PVBMT, CIRAD, INRAE, Université de La Réunion,
Saint-Pierre, France

³ The Mauritius Herbarium, Agricultural Services,
Ministry of Agro-Industry and Food Security, Port-Louis, Maurice

⁴ Institut de Systématique, Évolution, Biodiversité (ISYEB), CNRS,
Muséum national d'Histoire naturelle, Sorbonne Université, EPHE,
Université des Antilles, Paris, France

Botanists have generally neglected their cultivated varieties, as beneath their notice.

Charles Darwin,

The Variation of Animals and Plants under Domestication, 1868

Les micro-organismes pathogènes de plantes cultivées peuvent être responsables d'épidémies à fort impact socio-économique. Les herbiers des collections d'histoire

Les collections naturalistes dans la science du XX^e siècle,
coordonné par Roseli PELLENS. © ISTE Editions 2021.

naturelle constituent des sources d'informations pour l'étude de ces épidémies : les 200 dernières années sont couvertes, incluant la dernière vague des grands changements agricoles et complétant les données des échantillons modernes. Certains échantillons anciens constituent ainsi une « preuve directe » de la présence d'une maladie sur une plante hôte, à une date et localité données, ce qui peut fournir des informations sur l'origine du pathogène et ses routes d'invasion. Les développements méthodologiques actuels (séquençage à haut débit et épidémiologie moléculaire), appliqués à ces échantillons historiques, permettent d'obtenir des datations de temps de divergence avec une précision accrue, mais aussi d'élucider les mécanismes évolutifs microbiens associés aux dynamiques épidémiologiques.

L'exploration d'un herbier prend en compte la date et le lieu de prélèvement de l'échantillon, et l'existence de symptômes visibles. L'analyse de séquences exploitables à partir du prélèvement, suite à l'extraction d'acides nucléiques (ADN ou siRNA), est soumise à des contraintes spécifiques. Des processus de dégradation d'une part, et évolutifs d'autre part, sont caractérisés. Nos recherches s'appuient sur des modèles d'importance économique dans la zone de l'océan Indien ou de l'Afrique subsaharienne (*Xanthomonas citri* pv. *citri*, pour les agrumes ; géminivirus, pour le manioc et la patate douce) et sur la complémentarité des herbiers mondiaux. Ces résultats soulignent l'importance des collections d'herbiers pour une meilleure connaissance, et une meilleure gestion, de l'émergence et de l'histoire évolutive de phytopathogènes bactériens et viraux.

13.1. Épidémies, émergences et re-émergences

La sécurité alimentaire est un enjeu sociétal important et les pertes causées par les maladies et les ravageurs des cultures peuvent avoir des conséquences dramatiques pour les populations et les économies, avec des pertes cumulées pouvant dépasser 40 % des récoltes (Anderson *et al.* 2004 ; Savary *et al.* 2019). Les maladies des cultures sont causées par des agents pathogènes¹ le plus souvent microbiens pouvant être responsables d'épidémies récurrentes contre lesquelles les agriculteurs luttent tant bien que mal à l'aide de mesures de quarantaine, de produits phytopharmaceutiques, de variétés résistantes ou de pratiques culturales adaptées (figure 13.1). Dans ce contexte, l'utilisation excessive de produits phytopharmaceutiques représente une menace pour la biodiversité, la qualité de l'environnement ainsi que pour la santé. Une maladie peut par ailleurs avoir récemment émergé, c'est-à-dire être apparue sur une plante ou dans une région où elle n'avait jamais été décrite ou avoir re-émergé dans une région dont elle avait disparu. Les cas d'émergence sont délicats car on est en général particulièrement désarmé

1. Virus, bactéries, champignons, oomycètes et nématodes responsables de maladies (les ravageurs sont plutôt des organismes supérieurs, comme les arthropodes qui causent des dégâts en s'alimentant).

pour proposer rapidement des solutions de lutte (Almeida *et al.* 2019), notamment du fait qu'il est difficile de comprendre les mécanismes ayant permis cette émergence. Il peut s'agir de facteurs extrinsèques (changement climatique ou de conditions culturales, introduction d'inoculum, apparition d'un vecteur, etc.) ou intrinsèques au micro-organisme (adaptation aux conditions climatiques, changement d'hôte, acquisition de facteurs de virulence par mutation ou échanges de matériel génétique, contournement de résistance, etc.) (Anderson *et al.* 2004).



Figure 13.1. Dégâts de bactériose vasculaire du manioc en Colombie, causés par la bactérie *Xanthomonas phaseoli* pv. *manihotis*

COMMENTAIRE SUR LA FIGURE 13.1.– Cette bactérie entre au niveau des feuilles et se propage de façon systémique en envahissant les vaisseaux de la plante où elle bloque le transport d'eau et de nutriments, conduisant à un flétrissement généralisé. Elle se propage de façon aérienne (combinaison pluie et vent) ou lors du bouturage. Photo : C. Trujillo.

13.2. Développement de l'agriculture, domestication des plantes cultivées et de leurs maladies

Les théories de la domestication des plantes pour l'agriculture émettent en général l'hypothèse que celle-ci a permis l'émergence initiale d'agents pathogènes spécialisés causant des maladies ayant un impact socio-économique important. Si les agents pathogènes microbiens existent depuis longtemps (Achtman 2016), leur émergence et l'apparition d'épidémies serait contemporaine de la révolution néolithique² (Mira *et al.* 2006).

2. Période de l'histoire de l'humanité au cours de laquelle se sont développés l'agriculture et l'élevage ainsi que les premières sociétés organisées, il y a environ 10 000 ans.

Ceci serait dû au fait que la domestication végétale implique une diminution de la diversité génétique des plantes, un coût génétique, une augmentation de la densité des génotypes, des conditions de culture plutôt favorables, une augmentation des échanges de matériel végétal ou tout simplement du transport passif (par l'irrigation par exemple), la proximité avec des plantes d'une aire d'origine différente. Ces mécanismes constituent un ensemble de paramètres qui pourraient favoriser des micro-organismes non généralistes, à forte virulence, à reproduction plutôt asexuée mais aussi les mécanismes de transmission d'inoculum, les transferts horizontaux de matériel génétique, voire l'hybridation (Stukenbrock et McDonald 2008 ; Stukenbrock et Bataillon 2012). Si c'est au Néolithique que l'agriculture s'est développée avec la domestication (en général locale) de nombreuses plantes et de leurs micro-organismes associés (Larson *et al.* 2014 ; Zeder 2017), d'autres périodes plus récentes ont vu des changements importants dans l'agriculture, comme les XVIII^e et XIX^e siècles avec le développement d'une agriculture coloniale de rente, des échanges intercontinentaux fréquents de matériel végétal, les débuts de la sélection variétale consciente, ou encore le développement des techniques de greffage. Au XX^e siècle, avant puis pendant la révolution verte, la sélection (contre les maladies ou pour des caractéristiques agronomiques ou alimentaires) devient un outil majeur accompagné de changements agronomiques importants (engrais chimiques, produits phytosanitaires de synthèse, industrie semencière, mécanisation, augmentation des surfaces monocultivées), mais aussi de changements climatiques et d'une augmentation des échanges entre les différentes régions du monde. Toutes ces caractéristiques vont soit favoriser les épidémies dans l'absolu (par exemple en étendant l'aire de présence d'un micro-organisme), soit exercer sur les micro-organismes des pressions de sélection directionnelles ou non qui vont faire s'effondrer les moyens de lutte classiques (cycles *boom and bust*)³, soit encore donner l'opportunité à de nouvelles combinaisons hôte-agent pathogène de se réaliser (*host jumps* ou *host shifts*). On parle couramment de co-évolution entre les plantes et leurs parasites mais l'évolution des plantes cultivées est en fait artificielle, humaine et rapide, tandis que les micro-organismes possèdent un potentiel adaptatif encore plus important (McDonald et Linde 2002).

Si la révolution agricole néolithique peut être explorée par l'archéologie (et aujourd'hui par des approches génétiques), les travaux des naturalistes explorateurs entrepris à partir du XVIII^e siècle ont permis de cataloguer la diversité végétale planétaire avec, entre autres, l'objectif d'aller vers une meilleure agriculture. Les herbiers qu'ils ont constitués sont autant de témoignages de la diversité végétale et microbienne à un temps donné, sur une plante donnée, dans une région donnée, et qui jusqu'à récemment

3. Cette expression décrit le processus d'expansion d'une monoculture (« boom ») possédant par exemple une résistance à un agent pathogène, suivi de l'effondrement de cette résistance (« bust ») du fait de son contournement par l'agent pathogène ou de l'arrivée de nouvelles populations pathogènes non sensibles à cette résistance.

étaient les seules données exploitables de ces échantillons morts. Plus tard, l'intensification de la fréquence et de l'ampleur des épidémies, comme celle du mildiou de la pomme de terre au milieu du XIX^e siècle ou les premières épidémies de chancre bactérien des agrumes vers 1915 aux États-Unis ont orienté le développement de la phytopathologie et pour la première fois la mise en place de collectes rationalisées d'échantillons symptomatiques (figure 13.2) (Peltier et Frederich 1924 ; Hasse 2015).

13.3. La biologie moléculaire et la génomique comme un outil d'étude des micro-organismes phytopathogènes

De nos jours, l'étude des agents pathogènes des cultures et des maladies qu'ils causent bénéficie grandement de l'apport de la biologie moléculaire et de la génétique des populations (Stukenbrock et McDonald 2008 ; Goss 2015). Depuis près de 30 ans, les chercheurs ont développé des méthodes pour extraire, amplifier, détecter, typer et séquencer les acides nucléiques des agents phytopathogènes à partir d'échantillons prélevés sur des plantes infectées au champ. L'analyse de la diversité et de la structuration génétique des populations est une source d'informations directe pour la lutte contre les maladies : elle permet par exemple de connaître la diversité globale du pathogène pour adapter les déploiements de résistances et anticiper les contournements, ou encore de déchiffrer les mécanismes de pathogénie pour choisir des résistances durables.

Elle permet aussi dans certains cas l'identification des populations sources, des voies de dissémination ou encore l'estimation de paramètres biotiques et abiotiques associés à un agent phytopathogène. Plus récemment, l'essor des méthodes de séquençage à haut débit (séquençage parallèle massif, aussi appelé *High-Throughput Sequencing*, HTS) combinées à de nombreux développements méthodologiques dans les domaines de l'épidémiologie moléculaire ont permis de reconstruire l'histoire de l'émergence et de l'évolution des agents pathogènes dans le temps et dans l'espace avec une précision accrue (Croucher et Didelot 2015). Cependant, la grande majorité des études réalisées à ce jour n'a été généralement basée que sur l'analyse d'échantillons contemporains prélevés durant les 40 dernières années. Bien que potentiellement très informative, l'analyse de tels échantillons peut dans certaines situations montrer ses limites, notamment en termes de résolution temporelle.

13.4. Apports des échantillons d'herbier

Comme nous l'avons dit plus haut, les herbiers des collections d'histoire naturelle constituent des sources très précieuses de matériel biologique historique qui vont compléter les données issues d'échantillons modernes. Plus généralement, les collections d'échantillons biologiques des musées deviennent une source de plus en plus exploitée pour reconstituer l'histoire évolutive des êtres vivants. Bien que moins âgés que des

échantillons « anciens » au sens strict (issus de fouilles archéologiques ou paléontologiques par exemple), de tels spécimens possèdent des caractéristiques particulières et des limitations dont il faut tenir compte, voire qu'il est nécessaire de corriger (représentativité, dégradation, contamination, métadonnées incomplètes), mais sont également en général bien classifiés et caractérisés (Wandeler *et al.* 2007). Ils permettent de couvrir la période des 200 dernières années qui incluent la dernière vague des grands changements agricoles mondiaux.



Figure 13.2. Planche d'herbier avec symptômes : a) collecté par un botaniste en 1860 ; b) collecté par un phytopathologiste en 1917

COMMENTAIRE SUR LA FIGURE 13.2. a) Planche L.2019517 de *Citrus* symptomatique préservée à l'Herbier du Naturalis Biodiversity Center (Leiden, Pays-Bas), les symptômes potentiels sont indiqués par des pointes de flèches. b) Échantillon symptomatique de *Citrus* conservé par le service de quarantaine de l'USDA (Beltsville, MD, États-Unis). Échantillon identifié comme symptomatique (taches brunes avec un halo chlorotique) dès la récolte par Mme Clara Hesse. Photos : Naturalis Biodiversity Center et L. Gagnevin.

Les collections d'herbiers représentent une source d'information permettant d'améliorer notre compréhension des mécanismes d'émergence et d'évolution des agents pathogènes des cultures (Yoshida *et al.* 2014) grâce aux éléments suivants.

13.4.1. Des preuves directes

Tout d'abord, un spécimen d'herbier arborant des symptômes pathologiques caractéristiques d'une maladie représente une « preuve directe » de la présence de cette maladie sur une plante hôte, à une date et à une localité données. Une telle information permet de compléter les descriptions historiques des agents phytopathogènes et parfois de les affiner. Ainsi, Fawcett et Jenkins en 1933 (Fawcett et Jenkins 1933) ont fait usage d'échantillons de *Citrus* conservés dans des herbiers britanniques et américains pour dater des années 1830 les premiers échantillons de l'espèce infestés par la bactérie *Xanthomonas citri* pv. *citri* (Xci) et supposer une origine indienne ou indonésienne de la maladie du chancre des agrumes. Plus récemment, en étudiant des échantillons d'espèces de Silène (famille des Caryophyllacées) infectés par la maladie du charbon des anthères de fleurs (causée par le basidiomycète *Microbotryum violaceum*) au sein d'échantillons d'herbiers américains, des chercheurs ont réussi à reconstruire la distribution spatiotemporelle de cette maladie dans l'est des États-Unis ainsi qu'à mettre en évidence des possibles changements d'hôte entre plusieurs espèces végétales (Antonovics *et al.* 2003). De manière plus large, les échantillons d'herbiers même récents peuvent permettre d'obtenir des informations (et des génomes) de pathogènes dans des zones ou sur des hôtes peu ou pas accessibles sur le terrain (zones difficiles d'accès, hôtes ayant disparu, etc.).

13.4.2. Des analyses moléculaires

Il est possible d'extraire, amplifier et séquencer des acides nucléiques provenant d'agents pathogènes qui infectaient les plantes lors de leur échantillonnage et mise en collection. De telles données génétiques permettent tout d'abord, *via* une approche de « génomique comparative », de déterminer les différences d'architecture génétique (par exemple variation du contenu en gènes, présence d'insertions/délétions de fragments de séquences nucléotidiques, mutations ponctuelles, génomes accessoires, ploïdie) existant entre souches modernes et historiques et pouvant potentiellement être associées à l'émergence des agents phytopathogènes. Ainsi, dans une étude se focalisant sur l'oomyète⁴ *Phytophthora infestans* parasite de la pomme de terre (*Solanum tuberosum* L.) et responsable de la grande famine irlandaise en 1845, l'analyse d'échantillons d'herbiers a permis de mettre en évidence une augmentation de la ploïdie chez les souches modernes de l'agent pathogène (Yoshida *et al.* 2013). Un tel mécanisme évolutif est avantageux pour les espèces asexuées comme *P. infestans* puisqu'il permet d'augmenter la diversité génétique et peut induire l'assortiment de gènes ou d'allèles adaptatifs favorisant l'émergence et la virulence d'un clone pathogène.

4. Les oomyètes sont une classe de microorganismes eucaryotes ressemblant à des champignons (et longtemps classifiés avec eux) mais phylogénétiquement proches des plantes et algues brunes.

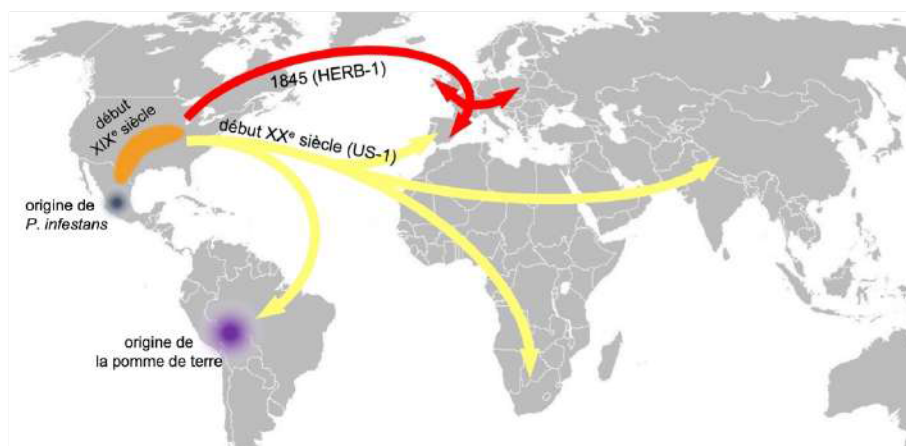


Figure 13.3. Exemple de reconstitution des routes d'invasion de *Phytophthora infestans* à partir de données de génomique issues d'échantillons d'herbier (d'après (Yoshida et al. 2013))

Les données génétiques historiques d'agents phytopathogènes obtenues à partir de spécimens d'herbiers offrent également une formidable opportunité pour réaliser des inférences phylogénétiques. En effet, pour établir de telles inférences, il est nécessaire d'avoir un « signal temporel », c'est-à-dire de pouvoir mesurer l'accumulation progressive des mutations dans les génomes des échantillons obtenus à différentes dates. Plus la fenêtre de temps entre les individus anciens et modernes est grande, plus la détection du signal temporel est favorisée et l'inférence phylogénétique de paramètres réalisable. Typiquement, les échantillons d'herbiers peuvent être utilisés comme des points de calibration pour conjointement estimer taux de substitution (la vitesse à laquelle les mutations s'accumulent dans les génomes) et temps de divergence au sein d'un arbre phylogénétique (Rieux et Balloux 2016).

Les datations de divergence génétique peuvent ensuite être comparées avec des informations historiques pour tenter d'élucider les causes associées aux dynamiques épidémiologiques. Ainsi, l'ancêtre commun des souches de *P. infestans* a été daté du début du XVI^e siècle, ce qui coïncide avec la période de la conquête espagnole du Mexique. Cependant, la pandémie qui a démarré en 1845 et les épidémies subséquentes sont soit le fait d'introductions-remplacements successifs par des populations de *P. infestans* contournant les variétés résistantes déployées pour contrôler la maladie, soit le résultat d'une diversification et adaptation rapides aux nouvelles variétés à partir d'un fonds génétique limité. En revanche les populations modernes seraient issues d'une introduction ultérieure (figure 13.3) mais ont également un potentiel évolutif important favorisé par la polyploïdie, une plasticité génétique accrue et des pressions de sélection intenses (Martin et al. 2013).

Finalement, les données génétiques obtenues à partir de spécimens d'herbiers permettent également d'accéder à des séquences provenant de la plante et de ce fait d'étudier spécifiquement les interactions hôtes-pathogènes. En conservant l'exemple emblématique des travaux menés sur *P. infestans*, les chercheurs ont ainsi pu caractériser la diversité des gènes de résistance présents chez les spécimens historiques de pomme de terre et montrer que la plupart d'entre eux ont été introgressés⁵ dans les variétés cultivées à partir d'espèces sauvages au début du XX^e siècle (Yoshida *et al.* 2013).

13.5. Comment explorer un herbier

Tels qu'ils se présentent en ce début du XXI^e siècle, les herbiers de muséums d'histoire naturelle ou de jardins botaniques sont des bâtiments où les planches botaniques sont le plus souvent rangées par famille botanique, genre, région géographique, puis espèce (figure 13.4). Les botanistes n'ont pas exclusivement choisi des échantillons sains et « beaux », mais des échantillons représentatifs de la réalité de terrain, y compris lorsqu'ils sont infectés ou en partie détériorés par des ravageurs. Par ailleurs un grand nombre de spécimens sont des plantes cultivées ou apparentées à des plantes cultivées, en relation avec l'objectif initial d'inventaire et de valorisation agricole. Le mode de rangement famille-genre-région-espèce est bien adapté à nos objectifs puisqu'il permet d'observer tous les spécimens d'une région (par exemple où la maladie a été décrite), puis rapidement d'élargir à d'autres régions et aux autres espèces du genre, ce qui est une source d'information pour étudier la gamme d'hôtes d'un agent pathogène et potentiellement reconstruire ses sauts d'hôte.

On peut distinguer grossièrement deux façons d'échantillonner : l'une « à l'aveugle », pour les maladies dont les symptômes ne sont pas visibles lorsque les plantes sont séchées, en orientant l'exploration à l'aide de critères de lieu, de date et d'espèce végétale. La seconde consiste simplement à rechercher des symptômes caractéristiques de la maladie, ce qui peut même être fait ante-prospection sur les photographies numérisées de plus en plus disponibles dans les bases de données des muséums. À titre d'exemple, le New York Botanical Garden Steere Herbarium contient 7 600 spécimens de rutacées, dont 330 spécimens du genre *Citrus*. Les symptômes de chancre bactérien des agrumes ont été observés sur 38 spécimens de *Citrus* et des symptômes similaires ont été observés sur une dizaine d'espèces de Rutacées non répertoriées comme hôte.

Les échanges d'informations et de matériel scientifiques au XIX^e siècle n'étaient pas moins intenses qu'aujourd'hui et les différents herbiers mondiaux ont constamment échangé des planches botaniques (les botanistes créant en général plusieurs planches

5. Le terme introgression désigne en biologie le transfert d'un gène (de résistance par exemple) dans une variété à fort intérêt agronomique par croisements successifs avec une autre variété naturellement résistante mais dont l'intérêt agronomique est moindre.

pour un même prélèvement), ce qui fait qu'il est possible d'augmenter la robustesse des analyses avec des échantillons redondants, mais aussi le cas échéant de maintenir une « copie de sauvegarde » d'un échantillon.



Figure 13.4. Une salle de l'Herbier national au Muséum national d'Histoire naturelle de Paris. Photo : R. Garrouste, MNHN



Figure 13.5. Planche d'herbier de la figure 13.2a après prélèvement (excision rectangulaire d'environ 1 cm² au centre de la planche). Photo : L. Gagnevin

La partie délicate du traitement des spécimens d'herbier est bien entendu le prélèvement proprement dit puisqu'il est nécessaire de découper des tissus végétaux potentiellement infectés en quantité suffisante pour purifier les acides nucléiques qui seront analysés (figure 13.5). Ceci ne pose en général pas de problème pour les spécimens cultivés qui sont multiples et redondants, mais peut s'avérer un facteur limitant pour des plantes dont les spécimens doivent être préservés ou pour des spécimens très anciens non redondants. Des tests sont en cours pour évaluer des techniques de prélèvement des acides nucléiques préservant au maximum l'intégrité du spécimen botanique (en minimisant la quantité de matériel prélevé ou même par seul frottis de l'échantillon ; pour exemple, (Shepherd 2017)).

13.6. Caractéristiques des acides nucléiques anciens et leur traitement

L'essor de l'étude de l'ADN ancien a été facilité par le séquençage parallèle massif (HTS, voir section 13.3). Cette technologie repose sur une première étape de construction de banques d'acides nucléiques qui consiste à lier, de part et d'autre de l'ADN d'intérêt, des petits segments d'ADN artificiels spécifiques (oligonucléotides de synthèse), permettant une amplification de cet ensemble (par PCR, *Polymerase Chain Reaction*) avant une seconde étape de séquençage à grande échelle. Dès les années 2000, grâce aux HTS, un ADN dégradé (tel que l'ADN ancien) pouvait ainsi être séquençé à partir d'une extraction de tissu entier (Briggs *et al.* 2007 ; Burrell *et al.* 2015). Un certain nombre de caractéristiques rend le traitement et l'étude des ADN anciens (notamment ceux conservés dans les plantes d'herbiers) difficiles : elles seront décrites ci-après pour des agents pathogènes classiques (micro-organismes à ADN) extraits à partir des feuilles des échantillons d'herbiers, mais aussi pour les virus, dont le support génétique peut ne pas être l'ADN.

L'analyse de nombreuses séquences anciennes d'origines diverses (Pääbo *et al.* 2004 ; Briggs *et al.* 2007) ont révélé :

- une fragmentation de l'ADN : cette dernière est partiellement expliquée par la présence d'une purine en aval, dont la dépurination (hydrolyse des bases puriques de l'ADN, G[uanine] ou A[dénine]) faciliterait la cassure du brin immédiatement en amont (Lindahl et Andersson 1972). Ainsi, les deux processus, dépurination et fragmentation, sont liés ; cependant, certaines cassures ont lieu indépendamment de toute dépurination ;
- une accumulation de substitutions nucléotidiques aux extrémités des fragments séquençés : cette accumulation est principalement due au processus naturel de désamination de la base pyrimidique C[ytosine], préférentiellement au niveau d'un ADN simple brin. Pour des raisons inhérentes à la préparation des banques et au séquençage, cette désamination se traduit le plus souvent aux extrémités 5' par une substitution de C vers T[hymine], et aux extrémités 3' du même brin par une substitution de G vers A.

À partir de l'analyse de 86 échantillons d'herbiers collectés au cours des trois siècles derniers (genres *Arabidopsis* et *Solanum*, infectés ou non par l'oomycète *P. infestans*), un panorama des caractéristiques d'ADN ancien a récemment été établi. Les deux types de dégradations majoritaires observées, fragmentation et désamination, rendent compte d'un processus graduel et cumulatif (une diminution significative de la taille des fragments en fonction de l'âge est observée). La fragmentation affecte en moyenne 1,7 % des liaisons nucléotidiques tous les 100 ans, et la désamination atteint près de 5,0 % des nucléotides à l'extrémité des fragments pour les échantillons les plus anciens (des années 1730). Une analyse comparative a été effectuée plus en détail pour deux catégories d'ADN (nucléaire et chloroplastique). Elle a montré que les mêmes dégradations avaient lieu pour les deux catégories d'ADN mais que leur taux est jusqu'à deux fois plus élevé pour l'ADN nucléaire (Weiss *et al.* 2016).

De par sa fragmentation et sa faible concentration, mais aussi la présence d'inhibiteurs dans les feuilles des plantes hôtes (composés phénoliques et polysaccharides, préjudiciables à l'étape de construction de banques), l'ADN ancien de phytopathogènes est difficile à purifier par les méthodes commerciales les plus courantes. Des protocoles d'extraction classiques ont été optimisés avec une réduction du nombre d'étapes, l'ajout de molécules neutralisant les inhibiteurs, et enfin l'adaptation de colonnes de purification aux petites tailles des fragments d'ADN (Kistler 2012 ; Healey *et al.* 2014 ; Gutaker *et al.* 2017). La qualité de cette extraction dépend également de la taxonomie de la plante hôte, ainsi que des conditions de séchage et de conservation des échantillons d'herbiers (Sarkinen *et al.* 2012 ; Choi *et al.* 2015). De la même manière, la construction de banques préalable au séquençage doit être adaptée pour les échantillons d'ADN dégradé (Sarkinen *et al.* 2012 ; Gansauge et Meyer 2013 ; Carøe *et al.* 2017). Enfin, des stratégies de capture (par hybridation) peuvent être développées afin d'enrichir les échantillons en acides nucléiques cibles (par exemple ceux du pathogène par rapport à celui de la plante), mais elles nécessitent d'avoir une bonne connaissance du génome du micro-organisme (Gasc *et al.* 2016).

13.6.1. Le cas particulier des acides nucléiques viraux

À l'instar des travaux cités ci-avant qui ont exploré les histoires évolutives de bactéries ou d'oomycètes, les progrès récents dans les technologies HTS ont aussi permis de reconstituer des génomes viraux datant de plusieurs siècles (Smith *et al.* 2014 ; Yoshida *et al.* 2014). Une des spécificités de l'application des HTS à la caractérisation des virus, qu'ils soient contemporains ou anciens, est la nature même des acides nucléiques. À la différence des études précitées, l'ADN ne constitue pas la cible unique des travaux en virologie. Par ailleurs, aucun gène universel permettant de reconstruire un arbre phylogénétique de tous les virus n'existe. De fait, c'est une palette de différentes approches sans *a priori* de métagénomique virale qui a été décrite à ce jour, chaque

approche visant une forme spécifique d'acide nucléique viral (ARN et/ou ADN, simple ou double brins) avec ses avantages et ses inconvénients (Massart *et al.* 2014 ; Roosinck *et al.* 2015).

Une première approche consiste à extraire les ADN ou ARN totaux à partir d'un échantillon et de séquencer les acides nucléiques. La faible proportion de séquences génomiques de virus à ADN circulaire peut être aussi enrichie par amplification en cercle roulant (Wyant *et al.* 2012), approche qui a été utilisée avec succès pour reconstituer le génome total d'un géminivirus responsable d'une maladie émergente aux États-Unis à partir d'un échantillon de vigne de l'herbier de l'Université de Davis en Californie (Al Rwahnih *et al.* 2015). De même, les préparations d'ARN peuvent être enrichies en séquences virales par élimination de grandes quantités d'ARN ribosomal de la plante (ribodéplétion) (Adams *et al.* 2009).

Une deuxième approche consiste à extraire les acides nucléiques viraux à partir d'une première étape de semi-purification des particules virales (*Virus Associated Nucleic Acids*, VANA) (Candresse *et al.* 2014).

Une troisième approche, qui est limitée à l'analyse des virus à ARN, consiste à purifier les molécules d'ARN double-brin (ARNdb) qui s'accumulent dans les cellules de la plante lors de la réplication du virus (Roosinck *et al.* 2010). Cette approche a permis d'identifier un chrysovirus à partir d'épis de maïs d'Arizona datant d'environ un millénaire (Peyambari *et al.* 2019).

Enfin, une quatrième approche consiste à isoler et à séquencer massivement les ARN de petite taille, issus de la voie de défense antivirale appelée ARN interférence chez les plantes (Pooggin 2018). Ces ARN interférents (*small interfering RNA* – siRNA) ont une longueur de 21, 22 ou 24 nucléotides et l'ensemble de leurs séquences recouvre la plupart, sinon la totalité des génomes viraux accumulés pendant l'infection (Ding et Voinnet 2007). Cette stratégie a l'avantage de pouvoir détecter un large spectre de virus indépendamment de leur nature (ARN ou ADN) et de leur structure (linéaire, circulaire, simple ou double brin).

Cette dernière approche est peut-être en train de devenir centrale dans la recherche et l'analyse des virus anciens. En effet, plusieurs études récentes l'ont appliquée sur du matériel végétal ancien, comme par exemple des restes végétaux archéologiques ou bien des spécimens d'herbiers. L'ARN a longtemps été considéré comme une cause perdue en archéovirologie⁶, du fait de sa vulnérabilité aux attaques hydrolytiques, mais ces travaux suggèrent qu'il peut en fait être préservé sous certaines conditions arides pendant des centaines d'années (Guy 2014). Par ailleurs, les siRNA seraient plus stables que les longs ARN ou ADN grâce notamment aux structures secondaires de l'ARN.

6. Discipline qui étudie les virus de la préhistoire à l'époque contemporaine.

Ainsi, Smith *et al.* (2014) ont utilisé cette approche et réussi à identifier et à reconstruire le génome ancien à ARN d'un isolat de l'espèce *Barley stripe mosaic virus* (*Hordeivirus*, *Virgaviridae*) à partir de semences d'orge âgées de 700 ans.

13.7. *Xanthomonas citri* pv. *citri* et son émergence dans l'océan Indien

Le chancre asiatique est une maladie bactérienne qui affecte les agrumes dans la quasi-totalité des zones tropicales et subtropicales. Elle est causée par la γ -protéobactérie *Xanthomonas citri* pv. *citri* (*Xci*) ; les symptômes de chancre sur feuilles, fruits et tiges ont un impact économique majeur. Une meilleure compréhension des mécanismes sous-jacents à l'émergence, l'évolution et la diffusion de *Xci* apparaît comme un prérequis important pour une meilleure gestion de cette maladie et pour anticiper de nouvelles émergences.

La bactérie *Xci* a probablement émergé en Asie, zone d'origine de son hôte (le genre *Citrus*) avant de connaître une expansion hors du continent (vers l'Amérique, l'Afrique et l'Océanie) durant la première moitié du XX^e siècle (Pruvost *et al.* 2014). Dans la région de l'océan Indien, cette maladie a été décrite pour la première fois en 1937 aux îles Maurice et Rodrigues puis à partir de la fin des années 1960 à la Réunion, dans l'archipel des Comores et aux Seychelles. Depuis près de 35 ans, des chercheurs du CIRAD ont collecté des souches de la bactérie *Xci* provenant des différentes îles de l'océan Indien ainsi que d'autres régions du monde. Les génomes d'un certain nombre de ces souches ont été séquencés dans l'objectif de mieux comprendre les mécanismes d'émergence de cette bactérie au sein des îles de l'océan Indien. Pour ce faire, un jeu de données constitué de 150 génomes échantillonnés entre 1980 et 2015 au sein des différentes îles de la région ainsi que du reste du monde (incluant bien entendu l'Asie, zone d'origine des *Citrus*) a été constitué. Dans le but d'augmenter la dimension temporelle de ce jeu de données, les herbiers de l'université de la Réunion et de l'île Maurice ont été explorés. Deux échantillons historiques de *Citrus* sp. isolés à l'île Maurice, respectivement en 1937 et 1974, arborant des symptômes typiques de chancre ont été sélectionnés afin d'en séquencer l'ADN global en suivant les protocoles spécifiques présentés en section 13.6. Les séquences historiques ainsi générées ont été analysées de plusieurs manières.

Dans un premier temps nous avons cherché à déterminer l'origine taxonomique des lectures⁷ de séquence par une approche classique d'assignation taxonomique, obtenue par recherche de séquences homologues d'organismes séquencés et déposés dans les bases de données publiques. Pour les deux échantillons d'herbiers analysés, la majorité (environ 60 %) des lectures sont homologues à des séquences d'agrumes, un résultat attendu puisque la majeure partie des tissus dont on a extrait l'ADN provient de la plante. Ensuite, une fraction (environ 15 %) des lectures sont homologues au génome

7. Ou *reads*, ensemble des séquences individuelles générées par le séquençage à haut débit.

humain, résultat s'expliquant par des contaminations de l'échantillon d'herbier ayant eu lieu depuis l'échantillonnage initial par l'explorateur jusqu'à la mise en collection et toutes les autres manipulations ultérieures, incluant celles au sein de l'herbier. Finalement, une minorité (moins de 5 %) des lectures sont attribuées à *Xci*, la bactérie responsable du chancre des agrumes et à un mélange de micro-organismes divers (bactérie, virus, champignons) présents sur ou dans l'échantillon d'herbier (5 % également).

Dans un second temps, les lectures assignées à *Xci* ont été analysées dans le but de mesurer leurs niveaux de dégradation et notamment de les comparer avec ceux observés sur des échantillons contemporains. Pour ce faire deux statistiques ont été estimées à l'aide de l'outil *mapDamage* (Jonsson *et al.* 2013) : i) la distribution de la taille des fragments et ii) l'intensité de désamination des Cytosines (C) en fonction de leur position dans la lecture. Les lectures issues des échantillons historiques ont révélé des tailles de fragments réduites (60 pb en moyenne). De plus, une désamination a été observée sur une fraction significative des C des cinq premières bases présentes aux extrémités des lectures historiques, contrairement aux lectures modernes qui n'en présentent aucune trace. Ces résultats permettent d'émettre deux conclusions : i) l'ADN issu d'échantillons d'herbiers pourtant relativement jeunes (datant de 1937 et 1974 dans ce cas) peut être partiellement dégradé (fragmentation et désamination) et ii) ces traces de dégradations attestent de l'authenticité des séquences historiques et permettent de prouver que ces dernières ne sont pas issues de contaminations contemporaines lors de la manipulation des échantillons.

Finalement, les relations de parentés entre les deux souches historiques et les 150 génomes contemporains de *Xci* ont été reconstruites par un arbre phylogénétique dont l'analyse détaillée a permis de faire l'hypothèse d'une première introduction de la bactérie à l'île Maurice durant la période suivant l'abolition de l'esclavage (1835). Ainsi, le chancre des agrumes aurait pu être introduit par des plants importés d'Asie, lors du recrutement de travailleurs engagés (souvent d'origine indienne) dans les plantations des îles de l'océan Indien. Une vitesse d'évolution pour la bactérie *Xci* a également pu être inférée pour la toute première fois (Richard *et al.* 2020). L'analyse future des échantillons historiques d'herbiers supplémentaires issus d'autres origines géographiques (incluant la zone asiatique d'origine ou d'autres régions où la maladie a été introduite), d'autres périodes (XIX^e siècle) et d'autres espèces végétales (*Rutacées* autres que *Citrus* sp.) devrait permettre de reconstruire l'histoire évolutive de la maladie à une échelle temporelle et géographique élargie.

13.8. Émergence et histoire évolutive des virus phytopathogènes : le modèle des géminivirus

Les géminivirus sont responsables de nombreuses maladies émergentes dans le monde avec un impact économique majeur sur des cultures maraichères et vivrières,

très importantes pour la sécurité alimentaire notamment en Afrique subsaharienne comme le manioc (*Manihot esculenta*) et la patate douce (*Ipomoea batatas*). Leur potentiel évolutif important, avec des taux élevés de mutation et de recombinaison, font de ces virus un modèle idéal pour la compréhension des processus épidémiologiques et évolutifs associés à l'émergence virale.

13.8.1. Cas d'un complexe d'espèces responsable d'une maladie émergente

Les géminivirus du manioc forment un complexe d'espèces⁸ responsables de la maladie de la mosaïque du manioc, la plus sévère et la plus dommageable de cette culture vivrière en Afrique subsaharienne. Afin d'étudier les processus évolutifs à l'origine du succès épidémiologique de ces virus, nous avons entrepris de séquencer des génomes viraux historiques à partir d'échantillons de manioc de l'Herbier du Muséum national d'Histoire naturelle (Paris) présentant des symptômes apparents de la maladie (figure 13.6).



Figure 13.6. Planche P04808771 de l'Herbier du MNHN avant prélèvement, comprenant une feuille de manioc (*Manihot glaziovii*) collectée à Bambari en Centrafrique en 1928 et présentant une suspicion de la maladie de la mosaïque du manioc avec des symptômes de déformation foliaire. Photo : MNHN.

8. Plusieurs espèces distinctes mais étroitement apparentées responsables (ensemble ou séparément) d'une maladie et difficilement distinguables autrement que par la génétique.

Comme décrit précédemment, les géminivirus du manioc déclenchent une réponse de défense de type immunitaire de la plante avec l'accumulation de siRNA de 21, 22 et 24 nucléotides (Pooggin 2018). Nous avons entrepris, dans le cadre d'une preuve de concept, d'évaluer si l'approche par séquençage à haut débit des siRNA permettait de reconstruire les génomes complets à ADN des géminivirus du manioc. Nos résultats préliminaires démontrent notre capacité à reconstituer la séquence presque complète de ces géminivirus à partir de spécimens d'herbier vieux de 50 et 90 ans. En complément des séquences contemporaines, ces séquences anciennes peuvent être utilisées dans des études phylogénétiques, phylogéographiques et de génomique comparative pour élucider l'émergence et l'évolution de l'agent de la mosaïque du manioc.

13.8.2. Cas d'un géminivirus cryptique

La patate douce, dont le centre d'origine semble être l'Amérique centrale et du Sud (Clark *et al.* 2012), se classe parmi les dix cultures vivrières les plus importantes, notamment pour l'Afrique subsaharienne. De par sa multiplication végétative par les tubercules ou les boutures, la patate douce est sujette à l'accumulation et au transport de pathogènes, notamment divers virus cryptiques⁹ qui sont transmis à la descendance et vont s'accumuler au fil des générations pour former un complexe viral (une « association » de divers éléments viraux d'origines différentes) pouvant déboucher sur des formes sévères de maladie grâce à des phénomènes de synergie avec des répercussions sur la croissance, le rendement et la qualité des plantes (Paprotka *et al.* 2010). Plus d'une trentaine de virus ont été décrits chez la patate douce. La moitié de ces virus sont des virus à ADN, décrits très récemment grâce aux approches HTS, dont la plupart sont associés à des infections asymptomatiques. C'est notamment le cas de l'espèce *Sweet potato symptomless virus 1* (SpSV-1, *Mastrevirus*, *Geminiviridae*), pour lequel 14 génomes complets ont été caractérisés en Asie, Afrique et Amérique du Sud (Cao *et al.* 2017 ; Souza *et al.* 2018), suite à la première description d'une séquence partielle dans un cultivar du Pérou (Kreuze *et al.* 2009). Ce travail précurseur, qui utilise l'approche des siRNA dans l'indexage¹⁰ du germplasm de patate douce, a en effet permis d'identifier de nombreux virus cryptiques jamais décrits auparavant mais pouvant potentiellement constituer des complexes viraux épidémiques. Afin d'étudier les processus évolutifs à l'origine de cette association asymptomatique entre un virus et une plante hôte, nous avons entrepris de reconstruire les génomes viraux historiques de SpSV-1, à partir d'une vingtaine d'échantillons historiques d'herbiers de patate douce datant pour les plus anciens de 1820, avec l'approche des siRNA. Les premières analyses confirment notre capacité à reconstruire des séquences quasi complètes de SpSV-1 âgées de près de deux cents ans, nous permettant d'avoir une image de la communauté virale hébergée

9. La présence du virus n'est pas associée à des symptômes de maladie.

10. Évaluation des espèces virales présentes dans une plante.

par cette plante et d'émettre des hypothèses sur la structure et la dynamique épidémiologique des complexes viraux.

13.9. Discussion

Les approches et résultats présentés au sein de ce chapitre illustrent l'immense potentiel des collections historiques d'herbiers dans la compréhension de l'émergence et de l'évolution des micro-organismes pathogènes des cultures, mais aussi du développement récent de nouvelles approches méthodologiques permettant d'exploiter robustement ces ressources. De nouvelles études, sur d'autres maladies et d'autres cultures, sont actuellement en cours et permettront d'améliorer les connaissances sur l'histoire et l'évolution des micro-organismes phytopathogènes, prérequis indispensable à une meilleure gestion de ces derniers.

Bien entendu, il est important d'avoir en tête que les informations et données génétiques qui ont été obtenues à partir de ces échantillons ont des caractéristiques et des limites dont il faut tenir compte tout au long de leur traitement.

D'une part, la répartition temporelle et géographique des collections des muséums d'histoire naturelle est biaisée par plusieurs facteurs : périodes de prospection plus ou moins intenses selon le contexte historique et économique par exemple, aspects de politique scientifique (souvent liée aux priorités coloniales ou économiques du pays), préférences botaniques des collecteurs de terrain, etc.

Cependant, le fait que nous travaillions sur des plantes d'intérêt agronomique offre certains avantages : les échantillons sont nombreux et ont une valeur botanique et conservatoire moindre, ils proviennent de régions où la diversité des plantes est valorisable (zone d'origine par exemple), et les campagnes de prospection peuvent être relativement systématiques et rigoureuses car leur objectif est souvent de rechercher de nouvelles ressources biologiques pour l'agriculture locale et mondiale.

Sur un plan strictement technique il est important de noter que l'étude génétique des échantillons d'herbier est encore bien loin d'être équivalente à celle des souches vivantes : en particulier, la fragmentation des acides nucléiques anciens empêche le séquençage de longs fragments et limite certaines approches de génomique comparative très informatives (architecture du génome, étude des régions répétées, identification de gènes ou régions ayant disparu des génomes modernes, événements de recombinaison, etc.). De plus, la grande majorité des études se sont à ce jour focalisées sur la fraction des acides nucléiques assignée au pathogène, traitant à part celle des autres micro-organismes ou de la plante hôte dont l'analyse apparaît pourtant prometteuse et nécessaire pour mieux comprendre les dynamiques épidémiologiques et les mécanismes de co-évolution de façon plus intégrée.

13.10. Remerciements et financements

Nous souhaitons remercier pour leur accueil et leur assistance les curateurs d'herbiers Deby Arifiani (Herbarium Bogoriense, Indonésie), Jan Wieringa (Naturalis Biodiversity Center, Dutch Herbaria, Pays-Bas), Timm Utteridge (Royal Botanical Garden, Kew, Royaume-Uni), Mark Carine (Natural History Museum Herbarium, Royaume-Uni), Shannon Dominick (National Fungus Collection, USDA, États-Unis), Matthew Pace (New York Botanical Garden, Steere Herbarium, États-Unis), Meghann Toner (US National Herbarium, Smithsonian, États-Unis), Terry H. Trinder-Smith (Bolus Herbarium, University of Cape Town, Afrique du Sud).

Nos travaux ont bénéficié de financements de l'action COST (*European Cooperation in Science and Technology*) CA16107 EuroXanth, du programme SYNTHESYS¹¹ financé par l'action Research Infrastructure du programme FP7 « Capacities » de la Communauté européenne, du CIRAD (AI-CRESI-3/2016), d'Agropolis Fondation (convention N° 1600-014), de l'Agence nationale pour la recherche (JCJC MUSEOBACT contrat ANR-17-CE35-0009-01), de l'ED 227, Muséum national d'Histoire naturelle et Sorbonne Université, ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, ainsi que de l'Union européenne (*European Regional Development Fund*, ERDF contract GURDT I2016-1731-0006632) et du Conseil régional de la Réunion.

13.11. Bibliographie

- Achtman, M. (2016). How old are bacterial pathogens?. *Proceedings of the Royal Society B*, 283(1836).
- Adams, I.P., Glover, R.H., Monger, W.A., Mumford, R., Jackeviciene, E., Navalinskiene, M., Samuitiene, M., Boonham, N. (2009). Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. *Molecular Plant Pathology*, 10(4), 537–545.
- Al Rwahnih, M., Rowhani, A., Golino, D. (2015). First report of grapevine red blotch-associated virus in archival grapevine material from Sonoma County, California. *Plant Disease*, 99(6), 895.
- Almeida, R.P.P., De La Fuente, L., Koebnik, R., Lopes, J.R.S., Parnell, S., Scherm, H. (2019). Addressing the new global threat of *Xylella fastidiosa*. *Phytopathology*, 109(2), 172–174.
- Anderson, P.K., Cunningham, A.A., Patel, N.G., Morales, F.J., Epstein, P.R., Daszak, P. (2004). Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. *Trends in Ecology & Evolution*, 19(10), 535–544.

11. <http://www.synthesys.info/>.

- Antonovics, J., Hood, M.E., Thrall, P.H., Abrams, J.Y., Duthie, G.M. (2003). Herbarium studies on the distribution of anther-smut fungus (*Microbotryum violaceum*) and *Silene* species (Caryophyllaceae) in the eastern United States. *American Journal of Botany*, 90(10), 1522–1531.
- Briggs, A.W., Stenzel, U., Johnson, P.L., Green, R.E., Kelso, J., Prufer, K., Meyer, M., Krause, J., Ronan, M.T., Lachmann, M., Paabo, S., (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences of the United States of America*, 104(37), 14616–14621.
- Burrell, A.S., Disotell, T.R., Bergey, C.M., (2015). The use of museum specimens with high-throughput DNA sequencers. *Journal of Human Evolution*, 79, 35–44.
- Candresse, T., Filloux, D., Muhire, B., Julian, C., Galzi, S., Fort, G., Bernardo, P., Daugrois, J.H., Fernandez, E., Martin, D.P., Varsani, A., Roumagnac, P. (2014). Appearances can be deceptive: revealing a hidden viral infection with deep sequencing in a plant quarantine context. *PLoS One*, 9(7), e102945.
- Cao, M., Lan, P., Li, F., Abad, J., Zhou, C., Li, R. (2017). Genome characterization of sweet potato symptomless virus 1: a mastrevirus with an unusual nonanucleotide sequence. *Archives of Virology*, 162(9), 2881–2884.
- Carøe, C., Gopalakrishnan, S., Vinner, L., Mak, S.S.T., Sinding, M.H.S., Samaniego, J.A., Wales, N., Sicheritz-Pontén, T., Gilbert, M.T. (2017). Single-tube library preparation for degraded DNA. *Methods in Ecology and Evolution*, 9, 410–419.
- Choi, J., Lee, H., Shipunov, A. (2015). All that is gold does not glitter? Age, taxonomy, and ancient plant DNA quality. *PeerJ*, 3, e1087.
- Clark, C.A., Davis, J.A., Abad, J.A., Cuellar, W.J., Fuentes, S., Kreuze, J.F., Gibson, R.W., Mukasa, S.B., Tugume, A.K., Tairo, F.D., Valkonen, J.P.T. (2012). Sweet potato viruses: 15 years of progress on understanding and managing complex diseases. *Plant Disease*, 96(2), 168–185.
- Croucher, N.J., Didelot, X. (2015). The application of genomics to tracing bacterial pathogen transmission. *Current Opinion in Microbiology*, 23, 62–67.
- Ding, S.W., Voinnet, O. (2007). Antiviral immunity directed by small RNAs. *Cell*, 130(3), 413–426.
- Fawcett, H.S., Jenkins, A.E. (1933). Records of citrus canker from herbarium specimens of the genus *Citrus* in England and the United States. *Phytopathology*, 23(10), 820–824.
- Gansauge, M.T., Meyer, M. (2013). Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nature Protocols*, 8(4), 737–748.

- Gasc, C., Peyretaillade, E., Peyret, P. (2016). Sequence capture by hybridization to explore modern and ancient genomic diversity in model and nonmodel organisms. *Nucleic Acids Research*, 44(10), 4504–4518.
- Goss, E.M. (2015). Genome-enabled analysis of plant-pathogen migration. *Annual Review of Phytopathology*, 53, 121–135.
- Gutaker, R.M., Reiter, E., Furtwangler, A., Schuenemann, V.J., Burbano, H.A. (2017). Extraction of ultrashort DNA molecules from herbarium specimens. *Biotechniques*, 62(2), 76–79.
- Guy, P.L. (2014). Prospects for analyzing ancient RNA in preserved materials. *Wiley Interdisciplinary Reviews: RNA*, 5(1), 87–94.
- Hasse, C.H. (1915). *Pseudomonas citri*, the cause of citrus canker. *Journal of Agricultural Research*, 4(1), 97–99.
- Healey, A., Furtado, A., Cooper, T., Henry, R.J. (2014). Protocol: a simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. *Plant Methods*, 10, 21.
- Jonsson, H., Ginolhac, A., Schubert, M., Johnson, P.L., Orlando, L. (2013). mapDamage 2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, 29(13), 1682–1684.
- Kistler, L. (2012). Ancient DNA extraction from plants. Dans *Ancient DNA. Methods and Protocols*, Shapiro, B., Hofreiter, M. (dir.). Humana Press, Hatfield.
- Kreuze, J.F., Perez, A., Untiveros, M., Quispe, D., Fuentes, S., Barker, I., Simon, R. (2009). Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. *Virology*, 388(1), 1–7.
- Larson, G., Piperno, D.R., Allaby, R.G., Purugganan, M.D., Andersson, L. *et al.* (2014). Current perspectives and the future of domestication studies. *Proceedings of the National Academy of Sciences of the United States of America*, 111(17), 6139–6146.
- Lindahl, T., Andersson, A. (1972). Rate of chain breakage at apurinic sites in double-stranded deoxyribonucleic acid. *Biochemistry*, 11(19), 3618–3623.
- Martin, M.D., Cappellini, E., Samaniego, J.A., Zepeda, M.L., Campos, P.F. *et al.* (2013). Reconstructing genome evolution in historic samples of the Irish potato famine pathogen. *Nature Communications*, 4, 2172.
- Massart, S., Olmos, A., Jijakli, H., Candresse, T. (2014). Current impact and future directions of high throughput sequencing in plant virus diagnostics. *Virus Research*, 188, 90–96.

- McDonald, B.A., Linde, C. (2002). Pathogen population genetics, evolutionary potential, and durable resistance. *Annual Review of Phytopathology*, 40, 349–379.
- Mira, A., Pushker, R., Rodriguez-Valera, F. (2006). The Neolithic revolution of bacterial genomes. *Trends in Microbiology*, 14(5), 200–206.
- Pääbo, S., Poinar, H., Serre, D., Jaenicke-Despres, V., Hebler, J., Rohland, N., Kuch, M., Krause, J., Vigilant, L., Hofreiter, M. (2004). Genetic analyses from ancient DNA. *Annual Review of Genetics*, 38, 645–679.
- Paprotka, T., Boiteux, L.S., Fonseca, M.E., Resende, R.O., Jeske, H. *et al.* (2010). Genomic diversity of sweet potato geminiviruses in a Brazilian germplasm bank. *Virus Research*, 149(2), 224–233.
- Peltier, G.L., Frederich, W.J. (1924). Further studies on the relative susceptibility to citrus canker of different species and hybrids of the genus *Citrus*, including the wild relatives. *Journal of Agricultural Research*, 28(3), 227–239.
- Peyambari, M., Warner, S., Stoler, N., Rainer, D., Roossinck, M.J. (2019). A 1,000-year-old RNA virus. *Journal of Virology*, 93, e01188-18.
- Pooggin, M.M. (2018). Small RNA-omics for plant virus identification, virome reconstruction, and antiviral defense characterization. *Frontiers in Microbiology*, 9, 2779.
- Pruvost, O., Magne, M., Boyer, K., Leduc, A., Tourterel, C. *et al.* (2014). A MLVA genotyping scheme for global surveillance of the citrus pathogen *Xanthomonas citri* pv. *citri* suggests a worldwide geographical expansion of a single genetic lineage. *PLoS One*, 9(6), e98129.
- Richard, D., Pruvost, O., Balloux, F., Boyer, C., Rieux, A., Lefeuvre, P. (2020). Time-calibrated genomic evolution of a monomorphic bacterium during its establishment as an endemic crop pathogen. *Molecular Ecology*, 00, 1–13.
- Rieux, A., Balloux, F. (2016). Inferences from tip-calibrated phylogenies: a review and a practical guide. *Molecular Ecology*, 25(9), 1911–1924.
- Roossinck, M.J., Martin, D.P., Roumagnac, P. (2015). Plant virus metagenomics: advances in virus discovery. *Phytopathology*, 105(6), 716–727.
- Roossinck, M.J., Saha, P., Wiley, G.B., Quan, J., White, J.D., Lai, H., Chavarria, F., Shen, G., Roe, B.A. (2010). Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. *Molecular Ecology*, 19, 81–88.
- Sarkinen, T., Staats, M., Richardson, J.E., Cowan, R.S., Bakker, F.T. (2012). How to open the treasure chest? Optimising DNA extraction from herbarium specimens. *PLoS One*, 7(8), e43808.

- Savary, S., Willocquet, L., Pethybridge, S.J., Esker, P., McRoberts, N., Nelson, A. (2019). The global burden of pathogens and pests on major food crops. *Nature Ecology and Evolution*, 3(3), 430–439.
- Shepherd, L.D. (2017). A non-destructive DNA sampling technique for herbarium specimens. *PLoS One*, 12(8), e0183555.
- Smith, O., Clapham, A., Rose, P., Liu, Y., Wang, J., Allaby, R.G. (2014). A complete ancient RNA genome: identification, reconstruction and evolutionary history of archaeological Barley Stripe Mosaic Virus. *Scientific Reports*, 4, 4003.
- Souza, C.A., Rossato, M., Melo, F.L., Boiteux, L.S., Pereira-Carvalho, R.C. (2018). First report of Sweet Potato Symptomless Virus 1 infecting *Ipomoea batatas* in Brazil. *Plant Disease*, 102, 2052.
- Stukenbrock, E.H., Bataillon, T. (2012). A population genomics perspective on the emergence and adaptation of new plant pathogens in agro-ecosystems. *PLoS Pathogens*, 8(9), e1002893.
- Stukenbrock, E.H., McDonald, B.A. (2008). The origins of plant pathogens in agro-ecosystems. *Annual Review of Phytopathology*, 46, 75–100.
- Wandeler, P., Hoeck, P.E., Keller, L.F. (2007). Back to the future: museum specimens in population genetics. *Trends in Ecology & Evolution*, 22(12), 634–642.
- Weiss, C.L., Schuenemann, V.J., Devos, J., Shirsekar, G., Reiter, E. *et al.* (2016). Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. *Royal Society Open Science*, 3(6), 160239.
- Wyant, P.S., Strohmeier, S., Schafer, B., Krenz, B., Assuncao, I.P. *et al.* (2012). Circular DNA genomics (circomics) exemplified for geminiviruses in bean crops and weeds of northeastern Brazil. *Virology*, 427(2), 151–157.
- Yoshida, K., Schuenemann, V.J., Cano, L.M., Pais, M., Mishra, B. *et al.* (2013). The rise and fall of the *Phytophthora infestans* lineage that triggered the Irish potato famine. *Elife*, 2, e00731.
- Yoshida, K., Burbano, H.A., Krause, J., Thines, M., Weigel, D., Kamoun, S. *et al.* (2014). Mining herbaria for plant pathogen genomes: back to the future. *PLoS Pathogens*, 10(4), e1004028.
- Zeder, M.A. (2017). Domestication as a model system for the extended evolutionary synthesis. *Interface Focus*, 7(5), 20160133.

Objectifs de la thèse

Cette thèse a pour objectif principal d'évaluer l'apport de l'ADN ancien dans l'étude de l'émergence et de l'évolution de pathogènes de cultures en prenant comme principal modèle biologique *Xanthomonas citri* pathovar *citri* (*Xci*), la bactérie responsable du chancre asiatique des agrumes.

Dans un premier temps, nous avons cherché à développer et optimiser les protocoles moléculaires et les pipelines bioinformatiques permettant de séquencer et reconstruire au mieux des génomes de *Xci* à partir de spécimens historiques d'herbiers. Dans le cadre de cette démarche, nous nous sommes spécifiquement intéressés à la mesure des patrons de dégradations de l'ADN et avons tenté de déterminer les facteurs qui structurent leurs variations (chapitre 2). Dans un second temps, nous avons voulu préciser l'histoire de l'émergence de *Xci* à une échelle locale, celle des îles du sud-ouest de l'océan Indien (SOOI) grâce à l'analyse détaillée du premier génome historique bactérien reconstruit à partir d'un échantillon d'herbier datant de 1937. Durant ce travail, nous avons également cherché à décrire la composition taxonomique de l'échantillon d'herbier analysé (chapitre 3). Finalement, nous avons tenté de reconstruire l'histoire de l'origine et de la diversification de *Xci* à l'échelle mondiale par l'analyse combinée de 13 génomes historiques générés durant cette thèse et d'une collection de génomes modernes représentative de la diversité globale du pathogène (chapitre 4).

Suite à la revue bibliographique présentée en chapitre 1, mes travaux de thèse sont présentés dans quatre chapitres dont deux sont exposés en anglais sous forme de manuscrits (publiés ou en cours de finalisation). Le chapitre 5 synthétise brièvement un projet de recherche associé à la thématique de mon travail principal mais ciblant un autre modèle biologique, auquel j'ai eu l'opportunité de participer dans le cadre d'une collaboration scientifique. Pour finir, une discussion générale tentera d'intégrer les différents résultats obtenus dans le cadre de cette thèse et de faire émerger quelques perspectives de recherche.

Chapitre 2 – Optimisation et application de protocoles innovants au laboratoire et en bioinformatique

Les spécimens historiques d'herbier sont des échantillons caractérisés par des acides nucléiques en relativement faible quantité et dégradés, comme cela est le cas pour la plupart des tissus biologiques historiques (voir chapitre 1.3.4.). De ce fait, ces derniers nécessitent un traitement spécifique, aussi bien d'un point de vue moléculaire (pour l'extraction et la manipulation des acides nucléiques au laboratoire) que bioinformatique (pour l'analyse des séquences nucléotidiques générées). Ce chapitre est consacré à la présentation des travaux d'optimisation des protocoles moléculaires et du pipeline bioinformatique réalisés durant cette thèse, travaux ayant permis de générer et analyser les données et les résultats présentés dans les chapitres suivants.

Au laboratoire, nous avons travaillé à l'amélioration des protocoles d'extraction des acides nucléiques, de leur amplification par qPCR ainsi que de leur transformation en bibliothèques en vue de leur séquençage par les machines de la technologie Illumina. Cela nous a permis de garder la mainmise sur les différentes étapes de la manipulation moléculaire des acides nucléiques historiques et par la même occasion d'en réduire les coûts (comparé à la stratégie précédemment employée qui visait à sous-traiter certaines de ces étapes auprès d'une plateforme de séquençage). Concernant l'analyse des lectures de séquençage, nous avons développé un pipeline bioinformatique adapté aux spécificités de l'ADN dégradé et permettant l'analyse combinée d'un grand nombre d'échantillons historiques et modernes. L'ensemble de ces optimisations nous a permis de générer un total de 13 génomes historiques (à partir de spécimens d'herbier datant de 1845 à 1974) de *Xanthomonas citri* pv. *citri*.

Une des étapes de notre pipeline bioinformatique concerne la mesure des patrons de dégradation des acides nucléiques issus des spécimens historiques d'herbier. Dans ce contexte, nous avons spécifiquement réalisé une analyse statistique visant à tester l'influence de différents facteurs (protocoles moléculaires, âge des échantillons, type d'ADN...) sur les patrons de dégradations de l'ADN. Le détail de cette analyse, les résultats obtenus ainsi que leur interprétation sont présentés en fin de ce premier chapitre.

2.1. Optimisation des protocoles au laboratoire

Au début de cette thèse, la collection d'échantillons historiques d'agrumes au laboratoire renfermait plus de 300 spécimens symptomatiques d'herbier (voir chapitre 1.5. pour plus de détails). Afin d'identifier lesquels parmi ces échantillons comportaient de l'ADN exploitable du pathogène d'intérêt, *Xanthomonas citri* pv. *citri* (*Xci*), nous avons appliqué une approche de criblage par séquençage. Elle consiste en une extraction de l'ADN d'échantillons sur quelques lésions foliaires, transformée en librairies de séquençage, produisant des données informatiques de séquençage dites lectures, ou *reads*, parmi lesquelles la présence de *Xci* est recherchée. Les librairies où la présence de *Xci* (dit ADN endogène) s'est montrée suffisante (seuil que nous avons fixé à environ ~1% des lectures totales assignées à notre pathogène d'intérêt) ont alors été re-séquencées en profondeur dans le but de reconstruire le génome complet du pathogène. Cette approche nous a permis de nous affranchir d'un criblage par qPCR dont les polymérases, différentes de celles utilisées pour construire des librairies de séquençage, peuvent être inhibées par des composés végétaux (Schrader *et al.*, 2012) présents chez les *Citrus* (Hartung *et al.*, 1996; Li, Brlansky and Hartung, 2006), résultant en de faux-négatifs.

Les premiers échantillons que nous avons traités dans le cadre de notre étude ont été extraits au laboratoire selon un protocole dit « CTAB » (cétyltriméthylammonium bromure) (Ausubel *et al.*, 2003) modifié (Annexe 2.1.A). La construction de leurs librairies à partir d'extraits ADNs totaux et le séquençage 2x150 paired-end ont été effectués par la compagnie FASTERIS (<https://www.fasteris.com/dna/>) selon un protocole Illumina TruSeq Nano (https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_truseq/truseqnanodna/truseq-nano-dna-library-prep-guide-15041110-d.pdf, pages 11-26) modifié (sans étape de fragmentation, purification MinElute PCR Purification kit (Qiagen), pas d'autres précisions)), pour une quantité minimale de 50 ng d'ADN. Afin de mieux maîtriser la préparation des librairies, nous avons décidé d'appliquer des protocoles spécifiques à l'ADN ancien au sein de notre propre laboratoire.

Nous avons choisi le protocole de construction de librairies BEST (Blunt-End Single Tube) (Carøe *et al.*, 2018) 2.0 (Annexe 2.1.B, pages 1-4), développée pour l'ADN ancien et efficace à partir de quelques picogrammes d'ADN, avec la collaboration de Nathan Wales, un chercheur spécialiste des analyses moléculaires des ADN dégradés, reçu à cet effet au laboratoire. Brièvement, le protocole BEST permet la conversion d'ADN en librairie à partir de fragments double-brins à bouts francs. Une préparation des extrémités des fragments est faite grâce à la T4 DNA polymérase (tolérante des bases uraciles, lues alors comme des thymines) qui complète les bouts sortants en 5' et digère ceux en 3', les fragments sont alors liés à des adaptateurs double-brins qui permettent ensuite l'amplification des ADNs.

L'amplification et l'indexage des ADN se déroulent en une même étape grâce à des amorces porteuses des index avec une polymérase U-tolérante.

Nous avons testé différentes conditions de traitement de l'extrait ADN (dilution au 10^{ème}, purification à l'aide de billes magnétiques (Sera-mag Speed Beads, protocole BEST 2.0), pas de traitement) servant à la construction de librairie et mis en évidence que la purification d'ADN améliore significativement le rendement de conversion des extraits ADN en librairie, cette étape permettant probablement de débarrasser l'extrait d'inhibiteurs enzymatiques. Nous avons alors décidé de remplacer la dernière étape de précipitation de l'ADN à l'éthanol du protocole d'extraction CTAB par une purification directe du surnageant d'extraction sur billes magnétiques (Annexe 2.1.D). Tout en réduisant la durée du protocole, de l'extraction à la purification, de deux à un jour, nous avons montré que cette modification permettait d'extraire de l'ADN de qualité.

Au total, l'ADN de 42 échantillons historiques a été extrait au laboratoire dont 38 échantillons d'agrumes et 4 échantillons de caféier, ces derniers servant de témoins. Parmi les 38 échantillons d'agrumes, 22 présentaient des extraits d'ADN exploitables et ont été convertis en librairies (9 selon le protocole TruSeq Nano, 13 en BEST), ainsi que l'ensemble des échantillons de café (Annexe 2.1.E). Ces 26 librairies ont été séquencées en 2x150 paired-end Illumina. Au sein des librairies, 13 librairies issues d'échantillons de *Citrus* datant de 1845 à 1974 et principalement collectés en Asie, zone supposée de l'origine du pathogène, ont été testées positives pour *Xci* (voir Tableau 2.3.A). Une quatorzième librairie provenant d'un échantillon d'*Aegle marmelos*, un agrume de la famille des *Rutaceae* non *Citrus* collecté en 1867 en Inde, a également été testée positive à *Xci* avant de finalement se révéler être infectée par un autre pathogène: *Xanthomonas citri* pv. *bilvae* (après certaines analyses génétiques supplémentaires). *Xanthomonas citri* pv. *bilvae* est une bactérie appartenant à la même espèce que *Xci*, bien que génétiquement assez éloignée (Mhedbi-Hajri *et al.*, 2013; Patané *et al.*, 2019), précédemment uniquement décrite en Inde et connue pour être pathogène de *Rutaceae* non cultivées, bien que démontrée pathogène de plusieurs *Citrus* cultivés par inoculation (Patel, Allayyanavaramath and Kulkarni, 1953; Chakravarti *et al.*, 1984). Les données de cette librairie ont été exclues des analyses phylogénétiques de *Xci* présentées dans les chapitres 3 et 4.

Parallèlement à l'optimisation du protocole d'extraction ADN et sa conversion en librairie, l'équipe a mis au point un protocole de qPCR spécifique à *Xci* permettant son diagnostic sur échantillons récoltés au champ mais aussi sur ceux issus d'herbiers grâce à l'amplification d'une région cible de 58 pb. Le protocole comprend également un amplicon contrôle sur l'ADN de la plante hôte, validant l'obtention d'extraits ADN amplifiables par la polymérase utilisée. Cette méthode est hautement sensible et spécifique et pourrait permettre une méthode de criblage des échantillons d'herbier par qPCR plutôt

que par séquençage. Ces résultats ont fait preuve d'un article publié en 2020 dans la revue BMC Microbiology intitulé « ***Development and comparative validation of genomic-driven PCR-based assays to detect Xanthomonas citri pv. citri in citrus plants*** » (Annexe 2.1.F) pour lequel j'ai réalisé des analyses bioinformatiques de caractérisation de l'ADN *Xci* d'échantillons d'herbier.

2.2. Optimisation d'un pipeline bioinformatique

Les données de séquençage obtenues ont été traitées bioinformatiquement selon un pipeline mis au point durant mes travaux de thèse (portant tout d'abord sur un premier échantillon « preuve de concept » (chapitre 3) avant d'appliquer ce pipeline à l'ensemble des échantillons historiques générés (chapitre 4), Figure 2.2.A). Ce pipeline, après de premières étapes de contrôle qualité et de sélection des lectures, permet de réaliser 1) une estimation de la composition taxonomique des ADNs issus d'échantillons d'herbier, 2) une analyse des patrons de dégradation *post-mortem* de l'ADN du pathogène d'intérêt *Xci*, 3) une reconstruction des génomes de *Xci* (en prenant en compte la dégradation de l'ADN) et 4) la reconstruction des relations phylogénétiques entre individus modernes et historiques et, lorsque les conditions requises sont rencontrées, différentes inférences de paramètres démographiques, génétiques et évolutifs.

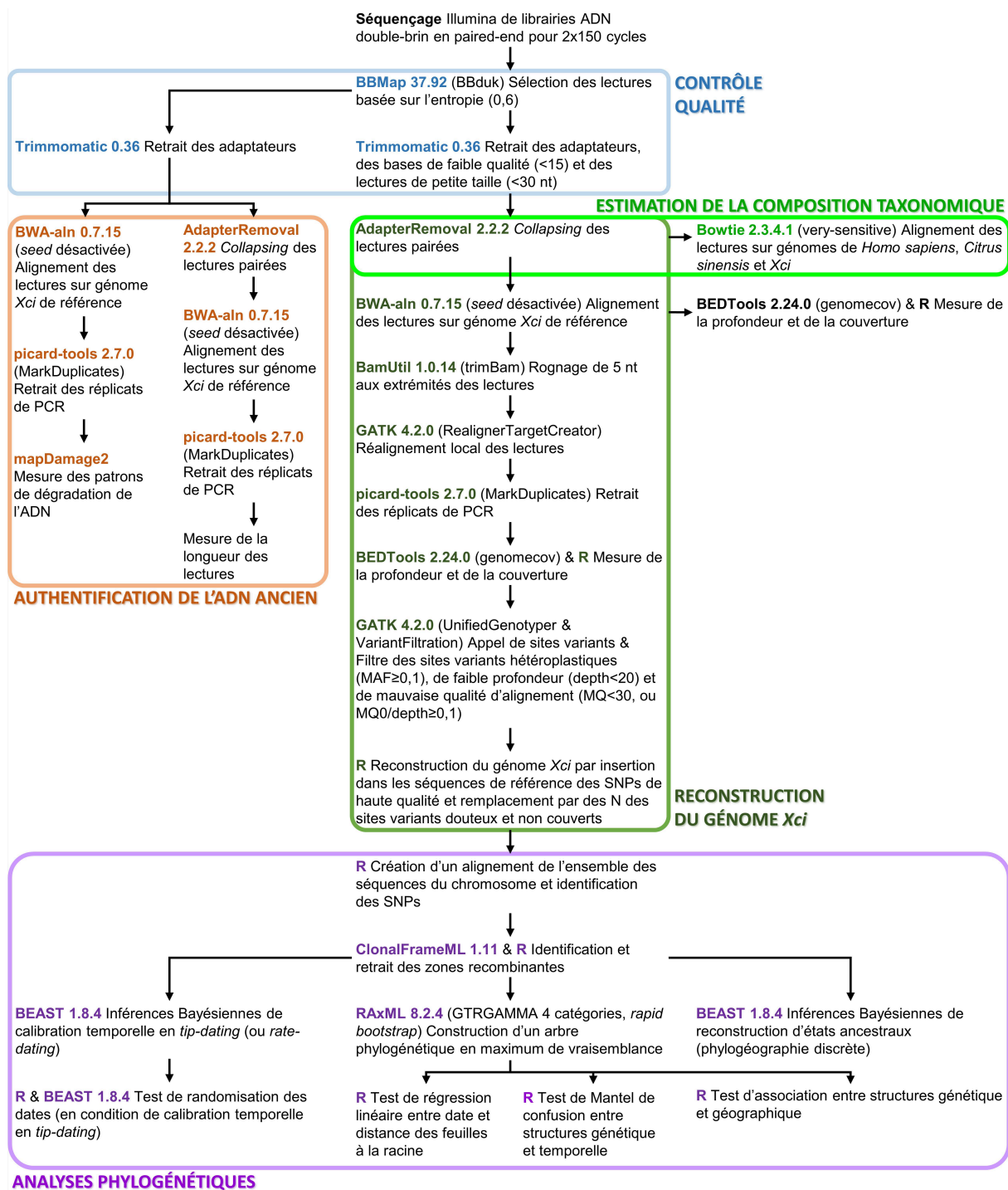


Figure 2.2.A. Pipeline bioinformatique utilisé incluant les bibliothèques issues d'échantillons d'herbier. Abréviations : nt, nucléotides ; *Xci*, *Xanthomonas citri* pv. *citri* ; PCR, *polymerase chain reaction* ; MQ, *mapping quality* ; SNPs, *single nucleotide polymorphisms*.

Ainsi, une première étape de sélection des lectures basée sur une mesure de l'entropie (degré de complexité de la séquence, valeur fixée à 0,6) est effectuée. Elle permet d'exclure les lectures composées majoritairement d'homopolymères causés par un « bégaiement » du séquenceur arrivé à l'extrémité de la lecture mais pas à la fin du nombre de cycle (et donc de nucléotides) demandé (150

cycles pour 150 nt lus sur la longueur totale des ADNs dans notre cas). Suite à cela, les adaptateurs Illumina sont retirés des lectures.

Un jeu de données est alors élaboré depuis ces lectures afin d'analyser les patrons de dégradation *post-mortem* de l'ADN. Telles quelles, elles permettent de mesurer le taux de désamination aux extrémités des lectures avec mapDamage2 (Jónsson *et al.*, 2013). Après fusion des lectures pairées (*collapsing* de la lecture issue du *run* R1 avec celle complémentaire dans le R2 en un insert unique, état des lectures correspondant le plus aux fragments ADNs initiaux), la distribution de la taille des inserts est mesurée.

En repartant des lectures dont les adaptateurs ont été retirés, un rognage additionnel, selon la qualité d'appel des bases lors du séquençage (valeur de 15), ainsi qu'une sélection sur une longueur minimale (30 nt), sont réalisés, assurant ainsi un jeu de données de bonne qualité des lectures avec des tailles permettant un alignement plus spécifique. Ces lectures sont également fusionnées par *collapsing* et constituent le jeu de données principal, pour les analyses d'estimation de la composition taxonomique, de reconstruction du génome et de phylogénie.

L'estimation de la composition taxonomique est réalisée par un alignement successif avec Bowtie 2 (Langmead and Salzberg, 2012) des lectures fusionnées sur le génome humain (GCF_000001405.39), le génome de *Citrus sinensis* (AJPS000000000.1) ainsi que le génome de *Xci* (souche de référence IAPAR 306, chromosome NC_003919.1, plasmides pXAC33 NC_003921.3 et pXAC64 NC_003922.1). Bowtie 2, de moindre stringence que BWA-aln, permet d'obtenir un aperçu plus rapide des proportions de lectures correspondant aux différents taxons étudiés lorsque le but n'est pas d'en reconstruire le génome.

La reconstruction du génome *Xci* est réalisée suite à l'alignement des lectures aux trois séquences *Xci* avec BWA-aln 0.7.15 (Li and Durbin, 2009) pour les échantillons d'herbier, dont les lectures sont relativement de petites tailles, et Bowtie 2 (Langmead and Salzberg, 2012) pour les échantillons modernes caractérisés par des lectures de plus grandes tailles. Une première estimation de la profondeur et de la couverture des séquences est effectuée. Les fichiers d'alignement des échantillons historiques sont alors modifiés pour rogner un certain nombre de nucléotides (5) à chaque extrémité des lectures, ce qui permet de masquer les nucléotides où de la désamination est observée et qui pourrait aboutir, dans le cas où ils ne sont pas masqués, à de l'appel de faux sites variants, dû à la dégradation de l'ADN et non pas de la mutation. La profondeur est alors recalculée, cette deuxième valeur sert ainsi lors des filtres qualité des SNPs. Les sites variants sont considérés douteux et exclus de la reconstruction des séquences dès qu'ils remplissent l'une des conditions de filtres suivantes :

profondeur < « profondeur moyenne sur l'ensemble du génome + 1 écart type », fréquence de l'allèle minoritaire $\geq 0,1$, qualité de l'alignement < 30 . La séquence reconstruite est obtenue en introduisant dans la séquence de référence les SNPs de haute qualité et en remplaçant les sites variants douteux et les sites non couverts par des N.

Une fois les séquences chromosomiques obtenues pour les 13 échantillons d'herbier mais également un ensemble de 361 souches modernes (Figure 2.2.B), un alignement des séquences du chromosome des souches historiques et modernes de *Xci* est construit, en incluant des échantillons *Xanthomonas* non *Xci* formant l'*outgroup* (Annexe 2.2.A). L'identification des SNPs est réalisée au sein de l'ingroup, échelle à laquelle les zones issues de transferts horizontaux sont identifiées à l'aide du logiciel ClonalFrameML (Didelot and Wilson, 2015) et exclues des analyses suivantes afin de prendre en compte la recombinaison lors des analyses phylogénétiques, basées sur l'héritabilité des caractères. Un arbre phylogénétique en maximum de vraisemblance (ML, *maximum likelihood*) est alors construit avec un modèle d'évolution des séquences *General Time-Reversible* (Tavaré and Miura, 1986) pour une distribution gamma avec quatre catégories de taux de substitutions. Cet arbre ML est ensuite utilisé pour tester la présence de signal temporel (accumulation progressive de mutations au cours du temps) et phylogénétique associé à la géographie, conditions requises pour pouvoir calibrer temporellement un arbre phylogénétique et en reconstruire les états ancestraux de localisation, respectivement (voir méthodes présentées en chapitre 1.2.2). Lorsque cela est possible, une telle calibration spatio-temporelle visant à reconstruire l'histoire évolutive des lignées portant ces signaux a été réalisée à l'aide du logiciel BEAST (Drummond and Rambaut, 2007).

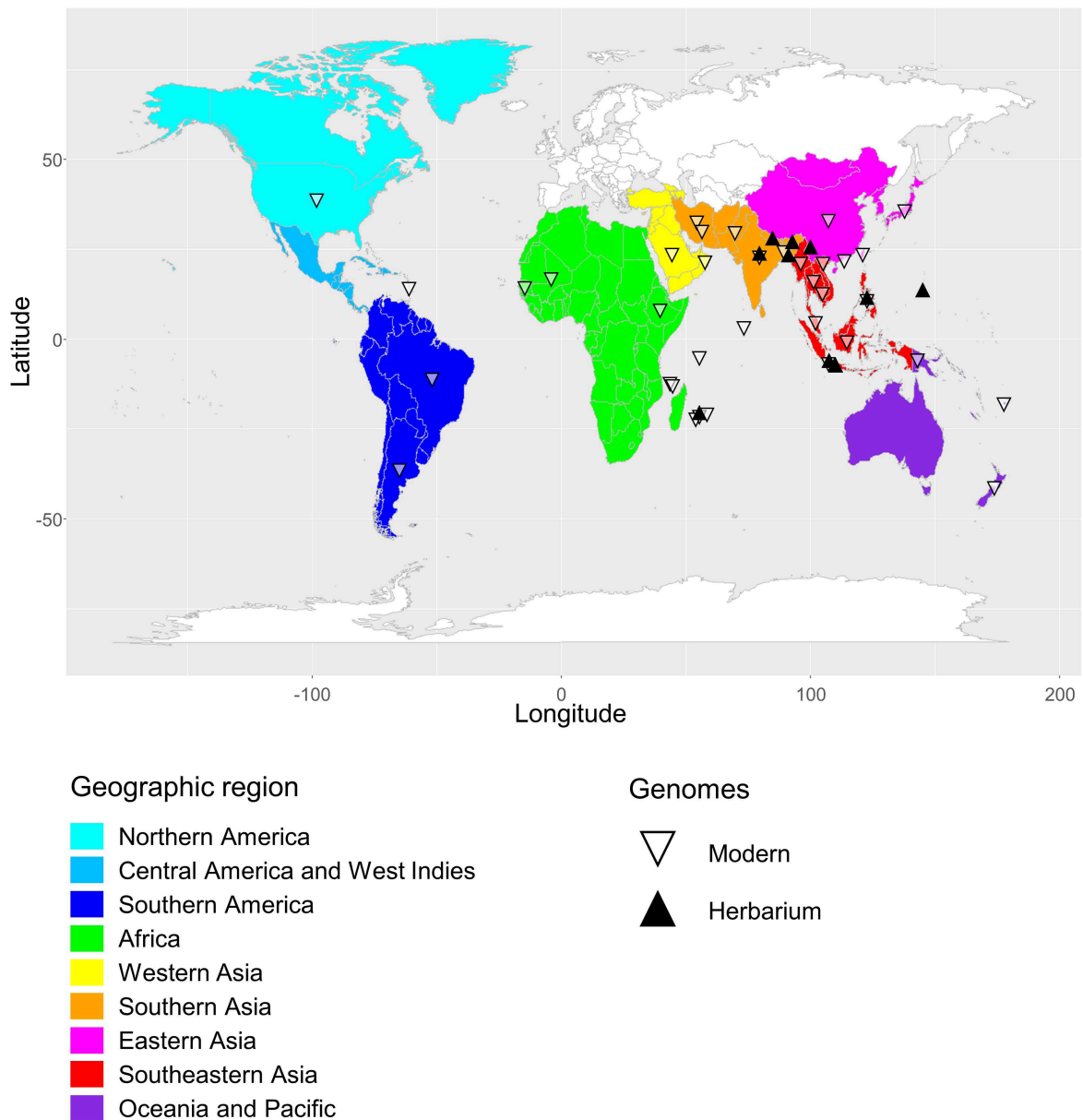


Figure 2.2.B. Distribution géographique des génomes historiques (issus d'herbiers) et modernes (issus de cultures bactériennes) de *Xci* utilisés durant la thèse. Les génomes modernes (triangles noirs vides inversés) et historiques (triangles noirs pleins) ont été placés au centroïde du pays ou de la localité d'origine, si celle-ci est renseignée. 28 génomes modernes ont été générés à partir de souches provenant du continent américain (1976-2017), 232 d'Afrique (incluant les îles du sud de l'océan Indien) (1978-2017), 95 d'Asie (1948-2017) et 6 d'Océanie et Pacifique (1949-1977). Parmi les 13 génomes historiques (1845-1974), 10 ont été générés à partir d'échantillons prélevés en Asie (4 en Asie du sud, 1 en Asie de l'est, 5 en Asie du sud-est), 2 en Afrique (île Maurice) et 1 en Océanie et Pacifique.

2.3. Caractérisation des échantillons et génération des génomes historiques

Nous avons appliqué notre pipeline bioinformatique sur les 13 librairies issues d'échantillons de *Citrus* testées positives pour la présence de *Xci*. Nous avons alors estimé une longueur moyenne des inserts de ces librairies comprise entre 42 ± 13 et 103 ± 45 nt (Tableau 2.3.A). Comme attendu d'ADNs issus de

lésions de feuilles, nous avons observé une source multiple d'ADN avec majoritairement de l'ADN de la plante hôte (entre 8,7 et 37,8% des ADNs totaux), de l'ADN du pathogène (entre 0,8 et 27,1%) mais également de l'ADN humain (entre 0,3 et 3,5%) (Tableau 2.3.A) qui pourrait provenir de la manipulation de la planche d'herbier au cours du temps.

Nous avons généré 13 génomes historiques de *Xci* quasi-complets. Sur le chromosome et pour une profondeur de 1X, la séquence est couverte entre 94,6 et 98,3% de sa longueur totale, avec des profondeurs moyennes allant de 6,2 à 96,2X; les deux plasmides, pXAC33 et pXAC64, présentent des couvertures à 1X plus faibles (49,7 à 97,1%) et des profondeurs moyennes plus élevées (17,9 à 130,3X). Sur la séquence du chromosome, nous avons identifié entre 122 et 2334 SNPs de haute qualité entre la séquence de référence IAPAR 306 et les séquences reconstruites des 13 échantillons d'herbier (Tableau 2.3.A).

Tableau 2.3.A. Caractéristiques générales de reconstruction et dégradation des génomes *Xci* issus de 13 échantillons historiques d'herbier. Statistiques d'alignement, couverture et profondeur moyenne ainsi que de dégradation (longueur et taux de désamination à l'extrémité des fragments d'ADN) du chromosome *Xci* et des plasmides (pXAC33 et pXAC64) pour chacun des 13 échantillons d'herbier. Abréviations : MNHN, Muséum national d'Histoire naturelle ; RB Gardens, Royal Botanical Gardens; NFC USDA, National Fungus Collection ; Chr, chromosome; nt, nucléotides ; SD, erreur standard.

ID	ID Herbarier	Herbier	Date	Localisation	Hôte	Protocole	N reads (en millions)	Longueur moyenne des inserts \pm SD	% reads Homosapiens	% reads Citrus	% reads Xci	Profondeur moyenne				Couverture à profondeur de 1X (%)				SNPs de haute qualité				Longueur moyenne des fragments \pm SD (nt)				Taux de désamination à l'extrémité des reads R1 (%)			
												Chr m	pXAC 33	pXAC 64	Chr m	pXA C33	pXA C64	sur Chr m	Chr m	pXAC 33	pXAC 64	Chr m	pXAC 33	pXAC 64	Chr m	pXAC 33	pXAC 64				
HERB_1845	P0529798	MNHN Paris	1845	Indonesia, Java	<i>Citrus aurantiifolia</i>	TruSeq Nano	414,3	57,5 \pm 28,4	0,7	21,7	10,4	82,1	122,2	120,4	98,3	93,3	95,7	149	50,4 \pm 23,0	52,2 \pm 22,8	51,9 \pm 22,6	3,68	3,91	3,90	3,90						
HERB_1884	1206	RB Gardens	1884	Philippines, Luzon	<i>Citrus medica</i>	TruSeq Nano	246,8	51,8 \pm 23,1	1,1	25,2	10,5	55,6	92,1	94,6	98,1	89,8	93,2	275	46,2 \pm 16,6	48,4 \pm 17,0	48,4 \pm 16,9	3,22	3,38	3,28	3,28						
HERB_1911	P0529799	MNHN Paris	1911	Indonesia, Java	<i>Citrus aurantiifolia</i>	TruSeq Nano	365,2	88,4 \pm 32,3	3,5	37,8	2,1	32,4	69,2	65,2	98,2	90,2	90,2	330	69,1 \pm 22,3	69,7 \pm 21,9	69,4 \pm 21,8	3,70	3,42	3,78	3,78						
HERB_1915	P0529799	MNHN Paris	1915	Philippines	<i>Citrus lime</i>	TruSeq Nano	217,3	51,1 \pm 22,2	1,3	27,0	6,0	39,3	66,2	65,8	98,1	88,0	92,9	293	47,9 \pm 17,9	48,5 \pm 17,1	48,7 \pm 17,1	2,81	2,93	2,98	2,98						
HERB_1937	MAU0015	Mauritius	1937	Mauritius	<i>Citrus</i> sp.	TruSeq Nano	220,9	65,5 \pm 32,9	0,4	12,5	0,8	6,2	22,6	17,9	94,6	82,9	88,5	122	42,7 \pm 12,7	44,7 \pm 13,8	44,6 \pm 13,6	3,65	3,99	4,09	4,09						
HERB_1946	USDA686	NFC USDA	1946	Guam, Tlofofo	<i>Citrus</i> sp.	TruSeq Nano	262,5	63,2 \pm 33,3	0,7	25,6	7,5	64,3	114,9	108,1	98,2	94,7	93,2	331	57,9 \pm 29,5	58,3 \pm 28,1	58,4 \pm 28,2	1,81	1,98	2,01	2,01						
HERB_1974	MAU0015	Mauritius	1974	Mauritius	<i>Citrus lime</i>	TruSeq Nano	260,8	42,1 \pm 12,8	0,7	19,1	6,8	35,9	42,6	25,5	97,9	82,5	49,7	316	41,1 \pm 9,3	41,5 \pm 9,5	41,6 \pm 9,5	2,09	2,46	2,56	2,56						
HERB_1852	Q1874	RB Gardens	1852	India, Khasi Hills	<i>Citrus medica</i>	BEST	314,9	75,8 \pm 38,7	0,7	32,8	4,8	63,7	130,3	117,9	97,8	95,9	96,5	1496	70,9 \pm 43,6	70,4 \pm 41,4	69,7 \pm 41,1	1,89	2,13	2,17	2,17						
HERB_1854	Q1954	RB Gardens	1854	Indonesia, Java	<i>Citrus aurantiifolia</i>	BEST	113,0	81,9 \pm 41,8	0,9	25,3	10,8	54,1	114,0	108,6	98,3	95,2	97,1	150	75,7 \pm 42,8	72,0 \pm 39,6	72,3 \pm 39,7	1,22	1,35	1,40	1,40						
HERB_1859	P0524071	MNHN Paris	1859	Bangladesh, Chittagong	<i>Citrus medica</i>	BEST	159,5	78,4 \pm 34,0	0,8	33,3	5,3	41,8	84,1	75,4	97,7	94,2	95,3	1466	73,9 \pm 33,8	71,7 \pm 31,6	71,4 \pm 31,4	1,76	1,93	2,03	2,03						
HERB_1865	Q1889	RB Gardens	1865	India	<i>Citrus medica</i>	BEST	156,4	102,9 \pm 45,1	0,4	37,5	4,9	49,8	95,1	101,5	98,2	77,9	96,6	309	88,1 \pm 39,2	85,1 \pm 36,8	85,1 \pm 36,7	1,57	1,81	1,81	1,81						
HERB_1922	1756364	US National	1922	China, Yunnan	<i>Citrus medica</i>	BEST	120,9	72,2 \pm 30,2	0,5	13,4	27,1	96,2	99,6	128,7	97,3	76,5	94,5	2334	67,9 \pm 29,3	67,7 \pm 29,2	67,2 \pm 28,5	1,22	1,62	1,54	1,54						
HERB_1963	630116	RB Gardens	1963	Nepal, Sanichare	<i>Citrus medica</i>	BEST	56,3	80,5 \pm 36,5	0,3	8,7	14,6	43,0	106,8	86,2	98,0	94,3	96,0	492	74,7 \pm 32,3	73,2 \pm 31,0	73,0 \pm 30,9	1,25	1,46	1,53	1,53						

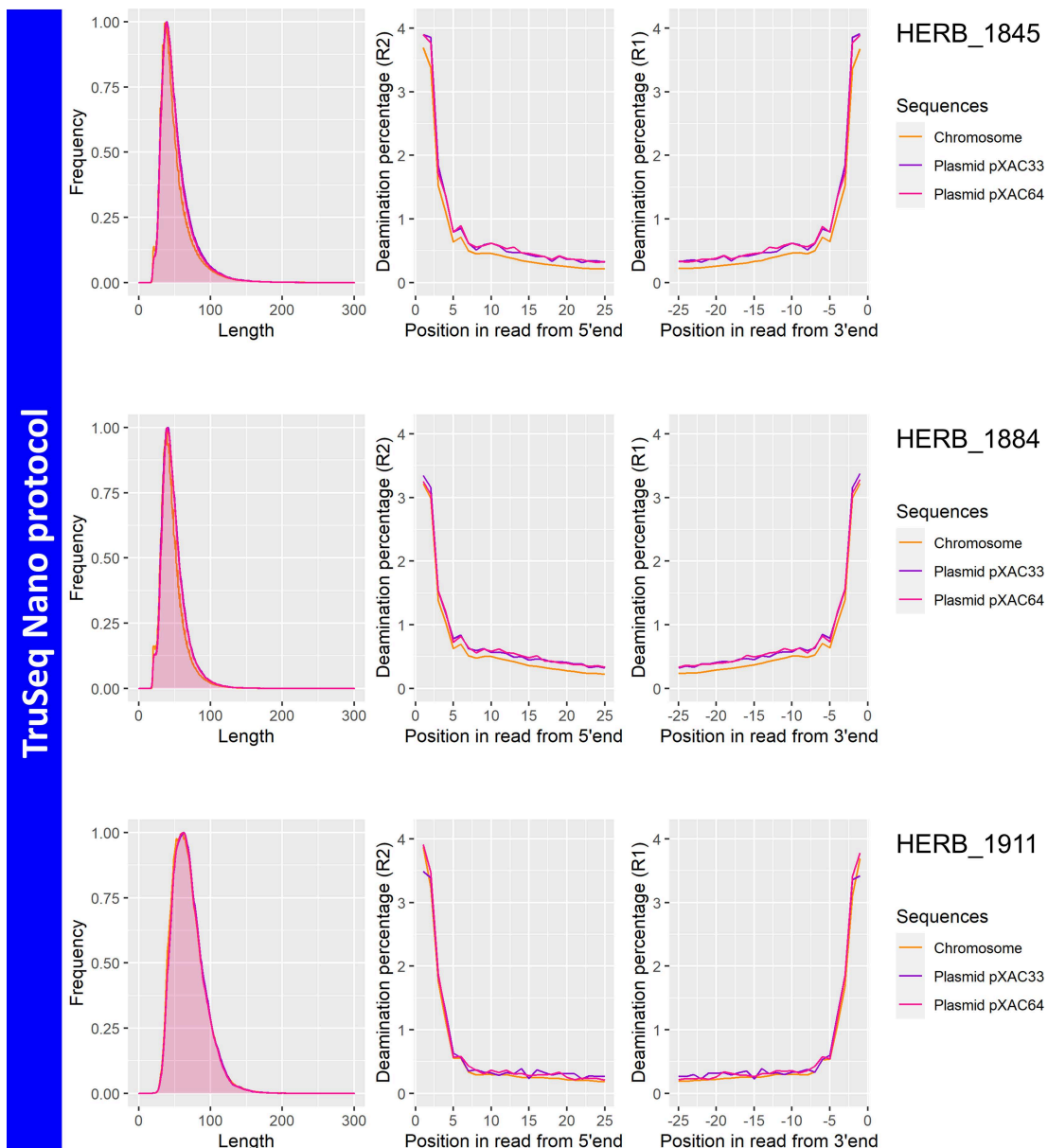
Dans le but d'estimer d'éventuelles contaminations entre échantillons, une librairie témoin issue d'un échantillon historique d'herbier d'une espèce de plante non hôte de *Xci*, le caféier (*Coffea arabica*), a été construite conjointement aux différentes séries de librairies. En mesurant la proportion d'ADN de *Coffea sp.*, de *Citrus sp.* et de *Xci* dans chaque librairie par alignement des lectures aux génomes de référence, nous pouvons tenter d'identifier de la contamination croisée. Les librairies témoins issues d'échantillons de *Coffea* présentent moins de 0,01% d'ADN de *Xci*, et entre 0,4 et 0,6% d'ADN de *Citrus sp.* contre 90,7 à 92,5% d'ADN de *Coffea sp.* alors que les librairies issues de *Citrus* ne présentent que 1,1 à 2,4% d'ADN de *Coffea sp.* pour les 8,7 à 37,8% d'ADN de *Citrus sp.* énoncés plus haut. Cela nous a permis d'exclure la présence de contamination importante entre librairies *Citrus* et témoins.

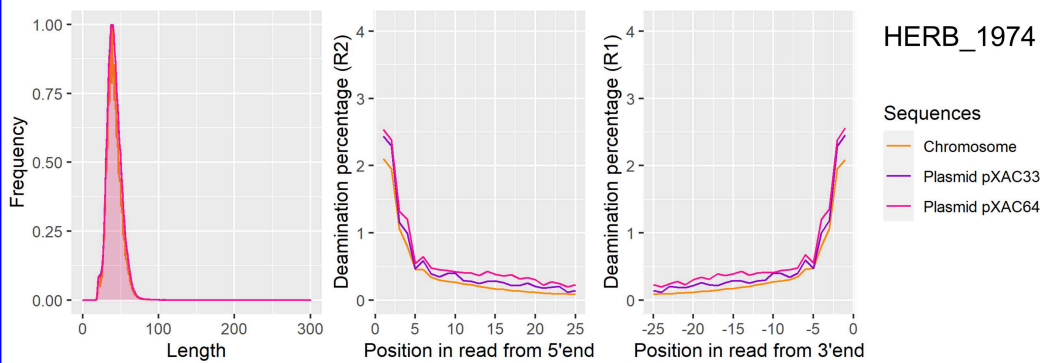
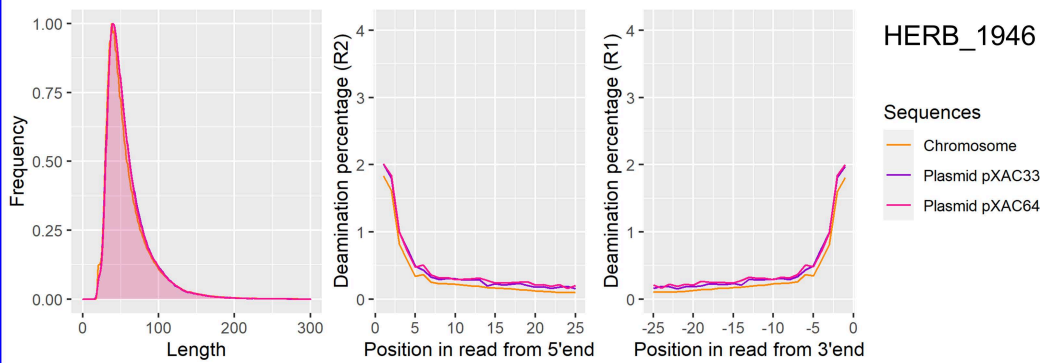
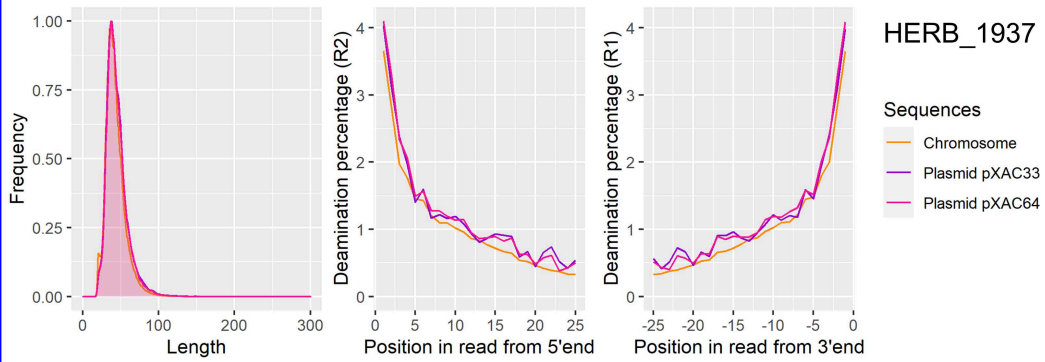
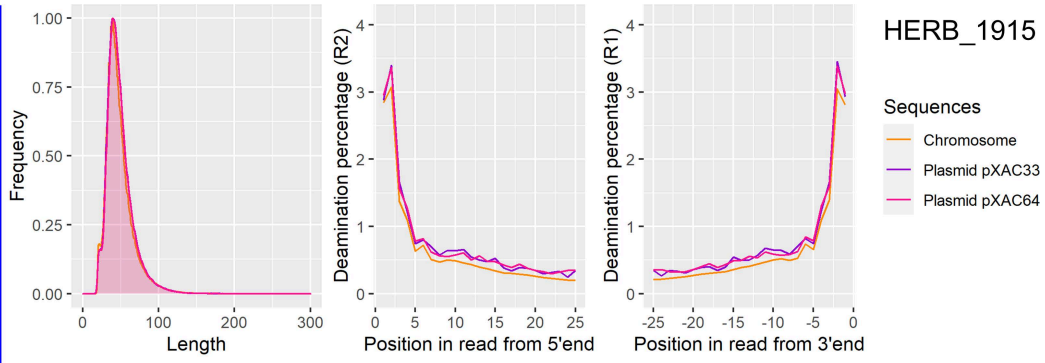
Une analyse sur la proportion de sites hétéroplastiques (site où l'allèle minoritaire est de fréquence supérieure à 0,1) à l'emplacement des sites variables identifiés sur le génome de *Xci* a également été réalisée afin de tester l'existence de contamination des librairies *Citrus* entre elles. Parmi les 13 génomes *Xci*, trois d'entre eux (HERB_1845, HERB_1852 et HERB_1937) présentent un taux relativement important de sites hétéroplastiques sur sites variants (entre 24,7 et 43,1%). Cependant, moins de 12,5% au maximum de ces sites sont partagés et correspondent à des sites variables chez d'autres échantillons manipulés simultanément, ce qui ne permet pas d'identifier une source sûre de contamination entre librairies. Si la mesure de l'hétéroplasmie permet de mettre en évidence une source multiple d'ADN *Xci*, il n'est pas à exclure que cette source provient des différentes lésions nécessaires à une même extraction ADN. En effet, des infections *Xci* au champ ont été montrées polyclonales à l'échelle des feuilles d'une branche et même de la lésion (Pruvost *et al.*, 2019). Nous avons alors décidé d'exclure les sites hétéroplastiques lors de la prise en compte de sites variables afin de ne conserver que des SNPs de haute qualité dans les génomes reconstruits.

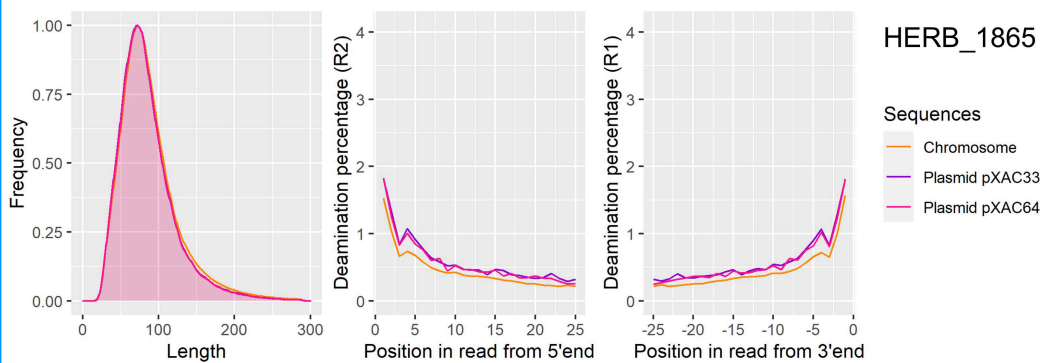
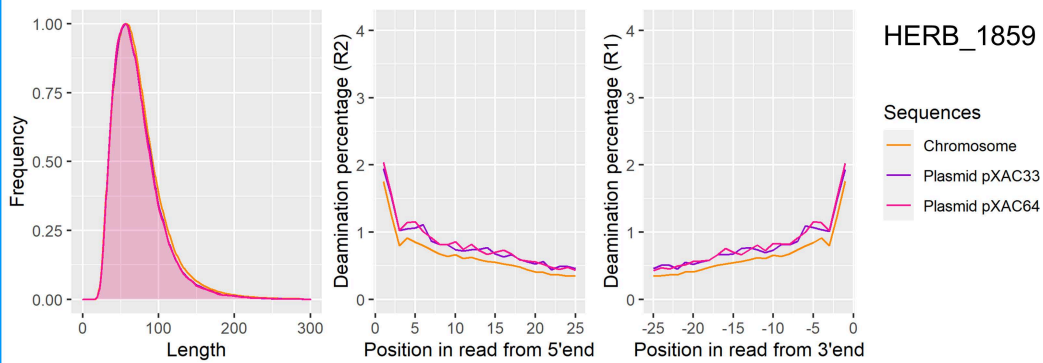
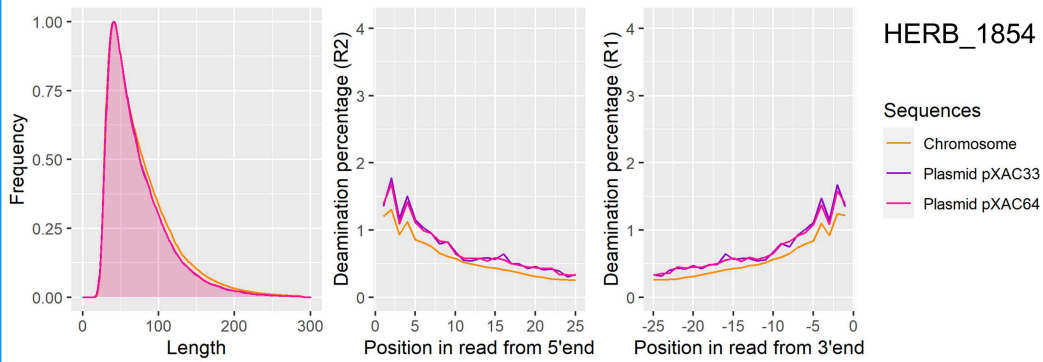
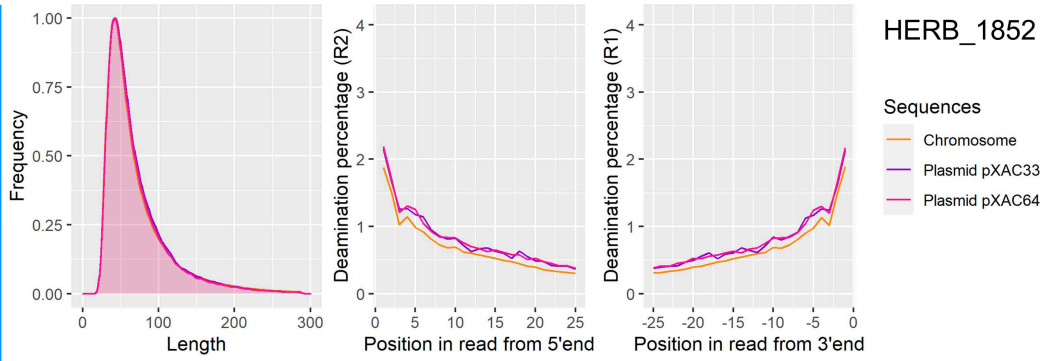
Comme décrit en chapitre 1.3.4, l'ADN ancien est typiquement dégradé, présentant des fragments de petites tailles dont les extrémités sont marquées par la désamination de leurs cytosines C en uraciles U (Pääbo *et al.*, 2004; Dabney, Meyer and Pääbo, 2013) lues comme des thymines T par les enzymes U-tolérantes. Ces désaminations auraient lieu plus rapidement sur les extrémités ADN simple-brins que celles double-brins, avec ainsi une accumulation sur les positions les plus externes (Briggs *et al.*, 2007). Elles peuvent être mesurées après alignement des lectures sur séquence de référence avec des outils comme mapDamage2 (Jónsson *et al.*, 2013). Ces différents patrons sont alors utilisés pour authentifier les ADNs anciens (Rohland *et al.*, 2015; Kistler *et al.*, 2017).

Après une étude fine des différentes sorties de mapDamage2 (Jónsson *et al.*, 2013), nous avons choisi d'analyser les débuts des lectures en R1 (avec vérification sur lectures complémentaires en R2) (Kistler *et al.*, 2017), qui nous donnaient les résultats les plus fiables. Pour les échantillons issus du protocole

TruSeq Nano, nous avons éliminé les données de la position 1 car siège de nombreuses autres substitutions, et analysé les données à partir de la position 2. Nous avons alors identifié que les ADNs *Xci* issus des 13 échantillons d'herbier (1845-1974) mesuraient entre 20 et 289 nt (cette dernière étant la limite de taille pour le *collapsing* des lectures), avec des moyennes de $41,1 \pm 9,3$ à $88,1 \pm 39,2$ nt et présentaient des signes de désamination (taux entre 1,2 et 3,7% à la position externe en 3' pour le chromosome, entre 1,4 et 4,1% pour les plasmides). Ces valeurs sont compatibles avec celles d'un oomycète phytopathogène échantillonné au XIX^{ème} siècle (Martin *et al.*, 2013; Yoshida *et al.*, 2013; Weiß *et al.*, 2016) et différent de celles que nous avons obtenues sur souches modernes (taille des fragments dictée par la longueur de séquençage demandée (150 nt) et taux maximal de 0,2%). Ces analyses de patrons de dégradation de l'ADN ont permis d'authentifier les différents génomes de *Xci* issus d'échantillons d'herbier comme historiques (Figure 2.3.A & Tableau 2.3.A).







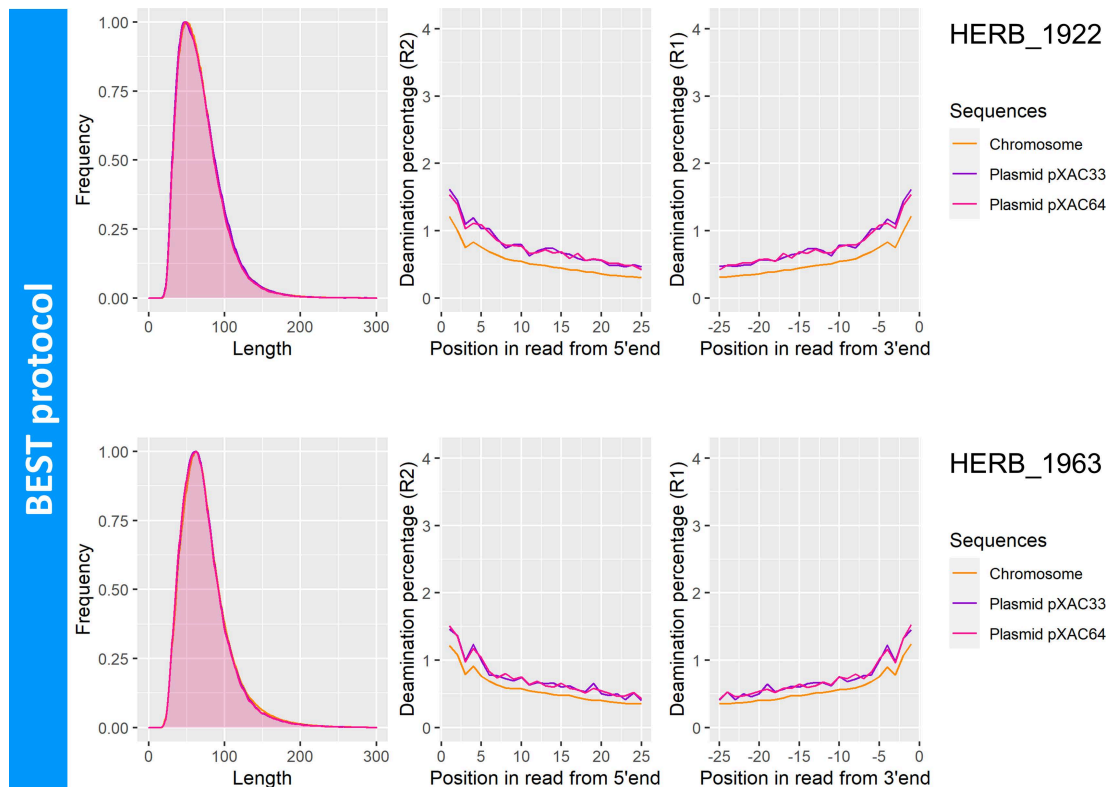


Figure 2.3.A. Patterns de dégradation *post-mortem* des 13 génomes *Xci* issus d'échantillons historiques d'herbier. Les patrons de dégradation *post-mortem* de l'ADN ont été mesurés sur le génome *Xci* (chromosome, plasmides pXAC33 et pXAC64 en jaune, violet et rose respectivement), avec les sept premiers échantillons obtenus selon le protocole de construction de bibliothèques TruSeq Nano et les six derniers avec le protocole BEST. (Gauche) Distribution de la longueur des fragments ADN (en nucléotides, fréquence relative en unité arbitraire). (Milieu et Droite) Taux de désamination aux 25 nucléotides des débuts des lectures (issues des R1 et R2), correspondant aux pourcentages de substitutions C vers T.

2.4. Analyses des patrons de dégradation de l'ADN

Weiß *et al.* (2016) ont observé, dans leur étude sur 71 échantillons de plantes historiques (datant de 1737 à 1993), un effet significatif de l'âge des échantillons sur le taux de dégradation de l'ADN (fragmentation et désamination), ce qui leur a permis d'estimer un taux de dégradation des spécimens d'herbier. Ces mêmes auteurs ont également observé un effet significatif du type d'ADN (nucléaire ou chloroplastique) sur l'intensité de la désamination aux extrémités des fragments qui pourrait être imputée à la structure circulaire de l'ADN chloroplastique, qui le rendrait moins accessible aux endonucléases que l'ADN nucléaire.

Au vu de cette étude et des profils de dégradation quantifiés dans notre jeu de données (Figure 2.3.A), nous avons examiné l'effet de trois paramètres sur la fragmentation et la désamination de l'ADN :

- 1) les conditions de protocole (condition « TruSeq Nano » avec protocoles d'extraction CTAB modifié, purification selon le kit MinElute PCR Purification (Qiagen) par FASTERIS, construction

de librairies TruSeq Nano; condition « BEST » avec protocoles d'extraction et purification CTAB et billes magnétiques, et de construction de librairies BEST 2.0) ;

- 2) le type d'ADN (chromosomique ou plasmidique) ;
- 3) l'âge (échantillons datant de 1845 à 1974, correspondant à des âges de 173 à 44 ans).

Les analyses statistiques présentées ci-dessous ont été réalisées sous R (version 2021) en collaboration avec Frédéric Chiroleu et Thuy Trang Cao, deux statisticiens de notre unité de recherche. Pour identifier la présence potentielle de certains de ces effets sur la distribution de tailles de fragments, 11 catégories de tailles régulières ont été déterminées (de 15 à 290 nt, tous les 25 nucléotides) et la contribution (nombre de fragments) à chacune de ces catégories a été mesurée. Une analyse factorielle des correspondances a été effectuée avec la fonction CA du paquet « FactoMineR » (Lê, Josse and Husson, 2008) afin de résumer le jeu de données à ses composantes principales et visualiser l'association entre individus et variables. Un test de χ^2 a été effectué et l'hypothèse d'indépendance entre les individus et les variables (catégories de tailles et âge) a été rejetée ($\chi^2=17\ 920\ 863$, $p\text{-value}<0,0001$), suggérant qu'il existait bien une association significative entre (certains) éléments lignes et (certains) éléments colonnes du jeu de données. Sur le plan à deux dimensions expliquant 98% de la variance totale du jeu de données (Figure 2.4.A), les éléments lignes sont groupés par individu dont ils sont issus, séquences chromosomique et plasmidiques présentant des profils similaires entre eux. Une dispersion différente des éléments lignes apparaît selon le protocole utilisé. Les éléments lignes de six des sept individus obtenus en conditions « TruSeq Nano » se trouvent du côté des éléments colonnes des plus petites catégories de tailles de fragments (15 à 40 nt et 41 à 65 nt) auxquelles ils s'associent le plus. Ils sont séparés par l'axe de la dimension 1 des autres éléments lignes (des six individus obtenus en conditions « BEST » et du septième individu en conditions « TruSeq Nano ») et colonnes (catégories de tailles de 66 à 290 nt et l'âge), avec ces éléments lignes situés proches des éléments colonnes des catégories de tailles de fragments de 66 à 140 nt. Parmi les éléments colonnes, celui « Âge » a une mauvaise projection (somme des cosinus² sur les dimensions 1 et 2 égale à 0,01) et son association par rapport aux autres éléments ne peut être interprétée.

Figure 2.4.A. Analyse factorielle des correspondances des catégories de tailles de fragments des génomes *Xci* issus de 13 échantillons historiques. Les modalités lignes (chromosome, plasmides pXAC33 et pXAC64 des 13 échantillons, respectivement symboles de grande, petite et moyenne taille), distinguées selon leur protocole (« TruSeq Nano » et « BEST », cercle et triangle respectivement, ellipses comprenant 95% de la dispersion de leurs points respectifs) et âge (gradient de bleu) sont projetées concomitamment aux modalités colonnes (catégories de tailles de fragments ADN et âge, croix rouges) sur un plan à deux dimensions expliquant 98,31% de la variance totale du jeu de données; leur qualité de projection au sein du plan à deux dimensions (\cos^2) est représentée par l'opacité des points.

Les effets « protocole », « type d'ADN » et « âge » ont été testés par une analyse de variance sur les catégories de tailles de fragments comprises entre 15 à 140 nt (contribution à chaque catégorie transformée en pourcentage de fragments) avec la fonction aov du paquet R de base « stats ». Lorsqu'une différence significative entre groupes apparaissait, un test de Student a été effectué avec la fonction t.test afin d'identifier en quel sens les groupes diffèrent. Ainsi, nous avons observé une différence significative de la contribution aux différentes catégories de tailles de fragments selon le protocole (p -value $<0,0001$ pour chaque catégorie analysée), avec les individus obtenus en conditions « TruSeq Nano » ayant en moyenne plus de fragments de 15 à 40 et de 41 à 65 nt que les individus obtenus en conditions « BEST » (p -value $<0,0001$ dans les deux cas), ces derniers ont en moyenne significativement plus de fragments de 66 à 140 nt que les premiers (p -value $<0,0001$ dans les trois cas). Le jeu de données a alors été traité séparément, selon les conditions de protocole utilisées, dans la suite des analyses. Pour chacune des deux conditions de protocole, aucune différence significative de la contribution aux différentes catégories de tailles par rapport au type de séquences n'a été observée (p -value de 0,924 à 0,997 pour le jeu de données « TruSeq Nano », 0,616 à 0,997 pour celui « BEST »). Il n'y avait pas non plus de différence significative quant à la contribution aux différentes catégories de tailles de fragments selon l'âge des échantillons (« TruSeq Nano », p -value de 0,288 à 0,908 ; « BEST », p -value de 0,051 à 0,909).

L'analyse du taux de désamination a été réalisée sur les données de l'extrémité 3' (obtenues depuis le R1, nombre de substitutions G vers A par rapport aux nombres de G dans la séquence de référence) à la position la plus externe. L'existence d'un effet entre le taux de désamination et les variables « protocole » et « âge » a été testée avec un modèle linéaire généralisé en utilisant la fonction glm du paquet de base « stats » de R et selon une distribution quasi-binomiale afin de prendre en compte la surdispersion significative (ratio du paramètre de dispersion >1 et p -value $<0,0001$) des points. La variable « type d'ADN » a été testée avec la fonction glmmPQL selon une distribution quasi-binomiale (modèle linéaire généralisé mixte avec quasi-vraisemblance pénalisée) afin de prendre en compte l'effet individu (valeurs d'un même individu appariées). Ainsi, un effet significatif du protocole a été identifié (p -value $<0,0001$, Figure 2.4.B) et les

les données ont été par la suite traitées séparément selon le protocole utilisé. Un effet significatif de l'âge sur le taux de désamination a été observé au sein de chacun de ces jeux de données (p -value $<0,0001$ dans les deux cas, Figure 2.4.B), avec des taux de désamination plus élevés mesurés chez les échantillons les plus âgés.

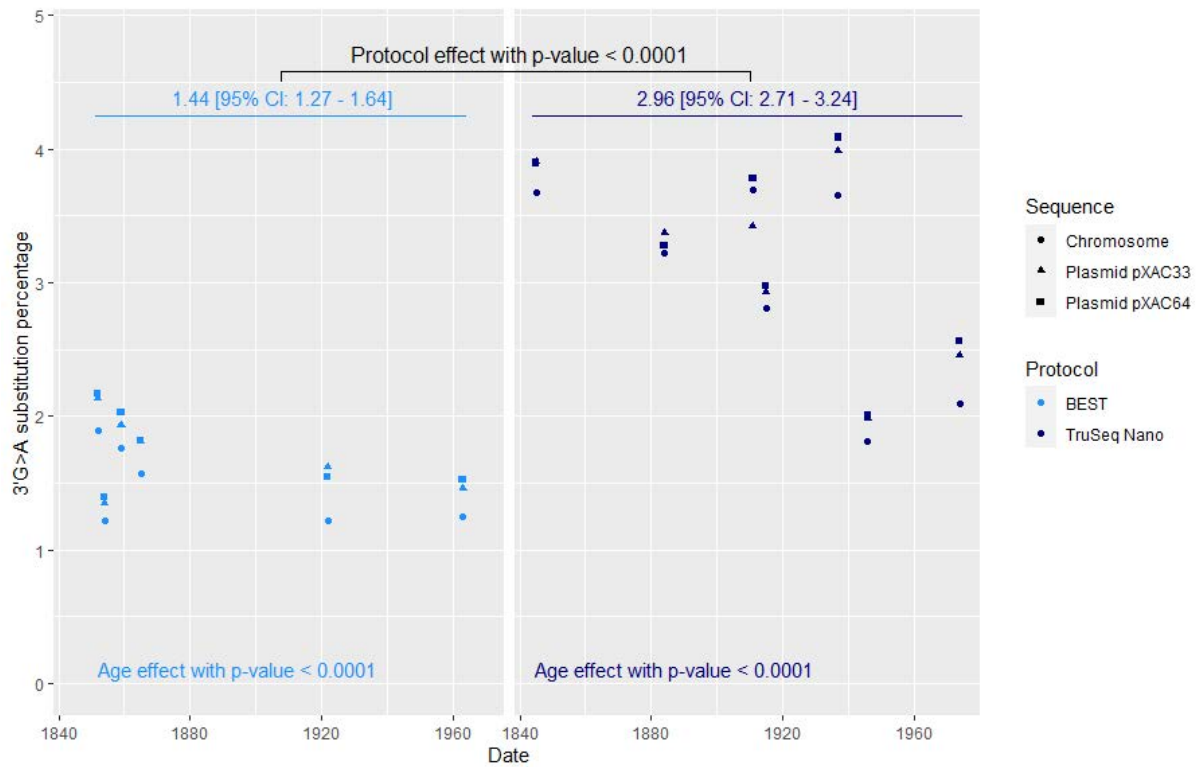


Figure 2.4.B. Pourcentage de désamination à la dernière position de l'extrémité 3' des génomes *Xci* issus de 13 échantillons historiques. Les génomes *Xci* (chromosome, plasmides pXAC33 et pXAC64, respectivement, cercle, triangle et carré) des 13 échantillons historiques sont distingués selon le protocole de construction de bibliothèques utilisé (BEST en bleu clair, TruSeq Nano en bleu foncé).

Enfin, une différence significative du taux de désamination par rapport au type d'ADN a également été observée entre le chromosome et chacun des plasmides (BEST avec p -value $<0,0001$ dans les deux cas, TruSeq Nano avec p -value=0,0259 entre chromosome et plasmide pXAC33 et 0,0046 entre chromosome et plasmide pXAC64, Figure 2.4.C). Aucune différence significative n'a été mesurée entre les deux plasmides.

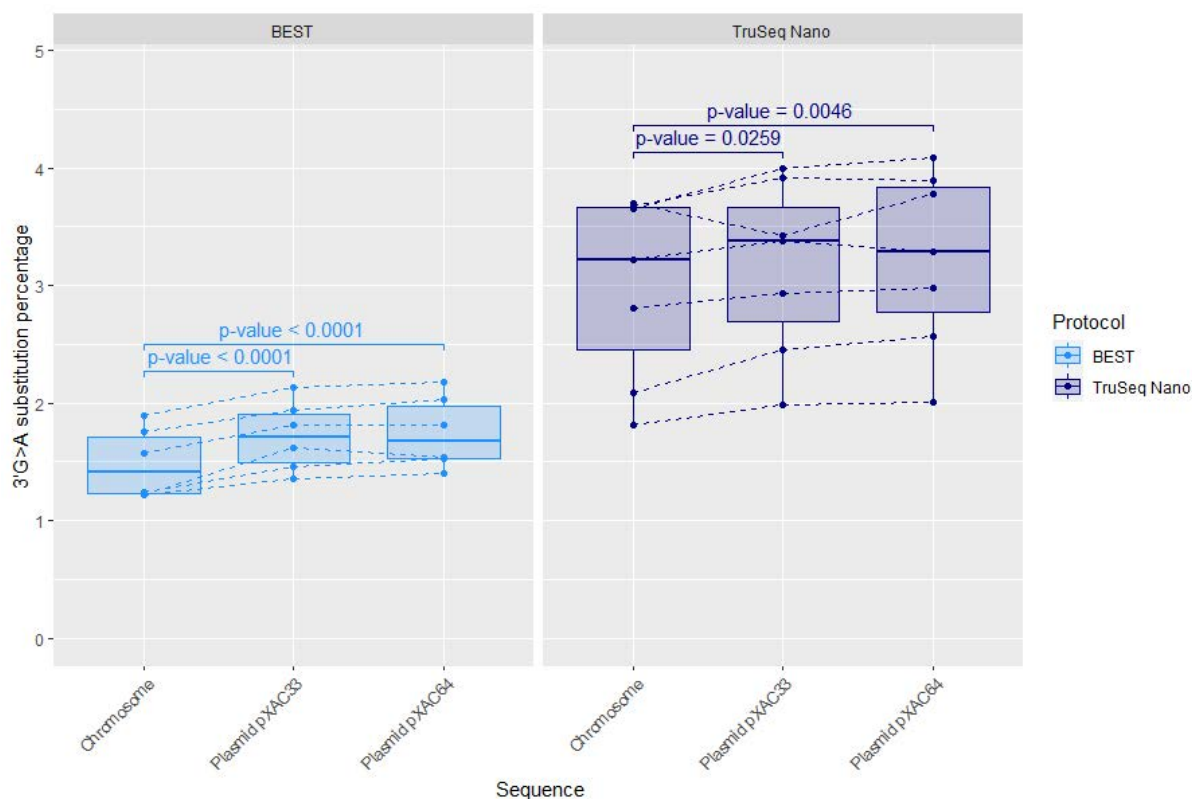


Figure 2.4.C. Pourcentage de désamination à la dernière position de l'extrémité 3' des ADN des trois séquences (chromosome, plasmides pXAC33 et pXAC64) des génomes *Xci* issus de 13 échantillons historiques. Les séquences des 13 échantillons historiques sont distinguées selon le protocole de construction de bibliothèques utilisé (BEST en bleu clair, TruSeq Nano en bleu foncé) et reliées par individu par des lignes en pointillés (valeurs d'un même individu appariées pour le test statistique glmpQL). Les boîtes représentent les 25^{èmes} aux 75^{èmes} percentiles, les lignes leurs valeurs maximales et minimales.

Cet échantillonnage réduit ne nous a cependant pas permis, de tester un effet potentiel des conditions de préservation au sein des différents herbiers sur la dégradation des acides nucléiques, comme cela a pu être montré pour le taux de désamination de 185 ADN (dont 15 échantillons d'herbier) datant du Pléistocène au XIX^{ème} siècle dans une étude récente (Kistler *et al.*, 2017). D'autres variables peuvent avoir un effet sur la dégradation de l'ADN, comme le traitement des échantillons d'herbier lors de la mise en collection, affectant la conservation (Staats *et al.*, 2011) et l'amplification de l'ADN (Särkinen *et al.*, 2012). Les conditions de conservation des échantillons d'herbier au cours du temps, mal référencées avant le XX^{ème} siècle, pourraient permettre de prédire le taux de dégradation de l'ADN de ce type d'échantillon, et donc de prédire la qualité d'une extraction sur un spécimen candidat, au cours du temps.

Cependant, les différentes analyses statistiques que nous avons appliquées sur notre jeu de données de seulement 13 échantillons, nous ont permis de mettre en évidence un effet du protocole sur la distribution des tailles de fragments d'ADN effectivement convertis en bibliothèques ainsi que sur le taux

de désamination à leur extrémité 3', avec les conditions « BEST » permettant d'obtenir des ADNs en moyenne de plus grandes tailles et moins désaminés que ceux obtenus en « TruSeq Nano ». Différents paramètres peuvent expliquer cela comme le protocole d'extraction, de précipitation, ou de construction de librairies. Ainsi, bien que nous ayons pu observer la présence des effets du type d'ADN ou de l'âge au sein des deux conditions, chacun des protocoles a pu avoir une influence sur la mesure des patrons de dégradation *post-mortem* de l'ADN et une attention toute particulière doit être donnée lors du *design* expérimental afin de limiter cet effet de protocole.

Enfin, si la différence testée entre types d'ADN n'était pas significative pour la fragmentation, elle l'était pour le taux de désamination, avec les séquences chromosomiques présentant des taux plus faibles que celles plasmidiques. Cette différence pourrait s'expliquer par des patrons de méthylation des cytosines différents entre chromosome et plasmides chez *Xanthomonas*. En effet, les N4-méthylcytosines, un patron propre aux bactéries (Sánchez-Romero, Cota and Casadesús, 2015), ont été trouvées en plus grande proportion sur le chromosome que sur certains des plasmides de deux espèces de *Xanthomonas* (Seong *et al.*, 2016). Or, les N4-méthylcytosines ont été identifiées chez des bactéries extrémophiles comme plus résistantes à la désamination que les cytosines non méthylées (Ehrlich *et al.*, 1986). Le taux de désamination plus faible observé sur le chromosome par rapport aux deux plasmides pourrait être dû à des cytosines chromosomiques chez *Xci* protégées de la désamination par ce type de méthylation, indépendamment d'autres mécanismes de dégradation de l'ADN comme la fragmentation. C'est à notre connaissance la première fois que cette observation différentielle de la désamination est faite entre chromosome et plasmide. Enfin, nous avons observé une corrélation significative entre le taux de désamination et l'âge des échantillons dès six (protocole « BEST ») et sept (protocole « TruSeq Nano ») échantillons historiques considérés. Cette observation significative, même à échantillonnage réduit, montre la puissance de notre méthode statistique, basée sur des décomptes de lectures portant l'information, là où d'autres études comme celle de Weiß *et al.* (2016), utilisant uniquement une information transformée (moyenne-log de la taille des fragments d'ADN, taux de substitution...) nécessitent un échantillonnage plus conséquent.

2.5. Annexes

Annexe 2.1.A. Protocole d'extraction ADN au CTAB (modifié de Ausubel et al., 2003)

Annexe 2.1.B Protocole de construction de librairies Illumina BEST 2.0

Annexe 2.1.C. Protocole d'extraction ADN au CTAB et billes magnétiques

Annexe 2.1.D. Liste des échantillons d'herbier convertis en librairies

Annexe 2.1.E. Robène *et al.* (2020) Development and comparative validation of genomic-driven PCR-based assays to detect *Xanthomonas citri* pv. *citri* in citrus plants

Annexe 2.1.A. Protocole d'extraction au CTAB (modifié de Ausubel *et al.*, 2003)

DNA extraction CTAB protocol

1. Sampling

Prepare 2.0 mL tube with 5 1.5mm glass beads and 1 4mm glass beads and autoclave

Use scalpel to cut around 5 lesions for approximately 10 mg and put them in tube

Change scalpel blade between each sample

Grind 5 times or until powdered in FastPred96 for 20 seconds at 1600 RPM

2. Extraction

Digestion

Resuspend powder in base buffer

Base buffer :

Reagent	Final concentration	Volume/Reaction (µL)
Tris-EDTA 2X	1X	500
H ₂ O		198
RNAse A (10 mg/mL)	0,05 mg/mL	5
N-lauroylsarcosine (10%)	0,50%	50
NaCl (5 M)	0,71M	142

Mix well

CTAB (10%)	1,00%	100
Proteinase K (20 mg/mL, QIAGEN blood & tissue kit)	0,1 mg/mL	5
Volume tot		1000
Volume by sample		900

Centrifuge samples before use and manipulate under Sorbonne

Reaction size: 900 µL

Put tube in VWR thermal shake until homogenous (5-6 hours at 600 RPM at 56°C, mix every hour)

Phase separation

Centrifugate 5 minutes at 5000G at room temperature

Collect supernatant into new 1.5mL tube (600 µL approximately)

Add equal volume of chloroform/isoamyl alcohol 24:1 and mix well

Centrifugate 20 minutes at 20000G at room temperature

Collect supernatant into new 1.5mL tube (500 µL approximately)

Add equal volume of chloroform/isoamyl alcohol 24:1 and mix well

Centrifugate 20 minutes at 20000G at room temperature

Collect supernatant into new 1.5mL tube (450 µL approximately)

DNA precipitation

Add 7/3 volumes (up to 8/2) of pure ethanol, mix by inverting or gently vortex 10 seconds

Leave at -20°C at least overnight (more if expected DNA concentration is slow)

Allow tube temperature to go up to 4°C on ice

Centrifuge 30 minutes at 20000G at 4°C

Rinse pellet twice with 400 µL chilled 70% ethanol

Centrifuge 15 minutes at 20000G at 4°C
Allow pellet to dry (on table)

DNA resuspension

Resuspend pellet in 50 µL Tris-HCl (10 mM, pH 8.0) and leave at 4°C overnight
Keep at 4°C overnight or -20°C for longer

Annexe 2.1.B. Protocole de construction de bibliothèques Illumina BEST 2.0

BEST library preparation Version 2.0 – Updated November 22th 2019, Christian Carøe, Copenhagen, Denmark

BEST protocol 2.0

Blunt-End Single-Tube Illumina library building for modern and ancient DNA

Carøe, C. et al. (2018) 'Single-tube library preparation for degraded DNA', *Methods in Ecology and Evolution*. Edited by S. Johnston, 9(2), pp. 410–419.
doi:10.1111/2041-210X.12871.

Mak et al. (2017) Comparative performance of the BGISEQ-500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. *GigaScience*, 6, 1–13

1. Library preparation

End Repair

Prior to library build dilute your samples to 32 μL using EB or EBT buffer, avoid using EDTA based buffers such TE. Then assemble the End-Repair master mix on ice or cooling block.

Sample Input	32 μL
--------------	------------------

End-Repair master mix

Reagent	Stock conc.	Final conc.	Total U	V/R μL	X reactions
T4 DNA polymerase	3 U/ μL	0.03	1.2	0.4	
T4 PNK	10 U/ μL	0.25	10	1	
dNTP	25 mM	0.25	-	0.4	
T4 DNA ligase buffer (NEB)	10x	1x	-	4	
Reaction booster/enhancer (see buffer preparation)				2.2	
Total:				8	

Reaction size:	40 μL
----------------	------------------

- Mix the mastermix by pipetting 10 times and add 8 μL mastermix to each reaction tube.
- Add the 32 μL sample to the mastermix and mix by pipetting 10 times. Spin down shortly and transfer the reaction tubes to a preprogrammed thermocycler with heated lid ($>75^{\circ}\text{C}$). If the thermocycler is warm or takes long to heat up the lid, let the reaction tubes wait on ice before placing them in the thermocycler.
- **Incubate: 30 min at 20 °C** directly followed by **30 min at 65 °C**, cool to 4 °C.

- When the above reaction is finished, add 2 μL adaptor solution to each reaction (with selected concentration, commonly 10-20 μM). This is done after the total end repair incubation of 60 minutes.
- **Important:** Mix the end repaired DNA and adaptor thoroughly by pipetting at least 10 times **before** adding Ligase master mix. Alternatively flick the tube and spin down.

Ligase master mix

Reagent	Stock conc.	Final conc.	Total U	V/R μL	X reactions
T4 DNA ligase buffer (NEB)	10x	1x		1	
PEG-4000	50%	6.25%		6	
T4 DNA ligase (NEB 400 U/ μL)	400 U/ μL	8	400	1	
Total:				8	

Reaction size:	50 μL
----------------	------------------

- Mix the mastermix by pipetting and add 8 μL to each reaction tube. The ligase mix is viscous and slow pipetting is necessary. Mix with the sample by pipetting (50 μL) or flicking the tube.
- Spin down shortly and transfer the reaction tubes to a preprogrammed thermocycler with heated lid ($>70^{\circ}\text{C}$). If the thermocycler is warm or takes long to heat up the lid, let the reaction tubes wait on ice before placing them in the thermocycler.

Incubate: 30 min at 20 °C followed by 10 min at 65 °C, cool to 4°C.

Fill-in master mix

Reagent	Stock conc.	Final conc.	Total U	V/R μL	X reactions
Isothermal amp. buffer	10x	0.33x		2	
dNTP	25 mM	0.33 mM		0.8	
Bst 2.0 Warmstart polymerase	8 U/ μL	0.21 U/ μL	12.8	1.6	
Mol. Grade Water				5.6	
Total:				10	

Reaction size:	60 μL
----------------	------------------

- Mix the mastermix by pipetting and add 10 μL to each reaction tube. The ligase mix is viscous and slow pipetting is necessary. Mix by pipetting (50 μL) or flicking the tube.
- Spin down shortly and transfer the reaction tubes to a preprogrammed thermocycler with heated lid ($>70^{\circ}\text{C}$). If the thermocycler is warm or

BEST library preparation Version 2.0 – Updated November 22th 2019, Christian Carøe, Copenhagen, Denmark
takes long to heat up the lid, let the reactions tubes wait on ice before placing them in the thermocycler.

Incubate: 15 min at 65 °C followed by 15 min at 80°C, cool to 4°C.

Clean up the library with either spin column (such as Qiagens MinElute or NEB's Monarch, following manufacturers recommendations) or speedbead purification according to Rohland & Reich (2012), see "buffer preparation" for guidelines on buffers.

1. Add 100 µL speedbead solution to the final library (60 µL) and mix 10 times with the pipette.
2. Incubate for 5 minutes at room temperature (not on the magnet)
3. Wash beads twice with fresh 80% ethanol (with the plate *on* the magnet)
4. Dry for exactly 5 minutes (do not overdry! - remove leftover ethanol with a pipette)
5. Elute in 5-100 µL EBT (30 µL recommended) by incubation for 10 minutes at 37 °C, before collecting supernatant by placing the beads on the magnet.

qPCR

Prepare diluted library samples by aliquoting 2 µL of 10x diluted library into each reaction tube (either PCR strip or 96 wells).

3) Prepare mastermix:

Reagent	stock	volume
Roche lightcycler 480 2x Mastermix	2x	5 µL
IS7/IS8 primermix	10 µM	0.3 µL
Mol. Grade water		2.7 µL

Master mix volume 8 µL per sample

4) Add 8µL mastermix to each reaction to obtain a full reaction volume of 10 µL. Mix reactions by pipetting at least 5 times or flick the tubes.
Spin down reactions to collect all liquid at the bottom of wells and avoid too many bubbles.

5) In the qPCR instrument, run the qPCR assay following the cycling conditions listed below using SYBR green detection mode:

Initial Denaturation at 95°C for 1 Minute

35-40 cycles of:

- 95°C for 15 seconds
- 63°C for 60 seconds

It is also advised to run a melting curve, to give an indication of the library content. Discard samples after qPCR without opening the tubes to avoid laboratory contamination.

Index PCR

In this setup, KAPA HIFI U+ is used, but other polymerases and buffers can be used (see guidelines).

Reagent	stock	volume
KAPA HIFI U+ 2x Mastermix	2x	25 µL
P7 primer	10 µM	1 µL
P5 primer	10 µM	1 µL

Mastermix volume 27 µL per sample

Add 23 µL sample to the above for a total volume of 50 µL. Mix reactions by pipetting at least 5 times or flick the tubes. Spin down reactions to collect all liquid at the bottom of wells and avoid too many bubbles.

In a thermocycler, run the following program:

Initial denaturation:

- 45 seconds at 98°C

X cycles of:

- 20 seconds at 98°C
- 30 seconds at 60°C
- 30 seconds at 72°C

Final extension

- 60 seconds at 72°C, hold at 4°C

Clean up the PCR using speedbeads as previously described. A 1.2x volume of speedbeads:sample is suitable to remove small size secondary products, but this may vary and should be tested with the specific batch of speedbead solution made.

1. Equilibrate SPRI/speedbead solution to room temperature and mix thoroughly.
2. Add 60 µL speedbead solution to the PCR reaction (50 µL) and mix 10 times with the pipette.
3. Incubate for 5 minutes at room temperature (not on the magnet)
4. Place on magnet for 2 minutes
5. Discard supernatant and wash beads twice with fresh 80% ethanol (with the plate *on* the magnet)
6. Dry for exactly 5 minutes (do not overdry! - remove leftover ethanol with a pipette)
7. Elute in 5-100 µL TET buffer (30 µL recommended) by incubation for 10 minutes at 37 °C, before collecting supernatant by placing the beads on the magnet.

Store eluted DNA in TET buffer at -20°C for long term storage or 5°C overnight. Quantification and subsequent pooling of PCR products should be done on Qubit and Bioanalyzer/fragment analyzer or absolute qPCR with standard curve.

Annexe 2.1.C. Protocole d'extraction ADN au CTAB et billes magnétiques

DNA extraction CTAB & Beads protocol

1. Sampling

Prepare 2.0 mL tube with 5 1.5mm glass beads and 1 4mm glass beads and autoclave

Use scalpel to cut around 5 lesions for approximately 10 mg and put them in tube

Change scalpel blade between each sample

Grind 5 times or until powdered in FastPred96 for 20 seconds at 1600 RPM

2. Extraction

Digestion

Resuspend powder in base buffer

Base buffer :

Reagent	Final concentration	Volume/Reaction (µL)
Tris-EDTA 2X	1X	500
H ₂ O		198
RNAse A (10 mg/mL)	0,05 mg/mL	5
N-lauroylsarcosine (10%)	0,50%	50
NaCl (5 M)	0,71M	142

Mix well

CTAB (10%)	1,00%	100
Proteinase K (20 mg/mL, QIAGEN blood & tissue kit)	0,1 mg/mL	5
Volume tot		1000
Volume by sample		900

Centrifuge samples before use and manipulate under Sorbonne

Reaction size: 900 µL

Put tube in VWR thermal shake until homogenous (5-6 hours at 600 RPM at 56°C, mix every hour)

Phase separation

Centrifugate 5 minutes at 5000G at room temperature

Collect supernatant into new 1.5mL tube (600 µL approximately)

Add equal volume of chloroform/isoamyl alcohol 24:1 and mix well

Centrifugate 20 minutes at 20000G at room temperature

Collect supernatant into new 1.5mL tube (500 µL approximately)

Add equal volume of chloroform/isoamyl alcohol 24:1 and mix well

Centrifugate 20 minutes at 20000G at room temperature

Collect supernatant into new 1.5mL tube (450 µL approximately)

DNA hybridisation on magnetic beads

Supernatant is hybridised on SpeedBeads™ magnetic carboxylate modified particles, Sigma-Aldrich, Cat#: GE45152105050250

3. Buffer preparation

TE buffer (50 mL)

5 mL EDTA (10 mM)

10 mL Tris-HCl (50 mM, pH 8.0)

35 mL H₂O

Washed and diluted beads in TE buffer (10%)

Mix Sera-mag Speed Beads and transfer 0,5 mL to a 1.5mL tube

Place tube on a magnet rack and wait 30 seconds for the liquid to clear

Discard supernatant

Remove tube from magnet rack

Add 1 mL TE buffer and bring the beads to solution by flicking the tube

Place tube on a magnet rack and wait 30 seconds for the liquid to clear

Discard supernatant

Remove tube from magnet rack

Add 1 mL TE buffer and bring the beads to solution by flicking the tube

Place tube on a magnet rack and wait 30 seconds for the liquid to clear

Discard supernatant

Resuspend the beads in 5 mL TE buffer and place on non-magnetic rack

Keep at 4°C in dark (foil around tube)

Bring at room temperature before use

Elution buffer (50 mL)

10 mL Tris-HCl (50 mM, pH 8.0)

250 µL Tween-20 (10%)


39.75 mL H₂O

Annexe 2.1.D. Liste des échantillons d'herbier convertis en librairies. Les échantillons sont identifiés par ID, herbier, date et localisation de collection, hôte et protocole et colorés selon leur plante hôte (*Citrus* (noir), *Aegle* (bleu), *Rutaceae* non *Citrus*) et *Coffea* (rouge, servant de témoin de contamination *Xci*).

ID Herbarier	Herbier	Date	Localisation	Hôte	Protocole de construction de librairies	% reads Xci (séquençage de titration)	ID Xci
P05297996	MNHN Paris	1911	Indonesia, Java	Citrus aurantiifolia	TruSeq Nano	1,77	HERB_1911
MAU00015159	Mauritius herbarium	1992	Rodrigues	Citrus sp.	TruSeq Nano	0,01	
MAU00015154	Mauritius herbarium	1974	Mauritius	Citrus lime	TruSeq Nano	3,30	HERB_1974
686249	National Fungus Collections, ARS, USDA	1946	Guam, Tlofofo	Citrus sp.	TruSeq Nano	7,60	HERB_1946
1206	Royal Botanic Gardens, Kew	1884	Philippines, Luzon	Citrus medica	TruSeq Nano	10,50	HERB_1884
P05297986	MNHN Paris	1845	Indonesia, Java	Citrus aurantiifolia	TruSeq Nano	10,40	HERB_1845
P05297992	MNHN Paris	1915	Philippines	Citrus aurantiifolia	TruSeq Nano	6,70	HERB_1915
16022	Royal Botanic Gardens, Kew	1927	China, Hainan	Citrus maxima	TruSeq Nano	0,01	
MAU00015151	Mauritius herbarium	1937	Mauritius	Citrus sp.	TruSeq Nano	1,21	HERB_1937
MAU00015092	Mauritius herbarium	1963	Mauritius	Citrus sp.	BEST	0,13	
K000466665	Royal Botanic Gardens, Kew	1995	Costa Rica	Coffea sp.	BEST	0,00	
K000466714	Royal Botanic Gardens, Kew	1982	Brazil	Coffea sp.	BEST	0,02	
Q1889	Royal Botanic Gardens, Kew	1865	India	Citrus medica	BEST	6,86	HERB_1865
Q1873	Royal Botanic Gardens, Kew	1852	India, Khasi Hills	Citrus medica	BEST	0,18	
6284	Royal Botanic Gardens, Kew	1914	Japan, Kagoshima	Citrus sp.	BEST	0,04	
25354	Royal Botanic Gardens, Kew	1916	Philippines, Luzon	Citrus hystrix	BEST	0,05	
630116	Royal Botanic Gardens, Kew	1963	Nepal, Sanichare	Citrus medica	BEST	15,27	HERB_1963
B635	NHM London	1864	China	Citrus medica	BEST	0,10	
Q1875	Royal Botanic Gardens, Kew	1852	India, Khasi Hills	Citrus medica	BEST	0,03	
K000466714	Royal Botanic Gardens, Kew	1982	Brazil	Coffea sp.	BEST	0,01	
1756364	US National Herbarium	1922	China, Yunnan	Citrus medica	BEST	28,68	HERB_1922
P05240716	MNHN Paris	1859	Bangladesh, Chittagong	Citrus medica	BEST	5,91	HERB_1859
Q1874	Royal Botanic Gardens, Kew	1852	India, Khasi Hills	Citrus medica	BEST	5,26	HERB_1852
K000466714	Royal Botanic Gardens, Kew	1982	Brazil	Coffea sp.	BEST	0,00	
Q1211	Royal Botanic Gardens, Kew	1867	India, Bihar	Aegle marmelos	BEST	4,83	
Q1954	Royal Botanic Gardens, Kew	1854	Indonesia, Java	Citrus aurantiifolia	BEST	11,62	HERB_1854



Development and comparative validation of genomic-driven PCR-based assays to detect *Xanthomonas citri* pv. *citri* in citrus plants

Isabelle Robène^{1*} , Véronique Maillot-Lebon¹, Aude Chabirand², Aurélie Moreau², Nathalie Becker³, Amal Moumène⁴, Adrien Rieux¹, Paola Campos^{1,3}, Lionel Gagnevin⁵, Myriam Gaudeul⁶, Claudia Baidier⁷, Frédéric Chiroleu¹ and Olivier Pruvost¹

Abstract

Background: Asiatic Citrus Canker, caused by *Xanthomonas citri* pv. *citri*, severely impacts citrus production worldwide and hampers international trade. Considerable regulatory procedures have been implemented to prevent the introduction and establishment of *X. citri* pv. *citri* into areas where it is not present. The effectiveness of this surveillance largely relies on the availability of specific and sensitive detection protocols. Although several PCR- or real-time PCR-based methods are available, most of them showed analytical specificity issues. Therefore, we developed new conventional and real-time quantitative PCR assays, which target a region identified by comparative genomic analyses, and compared them to existing protocols.

Results: Our assays target the *X. citri* pv. *citri* XAC1051 gene that encodes for a putative transmembrane protein. The real-time PCR assay includes an internal plant control (5.8S rDNA) for validating the assay in the absence of target amplification. A receiver-operating characteristic approach was used in order to determine a reliable cycle cut-off for providing accurate qualitative results. Repeatability, reproducibility and transferability between real-time devices were demonstrated for this duplex qPCR assay (XAC1051-2qPCR). When challenged with an extensive collection of target and non-target strains, both assays displayed a high analytical sensitivity and specificity performance: LOD_{95%} = 754 CFU ml⁻¹ (15 cells per reaction), 100% inclusivity, 97.2% exclusivity for XAC1051-2qPCR; LOD_{95%} = 5234 CFU ml⁻¹ (105 cells per reaction), 100% exclusivity and inclusivity for the conventional PCR. Both assays can detect the target from naturally infected citrus fruit. Interestingly, XAC1051-2qPCR detected *X. citri* pv. *citri* from herbarium citrus samples. The new PCR-based assays displayed enhanced analytical sensitivity and specificity when compared with previously published PCR and real-time qPCR assays.

(Continued on next page)

* Correspondence: isabelle.robene@cirad.fr

¹CIRAD, UMR PVBMT, Saint-Pierre, Reunion Island, France

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

(Continued from previous page)

Conclusions: We developed new valuable detection assays useful for routine diagnostics and surveillance of *X. citri* pv. *citri* in citrus material. Their reliability was evidenced through numerous trials on a wide range of bacterial strains and plant samples. Successful detection of the pathogen was achieved from both artificially and naturally infected plants, as well as from citrus herbarium samples, suggesting that these assays will have positive impact both for future applied and academic research on this bacterium.

Keywords: Asiatic Citrus canker, Surveillance, Real-time quantitative PCR, Diagnostics, Cycle cut-off, Ancient DNA

Background

Over the last half century, there has been a dramatic increase in biological invasions worldwide as a result of globalization and increased international trade and travel [1]. To limit the threat posed by the introduction of exotic plant pathogens and pests through trade and human transport, many countries have tightened border biosecurity surveillance, as well as phytosanitary inspection, and quarantine measures. This biosecurity effort has slowed down the introduction and establishment of pathogens despite the increase in trade and the international movement of people. However, biosecurity measures have adopted to differing degrees across the agricultural sector. Measures to protect annual crop and pasture species have had a positive impact. In contrast, the biosecurity effort should be enhanced for perennial crops such as forest and fruit tree species, which remain vulnerable [2]. When a plant disease outbreak is observed in a new area, quick and appropriate management measures, such as containment or eradication, are necessary to avoid the establishment and further spread of the pathogen. For efficient biosecurity surveillance and plant disease management, identifying regulated pathogens fast and accurately is essential. Misdiagnosis can have a severe economic impact, for example, if unwanted pathogens are introduced or inappropriate management options applied [3–5].

Xanthomonas citri pv. *citri*, the causal agent of Asiatic citrus canker (ACC), is an example of a high-concern regulated pathogen that threatens an economically important fruit crop (i.e., collectively, citrus fruit rank #1 fruit crop worldwide). This bacterial pathogen causes serious direct and indirect economic losses due to reduced crop yield and quality, the cost of eradication, containment (including the destruction of nursery, grove or backyard trees) or integrated management measures, as well as embargos on the movement of fruit. In addition, it raises concerns about environmental issues linked to increased pesticide use and the development of resistance to antimicrobials [4, 6]. For example, more than one billion US dollars was spent over a decade in Florida in an attempt to eradicate the pathogen [7]. Australia organized several successful ACC eradication campaigns in the past [8, 9], and is currently conducting

a response plan in the Northern Territory at a cost of millions of A\$.

Although at least four distinct xanthomonads are pathogenic to citrus, only two of them, *X. citri* pv. *citri* and *X. citri* pv. *aurantifolii*, cause visually indistinguishable canker-like symptoms. They are the causal agents of Asiatic and South American citrus canker, respectively. However, only the former bacterium has a major agricultural significance, because it is the only one associated with serious canker outbreaks even in countries where both pathogens occur concomitantly. Within *X. citri* pv. *citri*, strains differ in host range among citrus lines and can be classified into three distinct pathotypes. Pathotype A strains have the greatest global economic impact on the citrus industry. They are widely distributed and induce canker on a broad range of rutaceous hosts, including many *Citrus* species, hybrids or related genera such as trifoliate orange (*Poncirus trifoliata*) [4]. Pathotype A* strains are pathogenic to a restricted range of citrus hosts. Most outbreaks occur on Mexican lime (*Citrus x aurantiifolia*) in Asia, the Arabian Peninsula and Eastern Africa [10–12]. Pathotype A^w has been reported to date on the Indian subcontinent, in the Arabian Peninsula and the USA. Natural infections are restricted to Mexican lime and lemon (*C. x macrophylla*). It has the unique feature of causing a hypersensitive response when inoculated into some non-host citrus lines such as *C. x paradisi* and *C. x sinensis* [13, 14]. Pathotypes represent phylogenetically-coherent lineages based on whole-genome sequencing (WGS) data or genotyping data [15–18]. These techniques have made it possible to identify previously unreported sublineages within pathotypes A and A*, which are responsible for outbreaks both in the pathogen's area of origin or in regions where it has recently emerged [11, 16, 18].

Long-distance dissemination of *X. citri* pv. *citri* occurs primarily when humans transport diseased citrus plant material [19]. Since the early 2000s, there have been reports of several cases of geographical expansion and successful pathogen establishment in some Western and Eastern African countries, and the Caribbean [10, 20, 21]. *X. citri* pv. *citri* is a major threat to disease-free areas (e.g. New Zealand, Australia, and countries in Southern Africa and the Mediterranean) where it is listed as a quarantine organism. Preventing the establishment of *X. citri* pv. *citri* in new areas very much depends on the availability of

specific detection protocols and the implementation of surveillance and quarantine measures. Given the significance of ACC, numerous molecular detection methods have been developed, (i) conventional PCR assays [22–27], (ii) real-time quantitative qPCR assays [28, 29]; and (iii) Lamp assay [30]. Issues of inclusivity (i.e., the ability of the assay to detect all strains of the target organism) and/or exclusivity (i.e., the capacity to generate negative responses from non-target strains) occur with most of the previous assays developed for *X. citri* pv. citri [31]. Therefore, we evaluated recently published diagnostic tools and developed a new system.

Many microbial genomes that are publicly available constitute a valuable resource for identifying new, specific molecular markers. In this study, we describe the development of highly specific conventional and real-time PCR assays from a DNA marker selected using in silico comparative genomic analysis of *X. citri* pv. citri genomes and non-target *Xanthomonas* genomes. The real-time PCR assay amplified an endogenous plant DNA sequence present in the sample (5.8S rDNA) as a co-extracted and co-amplified internal control. It reveals any flaws in DNA extraction and the presence of PCR inhibitors [32–34]. These protocols were further validated using naturally infected citrus material collected in the field. These new molecular tools were independently evaluated and compared to existing protocols by the French Agency for Food, Environmental and Occupational Health & Safety (ANSES).

Results

Selection of a specific DNA fragment for PCR and qPCR assay design

The comparative genomic analysis of 30 *X. citri* pv. citri genomes against 30 other *Xanthomonas* genomes using the MicroScope platform [35] resulted in the selection of 33 coding DNA sequences (CDS), which were present in all the *X. citri* pv. citri genomes and absent in the non-target genomes used (Table S1). CDS with sizes <

100 bp ($n = 2$) and CDS corresponding to mobile elements (integrases, transposases, phage- and plasmid-borne genes) were not considered further ($n = 5$). The CDS XAC1051 (564 pb) encoding for a putative transmembrane protein displayed the best in silico specificity among the remaining candidates. A homologous sequence was only found in the *X. citri* pv. cajani strain LMG 558. It was split into two fragments located on two different contigs of its draft sequence. Thus, XAC1051 was used to design the qPCR Taqman® assay and conventional PCR assays (Table 1). The Primer Express® software used to design systems could not generate an efficient probe/primer system to allow the lack of amplification of *X. citri* pv. cajani strains. Conversely, primers were successfully designed to prevent the target amplification in *X. citri* pv. cajani for the conventional PCR assay.

Analytical specificity of XAC1051-2qPCR

The 58 bp targeted DNA region of *X. citri* pv. citri strain IAPAR 306, including the primers and probes perfectly matched (100% identity and 100% query coverage) with sequences belonging to all the 91 *X. citri* pv. citri genomes available on NCBI, three *X. citri* pv. citri historical genomes from herbarium samples used in this study and the *X. citri* pv. cajani strain LMG 558. Conversely, no significant similarity was found with sequences from the other 2790 non-target xanthomonads.

All *X. citri* pv. citri strains showed a FAM-positive signal when tested with the real-time PCR assay. The typical exponential amplification curves and Ct values ranged from 27.7 to 31.8 (mean of 29.7 and standard deviation of 1.0). Among the 101 non-target strains, only three strains of *X. citri* pv. cajani tested positive with the real-time PCR assay with Ct values ranging from 22.9 to 23.7.

Dynamic range

The dynamic range of the duplex quantitative real-time PCR was assayed with three independent 10-fold dilution series of strain IAPAR 306 in each of the five

Table 1 Primers and probes used in this study

Primers/probes	Sequence 5' > 3'	Amplicon size
XAC1051-2qPCR		
P-XAC1051-MGB (6-Fam™)	CGGTGAGAAGCTGTAC	58 bp
qPCR-XAC1051-F	AGAGGCCACTATGGCTTTC	
qPCR-XAC1051-R	CAACCCAGGACCTGCAAGAA	
P-citrus5.8S- MGB (Vic™)	ATCCCGTGAACCATCG	94 bp
citrus5.8S -F	GCGAAATGCGATACTGGTGTGA	
citrus5.8S-R	CGTGCCCTCGGCCTAATG	
XAC1051-F/R PCR		
XAC1051-F	AAATTCTGTGATCTGCTGGCT	499 bp
XAC1051-R	GCCGCCGCATAATCTCTCAC	

different plant matrices. Calibration curves were constructed for each plant matrix by plotting the obtained Ct values against the Log_{10} of the standard concentrations (Fig. 1 and Fig. S1). A very strong linear relationship was observed between Ct values and the logarithm of bacterial concentration down to 1×10^3 CFU ml⁻¹ for all plant matrices with $0.988 > r^2 > 0.992$. The PCR efficiency (E) ranged from 90 to 102% according to the different plant matrices. As target concentration decreased, some of the replicate samples appeared to be negative (undetermined results): 22 negative signals out of 45 and 39 out of 45 registered for the concentrations of 10^2 CFU ml⁻¹ and 10^1 CFU ml⁻¹, respectively. Nevertheless, when the calibration curves included data from the 10^2 CFU ml⁻¹ concentration, the coefficient and efficiency values were still acceptable ($0.976 > r^2 > 0.991$ and $92\% < E < 105\%$). In other words, the Ct values obtained for low target concentrations could be valid and may be considered as positive signals in routine diagnosis.

Plant internal control

A VIC-positive signal, with Ct values ranging from 22.6 to 30.9, was detected for all symptomless ($n = 90$) and spiked plant samples with bacterial concentrations ranging from 1×10^1 to 1×10^4 CFU ml⁻¹ ($n = 180$). For higher bacterial concentrations, the number of VIC-positive signals decreased proportionally to the increase in bacterial populations: 15/36, 3/36 and 0/36 positive

signals were registered for the concentrations of 1×10^5 CFU ml⁻¹, 1×10^6 CFU ml⁻¹ and 1×10^7 CFU ml⁻¹, respectively.

Analytical sensitivity in plant extracts

Before determining the analytical sensitivity, the first step was to implement a Ct cut-off, i.e., the value of Ct beyond which the real-time PCR signal was no longer considered as positive. Based on Youden's index, the cut-off was estimated at 35.4 (Fig. 2). This cycle cut-off value was used to convert the quantitative data into binary data.

The limit of detection (LOD95%) was then determined for the whole dataset by plotting the numbers of positive responses obtained at the different concentrations against the Log_{10} of bacterial concentrations (Fig. 3). Log-Log was the best fitting curve with an estimated LOD95% of 754 CFU ml⁻¹ (95% CI 600–949), which corresponds to 15 bacteria per reaction. The same samples tested with the XAC1051-PCR assay yielded to an estimated LOD95% of 5234 CFU ml⁻¹ (95% CI 3656–7482), which corresponds to 105 bacteria per test sample.

Repeatability, reproducibility and transferability

The intra-assay coefficients of variation obtained for all Ct value triplicates ranged from 0.035 to 2.482% with a median of 0.429%. The inter-assay coefficients of

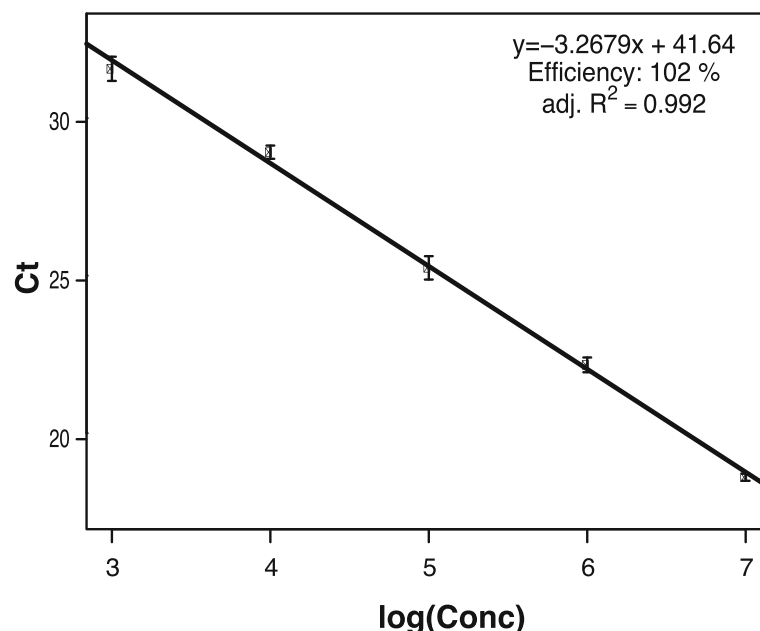


Fig. 1 Standard curve obtained for a dilution series of the *X. citri* pv. *citri* strain IAPAR 306 in a sweet orange matrix. XAC1051-2qPCR was run on total DNA extracted from sweet orange leaves spiked with serially suspensions obtained from 10-fold serial dilution (10^7 to 10^2 CFU ml⁻¹). The standard curve was constructed using linear regression analysis of the threshold cycle (Ct) values for the serial dilutions over the Log_{10} of the initial target concentrations (compilation of all series and runs, corresponding to 9 replicates at each concentration level). The linear regression equation, the efficiency value and the adjusted R^2 are indicated

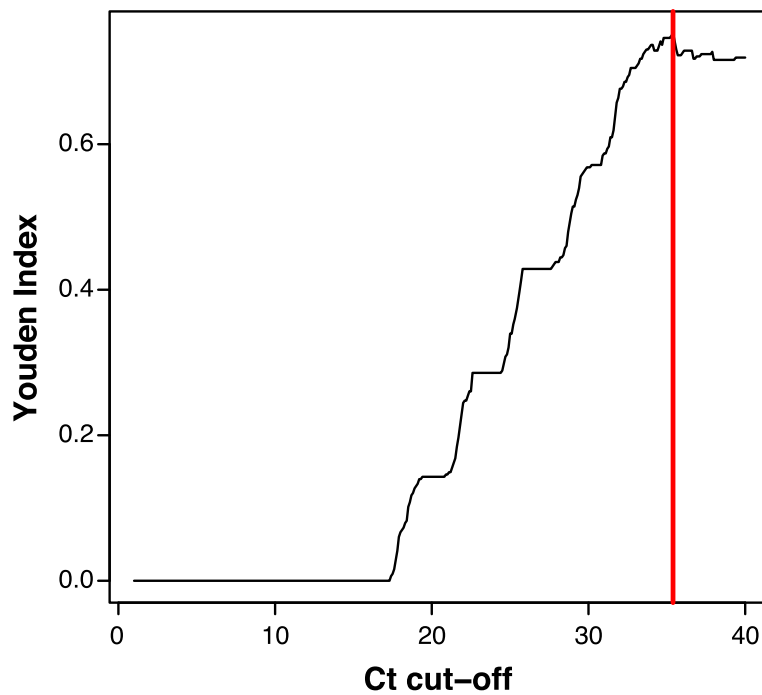


Fig. 2 Determining the Ct cut-off according to the Youden index J. The optimal cut-off point is the PCR cycle with the highest value of the Youden index value, which represents a trade-off between sensitivity and specificity. The Ct cut-off was estimated to be 35.4

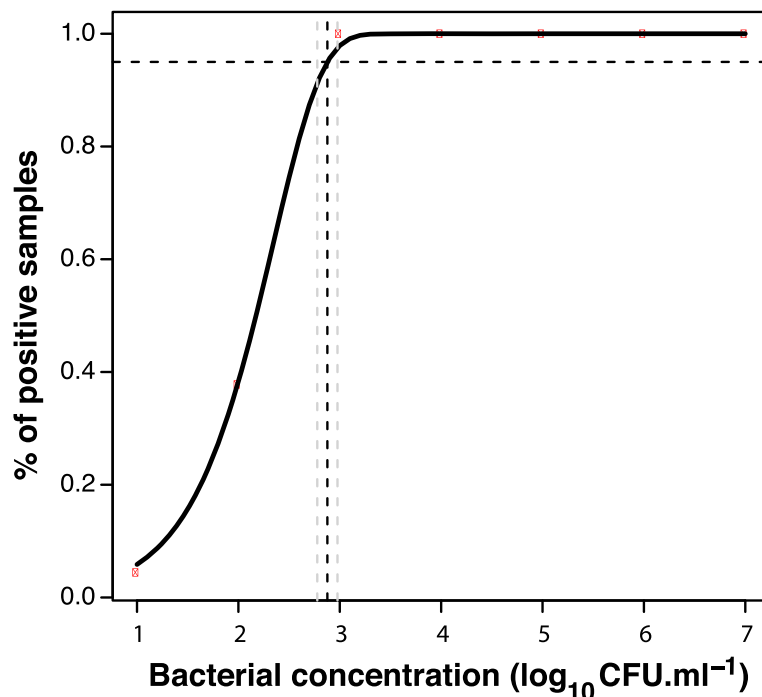


Fig. 3 Determining the limit of detection (LOD95%) of the real-time qPCR assay. The x-axis represents the log of the bacterial concentrations, and the y-axis represents the probability of detection (POD) of replicate samples with a Ct value below 35.4 (cut-off). Each red point on the graph corresponds to means of 45 data samples. The smooth fitting line represents the best fitting model to the data points, based on a least squares approach using a probit model. The dark dotted vertical line indicates the bacterial concentration corresponding to LOD 95% (754 CFU ml^{-1}) and the two clear dotted lines indicate the corresponding 95% confidence interval

variations were calculated using the Ct values obtained for the three independent dilution series in each plant matrix and ranged from 0.492 to 2.775% with a median of 1.558%. These low values reflected the repeatability and the reproducibility of the assay.

The qPCR efficiency values collected during the transferability assessment were all included between accepted range (90–110%) and all correlation coefficients were superior to 0.98 (Table 2). Low intra- and inter-assay C_v values were obtained for all experiments. Cut-off values were determined for each experiment involving a different qPCR device. These cut-off values were used to convert the quantitative data into qualitative data and to estimate the LOD95% values. LOD95% values were not statistically significant when the XAC1051-2qPCR protocol was tested on QS and LC480 devices.

Detection from naturally infected fruit

In order to detect *X. citri* pv. *citri* from symptomatic fruits collected in the field, we used XAC1051-PCR, XAC1051-2qPCR and enumeration of *Xanthomonas*-like colonies on semi-selective agar medium (Table S2). All samples showing canker lesions were tested positive using these methods for all four assayed citrus lines, while all the healthy citrus control samples tested negative. Non-repeatable low qPCR signals ($C_t > 35.4$), interpreted as negative results, were obtained for some replicates of a few symptomless samples. Some doubtful *Xanthomonas* colonies were observed on KC medium for some tangor samples. Nevertheless, suspensions from these colonies were tested negative by XAC1051-2qPCR. The plant signal was detected for all symptomless samples, which confirmed that the failure to detect *X. citri* pv. *citri* was not due to a technical problem during the step of DNA extraction and/or PCR amplification.

Detection from herbarium citrus samples

DNA extracted from the three herbarium specimens displayed substantial fragmentation (between 70 and 90 nt

on average, see Table S3), as expected for DNA obtained from this type of material [36]. Nevertheless, they all tested positive with the Xac-qPCR assay. The presence of *X. citri* pv. *citri* in these samples was further confirmed by analyzing the next generation sequencing data (Table S3) obtained from the same samples during the course of another study (Rieux, unpublished data).

Comparison of XAC1051-based conventional and real-time PCR assays with existing molecular tests

All *X. citri* pv. *citri* strains, irrespective of pathotype, tested positive with all five PCR assays and the three qPCR assays, with the exception of strain NCPPB 211 for J-Taqpth-qPCR (Table 3 and Table S4). The *X. citri* pv. *aurantifolii* B and C strains only tested positive only with the conventional Jpth1/2 and VM3/VM4 PCR assays and the VM-Syb-qPCR assay, which is consistent with previously published data [31].

In terms of exclusivity, the XAC1051-F/R PCR assay displayed 100% exclusivity whereas the other PCR assays picked up some non-target strains, with exclusivity values of 77.8% for both Jpth1/2 and VM 3/4 primers, 88.8% for XACF/R primers and 75.0% for XCF/R, respectively. Most of the PCR-positive non-target strains were phylogenetically close to *X. citri* [37]. Among the qPCR assays, XAC1051-2qPCR displayed the best exclusivity because only one non-target strain of *X. citri* pv. *cajani* was amplified, as seen above (97.2% exclusivity). J-Taqpth-qPCR showed an acceptable specificity (91.7% exclusivity) whereas VM-Syb-qPCR assay displayed only 77.8% exclusivity (i.e., the same value as for the conventional primer pairs).

The sensitivity of the different molecular assays was compared using different combinations of *X. citri* pv. *citri* strains and plant matrices (Table S5). Of the different conventional tests, the XAC1051-F/R assay was the most sensitive, with a detection threshold of 3×10^3 ml⁻¹ in most of the plant matrices. A few samples were

Table 2 Characteristics of the XAC1051-2qPCR assay performed using three different devices

	Plant matrix	StepOnePlus	QS	LC480
qPCR Efficiency	lemon	93%	94%	97%
	orange	102%	103%	106%
R ²	lemon	0.992	0.981	0.984
	orange	0.988	0.986	0.987
Intra-assay variation C_v range (median)		0.07–2.48% (0.34)	0.13–4.6% (0.60)	0.14–3.6% (0.81)
Inter-assay variation C_v range (median)		0.49–2.5% (1.50)	0.55–3% (1.30)	0.84–2.4% (1.50)
Cut-off values		35.9	38.43	38.89
LOD95% (CI)		2.90 (2.76–3.04)	3.04 (2.96–3.11)	3.09 (2.92–3.25)
Vic Ct range		25.8–31.2	22.3–27.9	23.4–29.7

Table 3 Comparison of the specificity of several PCR and real-time qPCR protocols

Assay		<i>X. citri</i> pv. citri			<i>X. citri</i> pv. aurantifolii (n = 5)	<i>X. euvesicatoria</i> pv. citrumelonis (n = 2)	Other <i>X. citri</i> ^a pathovars (n = 9)	Other <i>Xanthomonas</i> ^a species (n = 11)	Saprophytic xanthomonads ^b (n = 15)
		A (n = 63)	A* (n = 11)	A ^w (n = 4)					
PCR	Jpth1/2	62 ^c	11	4	5	0	7	1	0
PCR	VM3/4	63	11	4	5	0	7	1	0
PCR	XACF/R	63	11	4	0	0	4	0	0
PCR	XCF/R	63	11	4	0	0	9	0	0
PCR	XAC1051-F/R	63	11	4	0	0	0	0	0
qPCR	XAC1051-2qPCR	63	11	4	0	0	1	0	0
qPCR	J-Taqpht-qPCR	63	11	4	0	0	3	0	0
qPCR	VM-Syb-qPCR	63	11	4	5	0	7	1	0

^a Isolated from *Citrus* spp. but not pathogenic to *Citrus* spp.

^b Isolated from *Citrus* spp.

^c Number of positive samples

positive at $3 \times 10^4 \text{ ml}^{-1}$. The Miyoshi XCF/R protocol yielded similar results whereas all other conventional PCR assays were found less sensitive (3×10^4 to $3 \times 10^7 \text{ CFU ml}^{-1}$, depending on plant matrices and strains tested). Cut-off thresholds were determined for each real-time protocol using the ROC method. As expected, real-time PCR assays displayed a higher level of sensitivity than the conventional PCRs with almost identical sensitivity (a majority of detection limits at $3 \times 10^3 \text{ CFU ml}^{-1}$).

Discussion

Xanthomonas citri pv. citri is a major threat to global citrus production. The success of surveillance strategies and quarantine measures to control the international movement of *X. citri* pv. citri is highly dependent on the availability of rapid and reliable *in planta* detection tools. Previous studies have shown that a number of existing diagnostic protocols developed for *X. citri* pv. citri display insufficient exclusivity or in some cases inclusivity [31]. In this study, we developed new, highly specific and sensitive molecular assays to diagnose, detect and quantify *X. citri* pv. citri in citrus tissues. We compared the new assays to existing diagnostic tools using a broad collection of target and non-target strains and in different citrus matrices. Importantly, the present study evaluated how PCR protocols reacted to an extensive *X. citri* pv. citri strain collection, a feature that most earlier studies failed to achieve. Indeed, we assayed representative samples of all the lineages/sublineages of this bacterium, which have been reported to date throughout the world [11, 16–18, 38].

The selected target gene, XAC1051, encodes for a hypothetical transmembrane protein and is present on the chromosome of strains for which a complete genome sequence is available. This gene is part of a genomic

region previously considered to be specific to *X. citri* pv. citri, when compared to *X. citri* pv. aurantifolii B and C strains [39].

When a conventional PCR format was used, our assay appeared perfectly specific with 100% inclusivity and exclusivity, values that outcompete other conventional PCR assays.

Real-time quantitative PCR assays have significant benefits compared to conventional PCR, including shorter turnaround time, reproducibility and sensitivity. In addition, this method can be used for both qualitative and quantitative assessments (for recent examples, see [40, 41]). We therefore developed a duplex real-time PCR assay, targeting the same bacterial gene, and including a plant internal control. This control targets a plant 5.8S rDNA sequence conserved among *Citrus* species. We demonstrated successful amplification of the plant internal control for *X. citri* pv. citri concentrations $\leq 1 \times 10^4 \text{ CFU ml}^{-1}$. The plant signal was inhibited when higher bacterial concentrations were present in the extracts, which is consistent with previously published data [42]. Importantly, the plant control always yielded positive reactions in the absence of a bacterial signal. Therefore, it shows that a negative response for the bacterium is not due to a failure in the DNA extraction or PCR amplification process.

The XAC-1051-2qPCR assay was shown to be highly specific (100% inclusivity and 97.2% exclusivity). It displayed the best specificity when compared to the other real-time PCR assays available to date. Only strains of *X. citri* pv. cajani, responsible for bacterial leaf spot disease of pigeon pea (*Cajanus cajan*, Fabaceae) [43] tested positive with this molecular assay. This pathogen seems geographically restricted to India where pigeon pea is its sole known host species. *X. citri* pv. cajani was identified as the closest relative to *X. citri* pv. citri in a recent

phylogenomic analysis [38], suggesting that this DNA region may have been present in their most recent common ancestor. If necessary, suspect samples can be confirmed by performing the XAC1051-F/R PCR assay.

When real-time quantitative PCR is used as a qualitative method, a Ct cut-off, i.e., the PCR cycle number above which any sample response value (Ct) is considered to be a false positive, must be set. Indeed, depending on experimental conditions, some false positives with high Ct values may be registered. These values result from a spontaneous increase in fluorescence background emissions and/or low-level DNA cross-contaminations [44, 45]. The cut-off varies depending on the experiment context: inherent characteristics of the real-time PCR system, qPCR instrument, qPCR mix and DNA template. A statistical approach based on ROC analysis, where both false positive and false negative qPCR signals are considered, allowed us to establish an optimal Ct cut-off. The estimated Ct cut-off was applied to determine the level of sensitivity of our assay in different plant matrices. The XAC-1051-2qPCR assay was shown to be highly sensitive with the capacity to detect approximately 15 target cells per reaction. It demonstrated high repeatability and reproducibility. It also proved to be transferable between PCR cyclers with an optimization step, without compromising sensitivity and specificity. Importantly, the XAC-1051-2qPCR assay showed a good ability to detect the target from naturally infected citrus fruits. Interestingly, as predicted *in silico*, the XAC-1051-2qPCR assay was able to detect *X. citri* pv. *citri* from some herbarium samples dating back to 1911 (Table S3), despite the small quantities, the high fragmentation and the chemical modifications expected when ancient DNA (aDNA) is obtained from samples of this type [36]. The success of this assay is probably due to its high sensitivity and the small size of the target DNA (58pb). Interestingly, this result suggests that molecular tools, such as our XAC-1051-2qPCR assay, are useful for screening herbaria for plant pathogens. Screening is a prerequisite for further investigations such as metagenomics or population genetic analyses geared to reconstructing the evolutionary history of plant pathogens [46, 47].

Conclusions

Herein, we conducted a thorough comparative analysis of several conventional and quantitative PCR protocols using the same strain collection and plant samples. Thus, we hope to provide end-users with precise information with regard to the respective advantages and limitations of the different protocols in order to help them select one or more complementary methods for testing plant material or microbial cultures.

In agreement with previous studies [48–50], we conclude that genome-informed identification of targets is a powerful aid when it comes to developing highly specific diagnostic techniques for plant pathogens.

Methods

Bacterial strains and culture conditions

Ninety-eight strains of *X. citri* pv. *citri*, representing the currently known genetic diversity of this pathovar were used in this study (Table S6). This collection included the strain LMG 696. This *X. citri* pv. *citri* pathotype A* strain was recently authenticated by WGS data, after initially being mislabeled as *X. campestris* pv. *durantae* [38].

The present study also examined 101 non-target strains representing other bacterial genera, other pathogens of *X. citri*, other *Xanthomonas* strains pathogenic to rutaceous species and saprophytic *Xanthomonas* strains isolated from citrus (Table S7).

To compare the different PCR and real-time quantitative PCR protocols (see 2.10 and 3.4), a specific collection of strains was used, including some of the strains of Tables S6 and S7. This collection is listed separately in Table S4 to facilitate comprehension.

Strains were stored at -80°C on beads in cryovials (Microbank Prolab Diagnostics) or freeze-dried for long-term storage. All strains were streaked on yeast-peptone-glucose agar (YPGA; yeast extract 7 g l^{-1} , peptone 7 g l^{-1} , glucose 7 g l^{-1} , and agar 18 g l^{-1} ; pH 7.2) plates at 28°C for 3–4 days to check for purity. Subcultures were produced from single colonies on YPGA plates incubated at 28°C for 48 h. Bacterial suspensions were prepared and diluted in 0.01 M Tris buffer pH 7.2 (Sigma 7–9 Sigma-Aldrich, Saint-Quentin Fallavier, France) unless otherwise stated. Plant or canker lesion homogenates were prepared in the same buffer supplemented with 2% polyvinylpyrrolidone (PVP) with an average mol wt of 40,000 (Sigma-Aldrich).

Selection of a DNA target specific to *Xanthomonas citri* pv. *citri* and a plant internal control

A preliminary bioinformatics screening of candidate CDSs was performed using the “gene phyloprofile” tool in the MicroScope platform (Genoscope, Evry, France) [35] on 30 *X. citri* pv. *citri* genomes, including pathotype A, A* and A^W strains against 30 non-target genomes of *Xanthomonas* (other species and pathovars). The aim was to select CDSs conserved in all *X. citri* pv. *citri* genomes that had limited or no identity to CDSs from non-target genomes present in the database. Then, using the selected nucleotide sequences as query, we performed BLASTn and Megablast searches against NCBI databases (February 2020): nr/nt, draft ($n = 2225$) and complete genomes ($n = 565$) of *Xanthomonas* group

(taxid = 32,033), complete plasmids ($n = 17,302$) and complete bacteriophages ($n = 2717$). In silico presence and identity of chosen target regions was further confirmed in the three ancient genomes used in this study (see above, detection from herbarium citrus samples, and Table S3).

The 5.8S rDNA sequence from *C. x aurantiifolia* (MF797954) was selected to develop an endogenous plant internal control. This multicopy DNA region is conserved in the Rutaceae family, particularly among *Citrus* species.

Duplex real-time quantitative PCR (XAC1051-2qPCR) and PCR assays

Taqman[®] probe and primers were designed from the *X. citri* pv. *citri* XAC1051 gene and the plant 5.8S rDNA sequence using Primer Express[®] software Version 3.0 (Applied Biosystems, Courtaboeuf, France) and were provided by Applied Biosystems (Courtaboeuf, France). The different Taqman[®] probe and primer systems were checked in Oligo 7.6 (Molecular Biology Insights, Inc., Cascade, CO, USA) in order to minimize interactions between the different oligonucleotides. The selected primers and probes are listed in Table 1.

Amplifications were carried out in 15- μ l reaction volumes (in HPLC grade water) containing 7.5 μ l of 2 \times Mastermix (Applied Biosystems), 600 nM of qPCR-XAC1051-F and qPCR-XAC1051-R primers, 425 nM of 5'FAM-labeled XAC-1051 MGB probe (P-XAC1051-MGB), 50 nM of citrus5.8SF and citrus5.8SR primers, 50 nM of 5'VIC-labeled citrus5.8S MGB probe (P-citrus5.8S-MGB) and 2 μ l (pure bacterial suspensions) or 5 μ l (total plant DNA extract) of template DNA.

The real-time PCR cycling conditions included a step at 50 °C for 2 min, an initial denaturation step at 95 °C for 2 min followed by 40 cycles of denaturation and annealing/elongation for 15 s at 95 °C and 1 min at 60 °C, respectively. Analyses were performed using the StepOnePlus software version v2.2.2. Each sample was at least duplicated.

Conventional primers were also designed from the XAC1051 gene using Oligo 7.6 (Table 1). Amplifications were carried out in 25- μ l reaction volumes containing 5 μ l of Green GoTaq[®] Reaction Buffer, 2 mM MgCl₂, 0.5 μ M of XAC1051-F and XAC1051-R, 0.2 mM each dNTP, 1.25 U of GoTaq[®] DNA Polymerase (Promega), and 2 μ l of template DNA. PCR amplifications were performed using a Veriti[™] Thermal Cycler (Applied Biosystems, Courtaboeuf, France). The amplification program included denaturation at 95 °C for 2 min, 35 cycles consisting of denaturation at 95 °C for 45 s, annealing at 65 °C for 45 s, and extension at 72 °C for 1 min, and a final extension step at 72 °C for 5 min.

Specificity of XAC1051-2qPCR assay

The in silico-determined specificity of the 58 bp target region from *X. citri* pv. *citri* was subject to an additional experimental check following the guidelines in the EPPO PM 7/98 (4) standard protocol [51]. The real time PCR protocol was assayed on pure cultures of target ($n = 98$) and non-target ($n = 101$) strains (Tables S6 and S7). Spectrophotometrically adjusted suspensions containing approx. 1×10^8 CFU ml⁻¹ were diluted 100 or 10,000-fold for non-target and target strain assays, respectively. The suspensions were heated at 95 °C for 2 min and chilled on ice.

Dynamic range in the plant matrix

The dynamic range of the real-time PCR assay, i.e., the range of initial template concentrations for which accurate Ct values are obtained, was determined on the dilution series of the strain IAPAR 306 in different citrus matrices: sweet orange (*C. x sinensis*), clementine mandarin (*C. reticulata*), grapefruit (*C. x paradisi*), lemon (*C. x limon*) and makrut lime (*C. hystrix*). Overnight cultures of IAPAR 306 were adjusted spectrophotometrically to a concentration of approx. 1×10^8 CFU ml⁻¹ and serially 10-fold diluted. Fruit peel (0.1 g) was homogenized in 10 ml buffer using a grinder (Homex 6, Bioreba, Reinach, Switzerland) and spiked with bacterial suspensions at final concentrations ranging from 1×10^1 to 1×10^7 CFU ml⁻¹. Three replicated dilution series were performed in each citrus matrix. Total DNA was extracted from 2 ml homogenates using DNeasy Plant Mini kit (Qiagen, Courtaboeuf, France). Three qPCR replicates were carried out at each contamination level (nine Ct values were thus registered for each plant matrix and contamination level). Non-template controls (NTC) consisting of plant matrix and mix without DNA were included as negative samples ($n = 18$). Standard curves were generated for each citrus matrix by plotting Ct values against the logarithm of initial DNA concentrations. The reaction efficiency E was calculated according to the slope of the standard curves as follows: $E = 10^{(-\frac{1}{\text{slope}})} - 1$. The XAC1051-F/R PCR assay was also performed in duplicate using the same samples.

Cut-off Ct value and limit of detection (LOD)

A ROC (Receiver operating characteristic) was used in order to determine the Ct cut-off value, i.e., the PCR cycle number above which signals are no longer interpreted as positive [52, 53]. This analysis, based on the determination of the Youden J index, which considers false positives and negatives [42, 54], was performed on Ct values obtained (see § 2.5. above) for samples with a priori positive status, i.e., citrus spiked samples with different bacterial concentrations ($n = 135$) and for samples

with a priori negative status, i.e., NTC ($n = 110$). Samples with Ct values higher than the Ct cut-off value were then considered negative.

The analytical sensitivity of the XAC1051-2qPCR and the XAC1051-F/R PCR were estimated by determining the 95% limit of detection (LOD95%), i.e., the concentration at which a detection probability of 95% is expected [55–57] as explained previously [42].

Repeatability, reproducibility and transferability

Repeatability (i.e., the level of agreement between replicates of a sample tested under the same conditions) and reproducibility (i.e., the ability of a test to provide consistent results when applied to aliquots of the same sample tested under different conditions (time, personnel, equipment, location, etc.)) were estimated according to the EPPO standard PM 7/76 (5) [58]. Repeatability was evaluated by computing intra-assay coefficients of variation ($Cv = \frac{\sigma}{\mu}$) based on Ct mean values of qPCR triplicates obtained for the concentrations ranging from 1×10^3 CFU ml⁻¹ to 1×10^7 CFU ml⁻¹ (70 Cv values). Reproducibility was evaluated on three qPCR runs independently performed for each plant matrix at different times. Inter-assay Cv values were calculated from the Ct values (PCR triplicate means) obtained for concentrations ranging from 1×10^3 CFU ml⁻¹ to 1×10^7 CFU ml⁻¹ (25 Cv values).

The protocol's transferability was evaluated by reproducibility experiments performed by different operators, at different periods and in different laboratories (Cirad and ANSES). Dilution series of suspensions prepared from the IAPAR 306 strain in lemon and orange matrices already tested on the StepOnePlus device (Applied Biosystems) were also assayed using the Light Cycler LC 480 (Roche Life Science, Meylan, France) and the Quantstudio5 (QS5) (Applied Biosystems) real-time PCR systems. The application of the StepOnePlus master mix and real time cycling conditions for other devices gave poor results and required optimization (data not shown). Successful amplifications were obtained for both LC480 and QS5 devices when using the GoTaq[®] probe qPCR master mix kit (Promega) and the following cycling conditions: a step at 95 °C for 2 min followed by 45 cycles of denaturation and annealing/elongation for 15 s at 95 °C and 1 min at 60 °C, respectively. The concentrations of the different primers and probes remained the same as for the StepOnePlus. Efficiency and correlation coefficients were calculated for each fruit/real-time device data set. Cut-off values and LOD95% were estimated for each real-time qPCR device data set. Intra-assay and inter coefficients of variation were also calculated for each real-time qPCR device data set.

Detection from naturally infected fruit

Fifteen fruit (several species) showing typical ACC symptoms were collected in citrus groves in Reunion (Table S2). Three lesions per fruit (0.1 g each) were sampled and independently homogenized in 10 ml buffer (45 samples). Fifty microliters were plated in duplicates on KC medium to estimate target concentrations [59]. Total DNA was extracted from 2 ml homogenates using DNeasy Plant Mini kit (Qiagen, Courtaboeuf, France). Fifteen citrus fruits showing no canker symptoms were analyzed as well, with two or three samples (same size as diseased samples) collected independently per fruit and processed as described above (43 samples). Conventional and real time quantitative PCR assays were performed on the different samples as described in § 2.3.

Detection from herbarium citrus samples

The duplex PCR assay was also used to screen three herbarium citrus samples bearing typical citrus canker lesions, and collected between 1911 to 1992 in different areas (Table S3). They were provided by the Royal Mauritius Herbarium (acronym: MAU) and the National Herbarium of the Muséum National d'Histoire Naturelle, France (acronym: P). DNA extraction was performed in a bleach-cleaned facility room according to the protocol described in §2.5 using 0.01 g of leaf fragments (instead of fruit peel) as starting material. DNA concentration and fragment size were measured with Qubit (Invitrogen life Technologies) and TapeStation (Agilent Technologies) high sensitivity assays, respectively, according to the manufacturers' instructions. The XAC1051-dqPCR assay was performed as described in § 2.3.

Comparison of XAC1051-based conventional and real-time PCR assays with existing molecular tests

This comparison was performed by a laboratory (ANSES), which is different to the one where the qPCR XAC1051-based conventional and real-time PCR assays were developed (Cirad). We considered a selection of published PCR and real-time PCR protocols based on previously published data [31] or preliminary experimental and/or in silico data analyses (in the case of the most recent protocols). This excluded a recently published multiplex protocol designed to detect and differentiate between several citrus-associated xanthomonads [23], because the primers selected for *X. citri* pv. *citri* pathotype A perfectly matched in silico and reacted in preliminary assays with five other *Xanthomonas citri* pathovars. Table S8 presents the published protocols (and associated experimental conditions), which passed the first screen and were compared to the XAC1051-

based conventional and real-time PCR assays in terms of analytical specificity and sensitivity.

In the first assay, bacterial suspensions in sterile distilled water (containing approx. 1×10^4 or 1×10^6 CFU ml^{-1} for target and non-target strains, respectively) were used. A set of *X. citri* pv. *citri* strains ($n = 78$) and other xanthomonads ($n = 42$) (Table S4) was assayed in duplicate to compare the analytical specificity following the guidelines in EPPO PM 7/98 (4) standard protocol [51].

Then, the comparison of analytical sensitivity was carried out on different plant matrices spiked with tenfold-diluted bacterial suspensions. Plant matrices included the leaf or fruit peel of sweet orange grapefruit, lemon, Tahiti lime (*C. x latifolia*), clementine (*C. x clementina*) and makrut lime. They were spiked with 10-fold dilutions of the *X. citri* pv. *citri* strain CFBP 2525 with final concentrations ranging from 3×10^2 to 3×10^7 CFU ml^{-1} . In addition, leaves and fruit peels of Mexican lime (i.e., a host species susceptible to all *X. citri* pv. *citri* pathotypes) were spiked with 10-fold dilutions of the following strains: pathotype A strains JJ238–29 and LH001–1 (lineage 1 and 2, respectively), pathotype A^w strain LG115 (lineage 3), pathotype A* strain CFBP 2911 (lineage 4), the *X. citri* pv. *aurantifolii* pathotype B strain CFBP 2902 or the *X. citri* pv. *aurantifolii* pathotype C strain CFBP 2866 (same final concentrations as CFBP 2525). Homogenate production (0.1 mg plant matrix 5 ml buffer) and DNA extractions were performed as described above. PCR or real-time PCR assays were performed in duplicate.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12866-020-01972-8>.

Additional file 1.

Additional file 2.

Acknowledgements

We would like to express our thanks to the Plant Protection Platform (3P, IBISA) and to C. Boyer for her helpful contribution.

Authors' contributions

Conceived and designed the experiments: IR, AC, NB, AR, LG and OP. Provided biological material, performed the experiments and acquired the data: VML, AMor, NB, AMou, AR, PC, LG, MG and CB. Run bioinformatic analyses: IR, AR and PC. Analysed and interpreted the data: IR, VML, AC, NB, AMou, AR, PC, LG, CV, FC and OP. Wrote the manuscript: IR and OP. Revised and approved the manuscript: all authors. All authors read and approved the final manuscript.

Funding

The European Union Agricultural Fund for Rural Development (FEARFD contract 2014FR06RDRP004), the European Regional Development Fund (ERDF contract GURDT I2016–1731-0006632), the Région Réunion, the French Agropolis Foundation (Labex Agro – Montpellier, E-SPACE project number 1504–004), the Agence Nationale pour la Recherche (JCJC MUSEOBACT contract ANR-17-CE35–0009-01), ANSES and CIRAD provided financial support.

Availability of data and materials

Available as Supplementary Material. R scripts available upon request.

Ethics approval and consent to participate

Not Applicable.

Consent for publication

Not Applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹CIRAD, UMR PVBMT, Saint-Pierre, Reunion Island, France. ²Unit for Tropical Pests and Diseases, Plant Health Laboratory (LSV), French Agency for Food, Environmental and Occupational Health & Safety (ANSES), Saint-Pierre, Reunion Island, France. ³Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum National d'Histoire Naturelle, Sorbonne Université, EPHE, Université des Antilles, CNRS, Paris, France. ⁴Université de La Réunion, UMR PVBMT, Saint-Pierre, Reunion Island, France. ⁵CIRAD-UMR IPME, Montpellier, France. ⁶Herbier national (P), Muséum National d'Histoire Naturelle, Paris, France. ⁷Ministry of Agro Industry and Food Security, Mauritius Herbarium, R.E. Vaughan Building (MSIRI compound) Agricultural Services, Réduit, Mauritius.

Received: 23 June 2020 Accepted: 8 September 2020

Published online: 01 October 2020

References

- Paini DR, Sheppard AW, Cook DC, De Barro PJ, Worner SP, Thomas MB. Global threat to agriculture from invasive species. *Proc Natl Acad Sci U S A*. 2016;113(27):7575–9. <https://doi.org/10.1073/pnas.1602205113>.
- Sikes BA, Bufford JL, Hulme PE, Cooper JA, Johnston PR, Duncan RP. Import volumes and biosecurity interventions shape the arrival rate of fungal pathogens. *PLoS Biol*. 2018;16(5):e2006025. <https://doi.org/10.1371/journal.pbio.2006025>.
- Bull CT, Koike ST. Practical benefits of knowing the enemy: modern molecular tools for diagnosing the etiology of bacterial diseases and understanding the taxonomy and diversity of plant-pathogenic bacteria. *Annu Rev Phytopathol*. 2015;53:157–80. <https://doi.org/10.1146/annurev-phyto-080614-120122>.
- Graham JH, Gottwald TR, Cubero J, Achor DS. *Xanthomonas axonopodis* pv. *citri*: factors affecting successful eradication of citrus canker. *Mol Plant Pathol*. 2004;5(1):1–15.
- Miller SA, Beed FD, Harmon CL. Plant disease diagnostic capabilities and networks. *Annu Rev Phytopathol*. 2009;47:15–38. <https://doi.org/10.1146/annurev-phyto-080508-081743>.
- Richard D, Ravigné V, Rieux A, Facon B, Boyer C, Boyer K, et al. Adaptation of genetically monomorphic bacteria: evolution of copper resistance through multiple horizontal gene transfers of complex and versatile mobile genetic elements. *Mol Ecol*. 2017;26(7):2131–49.
- Gottwald TR, Irely M. Post-hurricane analysis of citrus canker II: Predictive model estimation of disease spread and area potentially impacted by various eradication protocols following catastrophic weather events. *Plant Health Prog*. 2007. <https://doi.org/10.1094/PHP-2007-0405-01-RS>.
- Broadbent P, Fahy PC, Gillings MR, Bradley JK, Barnes D. Asiatic citrus canker detected in a pummelo orchard in northern Australia. *Plant Dis*. 1992;76(8):824–9.
- Gambley CF, Miles AK, Ramsden M, Doogan V, Thomas JE, Parmenter K, et al. The distribution and spread of citrus canker in emerald, Australia. *Australas Plant Pathol*. 2009;38:547–57.
- Derso E, Vernière C, Pruvost O. First report of *Xanthomonas citri* pv. *citri*-A* causing citrus canker on lime in Ethiopia. *Plant Dis*. 2009;93(2):203.
- Pruvost O, Goodarzi T, Boyer K, Soltaninejad H, Escalon A, Alavi SM, et al. Genetic structure analysis of strains causing citrus canker in Iran reveals the presence of two different lineages of *Xanthomonas citri* pv. *citri* pathotype A*. *Plant Pathol*. 2015;64(4):776–84. <https://doi.org/10.1111/ppa.12324>.
- Vernière C, Hartung JS, Pruvost OP, Civerolo EL, Alvarez AM, Maestri P, et al. Characterization of phenotypically distinct strains of *Xanthomonas axonopodis* pv. *citri* from Southwest Asia. *Eur J Plant Pathol*. 1998;104(5):477–87.

13. Rybak M, Minsavage GV, Stall RE, Jones JB. Identification of *Xanthomonas citri* ssp. *citri* host specificity genes in a heterologous expression host. *Mol Plant Pathol.* 2009;10(2):249–62.
14. Sun XA, Stall RE, Jones JB, Cubero J, Gottwald TR, Graham JH, et al. Detection and characterization of a new strain of citrus canker bacteria from key Mexican lime and Alemow in South Florida. *Plant Dis.* 2004;88(11):1179–88.
15. Bui Thi Ngoc L, Vernière C, Jarne P, Brisse S, Guérin F, Boutry S, et al. From local surveys to global surveillance: three high throughput genotyping methods for the epidemiological monitoring of *Xanthomonas citri* pv. *citri* pathotypes. *Appl Environ Microbiol.* 2009;75(4):1173–84.
16. Gordon JL, Lefeuvre P, Escalon A, Barbe V, Cruveiller S, Gagnevin L, et al. Comparative genomics of 43 strains of *Xanthomonas citri* pv. *citri* reveals the evolutionary events giving rise to pathotypes with different host ranges. *BMC Genomics.* 2015;16(1):1098. <https://doi.org/10.1186/s12864-015-2310-x>.
17. Jeong K, Munoz-Bodnar A, Arias Rojas N, Poulin L, Rodriguez RL, Gagnevin L, et al. CRISPR elements provide a new framework for the genealogy of the citrus canker pathogen *Xanthomonas citri* pv. *citri*. *BMC Genomics.* 2019; 20(1):917. <https://doi.org/10.1186/s12864-019-6267-z>.
18. Pruvost O, Magne M, Boyer K, Leduc A, Tourterel C, Drevet C, et al. A MLVA genotyping scheme for global surveillance of the citrus pathogen *Xanthomonas citri* pv. *citri* suggests a worldwide geographical expansion of a single genetic lineage. *PLoS One.* 2014;9(6):e98129.
19. Schubert T, Rizvi S, Sun X, Gottwald T, Graham J, Dixon W. Meeting the challenge of eradicating citrus canker in Florida - again. *Plant Dis.* 2001; 85(4):341–56.
20. Leduc A, Traoré YN, Boyer K, Magne M, Grygiel P, Juhasz C, et al. Bridgehead invasion of a monomorphic plant pathogenic bacterium: *Xanthomonas citri* pv. *citri*, an emerging citrus pathogen in Mali and Burkina Faso. *Environ Microbiol.* 2015;17(11):4429–42.
21. Richard D, Boyer C, Javegny S, Boyer K, Grygiel P, Pruvost O, et al. First report of *Xanthomonas citri* pv. *citri* pathotype A causing Asiatic citrus canker in Martinique, France. *Plant Dis.* 2016;100(9):1946.
22. Cubero J, Graham JH. Genetic relationship among worldwide strains of *Xanthomonas* causing canker in citrus species and design of new primers for their identification by PCR. *Appl Environ Microbiol.* 2002;68(3):1257–64.
23. Fonseca NP, Felestrino EB, Caneschi WL, Sanchez AB, Cordeiro IF, Lemes CGC, et al. Detection and identification of *Xanthomonas* pathotypes associated with citrus diseases using comparative genomics and multiplex PCR. *PeerJ.* 2019;7:e7676. <https://doi.org/10.7717/peerj.7676>.
24. Hartung JS, Pruvost OP, Villemot I, Alvarez A. Rapid and sensitive colorimetric detection of *Xanthomonas axonopodis* pv. *citri* by immunocapture and a nested-polymerase chain reaction assay. *Phytopathology.* 1996;86(1):95–101.
25. Kingsley MT, Fritz LK. Identification of the citrus canker pathogen *Xanthomonas axonopodis* pv. *citri* A by fluorescent PCR assays. *Phytopathology.* 2000;90:542.
26. Miyoshi T, Sawada H, Tachibana Y, Matsuda I. Detection of *Xanthomonas campestris* pv. *citri* by PCR using primers from the spacer region between the 16S and 23S rRNA genes. *Ann Phytopath Soc Jpn.* 1998;64(4):249–54.
27. Park DS, Hyun JW, Park YJ, Kim JS, Kang HW, Hahn JH, et al. Sensitive and specific detection of *Xanthomonas axonopodis* pv. *Citri* by PCR using pathovar specific primers based on hrpW gene sequences. *Microbiol Res.* 2006;161:145–9.
28. Cubero J, Graham JH. Quantitative real-time polymerase chain reaction for bacterial enumeration and allelic discrimination to differentiate *Xanthomonas* strains on citrus. *Phytopathology.* 2005;95(11):1333–40.
29. Mavrodieva V, Levy L, Gabriel DW. Improved sampling methods for real-time polymerase chain reaction diagnosis of citrus canker from field samples. *Phytopathology.* 2004;94(1):61–8.
30. Rigano LA, Marano MR, Castagnaro AP, Do Amaral AM, Vojnov AA. Rapid and sensitive detection of citrus bacterial canker by loop-mediated isothermal amplification combined with simple visual evaluation methods. *BMC Microbiol.* 2010;10(1):176.
31. Delcourt S, Vernière C, Boyer C, Pruvost O, Hostachy B, Robène-Soustrade I. Revisiting the specificity of PCR primers for diagnostics of *Xanthomonas citri* pv. *citri* by experimental and *in silico* analyses. *Plant Dis.* 2013;97(3):373–8.
32. Muska A, Peck E, Palmer S. Standards and controls: concepts for preparation and use in real-time PCR application. In: Mackay IM, editor. *Real-time PCR in microbiology: from diagnosis to characterization.* Norfolk, UK: Caister Academic Press; 2007. p. 101–31.
33. Ios R, Fabre B, Saurat C, Fourrier C, Frey P, Marçais B. Development, comparison, and validation of real-time and conventional PCR tools for the detection of the fungal pathogens causing brown spot and red band needle blights of pine. *Phytopathology.* 2010;100(1):105–14.
34. Schenck N, Fourrier-Jeandel C, Ios R. A robust and specific real-time PCR tool for the detection of *Phytophthora lateralis* in plant tissues. *Eur J Plant Pathol.* 2016;146(2):231–44. <https://doi.org/10.1007/s10658-016-0909-7>.
35. Vallet D, Calteau A, Dubois M, Amours P, Bazin A, Beuvin M, et al. MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. *Nucleic Acids Res.* 2020;48(D1):D579–D89. <https://doi.org/10.1093/nar/gkz926>.
36. Weiss CL, Schuenemann VJ, Devos J, Shirsekar G, Reiter E, Gould BA, et al. Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. *R Soc Open Sci.* 2016;3(6):160239. <https://doi.org/10.1098/rsos.160239>.
37. Constantin EC, Cleenwerck I, Maes M, Baeyen S, Van Malderghem C, De Vos P, et al. Genetic characterisation of strains named as *Xanthomonas axonopodis* pv. *dieffenbachiae* leads to a taxonomic revision of the *X. axonopodis* species complex. *Plant Pathol.* 2016;65(5):792–806.
38. Patane JSL, Martins J Jr, Rangel LT, Belasque J, Digiampietri LA, Facincani AP, et al. Origin and diversification of *Xanthomonas citri* subsp. *citri* pathotypes revealed by inclusive phylogenomic, dating, and biogeographic analyses. *BMC Genomics.* 2019;20(1):700. <https://doi.org/10.1186/s12864-019-6007-4>.
39. Moreira AN, Laia ML, De Souza RF, Zaini PA, Da Silva AC, Da Silva AM, et al. Development and validation of a *Xanthomonas axonopodis* pv. *citri* DNA microarray platform (XACarray) generated from the shotgun libraries previously used in the sequencing of this bacterial genome. *BMC Res Notes.* 2010;3:150.
40. Abdulai M, Basim H, Saba CKS. Rapid identification and detection of *Xanthomonas phaseoli* pv. *manihotis*, causing bacterial blight disease in cassava by real-time PCR using LNA probe. *Int J Agric Biol.* 2020;23(2):259–68. <https://doi.org/10.17957/ijab/15.1284>.
41. Villela JGA, Ritschel P, Barbosa MAG, Baccin KMS, Rossato M, Maia JDG, et al. Detection of *Xanthomonas citri* pv. *viticola* on grapevine by real-time PCR and BIO-PCR using primers designed from pathogenicity and xanthomonadin gene sequences. *Eur J Plant Pathol.* 2019;155(2):445–59. <https://doi.org/10.1007/s10658-019-01779-y>.
42. Jouen E, Chiroleu F, Maillot-Lebon V, Chabirand A, Merion S, Boyer C, et al. A duplex quantitative real-time PCR assay for the detection and quantification of *Xanthomonas phaseoli* pv. *dieffenbachiae* from diseased and latently infected anthurium tissue. *J Microbiol Meth.* 2019;161:74–83.
43. Kulkarni YS, Patel MK, Abhyankar SG. A new bacterial leaf-spot and stem canker of pigeon pea. *Curr Sci.* 1950;19(12):384.
44. Burns DL. Biochemistry of type IV secretion. *Curr Opin Microbiol.* 1999; 2(1):25–9.
45. Caraguel CG, Stryhn H, Gagne N, Dohoo IR, Hammell KL. Selection of a cutoff value for real-time polymerase chain reaction results to fit a diagnostic purpose: analytical and epidemiologic approaches. *J Vet Diagn Invest.* 2011;23(1):2–15. <https://doi.org/10.1177/104063871102300102>.
46. Li W, Song Q, Brlansky RH, Hartung JS. Genetic diversity of citrus bacterial canker pathogens preserved in herbarium specimens. *Proc Natl Acad Sci U S A.* 2007;104(47):18427–32.
47. Yoshida K, Burbano HA, Krause J, Thines M, Weigel D, Kamoun S. Mining herbaria for plant pathogen genomes: back to the future. *PLoS Pathog.* 2014;10(4):e1004028. <https://doi.org/10.1371/journal.ppat.1004028>.
48. Lang JM, Hamilton JP, Diaz MGQ, Van Sluys MA, Burgos MRG, Vera Cruz CM, et al. Genomics-based diagnostic marker development for *Xanthomonas oryzae* pv. *oryzae* and *X. oryzae* pv. *oryzicola*. *Plant Dis.* 2010;94(3):311–9.
49. Langlois PA, Snelling J, Hamilton JP, Bragard C, Koebnik R, Verdier V, et al. Characterization of the *Xanthomonas translucens* complex using draft genomes, comparative genomics, phylogenetic analysis, and diagnostic LAMP assays. *Phytopathology.* 2017;107(5):519–27. <https://doi.org/10.1094/Phyto-08-16-0286-R>.
50. Triplett LR, Hamilton JP, Buell CR, Tisserat NA, Verdier V, Zink F, et al. Genomic analysis of *Xanthomonas oryzae* isolates from rice grown in the United States reveals substantial divergence from known *X. oryzae* pathovars. *Appl Environ Microbiol.* 2011;77(12):3930–7.
51. Anonymous. PM 7/98 (2) specific requirements for laboratories preparing accreditation for a plant pest diagnostic activity. *EPPD Bull.* 2014;44(2):117–47.

52. Grosdidier M, Aguayo J, Marçais B, Ios R. Detection of plant pathogens using real-time PCR: how reliable are late C-t values? *Plant Pathol.* 2017; 66(3):359–67. <https://doi.org/10.1111/ppa.12591>.
53. Nutz S, Doll K, Karlovsky P. Determination of the LOQ in real-time PCR by receiver operating characteristic curve analysis: application to qPCR assays for *Fusarium verticillioides* and *F. proliferatum*. *Anal Bioanal Chem.* 2011; 401(2):717–26.
54. Youden WJ. Index for rating diagnostic tests. *Cancer.* 1950;3(1):32–5.
55. Broeders S, Huber I, Grohmann L, Berben G, Taverniers I, Mazzara M, et al. Guidelines for validation of qualitative real-time PCR methods. *Trends Food Sci Tech.* 2014;37(2):115–26.
56. Burns M, Valvidia H. Modelling the limit of detection in real-time quantitative PCR. *Eur Food Res Technol.* 2008;226:1513–24.
57. Kolm C, Martzy R, Brunner K, Mach RL, Krška R, Heinze G, et al. A complementary isothermal amplification method to the U.S. EPA quantitative polymerase chain reaction approach for the detection of enterococci in environmental waters. *Environ Sci Technol.* 2017;51(12):7028–35. <https://doi.org/10.1021/acs.est.7b01074>.
58. Anonymous. PM7/76(5) use of EPPO diagnostic protocols. *EPPO Bull.* 2018; 48(3):373–7.
59. Pruvost O, Roumagnac P, Gaube C, Chiroleu F, Gagnevin L. New media for the semi-selective isolation and enumeration of *Xanthomonas campestris* pv. *mangiferaeindicae*, the causal agent of mango bacterial black spot. *J Appl Microbiol.* 2005;99(4):803–15.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



Chapitre 3 – Reconstruction de l’histoire d’une émergence à échelle locale

Les travaux présentés dans ce chapitre ont été menés dans l’objectif de préciser la reconstruction de l’histoire de l’émergence de *Xanthomonas citri* pv. *citri* (*Xci*) à une échelle locale, celle des îles du sud-ouest de l’océan Indien (SOOI).

Xci est présent dans l’ensemble des îles de la région du SOOI. La maladie du chancre asiatique des agrumes y a été décrite pour la première fois à Maurice en 1917 (Aubert, 2014), à Rodrigues en 1937 (Wiehe, 1941), en 1968 à La Réunion puis aux Comores (Brun, 1971). Depuis près de 35 ans, les chercheurs du CIRAD étudient cette maladie et ont collecté des souches de la bactérie *Xci* provenant des différentes îles du SOOI ainsi que d’autres régions du monde. Les génomes d’un certain nombre de ces souches ont précédemment été séquencés (près de 300) et analysés dans l’objectif de mieux comprendre les mécanismes d’émergence de cette bactérie au sein des îles du SOOI. Ainsi, dans une étude publiée par un précédent doctorant du laboratoire, il a été montré que les souches du SOOI formaient un groupe monophylétique au sein de la phylogénie mondiale de *Xci* (Richard *et al.*, 2020). Au sein des îles du SOOI, la variabilité des dates d’échantillonnage des souches (1978-2015) a permis à Richard *et al.* (2020) de co-estimer taux de substitution (0.43 [95%HPD 0.35-0.51] substitution/genome/an) et date de l’ancêtre commun le plus récent (MRCA, *Most Recent Common Ancestor*) des souches du SOOI (1818 [95%HPD 1762-1868]). Néanmoins, leurs résultats n’ont pas permis d’identifier la zone géographique d’émergence de la maladie au sein des îles du SOOI avec certitude.

A partir d’un spécimen d’herbier de *Citrus* sp. collecté en 1937 à Maurice, et grâce aux optimisations moléculaires et bioinformatiques présentées dans le chapitre 2, nous avons réussi à reconstruire le premier génome historique d’une souche de *Xci* (appelé HERB_1937_*Xci*, qui, à notre connaissance, constitue également le premier génome d’une bactérie pathogène des plantes reconstruit à partir d’un spécimen d’herbier). Nous avons analysé les données moléculaires ainsi générées de différentes façons. Dans un premier temps, nous avons cherché à décrire la composition taxonomique de notre échantillon historique d’herbier, que nous avons pu montrer être constituée d’ADN de *Citrus*, de *Xci*, de divers microorganismes connus pour être associés au microbiote des *Citrus* et également d’ADN humain. Deuxièmement, nous avons analysé les patrons de dégradation de l’ADN afin d’authentifier la nature historique de HERB_1937_*Xci* et d’exclure l’occurrence majeure de contaminations au laboratoire. De façon intéressante, cette analyse nous a permis de relever des différences significatives de désamination entre chromosome et plasmides. Ensuite, nous avons testé si l’ajout d’un échantillon

historique pouvait modifier les résultats des inférences phylogénétiques réalisées à partir de génomes contemporains seulement. Nos résultats ont montré que l'intégration de HERB_1937_Xci amenait à l'obtention d'un taux de substitution légèrement plus élevé et d'une date de diversification du MRCA des souches du SOOI légèrement plus récente, avec des intervalles de confiance associés plus réduits. De plus, la topologie de l'arbre phylogénétique obtenue, combinée avec l'analyse de la structuration de la diversité phylogénétique des souches au sein des différentes îles, nous a permis de suggérer une émergence à Maurice selon un (ou quelques) événement(s) d'introduction de souches génétiquement proches associés à des événements historiques majeurs dans la zone du SOOI. Pour finir, nous avons identifié le contenu en gènes de HERB_1937_Xci, en portant particulièrement attention aux facteurs de virulence, et avons montré qu'il était globalement très proche de celui de ces homologues modernes.

Ces résultats sont présentés dans un article publié en 2021 dans la revue PLOS Pathogens : « ***First historical genome of a crop bacterial pathogen from herbarium specimen: Insights into citrus canker emergence*** ». Il représente une « preuve de concept » illustrant la possibilité de générer un génome historique bactérien à partir de spécimen d'herbier, décrivant le panel d'analyses génétiques pouvant être réalisées et démontrant l'apport de tels échantillons dans la résolution et la précision des inférences associées à l'étude de l'histoire évolutive et d'émergence des pathogènes des cultures.

RESEARCH ARTICLE

First historical genome of a crop bacterial pathogen from herbarium specimen: Insights into citrus canker emergence

Paola E. Campos^{1,2}, Clara Groot Crego¹, Karine Boyer¹, Myriam Gaudeul^{2,3}, Claudia Baider⁴, Damien Richard¹, Olivier Pruvost¹, Philippe Roumagnac^{5,6}, Boris Szurek⁵, Nathalie Becker², Lionel Gagnevin^{5,6}, Adrien Rieux¹*

1 CIRAD, UMR PVBMT, Saint-Pierre, La Réunion, France, **2** Institut de Systématique, Évolution, Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, SU, EPHE, UA, Paris, France, **3** Herbarium national (P), Muséum national d'Histoire naturelle, Paris, France, **4** Ministry of Agro Industry and Food Security, Mauritius Herbarium, R.E. Vaughan Building (MSIRI compound), Agricultural Services, Réduit, Mauritius, **5** PHIM Plant Health Institute, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France, **6** CIRAD, UMR PHIM, Montpellier, France

✉ These authors contributed equally to this work.
* adrien.rieux@cirad.fr



OPEN ACCESS

Citation: Campos PE, Groot Crego C, Boyer K, Gaudeul M, Baider C, Richard D, et al. (2021) First historical genome of a crop bacterial pathogen from herbarium specimen: Insights into citrus canker emergence. *PLoS Pathog* 17(7): e1009714. <https://doi.org/10.1371/journal.ppat.1009714>

Editor: David Mackey, The Ohio State University, UNITED STATES

Received: November 2, 2020

Accepted: June 14, 2021

Published: July 29, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.ppat.1009714>

Copyright: © 2021 Campos et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The authors confirm that all data underlying the findings are fully available without restriction. HERB_1937 raw reads were deposited to the Sequence Read Archive

Abstract

Over the past decade, ancient genomics has been used in the study of various pathogens. In this context, herbarium specimens provide a precious source of dated and preserved DNA material, enabling a better understanding of plant disease emergences and pathogen evolutionary history. We report here the first historical genome of a crop bacterial pathogen, *Xanthomonas citri* pv. *citri* (*Xci*), obtained from an infected herbarium specimen dating back to 1937. Comparing the 1937 genome within a large set of modern genomes, we reconstructed their phylogenetic relationships and estimated evolutionary parameters using Bayesian tip-calibration inferences. The arrival of *Xci* in the South West Indian Ocean islands was dated to the 19th century, probably linked to human migrations following slavery abolishment. We also assessed the metagenomic community of the herbarium specimen, showed its authenticity using DNA damage patterns, and investigated its genomic features including functional SNPs and gene content, with a focus on virulence factors.

Author summary

Herbarium collections are a precious resource to plant pathologists, tracking crop diseases on specimens collected in the past centuries. In addition to indicating the presence of a disease at a specific time and locality, recent molecular technologies now allow extraction and microbial DNA sequencing from dead specimens. Despite challenges due to the degraded nature of DNA retrieved from historical samples, we were able to reconstruct the genome of a pathogenic bacterium from a 1937 herbarium specimen collected in Mauritius: *Xanthomonas citri* pv. *citri*, responsible for Asiatic citrus canker (ACC, an economically important agricultural disease controlled mostly through prophylactic and quarantine measures). Enhanced knowledge about the epidemiology and evolution of this

(SRR12792042). Consensus historical genome reconstructed for chromosome, plasmids pXAC33 and pXAC64 have also been deposited on GenBank database (CP072205-CP072207). The modern genomes used in this study have previously been published in the NCBI GenBank repository under accession numbers listed in S1 Table. Accession numbers of any previously published data used in this study are listed in [Supplementary information](#).

Funding: This work was financially supported by l'Agence Nationale pour la Recherche (AR: JCJC MUSEOBACT contrat ANR-17-CE35-0009-01; <https://anr.fr/>), the European Regional Development Fund (AR, KB, OP, NB: ERDF contract GURDT I2016-1731-0006632; <https://www.europe-en-france.gouv.fr/fr/fonds-europeens/fonds-europeen-de-developpement-regional-FEDER>), Région Réunion (AR, KB, OP, NB; <https://www.regionreunion.com/>), the French Agropolis Foundation Labex Agro—Montpellier (AR, OP, PR, BS, NB, LG: E-SPACE project number 1504-004) & (AR, PR, BS, LG: MUSEOVIR project number 1600-004; <https://www.agropolis-fondation.fr/?lang=en>), the SYNTHESYS Project (LG: grant GB-TAF-6437 & AR: grant GB-TAF-7130; <http://www.synthesys.info/>), the COST Action (LG, BS: grant CA16107 EuroXanth supported by COST; <https://www.cost.eu/>) & CIRAD/AI-CRESI (AR, PR, BS, LG : grant 3/2016; <https://www.cirad.fr/en/home-page>). PhD of PC. was co-funded by ED 227, Muséum national d'Histoire naturelle et Sorbonne Université, French Ministry of Higher Education, Research and Innovation, France. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

bacterial pathogen is valuable to improve these measures. We compared the genome of this 1937 bacterial strain to a collection of modern strains, included it in a tree representing their genetic relationships, and calculated both evolutionary mutation rate and divergence times. This “forensic investigation” informs us about how and when the disease developed in the South West Indian Ocean Islands. We hypothesize that there was a single (or a few related) introduction of ACC in Mauritius in the mid-19th century, followed by expansion to the neighbouring islands.

Introduction

Since the origins of agriculture, humanity has struggled with the incessant, devastating impact of plant diseases on food production [1]. As illustrated by the 19th century potato late blight epidemic caused by the oomycete *Phytophthora infestans* [2], crop pathogens have been responsible for tremendous losses, resulting in starvation for millions of human beings and massive migrations. Today, up to 40% of yield losses among major cultivated crops are associated with plant pathogens and pests with major economic impact [3]. Simultaneously, more than 800 million people remain chronically undernourished worldwide [4]. It is also widely acknowledged that the extensive use of pesticides against crop pathogens is detrimental to the environment, affects public health and threatens biodiversity [5].

In order to most effectively manage current infectious crop diseases and prevent future epidemics, a better understanding of the factors underlying pathogen emergence, adaptation and spread is necessary [1,6]. As sequencing technologies have become more accessible, genetic analyses have played an increasingly important role in infectious disease research. Whole genome sequencing of pathogens can confirm suspected cases of an infectious disease, discriminate between different strains, classify novel pathogens and reveal virulence mechanisms in a time- and cost-efficient manner [7,8]. In addition to examining individual pathogen sequences, multiple sequences can be combined within phylogenetic methods to assess population structure, elucidate evolutionary/transmission history and infer several demographic, evolutionary or epidemiological parameters [9,10]. Until recently, most studies were performed on field-sampled contemporary individuals over a time interval of maximum four decades. Although such studies grant a good understanding of the population structure and recent emergences of pathogens, such small temporal differences between samples do neither allow a thorough detection of measurable amounts of evolutionary changes, nor a reconstruction of deeper evolutionary timelines [11], leaving many questions on crop pathogen emergence unanswered. With the first studies on ancient DNA (hereafter aDNA) obtained from historical or ancient samples such as archaeological tissue remains or museum specimens [12], it became possible to explore the past from a genetic perspective.

The few studies performed on crop pathogens from herbarium specimens worldwide have emphasized the role of historical collections for understanding the evolution and epidemiology of plant pathogens [13–17]. First, the observation of disease symptoms associated with herbarium specimen information (collection date, geographic location, host species or other phenotypic traits) may allow a direct update of past disease occurrence, distribution and host range. For instance, Antonovics *et al.* [18] made use of infected *Silene* sp. historical specimens to survey the incidence of anther smut disease and showed a possible change of host range of this disease in the Eastern USA. Second, recent molecular developments have allowed a more efficient retrieval and sequencing of low quantity, short and degraded nucleic acids from historical desiccated plant tissues [19]. Historical and modern genomes can then be compared to detect changes in genetic contents and arrangement over time, such as the loss or gain of

functional genes or the change of ploidy levels, for both pathogens and their host plants [20]. Moreover, by expanding the temporal range between samples, the chance of detecting evolutionary changes, *i.e.* temporal signal, increases [21]. Compared to their most recent common ancestor, modern genomes are expected to have accumulated more mutations than their historical counterparts. These differences can be used to directly infer mutation rates, divergence time between lineages and sudden changes in genetic diversity [14,22], which can be correlated with historical and socioeconomic events. Such analyses performed on historical DNA sequences of *P. infestans* retrieved from 19th century herbarium specimens resolved the debated origin and identity of the strain that caused the 1840s late blight pandemic [14,17,23–25]. Although reconstructions of crop pathogen history using full genomic sequences have been successfully realized on historical oomycetes [14,23] and viruses [26–28], such an achievement has not yet been reported for a bacterial crop pathogen, for which only few genetic markers were previously exploited [29].

In this work, we focus on *Xanthomonas citri* pv. *citri* (*Xci*), the bacterium responsible for the Asiatic citrus canker (ACC) [29,30]. ACC causes important economic losses in most citrus-producing areas worldwide, both by decreasing fruit yield and quality, but also by limiting exportations due to its quarantine status [31]. The earliest records of ACC, dating back to 1812–1844, are in herbarium specimens from Indonesia and India [32], suggesting an Asiatic origin of *Xci* [33–35]. From there, although without direct evidence, *Xci* would have spread through multiple dispersal events over time, leading to its current broad distribution across continents and islands. In this context, a comparison of *Xci* multilocus genotypes retrieved from herbarium specimens suggested Japan as being the source of the 1911 ACC original outbreak in Florida [29]. Aiming for a refined chronology of these spreading events within the South West Indian Ocean (SWIO) area, where *Xci* diversity is well-documented [35–38], we focused on SWIO herbarium specimens.

We report the first genome of a historical bacterial pathogen retrieved from a citrus herbarium specimen collected in 1937 in Mauritius, 20 years after the first report of ACC on this island [39]. We studied the metagenomic composition of this herbarium sample and showed its authenticity as aDNA material by assessing damage patterns. Using tip-calibrated phylogenetic inferences performed with both the 1937 historical strain and a large set of modern genomes, we elucidated the emergence history of *Xci* in the SWIO islands and further analyzed its genomic characteristics, with a particular focus on virulence factors.

Results

Laboratory procedures & high-throughput sequencing

Herbarium specimen MAU 0015151 *Citrus* sp. from 1937, Mauritius (hereafter HERB_1937, Fig 1) was sampled from the Mauritius Herbarium collections (<https://herbaria.plants.ox.ac.uk/bol/mau>) and chosen for this study as the most ancient symptomatic herbarium specimen available from the SWIO area. It precedes the oldest culture available from this island by ~50 years. DNA was carefully extracted in a bleach-cleaned facility with no prior exposure to modern *Xci* DNA using an optimized protocol (see [Material & Methods](#)). Extracted DNA (yield of 0.75 ng per mg of leaf tissue) was shown with a specific and exclusive qPCR diagnostic assay [40] to contain *Xci* DNA, roughly equivalent to 3×10^5 CFU/cm² (average C_T of 32.0, C_T cut-off = 35.4 and no-C_T value for the negative control). Total DNA was then converted into an Illumina library, and sequencing generated 220.9 M paired-end reads with a base call accuracy of 99.90 to 99.96%. Following adaptor trimming and quality checking, insert reads were 59 ± 24 nt long and underwent four main analyses: metagenomic inference, ancient DNA authentication, comparative genomics and phylogenetic analyses, as summarized in Fig 2.



Fig 1. *Citrus* sp. specimen MAU 0015151 (HERB_1937), Mauritius Herbarium. MAU 0015151 *Citrus* sp. specimen (HERB_1937) was collected from Mauritius in 1937 and deposited in the Mauritius Herbarium. Leaf areas displaying typical symptoms of Asiatic citrus canker are highlighted with blue dotted frames.

<https://doi.org/10.1371/journal.ppat.1009714.g001>

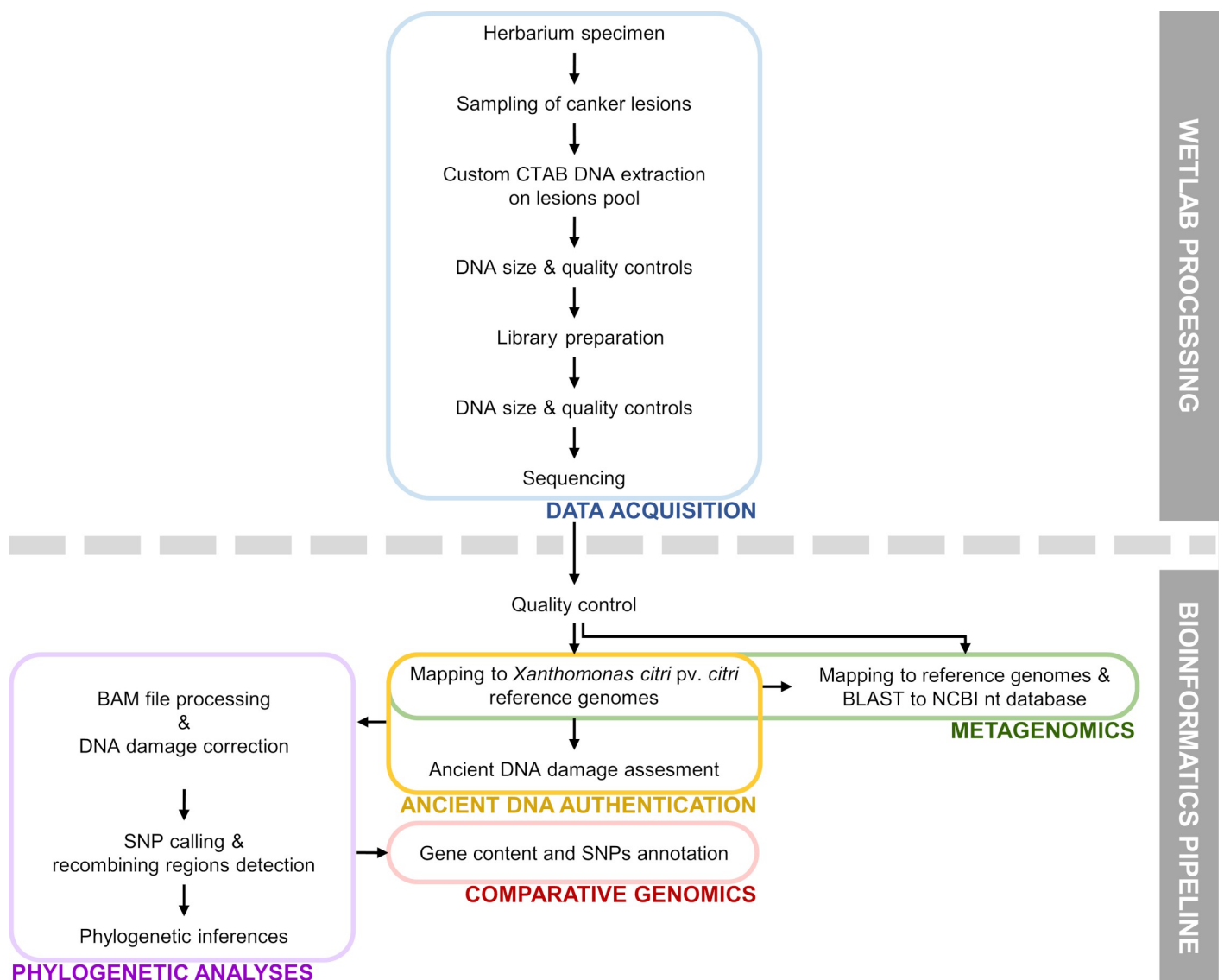


Fig 2. Major steps performed for characterization and integration of our herbarium sample into genomic analyzes. See [Material & Methods](#) for more details on the workflow processed for HERB_1937 in this study. Abbreviations: CTAB, cetyl trimethylammonium bromide; DNA, deoxyribonucleic acid; BAM file, binary alignment map file; BLAST, basic local alignment search tool; NCBI nt database: national center for biotechnology information nucleotide database; SNP, single-nucleotide polymorphism.

<https://doi.org/10.1371/journal.ppat.1009714.g002>

Importantly, no *Citrus* nor *Xci*-specific DNA fragments were found in our negative control, thus ruling out in-lab contamination.

Metagenomic composition

DNA extracted from leaf lesions is expected to originate from different sources. Using a combination of mapping- and BLAST-based approaches (as detailed in Material & Methods section), we studied the metagenomic composition of the reads obtained from HERB_1937. Identified sequences mostly consisted of the *Citrus* plant host genus (21.0%), followed by, at the species level, *Homo sapiens* (5.4%), and *Xci* found in 1.2% of the reads (Fig 3). Other reads were

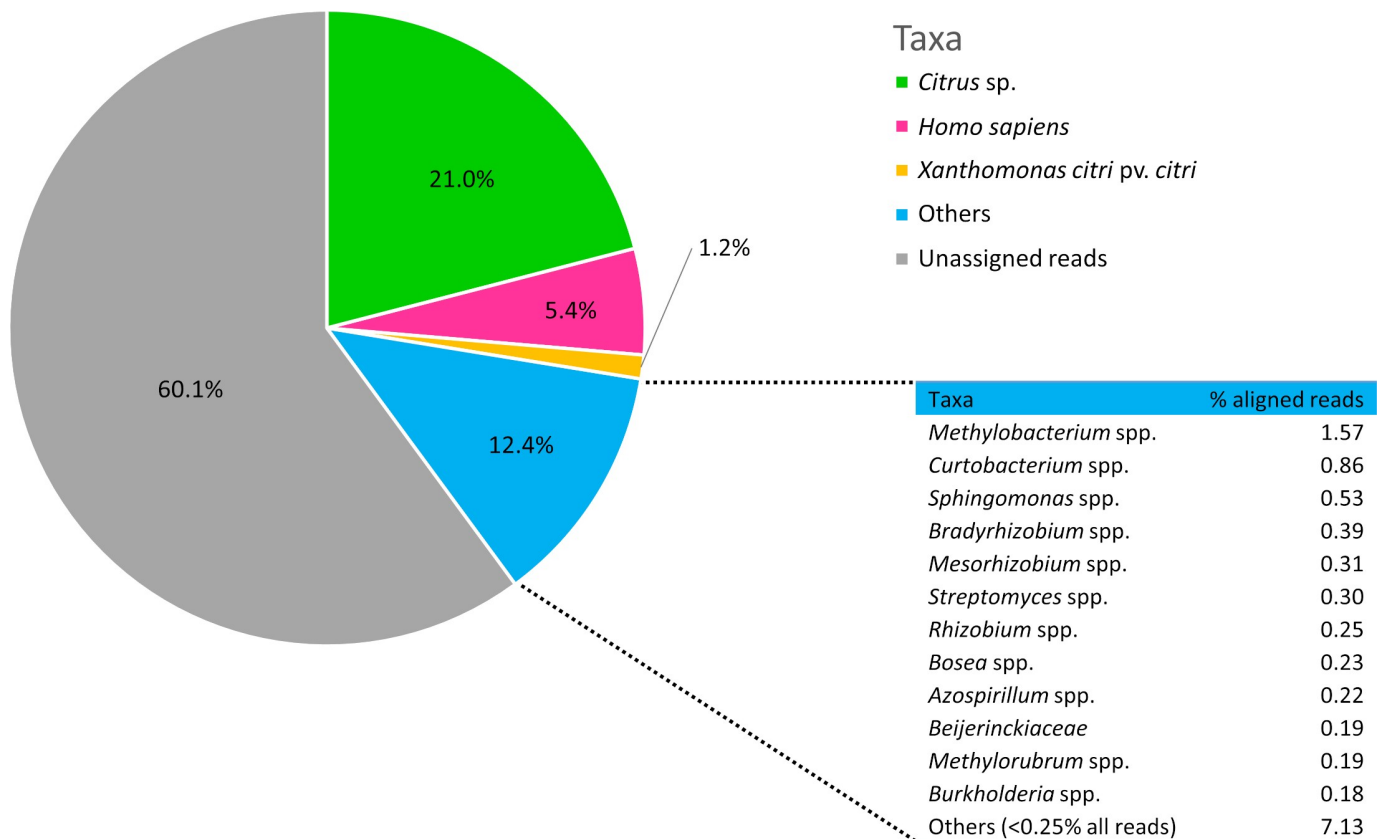


Fig 3. Metagenomic composition of HERB_1937 historical specimen. Proportions of reads assigned to *Homo sapiens* (5.4%), *Citrus* sp. (21.0%), *Xci* (1.2%). Others (12.4%): reads unassigned at the species level; unassigned reads (60.1%). Table: reads unassigned at the species level were assigned to the family (for *Beijerinckiaceae*) or genus level (for all others) and belong, for 0.18% to 1.57% of the aligned reads to the domain bacteria; “Others (<0.25% all reads)” include reads assigned to different plant, fungi, vertebrate, bacteria and phage genera (each identified genus totalizing less than 0.25% of all reads).

<https://doi.org/10.1371/journal.ppat.1009714.g003>

assigned to higher taxonomic levels, corresponding to one bacterial family and eleven different genera (from 0.18% (*Burkholderia*) to 1.57% (*Methylobacterium*) of aligned reads). Plant, fungi, vertebrate, bacteria and phage genera were marginally found (less than 0.25% of aligned reads for each genera). Altogether, reads unassigned to the species level added up to 12.4% of the reads. More than half (60.1%) of the reads were not assigned to any known taxa (Fig 3).

Historical genome reconstruction & characterization

A high quality *Xci* genome was reconstructed from HERB_1937 (hereafter called HERB_1937_*Xci*) by mapping the reads (discarding the 5 terminal nucleotides) to *Xci* IAPAR 306 reference sequences [41]. 0.74% (N = 1,628,776) of the total number of reads mapped to the *Xci* reference genome, a value unsurprisingly smaller than the 1.2% found with the “metagenomic pipeline” which combined both mapping and BLAST-n approaches. The reference chromosome was covered by a depth (the number of mapped reads at a given position) of at least 1X for 94% of its sequence, and displayed a mean depth of ~6X. Both pXAC33 and pXAC64 plasmids displayed a higher mean depth and larger non-covered regions (Fig 4 and Table 1). As non-covered positions can be caused by the absence of genes in the historical strain compared to the modern reference, but also by reads mapping ambiguously to multiple positions (repeated regions or replicated genes), we further characterized these loci (see gene content & virulence factors section).

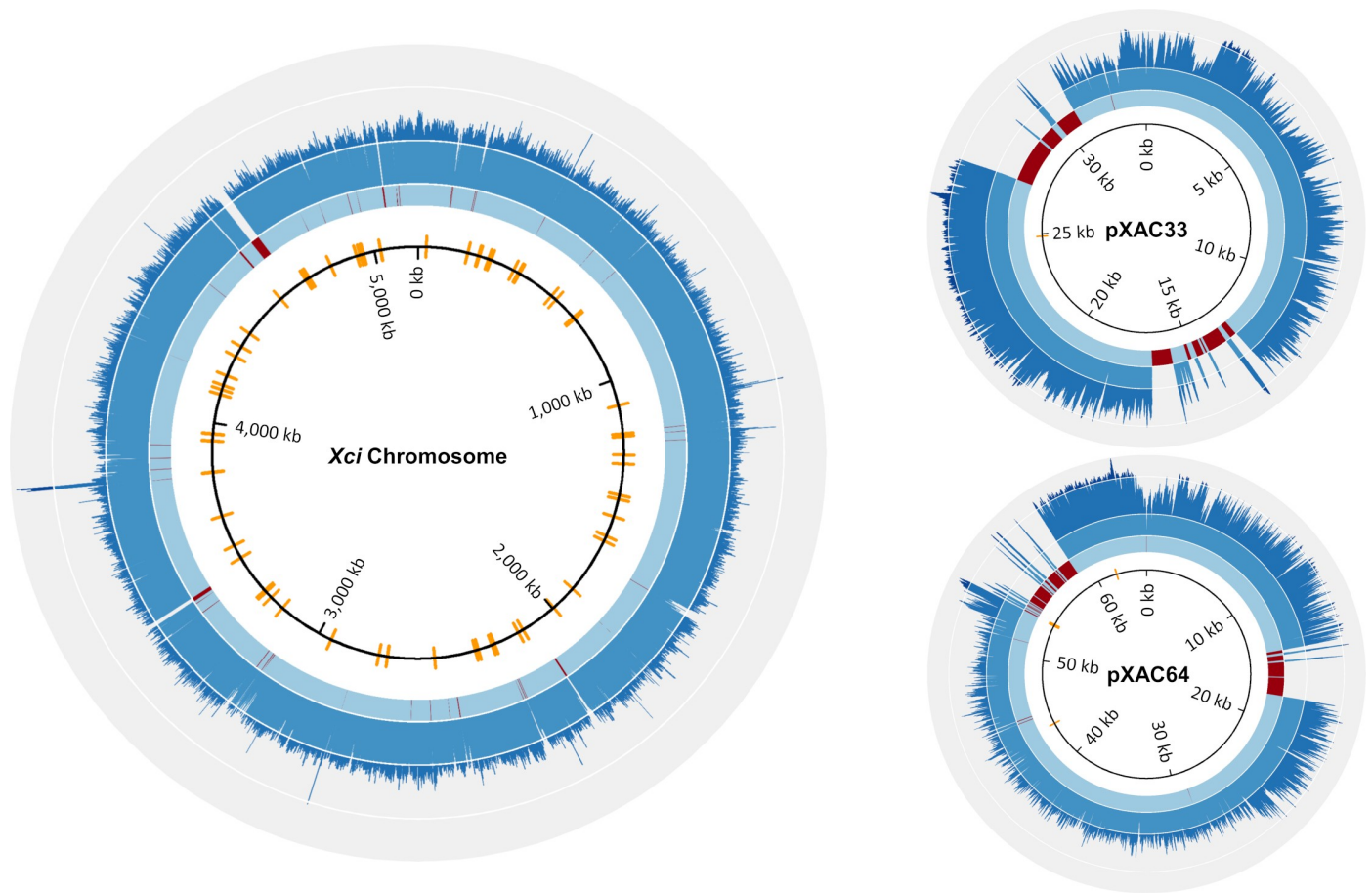


Fig 4. Coverage plots for the reconstructed HERB_1937_Xci chromosome and plasmids (pXAC33, pXAC64) sequences. From inside to outside, a light to dark blue scale (delimited by a white line) represents 1, 1–5, 5–15, 15–35-fold coverage (*Xci* chromosome) and 1, 1–5, 5–35, 35–90-fold coverage (plasmids). Red rings indicate no identified coverage (depth = 0). SNP positions between the respective reconstructed and reference sequences are indicated (orange line). Accession numbers for *Xci* reference strain IAPAR 306: NC_003919.1 (chromosome), NC_003921.3 (plasmid pXAC33) and NC_003922.1 (plasmid pXAC64).

<https://doi.org/10.1371/journal.ppat.1009714.g004>

Table 1. Summary of mapping, depth coverage and damage statistics for the reconstructed HERB_1937_Xci genome. Mapping, depth, coverage and damage statistics (read length, purine enrichment and deamination rate) are indicated for HERB_1937_Xci chromosome and plasmids (pXAC33, pXAC64). nt: nucleotides, SD: standard deviation.

Genome	Endogenous <i>Xci</i> DNA (%) [*]	Mean depth ^{**}	Coverage (%) ^{***}				Read length (nt)		Purine frequency enrichment at position -1		Deamination rate at terminal position (%)	
			0X	1X	5X	10X	Mean	SD	Mean	SD	5'C/T	3'G/A
Chromosome	0.71	5.9	5.8	94.2	53.0	7.0	42.75	12.64	1.79	0.00	2.25	2.35
pXAC33	0.01	21.9	17.4	82.6	80.2	75.1	45.43	14.21	1.76	0.05	2.91	2.96
pXAC64	0.02	17.3	11.5	88.5	82.3	63.6	45.20	13.85	1.77	0.01	2.65	2.73
Mean							44.46	13.57	1.77	0.03	2.60	2.68

^{*} Reads mapping to *Xci* reference genome/total reads before duplicate removal, expressed in %.

^{**} Average number of mapped reads at each base of the reference genome.

^{***} Percentage of reference genome covered at nX depth.

<https://doi.org/10.1371/journal.ppat.1009714.t001>

Ancient DNA damage assessment

Ancient DNA is typically degraded, presenting short fragments, excess of purine bases before DNA breaking points and cytosine deamination at fragment extremities [12,42]. We searched for such patterns of degradation in HERB_1937_Xci using the dedicated tool map-Damage2 [43]. The mean read length of HERB_1937_Xci reads was 44.5 ± 13.5 nt, showing substantial fragmentation (Fig 5A). DNA fragmentation being partially caused by depurination, we also examined the nucleotidic context surrounding 5' end DNA breakpoints. We found a mean relative purine enrichment of 1.77 ± 0.03 between upstream positions -1 and -5 of HERB_1937_Xci reads (Table 1). Modern strains, fragmented by enzymatic digestion prior to library construction, displayed no such enrichment (0.87 ± 0.01). Cytosine deamination was investigated by monitoring 5'C/T substitutions *versus* complementary 3'G/A substitutions, a classical analysis for double-stranded, blunt-ended libraries constructed prior to sequencing. Mean deamination rates of HERB_1937_Xci reads reached maximal values at the terminal nucleotide ($2.64 \pm 0.29\%$, Table 1 and Fig 5B). For statistical analyses, we took into account the five successive extreme positions of the reads, harbouring a significant increase between each nucleotide (outwards) along the five first or last positions of the reads (Wilcoxon matched-pairs signed rank test, 2-tailed p-value = 0.0313). The maximal rate among reads of three modern Xci controls displayed a significantly lower value of $0.10 \pm 0.07\%$ ($p < 0.0001$, unpaired 2-tailed Mann-Whitney test, Fig 5B). The apparent lower deamination rate for HERB_1937_Xci chromosome reads, as compared to both plasmids (Table 1), was analyzed similarly, along the five first or last positions of the reads. Interestingly, we found a significantly lower deamination rate for HERB_1937_Xci chromosome reads (1.6%) as compared to both plasmids (1.9%) (Wilcoxon matched-pairs signed rank test, 2-tailed p-value = 0.002). In

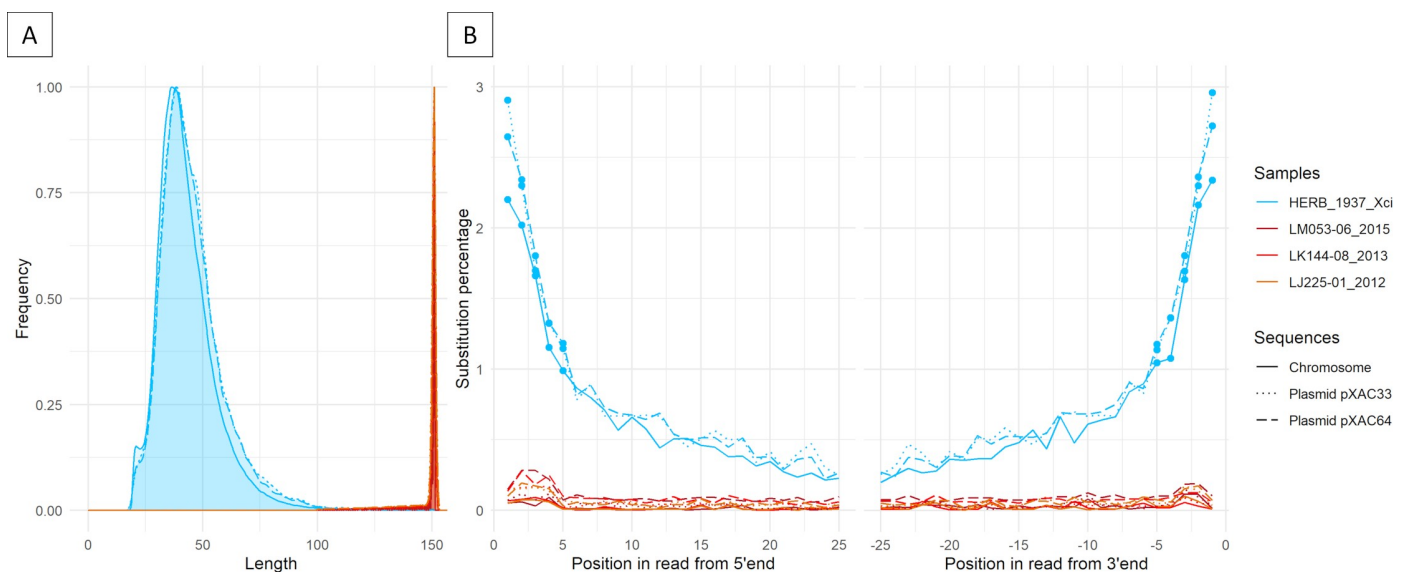


Fig 5. HERB_1937_Xci post-mortem DNA damage patterns. Post-mortem DNA damage patterns were measured on historical HERB_1937_Xci (full, dotted or dashed blue lines for chromosome, pXAC33 and pXAC64 respectively) and compared with three modern Xci strains isolated from SWIO in 2012, 2013 and 2015 respectively (red lines, see results and S1 Table for full description). (A) Fragment length distribution (nucleotides; relative frequency in arbitrary units). (B) Deamination percentages of the first 25 nucleotides from the 5' (C to T substitutions) and 3' (G to A substitutions) ends, respectively. Dots: five most extreme nucleotides of the reads, showing a significant increase (towards the extremity) between each nucleotide along the five first or last positions of all HERB_1937_Xci reads. Along the five extreme nucleotides, reads matching to HERB_1937_Xci harboured significantly higher values than modern controls, and reads matching to the HERB_1937_Xci chromosome harboured significantly lower deamination rates than sequences matching to either plasmid (see results for statistics).

<https://doi.org/10.1371/journal.ppat.1009714.g005>

contrast, we observed similar fragment lengths for chromosome and plasmid reads (Table 1 and Fig 5).

We performed the same analyses using the *Methylobacterium* reference genome (*M. organophilum*, strain DSM 760), since 1.57% of the reads unassigned at the species level were attributed to this particular genus (Fig 3). Mean read length was estimated at 66 ± 25 nt, relative enrichment of purine frequency reached 1.80 ± 0.52 (as previously, between position -1 and -5), and deamination rates at the terminal nucleotides averaged $2.18 \pm 0.07\%$.

Gene content & virulence factors

Out of 5,125 coding sequences (CDS) of strain IAPAR 306, only 139 were covered on less than 75% of their length by HERB_1937_Xci reads and will hereafter be designed as non-covered (S2 Table). Ninety-five of those CDS are present in multiple copies in IAPAR 306 with a strong nucleotide identity, leading to ambiguous mapping and poor coverage and can be considered as false absences. Among those 95 CDS, 77 are predicted to encode for full-length, or fragments of transposases. Four correspond to highly identical copies of Transcription Activator-Like Effector (TALE) genes (see specific paragraph below). The remaining multi-copy genes code for the elongation factor Tu, a xylose isomerase, a filamentous haemagglutinin and seven hypothetical proteins. Forty-four IAPAR 306 CDS were non-covered because no homologous reads were present in our dataset. Most of them hypothetically code for proteins of unknown function, with the exception of an identified type I restriction-modification system (including DNA methylase, endonuclease and specificity determinant), and six recombinases or integrases. Interestingly, 28 successive non-covered CDS correspond to a 27-kb block present only in IAPAR 306 and a few of its close relatives, which contains six transposases, three recombinases and 19 proteins of unknown function.

We verified the presence or absence of specific genes whose products have proven or are hypothetically involved in the pathogenicity of *Xanthomonas*. In particular, the type III secretion system (T3SS) is a syringe-like apparatus which injects “effectors” directly into the plant cell to inhibit plant defences and contribute to symptom development [44]. For this we investigated the presence in HERB_1937_Xci of a group of 82 genes found in *Xci* or other *Xanthomonas* species [45], encoding either for the T3SS or for type III effectors (T3E, which may participate in pathogenicity) [46]. Reads from HERB_1937_Xci covered 57 of those CDS on more than 94% of their sequence (S3 Table), including the entire set (24 CDS) of genes necessary for the T3SS and 33 potential T3E genes. The coverage of the remaining 25 CDS, all virulence factors from other *Xanthomonas* but not present in *Xci* [46] reached a maximum of 45.1% of their length (S3 Table), indicating the absence of the corresponding genes.

On the plasmids, most of the non-covered positions of HERB_1937_Xci were localized in four regions coding for TALE proteins (Fig 4 and S2 Table). These peculiar T3Es are responsible for the development of canker symptoms on citrus [47]; as our samples harbored such symptoms, we expected to find homologs in HERB_1937_Xci. *Xci* injects these TALE into the plant cell, activating the host’s transcriptional machinery to its benefit [48,49]. *Tale* genes encode for transcription activator-like proteins containing an N-terminal domain responsible for translocation from the bacterium to the plant cell, a C-terminal domain containing nuclear localization signals and a eukaryotic transcription activation domain, flanking tandem repeats of 33–34 conserved amino acids (S1 Fig). These repeats are highly homologous except for two amino-acid residues (called Repeat Variable Di-residues: RVD) responsible for DNA-binding specificity. We hypothesized that reads corresponding to *tale* genes were initially not mapped due to their particular structure and multicopy nature. Hence, we realized specific alignments using the sequences coding for either the N-terminal domain, the C-terminal domain and the

repeat domain (reduced to a three-repeat string) of *tale* gene *pthA4* as three independent references to test for the presence of *tale* sequences in HERB_1937_Xci. Almost 6,000 newly mapped reads corresponding to the *tale* gene were recovered (S1 Fig), with a mean depth of 44X for both 5' and 3' ends, about two times the mean depth of plasmid sequences outside *tale* gene positions (~23X). Moreover, two loci on the 5' end sequence were biallelic, presenting either T or C bases with a T/C ratio of 43/57 and 35/65, respectively. One of these loci translated into a conservative amino-acid substitution, found elsewhere in TALEs of proteic databases. Taken together, these results suggest the existence of two to four different 5' end sequences of *tale* genes, and therefore as many *tale* genes, in HERB_1937_Xci historical genome. Finally, among the remaining reads corresponding to the central repeat domain, we identified eight patterns of nucleotides coding for the RVD (S4 Table). Interestingly, although the most prevalent are found in modern *Xci tale* genes in similar proportions [50], three RVDs are unreported in modern TALE.

SNPs, phylogenetic reconstruction and tip-dating at the SWIO scale

We localized the SNPs between HERB_1937_Xci and the IAPAR 306 reference genome (Fig 4). After filtration of dubious SNPs (i.e. eliminated because of low depth, heteroplasmy and/or proximity to another SNP), HERB_1937_Xci displayed 83 high-quality SNPs on its chromosome sequence, one and four in sequences corresponding to pXAC33 and pXAC64, respectively. Forty-three SNPs were non-synonymous substitutions on the chromosome reference sequence, one on pXAC33 and three on pXAC64. The SNPs found between HERB_1937_Xci and IAPAR 306 were not characterized any further; a more meaningful analysis was performed for SNPs identified between HERB_1937_Xci and its related SWIO clade.

A total of 2,634 high confidence SNPs were found within the alignment of HERB_1937_Xci historical chromosome with 116 modern samples from the SWIO islands. The ClonalFrameML [51] analysis identified a single 5.9 kb recombinant region including two SNPs, which were removed from further inferences. For the 2,632 recombination-free SNPs, we estimated a ratio of non-synonymous (dN) to synonymous (dS) changes of 4.16. We identified 15 SNPs unique to HERB_1937_Xci and restricted to chromosome sequences, among which 14 were attributed to coding sequences. Analysis of these SNPs led to the identification of three synonymous and 11 non-synonymous mutations, which were characterized at the protein level. Interestingly, seven of those reveal unique amino-acids at these positions (among similar but non-redundant proteins of the *Xanthomonas* genus identified by BLASTp), thus harboring previously unknown proteic features (S5 Table). We added an outgroup and built a Maximum-Likelihood (ML) phylogeny with RAxML [52], which placed HERB_1937_Xci historical sequence outside of the “modern” SWIO clade (S2 Fig). The ML tree was well-supported and structured in three lineages: a Mauritius lineage (lineage A), sister-group of the rest of the modern strains of the SWIO comprising two lineages, the first with strains from Mauritius and Reunion (lineage B), and the second with strains from all SWIO islands (lineage C) (S2 Fig).

As a requirement to perform tip-based calibration, we tested the presence of temporal signal in our tree with both a linear regression between samples ages and root-to-tip distance, and a date-randomization test [22]. Both statistical tests revealed the presence of temporal signal (i.e. progressive accumulation of substitutions over time) within the SWIO tree. The linear regression test displayed a significant positive slope (value = 19.236×10^{-5} , adjusted $R^2 = 0.270$ with a p-value = 2.07×10^{-10}), with HERB_1937_Xci showing clear evidence of branch shortening (Fig 6B). The date-randomization test of the inferred root age of the real versus date-randomized datasets showed no overlap (95% Highest Posterior Density, S3 Fig). Therefore, we

built a time-calibrated tree with BEAST [53], which was globally congruent (similar topology and node supports) with the ML tree (Fig 6A). Phylogenetic diversity of Mauritius island strains (1530.4) was significantly higher (p -value = 2.2×10^{-16}) than those calculated from the other islands (Reunion strains = 1518.7, Rodrigues = 691.1, Comoros = 1024.0 and Mayotte = 319.0). We inferred a root date of 1843 [95% HPD: 1803–1881] and a mean substitution rate of 9.4×10^{-8} [95% HPD: 7.3×10^{-8} – 11.4×10^{-8}] per site per year, with a standard deviation for the uncorrelated log-normal relaxed clock of 0.271 [95% HPD: 0.182–0.366] suggesting low heterogeneity amongst branches (Figs 6B and S4). To specifically evaluate the contribution of HERB_1937_Xci, we considered modern strains only: although the dataset still displayed temporal signal (slope value = 9.885×10^{-5} , adjusted $R^2 = 0.077$, p -value = 0.0009) (Fig 6B), the BEAST analysis performed under the same parameters yielded significantly different values. An older tree root date of 1800 [95% HPD: 1745–1852] was inferred, together with a lower mean substitution rate of 8.2×10^{-8} substitutions per site per year [95% HPD: 6.4×10^{-8} – 9.9×10^{-8}] and a standard deviation for the uncorrelated log-normal relaxed clock of 0.188 [95% HPD: 0.082–0.289] among branches (S4B Fig). In summary, when comparing the estimates of root ages—with and without HERB_1937_Xci in the datasets—our results indicate that integrating the historical sequence significantly improves the accuracy of the temporal inferences, with a reduction of the 95% HPD from ~107 years to ~78 years (Fig 6C).

Discussion

We sequenced the genome of HERB_1937_Xci, an historical strain of the crop bacterial pathogen *Xanthomonas citri* pv. *citri* (*Xci*) from an infected herbarium specimen sampled in 1937 in Mauritius. To our knowledge, HERB_1937_Xci is the first historical genome of a pathogenic bacterium obtained from herbarium material. Similar achievement has been previously successfully realized on viruses [28,54], oomycetes such as *Phytophthora infestans* [14,23,24], and more recently on cyanobacteria [55]. But for plant pathogenic bacteria in general, and more specifically for *Xci*, only multilocus genotyping data could be exploited from such historical material [29].

Adopting a shotgun-based deep sequencing strategy allowed us to describe the metagenomic diversity contained within our historical herbarium specimen. Among assigned reads, HERB_1937 displayed 1.2% of *Xci* DNA for 21.0% of *Citrus* sp. DNA, a pathogen/plant ratio in the range of those previously observed for *P. infestans*, a nonvascular pathogen isolated from infected herbarium potato leaves [20,23,24]. The microbial community also contained several bacterial genera, all described in NGS studies as part of the citrus leaf [56] or root [56,57] microbiota. The three most prominent genera (*Methylobacterium*, *Curtobacterium* and *Sphingomonas*, >0.5% of aligned reads) belong to the core citrus leaf microbiome [58], and the relative abundance of *Methylobacterium* reads among bacteria (11.7%) is consistent with studies on modern samples (from 5 to 58% [56]). These bacterial genera were thus likely associated to the living citrus plant, and/or to HERB_1937 sample, colonized during collection and storage in the herbarium. Bieker *et al.* [59], using deamination studies, identified a fungal species proposed to have colonized herbarium specimens shortly after collection. As illustrated by the typical aDNA patterns we observed for *Methylobacterium* spp., we may exclude recent or laboratory contaminations [60]. Interestingly, up to 5.4% of the reads were assigned to human DNA resulting from contaminations during specimen manipulation (collection, mounting or storage). Finally, a substantial amount (60.1%) of HERB_1937 were unassigned reads, reflecting either the incompleteness of the reference database as compared to the microbial diversity of the sample [61] or the difficulty for short reads to be assigned taxonomically, a typical result in ancient DNA research [62].

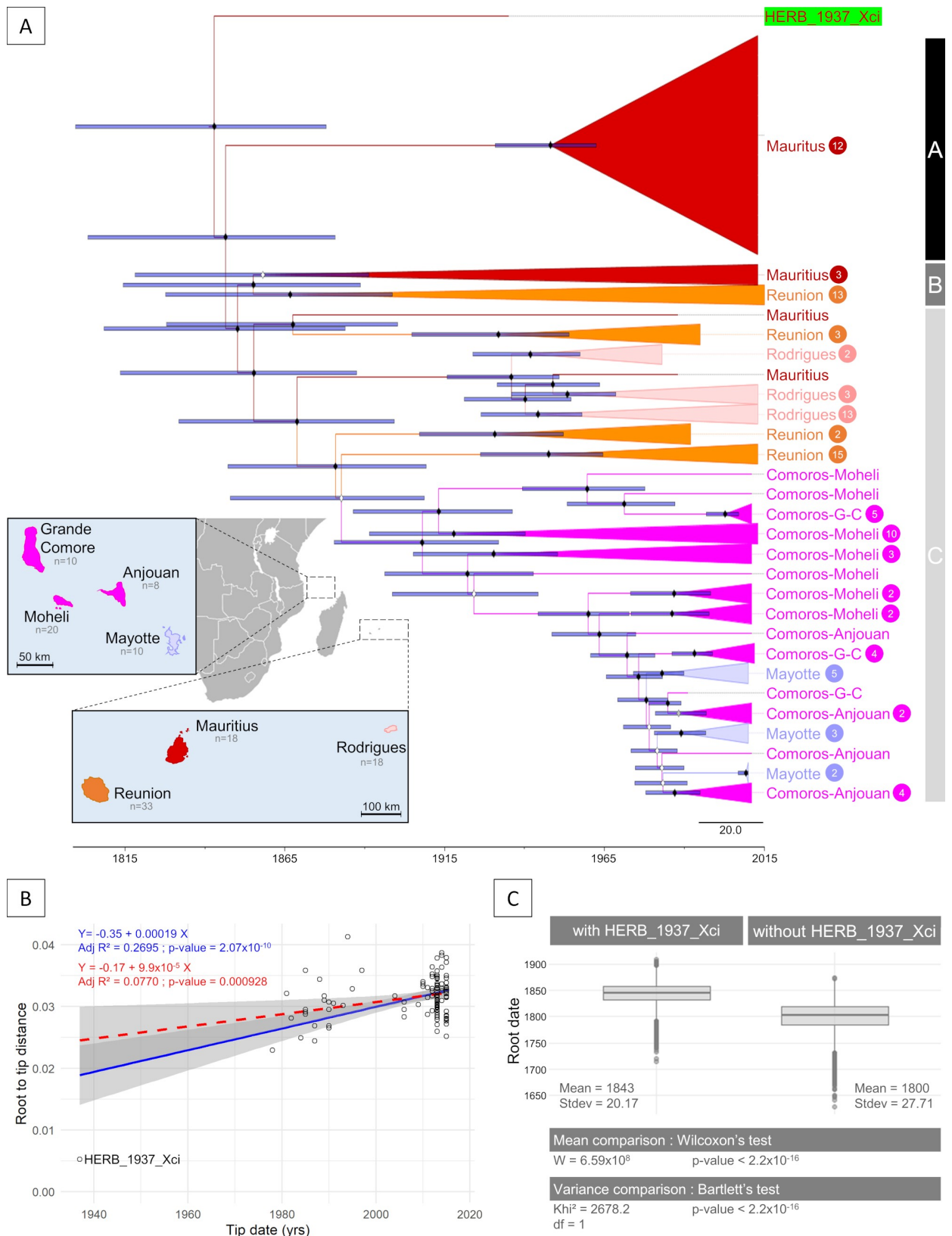


Fig 6. Tip-dating Bayesian inferences on historical and modern *Xci* genomes from the SWIO islands. (A) Dated BEAST tree of 116 *Xci* modern strains sampled from the SWIO islands between 1978 and 2015 with historical HERB_1937_*Xci* (highlighted in green) built from 2,632 non recombining SNPs. Node support values are displayed by diamonds, in white for Posterior Probabilities below 0.9, in black for values above 0.9; node bars cover 95% Highest Probability Density of node height. The tree is structured in three lineages (A, B & C). Branches are collapsed and colored, according to the sample's geographic origin, except lineage A, which is cartooned to help visualization. Tip labels include the geographic origin and, in cases of collapsed or cartooned branches, the number of samples. Map layer is from Natural Earth, available from <https://www.naturalearthdata.com>. (B) Linear regression of root-to-tip distance on year of sampling (tip date) test for temporal signal. Regression lines are plotted in blue when integrating historical HERB_1937_*Xci* genome and in red (dotted lines) when not. Grey areas correspond to their confidence interval. Associated values are the regression equation, adjusted R^2 (Adj R^2) and p-value. (C) Boxplot distribution of root age, with (left) and without (right) integrating historical HERB_1937_*Xci* in the dating inference, and associated statistical comparisons. Boxes represent 25th to 75th percentiles, Minimum-Maximum intervals are displayed by a vertical bar and outliers as circles.

<https://doi.org/10.1371/journal.ppat.1009714.g006>

Characterization of DNA degradation patterns specific to aDNA (fragmentation, depurination and deamination) combined with clear evidence of branch shortening confirmed the historical nature of our reconstructed *Xci* genome, a key point in any ancient DNA study [12,41]. Patterns of DNA degradation of HERB_1937_*Xci* appeared consistent with those measured on *P. infestans* from 19th century herbarium samples [23,24,61]. Interestingly, we observed significantly higher deamination rates of cytosine residues in reads mapping to either of the two plasmids, as compared to reads mapping to chromosomal DNA. Depurination rates and fragment sizes did not harbour such significant differences in our study. A possible explanation for our observation relates to differential methylation patterns of cytosines. In a recent study investigating epigenetic modifications in *Xanthomonas* species, N4-methylcytosines (N4meC, a bacteria-specific pattern) were identified in higher proportions in chromosomes *versus* plasmids [63]. Interestingly, N4meC have previously been found to be more resistant to deamination than unmethylated cytosines [64]. The lower deamination rate observed on HERB_1937_*Xci* chromosome (as compared to the plasmids) could thus be due to a better protection of the chromosomal cytosines from deamination, independently of depurination and fragmentation mechanisms. Further investigations, such as ancient methylome mappings [65], should refine our molecular understanding of the degradation patterns observed in this study.

Phylogenetic reconstruction confidently placed HERB_1937_*Xci* at the root of the modern SWIO lineages, a position that reveals its genetic relatedness with the SWIO *Xci* founding population. Modern Mauritian strains were found in all three main SWIO lineages and displayed the highest phylogenetic diversity, a typical pattern for source populations during biological invasions [66], which points Mauritius as the most likely entry point of ACC disease in the SWIO islands. Future studies including new historical genomes from other islands will be necessary to confirm this hypothesis. We estimated the age of the ancestor of all strains (*i.e.* the root), which is a proxy for *Xci* emergence date to 1843 [95% HPD: 1803–1881]. This predates the earliest record of the disease in the area (1917 in Mauritius [39]) and refines the recent estimation of 1818 [95% HPD: 1762–1868] obtained from modern strains only [36]. *Xci* and its main host genus, *Citrus*, originated in Asia [34,35] and were most likely disseminated out of their area of origin by human-mediated movements of plants or plant propagative material [31,67]. Richard *et al.* proposed two possible origins of the pathogen in the SWIO [36]. On the one hand, they hypothesized that a French botanist and colonial administrator, Pierre Poivre (1719–1786) could have introduced infected citrus plants from several Asian countries during his numerous peregrinations starting in mid-18th century [68]. Later, tens of thousands of indentured labourers arrived from several Asian countries (most numerous from India) after the abolition of slavery in Mauritius (1835) and Reunion (1848), mainly to work in agriculture [69]. This flow of people from the Asiatic continent, along with their possessions which consisted among other things of seeds, plants and fruits [70] may have led to the introduction of *Xci* in the SWIO area. The updated time frame of emergence inferred from our data favours

the second scenario. Future work including strains from the hypothetical Asian cradle of *Xci*, with some possibly obtained from herbarium specimens, will be required to investigate the geographic origin of the strains that first invaded SWIO.

Although both the root position of HERB_1937_ *Xci* and the monophyly of all SWIO strains suggest one or few successful historical introductions of genetically (and likely geographically) closely related *Xci* strains in this area, the structure of the phylogeny indicates multiple inter-island migration events, likely via infected plant material exchange. Such events may have first occurred between Mauritius and Reunion islands at the very beginning of the history of *Xci* in the area, as illustrated by the deepness of the most recent common ancestor (MRCA) shared between strains of those two islands that used to share tight historical and political links at the time. More recent migrations between *i*) Mauritius and Rodrigues (an island ca. 600 km east of Mauritius, part of its territory), *ii*) Reunion and the Comoros archipelago (Mayotte, 1,435 km distant from Reunion, part of the French overseas territories) and *iii*) the four islands of the Comoros archipelago (promoted by their historical and economic relationships, see insert Fig 6A). Altogether, our findings emphasise the influence of human-associated migratory events in shaping the global distribution and the emergence of preadapted crop pathogens, a well-known phenomenon [6,24,71,72]. Additionally, our results indicate that integrating historical genomes in phylogenetic analyses significantly refines divergence time estimates, as highlighted in previous ancient DNA studies [73,74].

Tip-date calibration of the SWIO phylogenetic tree also enabled us to estimate a mean mutation rate of 9.4×10^{-8} substitutions per site per year for *Xci*. This value is consistent with the recent estimation of 8.4×10^{-8} substitutions per site per year obtained by Richard *et al.* in the same area [36] and falls within estimations made over a similar time span (80 years) on several human-associated bacterial pathogens, spanning one order of magnitude (10^{-8} – 10^{-7}) [75]. This rate, among the first published for a crop pathogen, is averaged across all sites of the non-recombining portions of the *Xci* chromosome and appears to be homogeneous within the various SWIO lineages. Interestingly, we observed a relatively high dN/dS ratio as compared to other bacterial species [76], which might result from selection for diversification following *Xci* emergence and evolution within SWIO islands. In summary, our substitution rate estimate is crucial to further studies, since it can improve the prediction of the evolution of *Xci* using various modelling-based frameworks.

Finally, we aimed to compare the genomic features of HERB_1937_ *Xci* with its modern counterparts. Among the 15 SNPs unique to HERB_1937_ *Xci* and restricted to chromosome sequences, five non-synonymous SNPs are considered to induce conservative amino acid changes, and are thus not expected to alter the conformation or the active site of their respective proteins. Interestingly, among the six non-conservative SNPs, the location (next to the hinge and binding domain) of an amino acid substitution of the essential metabolic enzyme isocitrate dehydrogenase could modulate its adaptability, and thus the fitness of the pathogen [77]. Finally, seven non-synonymous SNPs account for unique amino acids at given positions of *Xanthomonas* sequence alignments, providing an exclusive signature for seven HERB_1937_ *Xci* specific protein homologs.

Our investigation of HERB_1937_ *Xci* gene content showed that it was globally similar to the one observed in reference strain IAPAR 306. The non-covered CDS corresponded to repeated CDS or to absent CDS. The former are mostly transposases, or other multicopy genes. Among actually absent CDS are mostly proteins of unknown function, recombinases, or notably a type I restriction-modification system, together with four adjacent CDS. The 27-kb block probably corresponds to a genomic island recently acquired by strain IAPAR 306 (and a few of its close relatives) but absent in other *Xci*. It is inserted in the middle of a CDS encoding for a putative competence protein [78]. However, as our gene content analysis

resulted from sequences reconstructed by mapping, we were unable to identify potential genome rearrangements. Furthermore, any genetic material present in HERB_1937_Xci but absent in the reference sequences used to reconstruct the historical genome would have been missed. Gene content investigation based on *de novo* assembly of historical reads would be a way to overcome this limitation but the short length of aDNA reads, their mixed origin as well as the relatively low coverage of HERB_1937_Xci hampered us from applying such strategy [79,80].

We showed that HERB_1937_Xci contains a complete set of genes for its type III secretion system, as well as the same assortment of effector genes as modern *Xci* strains [46]. In particular, Transcription Activator-Like Effectors (TALE) are crucial virulence factors for *Xci* [50]. We determined HERB_1937_Xci to possess between two and four paralogs of the functional *tale* gene *pthA4* present in strain IAPAR 306, a value consistent with modern *Xci* strains [81]. Although it was not possible to localize them in the genome or reconstitute their central repeat domain, sequences corresponding to their N- and C-terminal domains, as well as a repertoire of RVD sequences, were reconstructed, suggesting their functionality. Most of the essential RVD sequences were present in approximately the same proportion as in modern *Xci tale* genes. Three unique sequences encoding unreported RVDs could be mutational variants of the present RVD sequences: AAA—from AAT—(K* from N*), CACGAA and CAGGAT from CACGAT (respectively HE and QD from HD). A design of TALENs with artificial RVDs recently showed that HE and QD were functional and preferentially binding to C on target DNA *in vivo*, like HD [82]. This suggests that apart from undetectable loss of genes (in the case of effectors not identified in databases) and modification in the structure of the repeat region of TALE genes (which might have an important impact on virulence), the effector repertoire of *Xci* has been stable at least since the time of the last common ancestor of all SWIO strains.

In summary, our results show that herbarium specimens can provide a wealth of genomic information on bacterial pathogens, their associated microbial community, or their plant host (an aspect that we did not explore in this study). The present work focused on a single herbarium specimen in order to evaluate the feasibility of genetic analyzes and the added value such samples bring to phylogenetic and epidemiological approaches. Broader studies to reconstruct Asiatic citrus canker's worldwide propagation and evolutionary history would require additional, well-chosen, geographically and temporally representative samples. More generally, similar investigations could be and are performed on other important bacterial plant pathogens to elucidate their evolutionary history, investigate plant-pathogens interactions further, and study the temporal dynamic of plant-associated microbial communities. Such studies emphasise the interest of biological collections and will hopefully help to decipher the epidemiological and evolutionary factors leading to the emergence of plant pathogens. This, in turn, may provide clues to improve disease monitoring and achieve sustainable control.

Material & methods

Herbarium sampling

The collections of the Mauritius Herbarium (<https://herbaria.plants.ox.ac.uk/bol/mau>) were prospected in June 2017. Several citrus specimens displaying typical citrus canker lesions were sampled on site using gloves and sterile equipment and brought back to CIRAD laboratory in individual envelopes where they have been stored in vacuum-sealed boxes at 17°C until use. MAU 0015151 (Fig 1), a *Citrus* sp. specimen collected by Reginald E. Vaughan at Phoenix, Mauritius in 1937 was chosen as being the oldest specimen sampled from the SWIO area. The date and exact place of collection, which do not appear on the specimen itself, were found in the original collection book of the collector. MAU 0015151 was deposited in 1937 at the

collection of the Mauritius Institute (Port Louis, Mauritius). This collection was moved in 1960 to Réduit to form the core of The Mauritius Herbarium (MAU, acronym according Thiers 2021), where it has been preserved since under controlled temperature and humidity and regularly poisoned (e.g. fumigation and/or use of Kew mixture: solution of mercury, phenol, and ethanol).

DNA extraction, quality control and real-time quantitative PCR assay

HERB_1937 sample DNA extraction was performed in a bleach-cleaned facility room with no prior exposure to modern *Xci* DNA. DNA extraction was performed following a custom CTAB protocol modified from Ausubel (2003) [83]. Briefly, a pool of five canker lesions (to obtain approximately 10 mg) from a single leaf of HERB_1937 were cut. A 10 mg piece of a plant species that is not a host to *Xci*, a *Coffea arabica* herbarium 1965 specimen, was integrated as an aDNA negative control sample. Both samples were pulverised at room temperature and soaked in a CTAB extraction solution (1% CTAB, 700 mM NaCl, 0.1 mg/mL Proteinase K, 0.05 mg/mL RNase A, 0.5% N-lauroylsarcosine, 1X Tris-EDTA) under constant agitation and until tissue lysis at 56°C (up to six hours); an equal volume of 24:1 chloroform: isoamyl alcohol was added before centrifugation and recuperation of the aqueous phase (twice), followed by adding 7/3 volume of pure ethanol for an overnight precipitation at -20°C. Dried pellets were resuspended in 10 mM Tris buffer and stored at -20°C until further use. Quality assessment was performed for fragment size and concentration with Qubit (Invitrogen life Technologies) and TapeStation (Agilent Technologies) high sensitivity assays, according to the manufacturers' instructions. To confirm the specific presence of *Xci* in HERB_1937, we performed the *Xci*-exclusive Xac-qPCR diagnostic assay developed by Robène *et al.* on 3 replicates of 5 µL water-diluted (10 fold) DNA extract, our negative control and following the recommended amplification conditions [40].

Library preparation & sequencing

Library preparation and sequencing were outsourced (<https://www.fasteris.com/dna/>). Briefly, DNA was converted into a double-stranded library using a custom TruSeq DNA Nano Illumina protocol omitting the fragmentation step and using a modified bead ratio to keep small fragments. Sequencing of both HERB_1937 and the negative control sample was performed in a paired-end 2×150 cycles configuration on a single lane of the NextSeq flow cell.

Initial read trimming and merging

BBDuk from BBMap 37.92 [84] was first run with an entropy of 0.6 to remove artefactual homopolymer sequences. Illumina adaptors were trimmed out using the Illuminaclip option in Trimmomatic 0.36 [85]. Such roughly trimmed-reads were processed using the post-mortem DNA damage pipeline detailed below. Additional quality-trimming was performed with Trimmomatic 0.36 based on base-quality (LEADING:15; TRAILING:15; SLIDINGWINDOW:5:15) and read length (MINLEN:30). Paired reads were then merged using Adapter-Removal 2.2.2 [86] using default parameters before running both the metagenomic and the phylogenetic pipelines detailed below and in Fig 2.

Negative control sequences analysis

Reads generated from the negative control sample were sequentially mapped to reference sequence genomes of *Coffea arabica* (GCA_003713225.1), *Citrus sinensis* (AJPS00000000.1)

and *Xci* (strain IAPAR 306, chromosome NC_003919.1, plasmids pXAC33 NC_003921.3 and pXAC64 NC_003922.1) using BWA-aln 0.7.15 (default options and seed disabled).

Metagenomic pipeline

The metagenomic composition of historical HERB_1937 sample was assessed following a two-step procedure. First, reads were sequentially mapped to reference sequence genomes of human (GCF_000001405.39), *Citrus sinensis* (AJPS00000000.1) and *Xci* (strain IAPAR 306, chromosome NC_003919.1, plasmids pXAC33 NC_003921.3 and pXAC64 NC_003922.1) using the “very-sensitive” option (seed of -L 20) of Bowtie 2 [87]. In a second step, BLAST analysis was performed on 1,000,000 randomly chosen unmapped reads against the nucleotide database using the blastn command of NCBI BLAST 2.2.31 [88]. Only top hits with an e-value below 0.001 were saved. The proportion of each taxon in the sample was scaled over the total number of reads.

Ancient DNA damage assessment pipeline

Post-mortem DNA damage measured by DNA fragment length distribution, purine frequencies before DNA breakpoints and 5' C to T or 3' G to A misincorporation patterns were assessed with mapDamage2 [89] for both the historical specimen and three modern strains (strains LJ225-01, LK144-08 & LM053-06 isolated in 2012, 2013 and 2015, respectively—see S1 Table). Alignments were generated using BWA-aln 0.7.15 (default options and seed disabled) [90] as short-read aligner for the historical specimen and Bowtie 2 (options—non-deterministic—very-sensitive) [87] for the modern strains using IAPAR 306 *Xci* reference genome (plasmids pXAC33, pXAC44 and chromosome). PCR duplicates were removed using picard-tools 2.7.0 MarkDuplicates [91]. An independent damage assessment was performed using *Methylobacterium* reference sequence (*Methylobacterium organophilum* strain DSM 760 QEKZ01000001.1) with BWA-aln (same options as above). Statistical analyses were performed using GraphPad Prism version 6.00 for macOS, GraphPad Software, San Diego, California USA (www.graphpad.com) [92].

Historical genome reconstruction & characterization

Sequencing depths were computed using BEDTools genomecov 2.24.0 [93], and graphically represented with CIRCOS 0.69.9 [94]. BAM files were extremity-trimmed for 5 bp at each end with BamUtil 1.0.14 [95]. SNPs were called with GATK UnifiedGenotyper [96]. SNPs that met at least one of the following conditions: depth < average depth + 1sd [X = 9], allelic frequency < 0.9, distance from another SNP < 20 bp were considered as dubious and filtered out. Consensus historical sequences were then reconstructed by introducing the remaining high-quality SNPs in the reference genome and replacing both filtered-out variants and non-covered sites (depth = 0) by an N. Non-covered regions were identified with BEDTools 2.24.0 [93].

Gene content analysis

The presence (or absence) of a CDS was assumed when its sequence coverage was found to be above (or below) a 75% threshold. Their repeated nature, as well as their hypothetical functions (as predicted for strain IAPAR 306, chromosome NC_003919.1, plasmids pXAC33 NC_003921.3 and pXAC64 NC_003922.1) were assessed using the annotated reference sequences within the genome browser and synteny tool of the MicroScope platform [97] based on a small set of public strains from the SWIO and the rest of the world (C40, LH201, LB100-

1, JJ10-1, FDC217, LG115, LG97, LB302 [33]), and a few additional representatives of the genus *Xanthomonas* (*X. citri* pv. *bilvae* strain NCPPB 3213, *X. euvesicatoria* 85–10, *X. campestris* pv. *campestris* 8004, *X. perforans* 91–118).

To investigate the presence of virulence factors in HERB_1937_Xci, we used a list of 82 Type III effectors (see list in S3 Table) found in *Xanthomonas* [45,46]. The reference sequences used to assess homology were the IAPAR 306 CDS when available or other *Xanthomonas* CDS for genes not present in *Xci*. We assessed coverage for the 57 effectors found in *Xci* from the reconstructed historical genome. For the 25 effectors from other *Xanthomonas*, reads were realigned on reference sequences with BWA-aln as described above. Coverage data was recovered from BAM files with BAMStats 1.25 tool [98].

In a second step, we aimed to specifically retrieve reads that initially did not map to *tale* genes. We performed a BWA-aln alignment (options as above) on the sequences coding for the conserved N- and C-terminal domains of the *pthA4* CDS from strain IAPAR 306 (S1 Fig). For reads corresponding to the central repeat domain, we constructed a chimera sequence of three repeats (containing Ns at the variable nucleotide positions in RVD) as a reference for the mapping.

Phylogeny pipeline & tree-calibration

An alignment of HERB_1937_Xci and 116 modern genomes (date range: 1978–2015) from the SWIO islands was constructed for phylogenetic analyses, with the modern *Xci* strain LG117 from Bangladesh used as outgroup (CDAX01000000) (S1 Table). Variants from modern strains were independently called and filtered using the same parameters as for HERB_1937_Xci (except for the threshold on depth that was modified to a value of 15). Regions acquired via horizontal gene transfers were identified with ClonalFrameML [51] and removed to account for the effect of recombination on phylogenetic reconstruction and avoid incongruent trees. A Maximum Likelihood tree was constructed using RAxML 8.2.4 [52] using a rapid Bootstrap analysis, a General Time-Reversible model of evolution [99] following a Γ distribution with four rate categories (GTRGAMMA) and 1,000 alternative runs.

The existence of a temporal signal was investigated by two different tests. First, a linear regression test between sample age and root-to-tip distances (computed from the ML tree) was done using the `distRoot` function from the “`adephylo`” R package [100]. Temporal signal was considered present if a significant positive correlation was observed. Secondly, we performed a date-randomization test [101] with 20 independent date-randomized datasets using R package “`TipDatingBeast`” [102]. Temporal signal was considered present when there was no overlap between the inferred root height 95% Highest Posterior Density (95% HPD) of the initial dataset and that of 20 date-randomized datasets. Tip-dating calibration Bayesian inferences were performed with BEAST 1.8.4 [53]. For this, leaf heights were constrained to be proportional to sample ages. Flat priors (*i.e.*, uniform distributions) for the substitution rate (10^{-12} to 10^{-2} substitutions/site/year), as well as for the age of any internal node in the tree, were applied. We also considered a GTR substitution model with a Γ distribution and invariant sites (GTR+G+I), an uncorrelated relaxed log-normal clock to account for variations between lineages, and a tree prior for demography of exponential growth as best-fit parameters described in Richard *et al.* [36]. The Bayesian topology was conjointly estimated with all other parameters during the Markov Chain Monte-Carlo and no prior information from the ML tree was incorporated in BEAST. Three independent chains were run for 25 million steps and sampled every 2,500 steps with a burn-in of 2,500 steps. Convergence to the stationary distribution and sufficient sampling and mixing were checked by inspection of posterior samples (effective sample size >200) in Tracer 1.7.1 [103]. Parameter estimation was based on the samples combined from the different chains. The best-supported tree was estimated from the combined samples by

using the maximum clade credibility method implemented in TreeAnnotator [53]. In order to assess the effect of including our historical sample in the tree calibration, we computed the same inferences on a dataset excluding HERB_1937_Xci. Wilcoxon rank sum test with continuity correction and a Bartlett test of homogeneity of variances were performed on the posterior estimates of the tree root age, to respectively compare the mean and variance of this parameter from both datasets. Finally, phylogenetic diversity (PD) [104], calculated as the sum of branch lengths of the minimum spanning path between strains of the region (island, or group of islands in the case of the Comoros) was calculated on patristic distances from the reconstructed phylogeny using the distRoot function implemented in “adephylo” R package [100]. To account for heterogeneity in region samplings, they were down-sampled to the smallest sampling (Mayotte = 10) and PD by region averaged over 1,000 iterations. PD comparison was done using a Wilcoxon rank sum test with continuity correction.

Supporting information

S1 Fig. Reads depth of a Transcription Activator-Like Effector (TALE) gene of HERB_1937.

(PDF)

S2 Fig. Maximum Likelihood (ML) phylogenetic tree of *Xci* genomes.

(PDF)

S3 Fig. Date-randomization test results.

(PDF)

S4 Fig. Effect of integrating HERB_1937_Xci on substitution rate estimates in BEAST.

(PDF)

S1 Table. Published modern genomes included in the phylogenetic analyzes.

(PDF)

S2 Table. List of *Xci* reference strain IAPAR 306 coding sequences (CDS) covered on less than 75% of their length by HERB_1937_Xci reads and hence designed as non-covered.

(PDF)

S3 Table. List and coverage of 82 *Xanthomonas* virulence factors CDS (pthA4 not included) used in this study.

(PDF)

S4 Table. List and frequency of nucleotide patterns coding for RVD found in HERB_1937_Xci reads.

(PDF)

S5 Table. Description of the 14 SNPs found in coding regions between HERB_1937_Xci and modern strains of the SWIO clade.

(PDF)

Acknowledgments

We are grateful to F. Chiroleu, A. Doizy, A. Duvermy, P. Lefeuvre, F. Balloux, V. Llaurens, R. Debruyne, A. Pérez-Quintero & I. Robène for valuable comments and discussions. Computational work was performed on the CIRAD HPC data center of the South Green bioinformatics platform (<http://www.southgreen.fr/>)

Author Contributions

Conceptualization: Paola E. Campos, Philippe Roumagnac, Boris Szurek, Nathalie Becker, Lionel Gagnevin, Adrien Rieux.

Data curation: Paola E. Campos, Adrien Rieux.

Formal analysis: Paola E. Campos, Nathalie Becker, Lionel Gagnevin, Adrien Rieux.

Funding acquisition: Philippe Roumagnac, Boris Szurek, Nathalie Becker, Lionel Gagnevin, Adrien Rieux.

Investigation: Paola E. Campos, Clara Groot Crego, Karine Boyer, Damien Richard, Olivier Pruvost, Nathalie Becker, Lionel Gagnevin, Adrien Rieux.

Methodology: Paola E. Campos, Nathalie Becker, Lionel Gagnevin, Adrien Rieux.

Resources: Myriam Gaudeul, Claudia Baidier.

Supervision: Nathalie Becker, Lionel Gagnevin, Adrien Rieux.

Visualization: Paola E. Campos, Adrien Rieux.

Writing – original draft: Paola E. Campos, Clara Groot Crego, Nathalie Becker, Lionel Gagnevin, Adrien Rieux.

Writing – review & editing: Paola E. Campos, Clara Groot Crego, Claudia Baidier, Damien Richard, Olivier Pruvost, Philippe Roumagnac, Boris Szurek, Nathalie Becker, Lionel Gagnevin, Adrien Rieux.

References

1. Stukenbrock EH, McDonald BA. The origins of plant pathogens in agro-ecosystems. *Annu Rev Phytopathol.* 2008; 46:75–100. <https://doi.org/10.1146/annurev.phyto.010708.154114> PMID: 18680424
2. Turner RS. After the famine: plant pathology, *Phytophthora infestans*, and the late blight of potatoes, 1845–1960. *Hist Stud Phys Biol.* 2005; 35:341–70. <https://doi.org/10.1525/hsps.2005.35.2.341>
3. Savary S, Willocquet L, Pethybridge SJ, Esker P, McRoberts N, Nelson A. The global burden of pathogens and pests on major food crops. *Nat Ecol Evol.* 2019; 3:430–9. <https://doi.org/10.1038/s41559-018-0793-y> PMID: 30718852
4. FAO, IFAD, UNICEF, WFP & WHO. The state of food security and nutrition in the world—Building resilience for food and food security. Rome: FAO; 2017.
5. Bernades MFF, Pazin M, Pereira LC, Dorta DJ. Impact of pesticides on environmental and human health. *Toxicology studies: cells, drugs and environment.* Rijeka, Croatia; 2015. pp.195–233.
6. Anderson PK, Cunningham AA, Patel NG, Morales FJ, Epstein PR, Daszak P. Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. *Trends Ecol Evol.* 2004; 19:535–44. <https://doi.org/10.1016/j.tree.2004.07.021> PMID: 16701319
7. Pallen MJ, Loman NJ, Penn CW. High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. *Curr Opin Microbiol.* 2010; 13:625–31. <https://doi.org/10.1016/j.mib.2010.08.003> PMID: 20843733
8. Relman DA. Microbial genomics and infectious diseases. *N Engl J Med.* 2011; 365:347–57. <https://doi.org/10.1056/NEJMra1003071> PMID: 21793746
9. Croucher NJ, Didelot X. The application of genomics to tracing bacterial pathogen transmission. *Curr Opin Microbiol.* 2015; 23:62–7. <https://doi.org/10.1016/j.mib.2014.11.004> PMID: 25461574
10. Li LM, Grassly NC, Fraser C. Genomic analysis of emerging pathogens: methods, application and future trends. *Genome Biol.* 2014; 15:1–9. <https://doi.org/10.1186/s13059-014-0541-9> PMID: 25418281
11. Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. Measurably evolving populations. *Trends Ecol Evol.* 2003; 18:481–8. [https://doi.org/10.1016/S0169-5347\(03\)00216-7](https://doi.org/10.1016/S0169-5347(03)00216-7)

12. Pääbo S, Poinar H, Serre D, Jaenicke-Després V, Hebler J, Rohland N, et al. Genetic analyses from ancient DNA. *Annu Rev Genet.* 2004; 38:645–79. <https://doi.org/10.1146/annurev.genet.37.110801.143214> PMID: 15568989
13. Bieker VC, Martin MD. Implications and future prospects for evolutionary analyses of DNA in historical herbarium collections. *Bot Lett.* 2018; 165:409–18. <https://doi.org/10.1080/23818107.2018.1458651>
14. Yoshida K, Burbano HA, Krause J, Thines M, Weigel D, Kamoun S. Mining herbaria for plant pathogen genomes: back to the future. *PLoS Pathog.* 2014; 10:1–6. <https://doi.org/10.1371/journal.ppat.1004028> PMID: 24763501
15. Ristaino JB, Groves CT, Parra GR. PCR amplification of the Irish potato famine pathogen from historic specimens. *Nature.* 2001; 411:695–7. <https://doi.org/10.1038/35079606> PMID: 11395772
16. May KJ, Ristaino JB. Identity of the mtDNA haplotype(s) of *Phytophthora infestans* in historical specimens from the Irish potato famine. *Mycol Res.* 2004; 108:171–9. <https://doi.org/10.1017/s0953756204009876> PMID: 15229999
17. Saville AC, Martin MD, Ristaino JB. Historic late blight outbreaks caused by a widespread dominant lineage of *Phytophthora infestans* (Mont.) de Bary. *PLoS ONE.* 2016; 11:1–22. <https://doi.org/10.1371/journal.pone.0168381> PMID: 28030580
18. Antonovics J, Hood ME, Thrall PH, Abrams JY, Duthie GM. Herbarium studies on the distribution of anther-smut fungus (*Microbotryum violaceum*) and *Silene* species (*Caryophyllaceae*) in the eastern United States. *Am J Bot.* 2003; 90:1522–31. <https://doi.org/10.3732/ajb.90.10.1522> PMID: 21659105
19. Gutaker RM, Reiter E, Furtwängler A, Schuenemann VJ, Burbano HA. Extraction of ultrashort DNA molecules from herbarium specimens. *BioTechniques.* 2017; 62:1–4. <https://doi.org/10.2144/000114517> PMID: 28193151
20. Yoshida K, Sasaki E, Kamoun S. Computational analyses of ancient pathogen DNA from herbarium samples: challenges and prospects. *Front Plant Sci.* 2015; 6:1–6. <https://doi.org/10.3389/fpls.2015.00001> PMID: 25653664
21. Biek R, Pybus OG, Lloyd-Smith JO, Didelot X. Measurably evolving pathogens in the genomic era. *Trends Ecol Evol.* 2015; 30:306–13. <https://doi.org/10.1016/j.tree.2015.03.009> PMID: 25887947
22. Rieux A, Balloux F. Inferences from tip-calibrated phylogenies: a review and a practical guide. *Mol Ecol.* 2016; 25:1911–24. <https://doi.org/10.1111/mec.13586> PMID: 26880113
23. Martin MD, Cappellini E, Samaniego JA, Zepeda ML, Campos PF, Seguin-Orlando A, et al. Reconstructing genome evolution in historic samples of the Irish potato famine pathogen. *Nat Commun.* 2013; 4:1–7. <https://doi.org/10.1038/ncomms3172> PMID: 23863894
24. Yoshida K, Schuenemann VJ, Cano LM, Pais M, Mishra B, Sharma R, et al. The rise and fall of the *Phytophthora infestans* lineage that triggered the Irish potato famine. *eLife.* 2013; 2:1–25. <https://doi.org/10.7554/eLife.00731> PMID: 23741619
25. Ristaino JB. The importance of mycological and plant herbaria in tracking plant killers. *Front Ecol Evol.* 2020; 7:1–11. <https://doi.org/10.3389/fevo.2019.00521>
26. Al Rwahnih M, Rowhani A, Golino D. First report of Grapevine red blotch-associated virus in archival grapevine material from Sonoma County, California. *Plant Dis.* 2015; 99:895. <https://doi.org/10.1094/PDIS-12-14-1252-PDN>
27. Malmstrom CM, Shu R, Linton EW, Newton LA, Cook MA. Barley yellow dwarf viruses (BYDVs) preserved in herbarium specimens illuminate historical disease ecology of invasive and native grasses. *J Ecology.* 2007; 95:1153–66. <https://doi.org/10.1111/j.1365-2745.2007.01307.x>
28. Smith O, Clapham A, Rose P, Liu Y, Wang J, Allaby RG. A complete ancient RNA genome: identification, reconstruction and evolutionary history of archaeological Barley stripe mosaic virus. *Sci Rep.* 2015; 4:1–6. <https://doi.org/10.1038/srep04003> PMID: 24499968
29. Li W, Song Q, Bransky RH, Hartung JS. Genetic diversity of citrus bacterial canker pathogens preserved in herbarium specimens. *PNAS.* 2007; 104:18427–32. <https://doi.org/10.1073/pnas.0705590104> PMID: 17998540
30. Hasse CH. *Pseudomonas citri*, the cause of citrus canker. *J Agric Res.* 1915;97–100.
31. Graham JH, Gottwald TR, Cubero J, Achor DS. *Xanthomonas axonopodis* pv. *citri*: factors affecting successful eradication of citrus canker. *Mol Plant Pathol.* 2004; 5:1–15. <https://doi.org/10.1046/j.1364-3703.2004.00197.x> PMID: 20565577
32. Fawcett HS, Jenkins AE. Records of citrus canker from herbarium specimens of the genus *Citrus* in England and the United States. *Phytopathology.* 1933;820–4.
33. Gordon JL, Lefeuvre P, Escalon A, Barbe V, Cruveiller S, Gagnevin L, et al. Comparative genomics of 43 strains of *Xanthomonas citri* pv. *citri* reveals the evolutionary events giving rise to pathotypes with different host ranges. *BMC Genomics.* 2015; 16:1–20. <https://doi.org/10.1186/1471-2164-16-1> PMID: 25553907

34. Patané JSL, Martins J, Rangel LT, Belasque J, Digiampietri LA, Facincani AP, et al. Origin and diversification of *Xanthomonas citri* subsp. *citri* pathotypes revealed by inclusive phylogenomic, dating, and biogeographic analyses. *BMC Genomics*. 2019; 20:1–23. <https://doi.org/10.1186/s12864-018-5379-1> PMID: 30606130
35. Pruvost O, Magne M, Boyer K, Leduc A, Tourterel C, Drevet C, et al. A MLVA genotyping scheme for global surveillance of the citrus pathogen *Xanthomonas citri* pv. *citri* suggests a worldwide geographical expansion of a single genetic lineage. *PLoS ONE*. 2014; 9:1–11. <https://doi.org/10.1371/journal.pone.0098129> PMID: 24897119
36. Richard D, Pruvost O, Balloux F, Boyer C, Rieux A, Lefeuvre P. Time-calibrated genomic evolution of a monomorphic bacterium during its establishment as an endemic crop pathogen. *Mol Ecol*. 2020; 1–13. <https://doi.org/10.1111/mec.15328> PMID: 31916358
37. Richard D, Ravigné V, Rieux A, Facon B, Boyer C, Boyer K, et al. Adaptation of genetically monomorphic bacteria: evolution of copper resistance through multiple horizontal gene transfers of complex and versatile mobile genetic elements. *Mol Ecol*. 2017; 26:2131–49. <https://doi.org/10.1111/mec.14007> PMID: 28101896
38. Pruvost O, Boyer K, Ravigné V, Richard D, Vernière C. Deciphering how plant pathogenic bacteria disperse and meet: Molecular epidemiology of *Xanthomonas citri* pv. *citri* at microgeographic scales in a tropical area of Asiatic citrus canker endemicity. *Evol Appl*. 2019; 12:1523–38. <https://doi.org/10.1111/eva.12788> PMID: 31462912
39. Aubert B. Vergers de la Réunion et de l'Océan Indien. CIRAD. Hommes et fruits en pays du Sud. CIRAD. 2014. pp.111–65. French
40. Robène I, Maillot-Lebon V, Chabirand A, Moreau A, Becker N, Moumène A, et al. Development and comparative validation of genomic-driven PCR-based assays to detect *Xanthomonas citri* pv. *citri* in citrus plants. *BMC Microbiol*. 2020; 20:1–13. <https://doi.org/10.1186/s12866-019-1672-7> PMID: 31896348
41. da Silva ACR, Ferro JA, Reinach FC, Farah CS, Furlan LR, Quaggio RB, et al. Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. *Nature*. 2002; 417:459–63. <https://doi.org/10.1038/417459a> PMID: 12024217
42. Dabney J, Meyer M, Pääbo S. Ancient DNA damage. *Cold Spring Harb Perspect Biol*. 2013; 7:1–8. <https://doi.org/10.1101/cshperspect.a012567> PMID: 23729639
43. Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*. 2013; 29:1682–4. <https://doi.org/10.1093/bioinformatics/btt193> PMID: 23613487
44. Büttner D. Behind the lines—actions of bacterial type III effector proteins in plant cells. *FEMS Microbiol Rev*. 2016; 40:894–937. <https://doi.org/10.1093/femsre/fuw026> PMID: 28201715
45. The *Xanthomonas* Resource (<http://www.xanthomonas.org/t3e.html>). 2018. Available from <http://www.xanthomonas.org/t3e.html> (accessed in May 2020)
46. Escalon A, Javegny S, Vernière C, Noël LD, Vital K, Poussier S, et al. Variations in type III effector repertoires, pathological phenotypes and host range of *Xanthomonas citri* pv. *citri* pathotypes: type III effectors in *Xanthomonas citri* pv. *citri*. *Mol Plant Pathol*. 2013; 14:483–96. <https://doi.org/10.1111/mpp.12019> PMID: 23437976
47. Swarup S, De Feyter R, Brlansky RH, Gabriel DW. A pathogenicity locus from *Xanthomonas citri* enables strains from several pathovars of *X. campestris* to elicit canker-like lesions on citrus. *Phytopathology*. 1991; 81:802–9.
48. Duan YP, Castañeda A, Zhao G, Erdos G, Gabriel DW. Expression of a single, host-specific, bacterial pathogenicity gene in plant cells elicits division, enlargement, and cell death. *Mol Plant Microbe Interact*. 1999; 12:556–60. <https://doi.org/10.1094/MPMI.1999.12.6.556>
49. Hu Y, Zhang J, Jia H, Sosso D, Li T, Frommer WB, et al. *Lateral organ boundaries 1* is a disease susceptibility gene for citrus bacterial canker disease. *PNAS*. 2014; 111:521–9. <https://doi.org/10.1073/pnas.1318582111> PMID: 24367083
50. Perez-Quintero AL, Szurek B. A decade decoded: spies and hackers in the history of TAL effectors research. *Annu Rev Phytopathol*. 2019; 57:459–81. <https://doi.org/10.1146/annurev-phyto-082718-100026> PMID: 31387457
51. Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol*. 2015; 11:1–18. <https://doi.org/10.1371/journal.pcbi.1004041> PMID: 25675341
52. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014; 30:1312–3. <https://doi.org/10.1093/bioinformatics/btu033> PMID: 24451623

53. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol.* 2007; 7:214. <https://doi.org/10.1186/1471-2148-7-214> PMID: 17996036
54. Peyambari M, Warner S, Stoler N, Rainer D, Roossinck MJ. A 1,000-year-old RNA virus. *J Virol.* 2018; 93:1–30. <https://doi.org/10.1128/JVI.01188-18> PMID: 30305356
55. Dvořák P, Hašler P, Pouličková A. New insights into the genomic evolution of cyanobacteria using herbarium exsiccatae. *Eur J Phycol.* 2020; 55:30–8. <https://doi.org/10.1080/09670262.2019.1638523>
56. Blaustein RA, Lorca GL, Meyer JL, Gonzalez CF, Teplitski M. Defining the core citrus leaf- and root-associated microbiota: factors associated with community structure and implications for managing huanglongbing (citrus greening) disease. *Appl Environ Microbiol.* 2017; 83:1–14. <https://doi.org/10.1128/AEM.00210-17> PMID: 28341678
57. Xu J, Zhang Y, Zhang P, Trivedi P, Riera N, Wang Y, et al. The structure and function of the global citrus rhizosphere microbiome. *Nat Commun.* 2018; 9:1–10. <https://doi.org/10.1038/s41467-017-02088-w> PMID: 29317637
58. Zhang Y, Trivedi P, Xu J, Roper MC, Wang N. The citrus microbiome: from structure and function to microbiome engineering and beyond. *Phytobiomes J.* 2021; 1–40. <https://doi.org/10.1094/PBIOMES-11-20-0084-RVW>
59. Bieker VC, Sánchez Barreiro F, Rasmussen JA, Brunier M, Wales N, Martin MD. Metagenomic analysis of historical herbarium specimens reveals a postmortem microbial community. *Mol Ecol Resour.* 2020; 1–14. <https://doi.org/10.1111/1755-0998.13125> PMID: 31823489
60. Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathog.* 2016; 8:1–12. <https://doi.org/10.1186/s13099-015-0083-z> PMID: 26759607
61. Weiß CL, Gansauge M-T, Aximu-Petri A, Meyer M, Burbano HA. Mining ancient microbiomes using selective enrichment of damaged DNA molecules. *BMC Genomics.* 2020; 21:1–9. <https://doi.org/10.1186/s12864-020-06820-7> PMID: 32586278
62. Key FM, Posth C, Krause J, Herbig A, Bos KI. Mining metagenomic datasets for ancient DNA: recommended protocols for authentication. *Trends in Genetics.* 2017; 33:508–20. <https://doi.org/10.1016/j.tig.2017.05.005> PMID: 28688671
63. Seong HJ, Park H-J, Hong E, Lee SC, Sul WJ, Han S-W. Methylome analysis of two *Xanthomonas* spp. using single-molecule real-time sequencing. *Plant Pathol J.* 2016; 32:500–7. <https://doi.org/10.5423/PPJ.FT.10.2016.0216> PMID: 27904456
64. Ehrlich M, Norris KF, Wang RY, Kuo KC, Gehrke CW. DNA cytosine methylation and heat-induced deamination. *Biosci Rep.* 1986; 6:387–93. <https://doi.org/10.1007/BF01116426> PMID: 3527293
65. Hanghøj K, Renaud G, Albrechtsen A, Orlando L. DamMet: ancient methylome mapping accounting for errors, true variants, and post-mortem DNA damage. *GigaScience.* 2019; 8:1–6. <https://doi.org/10.1093/gigascience/giz025> PMID: 31004132
66. Estoup A, Guillemaud T. Reconstructing routes of invasion using genetic data: why, how and so what? *Mol Ecol.* 2010; 19:4113–30. <https://doi.org/10.1111/j.1365-294X.2010.04773.x> PMID: 20723048
67. Gottwald TR, Graham JH, Schubert TS. Citrus canker: the pathogen and its impact. *Plant Health Prog.* 2002; 1–34. <https://doi.org/10.1094/PHP-2002-0812-01-RV>
68. Du Pont de Nemours PS. Oeuvres complètes de P. Poivre, intendant des isles de France et de Bourbon, correspondant de l'académie des sciences, etc. 1797. French
69. Carter M, Torabully K. Coolitude, an anthology of the Indian labour diaspora. Anthem Press; 2002.
70. Beaujard P. The first migrants to Madagascar and their introduction of plants: linguistic and ethnological evidence. *Azania: Archaeological Research in Africa.* 2011; 46:169–89. <https://doi.org/10.1080/0067270X.2011.580142>
71. McDonald BA, Linde C. Pathogen population genetics, evolutionary potential, and durable resistance. *Annu Rev Phytopathol.* 2002; 40:349–79. <https://doi.org/10.1146/annurev.phyto.40.120501.101443> PMID: 12147764
72. Goss EM. Genome-enabled analysis of plant-pathogen migration. *Annu Rev Phytopathol.* 2015; 53:121–35. <https://doi.org/10.1146/annurev-phyto-080614-115936> PMID: 25938274
73. Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I, et al. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature.* 2014; 514:494–7. <https://doi.org/10.1038/nature13591> PMID: 25141181
74. Duggan AT, Perdomo MF, Piombino-Mascalci D, Marciniak S, Poinar D, Emery MV, et al. 17th Century variola virus reveals the recent history of smallpox. *Curr Biol.* 2016; 26:3407–12. <https://doi.org/10.1016/j.cub.2016.10.061> PMID: 27939314

75. Duchêne S, Holt KE, Weill F-X, Le Hello S, Hawkey J, Edwards DJ, et al. Genome-scale rates of evolutionary change in bacteria. *Microb Genom*. 2016; 2:1–12. <https://doi.org/10.1099/mgen.0.000094> PMID: 28348834
76. Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH, et al. Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J Theor Biol*. 2006; 239:226–35. <https://doi.org/10.1016/j.jtbi.2005.08.037> PMID: 16239014
77. Spaans SK, Weusthuis RA, van der Oost J, Kengen SWM. NADPH-generating systems in bacteria and archaea. *Front Microbiol*. 2015; 6:1–27. <https://doi.org/10.3389/fmicb.2015.00001> PMID: 25653648
78. Gwinn ML, Ramanathan R, Smith HO, Tomb J-F. A new transformation-deficient mutant of *Haemophilus influenzae* Rd with normal DNA uptake. *J Bacteriol*. 1998; 180:746–8. <https://doi.org/10.1128/JB.180.3.746-748.1998> PMID: 9457884
79. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. *Nat Methods*. 2011; 8:61–5. <https://doi.org/10.1038/nmeth.1527> PMID: 21102452
80. Nagarajan N, Pop M. Sequence assembly demystified. *Nat Rev Genet*. 2013; 14:157–67. <https://doi.org/10.1038/nrg3367> PMID: 23358380
81. Lee S, Lee J, Lee DH, Lee Y-H. Diversity of *pthA* gene of *Xanthomonas* strains causing citrus bacterial canker and its relationship with virulence. *Plant Pathol J*. 2008; 24:357–60. <https://doi.org/10.5423/PPJ.2008.24.3.357>
82. Juillerat A, Pessereau C, Dubois G, Guyot V, Maréchal A, Valton J, et al. Optimized tuning of TALEN specificity using non-conventional RVDs. *Sci Rep*. 2015; 5:1–7. <https://doi.org/10.1038/srep08150> PMID: 25632877
83. Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, et al. *Current protocols in molecular biology*. John Wiley & Sons, New York; 2003.
84. Joint Genome Institute. BBTools. 1997. Available from <https://jgi.doe.gov/data-and-tools/bbtools/bbtools-user-guide/bbduk-guide/> (accessed in May 2020)
85. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014; 30:2114–20. <https://doi.org/10.1093/bioinformatics/btu170> PMID: 24695404
86. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res Notes*. 2016; 9:1–7. <https://doi.org/10.1186/s13104-015-1837-x> PMID: 26725043
87. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012; 9:357–9. <https://doi.org/10.1038/nmeth.1923> PMID: 22388286
88. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990; 215:403–10. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2) PMID: 2231712
89. Ginolhac A, Rasmussen M, Gilbert MTP, Willerslev E, Orlando L. mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics*. 2011; 27:2153–5. <https://doi.org/10.1093/bioinformatics/btr347> PMID: 21659319
90. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009; 25:1754–60. <https://doi.org/10.1093/bioinformatics/btp324> PMID: 19451168
91. Broad Institute. Picard Tools (<http://broadinstitute.github.io/picard/>).
92. GraphPad Prism. San Diego, California, USA: GraphPad Software; Available from www.graphpad.com (accessed in May 2020)
93. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010; 26:841–2. <https://doi.org/10.1093/bioinformatics/btq033> PMID: 20110278
94. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. *Genome Res*. 2009; 19:1639–45. <https://doi.org/10.1101/gr.092759.109> PMID: 19541911
95. Jun G, Wing MK, Abecasis GR, Kang HM. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res*. 2015; 25:918–25. <https://doi.org/10.1101/gr.176552.114> PMID: 25883319
96. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011; 43:491–8. <https://doi.org/10.1038/ng.806> PMID: 21478889
97. Vallenet D, Engelen S, Mornico D, Cruveiller S, Fleury L, Lajus A, et al. MicroScope: a platform for microbial genome annotation and comparative genomics. *Database*. 2009;1–12. <https://doi.org/10.1093/database/bap021> PMID: 20157493

98. SourceForge. BAMStats. 2011. Available from <http://bamstats.sourceforge.net> (accessed in May 2020)
99. Tavaré S, Miura RM. Some probabilistic and statistical problems in the analysis of DNA sequences. Some mathematical questions in biology: DNA sequence analysis. Providence; 1986. pp.57–86.
100. Jombart T, Balloux F, Dray S. *adephylo*: new tools for investigating the phylogenetic signal in biological traits. *Bioinformatics*. 2010; 26:1907–9. <https://doi.org/10.1093/bioinformatics/btq292> PMID: [20525823](https://pubmed.ncbi.nlm.nih.gov/20525823/)
101. Duchêne S, Duchêne D, Holmes EC, Ho SYW. The performance of the date-randomization test in phylogenetic analyses of time-structured virus data. *Mol Biol Evol*. 2015; 32:1895–906. <https://doi.org/10.1093/molbev/msv056> PMID: [25771196](https://pubmed.ncbi.nlm.nih.gov/25771196/)
102. Rieux A, Khatchikian CE. TIPDATINGBEAST: an R package to assist the implementation of phylogenetic tip-dating tests using BEAST. *Mol Ecol Resour*. 2017; 17:608–13. <https://doi.org/10.1111/1755-0998.12603> PMID: [27717245](https://pubmed.ncbi.nlm.nih.gov/27717245/)
103. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst Biol*. 2018; 67:901–4. <https://doi.org/10.1093/sysbio/syy032> PMID: [29718447](https://pubmed.ncbi.nlm.nih.gov/29718447/)
104. Faith DP. Conservation evaluation and phylogenetic diversity. *Biological Conservation*. 1992; 61:1–10. [https://doi.org/10.1016/0006-3207\(92\)91201-3](https://doi.org/10.1016/0006-3207(92)91201-3)

Chapitre 4 – Reconstruction de l’histoire évolutive d’un pathogène à échelle mondiale

Les travaux présentés dans ce chapitre ont été réalisés dans le but d’améliorer la reconstruction de l’histoire évolutive globale de *Xanthomonas citri* pv. *citri* (*Xci*).

Xci présente une répartition mondiale: on la retrouve dans la plupart des pays producteurs d’agrumes des zones tropicales et subtropicales (voir Figure 1.4.1.C en chapitre 1.4.1). Le pathogène aurait cependant, comme son hôte, une origine asiatique supposée par les données historiques ainsi que la distribution des patrons de diversité génétique (voir détails en chapitre 1.4.2 et 1.5). Dans une étude publiée durant ma thèse, Patané et al. (2019) ont analysé les génomes de 73 souches de *Xci* et 22 souches provenant d’autres espèces ou pathovars de *Xanthomonas*, isolées entre 1950 et 2015. Les auteurs ont cherché à reconstruire les mécanismes sous-jacents à l’origine et à la diversification de *Xci* en réalisant deux grands types d’inférences : 1) une datation moléculaire de l’arbre phylogénétique par une approche combinée de *rate-* et de *node-dating*, l’analyse de génomes contemporains uniquement n’apportant pas le signal temporel requis pour appliquer une approche de *tip-dating* ; et 2) une reconstruction des états ancestraux pour le caractère « plante hôte » et « zone géographique ». Leurs résultats suggèrent une origine de *Xci* au sein du sous-continent indien (Inde, Bangladesh, Pakistan), probablement sur une plante de la famille des *Fabaceae* à partir de laquelle le pathogène aurait ensuite émergé (par un saut d’hôtes) sur *Rutaceae*. Les auteurs proposent pour la première fois une datation de certains nœuds internes de l’arbre phylogénétique de *Xci* associés à son origine (entre 14 000 et 6 000 ans avant le présent) et à sa diversification en trois pathotypes (entre 5 600 et 1 700 ans avant le présent).

Les calibrations d’arbres phylogénétiques effectuées par *rate-* et/ou *node-dating* étant associées à différents biais méthodologiques pouvant mener à des estimations imprécises, voire parfois erronées (Rieux *et al.*, 2014), nous avons cherché à évaluer si l’ajout des 13 génomes historiques de *Xci* générés dans le cadre de cette thèse (voir chapitre 2) pouvaient apporter un signal temporel permettant de tester les datations réalisées par Patané et al. (2019) par une analyse indépendante en *tip-dating*, supposée plus robuste. Pour cela, nous avons séquencé 57 nouveaux génomes modernes de *Xci* à partir de souches lyophilisées et construit un jeu de données (contenant 172 génomes modernes au total) améliorant la représentation de la diversité génétique mondiale du pathogène. L’analyse combinée des génomes modernes et historiques nous a permis de préciser les relations de parentés entre lignées évolutives de *Xci* ainsi que d’émettre de nouveaux *scenarii* associées à l’histoire évolutive passée du pathogène en précisant les dates et les aires géographiques associées à sa diversification et à son origine. Pour finir, nous avons caractérisé le contenu en gènes

associés à la pathogénie de *Xci* et investigué sa variation au sein et entre différents pathotypes. De façon intéressante, nous n'avons pas observé de variations majeures de ce même contenu en gènes au cours des deux derniers siècles chez le pathotype A.

Ces résultats sont présentés sous forme d'un article intitulé « ***Improved reconstruction of the crop pathogenic bacterium Xanthomonas citri pv. citri diversification history using historical herbarium genomes*** ». Ce dernier sera prochainement soumis à une des revues suivantes : Current Opinion in Microbiology, eLIFE, PLOS Pathogens, Molecular Ecology, Scientific reports, BMC genomics.

1 Improved reconstruction of the crop pathogenic
2 bacterium *Xanthomonas citri* pv. *citri* diversification
3 history using historical herbarium genomes
4

5 Campos PE^{1,2}, Pruvost O¹, Boyer K¹, Gaudeul M^{2,3}, Baider C⁴, Utteridge T⁵,
6 Shannon D⁶, Toner M⁷, Becker N², Rieux A^{1+*} & Gagnevin L^{8,9+*}

7
8 ¹ CIRAD, UMR PVBMT, F-97410 St Pierre, La Réunion, France

9 ² Institut de Systématique, Evolution, Biodiversité (ISyEB), Muséum national d'Histoire naturelle,
10 CNRS, Sorbonne Université, EPHE, Université des Antilles. 57 rue Cuvier, CP 50, 75005 Paris, France

11 ³ Herbar national (P), Muséum national d'Histoire naturelle, CP39, 57 rue Cuvier, 75005 Paris, France

12 ⁴ Ministry of Agro Industry and Food Security, Mauritius Herbarium, R.E. Vaughan Building (MSIRI
13 compound), Agricultural Services, Réduit, Mauritius.

14 ⁵ Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3AB, UK

15 ⁶ USDA, Mycology and Nematology Genetic Diversity and Biology Laboratory, BELTSVILLE USA

16 ⁷ US National Herbarium, Smithsonian Institution NMNH, Washington, USA

17 ⁸ PHIM Plant Health Institute, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France

18 ⁹ CIRAD, UMR PHIM, Montpellier, France

19 ⁺ These authors contributed equally to this work.

20

21 *Corresponding authors: lionel.gagnevin@cirad.fr and adrien.rioux@cirad.fr

22

23

24

25

26

27

28

29 Manuscript soon to be submitted to one of the following journals: Current Opinion in Microbiology,
30 eLIFE, PLOS Pathogens, Molecular Ecology, Scientific reports, BMC genomics...

31 Abstract

32 Over the past decade, the field of ancient genomics has triggered considerable progress in the study
33 of various pathogens, including those affecting crops. In this context, herbarium collections have been
34 shown to be an enormous source of dated, identified and preserved DNA material that can be used in
35 comparative genomic and phylogeographic studies to shed light into the emergence and evolutionary
36 history of plant pathogens. In this study, we reconstructed 13 historical genomes of the bacterial crop
37 pathogen *Xanthomonas citri* pv. *citri* (*Xci*) from infected citrus herbarium specimens using a shotgun-
38 based deep sequencing strategy. Following authentication of the historical genomes based on ancient
39 DNA damage patterns assessment, we compared them to a large set of modern genomes to
40 reconstruct their phylogenetic relationships, pathogeny-associated genes content and estimate
41 several evolutionary parameters, using Bayesian tip-dating calibration and phylogeography
42 inferences. Our results reveal that *Xci* originated in southern Asia ~11,500 years ago and diversified
43 during the beginning of the 13th century before spreading to the rest of the world, an updated
44 scenario associated with both global climatic warming event and human trade westward along the
45 Silk road. All together, our study emphasises the great potential hidden behind herbarium collections
46 to bring light on the evolutionary dynamics that drive pathogens invasion, ultimately helping us to
47 better control current and future crop epidemics.

48

49 Keywords:

50 Crop pathogen, evolution, ancient DNA, phylogenomic inferences, molecular dating

51 Introduction

52 Plant pathogens have plagued societies since the beginning of agriculture, and with it, the
53 development of plant domestication in agroecosystems (Stukenbrock and McDonald, 2008). In such
54 human-engineered ecosystem, the high density and low genetic diversity of hosts as well as the
55 environmental homogeneity caused by agricultural practices facilitated the rise of diseases and the
56 evolution of host-adapted, more virulent pathogens and their propagation (Mira, Pushker and
57 Rodríguez-Valera, 2006; Stukenbrock and McDonald, 2008). The intensification of agriculture, the rise
58 of monocultures size and the globalisation of trade are all contributing factors to the expansion and
59 emergence of pathogens and the augmentation of opportunities for pathogens to meet new naive
60 host populations on which host shift or host jump can be realised (Anderson *et al.*, 2004; Stukenbrock
61 and McDonald, 2008).

62 Today, plant pathogens and pests cause up to 40% yield loss in major crops, threatening food security
63 (Savary *et al.*, 2019), conservation and public health (Anderson *et al.*, 2004; Bernades *et al.*, 2015).
64 Understanding the factors underlying the origin, evolution and emergence of pathogens would help
65 assess the risks they pose to cultures and improve tools for surveillance and disease control. The
66 combination of genetic material obtained from historic biological collections such as herbaria and
67 modern samples provides heterochronous datasets which can improve phylogenetic estimates of
68 evolutionary parameters and the timelines of their emergence and spread by bringing robust time
69 component to inferences (Rieux *et al.*, 2014; Duchêne *et al.*, 2020). Indeed, adding ancient or historical
70 sequences expands the temporal range of the dataset, increasing the chance to detect evolutionary
71 change, *i.e.* temporal signal, which can be used to infer substitution rates and divergence time
72 between lineage, as well as sudden modifications in genetic diversity (Drummond *et al.*, 2002, 2003;
73 Rieux and Balloux, 2016).

74 The most well-studied crop pathosystem using historical herbarium genetic material is *Phytophthora*
75 *infestans*, the oomycete responsible for potato late blight (Martin *et al.*, 2013; Yoshida *et al.*, 2013,
76 2014; Yoshida, Sasaki and Kamoun, 2015; Saville, Martin and Ristaino, 2016; Ristaino, 2020). Through
77 the sequencing of 19th century infected specimens, the strain which caused the great potato famine
78 in 1845-1849 has been identified and its genome characterised (Martin *et al.*, 2013; Yoshida *et al.*,
79 2013; Yoshida, Sasaki and Kamoun, 2015). Phylogeny reconstruction showed the historical strain to
80 have originated from a secondary diversification area of the pathogen in America from where one or
81 a few dispersal events caused *P. infestans* emergence in Europe (Yoshida *et al.*, 2013; Saville, Martin
82 and Ristaino, 2016). Similar studies reconstructing the evolutionary history of crop pathogens from
83 full genomes have been successfully realised on viruses as well (Malmstrom *et al.*, 2007; Smith *et al.*,

84 2014; Al Rwahnih, Rowhani and Golino, 2015), but for bacterial crop pathogens, only the history of a
85 local emergence of *Xanthomonas citri* pv. *citri* (*Xci*) has been described using ancient samples (Campos
86 *et al.*, 2021).

87 *Xci* is a serious threat to citriculture, the first fruit production worldwide (FAO 2019 data), and no
88 definitive control measure is available. Control strategies involve eradication, drastic quarantine
89 measures and epidemicsurveillance, yet little is known about the epidemiological and evolutionary
90 history of the pathogen. In particular, determinants of its host range and pathogenicity are the object
91 of many functional and evolutionary studies. The pathosystem benefits from both an extensive
92 sampling covering the last 70 years for *Xci* and a good representation of *Citrus* specimens in herbaria.
93 *Xci* is responsible for Asiatic citrus canker (ACC), found in most subtropical citrus-producing regions.
94 The disease causes important economic losses, both by decreasing fruit yield and quality and because
95 of *Xci* quarantine organism status, which impedes exportations and entails costly disease management
96 and surveillance (Gottwald, Graham and Schubert, 2002; Talon, Caruso and Gmitter Jr., 2020). *Xci*
97 comprises three major lineages corresponding to groups defined on genetic diversity and host range
98 called pathotypes, with pathotype A as the most prevalent worldwide and with the broadest host-
99 range (nearly all *Citrus*) (Graham *et al.*, 2004), pathotype A* and pathotype A^W are found in parts of
100 Asia and are restricted to *Citrus aurantiifolia* and *Citrus macrophylla* (Vernière *et al.*, 1998; Schubert
101 *et al.*, 2001). However, pathotype A* is also found on *Citrus latifolia*, while pathotype A^W elicits an
102 hypersensitive response on *Citrus paradisi* (Sun *et al.*, 2004). The phylogenetic relationships between
103 the different pathotypes have been reconstructed, first with molecular markers such as minisatellites
104 (Pruvost *et al.*, 2014), then with more resolutive data from whole genomes (Gordon *et al.*, 2015; Zhang
105 *et al.*, 2015; Patané *et al.*, 2019), suggesting that pathotypes A and A^W are more closely related to each
106 other than they are to pathotype A*. Recently, a comparative genomics and evolutionary study of a
107 collection of contemporary strains allowed identifying pathotype-specific pathogeny-associated genes
108 and inferring a host of origin but also the probable date and location of origin and diversification of
109 the pathogen. It was demonstrated that *Xci* has a relatively recent origin and diversification compared
110 to its host as well as major climatic events such as the Last Glacial Maximum (Patané *et al.*, 2019).
111 However, the sole use of modern genomes impeded the detection of sufficient *de novo* evolutionary
112 change within the dataset (as referring to “measurably evolving populations” (Drummond *et al.*, 2003;
113 Biek *et al.*, 2015) and constrained Patané *et al.* (2019) to build a timeframe of evolution based on both
114 the extrapolation of rates from external measures (*i.e.* rate dating) and a constraint on the distribution
115 of a single external node age (*i.e.* node dating), two dating approaches known to yield potential
116 misleading estimates (Ho *et al.*, 2008; Rieux *et al.*, 2014).

117 In the present study, we sequenced historical bacterial genomes of *Xci* from 13 herbarium samples
118 showing typical canker symptoms originating from the putative center of origin of the pathogen.
119 Following authentication based on DNA degradation patterns analyses, we compared them to the
120 ones of 172 modern strains representative of the worldwide genetic diversity, 57 of which specifically
121 sequenced for the purpose of this study. We thus aimed to improve knowledge about *Xci* origin and
122 diversification history by 1) reconstructing a thorough time-calibrated phylogeny and inferring
123 evolutionary parameters based on a robust dating approach within the measurably evolving
124 populations framework, 2) inferring the ancestral geographical state of lineages and estimating source
125 populations of epidemics, 3) assessing the pathogenicity-associated genes content across all lineages.

126 Material & Methods

127 Herbarium material sampling

128 The collections of “Kew Royal Botanic Gardens,” ([https://www.kew.org/science/collections-and-](https://www.kew.org/science/collections-and-resources/collections/herbarium)
129 [resources/collections/herbarium](https://www.kew.org/science/collections-and-resources/collections/herbarium)), “Mauritius Herbarium”
130 (<https://herbaria.plants.ox.ac.uk/bol/mau>), “Muséum national d’Histoire naturelle”
131 (<https://www.mnhn.fr/fr/collections/ensembles-collections/botanique>), “U.S. National Fungus”
132 Collection Herbarium (<https://nt.ars-grin.gov/fungalatabases/specimens/specimens.cfm>) and “U.S.
133 National Herbarium” (<https://collections.nmnh.si.edu/search/botany/>) were prospected between
134 May 2016 and October 2017. Citrus specimens displaying typical asiatic citrus canker lesions were
135 sampled on site using sterile equipment and transported back to CIRAD laboratory inside individual
136 envelopes where they were stored at 17°C in vacuum-sealed boxes until use. Thirteen historic
137 specimens sampled between 1845 to 1974 and originating from five different herbaria were selected
138 for analysis (Table 1). Those were chosen as the oldest available from Asia, the supposed geographic
139 origin of *Xci*, as well as from Oceania and the Southwest Indian Ocean. Amongst the 13 herbarium
140 specimens, one (HERB_1937) was taken from a previously published work (Campos *et al.*, 2021) while
141 the 12 remaining ones were generated during the course of this study.

142 Ancient DNA extraction and purification

143 Herbarium samples DNA extraction was performed in a bleach-cleaned facility room with no previous
144 exposure to modern *Xci* DNA. Eight of them were achieved as described in (Campos *et al.*, 2021).
145 Briefly, pools of five canker lesions (10 to 30 mg) from a single leaf of each herbarium specimen were
146 cut, along with herbarium samples of *Coffea* sp., a non *Xci*-host species, dating back to 1982 and 1996
147 serving as aDNA negative control. Samples were pulverised at room temperature and soaked in CTAB
148 extraction solution (1% CTAB, 700 mM NaCl, 0.1 mg/mL Proteinase K, 0.05 mg/mL RNase A, 0.5% N-

149 lauroylsarcosine, 1X Tris-EDTA) under constant agitation and until tissue lysis at 56°C (four to six
150 hours); an equal volume of 24:1 chloroform/isoamyl alcohol was added before centrifugation and
151 retrieval of the aqueous phase (twice), followed by adding 7/3 volume of pure ethanol for an overnight
152 precipitation at -20°C. Dried pellets were resuspended in 10 mM Tris buffer and stored at -20°C until
153 further use. Two of these extractions were purified with SPRI (Solid Phase Reversible Immobilisation)
154 magnetic beads (10 mM Tris-HCl, 1.0 M NaCl, 18% PEG8000) according to (Carøe *et al.*, 2018). For the
155 five remaining herbarium samples, the ethanol precipitation was replaced by DNA binding and
156 purification with SPRI magnetic beads (10 mM Tris-HCl, 1.11 M NaCl, 20% PEG8000) and eluted in
157 elution buffer (10 mM Tris-HCl, 0.05% Tween-20). Quality assessment was realised for fragment size
158 and concentration with TapeStation (Agilent Technologies) high sensitivity assays according to the
159 manufacturers' recommendations.

160 Modern bacterial strains culture and DNA extraction

161 Fifty-seven bacterial strains isolated between 1963 and 2008, mainly from Asia (S1 Table) and stored
162 as lyophilisates at -80°C, were chosen to complete the collection of available modern genomes. Strains
163 were inoculated in YP (Yeast Peptone) broth tubes (7 g/L yeast extract, 7 g/L peptone, pH 7.2) and
164 then grown at 28°C on LPGA (Levure Peptone Glucose Agar) plates (7 g/L yeast extract, 7 g/L peptone,
165 7 g/L glucose, 18 g/L agar, supplemented by 20 mg/L propiconazole, pH 7.2). Single cultures were used
166 for DNA extraction using the Wizard® genomic DNA purification kit (Promega) following the
167 manufacturers' instructions. Quality assessment was realised for concentration with QuBit (Invitrogen
168 life Technologies) broad range assays and Nanodrop (Thermo Fisher Scientific) under the
169 manufacturers' instructions.

170 Library preparation & sequencing

171 Library preparation of six herbarium samples were outsourced to Fasteris
172 (<https://www.fasteris.com/dna/>). Briefly, DNA was converted into a double-stranded library using a
173 custom TruSeq DNA Nano Illumina protocol omitting the fragmentation step and using a modified
174 bead ratio to keep small fragments. The remaining seven herbarium samples were converted into
175 double-stranded libraries in a bleach-cleaned facility room using the aDNA-adapted BEST (Blunt-End-
176 Single-Tube) protocol from (Carøe *et al.*, 2018). Library preparation of the modern strains were
177 outsourced to Fasteris where classic TruSeq DNA Nano Illumina protocol following Nextera enzymatic
178 DNA fragmentation was applied. Sequencing was performed in a paired-end 2×150 cycles
179 configuration on a NextSeq machine in several batches, with historical and modern libraries being
180 independently treated.

181 Initial reads trimming and merging

182 Artefactual homopolymer sequences were removed from libraries when presenting entropy inferior
183 than 0.6 using BBDuk from BBMap 37.92 (Joint Genome Institute, 1997). Adaptors were trimmed using
184 the Illuminaclip option from Trimmomatic 0.36 (Bolger, Lohse and Usadel, 2014). Such reads were
185 processed into the post-mortem DNA damage assessment pipeline detailed in the section below.
186 Additional quality-trimming was realised with Trimmomatic based on base-quality (LEADING:15;
187 TRAILING:15; SLIDINGWINDOW:5:15) and read length (MINLEN:30). Paired reads were then merged
188 using AdapterRemoval 2.2.2 (Schubert, Lindgreen and Orlando, 2016), with default options.

189 Ancient DNA damage assessment

190 Post-mortem DNA damage was measured by DNA fragment length distribution and terminal
191 deamination patterns using mapDamage 2.2.1 (Jónsson *et al.*, 2013). Alignments required for
192 mapDamage were performed with an aligner adapted to short reads BWA-aln 0.7.15 (default options,
193 seed disabled) (Li and Durbin, 2009) for the herbarium samples and Bowtie 2 (options --non-
194 deterministic --very-sensitive) (Langmead and Salzberg, 2012) for modern strains, using *Xci* reference
195 strain IAPAR 306 genome (chromosome NC_003919.1, plasmids pXAC33 NC_003921.3 and pXAC64
196 NC_003922.1). MarkDuplicates in picardtools 2.7.0 (Broad Institute, no date) was run to remove PCR
197 duplicates.

198 Genome reconstruction

199 Genomes were reconstructed by mapping quality-trimmed reads to *Xci* reference strain IAPAR 306
200 genome with either BWA-aln (Li and Durbin, 2009) or Bowtie 2 (Langmead and Salzberg, 2012) aligners
201 as defined above. Sequencing depths were computed using BEDTools genomecov 2.24.0 (Quinlan and
202 Hall, 2010). For herbarium specimens, BAM (Binary Alignment Map) files were extremity-trimmed on
203 their 5 external nucleotides at each end using BamUtil 1.0.14 (Jun *et al.*, 2015). SNPs (Single Nucleotide
204 Polymorphisms) were called with GATK UnifiedGenotyper (DePristo *et al.*, 2011), they were
205 considered dubious and filtered out if they met at least one of the following conditions: “depth<20”,
206 “minor allelic frequency<0.9” and “mapping quality<30”. Consensus sequences were then
207 reconstructed by introducing the high-quality SNPs in *Xci* reference genome and replacing dubious
208 SNPs and non-covered sites (depth=0) by an N.

209 Phylogeny & tree-calibration

210 A dataset of 172 modern *Xci* genomes (date range: 1948 - 2017) representative of *Xci* global diversity
211 was built from 115 previously published genomes and 57 new genomes generated within the course
212 of this study (S1 Table). An alignment of the 13 historical chromosome sequences with the 172 modern

213 sequences was constructed for phylogenetic analyses, with strains of *Xanthomonas axonopodis* pv.
214 *vasculorum* NCPPB-796 from Mauritius (isolated in 1960, GCF_013177355.1), *Xanthomonas citri* pv.
215 *cajani* LMG558 from India (1950, GCF_002019105.1) and *Xanthomonas citri* pv. *clitoriae* LMG9045
216 from India (1974, GCA_002019345.1) used as outgroup. Outgroup sequences were realigned on IAPAR
217 306 chromosome sequence. Variants from modern strains were independently called and filtered
218 using the same parameters as for historical genomes. Regions acquired via horizontal gene transfers
219 were identified inside the *Xci* dataset with ClonalFrameML (Didelot and Wilson, 2015) and removed
220 to account for the effect of recombination on phylogenetic reconstruction and avoid incongruent
221 trees. Two SNPs datasets were constructed, either with variable positions identified inside the *Xci*
222 clade only, or with SNPs positions from across the whole dataset including the outgroups. Maximum
223 Likelihood tree was constructed on both those SNPs alignments using RAxML 8.2.4 (Stamatakis, 2014)
224 using a rapid Bootstrap analysis, a General Time-Reversible model of evolution (Tavaré and Miura,
225 1986) following a Γ distribution with four rate categories (GTRGAMMA) and 1,000 alternative runs.

226 As a requirement to build tip-calibrated phylogenies, the existence of a temporal signal was
227 investigated thanks to three different tests. First, a linear regression test between sample age and
228 root-to-tip distances was computed at each internal node of the ML tree using PhyloStems (Doizy *et*
229 *al.*, 2020). Temporal signal was considered present at nodes displaying a significant positive
230 correlation. Secondly, a date-randomisation test (DRT) (Duchêne *et al.*, 2015) was performed with 20
231 independent date-randomised datasets generated using the R package “TipDatingBeast” (Rieux and
232 Khatchikian, 2017). Temporal signal was considered present when there was no overlap between the
233 inferred root height 95% Highest Posterior Density (95% HPD) of the initial dataset and that of 20 date-
234 randomised datasets. Finally, a Mantel test with 1,000 date-randomised iterations investigating
235 whether closely related sequences were more likely to have been sampled at similar times was also
236 performed to ensure no confounding effect between temporal and genetic structure, as recent work
237 suggested that temporal signal investigation through root to-tip-regression and DRT could be misled
238 in such a case (Murray *et al.*, 2016).

239 Tip-dating calibration Bayesian inferences were performed on the primary SNPs alignment (SNPs
240 identified inside *Xci*) with BEAST 1.8.4 (Drummond and Rambaut, 2007). Leaf heights were constrained
241 to be proportional to sample ages. Flat priors (*i.e.*, uniform distributions) for the substitution rate (10^{-12}
242 to 10^{-2} substitutions/site/year) and for the age of all internal nodes in the tree were applied. We
243 also considered a GTR substitution model with a Γ distribution and invariant sites (GTR+G+I), an
244 uncorrelated relaxed log-normal clock to account for variations between lineages, and a tree prior for
245 demography of coalescent extended Bayesian skyline. The Bayesian topology was conjointly estimated

246 with all other parameters during the Markov chain Monte-Carlo (MCMC) and no prior information
247 from the tree was incorporated in BEAST. Five independent chains were run for 200 million steps and
248 sampled every 20,000 steps, discarding the first 20,000 steps as burn-in. BEAGLE (Broad-platform
249 Evolutionary Analysis General Likelihood Evaluator) library was used to improve computational speed
250 (Suchard and Rambaut, 2009; Ayres *et al.*, 2012). Convergence to the stationary, sufficient sampling
251 (effective sample size > 200) and mixing were checked by inspecting posterior samples with Tracer
252 1.7.1 (Rambaut *et al.*, 2018). Final parameters estimation was based on the combination of the
253 different chains. Maximum clade credibility method in TreeAnnotator (Drummond and Rambaut,
254 2007) was used to determine the best-supported tree of the combined chains.

255 Rate-dating calibration outside the *Xci* clade was performed on the secondary SNPs alignment (SNPs
256 identified across *Xci* and the outgroup) with BEAST 1.8.4 (Drummond and Rambaut, 2007). Instead of
257 using tip-dates, we applied a prior on the substitution rate by drawing values from a normal
258 distribution with mean and standard deviation values fixed as those inferred using tip-dating
259 calibration within *Xci* clade. All other parameters were applied as described previously.

260 Phylogeography & ancestral location state reconstruction

261 The presence of geographic structure in the ML tree was measured through calculation of the
262 Association index (AI) and the comparison of its value with the ones computed from 1,000 location-
263 randomised trees (Parker, Rambaut and Pybus, 2008). Non-random association between phylogeny
264 and location was assumed when less than 5% of AI values computed from the randomised trees were
265 smaller than the AI value of the real ML tree.

266 Ancestral location state was reconstructed using BEAST 1.8.4 (Drummond and Rambaut, 2007) under
267 the same parameters as tip-dating calibration but adding a partition for location character. We
268 modelled discrete location transitioning between areas throughout *Xci* phylogenetic history using a
269 continuous-time Markov chain (CTMC) process under an asymmetric substitution model with a
270 Bayesian stochastic search variable selection (BSSVS) procedure. States were recoded from countries
271 to greater areas: East Africa (Ethiopia), West Africa (Mali and Senegal), Central America and West
272 Indies (Martinique, a French territories), North America (United States of America), South America
273 (Argentina and Brazil), East Asia 1 (China, Hong Kong and Taiwan), East Asia 2 (Japan), South East Asia
274 1 (Cambodia, Malaysia, Myanmar, Thailand and Vietnam), South East Asia 2 (Indonesia and
275 Philippines), South Asia 1 (Bangladesh, India and Nepal), South Asia 2 (Iran and Pakistan), West Asia
276 (Oman and Saudi Arabia), North Indian ocean islands (Maldives and Seychelles), Oceania and Pacific
277 (Fiji, Guam, New Zealand and Papua New Guinea) and South West Indian Ocean islands (Comoros,
278 Mauritius and Rodrigues, Mayotte and La Réunion (the two latter being French territories)).

279 Pathogenicity-associated genes content analysis

280 The presence of pathogenicity-associated genes was investigated using a list of 82 Type III effectors
 281 found in *Xanthomonas* (Escalon *et al.*, 2013; *The Xanthomonas Resource*
 282 (<http://www.xanthomonas.org/t3e.html>), 2018) as well as 64 genes more distantly involved in
 283 pathogenicity (Patané *et al.*, 2019). Alignments were performed either with BWA-aln or Bowtie 2
 284 (same conditions as above), for herbarium samples and modern strains respectively. The sequences
 285 used to assess homology were the reference strain IAPAR 306 CDS when available or other
 286 *Xanthomonas* CDS for genes not present in *Xci*. Depth was recovered using BEDTools genomecov
 287 2.24.0 (Quinlan and Hall, 2010) and coverage was calculated with R. Genes were considered present
 288 if they presented a sequence coverage above a 75% threshold. The pathogenicity-associated genes
 289 content was then projected on ML phylogenetic tree using the gheatmap function of R “ggtree”
 290 package (Yu *et al.*, 2017).

291 Results

292 Laboratory procedure & high-throughput sequencing

293 Herbarium samples sequencing produced between 56.3 and 365.2 M paired-end reads. Following
 294 quality checking and adaptor trimming, reads were merged, presenting insert median length of 32 to
 295 92 nt (mean lengths of 42.1 ± 12.8 to 102.9 ± 45.1 nt).

296 **Table 1. General characteristics of the 13 herbarium specimens.** MNHN: Muséum national d’Histoire
 297 naturelle.

ID	ID herbarium	Herbarium	Collection year	Location	Host
HERB_1845	P05297986	MNHN	1845	Indonesia, Java	<i>Citrus aurantiifolia</i>
HERB_1852	Q1874	Royal Botanic Gardens, Kew	1852	India, Khasi hills	<i>Citrus medica</i>
HERB_1854	Q1954	Royal Botanic Gardens, Kew	1854	Indonesia, Java	<i>Citrus aurantiifolia</i>
HERB_1859	P05240716	MNHN	1859	Bangladesh	<i>Citrus medica</i>
HERB_1865	Q1889	Royal Botanic Gardens, Kew	1865	India	<i>Citrus medica</i>
HERB_1884	1206	Royal Botanic Gardens, Kew	1884	Philippines, Luzon	<i>Citrus medica</i>
HERB_1911	P05297996	MNHN	1911	Indonesia, Java	<i>Citrus aurantiifolia</i>
HERB_1915	P05297992	MNHN	1915	Philippines	<i>Citrus lime</i>
HERB_1922	1756364	U.S. National Herbarium	1922	China, Yunnan	<i>Citrus medica</i>
HERB_1937	MAU0015151	Mauritius Herbarium	1937	Mauritius	<i>Citrus sp.</i>
HERB_1946	686249	U.S. National Fungus Collection	1946	Guam, Tlofofo	<i>Citrus sp.</i>
HERB_1963	630116	Royal Botanic Gardens, Kew	1963	Nepal, Sanichare	<i>Citrus medica</i>
HERB_1974	MAU0015154	Mauritius Herbarium	1974	Mauritius	<i>Citrus lime</i>

298 Historical genomes reconstruction & ancient DNA damage pattern assessment

299 Thirteen historical draft *Xci* genomes were reconstructed by mapping merged reads (after discarding
300 the 5 terminal nucleotides) on reference strain IAPAR 306 genome (chromosome, plasmids pXAC33
301 and pXAC64) (da Silva *et al.*, 2002). Endogenous DNA, defined as the proportion of reads mapping to
302 *Xci* reference genomes varied between 0.8% (HERB_1937) to 27.1% (HERB_1922). Herbarium samples
303 of *Coffea* sp. displayed 90.7 to 92.5% of reads mapping to the *Coffea* reference genome and less than
304 0.5% of the reads mapping to the *Citrus* genome or 0.04 % of DNA mapping to *Xci* genome.

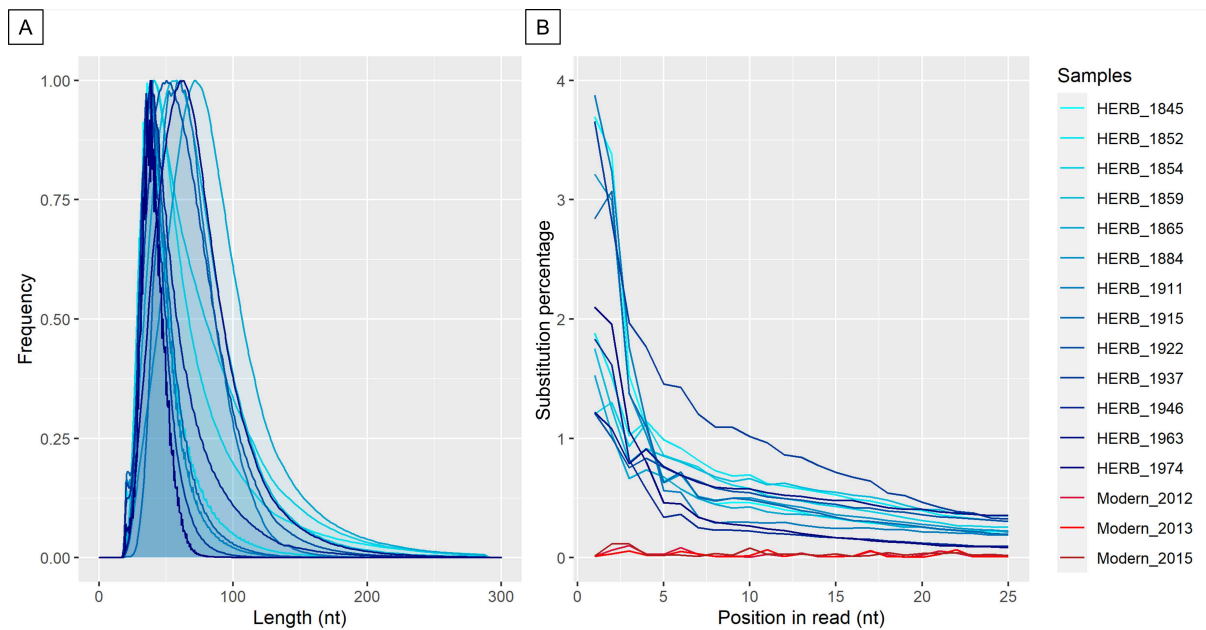
305 The chromosome sequence displayed a coverage (proportion of reference genome covered) at 1X
306 between 94.6 and 98.2%, with mean depth (average number of mapped reads at each base of the
307 reference genome) of 6.2 (HERB_1937) to 96.2X (HERB_1922) (Table 2). Plasmids displayed lower
308 coverages (49.7 to 97.1%) but higher mean depths (17.9 to 130.3X) (S2 Table).

309 **Table 2. Summary of mapping, depth, coverage and damage statistics for the 13 historic *Xci***
310 **genomes.** Mapping statistics are given for both all *Xci* DNA or only its chromosomal part while depth,
311 coverage and damage statistics are indicated for the *Xci* chromosome only. SD: standard deviation; nt:
312 nucleotides.

ID	Protocol	N reads (in millions)	Endogenous <i>Xci</i> DNA (%)	Chromosome			
				Depth (X)	Coverage at 1X (%)	Insert length (mean ± SD, in nt)	Deamination rate at terminal position (%)
HERB_1845	TruSeq Nano	414.3	10.39	82.1	98.3	50.4±23.0	3.68
HERB_1884	TruSeq Nano	246.8	10.48	55.6	98.1	46.2±16.6	3.22
HERB_1911	TruSeq Nano	365.2	2.05	32.4	98.2	69.1±22.3	3.70
HERB_1915	TruSeq Nano	217.3	6.04	39.3	98.1	47.9±17.9	2.81
HERB_1937	TruSeq Nano	220.9	0.82	6.2	94.6	42.7±12.7	3.65
HERB_1946	TruSeq Nano	262.5	7.52	64.3	98.2	57.9±29.5	1.81
HERB_1974	TruSeq Nano	260.8	6.80	35.9	97.9	41.1±9.3	2.09
HERB_1852	BEST	314.9	4.81	63.7	97.8	70.9±43.6	1.89
HERB_1854	BEST	113.0	10.76	54.1	98.3	75.7±42.8	1.22
HERB_1859	BEST	159.5	5.31	41.8	97.7	73.9±33.8	1.76
HERB_1865	BEST	156.5	4.91	49.8	98.2	88.1±39.2	1.57
HERB_1922	BEST	120.9	27.10	96.2	97.3	67.9±29.3	1.22
HERB_1963	BEST	56.3	14.63	43	98.0	74.7±32.3	1.25

313
314 Ancient DNA is typically degraded, presenting short fragments and cytosine deamination at fragment
315 extremities (Pääbo *et al.*, 2004; Dabney, Meyer and Pääbo, 2013). We searched for such patterns of
316 degradation in our historical genomes using the dedicated tool mapDamage2 (Jónsson *et al.*, 2013).
317 For the chromosome sequence, herbarium specimens displayed mean fragments length of 41 ± 9 to

318 88 ± 39 nt and 5'C/T substitution rates at terminal nucleotides of 1.21 to 3.88% (Fig 1). We observed
 319 an exponential decrease of both 5'C/T and complementary 3'G/A substitutions along the DNA
 320 molecule for all historical genomes. Modern DNA from three *Xci* strains serving as control displayed
 321 no such decay. Global similar pattern was detected for plasmid sequences (S1 Fig and S2 Table).



322

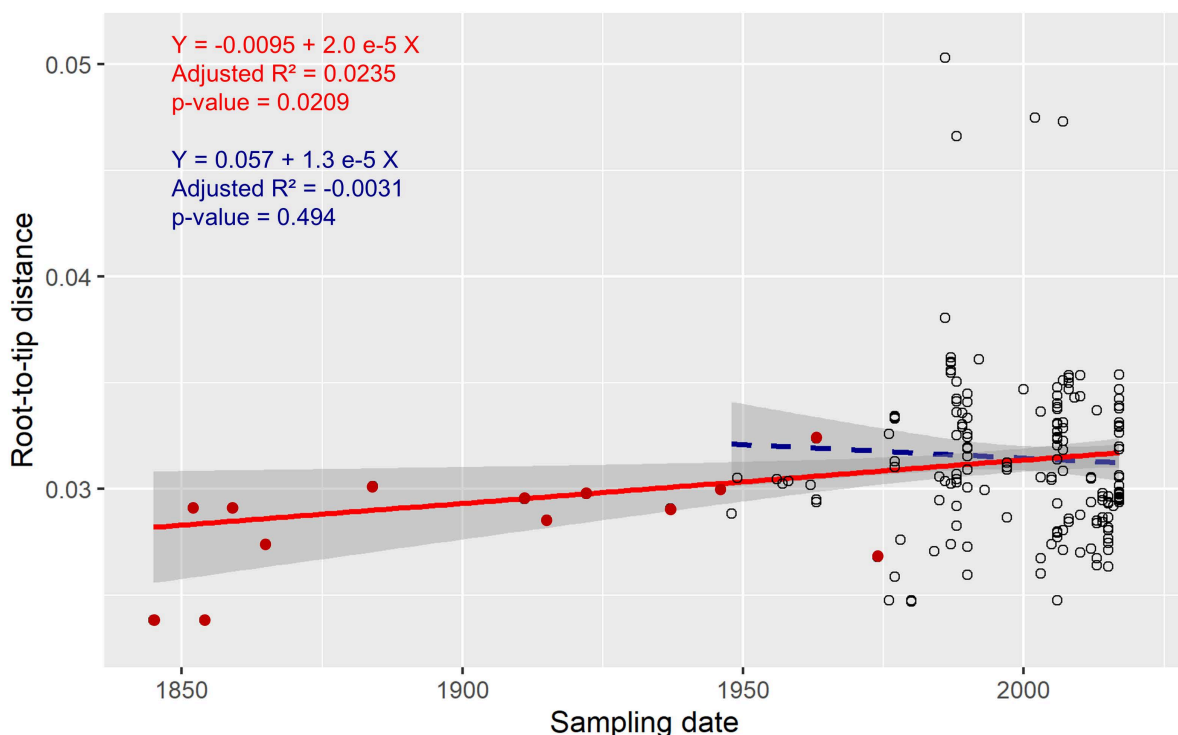
323 **Fig 1. Chromosomal post-mortem DNA damage patterns measured on *Xci* chromosome.** (A)
 324 Fragment length distribution (nucleotides; relative frequency in arbitrary units). (B) Deamination
 325 percentage of the first 25 nucleotides from the 5' end, respectively measured on the 13 historical
 326 genomes (blue lines) and on three modern *Xci* strains (red lines).

327 Phylogenetic reconstruction, dating and ancestral geographic state estimation

328 Alignment of the chromosome sequence of 13 historical genomes and 172 modern genomes (17
 329 modern strains from the A* pathotype, 4 from the A^W pathotype and 151 from the A pathotype)
 330 allowed for the identification of 15,292 high-quality SNPs (Single Nucleotide Polymorphisms).
 331 ClonalFrameML (Didelot and Wilson, 2015) identified four major recombining regions (S3 Table), from
 332 which 2,285 SNPs were removed from further inferences. On the 13,007 recombination-free SNPs
 333 alignment, a paraphyletic outgroup was added, formed of *Xanthomonas axonopodis* pv. *vasculorum*
 334 NCPPB-796 and two strains phylogenetically close to *Xci*, *Xanthomonas citri* pv. *cajani* LMG558 and
 335 *Xanthomonas citri* pv. *clitoriae* LMG9045. A Maximum-Likelihood (ML) phylogeny was built with
 336 RAxML (Stamatakis, 2014) and rooted with *Xanthomonas axonopodis* pv. *vasculorum* (S2 Fig). Strains
 337 from each pathotype grouped together and formed distinct clades. Clade A* was at the root of clade
 338 *Xci*, while clade A^W is a sister-group of clade A. This topology (A*, (A^W, A)) was highly supported with
 339 bootstraps values of 100. Clade A displayed three major, highly supported lineages (A3, (A2, A1)) :
 340 lineage A3 contained seven strains from Bangladesh as well as three historical specimens from

341 Bangladesh, India, and China (Yunnan). Lineage A2, contained strains from India and Senegal but also
 342 from Pakistan and Mali. Finally, lineage A1 contained ten historical specimens and the rest of the
 343 pathotype A strains, most of them in a polytomy. Historical specimens were mainly clustered with
 344 geographically close modern strains. Overall, clade A displayed a geographical clustering of strains,
 345 with a high genetic diversity of the Asian genomes.

346 The ML tree was used to test the presence of temporal signal (*i.e.* progressive accumulation of
 347 mutations over time) within the *Xci* clade using three different tests. The linear regression test
 348 between root-to-tip distances and sampling ages displayed a significantly positive slope
 349 (value= 36.8×10^{-6} , adjusted $R^2=0.0235$ with a p-value=0.0209) (Fig 2). Interestingly, such a pattern was
 350 also conserved at several other internal nodes (S3 Fig). Second, the BEAST inferred root age and
 351 substitution rates of the real *versus* date-randomised datasets exhibited no overlap of the 95% HPD
 352 (Highest Posterior Density) (Fig 2). Finally, the Mantel test displayed no confounding effect ($r=-0.73$,
 353 p-value=0.89) between temporal and genetic structures. To specifically evaluate the contribution of
 354 historical genomes to the magnitude of temporal signal, we repeated the above tests on a dataset
 355 containing modern genomes only, producing no temporal signal at the *Xci* clade scale (Fig 2 and S4
 356 Fig).

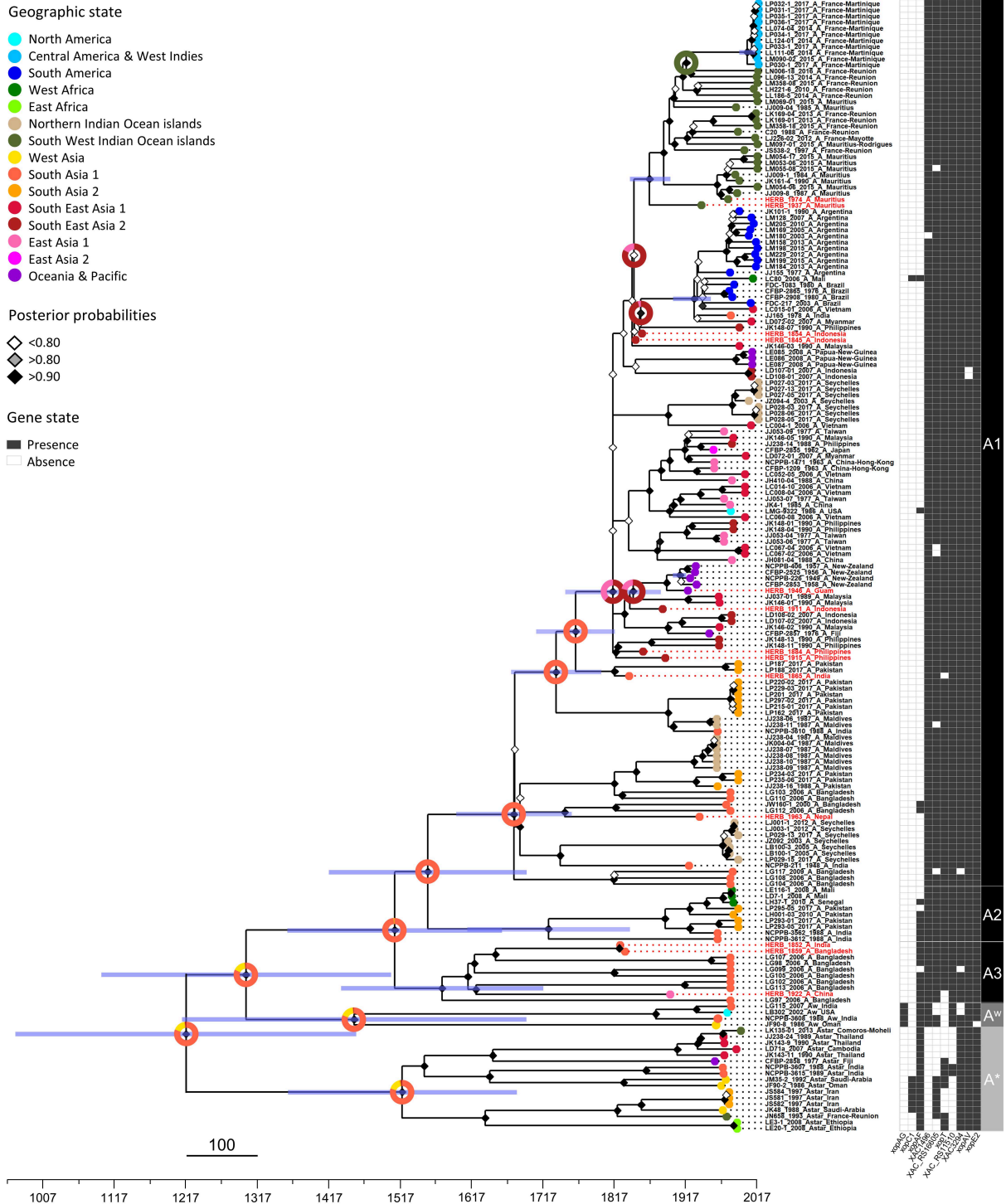


357

358 **Fig 2. Root-to-tip regression temporal test results performed either with or without including**
 359 **historical genomes in the dataset.** Regression lines are plotted in red when integrating historical

360 genomes (red dots) and in blue dotted lines when not. Grey areas indicate confidence intervals.
361 Associated values are the regression equation, adjusted R^2 (Adj R^2) and p-value.

362 A Bayesian time-calibrated tree was therefore built with BEAST (Drummond and Rambaut, 2007) (Fig
363 3), which was globally congruent (similar topology and node supports) with the ML tree. The root of
364 the *Xci* clade (node at which *Xci* diversified into numerous pathotypes) was inferred to date to 1218
365 [95% HPD: 962 - 1437]. We obtained a mean substitution rate of 14.30×10^{-8} [95% HPD: 12.47×10^{-8} -
366 16.14×10^{-8}] per site per year with a standard deviation for the uncorrelated log-normal clock of 0.507
367 [95% HPD: 0.428 - 0.594], suggesting low rate heterogeneity among tree branches. The inferred dates
368 of other internal nodes of interest, including the MRCA for each of the three pathotypes as well as of
369 some geographically structured clades are given in Table 3.



370

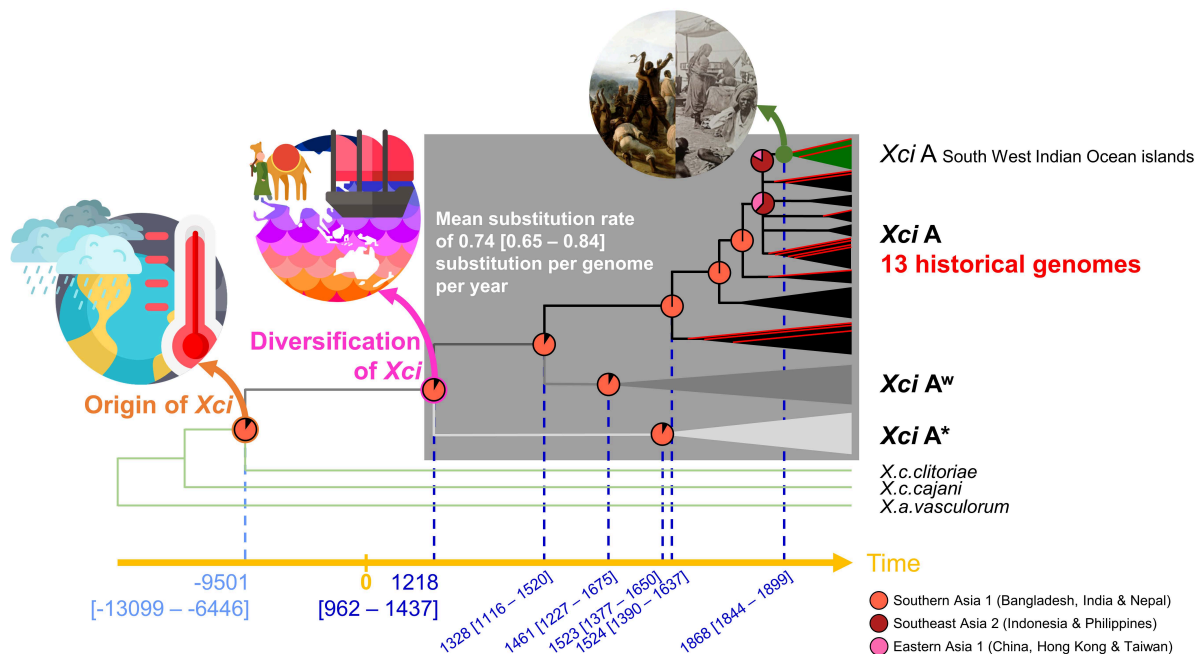
371 **Fig 3. Spatiotemporal Bayesian reconstruction of *Xci* evolutionary history.** Dated phylogenetic tree
 372 including 13 historical specimens (labelled in red) and 172 modern strains (black) built from 13,007
 373 recombination-free SNPs. Node support values are displayed by diamonds (white, grey and black for
 374 posterior probabilities below 0.80, above 0.80 and above 0.90 respectively); node bars cover 95%
 375 Highest Probability Density of node height. Tips points are coloured according to the sample's
 376 geographic area of origin. Ancestral geographic states are reconstructed at nodes with pie charts
 377 representing the posterior probability of each state. Tip labels include collection date, pathotype and
 378 country of origin. Heatmap indicates major clades (A*, A^W & A) as well as presence/absence of the
 379 eight pathogen-associated genes with variable pattern.

380 To date the origin of *Xci*, *i.e.* the date at which it diverged from its closest relatives, a secondary
381 alignment of the chromosome sequences was realised with the 13 historical sequences, the 172
382 modern sequences and the three outgroup sequences. A total of 209,306 high-quality SNPs was found
383 across the dataset, 198,249 of which were outside recombining regions. The presence of temporal
384 signal was tested as previously described. On this dataset, temporal signal was still present at the root
385 of *Xci* clade but the signal disappeared at the external node connecting the outgroup, with
386 *Xanthomonas citri* pv. *cajani* as *Xci* closest relative (root-to-tip regression test: slope value= -4.6×10^{-6} ,
387 adjusted $R^2=0.0006$ with a p-value=0.293; date-randomisation test: overlapping of the 95% HPD of the
388 root age (S5 Fig); Mantel's test: $r=0.04$, p-value=0.09). As the prerequisite for tip-dating was not met,
389 we realised a rate-dating analysis integrating the mean substitution rate inferred in the tip-dating
390 calibration. The node date at which *Xci* split from other *Xanthomonas citri* pathovars was inferred to -
391 9501 [95% HPD: -13099 - -6446].

392 In order to infer ancestral location state at nodes, geographic signal in the dataset must first be
393 detected. By measuring the association index (AI) between topology and the location trait data on
394 both the real tree and 1,000 location-randomised trees, our results highlighted the existence of a non-
395 random association between spatial and phylogenetic structure ($r=-0.73$, p-value=0.89). A discrete
396 phylogeographic analysis was therefore run with BEAST (Drummond and Rambaut, 2007) (Fig 3, Table
397 3), inferring a highly-supported South Asia 1 (Bangladesh, India and Nepal) origin where *Xci* split from
398 its *Xanthomonas citri* relatives and then diversified into its three lineages A*, A^W and A. Each of these
399 lineages were also inferred to have diversified in the same region, as well as the A3, A2 and A1 lineages
400 inside the clade of the pathotype A. In the lineage A1, the polytomy was inferred to have a Southeast
401 Asia 2 (Indonesia or Philippines) origin (node support of 1, state posterior probability of 0.62). The
402 same geographic origin was inferred for the lineage composed of all New Zealand strains which had
403 herbarium specimen HERB_1946_A_Guam at its root, the lineage comprising of South America strains
404 (with HERB_1854_A_Indonesia and HERB_1845_A_Indonesia branching closeby) and the lineage with
405 historical specimens from Mauritius and strains from the Southwest Indian ocean (SWIO) islands (the
406 Comoros, Mauritius and Rodrigues, Mayotte and La Réunion), which included all strains from the West
407 Indies (Martinique island).

408 **Table 3. Inferred spatiotemporal data at major nodes.** Results were compared with estimations
 409 inferred in Patané *et al.* (2019) (green) when possible. Abbreviations: *Xci*, *Xanthomonas citri* pathovar
 410 *citri*; SWIO, south west Indian ocean.

Node	Date[95% Highest Probability Density]	Location (posterior probability)
<i>Xci</i> origin	-9501 [-13099 - -6446] [-12052- -3851]	South Asia 1 (0.85) Bangladesh, India, Pakistan
<i>Xci</i> diversification	1218 [962 - 1437] [-3648 - 285]	South Asia 1 (0.81) Bangladesh, India, Pakistan
<i>Xci</i> A* (diversification)	1523 [1377 - 1650]	South Asia 1 (0.80) Iran, Oman, Saudi Arabia
<i>Xci</i> A ^W (diversification)	1461 [1227 - 1675]	South Asia 1 (0.80) Bangladesh, India, Pakistan
<i>Xci</i> A (diversification)	1524 [1390 - 1637]	South Asia 1 (0.99) Bangladesh, India, Pakistan
<i>Xci</i> A - New Zealand	1925 [1896 - 1949]	South East Asia 2 (0.79)
<i>Xci</i> A - South America	1930 [1907 - 1951]	South East Asia 2 (0.96)
<i>Xci</i> A - SWIO islands	1868 [1844 - 1899]	South East Asia 2 (0.83)
<i>Xci</i> A - West Indies	2000 [1991 - 2008]	SWIO (1.00)



411
 412 **Fig 4. Schematic illustration of the study main results.**

413 Pathogenicity-associated genes content

414 We investigated the presence or absence of 146 pathogenicity-associated genes (S4 Table) under a
 415 mapping approach: more particularly, the 92 genes coding for the type III secretion system (T3SS), a
 416 syringe-like apparatus, and the “effectors” (T3E) it injects inside the plant cell effectors that inhibit the
 417 plant defences and contribute to the development of symptoms among *Xanthomonas* species (Escalon
 418 *et al.*, 2013; Büttner, 2016; *The Xanthomonas Resource* (<http://www.xanthomonas.org/t3e.html>),

419 2018), as well as 54 genes more distantly involved (Patané *et al.*, 2019). Coding sequences (CDS)
420 covered on less than 75% of their length were considered absent. Among the 146 CDS, 108 were
421 always present, comprising the entire set of genes necessary for the T3SS (24 CDS) and potential T3E
422 (35 CDS) identified for *Xci* except for gene *xopAF* which was present in A* and A^W clades but absent in
423 most A strains, *xopAV* which was absent in a lineage of two strains from Indonesia, as well as *xopE2*
424 which was absent in a single A^W strain (JF90-8_1986_Aw_Oman) and *xopAG* present only among A^W
425 strains. There were 32 genes always absent across all *Xci* samples, all coding for T3E among
426 *Xanthomonas* species other than *Xci*. Six other genes displayed variable presence in the dataset:
427 *xopC1* was absent in all *Xci* but 2 lineages in the A* clade and 1 A strain, *xopT* was present across all
428 A strains except 3, present in 8 A* strains but absent in A^W strains, XAC1496 was found in all A^W and
429 A strains (except 1 A modern strain) but no A* strain, XAC3294 was present in all A* strains and A
430 (except 2) strains and 2 of the 4 A^W strains, XAC_RS11510 was present in all A and A^W strains but
431 only 2 A* strains, finally, XAC_RS16605 was present across most A samples (absent in 5 strains),
432 present in 2 of the 4 A^W strains and 7 A* strains. However, no pattern seemed to entirely distinguish
433 the three pathotypes clades inside *Xci* (Fig 3).

434 Discussion

435 The evolutionary history of *Xci* is a subject of great interest since it may help understanding how
436 bacterial pathogens specialise on a host and how they diversify while expanding their geographical
437 range. In this study, we successfully reconstructed the genome of 13 historical strains from herbarium
438 material collected between 1845 and 1974, which we compared with a set of 172 modern genomes
439 representative of the bacterial global diversity, 57 of them having been specifically generated within
440 the course of this study. The inclusion of historical genomes allowed us building a time-calibrated
441 phylogeny without making any underlying assumption on the age of any node in the tree, nor on the
442 rate of evolution and proposing new evolutionary *scenarii* for the origin and diversification of the
443 pathogen. To our knowledge, this is the first study attempting to elucidate the evolutionary history of
444 a bacterial crop pathogen at such a global scale using herbarium specimens. Previous ones did not
445 use historical strains (Patané *et al.*, 2019) or, when they did, either focused on a recent and local
446 emergence (Campos *et al.*, 2021) or were limited by the exploitation of a few partial genetic markers
447 only (Li *et al.*, 2007).

448 Adopting a shotgun-based deep sequencing strategy revealed between 0.8% to 25.5% of endogenous
449 *Xci* DNA amongst the 13 historical samples, a wide variation falling in the range of previous studies
450 that attempted to retrieve non vascular pathogen DNA from infected herbarium leaves (Martin *et al.*,
451 2013; Yoshida *et al.*, 2013; Yoshida, Sasaki and Kamoun, 2015). Importantly, assessment of *post-*

452 *mortem* DNA degradation patterns specific to ancient DNA, such as fragmentation and deamination,
453 confirmed the historical nature of the reconstructed genomes with degradation patterns also
454 consistent with expected values from the literature (Martin *et al.*, 2013; Yoshida *et al.*, 2013; Yoshida,
455 Sasaki and Kamoun, 2015; Weiß *et al.*, 2016).

456 Following authentication, we aligned the 13 historical genomes with 172 modern ones representative
457 of the bacterial global diversity, and after removal of the recombinant regions, built a phylogenetic
458 tree from the chromosome-wide vertically inherited SNPs of the genome. This tree confidently
459 associated the pathotypes to be monophyletic groups, and displayed a (A*, (A^w, A)) topology, as
460 previously reported from genome-wide (Gordon *et al.*, 2015) and unicopy gene families analyses
461 (Patané *et al.*, 2019), respectively. Interestingly, the relationships inside the pathotype A clade (A3,
462 (A2, A1)) agreed with the former analysis but not the latter, a discrepancy explained by the under-
463 representation of lineage A3 in Patané *et al.* (2019). Globally, the observed geographic clustering
464 inside the pathotype A clade is consistent with previous studies (Pruvost *et al.*, 2014; Patané *et al.*,
465 2019; Richard *et al.*, 2020).

466 The presence of temporal structure is an essential prerequisite to perform tip-calibrated inferences
467 (Drummond *et al.*, 2002, 2003; Rieux and Balloux, 2016). While the dataset only containing
468 contemporary genomes (1948 - 2017) did not reveal the existence of any measurably evolving
469 population, inclusion of the 13 historical genomes (1845 - 1974) brought the required temporal signal
470 within the *Xci* clade. We inferred a mean substitution rate of 14.30×10^{-8} [95% HPD: 12.47×10^{-8} -
471 16.14×10^{-8}] substitutions per site per year, a value $\sim 1.5 \times$ faster than the one (9.4×10^{-8} [95% HPD:
472 7.3×10^{-8} - 11.4×10^{-8}]) obtained by Campos *et al.* (2021) on a single lineage within the pathotype A clade
473 of *Xci*, at the local scale of the South West Indian Ocean islands. We dated the MRCA of all *Xci* strains,
474 the node leading to bacterial diversification, to the beginning of the 13th century (1218 [95% HPD:
475 962 - 1437]), a much more recent timespan than the one [-3648 - 285] previously inferred by Patané
476 *et al.* (2019). The discrepancy between those estimates probably arises from differences in the
477 considered molecular dating methodologies. Indeed, Patané *et al.* (2019) calibrated their molecular
478 clock by applying a prior on both the rate of evolution (estimated on house-keeping genes of non-*Xci*
479 *Xanthomonas* species) and the age of a node external to the *Xci* clade (indirectly deriving from the
480 same rate of evolution) while the methodology used in this study only makes use of the age of the
481 strains, a method shown to yield far more accurate and robust estimates than the former ones (Ho
482 *et al.*, 2008; Rieux *et al.*, 2014; Rieux and Balloux, 2016). In addition of dating the emergence of the
483 three *Xci* lineages A*, A^w & A, we also inferred the ones of geographically structured lineages such as
484 the one in New Zealand in 1925 [95% HPD: 1896 - 1949], in South-America in 1930 [95% HPD: 1907 -
485 1951] or in Martinique in 2000 [95% HPD: 1991 - 2008] with values always predating disease first

486 reports made in 1937 (Dye, 1969), 1972 (Rosseti, 1977) and 2014 (Richard *et al.*, 2016), respectively.
487 Finally, as the divergence between a pathogen and its closest known relative places a maximum bound
488 on the timing of its emergence (Duchêne *et al.*, 2020), we included three outgroup sequences, of
489 which *Xanthomonas citri* pv. *cajani*, a pathogen of the *Fabaceae* plant family, was the first to branch
490 out of the *Xci* clade. As the inclusion of divergent outgroup genomes precluded the application of tip-
491 dating methodology we extrapolated the rate of substitution previously estimated within the *Xci*
492 clade to date the split between *Xci* and *Xanthomonas citri* pv. *cajani* to -9501 [95% HPD: -13099 -
493 -6446], a value which partly overlap with the inferred interval of [-12052 - -3851] found by Patané *et*
494 *al.* (2019), although on a more restrained length of time.

495 Our phylogeographic analysis inferred an origin and diversification of *Xci* in an area of South Asia
496 neighbouring to Bangladesh, India and Nepal, consistently with previous reconstructions (Patané *et*
497 *al.*, 2019) and estimations based on genetic diversity (Das, 2003; Pruvost *et al.*, 2014; Gordon *et al.*,
498 2015). Altogether, our spatio-temporal calibrations found a maximum bound of 11.5 ky old for the
499 origin of *Xci* in the area of the Bangladesh, India and Nepal, a period which coincides with the
500 beginning of the Holocene (~9700 - present) (Walker *et al.*, 2009) following the end of the Last Glacial
501 Period and the Bølling-Allerød warming global event (~12700 - -10900) (Rasmussen *et al.*, 2006). Such
502 warmer and wetter climates could have facilitated plants expansion into new areas previously
503 occupied by ice such as the mountainous regions and the northern parts of South Asia (Staubwasser
504 and Weiss, 2006). There, *Fabaceae* and *Rutaceae* could have been sympatric, allowing for bacterial
505 host jump from the former to the latter and the speciation of *Xanthomonas citri* population into the
506 emerging pathogenic *Xci*, as previously proposed (Patané *et al.*, 2019). The diversification of *Xci* was
507 dated to the early 13th century in Bangladesh, India or Nepal, which were crossed at the time by the
508 southern Silk Road (Talon, Caruso and Gmitter Jr., 2020), linking East and West civilisations through
509 trading. The westward commerce of goods, including citrus, which have been found in mediterranean
510 countries since -500 (Zech-Matterne and Fiorentino, 2017), as well as the breeding of citrus varieties
511 for cooking and for raw eating (Talon, Caruso and Gmitter Jr., 2020), could have dispersed and isolated
512 the pathogen into its three known pathotypes. Furthermore, the younger lineages could have
513 dispersed internationally by trade as South and South East Asia had participated in the spice trade in
514 the 16th century and then under European colonial rules til the 20th century.

515 The presence/absence analysis of the 146 pathogeny-associated genes did not reveal any clear
516 pathotype-associated variation pattern as strains from each clade displayed some content variability,
517 especially A* strains. *Xci* strains all presented the full set of *Xci* genes necessary for the T3SS and
518 potential T3E except three strains for genes of putative protein or unknown function (*xopE2* and
519 *xopAV*, respectively). *XopAG* gene, involved in hypersensitive response-like symptoms formation

520 (Rybak et al., 2009), was only found in A^W strains (Escalon et al., 2013; Patané et al., 2019). XopC1,
521 coding for a phosphoribosyl transferase domain and haloacid dehalogenase-like hydrolase, was found
522 among A* strains, as previously described (Escalon *et al.*, 2013; Patané *et al.*, 2019), as well as a
523 pathotype A strain. The absence of this T3E gene does not affect pathogenicity (Escalon *et al.*, 2013).
524 This pathogeny-associated gene content lineage-to-lineage variability, especially marked for A*
525 strains, did not seem correlated with host specialisation observed in the different pathotypes, as has
526 been previously proposed (Escalon *et al.*, 2013), suggesting the involvement of other genes or versions
527 of genes. Historical strains did not display peculiar patterns compared to modern ones, hinting at a
528 rather time-conserved repertory of pathogenicity-associated genes over the last 200 years with local,
529 lineage variability inside pathotype A.

530 Our work presents two main limitations. First, although with 164 genomes our dataset displayed the
531 best representation of pathotype A genetic diversity published to date, the reconstructed
532 phylogenetic tree exhibits high level of imbalance (with only 17 and 4 A* and A^W genomes,
533 respectively), a property previously shown to lead to reduced accuracy or precision of phylogenetic
534 timescale estimates (Duchêne, Duchêne and Ho, 2015). Bias in representation of populations, such as
535 overrepresentation of one compared to the others or the absence of representants from the true
536 founder lineage, can also lead to the reconstruction of ancestral state tending to correspond to the
537 oversampled population rather than the true founder lineage (Rasmussen and Grünwald, 2020).
538 Although this feature arises from the fact that *Xci* worldwide expansion mostly involved pathotype A
539 strains (Pruvost *et al.*, 2014), future work should aim to better characterise the genomic diversity of
540 A* and A^W genomes. Secondly, as gene content variation analysis was performed by mapping reads to
541 reference sequences, we were unable to identify potential genomic rearrangements among strains, a
542 process known to be frequent within *Xanthomonas* species (Jacques *et al.*, 2016). Similarly, this
543 approach impeded us from identifying genetic content absent from the reference sequences. To
544 overcome those limitations and better recover pathotype-specific genes, comparative genomic
545 analysis based on *de novo* assembly and/or without *a priori* on the targeted genes would be interesting
546 to perform.

547 To conclude, our study emphasises how historical genomes from herbarium samples can provide a
548 wealth of genetic and temporal information on bacterial crop pathogens evolution. Similar studies
549 could be applied to other plant pathogens to infer the temporal dynamic of their populations and
550 elucidate their evolutionary history with more resolute estimations, which in turn may provide clues
551 to improve disease monitoring and achieve sustainable control.

552 References

- 553 Al Rwahnih, M., Rowhani, A. and Golino, D. (2015) 'First report of Grapevine red blotch-associated
554 virus in archival grapevine material from Sonoma County, California', *Plant Disease*, 99(6), p. 895.
555 doi:10.1094/PDIS-12-14-1252-PDN.
- 556 Anderson, P.K. *et al.* (2004) 'Emerging infectious diseases of plants: pathogen pollution, climate
557 change and agrotechnology drivers', *Trends in Ecology & Evolution*, 19(10), pp. 535–544.
558 doi:10.1016/j.tree.2004.07.021.
- 559 Ayres, D.L. *et al.* (2012) 'BEAGLE: An Application Programming Interface and High-Performance
560 Computing Library for Statistical Phylogenetics', *Systematic Biology*, 61(1), pp. 170–173.
561 doi:10.1093/sysbio/syr100.
- 562 Bernades, M.F.F. *et al.* (2015) 'Impact of pesticides on environmental and human health', in
563 *Toxicology studies: cells, drugs and environment*. Rijeka, Croatia (InTech), pp. 195–233.
- 564 Biek, R. *et al.* (2015) 'Measurably evolving pathogens in the genomic era', *Trends in Ecology &
565 Evolution*, 30(6), pp. 306–313. doi:10.1016/j.tree.2015.03.009.
- 566 Bolger, A.M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: a flexible trimmer for Illumina sequence
567 data', *Bioinformatics*, 30(15), pp. 2114–2120. doi:10.1093/bioinformatics/btu170.
- 568 Broad Institute (no date) *Picard Tools* (<http://broadinstitute.github.io/picard/>).
- 569 Büttner, D. (2016) 'Behind the lines—actions of bacterial type III effector proteins in plant cells', *FEMS
570 Microbiology Reviews*, 40(6), pp. 894–937. doi:10.1093/femsre/fuw026.
- 571 Campos, P.E. *et al.* (2021) 'First historical genome of a crop bacterial pathogen from herbarium
572 specimen: Insights into citrus canker emergence', *PLOS Pathogens*. Edited by D. Mackey, 17(7), pp.
573 1–25. doi:10.1371/journal.ppat.1009714.
- 574 Carøe, C. *et al.* (2018) 'Single-tube library preparation for degraded DNA', *Methods in Ecology and
575 Evolution*. Edited by S. Johnston, 9(2), pp. 410–419. doi:10.1111/2041-210X.12871.
- 576 Dabney, J., Meyer, M. and Pääbo, S. (2013) 'Ancient DNA damage', *Cold Spring Harbor Perspectives
577 in Biology*, 7, pp. 1–8. doi:10.1101/cshperspect.a012567.
- 578 Das, A.K. (2003) 'Citrus canker - a review', *Journal of Applied Horticulture*, 5(1), pp. 52–60.
- 579 DePristo, M.A. *et al.* (2011) 'A framework for variation discovery and genotyping using next-
580 generation DNA sequencing data', *Nature Genetics*, 43(5), pp. 491–498. doi:10.1038/ng.806.
- 581 Didelot, X. and Wilson, D.J. (2015) 'ClonalFrameML: efficient inference of recombination in whole
582 bacterial genomes', *PLoS Computational Biology*, 11(2), pp. 1–18. doi:10.1371/journal.pcbi.1004041.
- 583 Doizy, A. *et al.* (2020) 'Phylostems: a new graphical tool to investigate temporal signal of
584 heterochronous sequences at various evolutionary scales', *BioRxiv*, pp. 1–23.
585 doi:10.1101/2020.10.19.346429.
- 586 Drummond, A.J. *et al.* (2002) 'Estimating mutation parameters, population history and genealogy
587 simultaneously from temporally spaced sequence data', *Genetics*, 161(3), pp. 1307–1320.
588 doi:10.1093/genetics/161.3.1307.
- 589 Drummond, A.J. *et al.* (2003) 'Measurably evolving populations', *Trends in Ecology & Evolution*,

590 18(9), pp. 481–488. doi:10.1016/S0169-5347(03)00216-7.

591 Drummond, A.J. and Rambaut, A. (2007) ‘BEAST: Bayesian evolutionary analysis by sampling trees’,
592 *BMC Evolutionary Biology*, 7(1–8), p. 214. doi:10.1186/1471-2148-7-214.

593 Duchêne, D., Duchêne, S. and Ho, S.Y.W. (2015) ‘Tree imbalance causes a bias in phylogenetic
594 estimation of evolutionary timescales using heterochronous sequences’, *Molecular Ecology*
595 *Resources*, 15(4), pp. 785–794. doi:10.1111/1755-0998.12352.

596 Duchêne, S. *et al.* (2015) ‘The performance of the date-randomization test in phylogenetic analyses
597 of time-structured virus data’, *Molecular Biology and Evolution*, 32(7), pp. 1895–1906.
598 doi:10.1093/molbev/msv056.

599 Duchêne, S. *et al.* (2020) ‘The recovery, interpretation and use of ancient pathogen genomes’,
600 *Current Biology*, 30(19), pp. R1215–R1231. doi:10.1016/j.cub.2020.08.081.

601 Dye, D. (1969) ‘Eradicating citrus canker from New Zealand’, *New Zealand Journal of Agriculture*,
602 118, pp. 20–21.

603 Escalon, A. *et al.* (2013) ‘Variations in type III effector repertoires, pathological phenotypes and host
604 range of *Xanthomonas citri* pv. *citri* pathotypes: type III effectors in *Xanthomonas citri* pv. *citri*’,
605 *Molecular Plant Pathology*, 14(5), pp. 483–496. doi:10.1111/mpp.12019.

606 Gordon, J.L. *et al.* (2015) ‘Comparative genomics of 43 strains of *Xanthomonas citri* pv. *citri* reveals
607 the evolutionary events giving rise to pathotypes with different host ranges’, *BMC Genomics*,
608 16(1098), pp. 1–20. doi:10.1186/s12864-015-2310-x.

609 Gottwald, T.R., Graham, J.H. and Schubert, T.S. (2002) ‘Citrus canker: the pathogen and its impact’,
610 *Plant Health Progress*, pp. 1–34. doi:10.1094/PHP-2002-0812-01-RV.

611 Graham, J.H. *et al.* (2004) ‘*Xanthomonas axonopodis* pv. *citri*: factors affecting successful eradication
612 of citrus canker’, *Molecular Plant Pathology*, 5(1), pp. 1–15. doi:10.1046/j.1364-3703.2004.00197.x.

613 Ho, S.Y.W. *et al.* (2008) ‘The effect of inappropriate calibration: three case studies in molecular
614 ecology’, *PLoS ONE*. Edited by P. Bennett, 3(2), pp. 1–8. doi:10.1371/journal.pone.0001615.

615 Jacques, M.-A. *et al.* (2016) ‘Using ecology, physiology, and genomics to understand host specificity
616 in *Xanthomonas*’, *Annual Review of Phytopathology*, 54(1), pp. 163–187. doi:10.1146/annurev-
617 phyto-080615-100147.

618 Joint Genome Institute (1997) *BBTools*. Available at: [https://jgi.doe.gov/data-and-tools/bbtools/bb-
619 tools-user-guide/bbduk-guide/](https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/) (Accessed: 1 November 2018).

620 Jónsson, H. *et al.* (2013) ‘mapDamage2.0: fast approximate Bayesian estimates of ancient DNA
621 damage parameters’, *Bioinformatics*, 29(13), pp. 1682–1684. doi:10.1093/bioinformatics/btt193.

622 Jun, G. *et al.* (2015) ‘An efficient and scalable analysis framework for variant extraction and
623 refinement from population-scale DNA sequence data’, *Genome Research*, 25(6), pp. 918–925.
624 doi:10.1101/gr.176552.114.

625 Langmead, B. and Salzberg, S.L. (2012) ‘Fast gapped-read alignment with Bowtie 2’, *Nature Methods*,
626 9(4), pp. 357–359. doi:10.1038/nmeth.1923.

627 Li, H. and Durbin, R. (2009) ‘Fast and accurate short read alignment with Burrows-Wheeler
628 transform’, *Bioinformatics*, 25(14), pp. 1754–1760. doi:10.1093/bioinformatics/btp324.

- 629 Li, W. *et al.* (2007) 'Genetic diversity of citrus bacterial canker pathogens preserved in herbarium
630 specimens', *Proceedings of the National Academy of Sciences*, 104(47), pp. 18427–18432.
631 doi:10.1073/pnas.0705590104.
- 632 Malmstrom, C.M. *et al.* (2007) 'Barley yellow dwarf viruses (BYDVs) preserved in herbarium
633 specimens illuminate historical disease ecology of invasive and native grasses', *Journal of Ecology*,
634 95(6), pp. 1153–1166. doi:10.1111/j.1365-2745.2007.01307.x.
- 635 Martin, M.D. *et al.* (2013) 'Reconstructing genome evolution in historic samples of the Irish potato
636 famine pathogen', *Nature Communications*, 4(2172), pp. 1–7. doi:10.1038/ncomms3172.
- 637 Mira, A., Pushker, R. and Rodríguez-Valera, F. (2006) 'The Neolithic revolution of bacterial genomes',
638 *Trends in Microbiology*, 14(5), pp. 200–206. doi:10.1016/j.tim.2006.03.001.
- 639 Murray, G.G.R. *et al.* (2016) 'The effect of genetic structure on molecular dating and tests for
640 temporal signal', *Methods in Ecology and Evolution*. Edited by M. Gilbert, 7(1), pp. 80–89.
641 doi:10.1111/2041-210X.12466.
- 642 Pääbo, S. *et al.* (2004) 'Genetic analyses from ancient DNA', *Annual Review of Genetics*, 38(1), pp.
643 645–679. doi:10.1146/annurev.genet.37.110801.143214.
- 644 Parker, J., Rambaut, A. and Pybus, O.G. (2008) 'Correlating viral phenotypes with phylogeny:
645 Accounting for phylogenetic uncertainty', *Infection, Genetics and Evolution*, 8(3), pp. 239–246.
646 doi:10.1016/j.meegid.2007.08.001.
- 647 Patané, J.S.L. *et al.* (2019) 'Origin and diversification of *Xanthomonas citri* subsp. *citri* pathotypes
648 revealed by inclusive phylogenomic, dating, and biogeographic analyses', *BMC Genomics*, 20(700),
649 pp. 1–23. doi:10.1186/s12864-019-6007-4.
- 650 Pruvost, O. *et al.* (2014) 'A MLVA genotyping scheme for global surveillance of the citrus pathogen
651 *Xanthomonas citri* pv. *citri* suggests a worldwide geographical expansion of a single genetic lineage',
652 *PLoS ONE*, 9(6), pp. 1–11. doi:10.1371/journal.pone.0098129.
- 653 Quinlan, A.R. and Hall, I.M. (2010) 'BEDTools: a flexible suite of utilities for comparing genomic
654 features', *Bioinformatics*, 26(6), pp. 841–842. doi:10.1093/bioinformatics/btq033.
- 655 Rambaut, A. *et al.* (2018) 'Posterior summarization in Bayesian phylogenetics using Tracer 1.7',
656 *Systematic Biology*, 67(5), pp. 901–904. doi:10.1093/sysbio/syy032.
- 657 Rasmussen, D.A. and Grünwald, N.J. (2020) 'Phylogeographic approaches to characterize the
658 emergence of plant pathogens', *Phytopathology*, 111(1), pp. 68–77. doi:10.1094/PHYTO-07-20-0319-
659 Fl.
- 660 Rasmussen, S.O. *et al.* (2006) 'A new Greenland ice core chronology for the last glacial termination',
661 *Journal of Geophysical Research*, 111(D6), pp. 1–16. doi:10.1029/2005JD006079.
- 662 Richard, D. *et al.* (2016) 'First report of *Xanthomonas citri* pv. *citri* pathotype A causing Asiatic citrus
663 canker in Martinique, France', *Plant Disease*, 100(9), p. 1946. doi:10.1094/PDIS-02-16-0170-PDN.
- 664 Richard, D. *et al.* (2020) 'Time-calibrated genomic evolution of a monomorphic bacterium during its
665 establishment as an endemic crop pathogen', *Molecular Ecology*, pp. 1–13. doi:10.1111/mec.15770.
- 666 Rieux, A. *et al.* (2014) 'Improved calibration of the human mitochondrial clock using ancient
667 genomes', *Molecular Biology and Evolution*, 31(10), pp. 2780–2792. doi:10.1093/molbev/msu222.

- 668 Rieux, A. and Balloux, F. (2016) 'Inferences from tip-calibrated phylogenies: a review and a practical
669 guide', *Molecular Ecology*, 25(9), pp. 1911–1924. doi:10.1111/mec.13586.
- 670 Rieux, A. and Khatchikian, C.E. (2017) 'TIPDATINGBEAST: an R package to assist the implementation
671 of phylogenetic tip-dating tests using BEAST', *Molecular Ecology Resources*, 17(4), pp. 608–613.
672 doi:10.1111/1755-0998.12603.
- 673 Ristaino, J.B. (2020) 'The importance of mycological and plant herbaria in tracking plant killers',
674 *Frontiers in Ecology and Evolution*, 7(521), pp. 1–11. doi:10.3389/fevo.2019.00521.
- 675 Rosseti, V. (1977) 'Citrus caker in Latin America: a review', *Proceedings of the International Society of*
676 *Citriculture*, 3, pp. 918–924.
- 677 Rybak, M. *et al.* (2009) 'Identification of *Xanthomonas citri* ssp. *citri* host specificity genes in a
678 heterologous expression host', *Molecular Plant Pathology*, 10(2), pp. 249–262. doi:10.1111/j.1364-
679 3703.2008.00528.x.
- 680 Savary, S. *et al.* (2019) 'The global burden of pathogens and pests on major food crops', *Nature*
681 *Ecology & Evolution*, 3, pp. 430–439. doi:10.1038/s41559-018-0793-y.
- 682 Saville, A.C., Martin, M.D. and Ristaino, J.B. (2016) 'Historic late blight outbreaks caused by a
683 widespread dominant lineage of *Phytophthora infestans* (Mont.) de Bary', *PLoS ONE*, 11(12), pp. 1–
684 22. doi:10.1371/journal.pone.0168381.
- 685 Schubert, M., Lindgreen, S. and Orlando, L. (2016) 'AdapterRemoval v2: rapid adapter trimming,
686 identification, and read merging', *BMC Research Notes*, 9(88), pp. 1–7. doi:10.1186/s13104-016-
687 1900-2.
- 688 Schubert, T.S. *et al.* (2001) 'Meeting the challenge of eradicating citrus canker in Florida—again',
689 *Plant Disease*, 85(4), pp. 340–356. doi:10.1094/PDIS.2001.85.4.340.
- 690 da Silva, A.C.R. *et al.* (2002) 'Comparison of the genomes of two *Xanthomonas* pathogens with
691 differing host specificities', *Nature*, 417(6887), pp. 459–463. doi:10.1038/417459a.
- 692 Smith, O. *et al.* (2014) 'A complete ancient RNA genome: identification, reconstruction and
693 evolutionary history of archaeological Barley stripe mosaic virus', *Scientific Reports*, 4(1), pp. 1–6.
694 doi:10.1038/srep04003.
- 695 Stamatakis, A. (2014) 'RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
696 phylogenies', *Bioinformatics*, 30(9), pp. 1312–1313. doi:10.1093/bioinformatics/btu033.
- 697 Staubwasser, M. and Weiss, H. (2006) 'Holocene Climate and Cultural Evolution in Late Prehistoric–
698 Early Historic West Asia', *Quaternary Research*, 66(3), pp. 372–387. doi:10.1016/j.yqres.2006.09.001.
- 699 Stukenbrock, E.H. and McDonald, B.A. (2008) 'The origins of plant pathogens in agro-ecosystems',
700 *Annual Review of Phytopathology*, 46(1), pp. 75–100. doi:10.1146/annurev.phyto.010708.154114.
- 701 Suchard, M.A. and Rambaut, A. (2009) 'Many-core algorithms for statistical phylogenetics',
702 *Bioinformatics*, 25(11), pp. 1370–1376. doi:10.1093/bioinformatics/btp244.
- 703 Sun, X. *et al.* (2004) 'Detection and characterization of a new strain of citrus canker bacteria from
704 key/Mexican lime and alemow in South Florida', *Plant Disease*, 88(11), pp. 1179–1188.
705 doi:10.1094/PDIS.2004.88.11.1179.
- 706 Talon, M., Caruso, M. and Gmitter Jr., F.G. (eds) (2020) *The Genus Citrus*. Duxford: Woodhead

707 Publishing.

708 Tavaré, S. and Miura, R.M. (1986) 'Some probabilistic and statistical problems in the analysis of DNA
709 sequences', in *Some mathematical questions in biology: DNA sequence analysis*. Providence (New
710 York American Mathematical Society), pp. 57–86.

711 *The Xanthomonas Resource* (<http://www.xanthomonas.org/t3e.html>) (2018). Available at:
712 <http://www.xanthomonas.org/t3e.html> (Accessed: 1 June 2020).

713 Vernière, C. *et al.* (1998) 'Characterization of phenotypically distinct strains of *Xanthomonas*
714 *axonopodis* pv. *citri* from Southwest Asia', *European Journal of Plant Pathology*, 104, pp. 477–487.

715 Walker, M. *et al.* (2009) 'Formal definition and dating of the GSSP (Global Stratotype Section and
716 Point) for the base of the Holocene using the Greenland NGRIP ice core, and selected auxiliary
717 records', *Journal of Quaternary Science*, 24(1), pp. 3–17. doi:10.1002/jqs.1227.

718 Weiß, C.L. *et al.* (2016) 'Temporal patterns of damage and decay kinetics of DNA retrieved from
719 plant herbarium specimens', *Royal Society Open Science*, 3(6), p. 160239. doi:10.1098/rsos.160239.

720 Yoshida, K. *et al.* (2013) 'The rise and fall of the *Phytophthora infestans* lineage that triggered the
721 Irish potato famine', *eLife*, 2, pp. 1–25. doi:10.7554/eLife.00731.

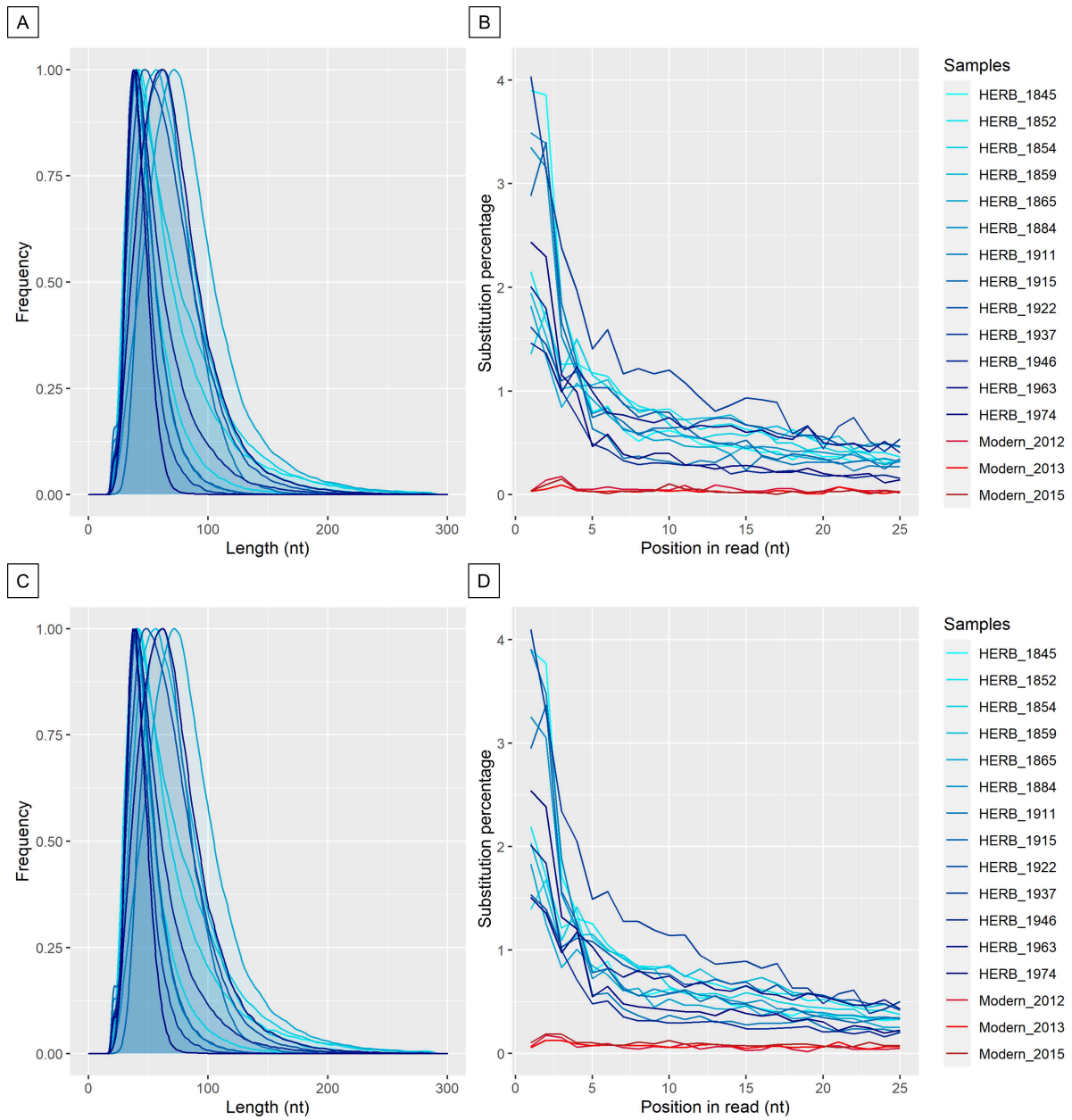
722 Yoshida, K. *et al.* (2014) 'Mining herbaria for plant pathogen genomes: back to the future', *PLoS*
723 *Pathogens*, 10(4), pp. 1–6. doi:10.1371/journal.ppat.1004028.

724 Yoshida, K., Sasaki, E. and Kamoun, S. (2015) 'Computational analyses of ancient pathogen DNA from
725 herbarium samples: challenges and prospects', *Frontiers in Plant Science*, 6, pp. 1–6.
726 doi:10.3389/fpls.2015.00771.

727 Yu, G. *et al.* (2017) '*ggtree*: an *r* package for visualization and annotation of phylogenetic trees with
728 their covariates and other associated data', *Methods in Ecology and Evolution*. Edited by G.
729 McInerney, 8(1), pp. 28–36. doi:10.1111/2041-210X.12628.

730 Zech-Matterne, V. and Fiorentino, G. (eds) (2017) *AGRUMED: Archaeology and history of citrus fruit*
731 *in the Mediterranean: Acclimatization, diversifications, uses*. Publications du Centre Jean Bérard.
732 doi:10.4000/books.pcb.2107.

733 Zhang, Y. *et al.* (2015) 'Positive selection is the main driving force for evolution of citrus canker-
734 causing *Xanthomonas*', *The ISME Journal*, 9(10), pp. 2128–2138. doi:10.1038/ismej.2015.15.
735



737

738 **S1 Fig. Post-mortem DNA damage patterns on *Xci* plasmids pXAC33 and pXAC64.** (A&C) Fragment
 739 length distribution (nucleotides; relative frequency in arbitrary units) of plasmids pXAC33 and pXAC64
 740 respectively. (B&D) Deamination percentage of the first 25 nucleotides from the 5' end, respectively
 741 measured on the 13 historical genomes (blue lines) and on three modern *Xci* strains (red lines) for
 742 plasmids pXAC33 and pXAC64 respectively.

Geographic state

- North America
- Central America & West Indies
- South America
- West Africa
- East Africa
- Northern Indian Ocean islands
- South West Indian Ocean islands
- West Asia
- South Asia 1
- South Asia 2
- South East Asia 1
- South East Asia 2
- East Asia 1
- East Asia 2
- Oceania & Pacific

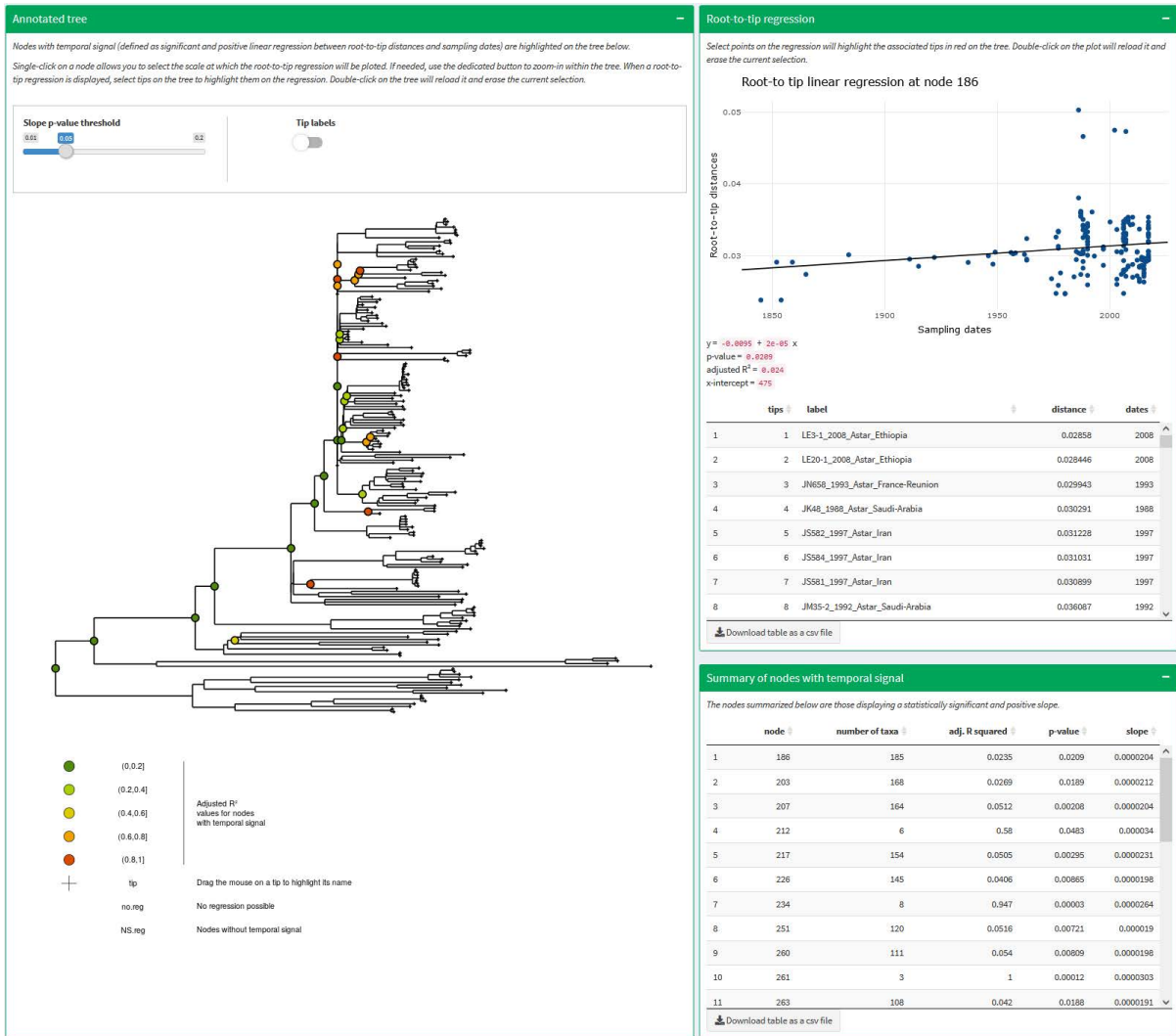
Bootstrap values

- ◇ <75
- ◊ >75
- ◆ >90



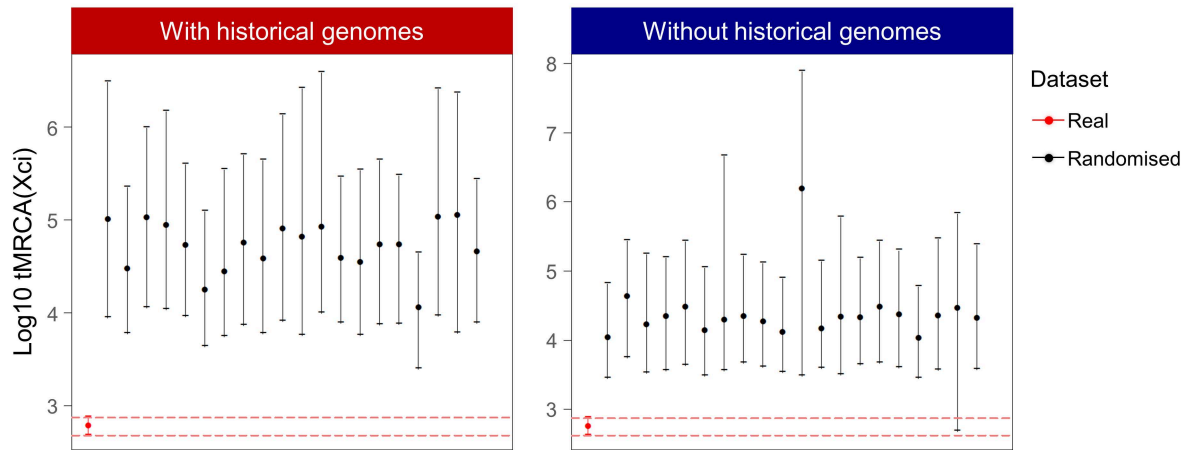
743

744 **S2 Fig. Maximum Likelihood (ML) phylogenetic tree of historical and modern *Xci* genomes.** ML tree
 745 including 13 historical specimens (labelled in red) and 172 modern strains (black) built from 13,007
 746 recombination-free SNPs. *Xanthomonas axonopodis* pv. *vasculorum* NCPPB-796 isolated in 1960 from
 747 Mauritius (GenBank accession number: GCF_013177355.1) was used to root the tree. Node values
 748 correspond to bootstrap values calculated on 1,000 iterations. Tip labels indicate sample reference ID,
 749 date of collection, pathotype and locality; colors differ according to the geographic origin of the
 750 sample. The tree is structured in three major clades corresponding to the three pathotypes A*, Aw
 751 and A (light grey, dark grey and black vertical bars, respectively) with the latter pathotype A clade split
 752 in three lineages (A3, A2 and A1).



754

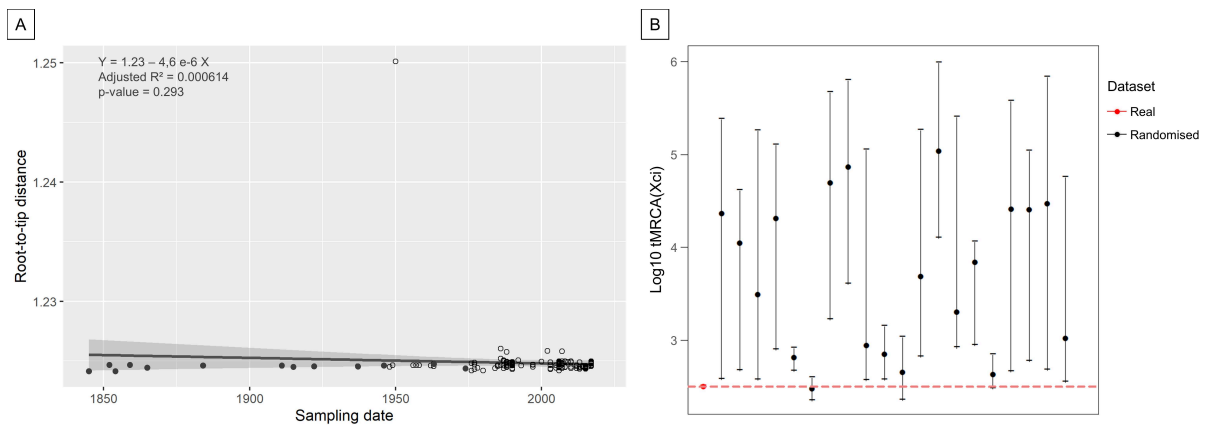
755 **S3 Fig. Root-to-tip regression temporal test visualised on online tool PhyloStemsS.** The tool is
 756 accessible online (<https://pvbmt-apps.cirad.fr/apps/phylostems/>).



757

758 **S4 Fig. Date-randomisation temporal test results with and without historical genomes in the**
 759 **dataset.** Evaluating temporal signal in the dataset by date-randomisation test showed no overlap
 760 between the age of the root estimated from the real dataset (red) vs 20 date-randomised datasets
 761 (black) in the analysis with historical genomes (red, left) whereas there was one in the analysis without
 762 (blue, right). Vertical bars represent 95% Highest Posterior Density intervals.

763



764

765 **S5 Fig. Root-to-tip regression and date-randomisation temporal test results when performed on the**
 766 **dataset including outgroups.** (A) Regression line is plotted in black with historical genomes in black
 767 filled dots and modern ones in black empty dots. Grey areas indicate the confidence interval.
 768 Associated values are the regression equation, adjusted R^2 (Adj R^2) and p-value. (B) Evaluating
 769 temporal signal in the dataset by date-randomisation test showed overlap between the age of the
 770 root estimated from the real dataset (red) vs 20 date-randomised datasets (black). Vertical bars
 771 represent 95% Highest Posterior Density intervals.

772

773 **S1 Table. General characteristics of the historical and modern strains of the study.** ID are composed
774 of the strains ID, the year, the *Xci* pathotype (or species when not *Xci*) and country of isolation.
775 *X.a.vasculorum*: *Xanthomonas axonopodis* pathovar *vasculorum*; *X.c.clitoriae*: *Xanthomonas citri*
776 pathovar *clitoriae*; *X.c.cajanie*: *Xanthomonas citri* pathovar *cajani*. Africa E: East Africa; Africa W: West
777 Africa; America C & W Indies: Central America and West Indies; America N: North America; America
778 S: South America; Asia E1: East Asia 1; Asia E2: East Asia 2; Asia S1: south Asia 1; Asia S2: south Asia 2;
779 Asia S-E1, South east Asia 1; Asia S-E2, South east Asia 2; NIO: North Indian Ocean; SWIO: South West
780 Indian Ocean.

ID	Geographic origin (area)	Isolation Host	GenBank accession ID	Sequence Read Archive accession ID	Publication
HERB_1845_A_Indonesia	Asia S-E2	Citrus x aurantiifolia			
HERB_1852_A_India	Asia S1	Citrus medica			
HERB_1854_A_Indonesia	Asia S-E2	Citrus javanica			
HERB_1859_A_Bangladesh	Asia S1	Citrus medica			
HERB_1865_A_India	Asia S1	Citrus medica			
HERB_1884_A_Philippines	Asia S-E2	Citrus medica			
HERB_1911_A_Indonesia	Asia S-E2	Citrus x aurantiifolia			
HERB_1915_A_Philippines	Asia S-E2	Citrus x aurantiifolia			
HERB_1922_A_China	Asia E1	Citrus medica			
HERB_1937_A_Mauritius	SWIO	Citrus sp.	CP072205-CP072207	SRR12792042	(Campos <i>et al.</i> , 2021)
HERB_1946_A_Guam	Oceania & Pacific	Citrus sp.			
NCPFB-211_1948_A_India	Asia S1	Citrus sp.	JAABAS000000000	SRR11234665	(Richard <i>et al.</i> , 2020)
NCPFB-226_1949_A_New-Zealand	Oceania & Pacific	Citrus sp.	JAABCW000000000	SRR11234664	(Richard <i>et al.</i> , 2020)
CFBP-2525_1956_A_New-Zealand	Oceania & Pacific	Citrus x limon	JAABCU000000000	SRR11234614	(Richard <i>et al.</i> , 2020)
NCPFB-406_1957_A_New-Zealand	Oceania & Pacific	Citrus sp.	JAABCX000000000	SRR11234661	(Richard <i>et al.</i> , 2020)
CFBP-2853_1958_A_New-Zealand	Oceania & Pacific	Citrus sp.	JAABCV000000000	SRR11234592	(Richard <i>et al.</i> , 2020)
CFBP-2855_1962_A_Japan	Asia E2	Citrus sp.	JAABAW000000000	SRR11234581	(Richard <i>et al.</i> , 2020)
CFBP-1209_1963_A_China-Hong-Kong	Asia E1	Citrus maxima	JAABAP000000000	SRR11234741	(Richard <i>et al.</i> , 2020)
HERB_1963_A_Nepal	Asia S1	Citrus medica			
NCPFB-1471_1963_A_China-Hong-Kong	Asia E1	Citrus x paradisi			
HERB_1974_A_Mauritius	SWIO	Citrus lime			
CFBP-2857_1976_A_Fiji	Oceania & Pacific	Citrus x aurantiifolia	JAABAO000000000	SRR11234570	(Richard <i>et al.</i> , 2020)
CFBP-2865_1976_A_Brazil	America S	Citrus x aurantiifolia	JAAAYX000000000	SRR11234559	(Richard <i>et al.</i> , 2020)
JJ053-04_1977_A_Taiwan	Asia E1	Poncirus trifoliata			
JJ053-06_1977_A_Taiwan	Asia E1	Citrus sunki	JAABJE000000000	SRR11234718	(Richard <i>et al.</i> , 2020)
JJ053-07_1977_A_Taiwan	Asia E1	Citrus x limonia			
JJ053-09_1977_A_Taiwan	Asia E1	Citrus x aurantiifolia			

JJ155_1977_A_Argentina	America S	Citrus x aurantiifolia	JAAAYI000000000	SRR11234707	(Richard <i>et al.</i> , 2020)
JJ165_1978_A_India	Asia S1	Citrus maxima	JAABAR000000000	SRR11234696	(Richard <i>et al.</i> , 2020)
CFBP-2908_1980_A_Brazil	America S	Citrus reticulata	JAAAYY000000000	SRR11234851	(Richard <i>et al.</i> , 2020)
FDC-1083_1980_A_Brazil	America S	Citrus reticulata	CCVZ00000000.1	SRR11234818	(Gordon <i>et al.</i> , 2015)
JJ009-1_1984_A_Mauritius	SWIO	Citrus sinensis x Poncirus trifoliata	JAABBU000000000	SRR11234763	(Richard <i>et al.</i> , 2020)
JJ009-04_1985_A_Mauritius	SWIO	Citrus sp. x bergamia	JAABBT000000000	SRR11234774	(Richard <i>et al.</i> , 2020)
JK4-1_1985_A_China	Asia E1	Citrus sp.	CDMR01000000	SRR11234656	(Gordon <i>et al.</i> , 2015)
LMG-9322_1986_A_USA	America N	Citrus x aurantiifolia	JPYD00000000.1	SRR11234695	(Gordon <i>et al.</i> , 2015)
JJ009-8_1987_A_Mauritius	SWIO	Citrus x sinensis	JAABBV000000000	SRR11234752	(Richard <i>et al.</i> , 2020)
JJ238-04_1987_A_Maldives	NIO	Citrus x aurantiifolia	JAABAX000000000	SRR11234685	(Richard <i>et al.</i> , 2020)
JJ238-06_1987_A_Maldives	NIO	Citrus x aurantiifolia	JAABAY000000000	SRR11234674	(Richard <i>et al.</i> , 2020)
JJ238-07_1987_A_Maldives	NIO	Citrus x aurantiifolia	JAABAZ000000000	SRR11234663	(Richard <i>et al.</i> , 2020)
JJ238-08_1987_A_Maldives	NIO	Citrus x aurantiifolia	JAABBA000000000	SRR11234652	(Richard <i>et al.</i> , 2020)
JJ238-09_1987_A_Maldives	NIO	Citrus x aurantiifolia	JAABBB000000000	SRR11234641	(Richard <i>et al.</i> , 2020)
JJ238-10_1987_A_Maldives	NIO	Citrus x aurantiifolia	CCWC00000000.1	SRR11234629	(Gordon <i>et al.</i> , 2015)
JJ238-11_1987_A_Maldives	NIO	Citrus x aurantiifolia	JAABBD000000000	SRR11234623	(Richard <i>et al.</i> , 2020)
JK004-04_1987_A_Maldives	NIO	Citrus x aurantiifolia	JAABBE000000000	SRR11234621	(Richard <i>et al.</i> , 2020)
C20_1988_A_France-Reunion	SWIO	Citrus x paradisi	JAABHN000000000	SRR11234852	(Richard <i>et al.</i> , 2020)
JH081-04_1988_A_China	Asia E1	Citrus sp.			
JH410-04_1988_A_China	Asia E1	Citrus reticulata			
JJ238-14_1988_A_Philippines	Asia S-E2	Citrus sp.			
JJ238-16_1988_A_Pakistan	Asia S2	Citrus x sinensis	JAABCY000000000	SRR11234622	(Richard <i>et al.</i> , 2020)
NCPPB-3562_1988_A_India	Asia S1	Citrus limon	CCXZ01000000	SRR11234662	(Gordon <i>et al.</i> , 2015)
NCPPB-3610_1988_A_India	Asia S1	Poncirus trifoliata	CDAO00000000.1	SRR11234642	(Gordon <i>et al.</i> , 2015)
NCPPB-3612_1988_A_India	Asia S1	Citrus x aurantiifolia	CDAQ00000000.1	SRR11234640	(Gordon <i>et al.</i> , 2015)
JJ037-01_1989_A_Malaysia	Asia S-E1	Citrus x sinensis			
JK101-1_1990_A_Argentina	America S	Citrus x paradisi	JAAAYJ000000000	SRR11234620	(Richard <i>et al.</i> , 2020)
JK146-01_1990_A_Malaysia	Asia S-E1	Citrus maxima			
JK146-02_1990_A_Malaysia	Asia S-E1	Citrus reticulata			
JK146-03_1990_A_Malaysia	Asia S-E1	Citrus reticulata			
JK146-05_1990_A_Malaysia	Asia S-E1	Citrus maxima			
JK148-01_1990_A_Philippines	Asia S-E2	Citrus x aurantiifolia			
JK148-04_1990_A_Philippines	Asia S-E2	Citrus x limon			
JK148-07_1990_A_Philippines	Asia S-E2	x Citrofortunella microcarpa			

JK148-11_1990_A_Philippines	Asia S-E2	Citrus maxima			
JK148-13_1990_A_Philippines	Asia S-E2	Citrus maxima			
JK161-4_1990_A_Mauritius	SWIO	Citrus x sinensis	JAABBW000000000	SRR11234619	(Richard <i>et al.</i> , 2020)
JS538-2_1997_A_France-Reunion	SWIO	Citrus x paradisi	JAABHE000000000	SRR11234598	(Richard <i>et al.</i> , 2020)
JW160-1_2000_A_Bangladesh	Asia S1	Citrus x aurantiifolia	CCWH000000000		(Gordon <i>et al.</i> , 2015)
FDC-217_2003_A_Brazil	America S	Citrus sinensis	CCWY000000000.1	SRR11234807	(Gordon <i>et al.</i> , 2015)
JZ092_2003_A_Seychelles	NIO	Citrus x limon	JAABIQ000000000	SRR11234595	(Richard <i>et al.</i> , 2020)
JZ094-4_2003_A_Seychelles	NIO	Citrus x limon	JAABIR000000000	SRR11234594	(Richard <i>et al.</i> , 2020)
LM180_2003_A_Argentina	America S	Citrus x paradisi	JAAAYM000000000	SRR11234708	(Richard <i>et al.</i> , 2020)
LB100-1_2005_A_Seychelles	NIO	Citrus sinensis x Poncirus trifoliata	CDAV010000000	SRR11234654	(Gordon <i>et al.</i> , 2015)
LB100-3_2005_A_Seychelles	NIO	Citrus sinensis x Poncirus trifoliata	JAABIT000000000	SRR11234591	(Richard <i>et al.</i> , 2020)
LM169_2005_A_Argentina	America S	Citrus x paradisi	JAAAYL000000000	SRR11234709	(Richard <i>et al.</i> , 2020)
LC004-1_2006_A_Vietnam	Asia S-E1	Citrus maxima	JAABJG000000000	SRR11234590	(Richard <i>et al.</i> , 2020)
LC008-04_2006_A_Vietnam	Asia S-E1	Citrus reticulata			
LC014-10_2006_A_Vietnam	Asia S-E1	Citrus reticulata			
LC015-01_2006_A_Vietnam	Asia S-E1	Citrus sp.			
LC052-05_2006_A_Vietnam	Asia S-E1	Citrus x sinensis			
LC060-08_2006_A_Vietnam	Asia S-E1	Citrus x sinensis			
LC067-02_2006_A_Vietnam	Asia S-E1	Citrus x sinensis			
LC067-04_2006_A_Vietnam	Asia S-E1	Citrus maxima			
LC80_2006_A_Mali	Africa W	Citrus reticulata x Citrus sinensis	CCWJ000000000.1	SRR11234586	(Gordon <i>et al.</i> , 2015)
LG099_2006_A_Bangladesh	Asia S1	Citrus x aurantiifolia			
LG102_2006_A_Bangladesh	Asia S1	Citrus sp.	CDAN010000000	SRR11234649	(Gordon <i>et al.</i> , 2015)
LG103_2006_A_Bangladesh	Asia S1	Citrus x aurantiifolia			
LG104_2006_A_Bangladesh	Asia S1	Citrus x aurantiifolia			
LG105_2006_A_Bangladesh	Asia S1	Citrus x aurantiifolia			
LG107_2006_A_Bangladesh	Asia S1	Citrus x aurantiifolia			
LG108_2006_A_Bangladesh	Asia S1	Citrus x aurantiifolia			
LG110_2006_A_Bangladesh	Asia S1	Citrus x aurantiifolia			
LG112_2006_A_Bangladesh	Asia S1	Citrus x aurantiifolia			
LG113_2006_A_Bangladesh	Asia S1	Citrus x aurantiifolia			
LG97_2006_A_Bangladesh	Asia S1	Citrus sp.	CDAK010000000	SRR11234647	(Gordon <i>et al.</i> , 2015)
LG98_2006_A_Bangladesh	Asia S1	Citrus x aurantiifolia	CDBA010000000	SRR11234646	(Gordon <i>et al.</i> , 2015)
LD072-01_2007_A_Myanmar	Asia S-E1	Citrus maxima			

LD072-02_2007_A_Myanmar	Asia S-E1	Citrus x sinensis			
LD107-01_2007_A_Indonesia	Asia S-E2	Citrus maxima			
LD107-02_2007_A_Indonesia	Asia S-E2	Citrus maxima			
LD108-01_2007_A_Indonesia	Asia S-E2	Citrus x aurantiifolia			
LD108-02_2007_A_Indonesia	Asia S-E2	Citrus x aurantiifolia			
LM128_2007_A_Argentina	America S	Citrus reticulata			
LD7-1_2008_A_Mali	Africa W	Citrus x aurantiifolia	CDAL00000000.1	SRR11234651	(Gordon <i>et al.</i> , 2015)
LE085_2008_A_Papua-New-Guinea	Oceania & Pacific	Citrus x limon			
LE086_2008_A_Papua-New-Guinea	Oceania & Pacific	Citrus x limon			
LE087_2008_A_Papua-New-Guinea	Oceania & Pacific	Citrus x aurantiifolia			
LE116-1_2008_A_Mali	Africa W	Citrus x aurantiifolia	CDHD01000000	SRR11234650	(Gordon <i>et al.</i> , 2015)
LG117_2009_A_Bangladesh	Asia S1	Citrus sp.	CDAX01000000	SRR11234648	(Gordon <i>et al.</i> , 2015)
LH001-03_2010_A_Pakistan	Asia S2	Citrus sp.			
LH221-6_2010_A_France-Reunion	SWIO	Citrus hystrix	JAABEL00000000	SRR11234582	(Richard <i>et al.</i> , 2020)
LH37-1_2010_A_Senegal	Africa W	Citrus paradisi	CDAS00000000.1	SRR11234645	(Gordon <i>et al.</i> , 2015)
LM205_2010_A_Argentina	America S	Citrus x paradisi	JAAAYQ00000000	SRR11234703	(Richard <i>et al.</i> , 2020)
LJ001-1_2012_A_Seychelles	NIO	Citrus sp.	JAABIU00000000	SRR11234557	(Richard <i>et al.</i> , 2020)
LJ003-1_2012_A_Seychelles	NIO	Citrus x sinensis	JAABIV00000000	SRR11234556	(Richard <i>et al.</i> , 2020)
LJ226-02_2012_A_France-Mayotte	SWIO	Citrus x sinensis	JAABCN00000000	SRR11234551	(Richard <i>et al.</i> , 2020)
LM229_2012_A_Argentina	America S	Citrus x sinensis	JAAAYR00000000	SRR11234702	(Richard <i>et al.</i> , 2020)
LK169-01_2013_A_France-Reunion	SWIO	Citrus hystrix	JAABFW00000000	SRR11234810	(Richard <i>et al.</i> , 2020)
LK169-04_2013_A_France-Reunion	SWIO	Citrus hystrix	JAABFY00000000	SRR11234808	(Richard <i>et al.</i> , 2020)
LM158_2013_A_Argentina	America S	Citrus x sinensis	JAAAYK00000000	SRR11234710	(Richard <i>et al.</i> , 2020)
LM184_2013_A_Argentina	America S	Citrus x limon	JAAAYN00000000	SRR11234706	(Richard <i>et al.</i> , 2020)
LL074-04_2014_A_France-Martinique	America C & W Indies	Citrus x paradisi	JAABBI00000000	SRR11234793	(Richard <i>et al.</i> , 2020)
LL096-13_2014_A_France-Reunion	SWIO	Citrus x limon	JAABHJ00000000	SRR11234772	(Richard <i>et al.</i> , 2020)
LL111-06_2014_A_France-Martinique	America C & W Indies	Citrus x sinensis	JAABBJ00000000	SRR11234768	(Richard <i>et al.</i> , 2020)
LL124-01_2014_A_France-Martinique	America C & W Indies	Citrus x sinensis	JAABBK00000000	SRR11234764	(Richard <i>et al.</i> , 2020)
LL186-5_2014_A_France-Reunion	SWIO	Citrus x sinensis	JAABHV00000000	SRR11234753	(Richard <i>et al.</i> , 2020)
LM053-06_2015_A_Mauritius	SWIO	Citrus x aurantiifolia	JAABCA00000000	SRR11234751	(Richard <i>et al.</i> , 2020)
LM054-06_2015_A_Mauritius	SWIO	Citrus x sinensis	JAABCC00000000	SRR11234749	(Richard <i>et al.</i> , 2020)
LM054-17_2015_A_Mauritius	SWIO	Citrus x sinensis	JAABCD00000000	SRR11234748	(Richard <i>et al.</i> , 2020)
LM055-08_2015_A_Mauritius	SWIO	Citrus sp.	JAABCE00000000	SRR11234747	(Richard <i>et al.</i> , 2020)

					<i>al.</i> , 2020)
LM069-01_2015_A_Mauritius	SWIO	Citrus x aurantiifolia	JAABCI000000000	SRR11234743	(Richard <i>et al.</i> , 2020)
LM090-02_2015_A_France-Martinique	America C & W Indies	Citrus sp.	JAABBL000000000	SRR11234726	(Richard <i>et al.</i> , 2020)
LM097-01_2015_A_Mauritius-Rodrigues	SWIO	Citrus x aurantiifolia	JAABIN000000000	SRR11234716	(Richard <i>et al.</i> , 2020)
LM198_2015_A_Argentina	America S	Citrus x sinensis	JAAAYO000000000	SRR11234705	(Richard <i>et al.</i> , 2020)
LM199_2015_A_Argentina	America S	Citrus x sinensis	JAAAYP000000000	SRR11234704	(Richard <i>et al.</i> , 2020)
LM358-08_2015_A_France-Reunion	SWIO	Citrus x limon	JAABFN000000000	SRR11234700	(Richard <i>et al.</i> , 2020)
LM358-18_2015_A_France-Reunion	SWIO	Citrus x limon	JAABFP000000000	SRR11234698	(Richard <i>et al.</i> , 2020)
LN006-18_2016_A_France-Reunion	SWIO	Citrus reticulata x sinensis	JAABDQ000000000	SRR11234689	(Richard <i>et al.</i> , 2020)
LP027-03_2017_A_Seychelles	NIO	Citrus x aurantiifolia	JAABIW000000000	SRR11234681	(Richard <i>et al.</i> , 2020)
LP027-05_2017_A_Seychelles	NIO	Citrus x aurantiifolia	JAABIX000000000	SRR11234680	(Richard <i>et al.</i> , 2020)
LP027-13_2017_A_Seychelles	NIO	Citrus x aurantiifolia	JAABIY000000000	SRR11234679	(Richard <i>et al.</i> , 2020)
LP028-03_2017_A_Seychelles	NIO	Citrus x aurantiifolia	JAABIZ000000000	SRR11234678	(Richard <i>et al.</i> , 2020)
LP028-05_2017_A_Seychelles	NIO	Citrus x aurantiifolia	JAABJA000000000	SRR11234677	(Richard <i>et al.</i> , 2020)
LP028-06_2017_A_Seychelles	NIO	Citrus x aurantiifolia	JAABJB000000000	SRR11234676	(Richard <i>et al.</i> , 2020)
LP029-13_2017_A_Seychelles	NIO	Citrus x sinensis	JAABJC000000000	SRR11234675	(Richard <i>et al.</i> , 2020)
LP029-15_2017_A_Seychelles	NIO	Citrus x sinensis	JAABJD000000000	SRR11234673	(Richard <i>et al.</i> , 2020)
LP030-1_2017_A_France-Martinique	America C & W Indies	Citrus x aurantiifolia	JAABBM000000000	SRR11234672	(Richard <i>et al.</i> , 2020)
LP031-1_2017_A_France-Martinique	America C & W Indies	Citrus x sinensis	JAABBN000000000	SRR11234671	(Richard <i>et al.</i> , 2020)
LP032-1_2017_A_France-Martinique	America C & W Indies	Citrus x aurantiifolia	JAABBO000000000	SRR11234670	(Richard <i>et al.</i> , 2020)
LP033-1_2017_A_France-Martinique	America C & W Indies	Citrus x paradisi	JAABBP000000000	SRR11234669	(Richard <i>et al.</i> , 2020)
LP034-1_2017_A_France-Martinique	America C & W Indies	Citrus x aurantiifolia	JAABBQ000000000	SRR11234668	(Richard <i>et al.</i> , 2020)
LP035-1_2017_A_France-Martinique	America C & W Indies	Citrus x aurantiifolia	JAABBR000000000	SRR11234667	(Richard <i>et al.</i> , 2020)
LP036-1_2017_A_France-Martinique	America C & W Indies	Citrus x aurantiifolia	JAABBS000000000	SRR11234666	(Richard <i>et al.</i> , 2020)
LP162_2017_A_Pakistan	Asia S2	Citrus reticulata			
LP187_2017_A_Pakistan	Asia S2	Citrus x sinensis			
LP188_2017_A_Pakistan	Asia S2	Citrus x sinensis			
LP201_2017_A_Pakistan	Asia S2	Citrus sp.			
LP215-01_2017_A_Pakistan	Asia S2	Citrus reticulata			
LP220-02_2017_A_Pakistan	Asia S2	Citrus x sinensis			
LP229-03_2017_A_Pakistan	Asia S2	Citrus reticulata			
LP234-03_2017_A_Pakistan	Asia S2	Citrus reticulata			
LP235-06_2017_A_Pakistan	Asia S2	Citrus x sinensis			
LP293-01_2017_A_Pakistan	Asia S2	Citrus x sinensis			
LP293-05_2017_A_Pakistan	Asia S2	Citrus x sinensis			

LP295-05_2017_A_Pakistan	Asia S2	Citrus x sinensis			
LP297-02_2017_A_Pakistan	Asia S2	Citrus x sinensis			
CFBP-2858_1977_Astar_Fiji	Oceania & Pacific	Citrus x aurantiifolia			
JF90-2_1986_Astar_Oman	Asia W	Citrus x aurantiifolia	CCWA00000000		(Gordon <i>et al.</i> , 2015)
JK48_1988_Astar_Saudi-Arabia	Asia W	Citrus x aurantiifolia	CDAJ00000000		(Gordon <i>et al.</i> , 2015)
NCPPB-3607_1988_Astar_India	Asia S1	Citrus x aurantiifolia	CDAT00000000		(Gordon <i>et al.</i> , 2015)
JJ238-24_1989_Astar_Thailand	Asia S-E1	Citrus x aurantiifolia	CCVX00000000		(Gordon <i>et al.</i> , 2015)
NCPPB-3615_1989_Astar_India	Asia S1	Citrus x aurantiifolia	CDAM00000000		(Gordon <i>et al.</i> , 2015)
JK143-11_1990_Astar_Thailand	Asia S-E1	Citrus sp.	CDMO00000000		(Gordon <i>et al.</i> , 2015)
JK143-9_1990_Astar_Thailand	Asia S-E1	Citrus sp.	CDMQ00000000		(Gordon <i>et al.</i> , 2015)
JM35-2_1992_Astar_Saudi-Arabia	Asia W	Citrus x aurantiifolia	CDMS00000000		(Gordon <i>et al.</i> , 2015)
JN658_1993_Astar_France-Reunion	SWIO	Citrus x aurantiifolia			
JS581_1997_Astar_Iran	Asia S2	Citrus limetta	CDAW00000000		(Gordon <i>et al.</i> , 2015)
JS582_1997_Astar_Iran	Asia S2	Citrus x latifolia	CDAP00000000		(Gordon <i>et al.</i> , 2015)
JS584_1997_Astar_Iran	Asia S2	Citrus sp.	CCWF00000000		(Gordon <i>et al.</i> , 2015)
LD71a_2007_Astar_Cambodia	Asia S-E1	Citrus sp.	CCWE00000000		(Gordon <i>et al.</i> , 2015)
LE20-1_2008_Astar_Ethiopia	Africa E	Citrus x aurantiifolia	CCWK00000000		(Gordon <i>et al.</i> , 2015)
LE3-1_2008_Astar_Ethiopia	Africa E	Citrus x aurantiifolia	CDAI00000000		(Gordon <i>et al.</i> , 2015)
LK135-01_2013_Astar_Comoros-Moheli	SWIO	Citrus sp.			
JF90-8_1986_Aw_Oman	Asia W	Citrus x aurantiifolia	CCWB00000000		(Gordon <i>et al.</i> , 2015)
NCPPB-3608_1988_Aw_India	Asia S1	Citrus x aurantiifolia	CCWG00000000		(Gordon <i>et al.</i> , 2015)
LB302_2002_Aw_USA	America N	Citrus x aurantiifolia	CDAU00000000		(Gordon <i>et al.</i> , 2015)
LG115_2007_Aw_India	Asia S1	Citrus sp.	CDAY00000000		(Gordon <i>et al.</i> , 2015)
NCPPB796_1960_X.a.vasculorum_Mauritius	SWIO	Saccharum sp.	SAMN14772551		
LMG558_1950_X.c.cajani_India	Asia S1	Cajanus cajan	LOKQ00000000		(Patané <i>et al.</i> , 2019)
LMG9045_1974_X.c.clitoriae_India	Asia S1	Clitoria sp.	LOKA00000000		(Patané <i>et al.</i> , 2019)

781

782

783 **S2 Table. Summary of mapping, depth, coverage and damage statistics for the 13 historic *Xci***
 784 **plasmids pXAC33 and pXAC64. SD: standard deviation; nt: nucleotides.**

ID	Protocol	Endogenous <i>Xci</i> DNA (%)		Mean depth		Coverage at 1X (%)		Insert length (mean \pm SD in nt)		Deamination rate at terminal position (%)	
		pXAC33	pXAC64	pXAC33	pXAC64	pXAC33	pXAC64	pXAC33	pXAC64	pXAC33	pXAC64
HERB_1 845	TruSeq Nano	0.36	0.48	122.2	120.4	93.3	95.7	52.2 \pm 22.8	51.9 \pm 22.6	3.91	3.90
HERB_1 884	TruSeq Nano	0.35	0.54	92.1	94.6	89.8	93.2	48.4 \pm 17.0	48.4 \pm 16.9	3.38	3.28
HERB_1 911	TruSeq Nano	0.06	0.08	69.2	65.2	90.2	90.2	69.7 \pm 21.9	69.4 \pm 21.8	3.42	3.78
HERB_1 915	TruSeq Nano	0.16	0.24	66.2	65.8	88	92.9	48.5 \pm 17.1	48.7 \pm 17.1	2.93	2.98
HERB_1 937	TruSeq Nano	0.03	0.04	22.6	17.9	82.9	88.5	44.7 \pm 13.8	44.6 \pm 13.6	3.99	4.09
HERB_1 946	TruSeq Nano	0.23	0.32	114.9	108.1	94.7	93.2	58.3 \pm 28.1	58.4 \pm 28.2	1.98	2.01
HERB_1 974	TruSeq Nano	0.12	0.14	42.6	25.5	82.5	49.7	41.5 \pm 9.5	41.6 \pm 9.5	2.46	2.56
HERB_1 852	BEST	0.17	0.23	130.3	117.9	95.9	96.5	70.4 \pm 41.4	69.7 \pm 41.1	2.13	2.17
HERB_1 854	BEST	0.39	0.54	114	108.6	95.2	97.1	72.0 \pm 39.6	72.3 \pm 39.7	1.35	1.40
HERB_1 859	BEST	0.16	0.22	84.1	75.4	94.2	95.3	71.7 \pm 31.6	71.4 \pm 31.4	1.93	2.03
HERB_1 865	BEST	0.13	0.20	95.1	101.5	77.9	96.6	85.1 \pm 36.8	85.1 \pm 36.7	1.81	1.81
HERB_1 922	BEST	0.60	0.99	99.6	128.7	76.5	94.5	67.7 \pm 29.2	67.2 \pm 28.5	1.62	1.54
HERB_1 963	BEST	0.55	0.67	106.8	86.2	94.3	96	73.2 \pm 31.0	73.0 \pm 30.9	1.46	1.53

785

786

787 **S3 Table. Recombining regions among 185 historical or modern *Xci* strains.**

Starting position	Ending position	Length
3,095,443	3,119,514	24,071
3,799,050	3,805,347	6,297
4,257,584	4,630,760	373,176
4,959,140	4,988,729	29,589

788

789 **S4 Table. List of 146 pathogeny-associated genes investigated among 185 historical or modern *Xci***
790 **strains. *Xci*: *Xanthomonas citri* pv. *citri*; *Xee*: *Xanthomonas campestris* pv. *vesicatoria*; *Xcc*:**
791 ***Xanthomonas campestris* pv. *campestris*; *Xoc*: *Xanthomonas oryzae* pv. *oryzicola*; *Xoo*: *Xanthomonas***
792 ***oryzae* pv. *oryzae*; T3E: Type III effector.**

Gene family	CDS	Function	Organism strain
avrBs2	XAC0076	Glycerophosphoryl diester phosphodiesterase	<i>Xci</i> IAPAR 306
avrBs3	XACb0065 (<i>PthA4</i>)	AvrBs3/PthA-type transcription activator; nuclear localisation	<i>Xci</i> IAPAR 306
clpA	XAC_RS10175	ATP-dependent Clp protease ATP-binding subunit ClpA	<i>Xci</i> IAPAR 306
dksA	XAC_RS23635	RNA polymerase-binding protein DksA	<i>Xci</i> IAPAR 306
exbD1	XAC_RS00055	Biopolymer transporter ExbD	<i>Xci</i> IAPAR 306
exsF	XAC_RS15890	Response regulator	<i>Xci</i> IAPAR 306
exsG	XAC_RS15895	PAS domain S-box protein	<i>Xci</i> IAPAR 306
galU	XAC_RS11675	UTP—glucose-1-phosphate uridylyltransferase GalU	<i>Xci</i> IAPAR 306
gltB	XAC_RS00175	Glutamate syntase large subunit	<i>Xci</i> IAPAR 306
gumF	XAC_RS13145	Acytransferase	<i>Xci</i> IAPAR 306
gumK	XAC_RS13120	Glycosyltransferase	<i>Xci</i> IAPAR 306
helD	XAC_RS11510	Helicase IV	<i>Xci</i> IAPAR 306
hemB	XAC_RS20350	Porphobilinogen synthase	<i>Xci</i> IAPAR 306
hpaA	XAC0400	Type III secretion control protein, maybe not a T3E	<i>Xci</i> IAPAR 306
hpaB	XAC0396	Type III secretion system chaperone	<i>Xci</i> IAPAR 306
hpaC	XAC0404	Type III secretion system export control protein	<i>Xci</i> IAPAR 306
hpaH	XAC0417	Type III secretion system putative transglycosylase HpaH	<i>Xci</i> IAPAR 306
hpaI	XANAC_0475	Type III effector HpaI protein (fragment)	<i>Xci</i> IAPAR 306
hrcC	XAC0415	Type III secretion system outer membrane pore protein	<i>Xci</i> IAPAR 306
hrcD	XAC0399	Type III secretion system protein	<i>Xci</i> IAPAR 306
hrcJ	XAC0409	Type III secretion bridge between inner and outer membrane lipoprotein	<i>Xci</i> IAPAR 306
hrcL	XAC0411	Type III secretion system cytoplasmic protein	<i>Xci</i> IAPAR 306
hrcN	XAC_RS02160	EscN/YscN/HrcN family type III secretion system ATPase	<i>Xci</i> IAPAR 306
hrcN	XAC0412	Type III secretion system ATP synthase	<i>Xci</i> IAPAR 306
hrcQ	XAC0403	Type III secretion system apparatus protein	<i>Xci</i> IAPAR 306
hrcR	XAC0402	Type III secretion system inner membrane protein	<i>Xci</i> IAPAR 306
hrcS	XAC0401	Type III secretion system inner membrane protein	<i>Xci</i> IAPAR 306
hrcT	XAC0414	Type III secretion system inner membrane protein	<i>Xci</i> IAPAR 306
hrcU	XAC0406	Type III secretion system inner membrane protein	<i>Xci</i> IAPAR 306
hrcV	XAC0405	Type III secretion system inner membrane channel protein	<i>Xci</i> IAPAR 306
hrpB1	XAC0407	Type III secretion system protein	<i>Xci</i> IAPAR 306
hrpB2	XAC0408	Type III secretion system protein	<i>Xci</i> IAPAR 306
hrpB4	XAC0410	Type III secretion system protein	<i>Xci</i> IAPAR 306
hrpB7	XAC0413	Type III secretion system protein	<i>Xci</i> IAPAR 306
hrpD6	XAC0398	Type III secretion system regulator	<i>Xci</i> IAPAR 306
hrpE	XAC0397	Type III secretion system pilin	<i>Xci</i> IAPAR 306
hrpF	XAC0394	Type III secretion system translocator protein	<i>Xci</i> IAPAR 306
hrpG	XAC1265	Type III secretion system OmpR-type response regulator	<i>Xci</i> IAPAR 306

hrpM	XAC_RS03215	Glucans biosynthesis glucosyltransferase MdoH	Xci IAPAR 306
hrpW	XAC2922	Pectate lyase, maybe not a T3E	Xci IAPAR 306
hrpX	XAC1266	Type III secretion system transcriptional activator	Xci IAPAR 306
htrA	XAC_RS20055	Do family serine endopeptidase	Xci IAPAR 306
ispF	XAC_RS08775	2-C-methyl-D-erythritol 2 4-cyclodiphosphate synthase	Xci IAPAR 306
leuC	XAC_RS17505	3-isopropylmalate dehydratase large subunit	Xci IAPAR 306
nadD	XAC_RS14105	Nicotinate-nucleotide adenyltransferase	Xci IAPAR 306
ostA	XAC_RS16280	Alpha alpha-trehalose-phosphate synthase (UDP-forming)	Xci IAPAR 306
peh-1	XAC_RS03430	Endopolygalacturonase	Xci IAPAR 306
pgi	XAC_RS09105	Glucose-6-phosphate isomerase	Xci IAPAR 306
pstB	XAC_RS08000	Phosphate ABC transporter ATP-binding protein PstB	Xci IAPAR 306
tatB	XAC_RS21270	Twin-arginine translocase subunit TatB	Xci IAPAR 306
tatC	XAC_RS21265	Twin-arginine translocase subunit TatC	Xci IAPAR 306
trpC	XAC_RS02505	Indole-3-glycerol phosphate synthase TrpC	Xci IAPAR 306
xopA	XAC0416	Harpin, maybe not a T3E	Xci IAPAR 306
xopAE	XAC0393	LRR protein	Xci IAPAR 306
xopAG	XAC3608v2_110003	Unknown	Xci IAPAR 306
xopAI	XAC3230	Putative ADP-ribosyltransferase	Xci IAPAR 306
xopAK	XAC3666	Unknown	Xci IAPAR 306
xopAP	XAC2990	Unknown	Xci IAPAR 306
xopAQ	XAC40v3_800003	Unknown	Xci C40
xopAU	XAC1171	Serine/threonine kinase	Xci IAPAR 306
xopAV	XAC1172	Unknown	Xci IAPAR 306
xopAW	XAC2949	Calcium-binding protein	Xci IAPAR 306
xopAY	XAC1172	Unknown	Xci IAPAR 306
xopAZ	XAC1358	SlpA superfamily, FKBP-type peptidyl-prolyl cis-trans isomerase	Xci IAPAR 306
xopC2	XAC1210_ψ; XAC1209_ψ	Haloacid dehalogenase-like hydrolase	Xci IAPAR 306
xopE1	XAC0286	Putative transglutaminase	Xci IAPAR 306
xopE2	XACb0011	Putative transglutaminase, plasmidic	Xci IAPAR 306
xopE3	XAC3224	Putative transglutaminase	Xci IAPAR 306
xopF1	XANAC_0476_ψ	Unknown	Xci IAPAR 306
xopF1	XANAC_0477_ψ	Unknown	Xci IAPAR 306
xopF2	XAC2785_ψ	Unknown	Xci IAPAR 306
xopI1	XAC0754	F-box protein	Xci IAPAR 306
xopK	XAC3085	Unknown	Xci IAPAR 306
xopL	XAC3090	LRR protein	Xci IAPAR 306
xopM	XAC0418	Unknown	Xci IAPAR 306
xopN	XAC2786	ARM/HEAT repeat	Xci IAPAR 306
xopP	XAC1208	Unknown	Xci IAPAR 306
xopQ	XAC4333	Putative inosine-uridine nucleoside N-ribohydrolase	Xci IAPAR 306
xopR	XAC0277	Unknown	Xci IAPAR 306
xopS	XAC0315	Unknown	Xci IAPAR 306

xopT	XANAC_p10046	Unknown	Xci IAPAR 306
xopV	XAC0601	Unknown	Xci IAPAR 306
xopX	XAC0543	Unknown	Xci IAPAR 306
xopZ1	XAC2009	Unknown	Xci IAPAR 306
xpsD	XAC_RS17865	Type II secretion system protein GspD	Xci IAPAR 306
xpsE	XAC_RS17915	Type II secretion system protein GspE	Xci IAPAR 306
xpsF	XAC_RS17910	Type II secretion system F family protein	Xci IAPAR 306
xpsG	XAC_RS17905	Type II secretion system protein GspG	Xci IAPAR 306
xpsM	XAC_RS17875	General secretion pathway protein GspM	Xci IAPAR 306
xpsN	XAC_RS17870	Hypothetical protein	Xci IAPAR 306
yapH	XAC_RS20725	Filamentous hemagglutinin N-terminal domain-containing protein	Xci IAPAR 306
	XAC_RS00125	Peptidase M23	Xci IAPAR 306
	XAC_RS00350	Hypothetical protein	Xci IAPAR 306
	XAC_RS01780	Helix-turn helix transcriptional regulator	Xci IAPAR 306
	XAC_RS02490	Aminodeoxychorismate/anthranilate synthase component II	Xci IAPAR 306
	XAC_RS03785	Hypothetical protein	Xci IAPAR 306
	XAC_RS04070	Nudix family hydrolase	Xci IAPAR 306
	XAC_RS05160	Cell wall hydrolase	Xci IAPAR 306
	XAC_RS05275	Amidophosphoribosyltransferase	Xci IAPAR 306
	XAC_RS05830	2-methylaconitate cis-trans isomerase PrpF	Xci IAPAR 306
	XAC_RS06130	HDOD domain-containing protein	Xci IAPAR 306
	XAC_RS06280	Polyketide cyclase	Xci IAPAR 306
	XAC_RS06460	Helix-turn helix transcriptional regulator	Xci IAPAR 306
	XAC_RS07880	GntR family transcriptional regulator	Xci IAPAR 306
	XAC_RS09355	Methylthioribulose 1-phosphate dehydratase	Xci IAPAR 306
	XAC_RS15480	Hypothetical protein	Xci IAPAR 306
	XAC_RS15775	Glycosyltransferase	Xci IAPAR 306
	XAC_RS16605	BLUF domain-containing protein	Xci IAPAR 306
	XAC_RS17065	M23 family metalloproteinase	Xci IAPAR 306
	XAC_RS18580	Response regulator	Xci IAPAR 306
	XAC_RS19665	M23 family metalloproteinase	Xci IAPAR 306
	XAC_RS22275	Lytic murein transglycosylase	Xci IAPAR 306 pXAC64
	XAC1496	Uncharacterised protein	Xci IAPAR 306
	XAC3294	Uncharacterised protein	Xci IAPAR 306
avrBs1	XCVd0104	Unknown	Xee 85-10
avrXccA1	XCC4229	Unknown, maybe not a T3E	Xcc ATCC 33913
avrXccA2	XCC2396	Unknown, maybe not a T3E	Xcc ATCC 33913
xopAA	XCV3785	Early chlorosis factor; Proteasome/cyclosome repeat	Xee 85-10
xopAB	XOO3150_MAFF	Unknown	Xoo MAFF 311018
xopAC	XCC2565_ATCC33913	LRR-Fic/DOC protein; vascular hypersensitivity in <i>Arabidopsis landrace</i> Col-0	Xcc ATCC 33913
xopAD	XAC4213	SKWP repeat protein	Xoc BLS256

xopAF	XCAW_b00003	Unknown	<i>Xci</i> A ^W 12879
xopAF	XOC_0445_BLS256	Unknown	<i>Xoc</i> BLS256
xopAH	XCC2109_ATCC33913	Unknown	<i>Xcc</i> ATCC 33913
xopAJ	XCV4428	Unknown	<i>Xee</i> 85-10
xopAL1	XCC1246_ATCC33913	Unknown	<i>Xcc</i> ATCC 33913
xopAL2	XCCB100_0616	Unknown	<i>Xcc</i> B100
xopAM	XCC1089_ATCC33913	Unknown	<i>Xcc</i> ATCC 33913
xopAX	XCVd0086	Unknown	<i>Xee</i> 85-10
xopB	XCV0581	Unknown	<i>Xee</i> 85-10
xopC1	XCV2435	Phosphoribosyl transferase domain and haloacid dehalogenase-like hydrolase	<i>Xee</i> 85-10
xopD	XCV0437	C48-family SUMO cysteine protease (Ulp1 protease family) (Clan CE); EAR motif; DNA binding; nuclear localisation	<i>Xee</i> 85-10
xopF1	XCV0414	Unknown	<i>Xee</i> 85-10
xopF2	XCV2942	Unknown	<i>Xee</i> 85-10
xopG1	XCV1298	M27-family peptidase (Clostridium toxin)	<i>Xee</i> 85-10
xopG2	XCC3258	M27-family peptidase (Clostridium toxin)	<i>Xcc</i> ATCC 33913
xopH1	XCVd0105	Putative tyrosine phosphatase	<i>Xee</i> 85-10
xopJ1	XCV2156	C55-family cysteine protease or Ser/Thr acetyltransferase (Clan CE)	<i>Xee</i> 85-10
xopJ2	<i>Aave</i> _2166	C55-family cysteine protease or Ser/Thr acetyltransferase (Clan CE)	<i>Acidovorax citrulli</i> AAC00-1
xopJ3	XCV0471	C55-family cysteine protease or Ser/Thr acetyltransferase (Clan CE)	<i>Xee</i> 85-10
xopJ4	<i>RSc</i> 0826	C55-family cysteine protease or Ser/Thr acetyltransferase (Clan CE)	<i>Ralstonia solanacearum</i> GMI1000
xopJ5	XCC3731_ATCC33913	Putative C55-family cysteine protease or Ser/Thr acetyltransferase (Clan CE)	<i>Xcc</i> ATCC 33913
xopO	XCV1055	Unknown	<i>Xee</i> 85-10
xopU	PXO_00236	Unknown	<i>Xoo</i> PX099A
xopW	PXO_03356	Unknown	<i>Xoo</i> PX099A
xopY	XOO1488_MAFF	Unknown	<i>Xoo</i> MAFF 311018

793

794

795 Acknowledgments

796 We are grateful to P. Lefeuvre, D. Richard, F. Balloux, V. Llaurens, R. Debruyne & Á. Pérez-Qintero for
797 valuable comments and discussions. Computational work was performed on the CIRAD HPC data
798 center of the south green bioinformatics platform (<http://www.southgreen.fr/>) and MESO@LR-
799 Platform at the University of Montpellier (<https://hal.umontpellier.fr/MESO>).

800

Chapitre 5 – Collaborations et travaux annexes

En plus de mes travaux de recherche principaux ciblant *Xci* et présentés dans les chapitres précédents, j'ai eu, dans le cadre de cette thèse, l'opportunité de développer des collaborations scientifiques annexes avec plusieurs chercheurs sur la thématique générale d'étude et de valorisation des ADN historiques issus des collections d'herbiers. Parmi ces travaux annexes, certains comme l'application de la démarche développée chez *Xci* à une autre bactérie phytopathogène ou encore l'étude spécifique des données nucléotidiques historiques issues de la plante hôte, encore à un stade préliminaire seront plutôt abordés dans la discussion générale. Dans ce dernier chapitre, je présente une étude plus aboutie, réalisée en collaboration avec deux chercheurs virologues de l'UMR PVBMT : Jean-Michel Lett et Pierre Lefeuvre.

Contribution des ARNs interférents de petites tailles historiques issus d'herbier dans l'étude d'une maladie virale de cultures

Les *cassava mosaic geminivirus* (CMGs) sont des virus à ADN appartenant au genre *Begomovirus* de la famille des *Geminiviridae*. Les CMGs se caractérisent par un génome circulaire bipartite d'ADN simple brin (ADN-A et ADN-B). Ils forment un complexe d'espèces responsable de la maladie de la mosaïque du manioc, maladie la plus dommageable en Afrique, caractérisée par des mosaïques et déformations foliaires. Depuis de nombreuses années, des recherches ont été menées au sein de notre laboratoire dans le but de mieux comprendre l'histoire de ces virus et de caractériser les forces évolutives influençant leur structuration génétique ainsi que l'évolution des génomes des CMGs (Lefeuvre *et al.*, 2010; Lefeuvre and Moriones, 2015; De Bruyn *et al.*, 2016). Ces recherches avaient été jusqu'alors uniquement menées à partir d'isolats contemporains échantillonnés sur le terrain. Dans le cadre de cette collaboration, nous avons eu accès à un spécimen historique d'herbier de manioc (*Manihot glaziovii*) collecté en 1928 en République centrafricaine présentant des symptômes typiques de mosaïque du manioc. Suite à l'échec d'une première stratégie visant à détecter la présence de CMGs par amplification d'ADN viral, l'originalité de ce travail a dans un second temps consisté à tester une approche indirecte ciblant les ARNs de petites tailles, approche ayant précédemment fait ses preuves pour reconstruire des génomes complets de CMGs à partir d'échantillons contemporains (Kreuze *et al.*, 2009). Parmi les ARNs de petites tailles, les ARNs interférents de petites tailles (siRNAs, *small interfering RNAs*) sont des molécules issues de la voie de défense antivirale des plantes, champignons et invertébrés, appelée « interférence de l'ARN ». L'ARN messager du pathogène est rendu inaccessible à la traduction par hybridation avec de l'ARN polymérisé par la plante puis découpe des ARNs duplexés. L'une des conséquences de cette réaction de défense est alors la production de siRNAs de 21, 22 et 24 nt couvrant tout le génome viral (Pooggin, 2018) et permettant d'accéder *a posteriori*

aux génomes viraux accumulés au sein d'une plante pendant une infection. Ainsi, grâce à l'extraction et au séquençage des ARNs de petites tailles à partir de l'échantillon d'herbier analysé, nous avons réussi à reconstruire le génome historique d'un isolat d'*African cassava mosaic virus* (ACMV). Après authentification par analyse des patrons de dégradation de l'ARN, le génome a été comparé phylogénétiquement à un large jeu de données représentatif de la diversité génétique moderne de ce virus. Nos estimations de taux de substitutions et de temps de divergence réalisées en présence du génome historique diffèrent de celles réalisées sur données modernes seulement et présentent une meilleure concordance des dates avec les relevés d'occurrences historiques.

Ces résultats sont présentés dans un article publié en 2021 dans la revue Scientific Reports et intitulé « ***Contribution of historical herbarium small RNAs to the reconstruction of a cassava mosaic geminivirus evolutionary history*** ». Ma contribution à ce travail a consisté en l'application d'une partie de mon pipeline d'analyse bioinformatique (alignement des lectures et reconstruction du génome historique) et de la réalisation de certaines des analyses phylogénétiques.



OPEN

Contribution of historical herbarium small RNAs to the reconstruction of a cassava mosaic geminivirus evolutionary history

Adrien Rieux^{1✉}, Paola Campos^{1,2}, Arnaud Duvermy¹, Sarah Scussel¹, Darren Martin³, Myriam Gaudeul^{2,4}, Pierre Lefeuvre¹, Nathalie Becker² & Jean-Michel Lett^{1✉}

Emerging viral diseases of plants are recognised as a growing threat to global food security. However, little is known about the evolutionary processes and ecological factors underlying the emergence and success of viruses that have caused past epidemics. With technological advances in the field of ancient genomics, it is now possible to sequence historical genomes to provide a better understanding of viral plant disease emergence and pathogen evolutionary history. In this context, herbarium specimens represent a valuable source of dated and preserved material. We report here the first historical genome of a crop pathogen DNA virus, a 90-year-old African cassava mosaic virus (ACMV), reconstructed from small RNA sequences bearing hallmarks of small interfering RNAs. Relative to tip-calibrated dating inferences using only modern data, those performed with the historical genome yielded both molecular evolution rate estimates that were significantly lower, and lineage divergence times that were significantly older. Crucially, divergence times estimated without the historical genome appeared in discordance with both historical disease reports and the existence of the historical genome itself. In conclusion, our study reports an updated time-frame for the history and evolution of ACMV and illustrates how the study of crop viral diseases could benefit from natural history collections.

Crop pests and diseases have plagued farmers since the dawn of agriculture¹. Today they continue to be major threats to agro-ecosystems worldwide, significantly reducing yields, incurring economic losses and threatening food security^{2,3}. Amongst the different taxonomic groups of crop pathogens, viruses account for almost half of emerging infectious diseases⁴ and, as such, are a major focus of ongoing scientific investigation⁵.

The effective management of infectious viral crop diseases requires understanding the factors underlying virus emergence, adaptation and spread⁶. Elucidating the history of a pathogen's emergence is a prerequisite to inferring the evolutionary, ecological and anthropogenic factors that have driven the past epidemiological dynamics of the pathogen; inferences which could in turn be used to design efficient future disease control strategies⁷. As sequencing technologies have become more accessible, pathogen genomic analyses have played an increasingly important role in infectious disease research⁸. Concomitantly, recent methodological developments in molecular phylogeography can now be applied to study the emergence and evolution of viral pathogens in space and time with an unprecedented degree of accuracy⁹. Examples of such recent inferences performed on field-sampled viruses include the reconstruction of the spread and evolution of tomato yellow leaf curl virus (TYLCV)¹⁰, maize streak virus (MSV)¹¹ or rice yellow mottle virus (RYMV)^{12,13}. Interestingly, analyses of ancient DNA and RNA virus genomic sequence data obtained from herbaria or archaeological material have demonstrated that historical samples can be leveraged to substantially improve phylogenetic based molecular dating studies^{14–16}. By

¹CIRAD, UMR PVBMT, 97410 St Pierre, La Réunion, France. ²Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles, 57 Rue Cuvier, CP 50, 75005 Paris, France. ³Computational Biology Division, Department of Integrative Biomedical Sciences, Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Observatory, Cape Town, South Africa. ⁴Herbier national (P), Muséum national d'Histoire Naturelle, CP39, 57 Rue Cuvier, 75005 Paris, France. ✉email: adrien.rioux@cirad.fr; jean-michel.lett@cirad.fr

countering the molecular-clock calibration biases that occur when using modern genomes to infer ancient lineage divergence times, the addition of ancient genomes with known sampling dates commonly yields estimates of viral lineage divergence times that are older and more in accordance with historical reports than when the ancient sequences are not included in molecular dating studies^{17,18}. In this context, the oldest historical crop-associated virus genome published to date is a member of the *Chrysovirus* genus isolated from a 1,000 year old maize cob¹⁹.

High throughput sequencing (HTS) and bioinformatic analyses have already contributed to a paradigm shift in the fields of virus discovery and diagnosis^{20–22}. Among various possible targets, such as virion-associated nucleic acids, double-stranded RNAs, total RNAs, ribosomal-RNA-depleted RNAs or messenger RNAs, the sequencing of small RNAs (sRNA) offers several advantages²³. First, since plant viruses are targeted by host silencing mechanisms, the sequencing of small interfering RNAs (siRNAs) should enable the identification of all types of plant viral agents, whatever the nature or structure of their genomes (DNA or RNA, single or double stranded). In this context, the pioneering work of Kreuze et al.²⁴ demonstrated the universal power of targeting, sequencing and analysing sRNAs for the comprehensive reconstruction of viral genomes from fresh material of cultivated and non-cultivated plants (as reviewed in²⁵). Moreover, viral sRNAs were reported as more stable than long RNA and DNA molecules, and proved to be suitable for deep sequencing, including paleovirology applications for several plant RNA phytoviruses^{17,26}. As an illustration, Smith et al.¹⁷ have reported the identification and reconstruction of an ancient isolate of barley stripe mosaic virus (Genus *Hordeivirus*, family *Virgaviridae*) by sequencing sRNAs extracted from 700 years-old barley seeds, with 99.4% of the contemporary virus reference genome being covered by sRNA contigs. In a recent study reconstructing RNA phytovirus genomes, a detailed characterisation (using size distribution and coverage data) underscored the preservation of siRNAs among viral sRNAs from dried, modern samples, yet to be shown from historical samples²⁷.

Cassava cultivation is associated with a wide range of diseases that seriously undermine the food and economic security in sub-Saharan African countries, the most notable of which is cassava mosaic disease (CMD), caused by a complex of cassava mosaic geminiviruses (CMGs, genus *Begomovirus*, family *Geminiviridae*)²⁸. CMD is currently the most damaging plant virus disease in the world (estimates of US\$1.9–2.7 billion annual loss) and was associated with an East African famine in the late 1990s that likely caused the deaths of thousands of people²⁹. As an expanding global threat, CMD is currently under surveillance in Southeast Asia since its first description in Eastern Cambodia in 2016^{30,31}. CMGs are transmitted by whiteflies of the *Bemisia tabaci* species complex or by the use of infected cuttings (for review see²⁸). In sub-Saharan Africa cassava growing areas, several native species of the *B. tabaci* species complex referred as sub-Saharan African species (SSA) have been reported as the prevalent whiteflies associated with the spread of viruses that cause cassava mosaic disease (CMD)³². However, several cassava surveys suggest that the use of infected cassava cuttings for the establishment of new plantations appears to be largely responsible for the high incidence of CMD in sub-Saharan Africa^{33,34}. CMGs possess bipartite genomes, with genome components, called DNA-A and DNA-B, comprising 2.7 kb circular single-stranded DNA molecules. Both components are necessary for successful infection of cassava. While DNA-A encodes proteins and regulatory elements responsible for replication, encapsidation functions and the control of gene expression, DNA-B encodes proteins enabling viral movement³⁵. In plant cells infected by geminiviruses, bidirectional read-through transcription of the circular viral dsDNA generates sense and antisense transcripts²⁶. These dsRNA overlapping transcripts are processed by Dicer-like (DCL) family proteins from the RNA interference machinery, generating 21, 22 and 24 nt siRNAs and covering the entire circular virus genome (including coding sequences, as well as the intergenic region that contains the promoter^{36,37}).

Interestingly, whereas cassava originates from South America³⁸, the African CMGs are endemic to Africa and are likely recent descendants of geminiviruses adapted to infect indigenous uncultivated African plant species³⁹. Therefore the adaptation of CMGs to cassava could have only started, either after cassava was introduced to West Africa in the Gulf of Guinea during the 16th century, or after it was introduced to East Africa and the South West Indian Ocean islands in the 18th century. Since the initial characterisation in the early 1980s of the first known CMG species, African cassava mosaic virus (ACMV), several others have been described in sub-Saharan Africa, surrounding islands and the Indian subcontinent⁴⁰. The distribution of ACMV on the African continent has enabled the use of phylogeographic studies to investigate its evolutionary and epidemiological dynamics. Based on time-scaled phylogeographic analyses of modern ACMV isolates sampled between 1982 and 2012, it has been inferred that ACMV-driven CMD began disseminating in the 1980s only, with a single discreet movement of the virus from East Africa to Madagascar between 1996 and 2003⁴¹.

Here we report the genome of a 90-year-old ACMV isolate reconstructed from sRNAs, characteristic of *bona fide* siRNAs and whose damage patterns prove its authenticity. Using tip-calibrated phylogenetic inferences, we estimate both rates of molecular evolution and divergence times, underscoring the contribution of the historical genome in this calculation. Finally, we demonstrate how this single genome significantly improves our understanding of the history of ACMV in Africa.

Results and discussion

Nucleic acids isolation and high-throughput sequencing. From a small leaf fragment of a *Manihot glaziovii* (cassava) herbarium leaf specimen (Fig. 1) collected in the Central African Republic in 1928 and displaying typical symptoms of CMD, 185 ng of total DNA and 215 ng of total RNA were carefully extracted in a bleach-cleaned hospital laboratory with no prior exposure to plant material. Our first attempt to amplify and sequence viral DNA following Rolling Circle Amplification (RCA) failed (data not shown), likely due to substantial fragmentation of DNA, as previously described for herbarium specimens of similar age⁴². Hence, based on the pioneering work of Kreuze et al.²⁴ and Smith et al.¹⁷, we decided to target sRNAs. After library construction, high throughput sequencing of the sRNA fraction on an Illumina Hi-Seq High Output platform generated 8.6 M



Figure 1. Leaf of *Manihot glaziovii* specimen P04808771 collected in Bambari, Central African Republic, in June 1928 and preserved at the Herbarium of the Muséum national d'Histoire naturelle, Paris, France. The original annotation (hand-written in French on bottom left) states "Leaf from a young diseased plant". Typical symptoms of cassava mosaic disease such as chlorotic mosaic and deformation of the leaf can be distinguished.

single-end reads with a base call accuracy of 99.90 to 99.96%. Following adaptor trimming and quality checking, reads ranging from 18 to 26 nt in length were selected for further analyses (Fig. 2a).

Detection of genuine historical ACMV in herbarium cassava specimen. The analysis of sRNA reads with VirusDetect software revealed the presence of both DNA-A and DNA-B ACMV segments within the historical cassava specimen, with one contig covering 99.3% of the reference DNA-A sequence and fourteen contigs covering 88.7 % of the reference DNA-B sequence (Figure S1). We hence attempted to PCR amplify ACMV-specific DNA but no amplicons were successfully generated (data not shown). This result further highlights the promising potential underlying sRNAs sequencing to reconstruct historical viral pathogen genomes, as compared to classical approaches targeting DNA. Both DNA-A and DNA-B contigs harboured all eight typical open reading frames (ORFs) and inverted repeats (IRs) described for bipartite cassava geminiviruses (as depicted in⁴³). No other viruses were detected by VirusDetect from this sample. Running BWA-aln, a dedicated tool optimised for small read mapping, 1.45% of reads aligned to ACMV reference sequences and 87.55% of reads mapped to the *M. glaziovii* (cassava) reference sequence (Fig. 2b). Interestingly, among the 18–26 nt sRNAs mapping to ACMV or cassava, a predominance of ACMV-mapping sRNAs was observed for 21 and 22 nt sRNAs (Fig. 2a). These viral sRNAs may represent siRNAs, among the 21, 22 and 24 nt siRNA size classes described for geminiviruses²⁵.

To authenticate the historical nature of the ACMV siRNA reads and rule out the possibility of them being derived from lab contaminations, they were assessed for the presence of postmortem RNA damage. We found a clear pattern of C to U deamination reaching maximum values ($\pm 4\%$) at read extremities and declining exponentially inwards (Fig. 3C), as expected and previously shown for historical RNA^{17,44}. In addition, the examined modern control displayed no such pattern. We found no difference in deamination patterns between DNA-A and DNA-B segments (not shown). The historical consensus sequences of ACMV DNA-A and DNA-B were

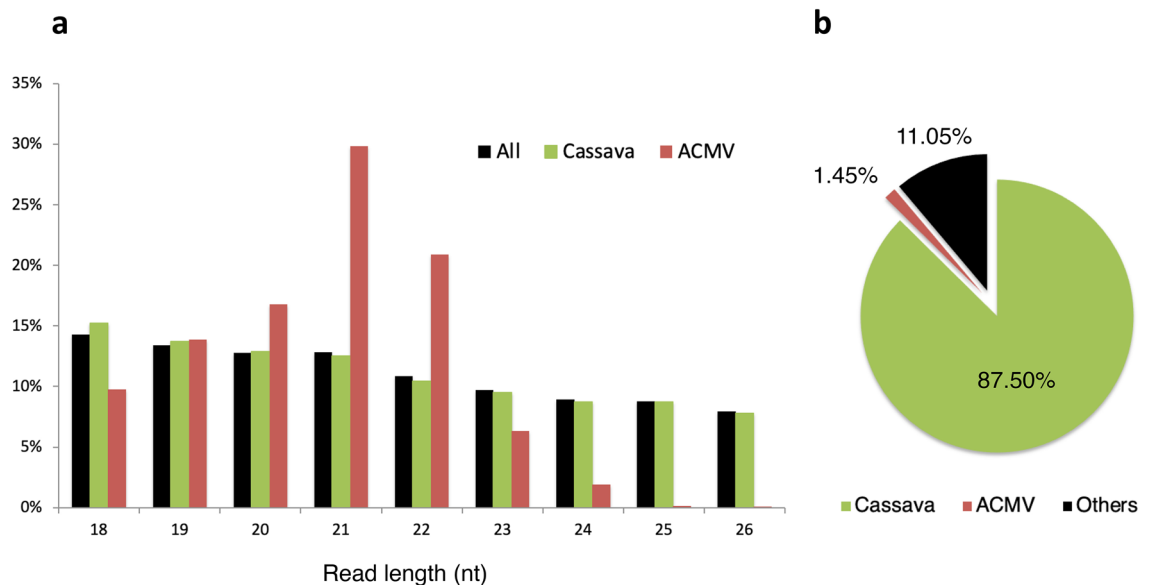


Figure 2. Main characteristics of small RNA (sRNA) isolated from historical specimen P04808771 (Herbarium of the Muséum national d'Histoire naturelle, Paris, France). **(a)** Size distribution of all, cassava-mapping and ACMV (DNA-A & DNA-B) genome-mapping reads. **(b)** Proportion of reads mapping to cassava and ACMV reference genomes.

reconstructed and covered 97.2% and 82.7% of the reference sequences (at 1X-fold) with a mean depth of 787.8 and 21.7 fold, respectively (Fig. 3A, B). Importantly, our mapping strategy aiming to reconstruct ACMV DNA-A and DNA-B consensus sequences was shown robust to (i) the choice of the short-read aligner, (ii) the presence of shared genomic regions between DNA-A & DNA-B segments and (iii) the choice of the reference sequences (Figure S2). The difference in sequencing depth between DNA-A and DNA-B could be explained by a difference in the abundance of these components in the plant tissues, and/or by higher host plant's RNAi-based antiviral defences targeting the DNA-A. In line with this latter observation, analyses of siRNA in ACMV-infected plants (*Nicotiana benthamiana* and cassava)^{36,43} showed a majority of siRNAs derived from the DNA-A component. A more detailed analysis of sRNA read coverage (Figure S3) locates a hotspot on ACMV-A, corresponding to overlapping transcripts coding for AC1, AC2 and AC3, consistent with previous siRNA analyses derived from ACMV^{36,43}.

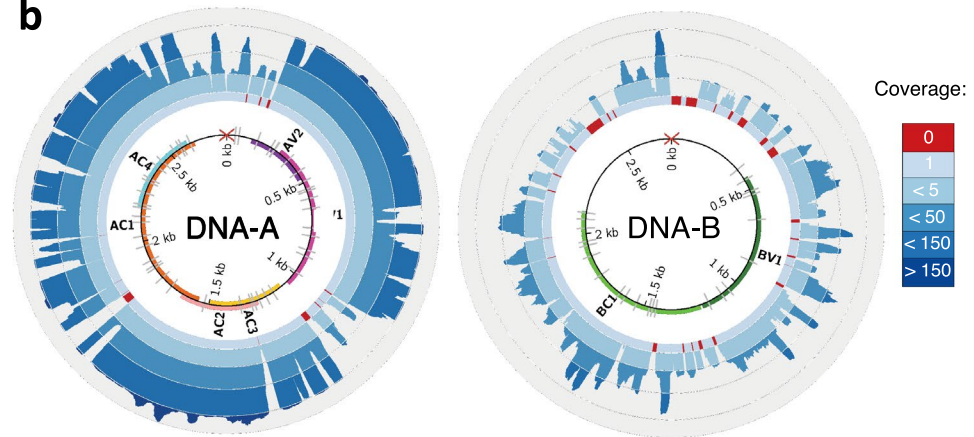
Recent large-scale surveys have revealed pervasiveness of transcriptionally active endogenous geminiviral sequences (EGSs) in several plant genomes^{45,46}. The hypothetical presence of sRNAs deriving from EGVs and their use in our analyses could potentially impact the reconstruction of our ancient viral sequences. However, for all the arguments developed below, we are convinced that the sRNAs sequenced in this study are from episomal viral DNA rather than EGS. First, to date only small portions of endogenous geminiviral sequences were proposed to be transcribed (homologous to ren and rep genes^{45,46}) while we were able to reconstruct a nearly complete ACMV genome from sRNA sequences. Second, in their recent study, Sharma et al.⁴⁶ did not find any trace of EGSs within the genome of *Manihot esculenta*. In this work, we analysed the currently publicly available genomic resources of *Manihot glaziovii*⁴⁷. Importantly, of the contigs that displayed similarities with geminiviruses (of length ranging between 163 and 2929 nt), all the hits covered 99 to 100% of the contigs. No chimeric contigs (containing both viruses and cassava sequences, that would indicate the presence of EGS), were detected (Table S3). This observation suggests that the analysed *M. glaziovii* genomes were generated from plants contaminated with episomal geminiviruses. Third, the herbarium specimen analysed displayed typical symptoms of Cassava Mosaic Disease. Although symptoms promoted by integrated viral sequences are theoretically possible, they wouldn't be expected for endogenous virus sequences, whose partial integration is unlikely to promote any infection⁴⁶, even for the longest integrated EGSs described so far⁴⁵. In addition, geminiviral endogenous elements have not so far been reported to give rise to episomal viruses^{25,48}. Finally, our reconstructed ACMV genomes showed a very high pairwise genetic identity (>99%) with their modern counterparts, a value that we would predict to be smaller in case of non-functional geminivirus sequences integrated in plant genomes for long periods⁴⁹.

Phylogenetic inferences and dating using both historical and modern sequences. In order to investigate the phylogenetic relationship of our historical sequences to those already available from recent samples, we built nucleotide alignments of our historical genome along with 134 and 99 public modern African ACMV DNA-A and DNA-B sequences, respectively. The historical and modern sequences displayed an average nucleotide divergence of 2.3% for DNA-A and 2.9% for DNA-B. Two recombinant events were detected in the ACMV sequences analysed in this study (Table S1). Recombinant ACMV regions (positions 631-781 & 1901-1933 relative to AY211884 sequence for ACMV DNA-A) were identified with RDP4⁵⁰ and removed from further inferences to avoid the potentially confounding effects these could have on the accuracy of inferred phylogenies. Note that as a precaution, recombinant region 2 was removed from the analysis, despite being detected in the

a

ACMV	Fraction of total reads mapping (%)	Read length mean [sd] (nt)	Mean depth (X)	% of reference genome covered at nX depth		
				0X	1X	10X
DNA-A	1.35	20.64 [1.49]	787.8	2.8	97.2	90.3
DNA-B	0.10	20.62 [1.60]	21.7	17.3	82.7	51.7

b



c

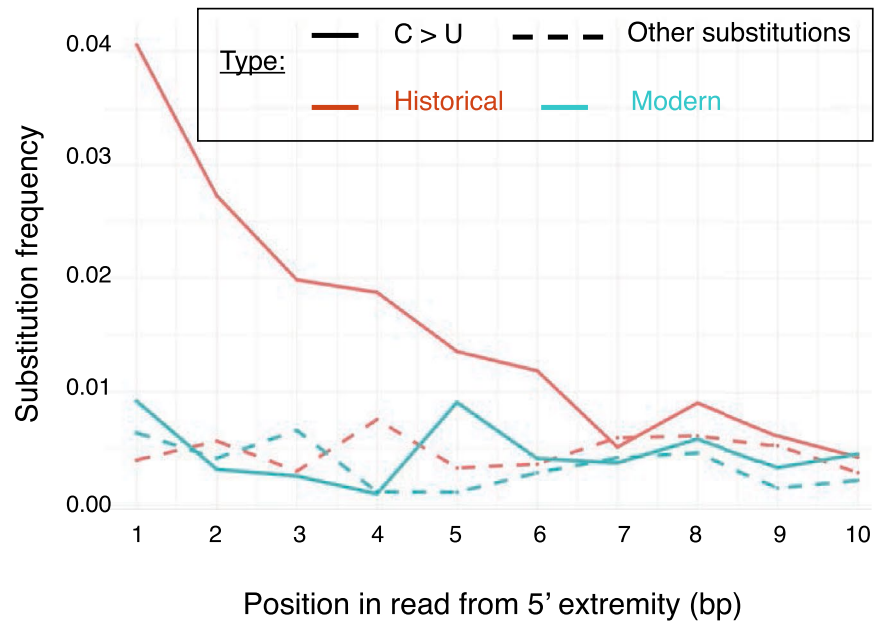


Figure 3. Reconstruction and authentication of historical ACMV genome. **(a)** Summary of mapping statistics to reference genomes for both ACMV DNA-A and DNA-B molecules. **(b)** Coverage plots (blue scale). Red arrays indicate regions that are not covered with siRNA reads (depth=0). Inner circle represents the genome and coding regions, as follows: AC1, AC3, AC3, AC4, AV1 and AV2 for DNA-A ; BC1 and BV1 for DNA-B; C: complementary strand; V: viral strand. Red cross symbolizes the geminivirus replication initiation site and grey ticks the SNPs between historical and reference sequences. **(c)** Post-mortem RNA damage patterns measured on historical (red) and modern ACMV sample isolated in 2017 (green). Straight and dotted lines represent C to U vs all other substitutions of the first 10 nucleotides from the 5'end, respectively.

historical DNA-A sequence with a single method only. The 1081 and 850 non-recombining SNPs obtained for ACMV DNA-A and DNA-B respectively were used to build Maximum-Likelihood (ML) phylogenies, using a cassava mosaic Madagascar virus (CMMGV) isolate (belonging to another species of CMG) as outgroup (Figure S4). The resulting ML trees were globally well supported (most bootstrap values >0.7) and appeared to be geographically structured. Interestingly, the historical ACMV genome sampled in 1928 in the Central African Republic clustered within a clade composed of modern isolates from the same country in both the ACMV DNA-A and DNA-B trees.

In order to date the evolutionary history of ACMV, we used the ACMV DNA-A dataset, as the historical DNA-A sequence displayed a much higher depth and coverage than the ACMV DNA-B. As a prerequisite to perform tip-based calibration, we tested the presence of temporal signal in our tree with both a linear regression between sample ages and root-to-tip distance, and a date-randomisation test. Both statistical tests revealed the presence of a temporal signal (i.e. progressive accumulation of substitutions over time) within the ACMV DNA-A tree. The linear regression test displayed a significant positive slope (slope value = 0.00017, adjusted R^2 = 0.0136 with a p -value = 0.038) and the date-randomisation test of the inferred root age of the real versus date-randomised dataset showed no overlap (Fig. 4). Additionally, our results showed no evidence of confounding between temporal and genetic structures (Mantel test: r = 0.001, p -value = 0.481), suggesting that the temporal signal detected is reliable and robust⁵¹. We therefore built a time-calibrated tree with BEAST⁵², which was globally congruent with the ML tree (similar topology and node supports; Figure 4). As in the ML-tree, the historical ACMV DNA-A sequence clustered within a clade composed of modern isolates sampled in the Central African Republic. This observation emphasises the value of historical samples in improving our understanding of the epidemiology of crop pathogens⁵³. Indeed, our historical ACMV genome constitutes “fossil” evidence that CMD has occurred in the Central African Republic since at least 1928, consistently with the very first historical report of a disease resembling CMD that was made in this country in 1924⁵⁴.

We inferred that the most recent common ancestor (MRCA) of all the analysed African ACMV DNA-A isolates most likely existed in 1849 [95% HPD: 1810–1880], a date that predates by more than 100 years the estimate of 1980 [95% HPD: 1990–1975] obtained by De Bruyn et al.⁴¹. The earlier estimate of the ACMV MRCA is more consistent with historical descriptions of the disease. Indeed, the earliest report of CMD-like symptoms in Africa was made in 1894 in what is now Tanzania⁴⁰. Subsequent reports were made in the 1920s in relation to CMD epidemics in Sierra Leone, Ivory Coast, Ghana, Nigeria, Madagascar and Uganda⁴⁰. By the end of the 1930s, CMD was reported from virtually all cassava-growing regions of the African mainland and surrounding islands.

We estimated a mean ACMV DNA-A substitution rate of 1.27×10^{-4} [95% HPD: 0.8×10^{-4} – 1.7×10^{-4}] per site per year, with a standard deviation for the uncorrelated log-normal relaxed clock of 0.26 [95% HPD: 0.18–0.33], suggesting low substitution-rate heterogeneity amongst branches. This rate estimate is $\sim 20 \times$ and $\sim 12.5 \times$ lower than that the ones previously obtained using modern isolates only of ACMV⁴¹ and EACMV⁵⁵, respectively.

Although our reconstructed evolutionary history of ACMV appears broadly inconsistent with the latter study using only modern isolates, the two analyses are not directly comparable because of differences in dataset composition and other methodological choices. To specifically evaluate the contribution of the historical ACMV DNA-A sequence to ACMV DNA-A MRCA date and substitution rate estimates, we reanalysed our dataset after removing the historical sequence. As anticipated, this reanalysis under the exact same parameters still yielded significantly different results, while belonging to the same order of magnitude. Excluding the historical sequence yielded a five times higher substitution rate estimate (Fig. 5A). The standard deviation of substitution rates amongst branches for the uncorrelated log-normal relaxed clock did not change significantly from the analysis including the historical sequence (not shown). Excluding the historical sequence also yielded a significantly later estimate date for the MRCA of the analysed ACMV DNA-A sequences (1957 [95% HPD: 1934–1976], Fig. 5B). Similarly, the MRCA age for Malagasy island isolates (believed to have arisen from a single introduction) was estimated to 1936 [95% HPD: 1900 – 1964] and 1990 [95% HPD: 1983–1998] when including or excluding it, respectively (Fig. 5C).

The timeline of ACMV DNA-A evolution that we have inferred when including the historical sequence is likely to be more accurate than that determined without this sequence for two main reasons. First, this estimated timeline fits better with historical reports of CMD disease, dating back to 1894 in Africa and to the 1930s in Madagascar⁴⁰. Second, the 95% credibility intervals of the estimated date of the ACMV DNA-A MRCA that was inferred without the historical sequence excludes 1928 and it therefore cannot be reconciled with the fact that a sequence sampled in 1928 clusters within the ACMV tree (i.e. it is not an outgroup) (Figure S5). Such striking lower substitution rate and hence higher divergence time estimates, when including ancient viral genome sequences, have been previously described in molecular dating studies focusing on different virus group representatives: barley stripe mosaic virus (BSMV)¹⁷, Human immunodeficiency virus⁵⁶, hepatitis B virus⁵⁷, as well as parvovirus B19⁵⁸ (a ssDNA Baltimore group II virus to whom ACMV belongs), as recently reviewed in¹⁵.

In summary, our results illustrate that high-quality historical genomes of DNA viruses can be both reconstructed by sequencing the small RNA fraction of a plant herbarium specimen, harbouring siRNA characteristics and authenticated by analysing post-mortem RNA damage patterns. Such historical genomes represent “fossil” records of past viral diversity that have the potential to shed light on the spatiotemporal history of plant diseases. Indeed, our results demonstrate that CMD-causing ACMV variants were already present in the Central African Republic in 1928, supporting the accuracy of the description of a historical record of CMD made in 1924 from visual inspection of cassava leaves. Second, phylogenetic inferences performed with the inclusion of our historical ACMV DNA-A sequence significantly altered the inferred date at which the MRCA of all currently sampled ACMV variants likely existed, providing a better fit with historical reports than previous estimates and yielding a lower rate of ACMV DNA-A molecular evolution. Future studies including additional historical ACMV genome sequences that are more geographically/temporally dispersed will help us to refine the evolutionary parameters inferred herein. The presence of ACMV should also be tested in other herbarium plant species/families if one

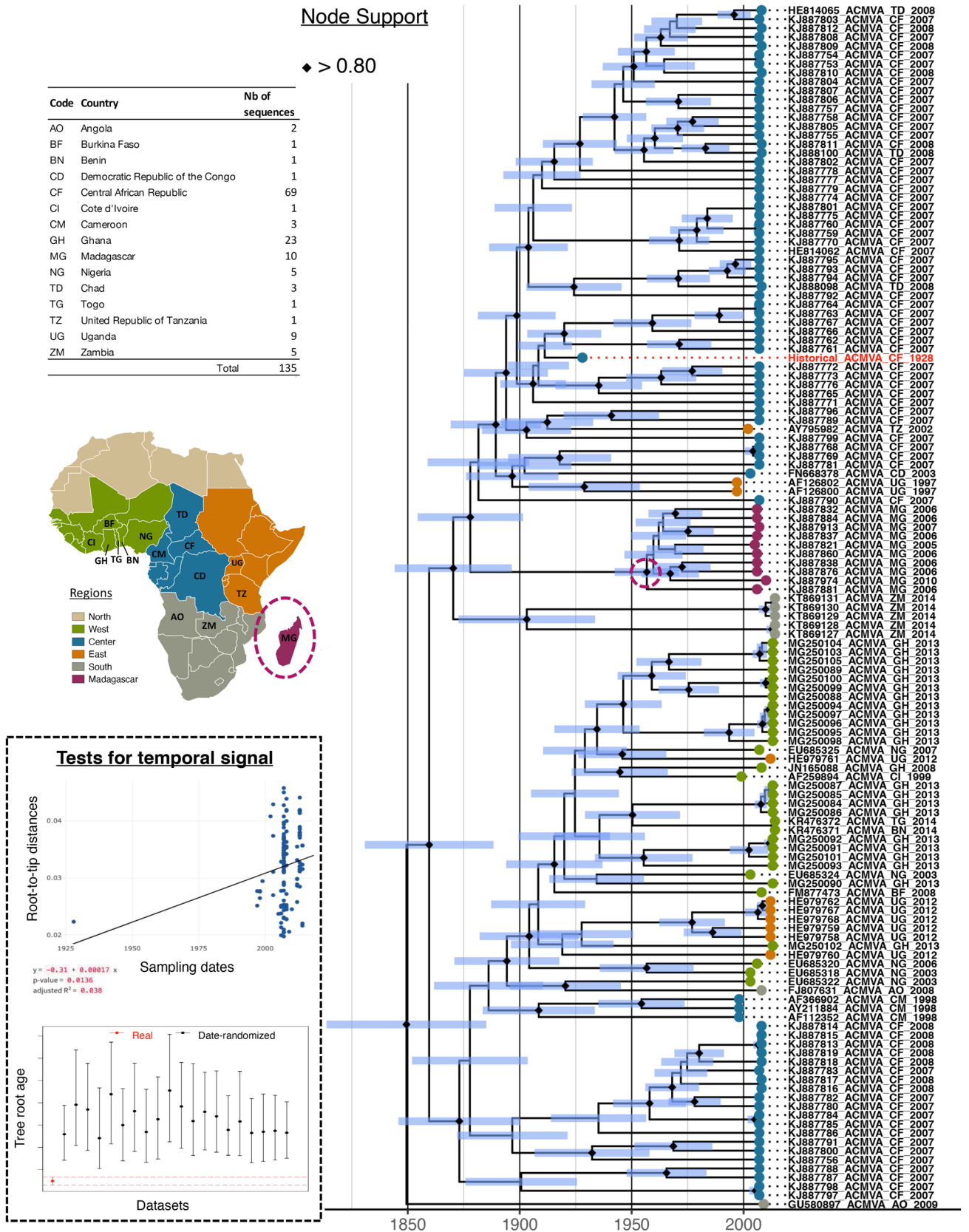


Figure 4. Bayesian dated tree of 134 sequences of ACMV DNA-A built from 1081 non-recombining SNPs. The historical DNA-A sequence is highlighted in red. Node support values with posterior probabilities above 0.8 are displayed by black diamonds. Node bars cover 95% Highest Probability Density of node height. Tips are colored according to the sample's geographic origin, according to the map on top left. The node corresponding to the common ancestor of all Malagasy isolates is circled in purple. Both tests of temporal signal (top: linear regression of root-to-tip distance on year of sampling date and bottom: date-randomization test) are presented in the dotted box.

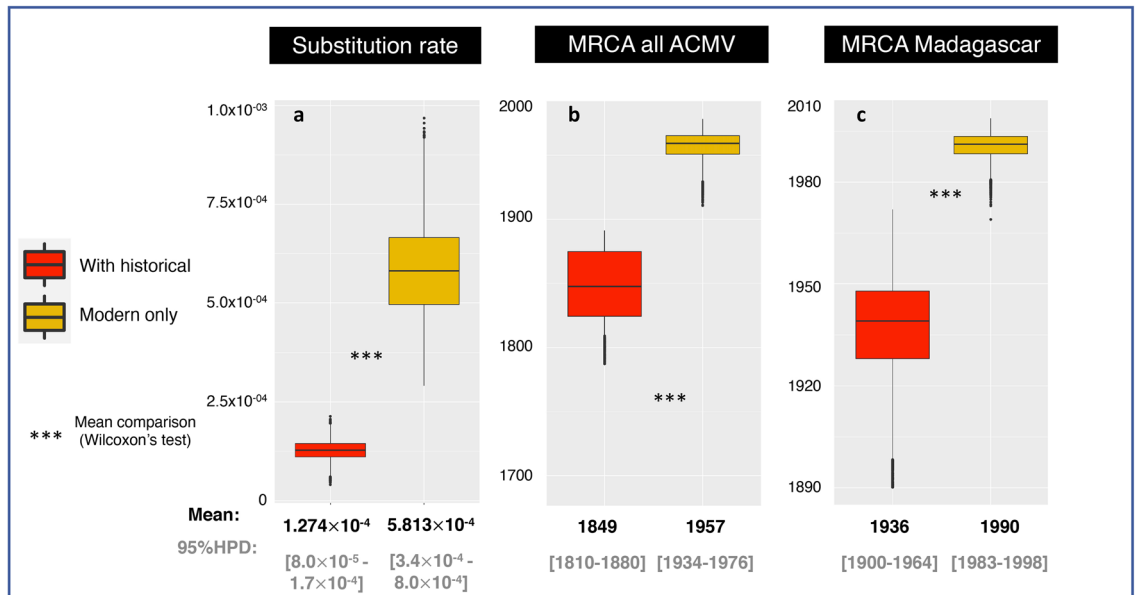


Figure 5. Bayesian estimations performed with or without including the historical genome. Substitution rate (a), MRCA of all (b) and from Madagascar (c) isolates, inferred with (red boxplot) and without (orange boxplot) the historical ACMV DNA-A component. *** $p < 0.001$.

aims to investigate possible host-switching events that may have led to the emergence of CMD in cassava. More generally, similar investigations on other important viral crop pathogens will improve disease monitoring and sustainable control, while highlighting the importance of natural history collections.

Material and methods

Herbarium sampling. In 2014, the historical collection of cassava specimens of the National Herbarium of the Muséum National d'Histoire Naturelle, Paris (<https://www.mnhn.fr/en>) was searched for in 2014 for leaves displaying symptoms of CMD. Sample P04808771 (Fig. 1), a *Manihot glaziovii* specimen collected by C. Tisserant at Bambari, Central African Republic in 1928, displayed chlorotic mosaic and leaf distortion, two typical symptoms of CMD. A small leaf fragment ($\approx 1\text{cm}^2$ / 12mg of dry material) was excised from this specimen using a disinfected blade and gloves, sealed in a clean envelope, transported to Reunion Island and stored in a vacuum-sealed box at 14°C until use. Permission to sample and perform destructive analysis on historical specimen P04808771 was obtained from the Muséum national d'Histoire naturelle (Paris, France). Collection of any plant material used in this study complies with institutional, national, and international guidelines.

DNA extraction, amplification and sequencing. DNA isolation was performed in a bleach-cleaned molecular biology laboratory at the Centre Hospitalier Universitaire Sud Réunion that met the authenticity criteria for the extraction of ancient biomolecules⁵⁹: a laboratory in which no plant samples had been manipulated before. Total DNA was extracted from the herbarium sample following manufacturer's instructions of the Qiagen DN easy plant kit. We attempted to detect both viral and ACMV specific DNA using the standard RCA-Cloning-Sanger sequencing protocol⁶⁰ and amplification of overlapping ACMV-specific PCR amplicons (ranging from 54 to 381nt), using validated primers (Harimalala, personal communication) listed in Table S2, respectively.

RNA extraction, library preparation, sequencing and quality control. RNA isolation was also performed at the Centre Hospitalier Universitaire Sud Réunion. Total RNA was extracted from the herbarium sample using a PureLink Plant RNA Reagent kit (Ambion) and quantified using an Agilent 2200 TapeStation system (Agilent, France). Purification of siRNA, library preparation and sequencing were carried out by FASTERIS NGS service team in Geneva, Switzerland. Using polyacrylamide gel electrophoresis, fragments of 18-30nt long were selected and converted into sequencing library using the Illumina TruSeq Small RNA Library Preparation kit. Sequencing was performed in a 1×50 cycle mode on a HiSeq instrument. Adaptors were trimmed from raw reads using the Illuminaclip option in Trimmomatic 0.36⁶¹. Additional quality-trimming was performed using the same tool to remove low Illumina quality score-associated bases (SLIDINGWINDOW:5:20) and reads shorter than 15nt (MINLEN:15). Those of size 18-30nt were retained as clean reads.

Virus detection and taxonomic classification. To identify viruses from our historical sample, we first used VirusDetect⁶², a bioinformatic pipeline built to efficiently analyse large-scale small RNA (sRNA) datasets. We fixed all parameters to their default values and used the Sept 2019 GenBank reference virus genome database. In a second step, we used the dedicated short read aligner BWA-aln⁶³ (with the following optimised options fixed

as in VirusDetect pipeline: -n 1 -o 1 -e 1 -i 0 -l 15 -k 1) to map quality-trimmed reads to both viral (i.e. the species detected by VirusDetect) and host plant (*Manihot glaziovii* specimen GISH—SRA: SRS597345) reference genomes. Our reads-mapping strategy was further assessed for the three following aspects. First, we evaluated the performance of another short-read aligner, Bowtie⁶⁴, allowing one mismatch. Second, we compared the effect of mapping reads either independently or simultaneously to both ACMV DNA-A and DNA-B segments, in order to evaluate the influence of shared genomic regions. Finally, we assessed the effect of reference choice on mapping statistics and variant calling/filtering. To this aim, reads were mapped to three supplementary reference sequences (selected for their close, intermediate and distant phylogenetic proximity with the historical genome).

Historical viral genome authentication and reconstruction. We examined the sequences for cytosine deamination patterns—a typical proxy of postmortem RNA damage—to authenticate the historical nature of the siRNA ACMV sequences obtained. Distributions of C to U vs other transitions along the siRNA reads were assessed from raw untrimmed reads using the dedicated mapDamage2 tool⁶⁵. Postmortem RNA damage was compared between the historical specimen and RNA isolated from an ACMV infected *Manihot esculenta* leaf sample collected in Madagascar in 2017. The modern RNA sample was obtained using the exact same wet-lab protocol used to obtain RNA from the 1928 sample. Quality scores of post-mortem damaged bases were down-scaled using the rescale parameter to correct for the effect of deamination and avoid generating artifactual SNPs in subsequent analyses. Historical ACMV DNA-A and DNA-B sequences were reconstructed from rescaled-BWA-aln generated BAM files for both DNA-A (JX658682) and DNA-B (KJ887590) GeneBank segment references. In brief, PCR duplicates were removed using picardtools 2.7.0 MarkDuplicates⁶⁶ and depth statistics were computed with the genomecov option of BEDTools 2.24.0⁶⁷, which were then graphically represented with CIRCOS 0.69.9⁶⁸. SNPs were called with GATK UnifiedGenotyper⁶⁹ and filtered out when their sequenced depth was <10 or their allelic frequency was < 0.6. Consensus historical sequences were then reconstructed by editing the reference DNA-A and DNA-B sequences with the remaining high-quality SNPs while replacing both filtered-out variants and unsequenced nucleotide sites (i.e. sites with a sequencing depth= 0) with “Ns”. Genes coding for AC3 and AC4 were deduced from other known ACMV sequences; all sequences were checked for open reading frame features.

In order to investigate the persistence of endogenous geminiviral sequences (EGSs) within *Manihot glaziovii* genomes, we downloaded raw reads of the two only available African *M. glaziovii* samples⁴⁷ at the date of search (01/08/2021) within the SRA database (SRR2847420 & SRR2847424). After *de novo* assembly of the reads into contigs with SPAdes V3.15.2⁷⁰ using default parameters, all reconstructed contigs were blasted (using BLASTN) on a custom-built database containing all described species of cassava mosaic geminiviruses. We predicted that the identification of chimeric contigs (composed of both cassava and virus sequences) would indicate the presence of EGSs. Instead, the detection of contigs displaying hits with virus sequences on their whole length would suggest plant infection by episomal viruses. Finally, the absence of any hits would reveal the absence of viral DNA, both from episomal and integrated forms, within *M. glaziovii* genomes.

Phylogenetic inferences using both historical and modern sequences. Alignments of our historical ACMV DNA-A and DNA-B components with 134 (for DNA-A) and 99 (for DNA-B) publicly available ACMV genome component sequences sampled between 1978 and 2014 (Table S4) were constructed with MAFFT⁷¹ for phylogenetic analyses. Each of these alignments also included a CMMGV sequence as an out-group (accession number HE617299 and HE617300 for DNA-A and DNA-B, respectively). Regions acquired via recombination were identified with RDP4⁵⁰ with default settings. Events that were detected by three or more methods with P-values <0.05 were accepted as credible and removed to avoid the potentially biasing impacts of recombination on phylogenetic reconstruction. Note that the historical sequence was analysed with particular scrutiny and recombination events detected with a single method were taken into account. Maximum likelihood trees for each of these alignments were constructed using RAxML 8.2.4⁷² using a rapid bootstrap test and the GTR+G+I model of nucleotide substitution was chosen as best-fitted model based on the Bayesian Information Criterion (BIC) computed with JModelTest2.0⁷³.

The existence of a temporal signal in this dataset was investigated using two different tests. First, a linear regression was fitted between sample age and root-to-tip distance using the distRoot function of the adephylo R package⁷⁴. Temporal signal was considered present if a significant positive correlation was observed. Secondly, we performed a date-randomisation test (DRT)⁷⁵ with 20 independent date-randomised datasets using the R package, TipDatingBeast⁷⁶. Temporal signal was considered present when there was no overlap between the inferred root height 95% highest posterior density (95% HPD) of the initial dataset and that of 20 date-randomised datasets. Finally, we also investigated whether our dataset showed confounding effects between temporal and genetic structures using a Mantel confounding test which investigate whether closely related sequences were more likely to have been sampled at similar times. This additional test is important because both the root-to-tip regression and the DRT can be confounded in such a situation⁵¹.

Tip-dating was performed with BEAST 1.8.4⁵² considering a GTR substitution model with a Γ distribution and invariant sites (GTR+G+I) along with an uncorrelated log-normal relaxed (UCLNR) clock to account for minor variations between lineages. Bayes factors calculated from the marginal likelihoods using both path and stepping-stone sampling methods shown “very strong” support (BF>10⁷⁷) for UCLNR over strict (S) and random local (RL) clocks. To minimise prior assumptions about demographic history, an extended Bayesian skyline plot (EBSP) approach was adopted to integrate data over different coalescent histories⁷⁸. Three independent chains were run for 25 million steps and sampled every 2500 steps with a burn-in of the first 2500 steps. Convergence to the stationary distribution and sufficient sampling and mixing were checked by inspection of posterior samples (effective sample size >200) in Tracer 1.7.1⁷⁹. Parameter estimation was based on the samples combined from the

different chains. The best-supported tree was estimated from the combined samples using the maximum clade credibility method implemented in TreeAnnotator. In order to specifically assess the effect of including our historical genome in the dating calibration, we computed the same inferences on a dataset where the 1928 DNA-A sequence was excluded (i.e. using only sequences sampled after 1977). Wilcoxon rank sum tests with continuity correction were performed to compare the means of the posterior estimates obtained from both datasets.

Data availability

Raw reads were deposited to the Sequence Read Archive (SRR13608699). Consensus historical genome reconstructed for ACMV DNA-A and DNA-B molecules have also been deposited on GenBank database (MW788219 & MW788220). The modern genomes used in this study have previously been published in the NCBI GenBank repository under accession numbers listed in Table S4.

Received: 8 March 2021; Accepted: 13 October 2021

Published online: 28 October 2021

References

1. Stukenbrock, E. H. & McDonald, B. A. The origins of plant pathogens in agro-ecosystems. *Annu. Rev. Phytopathol.* <https://doi.org/10.1146/annurev.phyto.010708.154114> (2008).
2. Savary, S., Ficke, A., Aubertot, J. N. & Hollier, C. Crop losses due to diseases and their implications for global food production losses and food security. *Food Secur.* <https://doi.org/10.1007/s12571-012-0200-5> (2012).
3. Strange, R. N. & Scott, P. R. Plant disease: a threat to global food security. *Annu. Rev. Phytopathol.* <https://doi.org/10.1146/annurev.phyto.43.113004.133839> (2005).
4. Anderson, P. K. *et al.* Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. *Trends Ecol. Evol.* <https://doi.org/10.1016/j.tree.2004.07.021> (2004).
5. Scholthof, K. B. G. *et al.* Top 10 plant viruses in molecular plant pathology. *Mol. Plant Pathol.* <https://doi.org/10.1111/j.1364-3703.2011.00752.x> (2011).
6. Stukenbrock, E. H. & Bataillon, T. A population genomics perspective on the emergence and adaptation of new plant pathogens in agro-ecosystems. *PLoS Pathog.* <https://doi.org/10.1371/journal.ppat.1002893> (2012).
7. Gilligan, C. A. Sustainable agriculture and plant diseases: an epidemiological perspective. *Philos. Trans. R. Soc. B: Biol. Sci.* <https://doi.org/10.1098/rstb.2007.2181> (2008).
8. Li, L. M., Grassly, N. C. & Fraser, C. Genomic analysis of emerging pathogens: methods, application and future trends. *Genome Biol. Evol.* <https://doi.org/10.1186/s13059-014-0541-9> (2014).
9. Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian phylogeography finds its roots. *PLoS Comput. Biol.* <https://doi.org/10.1371/journal.pcbi.1000520> (2009).
10. Lefevre, P. *et al.* The spread of tomato yellow leaf curl virus from the middle east to the world. *PLoS Pathog.* <https://doi.org/10.1371/journal.ppat.1001164> (2010).
11. Monjane, A. L. *et al.* Reconstructing the history of maize streak virus strain A dispersal to reveal diversification hot spots and its origin in southern Africa. *J. Virol.* <https://doi.org/10.1128/jvi.00640-11> (2011).
12. Trovao, N. S. *et al.* Host ecology determines the dispersal patterns of a plant virus. *Virus Evol.* <https://doi.org/10.1093/ve/vev016> (2015).
13. Rakotomalala, M. *et al.* Comparing patterns and scales of plant virus phylogeography: rice yellow mottle virus in Madagascar and in continental Africa. *Virus Evol.* <https://doi.org/10.1093/ve/vez023> (2019).
14. Gibbs, A. J., Fargette, D., García-Arenal, F. & Gibbs, M. J. Time - The emerging dimension of plant virus studies. *J. General Virol.* <https://doi.org/10.1099/vir.0.015925-0> (2010).
15. Simmonds, P., Aiewsakun, P. & Katzourakis, A. Prisoners of war: host adaptation and its constraints on virus evolution. *Nat. Rev. Microbiol.* <https://doi.org/10.1038/s41579-018-0120-2> (2019).
16. Jones, R. A. C., Boonham, N., Adams, I. P. & Fox, A. Historical virus isolate collections: an invaluable resource connecting plant virology's pre-sequencing and post-sequencing eras. *Plant Pathol.* **70**, 235–248 (2021).
17. Smith, O. *et al.* A complete ancient RNA genome: Identification, reconstruction and evolutionary history of archaeological Barley Stripe Mosaic Virus. *Sci. Rep.* <https://doi.org/10.1038/srep04003> (2014).
18. Malmstrom, C. M., Shu, R., Linton, E. W., Newton, L. A. & Cook, M. A. Barley yellow dwarf viruses (BYDVs) preserved in herbarium specimens illuminate historical disease ecology of invasive and native grasses. *J. Ecol.* <https://doi.org/10.1111/j.1365-2745.2007.01307.x> (2007).
19. Peyambari, M., Warner, S., Stoler, N., Rainer, D. & Roossinck, M. J. A 1000-Year-old RNA virus. *J. Virol.* **93**, e01188-18 (2019).
20. Adams, I. P. *et al.* Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. *Mol. Plant Pathol.* <https://doi.org/10.1111/j.1364-3703.2009.00545.x> (2009).
21. Vayssier-Taussat, M. *et al.* Shifting the paradigm from pathogens to pathobiome new concepts in the light of meta-omics. *Front. Cell. Infect. Microbiol.* <https://doi.org/10.3389/fcimb.2014.00029> (2014).
22. Massart, S., Olmos, A., Jijakli, H. & Candresse, T. Current impact and future directions of high throughput sequencing in plant virus diagnostics. *Virus Res.* <https://doi.org/10.1016/j.virusres.2014.03.029> (2014).
23. Roossinck, M. J., Martin, D. P. & Roumagnac, P. Plant virus metagenomics: advances in virus discovery. *Phytopathology* <https://doi.org/10.1094/PHYTO-12-14-0356-RVW> (2015).
24. Kreuze, J. F. *et al.* Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. *Virology* <https://doi.org/10.1016/j.virol.2009.03.024> (2009).
25. Pooggin, M. M. Small RNA-omics for plant virus identification, virome reconstruction, and antiviral defense characterization. *Front. Microbiol.* <https://doi.org/10.3389/fmicb.2018.02779> (2018).
26. Hartung, J. S. *et al.* History and diversity of Citrus Leprosis virus recorded in herbarium specimens. *Phytopathology* <https://doi.org/10.1094/PHYTO-03-15-0064-R> (2015).
27. Golyaev, V., Candresse, T., Rabenstein, F. & Pooggin, M. M. Plant virome reconstruction and antiviral RNAi characterization by deep sequencing of small RNAs from dried leaves. *Sci. Rep.* <https://doi.org/10.1038/s41598-019-55547-3> (2019).
28. Patil, B. L. & Fauquet, C. M. Cassava mosaic geminiviruses: actual knowledge and perspectives. *Mol. Plant Pathol.* <https://doi.org/10.1111/j.1364-3703.2009.00559.x> (2009).
29. Legg, J. P., Owor, B., Sseruwagi, P. & Ndunguru, J. Cassava mosaic virus disease in east and central Africa: epidemiology and management of a regional pandemic. *Adv. Virus Res.* [https://doi.org/10.1016/S0065-3527\(06\)67010-3](https://doi.org/10.1016/S0065-3527(06)67010-3) (2006).
30. Wang, H. L. *et al.* First report of Sri Lankan cassava mosaic virus infecting cassava in Cambodia. *Plant Dis.* <https://doi.org/10.1094/PDIS-10-15-1228-PDN> (2016).
31. Minato, N. *et al.* Surveillance for sri lankan cassava mosaic virus (SLCMV) in Cambodia and Vietnam one year after its initial detection in a single plantation in 2015. *PLoS One* <https://doi.org/10.1371/journal.pone.0212780> (2019).

32. Mugerwa, H., Wang, H. L., Sseruwagi, P., Seal, S. & Colvin, J. Whole-genome single nucleotide polymorphism and mating compatibility studies reveal the presence of distinct species in sub-Saharan Africa Bemisia tabaci whiteflies. *Insect Sci.* <https://doi.org/10.1111/1744-7917.12881> (2020).
33. Ntawuruhunga, P. *et al.* Incidence and severity of cassava mosaic disease in the Republic of Congo. *African Crop Sci. J.* <https://doi.org/10.4314/acsj.v15i1.54405> (2010).
34. Zinga, I. *et al.* Epidemiological assessment of cassava mosaic disease in Central African Republic reveals the importance of mixed viral infection and poor health of plant cuttings. *Crop Prot.* <https://doi.org/10.1016/j.cropro.2012.10.010> (2013).
35. Jeske, H. Geminiviruses. *Curr. Topics Microbiol. Immunol.* https://doi.org/10.1007/978-3-540-70972-5_11 (2009).
36. Vanitharani, R., Chellappan, P. & Fauquet, C. M. Geminiviruses and RNA silencing. *Trends Plant Sci.* <https://doi.org/10.1016/j.tplants.2005.01.005> (2005).
37. Aregger, M. *et al.* Primary and secondary siRNAs in geminivirus-induced gene silencing. *PLoS Pathog.* <https://doi.org/10.1371/journal.ppat.1002941> (2012).
38. Olsen, K. M. & Schaal, B. A. Evidence on the origin of cassava: Phylogeography of *Manihot esculenta*. *Proc. Natl. Acad. Sci. USA* <https://doi.org/10.1073/pnas.96.10.5586> (1999).
39. Fauquet, C. African cassava mosaic virus: etiology, epidemiology, and control. *Plant Dis.* <https://doi.org/10.1094/pd-74-0404> (1990).
40. Legg, J. P. & Fauquet, C. M. Cassava mosaic geminiviruses in Africa. *Plant Mol. Biol.* <https://doi.org/10.1007/s11103-004-1651-7> (2004).
41. De Bruyn, A. *et al.* Divergent evolutionary and epidemiological dynamics of cassava mosaic geminiviruses in Madagascar. *BMC Evol. Biol.* <https://doi.org/10.1186/s12862-016-0749-2> (2016).
42. Weiß, C. L. *et al.* Temporal patterns of damage and decay kinetics of dna retrieved from plant herbarium specimens. *R. Soc. Open Sci.* <https://doi.org/10.1098/rsos.160239> (2016).
43. Chellappan, P., Vanitharani, R., Ogbe, F. & Fauquet, C. M. Effect of temperature on geminivirus-induced RNA silencing in plants. *Plant Physiol.* <https://doi.org/10.1104/pp.105.066563> (2005).
44. Smith, O. & Gilbert, M. T. P. Ancient RNA. in (2018). doi:https://doi.org/10.1007/13836_2018_17.
45. Filloux, D. *et al.* The genomes of many yam species contain transcriptionally active endogenous geminiviral sequences that may be functionally expressed. *Virus Evol.* <https://doi.org/10.1093/ve/vev002> (2015).
46. Sharma, V. *et al.* Large-scale survey reveals pervasiveness and potential function of endogenous geminiviral sequences in plants. *Virus Evol.* <https://doi.org/10.1093/ve/veaa071> (2020).
47. Bredeson, J. V. *et al.* Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nat. Biotechnol.* <https://doi.org/10.1038/nbt.3535> (2016).
48. Serfraz, S. *et al.* Insertion of Badnaviral DNA in the Late Blight Resistance Gene (R1a) of Brinjal Eggplant (*Solanum melongena*). *Front. Plant Sci.* <https://doi.org/10.3389/fpls.2021.683681> (2021).
49. Lefeuvre, P. *et al.* Evolutionary time-scale of the begomoviruses: evidence from integrated sequences in the Nicotiana genome. *PLoS One* <https://doi.org/10.1371/journal.pone.0019193> (2011).
50. Martin, D. P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4: detection and analysis of recombination patterns in virus genomes. *Virus Evol.* <https://doi.org/10.1093/ve/vev003> (2015).
51. Murray, G. G. R. *et al.* The effect of genetic structure on molecular dating and tests for temporal signal. *Methods Ecol. Evol.* **7**, 80–89 (2016).
52. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* <https://doi.org/10.1186/1471-2148-7-214> (2007).
53. Yoshida, K. *et al.* Mining herbaria for plant pathogen genomes: back to the future. *PLoS Pathog.* <https://doi.org/10.1371/journal.ppat.1004028> (2014).
54. Dufrenoy, J. & Hédin, L. La. Mosaïque des feuilles du Manioc au Cameroun. *J. d'Agriculture Tradit. Bot. appliquée* **94**, 361–365 (1929).
55. Duffy, S. & Holmes, E. C. Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. *J. Gen. Virol.* **90**, 1539–1547 (2009).
56. Worobey, M. *et al.* Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* <https://doi.org/10.1038/nature07390> (2008).
57. Mühlemann, B. *et al.* Ancient hepatitis B viruses from the Bronze Age to the Medieval period. *Nature* <https://doi.org/10.1038/s41586-018-0097-z> (2018).
58. Toppinen, M. *et al.* Bones hold the key to DNA virus history and epidemiology. *Sci. Rep.* <https://doi.org/10.1038/srep17226> (2015).
59. Gilbert, M. T. P., Bandelt, H. J., Hofreiter, M. & Barnes, I. Assessing ancient DNA studies. *Trends Ecol. Evol.* <https://doi.org/10.1016/j.tree.2005.07.005> (2005).
60. Inoue-Nagata, A. K., Albuquerque, L. C., Rocha, W. B. & Nagata, T. A simple method for cloning the complete begomovirus genome using the bacteriophage ϕ 29 DNA polymerase. *J. Virol. Methods* <https://doi.org/10.1016/j.jviro.2003.11.015> (2004).
61. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btu170> (2014).
62. Zheng, Y. *et al.* VirusDetect: An automated pipeline for efficient virus discovery using deep sequencing of small RNAs. *Virology* <https://doi.org/10.1016/j.virol.2016.10.017> (2017).
63. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btp324> (2009).
64. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* <https://doi.org/10.1186/gb-2009-10-3-r25> (2009).
65. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F. & Orlando, L. MapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. in *Bioinformatics* (2013). doi:<https://doi.org/10.1093/bioinformatics/btt193>.
66. Broad Institute. Picard Tools - By Broad Institute. *GitHub* (2009).
67. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btq033> (2010).
68. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* <https://doi.org/10.1101/gr.092759.109> (2009).
69. Depristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* <https://doi.org/10.1038/ng.806> (2011).
70. Bankevich, A. *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* <https://doi.org/10.1089/cmb.2012.0021> (2012).
71. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* <https://doi.org/10.1093/molbev/mst010> (2013).
72. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
73. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. JModelTest 2: More models, new heuristics and parallel computing. *Nat. Methods* <https://doi.org/10.1038/nmeth.2109> (2012).

74. Jombart, T. & Dray, S. Adephylo: Exploratory analyses for the phylogenetic comparative method. *Bioinformatics* (2010).
75. Duchêne, S., Duchêne, D., Holmes, E. C. & Ho, S. Y. W. The performance of the date-randomization test in phylogenetic analyses of time-structured virus data. *Mol. Biol. Evol.* **32**, 1895–1906 (2015).
76. Rieux, A. & Khatchikian, C. E. Tipdatingbeast: an R package to assist the implementation of phylogenetic tip-dating tests using beast. *Mol. Ecol. Resour.* <https://doi.org/10.1111/1755-0998.12603> (2017).
77. Raftery, A. E. Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika* <https://doi.org/10.1093/biomet/83.2.251> (1996).
78. Ho, S. Y. W. & Shapiro, B. Skyline-plot methods for estimating demographic history from nucleotide sequences. *Mol. Ecol. Resour.* <https://doi.org/10.1111/j.1755-0998.2011.02988.x> (2011).
79. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* (2018) doi:<https://doi.org/10.1093/sysbio/syy032>.

Acknowledgements

We thank the Herbarium of the Muséum national d'Histoire naturelle (Paris, France) for allowing us to sample and perform destructive analysis on the *Manihot glaziovii* historical specimen P04808771. Collection of any plant material used in this study complies with institutional, national, and international guidelines. This work was financially supported by l'Agence Nationale pour la Recherche (JCJC MUSEOBACT contrat ANR-17-CE35-0009-01), the European Regional Development Fund (ERDF contract GURDT I2016-1731-0006632), Région Réunion, the French Agropolis Foundation (Labex Agro – Montpellier, E-SPACE Project Number 1504-004, MUSEOVIR project number 1600-004), the SYNTHESYS Project <http://www.synthesys.info/> (Grants GB-TAF-6437 and GB-TAF-7130) which is financed by European Community Research Infrastructure Action under the FP7 "Capacities" Program & CIRAD/AI-CRESI- 3/2016. PhD of P.C. was co-funded by ED 227, Muséum national d'Histoire naturelle et Sorbonne Université, French Ministry of Higher Education, Research and Innovation, France. Computational work was performed on the CIRAD - UMR AGAP HPC data center of the south green bioinformatics platform (<http://www.southgreen.fr/>). This work was conducted on the Plant Protection Platform (3P, IBISA). The authors thank the Centre Hospitalier Sud Réunion and Dr Julien Jaubert for hosting us in their laboratory, Denis Filloux, Philippe Roumagnac, Mikhail Pooggin, François Balloux, Violaine Llaurens, Regis Debruyne for fruitful discussions during this study and Dr. James Legg for his assistance with the history of the cassava mosaic disease in Africa.

Author contributions

This project was globally led by J.-M.L., N.B. & A.R. M.G. provided historical material and insights on herbarium specimen sampling. S.S. performed the wetlab processing of the historic sample under the supervision of N.B., P.L. & J.-M.L. A.R., P.C., A.D., S.S., D.M., N.B. & J.-M.L. analyzed the data and performed genetic analyses. A.R. & J.-M.L. wrote the first draft and all authors contributed to the final version.

Competing interest

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-00518-w>.

Correspondence and requests for materials should be addressed to A.R. or J.-M.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



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

© The Author(s) 2021

Discussion générale

Cette thèse avait pour objectif premier d'évaluer l'apport de l'ADN ancien dans l'étude de l'émergence et de l'évolution de pathogènes des cultures en prenant *Xanthomonas citri* pathovar *citri* (*Xci*), bactérie responsable du chancre asiatique des agrumes, comme modèle biologique d'étude. Pour ce faire, nous avons tout d'abord développé et optimisé la démarche au laboratoire et en bioinformatique afin d'obtenir et caractériser des données moléculaires issues d'échantillons historiques (chapitre 2). Nous avons ensuite montré l'intérêt des différents paléogénomes reconstruits dans l'étude de l'histoire évolutive de *Xci* à deux échelles évolutives distinctes (chapitres 3 et 4) ainsi que pour celle d'un autre agent pathogène (chapitre 5). Nous discutons ici des apports et limites des résultats de cette thèse et proposons des perspectives de recherche pour poursuivre et améliorer ces travaux.

Apport de l'ADN ancien dans l'étude de l'émergence et de l'évolution de pathogènes de cultures

Les résultats obtenus durant cette thèse illustrent le triple intérêt des données moléculaires issues d'échantillons historiques dans l'étude d'agents pathogènes de cultures.

Tout d'abord, un spécimen d'herbier infecté par un pathogène peut permettre d'actualiser la distribution spatio-temporelle d'une maladie, incluant la date de sa première description. Ainsi, l'origine du chancre asiatique des agrumes dont la plus ancienne date de description au champ remontait en 1899 au Japon (Kuhara, 1978) a été réévaluée par Fawcett & Jenkins (1933) sur la base de symptômes caractéristiques de la maladie observés sur des échantillons d'herbier de Java (1842-1844) et d'Inde (1827-1831). Cependant, l'identification moléculaire (par méthodes de diagnostic ou par séquençage) est primordiale pour confirmer la présence d'un pathogène suspectée sur la seule base de symptômes. Ainsi, HERB_1845 qui constitue le génome le plus ancien de *Xci* que nous avons reconstruit à partir d'un spécimen échantillonné également à Java en 1845 est en accord avec les données historiques. D'une autre part, le génome de *Xanthomonas citri* pv. *bilvae* (*Xcb*) que nous avons reconstruit à partir d'un spécimen d'herbier d'Inde datant de 1867 (chapitre 2.1, page 70) nous permet de repousser la date de première description de ce pathogène et de la maladie qu'il cause de près d'un siècle, la précédente datant de 1953 en Inde (Patel, Allayyanavaramath and Kulkarni, 1953).

Ensuite, les échantillons historiques permettent d'augmenter l'intervalle de temps couvert par l'échantillonnage et par là même d'augmenter la probabilité de détecter des populations dites en « évolution mesurable » (Spyrou *et al.*, 2019; Ho and Duchêne, 2020; Duchêne *et al.*, 2020). Dans les situations où un signal temporel est déjà présent sur un jeu de données composé de génomes « modernes » uniquement, nos résultats sur l'analyse de l'émergence de *Xci* dans les îles du sud-

ouest de l'océan Indien (chapitre 3) ou de l'ACMV en Afrique (chapitre 5) indiquent que l'ajout de génomes historiques permet d'affiner les estimations de paramètres évolutifs, avec des intervalles de confiance réduits et des dates estimées plus conformes avec les relevés d'occurrences historiques. Pour l'analyse de l'histoire évolutive globale de *Xci* (chapitre 4), l'intégration de génomes historiques a apporté le signal temporel requis pour réaliser des inférences par *tip-dating*, une approche plus robuste et faisant moins d'*a priori* que les autres méthodes de calibration, nous permettant d'émettre de nouveaux *scenarii* associés à l'histoire évolutive passée du pathogène en précisant les dates associées à sa diversification et à son origine.

Enfin, l'analyse comparative de génomes historiques et modernes offre théoriquement l'opportunité de mieux comprendre l'évolution des génomes au cours du temps, comme illustré avec l'augmentation de ploïdie et la variation en effecteurs de pathogénicité mise en évidence chez *Phytophthora infestans*, l'agent causal du mildiou de la pomme de terre (Yoshida *et al.*, 2013). Les analyses de ce type menées dans le cadre de cette thèse n'ont pas permis d'identifier des différences majeures dans le contenu en gènes ou en facteurs de virulence entre génomes modernes et historiques de *Xci*. Cependant, comme discuté ci-après, ce résultat a pu être biaisé par notre choix d'utiliser un génome de référence pour reconstruire et analyser les génomes historiques.

Caractérisation de la dégradation *post-mortem* de l'ADN d'une bactérie de plante

L'étude de l'ADN ancien a permis d'identifier des patrons moléculaires caractéristiques de la dégradation *post-mortem* tels que la faible quantité d'ADN extractible, la réduction de la taille des fragments ADN et la présence de lésions causées par désamination (Pääbo *et al.*, 2004; Dabney, Meyer and Pääbo, 2013; Kistler *et al.*, 2020). Ces différents patrons de dégradation *post-mortem* de l'ADN sont classiquement utilisés pour authentifier les ADNs anciens (Rohland *et al.*, 2015; Kistler *et al.*, 2017) et nous ont permis de confirmer la nature historique des génomes reconstruits durant cette thèse. De plus, l'utilisation concomitante de témoins « blancs » et d'échantillons non hôtes de *Xci* (spécimens d'herbier de *Coffea arabica*) nous a permis de vérifier l'absence de contamination croisée entre échantillons au laboratoire.

En tentant d'identifier des facteurs structurant les patrons de dégradation *post-mortem* de l'ADN de nos génomes historiques de *Xci*, nous avons identifié un effet significatif du temps sur le taux de désamination de l'ADN et ce à partir de 13 échantillons seulement. Cette observation a été possible grâce à une analyse statistique réalisée sur les décomptes de lectures de séquençage portant l'information de mésincorporation nucléotidique plutôt que sur des mesures transformées (*e.g.* le taux de désamination), nécessitant un plus large échantillonnage comme illustré par l'étude de Weiß *et al.*

(2016). Grâce à cette approche, nous avons également pu observer un effet significatif du protocole sur la mesure de ces patrons de dégradation, soulignant l'importance de la standardisation expérimentale afin de pouvoir correctement comparer les résultats entre eux.

Enfin, nous avons mis en évidence des profils de désamination différentiels entre chromosome et plasmides chez *Xci*. L'existence de telles dissimilitudes entre différents supports génétiques a précédemment été décrite entre ADN nucléaire et ADN chloroplastique chez trois espèces de plantes (Weiß *et al.*, 2016) mais c'est à notre connaissance la première description d'un tel patron entre ADN chromosomique et plasmidique chez une bactérie. Il serait intéressant de tester la conservation de ce patron chez d'autres espèces bactériennes ainsi que d'investiguer les facteurs qui pourraient en être à l'origine.

Limites de la reconstruction des génomes par alignement sur séquences de référence

L'approche par alignement des lectures de séquençage sur séquences de référence est communément utilisée pour reconstruire les génomes historiques (Hofreiter *et al.*, 2014). Cette dernière permet notamment la mesure de patrons de dégradation assurant leur authentification avec des outils comme mapDamage2 (Jónsson *et al.*, 2013), ainsi que la correction ou le masquage des nucléotides ayant subi des dégradations. Bien que particulièrement bien adaptée à la petite taille des lectures issus d'acides nucléiques dégradés, cette approche ne permet cependant ni l'identification de réarrangements génomiques entre le ou les génomes historiques et celui servant de référence, ni l'identification de matériel génétique présent chez le génome historique mais absent du génome de référence. L'importance des îlots de pathogénicité, des réarrangements chromosomiques et de la réduction génomique dans le déterminisme de la virulence chez les bactéries (Hacker and Kaper, 2000; Jackson *et al.*, 2011; Murray *et al.*, 2021) implique de devoir les identifier par des méthodes alternatives permettant une reconstruction optimale des paléogénomes.

Un premier moyen de contourner ce biais de méthode est d'augmenter le nombre et la diversité des séquences de référence utilisées (Valiente-Mullor *et al.*, 2021) en élaborant une base de données de séquences de références, gènes ou génomes, issues des différentes lignées évolutives au sein de l'espèce d'intérêt (mais aussi d'espèces proches). Cette stratégie, utilisée dans l'analyse de la variation en gènes impliqués dans la virulence de *Xci* (chapitre 4) nous a, par exemple, permis de confirmer que le gène *xopAG* est exclusif aux souches du clade du pathotype A^W et absent des clades A et A* (Escalon *et al.*, 2013; Gordon *et al.*, 2015; Patané *et al.*, 2019), y compris pour les souches historiques. Une telle approche pourrait être appliquée pour l'ensemble des gènes en tentant de représenter le pangénome

de l'espèce, comme précédemment réalisé par Richard *et al.* (2020) sur une collection de génomes modernes de *Xci*.

Une seconde approche consiste à s'affranchir des séquences de référence en réalisant un assemblage *de novo* du génome historique, permettant d'assembler les lectures chevauchantes entre elles par similarité de séquence en morceaux (contigs) plus longs. Bien que nous n'ayons pas tenté de réaliser de telles analyses sur nos lectures obtenues à partir des échantillons d'herbiers, la performance de ces dernières est connue pour être altérée dans les 3 conditions suivantes : 1) la présence de portions génomiques répétées, 2) le mélange de lectures issues de plusieurs unités taxonomiques (échantillon dit métagénomique) et 3) la petite taille des lectures (mais également la variation de cette dernière) et leur faible profondeur (Nagarajan and Pop, 2013; Seitz and Nieselt, 2017). D'une manière générale, les jeux de données paléogénomiques sont souvent caractérisés par les deux dernières conditions, même si le mélange de plusieurs unités taxonomiques peut être en partie évité grâce à l'utilisation de méthodes de capture lors de la préparation des bibliothèques. Comme illustré par l'analyse de la composition taxonomique d'HERB_1937 présentée dans le chapitre 3, les données que nous avons générées proviennent effectivement d'un mélange d'ADNs de plusieurs organismes qui pourraient à tort s'assembler entre eux à cause de motifs semblables ou sur des régions homologues conservées entre espèces, formant ainsi des séquences chimériques. Une possibilité pourrait être de pré-traiter le jeu de données en éliminant les lectures s'alignant à des génomes de références (celles de l'hôte ou de l'homme, par exemple), comme précédemment proposé (Lischer and Shimizu, 2017). Afin de dépasser ces limites, nombreuses autres optimisations ont récemment été apportées aux algorithmes d'assemblage *de novo*, bénéficiant parfois d'astuces bioinformatiques, pour gagner en efficacité (Seitz and Nieselt, 2017), ce qui a récemment permis la reconstruction de génomes microbiens à partir de données paléogénomiques en s'affranchissant de séquences de référence (Brealey *et al.*, 2020; Granehäll *et al.*, 2021; Wibowo *et al.*, 2021). Dans ce contexte, il paraît important de souligner qu'un outil de détection et de caractérisation de la dégradation d'acides nucléiques anciens à partir de données d'assemblage *de novo* a récemment été développé (Borry *et al.*, 2021). Au vu de la très bonne profondeur de la plupart des génomes anciens reconstruits dans ce travail (maximum de 96,2X, voir Tableau 2.3.A.), l'utilisation de pipelines optimisés d'assemblage *de novo* semble constituer une perspective prometteuse pour mieux caractériser l'évolution temporelle des génomes de *Xci* dans leur ensemble, même si la reconstruction des gènes déterminants du pouvoir pathogène TAL, connus pour être composés de longs motifs répétés me paraît à ce jour encore trop difficile *via* cette approche.

Avantages et contraintes de l'approche moléculaire non ciblée

Lors de cette thèse, nous avons privilégié une approche « *shotgun* » au laboratoire. Afin d'obtenir de l'ADN du pathogène d'intérêt, nous avons extrait les acides nucléiques à partir de tissus présentant des symptômes de la maladie mais n'avons pas cherché à sélectionner son ADN par capture par sondes spécifiques. La capture permet l'enrichissement de l'ADN ciblé mais *via* l'élaboration des sondes (Bahcall, 2013) qui induisent un risque de manquer le matériel génétique propre au génome historique et absent des séquences utilisées pour leur composition, un biais similaire à celui de l'utilisation de séquences de références tel que décrit plus haut. Afin de déterminer si, une fois convertis en librairie, les acides nucléiques extraits contiennent une proportion d'ADN endogène suffisante, nous avons opté pour une stratégie basée sur un criblage par séquençage de faible profondeur. Bien que cette stratégie se soit montrée efficace dans le cadre de ce travail (nous avons réussi à reconstruire des génomes historiques complets de *Xci* dès 0,8% d'ADN endogène), d'autres méthodes de criblages pourront dans le futur être adoptées, comme la détection d'un court fragment diagnostique de l'espèce recherchée par qPCR (Robène *et al.*, 2020). Une telle approche serait par ailleurs plus adaptée pour des pathogènes dont les symptômes ne sont pas spécifiques et peuvent être facilement confondus avec des signes de carences ou de chloroses, comme cela est le cas avec la bactérie pathogène *Xylella fastidiosa* (Sicard *et al.*, 2018). Dans ce contexte, notre protocole de purification des acides nucléiques nous a permis de rechercher la présence de *Xanthomonas citri* pathovar *fuscans* (bactérie responsable de la gousse commune du haricot) par qPCR à partir de semences et de gousses historiques, deux types de substrats également conservés dans certaines collections d'herbier mais que nous n'avions pas manipulés auparavant. L'étude, en collaboration avec l'UMR Pathologies Végétales (PaVé), INRAE, est prometteuse et a permis de reconstruire un premier génome historique en cours d'analyse.

De plus, les échantillons d'herbier étant une ressource historique finie, que nous détruisons en partie lors de l'échantillonnage, l'approche moléculaire « non-ciblée » permet de générer des données moléculaires provenant d'une large diversité d'organismes, comme illustré par l'analyse métagénomique de l'échantillon d'herbier de Maurice de 1937 de *Citrus* sp. (chapitre 3). Ainsi les données brutes que nous avons générées pourraient être utilisées de nombreuses autres manières. Tout d'abord, tout comme nous avons cherché à caractériser le contenu en gènes liés à la pathogénie chez *Xci* (Escalon *et al.*, 2013; Gordon *et al.*, 2015; Patané *et al.*, 2019) et estimer son évolution au cours du temps (chapitre 4), il serait intéressant d'étudier les gènes de la plante impliqués dans le développement de la maladie. Parmi ces derniers, certains sont directement activés par les facteurs de transcription bactériens TALEs (identifiés dans notre premier échantillon historique, chapitre 3), *via* des éléments de réponse spécifiques EBEs (*effector binding elements*). Différents mécanismes de

résistance (perte d'EBEs, nouveaux gènes-cibles de résistance ou interférence directe avec les TALEs) ont été décrits chez la plante (Büttner, 2016). L'ensemble des promoteurs des gènes-cibles des TALEs (préalablement caractérisés chez des gènes homologues modernes) pourrait être analysé *in silico* dans nos échantillons historiques (présence/absence d'EBEs...), afin de mieux comprendre la mise en place de ce système de régulation au cours du temps.

De manière moins focalisée sur le pathogène, des séquences d'intérêt, voire le génome, de la plante hôte pourraient être reconstruits afin de mesurer la diversité génétique passée de *Citrus* et dresser les relations de parenté entre variétés anciennes et modernes. En effet, les agrumes ont une généalogie complexe, une grande capacité d'hybridation et une longue histoire de reclassification (Ollitrault, Curk and Krueger, 2020) (Figure 6.A), rendant parfois difficile l'identification, spécifique ou variétale, des échantillons. Ainsi, les échantillons d'herbier, de plus en plus utilisés pour générer des marqueurs génétiques diagnostiques dits barcodes ADN, permettraient d'améliorer les bases de données de référence des *Citrus* mais également de vérifier l'identification d'un spécimen (Kistler *et al.*, 2020).

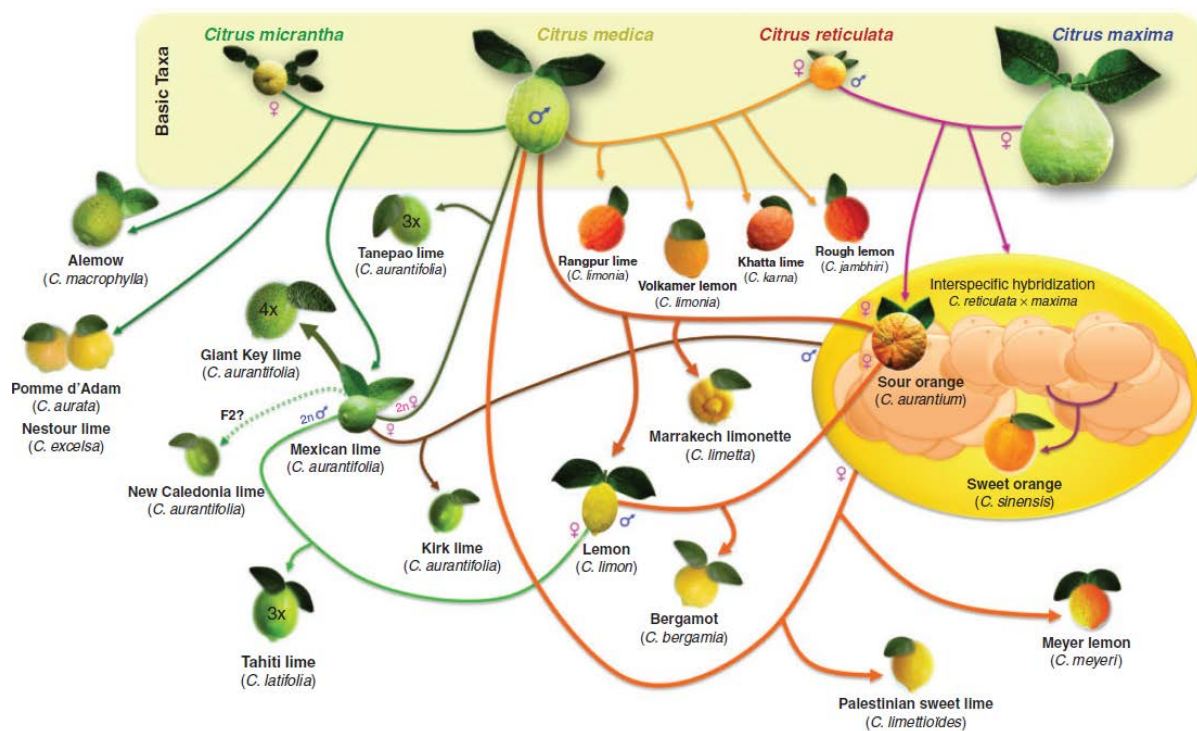


Figure 6.A. Origine généalogique des principaux groupes variétaux des *Citrus* (Curk *et al.*, 2016, Figure 8).

Dans le cadre d'une collaboration avec Patrick Ollitrault et Franck Curk (UMR Agap, CIRAD), nous avons cherché à exploiter la part d'ADN de plante de nos données de séquençage historiques dans le but d'identifier des SNPs diagnostiques ainsi que des mosaïques génomiques interspécifiques chez les différentes espèces et variétés connues et, si nécessaire, de redéfinir ces marqueurs moléculaires pour

prendre en compte la diversité génétique des échantillons historiques. Leur analyse d'un premier échantillon a permis d'identifier un échantillon de *Citrus* sp. comme *Citrus medica* avec une introgression de *Citrus micrantha* pouvant correspondre à une variété moderne de *Chinese citron*. L'analyse des 12 autres échantillons d'herbier est en cours. Outre cette validation des marqueurs moléculaires et révision de l'identification d'échantillons historiques, ces derniers pourraient aider à dater des événements de croisement entre lignées et amener une meilleure visualisation de l'histoire des *Citrus* durant leur domestication ces quatre derniers millénaires (Talon, Caruso and Gmitter Jr., 2020).

Quel(s) apport(s) des 250+ autres échantillons d'agrumes symptomatiques collectés mais non utilisés dans l'étude?

Au cours de cette thèse, j'ai manipulé une quarantaine d'échantillons symptomatiques d'herbier sur les plus de 300 constituant notre collection au laboratoire, et généré 13 génomes complets de *Xci* et un de *Xcb*. Il reste ainsi une grande partie de notre échantillonnage historique que nous n'avons pas exploré ou analysé et qui pourrait, pour différentes raisons, aider à l'amélioration de l'étude de *Xci*.

Tout d'abord, les 13 génomes historiques de *Xci* générés durant cette thèse tombent dans le clade du pathotype A. Des échantillons historiques des clades A^W et A* pourraient permettre de mieux représenter la diversité génétique au sein de ces deux clades et à mesurer leur évolution au cours du temps, notamment sur des gènes d'intérêt comme *xopAG*. Ils pourraient également apporter un signal temporel spécifiquement dans leur clades respectifs, ce qui permettrait de réaliser des inférences indépendantes au niveau de chacun des clades et ainsi tester l'existence de grosses différences dans les taux de substitutions pouvant être liées à des changements de pression de sélection et/ou d'adaptation du pathogène à ses hôtes au cours du temps. De telles différences de taux inférés sur deux clades d'une même espèce et au niveau de l'espèce ont été mesurées chez *Salmonella enterica* serovar Paratyphi A (Duchêne *et al.*, 2016), une bactérie responsable de la fièvre typhoïde chez l'homme. Afin d'augmenter nos chances d'obtenir des génomes historiques A^W et A*, il faudrait privilégier les échantillons de *Citrus aurantiifolia* et *Citrus macrophylla*, hôtes de ces deux clades, provenant d'Inde, du Bangladesh, du Népal et du Pakistan, leurs zones d'origine inférées (Patané *et al.*, 2019; chapitre 4).

Ensuite, nous pourrions utiliser le génome *Xcb* datant de 1867 comme outgroup dans les analyses phylogénétiques de *Xci* et tester s'il apporte un signal temporel hors du clade *Xci*. Si cela est le cas, une telle analyse donnerait l'opportunité de tester empiriquement si (et comment) l'ajout d'un génome historique « externe » à un groupe d'intérêt (ici, l'ingroup, *Xci*) influence la calibration de l'horloge

moléculaire de ce dernier. Bien que l'effet de la position des points de calibration (internes vs externes) par rapport à un groupe d'intérêt ait déjà été testé dans le cadre de calibrations par « *node-dating* » (Ho *et al.*, 2008), cela n'a, à ma connaissance, jamais été formellement évalué pour celles en « *tip-dating* » utilisant des génomes historiques.

Enfin, les échantillons symptomatiques historiques pourraient également permettre de tester l'hypothèse posée par Patané *et al.* (2019) selon laquelle *Xci* aurait émergé sur *Citrus* par saut d'hôte à partir d'une plante de la famille des *Fabaceae*. Pour cela, il pourrait être intéressant de rechercher des spécimens historiques de *Fabaceae* provenant du sous-continent indien et présentant des symptômes de bactérioses. Alternativement, si des spécimens historiques de *Rutaceae* non *Citrus*, également présents dans cette même région, se trouvaient être infectés par une bactérie branchant au clade *Xci* avant *Xanthomonas. citri* pv. *cajani*, plus proche parent actuellement connu, alors cela permettrait de proposer un nouveau modèle n'impliquant pas directement les *Fabaceae*.

Conclusions et perspectives

Formellement débutée dans les années 1980, la paléogénomique demeure à ce jour une science relativement récente, en plein essor depuis l'avènement des technologies de séquençage à haut débit, présentant aujourd'hui une multitude d'applications possibles et aspirant à la transdisciplinarité. Récemment, des optimisations au laboratoire ont permis d'améliorer l'accessibilité aux échantillons anciens et la qualité des études de paléogénomique. Par exemple, le protocole BEST, comparé à d'autres protocoles de construction de bibliothèques Illumina, limite la perte de matériel et permet l'obtention d'une complexité de bibliothèque supérieure tout en diminuant les temps et coûts de préparation (Carøe *et al.*, 2018). De futurs développements permettront peut-être de limiter au maximum la destruction des échantillons historiques, comme avec le protocole récemment proposé par Shepherd (2017) pour extraire de l'ADN à partir de gomme frottée sur spécimens d'herbiers. De telles améliorations au laboratoire pourraient aller de pair avec l'élaboration de nouveaux pipelines bioinformatiques de reconstruction des génomes anciens, notamment pour la reconstruction *de novo* de génomes historiques. Finalement, les inférences phylogénétiques devront probablement être complétées par le développement et l'application de nouveaux modèles de génétique des populations permettant une représentation plus complexe et réaliste des processus évolutifs ayant affecté les populations du passé, notamment *via* une modélisation plus explicite de leur isolation et admixture génétique (Hofreiter *et al.*, 2014). En épidémiologie moléculaire, l'application de la paléogénomique permet une meilleure compréhension de l'histoire évolutive des pathogènes ainsi que des risques qu'ils posent. Les taux de substitution, taux de migration, moyens de propagation, influence des activités humaines dans leur transport... sont autant de facteurs dont la compréhension est requise

pour améliorer notre capacité à prédire la dynamique des maladies actuelles et futures ainsi que leur gestion. Grâce à la richesse et l'accessibilité des collections d'herbiers, la paléogénomique peut ainsi contribuer à leur valorisation scientifique.

Bibliographie

- Anderson, P.K. *et al.* (2004) 'Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers', *Trends in Ecology & Evolution*, 19(10), pp. 535–544. doi:10.1016/j.tree.2004.07.021.
- Andrades Valtueña, A. *et al.* (2017) 'The Stone Age plague and its persistence in Eurasia', *Current Biology*, 27(23), pp. 3683–3691.e8. doi:10.1016/j.cub.2017.10.025.
- Antonovics, J. *et al.* (2003) 'Herbarium studies on the distribution of anther-smut fungus (*Microbotryum violaceum*) and *Silene* species (*Caryophyllaceae*) in the eastern United States', *American Journal of Botany*, 90(10), pp. 1522–1531. doi:10.3732/ajb.90.10.1522.
- Arenas, M. (2015) 'Trends in substitution models of molecular evolution', *Frontiers in Genetics*, 6, p. 9.
- Arning, N. and Wilson, D.J. (2020) 'The past, present and future of ancient bacterial DNA', *Microbial Genomics*, 6(7). doi:10.1099/mgen.0.000384.
- Arriola, L.A., Cooper, A. and Weyrich, L.S. (2020) 'Palaeomicrobiology: application of ancient DNA sequencing to better understand bacterial genome evolution and adaptation', *Frontiers in Ecology and Evolution*, 8, p. 40. doi:10.3389/fevo.2020.00040.
- Aubert, B. (2014) 'Vergers de la Réunion et de l'Océan Indien', in *Hommes et fruits en pays du Sud*. Mémoire des Hommes. (CIRAD), pp. 111–165.
- Ausubel, F.M. *et al.* (2003) *Current protocols in molecular biology*. John Wiley & Sons, New York.
- Awadalla, P. (2003) 'The evolutionary genomics of pathogen recombination', *Nature Reviews Genetics*, 4(1), pp. 50–60. doi:10.1038/nrg964.
- Bacigalupe, R. *et al.* (2019) 'A multihost bacterial pathogen overcomes continuous population bottlenecks to adapt to new host species', *Science Advances*, 5(11), pp. 1–14. doi:10.1126/sciadv.aax0063.
- Bahcall, O. (2013) 'Capturing ancient DNA', *Nature Genetics*, 45(12), pp. 1417–1417. doi:10.1038/ng.2842.
- Bakker, F.T. (2017) 'Herbarium genomics: skimming and plastomics from archival specimens', *Webbia*, 72(1), pp. 35–45. doi:10.1080/00837792.2017.1313383.
- Balter, M. (2007) 'Seeking agriculture's ancient roots', *Science*, 316(5833), pp. 1830–1835. doi:10.1126/science.316.5833.1830.
- de Barros Damgaard, P. *et al.* (2018) '137 ancient human genomes from across the Eurasian steppes', *Nature*, 557(7705), pp. 369–374. doi:10.1038/s41586-018-0094-2.
- 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), pp. 303–310. doi:10.1111/mpp.12284.
- Beaumont, M.A., Zhang, W. and Balding, D.J. (2002) 'Approximate Bayesian computation in population genetics', *Genetics*, 162(4), pp. 2025–2035. doi:10.1093/genetics/162.4.2025.
- Bernades, M.F.F. *et al.* (2015) 'Impact of pesticides on environmental and human health', in *Toxicology studies: cells, drugs and environment*. Rijeka, Croatia (InTech), pp. 195–233.
- Biek, R. *et al.* (2015) 'Measurably evolving pathogens in the genomic era', *Trends in Ecology & Evolution*, 30(6), pp. 306–313. doi:10.1016/j.tree.2015.03.009.
- Bieker, V.C. *et al.* (2020) 'Metagenomic analysis of historical herbarium specimens reveals a

- postmortem microbial community', *Molecular Ecology Resources*, pp. 1–14. doi:10.1111/1755-0998.13174.
- Bieker, V.C. and Martin, M.D. (2018) 'Implications and future prospects for evolutionary analyses of DNA in historical herbarium collections', *Botany Letters*, 165(3–4), pp. 409–418. doi:10.1080/23818107.2018.1458651.
- Boch, J. and Bonas, U. (2010) '*Xanthomonas* AvrBs3 family-type III effectors: discovery and function', *Annual Review of Phytopathology*, 48(1), pp. 419–436. doi:10.1146/annurev-phyto-080508-081936.
- Borry, M. *et al.* (2021) 'PyDamage: automated ancient damage identification and estimation for contigs in ancient DNA *de novo* assembly', *PeerJ*, 9, p. e11845. doi:10.7717/peerj.11845.
- Bos, K.I. *et al.* (2011) 'A draft genome of *Yersinia pestis* from victims of the Black Death', *Nature*, 478(7370), pp. 506–510. doi:10.1038/nature10549.
- Bos, K.I. *et al.* (2014) 'Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis', *Nature*, 514(7523), pp. 494–497. doi:10.1038/nature13591.
- Bos, K.I. *et al.* (2016) 'Eighteenth century *Yersinia pestis* genomes reveal the long-term persistence of an historical plague focus', *eLife*, 5, pp. 1–11.
- Brealey, J.C. *et al.* (2020) 'Dental calculus as a tool to study the evolution of the mammalian oral microbiome', *Molecular Biology and Evolution*. Edited by D. Falush, 37(10), pp. 3003–3022. doi:10.1093/molbev/msaa135.
- Briggs, A.W. *et al.* (2007) 'Patterns of damage in genomic DNA sequences from a Neandertal', *Proceedings of the National Academy of Sciences*, 104(37), pp. 14616–14621. doi:10.1073/pnas.0704665104.
- Brun, J. (1971) 'Le chancre bactérien des *Citrus*', *Fruits*, 26(7–8), pp. 533–540.
- Bui Thi Ngoc, L. *et al.* (2009) 'From local surveys to global surveillance: three high throughput genotyping methods for the epidemiological monitoring of *Xanthomonas citri* pv. *citri* pathotypes', *Applied and Environmental Microbiology*, 75(4), pp. 1173–1184. doi:10.1128/AEM.02245-08.
- Buonagurio, D.A. *et al.* (1986) 'Evolution of human influenza A viruses over 50 years: rapid, uniform rate of change in NS gene', *Science*, 232(4753), pp. 980–982. doi:10.1126/science.2939560.
- Büttner, D. (2016) 'Behind the lines—actions of bacterial type III effector proteins in plant cells', *FEMS Microbiology Reviews*, 40(6), pp. 894–937. doi:10.1093/femsre/fuw026.
- Carøe, C. *et al.* (2018) 'Single-tube library preparation for degraded DNA', *Methods in Ecology and Evolution*. Edited by S. Johnston, 9(2), pp. 410–419. doi:10.1111/2041-210X.12871.
- Chakravarti, B.P. *et al.* (1984) 'A bacterial spot of bael (*Aegle marmelos* Correa) in Rajasthan and a revived name of the bacterium', *Current Science*, 53, pp. 48–489.
- Chan, J.Z. *et al.* (2013) 'Metagenomic analysis of tuberculosis in a mummy', *New England Journal of Medicine*, 369(16), pp. 289–290. doi:10.1056/nejmc1302295.
- Croucher, N.J. and Didelot, X. (2015) 'The application of genomics to tracing bacterial pathogen transmission', *Current Opinion in Microbiology*, 23, pp. 62–67. doi:10.1016/j.mib.2014.11.004.
- Curk, F. *et al.* (2016) 'Phylogenetic origin of limes and lemons revealed by cytoplasmic and nuclear markers', *Annals of Botany*, 117(4), pp. 565–583. doi:10.1093/aob/mcw005.
- Dabney, J., Meyer, M. and Pääbo, S. (2013) 'Ancient DNA damage', *Cold Spring Harbor Perspectives in Biology*, 7, pp. 1–8. doi:10.1101/cshperspect.a012567.

- Darlu, P. and Tassy, P. (1993) *La reconstruction phylogénétique: concepts et méthodes*. Paris: Masson.
- De Bruyn, A. *et al.* (2016) 'Divergent evolutionary and epidemiological dynamics of cassava mosaic geminiviruses in Madagascar', *BMC Evolutionary Biology*, 16(182), pp. 1–21. doi:10.1186/s12862-016-0749-2.
- Devault, A.M. *et al.* (2014) 'Second-pandemic strain of *Vibrio cholerae* from the Philadelphia cholera outbreak of 1849', *New England Journal of Medicine*, 370(4), pp. 334–340. doi:10.1056/NEJMoa1308663.
- Devault, A.M. *et al.* (2017) 'A molecular portrait of maternal sepsis from Byzantine Troy', *eLife*, 6, pp. 1–31. doi:10.7554/eLife.20983.
- Didelot, X. and Maiden, M.C.J. (2010) 'Impact of recombination on bacterial evolution', *Trends in Microbiology*, 18(7), pp. 315–322. doi:10.1016/j.tim.2010.04.002.
- Didelot, X. and Wilson, D.J. (2015) 'ClonalFrameML: efficient inference of recombination in whole bacterial genomes', *PLoS Computational Biology*, 11(2), pp. 1–18. doi:10.1371/journal.pcbi.1004041.
- Doizy, A. *et al.* (2020) 'Phylostems: a new graphical tool to investigate temporal signal of heterochronous sequences at various evolutionary scales', *BioRxiv*, pp. 1–23. doi:10.1101/2020.10.19.346429.
- Domingues, M.N. *et al.* (2010) 'The *Xanthomonas citri* effector protein PthA interacts with citrus proteins involved in nuclear transport, protein folding and ubiquitination associated with DNA repair: PthA interaction with citrus proteins', *Molecular Plant Pathology*, p. no-no. doi:10.1111/j.1364-3703.2010.00636.x.
- van Dorp, L. *et al.* (2021) 'COVID-19, the first pandemic in the post-genomic era', *Current Opinion in Virology*, 50, pp. 40–48. doi:10.1016/j.coviro.2021.07.002.
- Drummond, A.J. *et al.* (2002) 'Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data', *Genetics*, 161(3), pp. 1307–1320. doi:10.1093/genetics/161.3.1307.
- Drummond, A.J. and Rambaut, A. (2007) 'BEAST: Bayesian evolutionary analysis by sampling trees', *BMC Evolutionary Biology*, 7(1–8), p. 214. doi:10.1186/1471-2148-7-214.
- Duan, Y.P. *et al.* (1999) 'Expression of a single, host-specific, bacterial pathogenicity gene in plant cells elicits division, enlargement, and cell death', *Molecular Plant-Microbe Interactions*, 12(6), pp. 556–560. doi:10.1094/MPMI.1999.12.6.556.
- Duchêne, S. *et al.* (2015) 'The performance of the date-randomization test in phylogenetic analyses of time-structured virus data', *Molecular Biology and Evolution*, 32(7), pp. 1895–1906. doi:10.1093/molbev/msv056.
- Duchêne, S. *et al.* (2016) 'Genome-scale rates of evolutionary change in bacteria', *Microbial Genomics*, 2(11), pp. 1–12. doi:10.1099/mgen.0.000094.
- Duchêne, S. *et al.* (2020) 'Bayesian evaluation of temporal signal in measurably evolving populations', *Molecular Biology and Evolution*. Edited by K. Crandall, 37(11), pp. 3363–3379. doi:10.1093/molbev/msaa163.
- Duchêne, S. *et al.* (2020) 'The recovery, interpretation and use of ancient pathogen genomes', *Current Biology*, 30(19), pp. R1215–R1231. doi:10.1016/j.cub.2020.08.081.
- Duggan, A.T. *et al.* (2016) '17th Century variola virus reveals the recent history of smallpox', *Current Biology*, 26(24), pp. 3407–3412. doi:10.1016/j.cub.2016.10.061.

- Dye, D.W. *et al.* (1980) 'International standards for naming pathovars of phytopathogenic bacteria and a list of pathovar names and pathotype strains', *Review of Plant Pathology*, 59(4), pp. 153–168.
- Ehrlich, M. *et al.* (1986) 'DNA cytosine methylation and heat-induced deamination', *Bioscience Reports*, 6(4), pp. 387–393. doi:10.1007/BF01116426.
- Escalon, A. *et al.* (2013) 'Variations in type III effector repertoires, pathological phenotypes and host range of *Xanthomonas citri* pv. *citri* pathotypes: type III effectors in *Xanthomonas citri* pv. *citri*', *Molecular Plant Pathology*, 14(5), pp. 483–496. doi:10.1111/mpp.12019.
- Estoup, A. and Guillemaud, T. (2010) 'Reconstructing routes of invasion using genetic data: why, how and so what?', *Molecular Ecology*, 19, pp. 4113–4130. doi:10.1111/j.1365-294X.2010.04773.x.
- Fawcett, H.S. and Jenkins, A.E. (1933) 'Records of citrus canker from herbarium specimens of the genus *Citrus* in England and the United States', *Phytopathology*, (23), pp. 820–824.
- Feldman, M. *et al.* (2016) 'A high-coverage *Yersinia pestis* genome from a 6th-century Justinianic plague victim', *Molecular Biology and Evolution*, 33(11), pp. 2911–2923. doi:10.1093/molbev/msw170.
- Felsenstein, J. (1996) 'Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods', *Methods in Enzymology*, 266, pp. 418–427. doi:10.1016/S0076-6879(96)66026-1.
- Fulton, T.L. (2012) 'Setting up an ancient DNA laboratory', in *Ancient DNA*. Totowa, NJ: Humana Press (Methods in Molecular Biology), pp. 1–11. doi:10.1007/978-1-61779-516-9.
- Gandon, S. *et al.* (2013) 'What limits the evolutionary emergence of pathogens?', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1610), pp. 1–10. doi:10.1098/rstb.2012.0086.
- Gelabert, P. *et al.* (2016) 'Mitochondrial DNA from the eradicated European *Plasmodium vivax* and *P. falciparum* from 70-year-old slides from the Ebro delta in Spain', *Proceedings of the National Academy of Sciences*, 113(41), pp. 11495–11500. doi:10.1073/pnas.1611017113.
- Glassing, A. *et al.* (2016) 'Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples', *Gut Pathogens*, 8(24), pp. 1–12. doi:10.1186/s13099-016-0103-7.
- Gochez, A.M. *et al.* (2018) 'Pacbio sequencing of copper-tolerant *Xanthomonas citri* reveals presence of a chimeric plasmid structure and provides insights into reassortment and shuffling of transcription activator-like effectors among *X. citri* strains', *BMC Genomics*, 19(16), pp. 1–14. doi:10.1186/s12864-017-4408-9.
- Goodwin, S.B., Cohen, B.A. and Fry, W.E. (1994) 'Panglobal distribution of a single clonal lineage of the Irish potato famine fungus', *Proceedings of the National Academy of Sciences*, 91(24), pp. 11591–11595. doi:10.1073/pnas.91.24.11591.
- Gordon, J.L. *et al.* (2015) 'Comparative genomics of 43 strains of *Xanthomonas citri* pv. *citri* reveals the evolutionary events giving rise to pathotypes with different host ranges', *BMC Genomics*, 16(1098), pp. 1–20. doi:10.1186/s12864-015-2310-x.
- Gottwald, T.R., Graham, J.H. and Schubert, T.S. (2002) 'Citrus canker: the pathogen and its impact', *Plant Health Progress*, pp. 1–34. doi:10.1094/PHP-2002-0812-01-RV.
- Graham, J.H. *et al.* (2004) '*Xanthomonas axonopodis* pv. *citri*: factors affecting successful eradication of citrus canker', *Molecular Plant Pathology*, 5(1), pp. 1–15. doi:10.1046/j.1364-3703.2004.00197.x.
- GraneHäll, L. *et al.* (2021) 'Metagenomic analysis of ancient dental calculus reveals unexplored diversity of oral archaeal *Methanobrevibacter*', *Microbiome*, 9(197), pp. 1–18. doi:10.1186/s40168-021-01132-

8.

Green, R.E. *et al.* (2010) 'A draft sequence of the neandertal genome', *Science*, 328(5979), pp. 710–722. doi:10.1126/science.1188021.

Grünwald, N.J. and Goss, E.M. (2011) 'Evolution and population genetics of exotic and re-emerging pathogens: novel tools and approaches', *Annual Review of Phytopathology*, 49(1), pp. 249–267. doi:10.1146/annurev-phyto-072910-095246.

Guellil, M. *et al.* (2018) 'Genomic blueprint of a relapsing fever pathogen in 15th century Scandinavia', *Proceedings of the National Academy of Sciences*, 115(41), pp. 10422–10427. doi:10.1073/pnas.1807266115.

Hacker, J. and Kaper, J.B. (2000) 'Pathogenicity Islands and the Evolution of Microbes', *Annual Review of Microbiology*, 54(1), pp. 641–679. doi:10.1146/annurev.micro.54.1.641.

Hagelberg, E., Hofreiter, M. and Keyser, C. (2015) 'Ancient DNA: the first three decades', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1660), p. 20130371. doi:10.1098/rstb.2013.0371.

Hartfield, M., Murall, C.L. and Alizon, S. (2014) 'Clinical applications of pathogen phylogenies', *Trends in Molecular Medicine*, 20(7), pp. 394–404. doi:10.1016/j.molmed.2014.04.002.

Hartung, J.S. *et al.* (1996) 'Rapid and sensitive detection of *Xanthomonas axonopodis* pv. *citri* by immunocapture and a nested-polymerase chain reaction assay', *Phytopathology*, 86(1), pp. 95–101.

Hartung, J.S. *et al.* (2015) 'History and diversity of *Citrus leprosis virus* recorded in herbarium specimens', *Phytopathology*, 105(9), pp. 1277–1284. doi:10.1094/PHYTO-03-15-0064-R.

Hasse, C.H. (1915) '*Pseudomonas citri*, the cause of citrus canker', *Journal of Agricultural Research*, (4), pp. 97–100.

Higuchi, R. *et al.* (1984) 'DNA sequences from the quagga, an extinct member of the horse family', *Nature*, 312(5991), pp. 282–284. doi:10.1038/312282a0.

Ho, S.Y.W. *et al.* (2008) 'The effect of inappropriate calibration: three case studies in molecular ecology', *PLoS ONE*. Edited by P. Bennett, 3(2), pp. 1–8. doi:10.1371/journal.pone.0001615.

Ho, S.Y.W. *et al.* (2011) 'Time-dependent rates of molecular evolution', *Molecular Ecology*, 20(15), pp. 3087–3101. doi:10.1111/j.1365-294X.2011.05178.x.

Ho, S.Y.W. and Duchêne, S. (2014) 'Molecular-clock methods for estimating evolutionary rates and timescales', *Molecular Ecology*, 23(24), pp. 5947–5965. doi:10.1111/mec.12953.

Ho, S.Y.W. and Duchêne, S. (2020) 'Dating the emergence of human pathogens', *Science*, 368(6497), pp. 1310–1311. doi:10.1126/science.abc5746.

Hofreiter, M. *et al.* (2014) 'The future of ancient DNA: technical advances and conceptual shifts', *BioEssays*, 37(3), pp. 284–293. doi:10.1002/bies.201400160.

Hu, Y. *et al.* (2014) '*Lateral organ boundaries 1* is a disease susceptibility gene for citrus bacterial canker disease', *Proceedings of the National Academy of Sciences*, 111(4), pp. 521–529. doi:10.1073/pnas.1313271111.

Ismail, S.I. *et al.* (2016) 'Ancestral state reconstruction infers phytopathogenic origins of sooty blotch and flyspeck fungi on apple', *Mycologia*, 108(2), pp. 292–302. doi:10.3852/15-036.

Jackson, R.W. *et al.* (2011) 'Bacterial pathogen evolution: breaking news', *Trends in Genetics*, 27(1), pp. 32–40. doi:10.1016/j.tig.2010.10.001.

- Jalan, N. *et al.* (2013) 'Comparative genomic and transcriptome analyses of pathotypes of *Xanthomonas citri* subsp. *citri* provide insights into mechanisms of bacterial virulence and host range', *BMC Genomics*, 14(1), p. 551. doi:10.1186/1471-2164-14-551.
- Jónsson, H. *et al.* (2013) 'mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters', *Bioinformatics*, 29(13), pp. 1682–1684. doi:10.1093/bioinformatics/btt193.
- Kahila Bar-Gal, G. *et al.* (2012) 'Tracing hepatitis B virus to the 16th century in a Korean mummy', *Hepatology*, 56(5), pp. 1671–1680. doi:10.1002/hep.25852.
- Kay, G.L. *et al.* (2014) 'Recovery of a medieval *Brucella melitensis* genome using shotgun metagenomics', *mBio*. Edited by P.S. Keim, 5(4), pp. 1–6. doi:10.1128/mBio.01337-14.
- Kay, G.L. *et al.* (2015) 'Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe', *Nature Communications*, 6(6717), pp. 1–9. doi:10.1038/ncomms7717.
- Kingman, J.F.C. (1982) 'The coalescent', *Stochastic Processes and their Applications*, 13(3), pp. 235–248. doi:10.1016/0304-4149(82)90011-4.
- Kistler, L. *et al.* (2017) 'A new model for ancient DNA decay based on paleogenomic meta-analysis', *Nucleic Acids Research*, 45(11), pp. 6310–6320. doi:10.1093/nar/gkx361.
- Kistler, L. *et al.* (2020) 'Ancient Plant Genomics in Archaeology, Herbaria, and the Environment', *Annual Review of Plant Biology*, 71(1), pp. 605–629. doi:10.1146/annurev-arplant-081519-035837.
- Krause-Kyora, B., Nutsua, M., *et al.* (2018) 'Ancient DNA study reveals HLA susceptibility locus for leprosy in medieval Europeans', *Nature Communications*, 9(1569), pp. 1–11. doi:10.1038/s41467-018-03857-x.
- Krause-Kyora, B., Susat, J., *et al.* (2018) 'Neolithic and medieval virus genomes reveal complex evolution of hepatitis B', *eLife*, 7, pp. 1–15. doi:10.7554/eLife.36666.
- Kreuze, J.F. *et al.* (2009) 'Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses', *Virology*, 388(1), pp. 1–7. doi:10.1016/j.virol.2009.03.024.
- Kuhara, S. (1978) 'Present epidemic status and control of the citrus canker disease (*Xanthomonas citri* (Hase) Dowson) in Japan', *Review of Plant Protection Research*, 11, pp. 132–142.
- Lang, P.L.M. *et al.* (2019) 'Using herbaria to study global environmental change', *New Phytologist*, 221(1), pp. 110–122. doi:10.1111/nph.15401.
- Langmead, B. and Salzberg, S.L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nature Methods*, 9(4), pp. 357–359. doi:10.1038/nmeth.1923.
- Larson, G. *et al.* (2014) 'Current perspectives and the future of domestication studies', *Proceedings of the National Academy of Sciences*, 111(17), pp. 6139–6146. doi:10.1073/pnas.1323964111.
- Lê, S., Josse, J. and Husson, F. (2008) 'FactoMineR : An R Package for Multivariate Analysis', *Journal of Statistical Software*, 25(1). doi:10.18637/jss.v025.i01.
- Lefevre, P. *et al.* (2010) 'The spread of tomato yellow leaf curl virus from the Middle East to the world', *PLoS Pathogens*. Edited by C. Fauquet, 6(10), pp. 1–12. doi:10.1371/journal.ppat.1001164.
- Lefevre, P. and Moriones, E. (2015) 'Recombination as a motor of host switches and virus emergence: geminiviruses as case studies', *Current Opinion in Virology*, 10, pp. 14–19. doi:10.1016/j.coviro.2014.12.005.
- Lemey, P. *et al.* (2014) 'Unifying viral genetics and human transportation data to predict the global

- transmission dynamics of human influenza H3N2', *PLoS Pathogens*. Edited by N.M. Ferguson, 10(2), pp. 1–10. doi:10.1371/journal.ppat.1003932.
- Lenormand, T. (2002) 'Gene flow and the limits to natural selection', *Trends in Ecology & Evolution*, 17(4), pp. 183–189. doi:10.1016/S0169-5347(02)02497-7.
- Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics*, 25(14), pp. 1754–1760. doi:10.1093/bioinformatics/btp324.
- Li, L.M., Grassly, N.C. and Fraser, C. (2014) 'Genomic analysis of emerging pathogens: methods, application and future trends', *Genome Biology*, 15(541), pp. 1–9. doi:10.1186/s13059-014-0541-9.
- Li, W. *et al.* (2007) 'Genetic diversity of citrus bacterial canker pathogens preserved in herbarium specimens', *Proceedings of the National Academy of Sciences*, 104(47), pp. 18427–18432. doi:10.1073/pnas.0705590104.
- Li, W., Brlansky, R.H. and Hartung, J.S. (2006) 'Amplification of DNA of *Xanthomonas axonopodis* pv. *citri* from historic citrus canker herbarium specimens', *Journal of Microbiological Methods*, 65(2), pp. 237–246. doi:10.1016/j.mimet.2005.07.014.
- Lin, H.-C., Hsu, S.-T. and Tzeng, K.-C. (2009) 'Histopathology and bacterial populations of atypical symptoms-inducing *Xanthomonas axonopodis* pv. *citri* strains in leaves of grapefruit and Mexican lime', *Plant Pathology Bulletin*, 18, pp. 125–134.
- Lindahl, T. and Andersson, A. (1972) 'Rate of chain breakage at apurinic sites in double-stranded deoxyribonucleic acid', *Biochemistry*, 11(19), pp. 3618–3623. doi:10.1021/bi00769a019.
- Lischer, H.E.L. and Shimizu, K.K. (2017) 'Reference-guided de novo assembly approach improves genome reconstruction for related species', *BMC Bioinformatics*, 18(474), pp. 1–12. doi:10.1186/s12859-017-1911-6.
- Lorenzen, E.D. *et al.* (2011) 'Species-specific responses of Late Quaternary megafauna to climate and humans', *Nature*, 479(7373), pp. 359–364. doi:10.1038/nature10574.
- Maixner, F. *et al.* (2016) 'The 5,300-year-old *Helicobacter pylori* genome of the Iceman', *Science*, 351(6269), pp. 162–165. doi:10.1126/science.aad2545.
- Malmstrom, C.M. *et al.* (2007) 'Barley yellow dwarf viruses (BYDVs) preserved in herbarium specimens illuminate historical disease ecology of invasive and native grasses', *Journal of Ecology*, 95(6), pp. 1153–1166. doi:10.1111/j.1365-2745.2007.01307.x.
- Marciniak, S. *et al.* (2016) '*Plasmodium falciparum* malaria in 1st –2nd century CE southern Italy', *Current Biology*, 26(23), pp. 1220–1222. doi:10.1016/j.cub.2016.10.016.
- Martin, M.D. *et al.* (2013) 'Reconstructing genome evolution in historic samples of the Irish potato famine pathogen', *Nature Communications*, 4(2172), pp. 1–7. doi:10.1038/ncomms3172.
- May, K.J. and Ristaino, J.B. (2004) 'Identity of the mtDNA haplotype(s) of *Phytophthora infestans* in historical specimens from the Irish potato famine', *Mycological Research*, 108(5), pp. 171–179. doi:10.1017/S0953756204009876.
- McCann, H.C. (2020) 'Skirmish or war: the emergence of agricultural plant pathogens', *Current Opinion in Plant Biology*, 56, pp. 147–152. doi:10.1016/j.pbi.2020.06.003.
- McDonald, B.A. (2004) 'Population genetics of plant pathogens', *American Phytopathological Society* [Preprint]. doi:10.1094/PHI-A-2004-0524-01.
- McDonald, B.A. and Linde, C. (2002) 'Pathogen population genetics, evolutionary potential, and durable resistance', *Annual Review of Phytopathology*, 40(1), pp. 349–379.

doi:10.1146/annurev.phyto.40.120501.101443.

Mendum, T.A. *et al.* (2014) 'Mycobacterium leprae genomes from a British medieval leprosy hospital: towards understanding an ancient epidemic', *BMC Genomics*, 15(270), pp. 1–8. doi:10.1186/1471-2164-15-270.

Mhedbi-Hajri, N. *et al.* (2011) 'Sensing and adhesion are adaptive functions in the plant pathogenic xanthomonads', *BMC Evolutionary Biology*, 11(67), pp. 1–12. doi:10.1186/1471-2148-11-67.

Mhedbi-Hajri, N. *et al.* (2013) 'Evolutionary history of the plant pathogenic bacterium *Xanthomonas axonopodis*', *PLoS ONE*. Edited by K.A. Crandall, 8(3), pp. 1–15. doi:10.1371/journal.pone.0058474.

Miller, W. *et al.* (2008) 'Sequencing the nuclear genome of the extinct woolly mammoth', *Nature*, 456(7220), pp. 387–390. doi:10.1038/nature07446.

Mira, A., Pushker, R. and Rodríguez-Valera, F. (2006) 'The Neolithic revolution of bacterial genomes', *Trends in Microbiology*, 14(5), pp. 200–206. doi:10.1016/j.tim.2006.03.001.

Mühlemann, B., Jones, T.C., *et al.* (2018) 'Ancient hepatitis B viruses from the Bronze Age to the Medieval period', *Nature*, 557(7705), pp. 418–423. doi:10.1038/s41586-018-0097-z.

Mühlemann, B., Margaryan, A., *et al.* (2018) 'Ancient human parvovirus B19 in Eurasia reveals its long-term association with humans', *Proceedings of the National Academy of Sciences*, 115(29), pp. 7557–7562. doi:10.1073/pnas.1804921115.

Murray, G.G.R. *et al.* (2016) 'The effect of genetic structure on molecular dating and tests for temporal signal', *Methods in Ecology and Evolution*. Edited by M. Gilbert, 7(1), pp. 80–89. doi:10.1111/2041-210X.12466.

Murray, G.G.R. *et al.* (2021) 'Genome reduction is associated with bacterial pathogenicity across different scales of temporal and ecological divergence', *Molecular Biology and Evolution*. Edited by D. Falush, 38(4), pp. 1570–1579. doi:10.1093/molbev/msaa323.

Nagarajan, N. and Pop, M. (2013) 'Sequence assembly demystified', *Nature Reviews Genetics*, 14(3), pp. 157–167. doi:10.1038/nrg3367.

Namouchi, A. *et al.* (2018) 'Integrative approach using *Yersinia pestis* genomes to revisit the historical landscape of plague during the Medieval period', *Proceedings of the National Academy of Sciences*, 115(50), pp. 11790–11797. doi:10.1073/pnas.1812865115.

Navascués, M., Depaulis, F. and Emerson, B.C. (2010) 'Combining contemporary and ancient DNA in population genetic and phylogeographical studies', *Molecular Ecology Resources*, 10(5), pp. 760–772. doi:10.1111/j.1755-0998.2010.02895.x.

Ollitrault, P., Curk, F. and Krueger, R. (2020) 'Citrus taxonomy', in *The Genus Citrus*. Duxford: Woodhead Publishing, pp. 57–82.

Orlando, L. and Cooper, A. (2014) 'Using ancient DNA to understand evolutionary and ecological processes', *Annual Review of Ecology, Evolution, and Systematics*, 45, pp. 573–598. doi:10.1146/annurev-ecolsys-120213-091712.

Pääbo, S. (1985) 'Molecular cloning of Ancient Egyptian mummy DNA', *Nature*, 314(6012), pp. 644–645. doi:10.1038/314644a0.

Pääbo, S. (1989) 'Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification.', *Proceedings of the National Academy of Sciences*, 86(6), pp. 1939–1943. doi:10.1073/pnas.86.6.1939.

Pääbo, S. *et al.* (2004) 'Genetic analyses from ancient DNA', *Annual Review of Genetics*, 38(1), pp. 645–

679. doi:10.1146/annurev.genet.37.110801.143214.

Pallen, M.J. and Wren, B.W. (2007) 'Bacterial pathogenomics', *Nature*, 449(7164), pp. 835–842. doi:10.1038/nature06248.

Parker, J., Rambaut, A. and Pybus, O.G. (2008) 'Correlating viral phenotypes with phylogeny: Accounting for phylogenetic uncertainty', *Infection, Genetics and Evolution*, 8(3), pp. 239–246. doi:10.1016/j.meegid.2007.08.001.

Patané, J.S.L. *et al.* (2019) 'Origin and diversification of *Xanthomonas citri* subsp. *citri* pathotypes revealed by inclusive phylogenomic, dating, and biogeographic analyses', *BMC Genomics*, 20(700), pp. 1–23. doi:10.1186/s12864-019-6007-4.

Patel, M.K., Allayyanavaramath, S.B. and Kulkarni, Y.S. (1953) 'Bacterial shot-hole and fruit canker of *Aegle marmelos* Correa', *Current Science*, 22, pp. 216–217.

Patterson Ross, Z. *et al.* (2018) 'The paradox of HBV evolution as revealed from a 16th century mummy', *PLOS Pathogens*. Edited by S. Duffy, 14(1), pp. 1–25. doi:10.1371/journal.ppat.1006750.

Pooggin, M.M. (2018) 'Small RNA-omics for plant virus identification, virome reconstruction, and antiviral defense characterization', *Frontiers in Microbiology*, 9, pp. 1–20. doi:10.3389/fmicb.2018.02779.

Pruvost, O. *et al.* (2002) 'Survival of *Xanthomonas axonopodis* pv. *citri* in leaf lesions under tropical environmental conditions and simulated splash dispersal of inoculum', *Phytopathology*, 92(4), pp. 336–346. doi:10.1094/PHYTO.2002.92.4.336.

Pruvost, O. *et al.* (2014) 'A MLVA genotyping scheme for global surveillance of the citrus pathogen *Xanthomonas citri* pv. *citri* suggests a worldwide geographical expansion of a single genetic lineage', *PLoS ONE*, 9(6), pp. 1–11. doi:10.1371/journal.pone.0098129.

Pruvost, O. *et al.* (2019) 'Deciphering how plant pathogenic bacteria disperse and meet: Molecular epidemiology of *Xanthomonas citri* pv. *citri* at microgeographic scales in a tropical area of Asiatic citrus canker endemicity', *Evolutionary Applications*, 12(8), pp. 1523–1538. doi:10.1111/eva.12788.

Rambaut, A. (2000) 'Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies', *Bioinformatics*, 16(4), pp. 395–399. doi:10.1093/bioinformatics/16.4.395.

Rambaut, A. *et al.* (2016) 'Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen)', *Virus Evolution*, 2(1), pp. 1–7. doi:10.1093/ve/vew007.

Rascovan, N. *et al.* (2019) 'Emergence and spread of basal lineages of *Yersinia pestis* during the Neolithic Decline', *Cell*, 176(1–2), pp. 295–305. doi:10.1016/j.cell.2018.11.005.

Rasmussen, D.A. and Grünwald, N.J. (2020) 'Phylogeographic approaches to characterize the emergence of plant pathogens', *Phytopathology*, 111(1), pp. 68–77. doi:10.1094/PHYTO-07-20-0319-Fl.

Rasmussen, S. *et al.* (2015) 'Early divergent strains of *Yersinia pestis* in Eurasia 5,000 years ago', *Cell*, 163(3), pp. 571–582. doi:10.1016/j.cell.2015.10.009.

Richard, D. *et al.* (2020) 'Time-calibrated genomic evolution of a monomorphic bacterium during its establishment as an endemic crop pathogen', *Molecular Ecology*, pp. 1–13. doi:10.1111/mec.15770.

Rieux, A. *et al.* (2014) 'Improved calibration of the human mitochondrial clock using ancient genomes', *Molecular Biology and Evolution*, 31(10), pp. 2780–2792. doi:10.1093/molbev/msu222.

Rieux, A. and Balloux, F. (2016) 'Inferences from tip-calibrated phylogenies: a review and a practical

- guide', *Molecular Ecology*, 25(9), pp. 1911–1924. doi:10.1111/mec.13586.
- Rigano, L.A. *et al.* (2007) 'Biofilm formation, epiphytic fitness, and canker development in *Xanthomonas axonopodis* pv. *citri*', *Molecular Plant-Microbe Interactions*[®], 20(10), pp. 1222–1230. doi:10.1094/MPMI-20-10-1222.
- Ristaino, J.B. (2020) 'The importance of mycological and plant herbaria in tracking plant killers', *Frontiers in Ecology and Evolution*, 7(521), pp. 1–11. doi:10.3389/fevo.2019.00521.
- Ristaino, J.B., Groves, C.T. and Parra, G.R. (2001) 'PCR amplification of the Irish potato famine pathogen from historic specimens', *Nature*, 411, pp. 695–697.
- Robène, I. *et al.* (2020) 'Development and comparative validation of genomic-driven PCR-based assays to detect *Xanthomonas citri* pv. *citri* in citrus plants', *BMC Microbiology*, 20(296), pp. 1–13. doi:10.1186/s12866-020-01972-8.
- Rohland, N. *et al.* (2015) 'Partial uracil–DNA–glycosylase treatment for screening of ancient DNA', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1660), pp. 1–11. doi:10.1098/rstb.2013.0624.
- Ryan, R.P. *et al.* (2011) 'Pathogenomics of *Xanthomonas*: understanding bacterium–plant interactions', *Nature Reviews Microbiology*, 9, pp. 344–355. doi:10.1038/nrmicro2558.
- Rybak, M. *et al.* (2009) 'Identification of *Xanthomonas citri* ssp. *citri* host specificity genes in a heterologous expression host', *Molecular Plant Pathology*, 10(2), pp. 249–262. doi:10.1111/j.1364-3703.2008.00528.x.
- Sánchez-Romero, M.A., Cota, I. and Casadesús, J. (2015) 'DNA methylation in bacteria: from the methyl group to the methylome', *Current Opinion in Microbiology*, 25, pp. 9–16. doi:10.1016/j.mib.2015.03.004.
- Särkinen, T. *et al.* (2012) 'How to Open the Treasure Chest? Optimising DNA Extraction from Herbarium Specimens', *PLoS ONE*. Edited by D. Caramelli, 7(8), p. e43808. doi:10.1371/journal.pone.0043808.
- Savary, S. *et al.* (2019) 'The global burden of pathogens and pests on major food crops', *Nature Ecology & Evolution*, 3, pp. 430–439. doi:10.1038/s41559-018-0793-y.
- Saville, A.C., Martin, M.D. and Ristaino, J.B. (2016) 'Historic late blight outbreaks caused by a widespread dominant lineage of *Phytophthora infestans* (Mont.) de Bary', *PLoS ONE*, 11(12), pp. 1–22. doi:10.1371/journal.pone.0168381.
- Schrader, C. *et al.* (2012) 'PCR inhibitors - occurrence, properties and removal', *Journal of Applied Microbiology*, 113(5), pp. 1014–1026. doi:10.1111/j.1365-2672.2012.05384.x.
- Schubert, T.S. *et al.* (2001) 'Meeting the challenge of eradicating citrus canker in Florida—again', *Plant Disease*, 85(4), pp. 340–356. doi:10.1094/PDIS.2001.85.4.340.
- Schuenemann, V.J. *et al.* (2013) 'Genome-wide comparison of medieval and modern *Mycobacterium leprae*', *Science*, 341(6142), pp. 179–183. doi:10.1126/science.1238286.
- Schuenemann, V.J., Avanzi, C., *et al.* (2018) 'Ancient genomes reveal a high diversity of *Mycobacterium leprae* in medieval Europe', *PLOS Pathogens*. Edited by D.M. Monack, 14(5), pp. 1–17. doi:10.1371/journal.ppat.1006997.
- Schuenemann, V.J., Kumar Lankapalli, A., *et al.* (2018) 'Historic *Treponema pallidum* genomes from colonial Mexico retrieved from archaeological remains', *PLOS Neglected Tropical Diseases*. Edited by S.J. Norris, 12(6), pp. 1–20. doi:10.1371/journal.pntd.0006447.
- Seitz, A. and Nieselt, K. (2017) 'Improving ancient DNA genome assembly', *PeerJ*, 5, pp. 1–20.

doi:10.7717/peerj.3126.

Seong, H.J. *et al.* (2016) 'Methylome analysis of two *Xanthomonas* spp. using single-molecule real-time sequencing', *The Plant Pathology Journal*, 32(6), pp. 500–507. doi:10.5423/PPJ.FT.10.2016.0216.

Shapiro, B. *et al.* (2004) 'Rise and Fall of the Beringian Steppe Bison', *Science*, 306(5701), pp. 1561–1565. doi:10.1126/science.1101074.

Shapiro, B. and Hofreiter, M. (2014) 'A paleogenomic perspective on evolution and gene function: new insights from ancient DNA', *Science*, 343(6169), pp. 1–7. doi:10.1126/science.1236573.

Shepherd, L.D. (2017) 'A non-destructive DNA sampling technique for herbarium specimens', *PLoS ONE*. Edited by M. Knapp, 12(8), pp. 1–7. doi:10.1371/journal.pone.0183555.

Sicard, A. *et al.* (2018) '*Xylella fastidiosa*: insights into an emerging plant pathogen', *Annual Review of Phytopathology*, 56(1), pp. 181–202. doi:10.1146/annurev-phyto-080417-045849.

da Silva, A.C.R. *et al.* (2002) 'Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities', *Nature*, 417(6887), pp. 459–463. doi:10.1038/417459a.

Skaria, M. and da Graça, J.V. (2012) 'History lessons towards proactive citrus canker efforts in Texas', *Subtropical Plant Science*, 64, pp. 29–33.

Spagnoletti, M. *et al.* (2014) 'Acquisition and Evolution of SXT-R391 Integrative Conjugative Elements in the Seventh-Pandemic *Vibrio cholerae* Lineage', *mBio*. Edited by J.E. Davies, 5(4). doi:10.1128/mBio.01356-14.

Spigelman, M. and Lemma, E. (1993) 'The use of the polymerase chain reaction (PCR) to detect *Mycobacterium tuberculosis* in ancient skeletons', *International Journal of Osteoarchaeology*, 3(2), pp. 137–143. doi:10.1002/oa.1390030211.

Spratt, B. (2001) 'The relative contributions of recombination and point mutation to the diversification of bacterial clones', *Current Opinion in Microbiology*, 4(5), pp. 602–606. doi:10.1016/S1369-5274(00)00257-5.

Spyrou, M.A. *et al.* (2016) 'Historical *Y. pestis* genomes reveal the European Black Death as the source of ancient and modern Plague pandemics', *Cell Host & Microbe*, 19(6), pp. 874–881. doi:10.1016/j.chom.2016.05.012.

Spyrou, M.A. *et al.* (2018) 'Analysis of 3,800-year-old *Yersinia pestis* genomes suggests Bronze Age origin for bubonic plague', *Nature Communications*, 9(2234), pp. 1–10. doi:10.1038/s41467-018-04550-9.

Spyrou, M.A. *et al.* (2019) 'Ancient pathogen genomics as an emerging tool for infectious disease research', *Nature Reviews Genetics*, 20(6), pp. 323–340. doi:10.1038/s41576-019-0119-1.

Staats, M. *et al.* (2011) 'DNA damage in plant herbarium tissue', *PLoS ONE*. Edited by C. Lalueza-Fox, 6(12), pp. 1–9. doi:10.1371/journal.pone.0028448.

Stukenbrock, E.H. and McDonald, B.A. (2008) 'The origins of plant pathogens in agro-ecosystems', *Annual Review of Phytopathology*, 46(1), pp. 75–100. doi:10.1146/annurev.phyto.010708.154114.

Sun, X. *et al.* (2004) 'Detection and characterization of a new strain of citrus canker bacteria from key/Mexican lime and alemow in South Florida', *Plant Disease*, 88(11), pp. 1179–1188. doi:10.1094/PDIS.2004.88.11.1179.

Tajima, F. (1983) 'Evolutionary relationship of DNA sequences in finite populations', *Genetics*, 105(2), pp. 437–460. doi:10.1093/genetics/105.2.437.

- Talon, M., Caruso, M. and Gmitter Jr., F.G. (eds) (2020) *The Genus Citrus*. Duxford: Woodhead Publishing.
- Taubenberger, J.K. *et al.* (2005) 'Characterization of the 1918 influenza virus polymerase genes', *Nature*, 437(7060), pp. 889–893. doi:10.1038/nature04230.
- Tavaré, S. and Miura, R.M. (1986) 'Some probabilistic and statistical problems in the analysis of DNA sequences', *American Mathematical Society*, 17, pp. 57–86.
- Thomas, W.K. *et al.* (1990) 'Spatial and temporal continuity of kangaroo rat populations shown by sequencing mitochondrial DNA from museum specimens', *Journal of Molecular Evolution*, 31(2), pp. 101–112. doi:10.1007/BF02109479.
- Trovão, N.S. *et al.* (2015) 'Host ecology determines the dispersal patterns of a plant virus', *Virus Evolution*, 1(1), p. vev016. doi:10.1093/ve/vev016.
- Tumpey, T.M. (2005) 'Characterization of the reconstructed 1918 Spanish influenza pandemic virus', *Science*, 310(5745), pp. 77–80. doi:10.1126/science.1119392.
- Vågene, Å.J. *et al.* (2018) '*Salmonella enterica* genomes from victims of a major 16th century epidemic in Mexico', *Nature Ecology & Evolution*, 2, pp. 520–528. doi:10.1038/s41559-017-0446-6.
- Valiente-Mullor, C. *et al.* (2021) 'One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads', *PLOS Computational Biology*. Edited by K.F. Au, 17(1), p. e1008678. doi:10.1371/journal.pcbi.1008678.
- van der Valk, T. *et al.* (2021) 'Million-year-old DNA sheds light on the genomic history of mammoths', *Nature*, 591(7849), pp. 265–269. doi:10.1038/s41586-021-03224-9.
- Vernière, C. *et al.* (1998) 'Characterization of phenotypically distinct strains of *Xanthomonas axonopodis* pv. *citri* from Southwest Asia', *European Journal of Plant Pathology*, 104, pp. 477–487.
- Wagner, D.M. *et al.* (2014) '*Yersinia pestis* and the Plague of Justinian 541–543 AD: a genomic analysis', *The Lancet Infectious Diseases*, 14(4), pp. 319–326. doi:10.1016/S1473-3099(13)70323-2.
- Wandeler, P., Hoeck, P.E.A. and Keller, L.F. (2007) 'Back to the future: museum specimens in population genetics', *Trends in Ecology & Evolution*, 22(12), pp. 634–642. doi:10.1016/j.tree.2007.08.017.
- Wang, R. *et al.* (2018) 'The global distribution and spread of the mobilized colistin resistance gene *mcr-1*', *Nature Communications*, 9(1), p. 1179. doi:10.1038/s41467-018-03205-z.
- Warinner, C. *et al.* (2014) 'Pathogens and host immunity in the ancient human oral cavity', *Nature Genetics*, 46(4), pp. 336–344. doi:10.1038/ng.2906.
- Weinert, L.A. *et al.* (2012) 'Molecular dating of human-to-bovid host jumps by *Staphylococcus aureus* reveals an association with the spread of domestication', *Biology Letters*, 8(5), pp. 829–832. doi:10.1098/rsbl.2012.0290.
- Weiβ, C.L. *et al.* (2016) 'Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens', *Royal Society Open Science*, 3(6), p. 160239. doi:10.1098/rsos.160239.
- Wibowo, M.C. *et al.* (2021) 'Reconstruction of ancient microbial genomes from the human gut', *Nature*, 594(7862), pp. 234–239. doi:10.1038/s41586-021-03532-0.
- Wiehe, P.O. (1941) 'Report on plant pathologist's visit to Rodrigues', *Ministry of Agriculture, Fisheries and Food*, 16, pp. 1–22.
- Worobey, M. *et al.* (2016) '1970s and "Patient 0" HIV-1 genomes illuminate early HIV/AIDS history in North America', *Nature*, 539(7627), pp. 98–101. doi:10.1038/nature19827.

- Wu, G.A. *et al.* (2018) 'Genomics of the origin and evolution of Citrus', *Nature*, 554(7692), pp. 311–316. doi:10.1038/nature25447.
- Yoshida, K. *et al.* (2013) 'The rise and fall of the *Phytophthora infestans* lineage that triggered the Irish potato famine', *eLife*, 2, pp. 1–25. doi:10.7554/eLife.00731.
- Yoshida, K. *et al.* (2014) 'Mining herbaria for plant pathogen genomes: back to the future', *PLoS Pathogens*, 10(4), pp. 1–6. doi:10.1371/journal.ppat.1004028.
- Yoshida, K., Sasaki, E. and Kamoun, S. (2015) 'Computational analyses of ancient pathogen DNA from herbarium samples: challenges and prospects', *Frontiers in Plant Science*, 6, pp. 1–6. doi:10.3389/fpls.2015.00771.
- Zeder, M.A. (2017) 'Domestication as a model system for the extended evolutionary synthesis', *Interface Focus*, 7, pp. 1–23. doi:10.1098/rsfs.2016.0133.
- Zhang, Y. *et al.* (2015) 'Positive selection is the main driving force for evolution of citrus canker-causing *Xanthomonas*', *The ISME Journal*, 9(10), pp. 2128–2138. doi:10.1038/ismej.2015.15.
- Zhou, Z. *et al.* (2018) 'Pan-genome analysis of ancient and modern *Salmonella enterica* demonstrates genomic stability of the invasive Para C lineage for millennia', *Current Biology*, 28(15), pp. 2420–2428. doi:10.1016/j.cub.2018.05.058.
- Ziebuhr, W. *et al.* (1999) 'Evolution of bacterial pathogenesis', *Cellular and Molecular Life Sciences (CMLS)*, 56(9–10), pp. 719–728. doi:10.1007/s000180050018.

Résumé

Les pathogènes de cultures représentent une menace pour l'Homme depuis les débuts de l'Agriculture. Afin de mieux comprendre les maladies actuelles de cultures et prévenir les épidémies futures, il est indispensable de comprendre les facteurs sous-jacents à l'émergence, l'adaptation et la diffusion des agents pathogènes. De récents développements méthodologiques dans le domaine de l'épidémiologie moléculaire permettent désormais de reconstruire les dynamiques spatio-temporelles des maladies avec une précision accrue. Alors que la majorité des études précédemment réalisées se sont entièrement appuyées sur l'échantillonnage d'individus contemporains datant des quelques dernières décennies, l'avènement de la paléogénomique permet désormais de reconstruire les génomes historiques de pathogènes datant de plusieurs siècles et ainsi d'étudier leur histoire évolutive avec une plus grande précision. Afin d'évaluer l'apport de la paléogénomique dans l'étude de l'émergence et de l'évolution de pathogènes de cultures, nous avons choisi *Xanthomonas citri* pathovar *citri* (*Xci*), la bactérie pathogène responsable du chancre asiatique des agrumes, comme modèle d'étude. Dans un premier temps nous avons optimisé les protocoles moléculaires et les pipelines bio-informatiques permettant de séquencer et reconstruire au mieux des génomes historiques de *Xci* à partir de spécimens d'herbiers. Dans ce contexte, nous nous sommes particulièrement attaché à étudier les patrons de dégradations de l'ADN dont la mesure est indispensable pour l'authentification des génomes historiques. Dans un second temps, nous avons précisé l'histoire de l'émergence de *Xci* à une échelle locale, celle des îles du sud-ouest de l'océan Indien (SOOI) grâce à l'analyse détaillée du premier génome historique de bactérie pathogène reconstruit à partir d'un échantillon d'herbier datant de 1937. Finalement, nous avons significativement amélioré la reconstruction de l'origine et de la diversification de *Xci* à l'échelle mondiale par l'analyse combinée des 13 génomes historiques générés durant cette thèse et d'une collection de génomes modernes représentative de la diversité génétique globale du pathogène. Nos travaux soulignent l'importance des données historiques dans la reconstruction de l'histoire évolutive d'agents pathogènes de cultures, valorisant les collections naturalistes et générant des connaissances ayant le potentiel d'optimiser les stratégies de lutte et de surveillance des épidémies actuelles et de mieux prédire les épidémies futures.

Mots-clefs : ADN ancien, paléogénomique, phylogénétique, datation moléculaire, histoire évolutive, phytopathogène, *Xanthomonas citri* pathovar *citri*, chancre asiatique des agrumes

Deciphering the emergence and evolutionary history of crop pathogens: insights from historical herbarium specimens

Abstract:

Crop pathogens have been a threat to mankind since the beginnings of agriculture. In order to better understand current crop diseases and prevent future epidemics, it is essential to appreciate the factors underlying the emergence, adaptation and spread of pathogens. Recent methodological developments in the field of molecular epidemiology now allow reconstructing disease dynamics in space and time. While the majority of studies previously carried out are entirely based on the sampling of contemporary individuals dating from the last few decades, the advent of paleogenomics now makes it possible to reconstruct historical genomes from several centuries and to study their evolutionary history with a greater precision. In order to assess the contribution of paleogenomics to the study of crop pathogen emergence and evolution, we chose *Xanthomonas citri* pathovar *citri* (*Xci*), the pathogenic bacterium responsible for Asiatic citrus canker, as study model. First, we optimised molecular protocols and bioinformatics pipelines to reconstruct historical *Xci* genomes from herbarium specimens. In this context, we specifically investigated DNA degradation patterns which measurement is essential for historical genomes authentication. Secondly, the detailed analysis of the first historical pathogenic bacterial genome reconstructed from a herbarium sample dating from 1937 allowed us to precise *Xci* emergence history at a local scale of the Southwest Indian Ocean (SWIO) islands. Finally, we significantly improved the reconstruction of the origin and diversification of *Xci* on a global scale by the combined analysis of the 13 historical genomes generated during this thesis and a collection of modern genomes representative of the worldwide genetic diversity of the pathogen. Our works emphasise the importance of historical data in the reconstruction of crop pathogens evolutionary history, valourising naturalist collections and generating knowledge bearing the potential of improving disease monitoring and sustainable control of current and future epidemics.

Key-words: ancient DNA, paleogenomics, phylogenetics, molecular dating, evolutive history, phytopathogen, *Xanthomonas citri* pathovar *citri*, Asiatic citrus canker