Évolution des génomes et modes de reproduction de Cryphonectria parasitica, l'agent causal du chancre du châtaignier, dans le contexte d'une double introduction en Europe.

Arthur Demene

université de BORDEAUX

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEUR DE

L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE SCIENCES ET ENVIRONNEMENTS

SPÉCIALITÉ ÉCOLOGIE ÉVOLUTIVE, FONCTIONNELLE ET DES COMMUNAUTÉS

Par Arthur DEMENÉ

Évolution des génomes et modes de reproduction de *Cryphonectria parasitica*, l'agent causal du chancre du châtaignier, dans le contexte d'une double introduction en Europe.

Sous la direction de : Cyril Dutech

Soutenue le 19/12/2019

Membres du jury :

Mme FOURNIER, Elisabeth	INRA, UMR BGPI, MONTPELLIER DR,	Rapporteure
M. DUPLESSIS, Sébastien	INRA, UMR 1136, Université de Lorraine	DR, Rapporteur
Mme AMSELEM, Joëlle	INRA, URGI, VERSAILLES	IR, Examinatrice
Mme SCHURDI-LEVRAUD, Valérie	Université de Bordeaux, UMR 1332	MdC, Examinatrice
M. SAUPE, Sven	CNRS, UMR 5095 IBGC, Bordeaux	DR, Examinateur
M. DUTECH, Cyril	INRA, UMR 1202 BIOGECO, Bordeaux	Directeur de thèse
Mme FOULONGNE-ORIOL , Marie	INRA, UR 1264 MYCSA , Bordeaux	CR, invitée

Titre : Évolution des génomes et modes de reproduction de *Cryphonectria parasitica*, l'agent causal du chancre du châtaignier, dans le contexte d'une double introduction en Europe.

Résumé :

Les invasions biologiques sont une composante majeure des changements globaux, causant de nombreuses perturbations dans le fonctionnement des écosystèmes et les activités humaines. Leur fréquence augmente depuis la fin du XXème siècle, notamment dans le cas des champignons pathogènes des arbres forestiers. Dans ce contexte d'introduction, l'adaptation des agents pathogènes à de nouveaux environnements et de nouveaux hôtes constitue un paradoxe évolutif, du fait d'une faible diversité génétique généralement introduite. Des mécanismes évolutifs permettant l'adaptation de ces organismes ont été proposés, mais à l'exception de quelques espèces modèles, souvent agents pathogènes de plantes cultivées, ces mécanismes restent peu étudiés chez les champignons.

Cette thèse a pour but d'étudier l'évolution des populations de Cryphonectria parasitica, l'agent causal du chancre du châtaignier, dans le contexte d'une double introduction en Europe. Originaire d'Asie, il a été introduit en Amérique du Nord à la fin du XIXème siècle, entraînant la disparition quasi-totale des populations naturelles de châtaigniers américains. Il a ensuite été introduit en Europe depuis ces populations nord-américaines et des populations asiatiques au début du XXème siècle. En Europe, les populations sont majoritairement structurées en lignées clonales contrairement aux populations d'origine. Ceci suggère un changement de mode de reproduction pouvant être impliqué dans le succès invasif de ces populations. Pourtant, les études précédentes ont montré que plusieurs génotypes n'appartenant pas aux principales lignées clonales se sont maintenus lors de la colonisation et que des croisements entre ces lignées clonales existent, même s'ils semblent limités. Les objectifs de cette thèse ont été de mieux décrire les croisements entre les lignées clonales et d'identifier de possibles barrières aux croisements entre celles-ci.

Dans une première partie, le génome de 50 isolats français, nord-américains et asiatiques ont permis de confirmer la forte similarité des isolats appartenant à une même lignée clonale européenne, à l'exception de petites régions divergentes échangées entre ces génomes, plus fréquemment observées entre lignées introduites d'une même origine. Par ailleurs, 5 des 6 lignées étudiées portent une signature d'échange récent de la région portant le gène du type sexuel, conférant aux lignées clonales la capacité de s'autoféconder (haploid-selfing). Ces résultats soulignent que le maintien de lignées clonales chez un champignon hétérothallique comme C. parasitica (cad que les croisements impliquent des génotypes avec des types sexuels différents) n'implique pas toujours et uniquement la reproduction asexuée.

Dans une seconde partie, l'assemblage de génomes d'isolats provenant des aires d'origine et introduites, utilisant des méthodes de séquençage long brin, a permis de comparer leur structure et composition chromosomique, et d'identifier de possibles barrières à la recombinaison. Aucun réarrangement majeur n'a cependant été détecté à l'exception d'une région d'1Mb, adjacente au locus du type sexuel, très riche en éléments transposables et variable entre les génomes. Ces résultats semblent infirmer l'hypothèse d'isolement reproducteur par réarrangement du génome. En revanche, les zones contenant de nombreux éléments mobiles pourraient permettre une évolution rapide de C. parasitica lors de l'introduction.

Une dernière partie aborde l'étude des génotypes semblant provenir de croisements entre les lignées clonales majoritaires afin d'explorer les processus de recombinaisons entre les pools génétiques des deux introductions et d'identifier de possible barrières à l'hybridation potentiellement associées à des combinaisons génétiques défavorables.

Ce travail souligne l'apport des données génomiques dans la compréhension des processus de recombinaison et la détection des variations génétiques d'un champignon pathogène invasif affectant ses capacités évolutives.

Mots clés : Génomique ; Recombinaison génétique ; Généalogie ; Éléments transposables ; Structure des populations ; Inférences bayesiennes ; Variations structurales du génome

Title: Evolution of genome and mating system of Cryphonectria parasitica, the chestnut blight fungus, during a multiple introduction contexte in Europe.

Abstract:

Biological invasions are a major component of global change, causing many disruptions in ecosystem functioning and human activities. Their frequency has been increasing since the end of the 20th century, particularly in the case of forest pathogenic fungi. In this introduction context, the adaptation of pathogens to new environments and hosts is an evolutionary paradox, due to the low genetic diversity generally introduced. Evolutionary mechanisms allowing the adaptation of these organisms have been proposed, but with the exception of a few model species, often pathogenic agents of crops plants, these mechanisms remain poorly studied in fungi.

The purpose of this thesis is to study the evolution of Cryphonectria parasitica populations, the causal agent of chestnut blight, in the context of a dual introduction into Europe. Native of Asia, it was introduced into North America at the end of the 19th century and almost caused the extinction of the American chestnut. It was then introduced into Europe from these North American and Asian populations at the beginning of the 20th century. In Europe, populations are mainly structured in clonal lineages, unlike the original populations. This suggests a change in reproductive mode that may be involved in the invasive success of these populations. However, previous studies have shown that several genotypes not belonging to the main clonal lineages have been maintained during colonization and that crosses between these clonal lineages exist, even if they seem to be limited. The objectives of this thesis were to better describe crosses between clonal lineages and to identify possible barriers to crosses between them.

In a first part, the genome of 50 French, North American and Asian isolates confirmed the strong similarity of isolates belonging to the same European clonal lineage, with the exception of small divergent regions exchanged between these genomes, more frequently observed between

introduced lineages from the same origin. In addition, 5 of the 6 lineages studied carry a signature of recent recombination of the region carrying the mating type locus. This gives the clonal lineages the ability to self-fertilize (haploid-selfing). These results underline that the maintenance of clonal lineages in a heterothallic fungus such as C. parasitica (i.e. crosses involve genotypes with different sexual types) does not always and solely involve asexual reproduction.

In a second part, the assembly of genomes of isolates from the native and introduced areas, using long-read sequencing methods, made it possible to compare their structure and chromosome composition, and to identify possible barriers to recombination. However, no major rearrangements were detected, except for a region of 1Mb adjacent to the sexual locus, which is very rich in transposable elements and variable between genomes. These results seem to refute the hypothesis of reproductive isolation by chromosomal rearrangement. On the other hand, areas containing many transposable elements could allow a rapid evolution of C. parasitica genomes during introduction.

A final part deals with the study of genotypes that appear to originate from crosses between the main clonal lineages in order to explore the recombination processes between the gene pools of the two introductions and to identify possible barriers to hybridization, potentially associated with unfavorable genetic combinations.

This work underlines the contribution of genomic data to the understanding of recombination processes and the detection of genetic variations in an invasive pathogenic fungus affecting its evolutionary capacities.

Keywords: Genomic; Genetic recombination; Genealogy; Transposable elements; Population structure; Bayesian inferences; Genome structural variations

Unité de recherche

INRA – Institut National de la Recherche Agronomique

UMR 1202 – Biodiversité, Gènes et Écosystèmes

Site de Recherches Forêt Bois de Pierroton

69, route d’Arcachon

33612 CESTAS Cedex-France

Remerciements

Tout d’abord, j’aimerais remercier vivement mes deux rapporteurs, Élisabeth et Sébastien, d’avoir accepté de lire mon manuscrit de thèse et pour m’avoir apporté des commentaires pertinents et encourageants. Lors de ces trois années de travail scientifique, je suis passé tant par des phases enrichissantes et motivantes que par des moments plus difficiles de remise en question, notamment lors de ces derniers mois de rédaction. Merci à vous deux pour votre regard bienveillant. Vous avez su mettre en relief les forces et les faiblesses de mon travail scientifique d’une manière constructive. Dans ce même élan, je remercie Joëlle, Valérie, Marie et Sven, membres de mon jury de thèse, pour ces discussions et suggestions éclairantes sur ma première production scientifique. Merci Cyril d’avoir mis en place ce jury aux sensibilités scientifiques diverses et complémentaires, ce qui a mené à des échanges très enrichissants, mettant en valeur mon travail de thèse dans un espace de réflexion multi-dimensionnel.

Évidemment, merci à Cyril pour ton encadrement. J’ai trouvé ton encadrement très complet et pertinent. Mes questionnements étaient nombreux et mes choix de direction à prendre dans ma thèse hasardeux, tu y as répondu d’une manière équilibrée en me montrant le chemin qui te paraissait le meilleur tout en me laissant une pleine liberté de mouvement. Merci.

Je tiens à remercier chaudement tous mes collègues de l’équipe GEMFor, pour votre soutient exceptionnel, votre aide et ces bons moments partagés. Je pense en particulier à Agathe, ma partenaire de thèse ; merci pour ces moments de galère et de joie partagés. C’était pas not’ guerre… Mais on l’a fait, on peut être fiers de nous. Merci Jack, c’était vraiment cool de refaire le monde avec toi pendant nos pauses. Et merci pour ton soutient dans mes derniers mois difficiles… Merci à Benoît, Jean-Paul, et Laure pour votre soutient et votre regard pertinent sur mon travail. Merci Gilles et Xavier, d’être venus vous geler avec moi pour échantillonner des chancres en plein février. Merci

Olivier pour ton expertise si précise en extraction d'ADN, tu es parti mais tu resteras un GEMFor à mes yeux... Merci Andrin, Christophe, Julie, Céline, pour ces efforts partagés à tenter d'extraire l'ADN de ce champignon capricieux. Un énorme merci à toi Sandrine, de m'avoir permis d'en finir avec ces extractions et pour tes conseils avisés. Merci Cécile et Marie-Laure d'avoir répondu à mes questions naïves de « patho », sans juger le bio-informaticien que vous aviez en face de vous qui assemble un génome de champignon sans savoir à quoi il ressemble. Merci Martine, pour nos discussions éternelles dans la laverie sur la Science et la Société. Merci beaucoup Ludovic, Benjamin(s) (Brachi et Penaud), Isabelle et Franck, pour ces discussions assemblages de génomes, n'hésitez pas si vous avez besoin de « Manuel l'assembleur de génomes à la main ».

Je tiens aussi à remercier vivement les membres de mon comité de thèse, pour votre appui scientifique. Merci Bertrand, Christophe, Pascal, Mathieu, Sophie et François pour votre regard constructif sur l'évolution de ma thèse.

J'aimerais aussi remercier les compagnons, amies et amis pictes, natifs ou adoptés. Pas besoin de faire une thèse pour faire de la science avec vous, et ce avec toute chose et à tout moment. C'est avec et grâce à vous aussi que j'en suis arrivé là. Merci…

Merci aux Bordelais, pour les pétanques, les virées au café des moines et ces soirées. Merci d'avoir fait semblant de vous intéresser à mon sujet de thèse pour me faire plaisir. Merci à vous tous pour votre soutient...

Un énorme merci à toi Minthé, de m'avoir supporté et soutenu tout au long de ma thèse, surtout sur ces derniers mois difficiles où ma lucidité et mon sommeil étaient au ras des pâquerettes. Je crois que je n'y serais pas arrivé sans toi…

Et enfin à vous ma famille, de m'avoir soutenu et encouragé du début à la fin de ma thèse. Papa et maman, je sais pas ce que vous avez mis dans nos biberons mais ça y est, le petit dernier aussi est devenu docteur.

Table des matières

Introduction

Lexique

Agressivité	Composante quantitative de la pathogénicité d'un génotype pathogène qui mesure le taux de dégâts provoqués par l'infection sur un individu ou une espèce donnée.
Allopatrie	Isolement géographique entre plusieurs populations.
Auto-stop (effet de)	Effet d'entraînement d'un allèle, variant structural ou variation nucléotique dû à sa liaison génétique avec une partie du génome soumise à la sélection.
Balayage sélectif	Réduction de la variation nucléotidique autour d'un allèle due à une pression de sélection ayant entrainé la fixation de cet allèle dans une population.
Changement d'hôte	Colonisation de nouvelle(s) espèce(s) hôte(s) associée à la perte de la capacité d'infecter l'hôte ancestral
Chromosomes accessoires	Chromosomes qui ne sont pas retrouvés dans les génomes de tous les individus d'une espèce ou d'une population. Ces chromosomes sont supposés ne pas être indispensables aux fonctions vitales (croissance, reproduction, survie) de l'individu.
Compatibilité végétative	Mécanisme génétique qui restreint la fusion des cytoplasmes entre deux génotypes qui possèdent des allèles différents sur un ou plusieurs locus impliqués dans cette compatibilité.
Déséquilibre de liaison	Association d'allèles de plusieurs locus dont la fréquence dans une population n'est pas aléatoire.

Effet pléiotropique	Influence d’un gène sur l’expression de plusieurs caractères phénotypiques.
Épistasie	Effet d’interaction entre les produits de plusieurs gènes.
Expansion de la gamme d’hôtes	Colonisation de nouvelle(s) espèce(s) d’hôte(s) tout en conservant la capacité d’infecter l’hôte ancestral.
Fardeau génétique	Accumulation de mutations délétères dans un génome.
Résistance (chez l’hôte)	Mesure de la capacité à contenir une infection.
Tolérance (chez l’hôte)	Mesure la capacité d'un hôte d'être peu affecté dans ses composantes de fitness (croissance, reproduction) mesuré à niveau d'infection égale.
ITS	De l’anglais Internal transcribed spacer. Région génomique située entre les gènes codants pour la petite sous-unité et la grande sous-unité de l’ARN ribosomique. Elle est utilisée comme marqueur taxonomique car ayant une haute probabilité d’identifier des champignons au niveau de l’espèce.
Métabolites secondaires	Molécules bio-actives non essentielles à la vie de l’organisme. Elles jouent parfois un rôle dans les interactions avec l’hôte chez les champignons. Une des signatures de ces gènes codant pour ces interactions sont qu'elles sont souvent regroupés en clusters.
Petits peptides secrétés	Ou effecteurs sécrétés. Molécules sécrétées par les champignons pathogènes asociées à leur capacité d'infecter et de coloniser leur(s) hôte(s).

Plasticité phénotypique	Aptitude d’un génotype à exprimer des phénotypes distincts dans des conditions environnementales différentes.
Pouvoir pathogène	Capacité d’un agent pathogène à causer une maladie sur une gamme d’hôte donnée.
Recombinaison ectopique	Échange de matériel génétique entre deux chromosomes non homologues.
Recombinaison parasexuée	Recombinaison génétique entre chromosomes homologues ou non hors de la méiose.
Séquençage de deuxième génération	L’ADN ou l’ARN d’un organisme est fragmenté aléatoirement, amplifié, puis séquençé produisant des millions de séquences, appelé “short-reads”, de tailles inférieures à 1 000pb. Très faibles taux d’erreur (~0.1%).
Séquençage de troisième génération	L’ADN ou l’ARN d’un organisme est séquencé sans étape d’amplification, et permet d’obtenir des “reads” mesurant jusqu’à 100 000pb. Taux d’erreur élevés (~10-15%).
Sympatrie	Co-existence de plusieurs populations dans une même aire géographique.
Synténie	Conservation de l'ordre et du sens des séquences nucléotidiques le long des chromosomes entre plusieurs génomes.
Suivi de l’hôte (host-tracking)	Co-évolution d’un hôte et d’un pathogène durant le processus de domestication de l’hôte, entraînant la spécialisation du pathogène à l’hôte domestiqué.

Trait quantitatif	Caractéristique phénotypique d’un organisme qu’il est possiblede mesurer par des nombres, dans une certaine unité de mesure.
Trait d'histoire de vie	Caractère dont l’expression est contingente à une ou plusieurs des étapes clés du développement d’un organisme. L’ensemble de ces étapes clés constitue le cycle de vie de cet organisme.
Valeur adaptative moyenne	Appelé fitness en anglais. Nombre moyen de descendants produits par un génotype. En général impossible à mesurer directement.
Variance génétique additive	Ampleur de la variabilité des allèles aux QTLs associés à la variabilité d’un caractèré.
Virulence	Composante qualitative de la pathogénicité, traduisant la capacité d’un agent pathogène à infecter ou non une plante.

I Les espèces invasives comme contexte d'étude de l'évolution

En 1859, Charles Darwin, motivé par le naturaliste Alfred Russel Wallace, publiait *De l'origine des espèces par le moyen de la sélection naturelle*, un ouvrage qui définit la sélection naturelle comme une force créatrice de nouvelles espèces à partir d'espèces plus anciennes. Il décrit la sélection naturelle comme un processus agissant sur les différences entre les individus d'une espèce, en préservant les variations favorables dans le milieu et éliminant les variations défavorables. Les notions de gènes et de mutations ne sont pas encore établies, mais cette théorie porte une attaque frontale à la conception fixiste qui postulait que les espèces ne changeaient pas au cours du temps. Elle introduit le terme d'évolution des espèces, qui décrit que les organismes se modifient au cours des générations, ce qui entraîne à terme la production de nouvelles espèces. Cette théorie conceptuellement simple et robuste dans son application aux données collectées sur les organismes vivants, constitue un nouveau paradigme permettant l'émergence de nouveaux concepts évolutionnistes. Au début du XXème siècle, la redécouverte des travaux de Gregor Mendel sur l'hérédité, la notion de gène et la transmission de leurs variations (les allèles) au fil des générations, sont revisités au regard de la théorie de Darwin de la sélection naturelle. Sewall Wright et Ronald A. Fisher introduisent ces concepts d'hérédité dans un modèle neutre pour étudier l'évolution de la répartition des allèles dans une population au cours du temps. Ces modèles marquent le début de la génétique des populations et permettent de modéliser les processus qui modifient la répartition des allèles dans une population par rapport aux attendus sous ce modèle neutre. L'évolution d'une population y est donc définie par la modification de la fréquence d'un ou plusieurs allèles dans cette population avec la sélection naturelle comme moteur principal de cette évolution (R.A. Fisher, *The Genetical Theory of Natural Selection*). Les travaux de J. B. S. Haldane confirment l'importance de la sélection naturelle dans l'évolution (Haldane, J. B. S., *A mathematical theory of natural and artificial selection*, 1924 -

1934), en s'appuyant notamment sur le cas de la Phalène du bouleau qui illustre qu'une pression de sélection extrême exercée par la prédation d'une forme biologique d'une espèce (un phénotype) peut mener à un remplacement quasi-total (98%) dans la population par une autre forme qui n'est pas contre-sélectionnée. Ces théories s'inscrivent dans le paradigme néo-darwinien qui a émergé à la fin de la première moitié du XX^{ème} siècle, dite « théorie synthétique de l'évolution » (J. Huxley, *The Modern Synthesis*, 1942).

Au cours de la deuxième partie du XX^{ème} siècle, le développement des méthodes d'amplification et de migration de l'ADN, support de l'information génétique, et des protéines, expression de cette information, sur gel d'électrophorèse ont permis de découvrir que la diversité génétique au sein des organismes était beaucoup plus importante que ce que suggérait le modèle synthétique (Lewontin et Hubby 1966), où chaque mutation a comme avenir d'être rapidement fixée ou éliminée. Ces avancées technologiques permettant d'accéder à l'information génétique portée par ADN conduisent à reconsidérer le paradigme néodarwinien et accompagnent l'émergence de la théorie neutraliste de l'évolution de Motoo Kimura (Motoo Kimura, 1968). Kimura postule que la cause principale de l'évolution des espèces n'est pas la sélection naturelle, mais la fixation aléatoire de mutations neutres ou faiblement délétères, la dérive génétique (Kimura, 1983). Cette théorie devient et demeure un pilier de la biologie évolutive, même si elle reste actuellement débattue (Kern et Hahn, 2018 ; Jensen et al., 2019). Le génome serait un paysage génétique diversifié évoluant principalement de manière aléatoire par dérive génétique et où la sélection purifiante (« background selection ») des nouvelles mutations souvent neutres ou faiblement délétères serait le mode de sélection naturel prédominant (Jensen et al., 2019). La sélection positive ou négative des mutations bénéfiques ou délétères s'ajouterait comme une sur-couche à ces deux processus décrits précédemment (Stephan, 2010). Ce qui unifie l'ensemble de ces concepts in fine, c'est que la mutation nucléotidique est la source essentielle de la variation génétique et de l'évolution des espèces. Cependant, les populations n'étant pas en équilibre démographique constant dans la nature, d'autres facteurs sont à considérer. La taille et les variations de taille des populations (F. Tajima 1989), leurs compositions

génétiques (Thornton et Jensen 2007) et la migration (Kimura et Weiss, 1964) doivent donc être pris en compte dans les modèles d'évolution pour comprendre leur impact sur la variabilité génétique d'une population et l'effet combiné de la sélection naturelle. En particulier, l'expansion de l'aire de distribution d'une espèce peut s'accompagner de changements démographiques dans les populations et avoir d'importantes conséquences sur son évolution, notamment en modifiant l'équilibre entre processus neutres et sélection naturelle (Eckert et al., 1996). Ainsi, il a été montré que suite à la colonisation de nouveaux environnements, une succession d'introductions peut limiter les effets de goulots d'étranglement démographiques et favoriser l'adaptation et la propagation d'une espèce (par exemple : Lavergne et Molofsky, 2007). Les populations qui ont réussi à s'établir au-delà des limites de leur aire de répartition initiale, en faisant face à des contraintes biotiques et abiotiques nouvelles, constituent donc un sujet propice à l'étude des processus qui dirigent l'évolution des populations et des espèces. Car plus généralement, certains auteurs comme Gould et Eldredge considèrent que l'évolution comprend de longues périodes impliquant peu de changements évolutifs, ponctuées par des périodes de changements majeurs d'environnement lors desquelles d'importants changements évolutifs ont lieu (Théorie des équilibres ponctués, Eldredge & Gould, 1972 ; Herrel et al., 2008), en particulier lors des grandes crises d'extinction du vivant comme nous le vivons actuellement.

I.1 Les invasions biologiques et leurs estimations via des paramètres démographiques

La définition d'espèce invasive, telle qu'utilisée dans ce manuscrit, correspond à celle proposée par Blackburn et ses collaborateurs qui ont proposé en 2011 un cadre synthétique de travail à partir des travaux de Williamson (1996) et Richardson et al., (2000). Selon ce cadre, une espèce invasive est un cas particulier d'espèce exotique qui a franchi plusieurs barrières et étapes qui auraient pu mener à son extinction. Ainsi, les individus des populations d'espèces invasives partagent la particularité d'avoir été transportés et introduits dans un nouvel environnement et ont été capables de survivre, se reproduire et se disperser dans ce nouvel environnement.

Il a été estimé que 480 000 espèces exotiques ont été introduites dans le monde (Pimentel et al., 2001). Une grande partie de ces espèces est bénéfique pour l'homme comme source de nourriture ou de matières premières, restaurateurs écologiques, ou bien en tant qu'animaux domestiques. Lorsque ces espèces deviennent invasives, elles peuvent avoir bouleversé le fonctionnement d'écosystèmes et constituent une composante importante des changements globaux (Vitousek et al. 1996), représentant la deuxième cause de perte de biodiversité (Alonso et al., 2001). Elles peuvent constituer une menace extrême pour l'agriculture (Paini et al., 2016) et les coûts environnementaux, sociaux et économiques liés à certaines espèces invasives sont importants (Pimentel et al., 2005). Un cas emblématique de l'impact négatif extrême que peuvent avoir ces espèces invasives concerne *Phytophthora infestans*, un oomycète causant le mildiou de la pomme de terre. Introduit en Europe depuis le Mexique au milieu des années 1840 (Goss et al., 2014), cet agent pathogène a causé une épidémie dévastatrice des cultures de pommes de terre en Irlande, principale denrée alimentaire en Irlande à cette époque (Birch et al., 2001). L'incapacité à déterminer l'origine de cette maladie et les réponses inadéquates apportées à cette crise alimentaire ont directement entraîné une terrible famine et la mort de plus d'un million

d’irlandais, ainsi que le déplacement de trois millions de personnes sur une population de huit millions de personnes (Duncan, 1999).

Les approches de génétique des populations permettent aujourd’hui d’analyser la diversité génétique des populations d’espèces invasives et d’identifier rapidement l’espèce, son origine géographique, ainsi que ses modes de reproduction et de dispersion. Ce sont autant de cibles potentielles de lutte contre ces espèces invasives. En effet, identifier l’origine de la population invasive peut permettre d’introduire des hôtes résistants ou des ennemis naturels aux agents pathogènes introduits, et comprendre les processus de dispersion permet d’agir sur les voies de dissémination de ces espèces invasives (Rouxel et al., 2012). Par ailleurs, estimer la variabilité génétique introduite permet aussi d'évaluer les capacités évolutives de ces espèces dans leur nouvel environnement et de mesurer les risques potentiels associés à l'introduction de diversité supplémentaire de la zone d'origine (Lavergne et Molfsky, 2007 ; Mariette et al., 2016)

I.2 Le paradoxe génétique des introductions biologiques

Les espèces invasives sont souvent exposées à des contraintes biotiques et abiotiques différentes par rapport à leur aire initiale de répartition. Les populations invasives doivent à priori s’adapter aux nouvelles conditions du milieu, et la variabilité génétique introduite de ces populations est supposée augmenter le succès de la colonisation (Sakai et al., 2001). En effet, une diversité génétique intra-population élevée devrait faciliter l’établissement de cette population car elle sous-entend des variations phénotypiques élevées, ce qui augmente les chances que certains individus supportent les nouvelles conditions du milieu (Frankham, 2005). Une meta-analyse regroupant une trentaine de jeu de données de plantes, de vertébrés et d’invertébrés a permis de montrer une corrélation positive significative entre la diversité génétique dans une population et sa valeur adaptative* moyenne (Reed et Frankham, 2003) [Les étoiles * signalent que le terme est définit dans le lexique de la thèse, page 10-13]. Pourtant, le processus d’introduction s’accompagne

souvent d’une diminution de la taille de la population car seulement une fraction des individus de la population d’origine est introduite dans le nouveau milieu. Ce phénomène, appelé goulot d’étranglement démographique, entraîne dans la plupart des cas une diversité génétique réduite dans la population introduite. Dans une étude réalisée sur 80 espèces de plantes, animaux et champignons, Dlugosch et Parker (2008) ont montré une diminution significative de 15.5 % de la richesse allélique dans les populations introduites en Amérique du Nord et dans le Pacifique par rapport aux populations d’origine, ainsi qu’une diminution de 18.7 % de l’hétérozygotie. De plus, et par effet d’entraînement, la reproduction d'un nombre réduit de génotypes peut induire une propension à la consanguinité qui va augmenter le fardeau génétique (l’accumulation des mutations délétères récessives, et l’augmentation de leur l’homozygotie pour les diploïdes), menant à une diminution de la valeur adaptative moyenne des individus, appelée fitness par la suite (Charlesworth et Willis, 2009). Enfin, dans une population d’effectif réduit à cause d’un goulot d’étranglement démographique, l’effet aléatoire de la dérive génétique sera plus important et pourra conduire à une diminution de la fitness de cette population (Willi et al., 2013).

Pourtant, malgré ce goulot d’étranglement démographique et génétique, les populations peuvent parfois s’adapter très rapidement à de nouvelles conditions biotiques et abiotiques. Dluglosch et Parker (2008b) ont étudié les conséquences évolutives d’une telle perte de diversité génétique sur plusieurs populations d’*Hypericum canariense*, une espèce de millepertuis. Cette espèce endémique et native des îles Canaries, a été introduite dans de nombreuses nouvelles localités en tant que plante ornementale. Malgré une diminution de 45 % de l’hétérozygotie dans trois populations introduites par rapport à la population d‘origine, ces populations invasives ont montré des signes rapides de changements adaptatifs : une croissance accrue et une adaptation locale de la période de floraison en fonction d’un gradient latitudinal (Dluglosch et Parker, 2008b). Cette adaptation rapide à un nouvel environnement semble être une caractéristique commune chez les espèces invasives. Une étude basée sur 38 espèces de plantes et d’animaux invasifs a montré que des traits associés au potentiel invasif des populations (taux de croissance, tolérance

climatique, temps de génération court etc.) avait subit des changements évolutifs suite à une introduction dans un nouvel environnement, en une période inférieure à une dizaine d'années dans certains cas (Whitney et Gabler, 2008). À la lumière de ces résultats, le fait qu'une population ayant une diversité génétique réduite par rapport à la population d'origine arrive à s'implanter et s'adapter rapidement aux nouvelles contraintes biotiques et abiotiques du milieu dans lequel elle est introduite constitue donc un paradoxe génétique (Sax et Brown, 2000).

I.3 Réponses possibles au paradoxe génétique de l'invasion

La diversité génétique peut parfois augmenter dans les populations introduites par rapport aux populations d'origine du fait de multiples événements d'introduction impliquant des populations sources génétiquement différenciées (Rius et Darling 2014). Ces introductions multiples peuvent permettre à la population introduite de surmonter l'effet délétère d'une perte de diversité génétique due à un goulot d'étranglement démographique et ainsi augmenter son potentiel évolutif dans le nouvel environnement. En effet, si la sélection naturelle dans l'aire d'introduction est forte, une population génétiquement diversifiée pourra répondre rapidement à ces pressions de sélection. Une étude réalisée sur *Phalaris arundinacea*, une espèce de plante de la famille des Poaceae, a mis en évidence l'importance de ces introductions multiples dans l'adaptation et le succès invasif des populations introduites (Lavergne et Molofsky, 2007). Cette plante vivace native d'Europe a été introduite en Amérique du nord où elle est devenue invasive. En se basant sur un large échantillonnage de l'ensemble de son aire de répartition (350 rhizomes) et à l'aide de 12 marqueurs neutres (allozymes), les auteurs ont montré que son potentiel invasif élevé en Amérique du nord était associé à une diversité génétique élevée dans les populations invasives, grâce à l'introduction de souches depuis plusieurs localités en Europe (Lavergne et Molofsky, 2007). Une deuxième caractéristique de ces introductions multiples est la possibilité de produire des génotypes nouveaux à partir de la

recombinaison génétique entre des génotypes d'origine génétique fortement différenciée (admixture). Une étude réalisée sur des populations de *Cottus gobio*, des poissons d'eaux douces, a permis de mettre en évidence une relation entre le phénomène d'admixture entre des populations allopatriques* et des changements phénotypiques permettant à l'espèce de coloniser les grands fleuves (Nolte et al., 2005). Les auteurs ont montré à partir de l'ADN mitochondrial et de SNPs (Single nucleotide polymorphism : variations de sites nucléotidiques) que la population invasive du nouveau milieu résultait de l'hybridation génétique entre des groupes génétiques parentaux différents. Seules les populations invasives de *C. gobio* ont pu coloniser les grands fleuves alors que les cours d'eau où sont établies les populations d'origine sont connectés à ces fleuves. Ainsi, bien que la nature de l'adaptation des hybrides reste inconnue, ce schéma ne peut s'expliquer que par un avantage des populations invasives vis-à-vis des facteurs biotiques et abiotiques des grands fleuves par rapport aux populations d'origine (Nolte et al., 2005). L'admixture de groupes génétiques différents peut donc permettre la création de nouveaux génotypes pouvant répondre à de nouvelles pressions de sélection en dehors de l'environnement natif d'une espèce. Pourtant, la présence de génotypes issus de groupes génétiques divergents en sympatrie* dans la zone d'introduction n'implique pas nécessairement leur mélange (Rius et Darling, 2014). Par exemple, l'analyse de six marqueurs microsatellites dans une population introduite de *Microcosmus squamiger*, un invertébré marin, a permis de montrer que des populations génétiquement différenciées associées à au moins deux sources génétiques ancestrales, portaient un signal très faible d'admixture chez les individus (~15%; Ordonez et al., 2013) alors que les deux groupes génétiques sont mélangés au niveau population (Figure 1).

La réussite de l'établissement d'espèces en dehors de leur aire de répartition naturelle peut aussi s'expliquer sans adaptation au nouveau milieu. L'espèce peut présenter un certain niveau de pré-adaptation aux nouvelles contraintes biotiques ou abiotiques, comme la capacité d'infecter un large spectre d'hôtes pour les agents pathogènes des plantes ou des animaux (Parker & Gilbert, 2004 ; Agosta & Klemens, 2008). Aussi, le succès des invasions a parfois été associé à un effet de fondation qui peut augmenter la

variance génétique additive* de la population invasive (Naciri-Graven & Goudet 2003) ou diminuer le fardeau génétique* (Facon et al. 2006). Enfin la plasticité phénotypique* pourrait permettre à un champignon pathogène d'infecter une gamme d'hôtes variée, sans impliquer de nouvelles combinaisons génotypiques (Lande, 2009). Ces exemples montrent qu'il peut y avoir plusieurs voies qui mènent au succès d'une invasion biologique qui n'impliquent pas toujours l'admixture pour répondre au paradoxe génétique de l'introduction. Ainsi, faute de trajectoire consensus, il semble important de mieux caractériser l'importance de la recombinaison génétique entre individus pour détecter des signaux d'admixture dans les populations introduites et mieux appréhender les mécanismes évolutifs qui mènent au succès d'une invasion biologique.

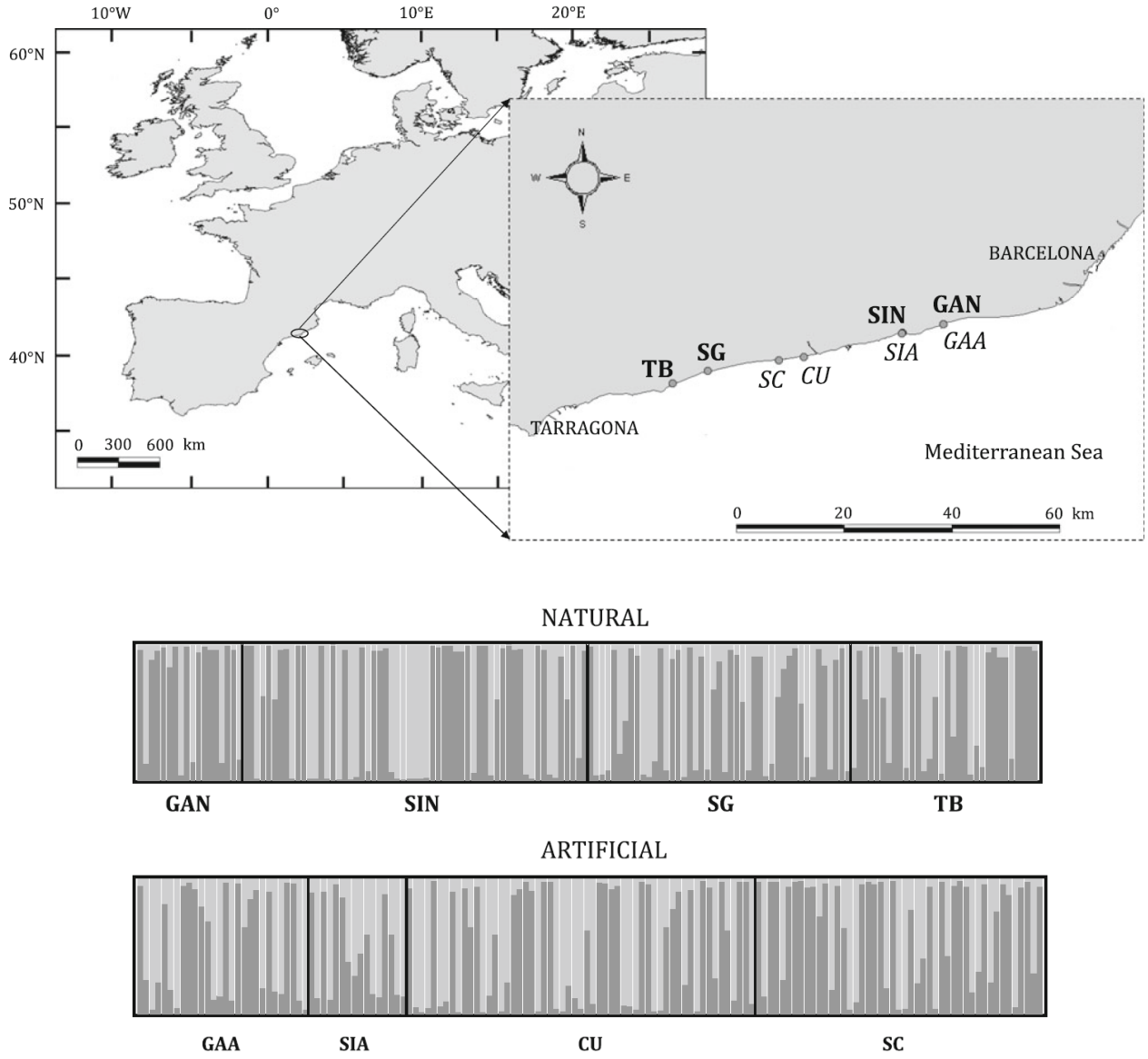


Figure 1, extraite de Ordóñez et al. (2013) : a) Localités échantillonnées de *Microcosmus squamiger*. En gras, les individus ont été échantillonés sur un substrat naturel, et en italique, échantillonnés sur un substrat artificiel b) Affectation de 302 individus à chacun des deux groupes génétiques (K=2) identifiés par le logiciel STRUCTURE. Chaque individu est représenté par une barre, les couleurs gris clair et gris foncé représentent la probabilité relative d'affectation à chaque cluster. Les populations sont séparées par des lignes verticales et leur identification correspond à la carte d'échantillonnage.

I.4 Limites des inférences démographiques et biais d'estimation et d'interprétation de la diversité génétique neutre

Malgré leur rôle potentiellement important dans le succès invasif, il peut être parfois difficile de détecter des phénomènes d'admixture, et le type de marqueurs moléculaires* utilisé pour mesurer la diversité génétique des populations peut avoir un impact sur cette détection (Estoup et al., 2016). Dluglosh et Parker (2008a) soulignent qu'il est indispensable de déterminer précisément la ou les population(s) source(s) à l'origine de populations introduites pour ne pas entraîner de fausses hypothèses de goulot d'étranglement ou d'admixture dans les populations introduites. Par exemple si la population d'origine est à la base faiblement diversifiée génétiquement, la faible diversité génétique des populations introduites n'implique pas forcément un goulot d'étranglement démographique. De plus, il est difficile de savoir si la perte de diversité génétique observée est due au goulot d'étranglement démographique, ou à un balayage sélectif* causé par la sélection d'allèles favorables pendant le processus d'invasion. Afin de déterminer précisément l'importance des effets de goulot d'étranglement et d'admixture, il est donc essentiel de rendre compte de la diversité génétique globale des populations introduites et natives, et d'utiliser des marqueurs génétique donnant accès à une résolution suffisante pour détecter les signaux de recombinaison génétiques et de balayage sélectif (Dlugosch et Parker, 2008; Estoup et al., 2016).

L'utilisation de marqueurs génétiques neutres (i.e. non soumis à la sélection) apportent une information essentielle pour identifier les processus évolutifs tels que les goulots d'étranglements génétiques et l'admixture qui modifient la variation génétique dans les populations introduites. Cependant, il est parfois évoqué que le paradoxe génétique est infondé car les marqueurs génétiques neutres utilisés en génétique des populations sont trop faiblement corrélés à la variation des traits quantitatifs (Reed & Frankham 2001) et à la fitness des individus (Roman et Darling 2007). Dans leur méta-analyse montrant une baisse de diversité génétique entre les populations introduites et les populations d'origine, Dlugosch et Parker (2008) ont souligné que le type de

marqueurs moléculaires utilisés pour mesurer la diversité génétique pouvait engendrer un biais dans son estimation. Les locus microsatellites révèlent des diminutions en richesses alléliques plus importantes que les protéines ($P < 0.0001$). De plus, la perte de variation quantitative des traits liés à la fitness est souvent plus faible que la perte de diversité des marqueurs neutres (Dlugosch et Parker 2008). En 2002, deux études se sont attardées à estimer la diversité génétique et la variation de sept traits quantitatifs d'histoire de vie* (traits étudiés détaillés dans Koskinen et al., 2002a) de trois populations isolées introduites depuis une même population source de *Thymallus arcticus*, un poisson de la famille des salmonidés. Les auteurs ont montré que les marqueurs moléculaires neutres utilisés, 17 marqueurs microsatellites, ont permis d'estimer une diminution de 50 % de la variabilité génétique dans les populations isolées, montrant un sévère goulot d'étranglement génétique, alors que les traits quantitatifs impliqués dans la fitness n'avaient pas été affectés (Koskinen et al., 2002a, 2002b). La perte de diversité génétique sur quelques marqueurs neutres n'est donc pas forcément corrélée à une perte de variance génétique sur des traits essentiels à l'établissement d'une espèce invasive, et le potentiel adaptatif d'une espèce peut être préservé même si l'on observe une perte de diversité génétique sur des marqueurs neutres.

Un autre problème majeur dans l'utilisation des locus microsatellites largement utilisés en génétique des populations est lié à leur taux de mutation élevé. Chez l'homme par exemple, le taux de mutation nucléotidique entre deux générations a été estimé à 10^{-9} (Crow 1993), alors que les taux de mutation estimés dans les allèles microsatellites sont estimés entre 10^{-3} et 10^{-4} (Weber et Wong, 1993). Ces taux de mutation très élevés ont aussi été estimés chez des espèce d'oiseaux, la drosophile et les fourmis (Ellegren 2000). De plus, il a été montré qu'un seul événement de mutation d'un locus microsatellite peut se traduire par le gain ou la perte de plusieurs unités répétées dans 4 % à 74 % des cas de mutation spontanée selon les organismes (Ellegren 2000). Il est donc difficile de déterminer si deux allèles d'un même locus microsatellite qui diffèrent de plusieurs répétitions sont issus d'un ou plusieurs événements de mutations récents ou plus probablement d'une migration récente d'un groupe génétique divergent. Ceci illustre un problème

majeur en génétique des populations, exacerbé par l'utilisation des locus microsatellites dans les études de génétique des populations, qui est l'impossibilité de différencier si la présence d'un allèle nouveau dans une population est dû à un événement de mutation ou à un événement de recombinaison génétique avec une source inconnue.

I.5 L’avènement du séquençage haut débit dans l’étude des invasions biologiques

L’amélioration des techniques de séquençage d’ADN depuis plusieurs dizaines d’années a permis d’obtenir une vision plus complète du génome et de son évolution neutre ou sélectionnée, et de s’affranchir des biais entraînés par un nombre de marqueurs limités. En 2003, Luikart et ses collaborateurs écrivaient « L’approche moléculaire idéale en génétique des populations permettrait l’étude de centaines de marqueurs polymorphes, recouvrant l’ensemble du génome, via une seule et simple expérimentation. Cette approche n’existe malheureusement pas encore ». Désormais, il est possible d’obtenir des milliers de marqueurs grâce à des techniques de génotypage par séquençage d’ADN (Davey et al., 2011), d’étudier le profil de transcription des gènes d’un organisme (*RNA-seq*, Ozsolak et Milos, 2011) ou même de séquencer l’intégralité du génome d’un organisme afin de reconstruire sa séquence d’ADN complète pouvant aller jusqu’au niveau des chromosomes entiers (*Whole genome sequencing*, Park et Kim, 2016). Ainsi, le séquençage d’ADN permet d’accéder à des milliers de marqueurs génétiques comme les SNPs qui permettent une couverture de l’ensemble du génome avec un taux d’erreur faible et une utilisation possible sur des espèces non modèles (Helyar et al., 2011). Cette couverture de l’ensemble du génome permet d’identifier des traces très réduites d’hybridations, des barrières aux échanges génétiques dans certaines parties du génome et des zones sous sélection pouvant être impliquées dans le succès invasif. Le séquençage massif de SNPs apporte aussi une lumière nouvelle sur le paradoxe des invasions biologiques, en permettant d’identifier de manière plus précise les populations à l’origine des populations introduites et donc de mieux caractériser les effets de goulot d’étranglement génétique. A partir d’un jeu de donnée de 65,195 SNPs, Selechnik et ses colaborateurs (2019) ont confirmé que des populations de crapauds buffle invasives qui se sont rapidement répandues, et peut-être adaptées, en Australie et à Hawaï avaient subi une diminution de diversité génétique par rapport aux populations d’origine en Guyane Française, confirmant ce qui avait été suggéré par le polymorphisme de 50 SNPs détectés sur une région de 468

paires de bases de l’ADN mitochondrial (Slade et Moritz, 1998) et de locus microsatellites (Estoup et al., 2001). Cependant, l’utilisation de SNPs couvrant l’intégralité du génome a permis de montrer que certains locus sous sélection ne présentaient pas une diminution de diversité génétique. Ce signal pourrait être la trace d‘une admixture entre plusieurs introductions génétiquement différentes. Dans ce cas ci, le paradoxe génétique suggéré par l’utilisation de marqueurs peu résolutifs (SNPs sur une séquence courte d’ADN mitochondrial et locus microsatellites) montrant une perte de diversité malgré l’adaptation rapide des populations, a été démontré comme infondé par l’utilisation d’une résolution accrue de SNPs ; la diversité génétique étant en fait maintenue pour certains locus adaptatifs. Ainsi, le développement des ces technologies alimente le débat sur la réelle importance du paradoxe génétique des invasions biologiques. En outre, le séquençage haut débit d’ADN s’accompagne de méthodes permettant de détecter et caractériser finement la recombinaison génétique dans les populations, dont l’impact sur le processus d’adaptation et la structure des populations au cours des invasions devrait être considéré prioritairement dans les études de génétique des populations (Martin et al., 2011).

II La dimension fongique des invasions biologiques

Il a été estimé qu'entre 65 et 85 % des épidémies émergentes sont causées par des agents phytopathogènes exotiques (Pimentel et al., 2001). Les agents pathogènes causant ces maladies infectieuses émergentes (*Emerging infections deseases*, EIDs) sont une proportion importante des invasions biologiques décrites à ce jour, en particulier les champignons pathogènes de plantes qui représentent la deuxième cause des EIDs des plantes (30%) causées par les micro-organismes, juste après les virus (Anderson et al., 2004). Dans le cas des écosystèmes forestiers, leur nombre a augmenté de manière exponentielle depuis plusieurs centaines d'années (Figure 2 a) avec une accélération ces vingt dernières années et les ascomycètes représentent 70 % de ces invasions (Fig 2 b ; Santini et al., 2013). Si les champignons pathogènes ont longtemps été peu étudiés par rapport aux animaux et aux plantes dans l'étude des invasions biologiques, ils sont beaucoup plus étudiés depuis une dizaine d'années, notamment parce qu'ils représentent une menace élevée pour la production de ressources des populations humaines (nourriture, matériaux, énergie) et la stabilité des écosystèmes (Strange et Scott, 2005 ; Deprez-Lousteau et al., 2007). Par ailleurs, les génomes des champignons sont généralement de petite taille (<100Mb, Cornell et al., 2007), comparés aux génomes des plantes (jusqu'à plus de 100Gb, Michael, 2014) ou des animaux (3.4Gb chez l'humain par exemple). Ils représentent donc des modèles idéaux

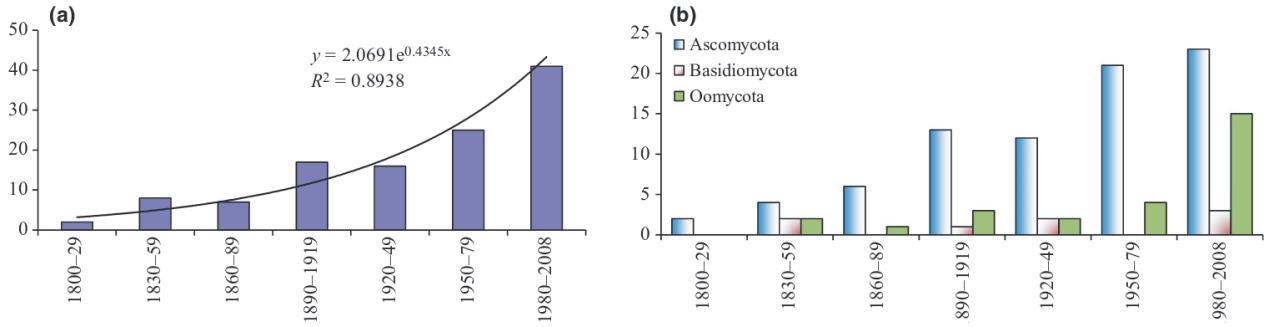


Figure 2, extraite de Santini et al. (2013) : a) Graphique montrant le nombre total de pathogènes forestiers invasifs en fonction de la date d'introduction en Europe. b) Division taxonomique des pathogènes forestiers invasifs en fonction de leur date d'introduction en Europe.

pour l’étude des processus évolutifs associés aux invasions biologiques grâce au séquençage de l’ADN génomique.

II.1 Evolution hôte-pathogène dans le cadre des introductions

Les épidémies émergentes causées par les champignons invasifs pathogènes des plantes sont en général le résultat d'une interaction nouvelle entre la plante hôte et le pathogène, impliquant soit un nouvel hôte soit des populations génétiquement divergentes de la même espèce d'hôte. La résultante de l'interaction entre l'hôte et le pathogène est déterminée par les gènes de résistance* et de tolérance* chez la plante et les gènes associés à la virulence* et d'agressivité* chez le champignon pathogène (Parker et Gilbert, 2004). Trois modes principaux de nutrition à partir de la plante infectée sont classiquement utilisés pour décrire les champignons pathogènes : la biotrophie (le champignon vit dans les tissus vivants de la plante), la nécrotrophie (le champignon tue les tissus cellulaires de la plante pour se nourrir) et l'hémi-biotrophie (le champignon alterne entre les deux stratégies précédentes) (Howlett, 2006). Cette interaction se produit par le biais d'un dialogue moléculaire dont les principaux acteurs fongiques sont des petits peptides secrétés* et des métabolites secondaires*. Du fait de leur rôle dans cette interaction, la diversité de ces molécules et l'évolution des gènes amenant à leur production, entraînent des modifications du pouvoir pathogène du champignon qui pourra conduire à l'adaptation à un nouvel hôte, notamment lors des processus d'introduction, ou au contournement des résistances de la plante. Les champignons pathogènes introduits doivent donc à la fois s'adapter aux conditions abiotiques du nouvel environnement, et éventuellement aux défenses mises en place par leur nouvel hôte. A l'aide d'inoculations expérimentales de sept espèces de champignons pathogènes sur six espèces de plantes herbacées, de Vienne et ses collaborateurs (2009) ont pu montrer deux corrélations qui témoignent que leur capacité à étendre leur gamme d'hôte est limitée à des espèces proches phylogénétiquement de l'hôte d'origine : 1) Le succès d'infection d'un nouvel hôte était corrélé négativement avec la distance génétique entre ce nouvel hôte et l'hôte d'origine de l'agent pathogène inoculé, et 2) le succès d'infection d'un nouvel hôte était corrélé négativement avec la distance génétique entre le pathogène inoculé et le

pathogène naturellement trouvé sur cet hôte. Pour comprendre le succès invasif des champignons pathogènes à travers des modifications génétiques adaptatives, il est important de déterminer les causes et les conséquences de ces adaptations. Les exemples suivants permettent d'illustrer les possibles causes des ces changements adaptatifs et les conséquences qu'ils peuvent entraîner sur la biologie de l'agent pathogène : changement d'hôte (host shift)*, expansion de la gamme d'hôtes (host range expansion)* ou co-évolution avec l'hôte (host tracking)*. Ils soulignent l'apport majeur de la génomique dans la détermination des mécanismes à la base de ces processus et l'importance de déterminer précisément la distribution géographique et la structure génétique des populations invasives et d'origine afin d'étudier les modifications possibles des traits d'histoire de vie de ces pathogènes.

II.1.a Host shift et isolement reproducteur partiel : le cas de *Magnaporthe oryzae*

Magnaporthe oryzae est un champignon de la division des ascomycètes, agent pathogène de plus de 50 espèces de plantes sauvages et cultivées (Ou, 1980). Des études basées sur le séquençage de 10 locus polymorphes ont montré que l'espèce était structurée en plusieurs clades (ou lignées génétiques), chaque clade ayant une capacité d'infection limitée à seulement quelques espèces hôtes (Crouch et al., 2005). Plus récemment, le séquençage du génome entier de huit isolats ayant des spécificités d'hôtes différentes a confirmé que cette spécificité était associée à des divergences génétiques entre les isolats et que le flux de gènes entre les différentes lignées génétiques était limité (Chiapello et al., 2015). La modification de la gamme d'hôtes de ce champignon pathogène pourrait donc être la cause de possibles barrières à la reproduction entre les clades. Cette diminution du flux de gènes entre des populations sympatrique adaptées à différents hôtes a été constatée chez d'autres champignons pathogènes (*Venturia inaequalis*, Leroy et al., 2013; *Botrytis cinerea,* Fournier & Giraud, 2008). Cependant, plus récemment, le séquençage de 81 isolats des différentes lignées génétiques de *M. oryzae* a permis de montrer plusieurs évènements récents d'admixture entre ces lignées (Gladieux et al., 2018). Par cette approche de génomique, les auteurs

suggèrent donc que ces lignées ne sont pas entièrement isolées par des barrières au flux de gènes, et que des échanges génétiques pourraient se réaliser lorsque des souches de lignées différentes se rencontrent sur des hôtes qu'ils peuvent communément infecter. Aussi, des croisements entre des isolats provenant de ces différents clades montrent que la perte de certains gènes de virulence reconnus par les plantes hôtes permettrait l'infection de nouveaux hôtes (Takabayashi et al., 2002). L'utilisation d'un génome de haute qualité obtenu à partir du séquençage de longs brins d'ADN montre qu'une grande proportion des séquences codant pour des petits peptides sécrétés était concentrée dans des régions particulièrement riches en éléments transposables (Bao et al., 2017). Cette étude de génomique a permis d'identifier une implication très probable des éléments transposables dans les gains et pertes de certains gènes de virulence suggérés antérieurement par Takabayashi et ses collaborateurs en 2002.

II.1.b Host tracking et isolement reproducteur : le cas de *Zymoseptoria tritici*

Zymoseptoria tritici est un champignon de la division des ascomycètes. Agent pathogène du blé, il est identifié comme la maladie entraînant les plus lourdes pertes de rendement dans les systèmes agricoles européens (Jørgensen et al., 2014). Des études de génétique des populations basées sur le séquençage de six séquences d'ADN (3,080pb) comprenant 464 sites polymorphes ont permis de montrer que l'origine de ce pathogène coïncide avec la période de domestication du blé dans le croissant fertile il y a 10 000 ans environ, suggérant que la domestication du blé s'est accompagné de la domestication de ce champignon pathogène (Stukenbrock et al., 2007). Ce phénomène est appelé *host tracking* dans la littérature scientifique, et se définit comme la co-évolution d'un hôte et d'un pathogène suite à la domestication d'une plante cultivée, ce qui entraîne une spécialisation du pathogène sur la plante cultivée et rend possible l'invasion de ces cultures exportées à travers le monde (Gladieux et al., 2014). Dans l'étude de Stukenbrock et ses collaborateurs (2007), l'utilisation de seulement six séquences d'ADN n'avait pas permis de mettre en évidence des échanges

génétiques entre les populations de pathogènes s'attaquant au blé cultivé et les populations de pathogène s'attaquant à des herbacées sauvages. En revanche, le séquençage du génome complet de 10 isolats collectés sur des plantes sauvages poussant à proximité de blés cultivés et de deux isolats de *M. graminicola* a permis d'identifier des espèces sœurs nommées *Z. pseudotritici* et *Z. ardabiliae*, qui montraient des niveaux d'infection plus faibles sur le blé que *Z. tritici* (Stukenbrock et al., 2012). Ces espèces sœurs, nommées respectivement S1 et S2, avaient précédemment été identifiées comme deux clades différents sur la base du locus ITS* (Internal transcribed spacer) dans les populations échantillonnées sur plantes sauvages (Stukenbrock et al., 2007). L'étude du génome entier de ces trois espèces sœurs a permis de montrer que de nombreux gènes partagés entre les trois espèces sœurs présentaient un signal de sélection positive, suggérant une possible implication dans l'adaptation de *Z. tritici* au blé cultivé (Stukenbrock et al., 2012). Enfin, l'assemblage du génome de référence de son espèces sœur la plus proche génétiquement (S1) suggère que l'adaptation entraînant la divergence de *Z. tritici* s'est accompagnée de modifications structurales du génome (Stukenbrock et al., 2010). Les auteurs ont confirmé l'absence de flux de gène entre *Z. tritici* et S1, ce qui semble confirmer que la spécialisation de ce pathogène s'accompagne de la mise en place de barrières au flux de gènes et suggère un processus de spéciation rapide. Dans ce cas, le phénomène d'host tracking a mené à un isolement reproducteur entre l'espèce pathogène de la plante cultivée et les espèces sœurs des plantes sauvages. En effet, dans certains agrosystèmes, la densité élevée des individus de la plante hôte et leur diversité génétique réduite pourraient favoriser une adaptation rapide des agents pathogènes à la plante hôte et donc favoriser l'apparition de barrières au flux de gènes avec des espèces proches d'agents pathogènes infectant les plantes sauvages (Stukenbrock & McDonald, 2008).

II.1.c Hybridation récente dans le cadre d'une introduction : le cas d'*Ophiostoma novo-ulmi* et d'*Heterobasidion annosum s.l.*

Nos connaissances sur les processus évolutifs associés au succès invasif de champignons phytopathogènes des écosystèmes naturels sont plus éparses. L'étude de l'espèce *Ophiostoma novo-ulmi* rapporte des cas fréquents d'hybridation entre deux sous-espèces en Europe. Ces cas d'hybridation sont associés à une épidémie dévastatrice de graphiose de l'orme depuis la deuxième partie du XX^{ème} siècle jusqu'à maintenant (Brasier & Buck, 2001). *Ophiostoma novo-ulmi* se sépare en deux sous-espèces, la sous-espèce *novo-ulmi* supposée introduite en Europe pour la première fois en Ukraine dans les années 1970, et la sous-espèce *americana* probablement introduite en Grande-Bretagne dans les années 1960 depuis les États-Unis. En Europe, l'espèce résidente, *O. ulmi*, a été remplacée par l'espèce invasive O. *novo-ulmi* (Brasier, 1986). Malgré la présence de barrières au flux de gènes entre ces deux espèces (Brasier et al., 1998), les allèles nécessaires à la reproduction sexuée (MAT-1, voir encadré 1) et des allèles de compatibilité végétative* ont été échangés entre ces dernières, permettant potentiellement la reproduction sexuée chez l'espèce *O. novo-ulmi* (Paoletti et al., 2006). De plus, de nombreuses études ont permis d'identifier des hybrides entre les deux sous-espèces d'*O. novo-ulmi* présentes en Europe à partir de différents marqueurs moléculaires (synthèse dans Brasier & Kirk, 2010). Dans les régions où les deux sous-espèces *novo-ulmi* et *americana* sont en sympatrie, les proportions d'hybrides entre ces sous-espèces sont élevées (Brasier & Kirk, 2010). Pourtant, si le premier génome d'*O. novo-ulmi* a été publié en 2013 (Forgetta et al., 2013), il n'y a pas à ma connaissance d'étude de génomique comparative entre les deux sous-espèces et les hybrides. Ce genre d'étude pourrait permettre d'étudier quels sont les mécanismes sous-jacents à cette hybridation massive et déterminer les régions génomiques qui donnent un avantage aux hybrides par rapport aux deux sous-espèces en Europe.

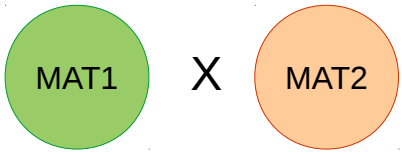
Encadré 1 – Modes de reproduction des champignons

Chez les champignons, des mécanismes moléculaires déterminent si deux isolats haploïdes peuvent fusionner leurs cellules (syngamie) et leurs noyaux (caryogamie). Ces mécanismes sont contrôlés par des locus de type sexuel (Matin type loci = MAT) qui peuvent présenter un ou plusieurs allèles. Il existe deux systèmes d'accouplement principaux : l'hétérothallisme et l'homothallisme (Figure 1). La reproduction sexuée des champignons hétérothalliques implique deux partenaires sexuellement compatibles. Chez les champignons homothalliques, la reproduction sexuée est possible entre deux isolats génétiquement identiques et même sur un unique isolat.

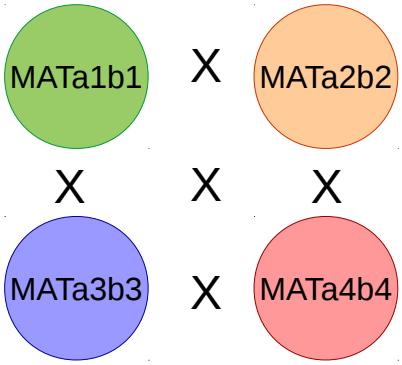
Tous les champignons sont capable de reproduction asexuée par croissance végétative ou synthèse de spores asexuées (Figure 2). Il est important de bien définir le terme d'autofécondation (selfing) chez les champignons ayant un cycle de vie majoritairement haploïde, et de différencier ce mécanisme de l'intra-haploid selfing. Ici, je définis l'autofécondation comme l'accouplement entre deux cellules issues d'une seule méiose, ou à minima entre deux méioses d'un même individu diploïde (Billiard et al., 2011, 2012). L'hétérothallisme et l'homothallisme régulent la possibilité de reproduction sexuée au niveau haploïde, et n'implique pas de restriction d'accouplement au niveau diploïde. Tous les champignons sont donc capables d'autofécondation selon cette définition. L'homothallisme permet de réaliser l'intra-haploid selfing, une autofécondation entre deux cellules issues de la même mitose. A l'inverse, l'hétérothallisme constitue une barrière à l'intra-halpoid selfing, car nécessite deux allèles différents sur le même locus MAT.

Encadré 1 – Modes de reproduction des champignons

Hétérothallisme
La syngamie entre deux isolats est possible s'ils possèdent deux allèles MAT opposés

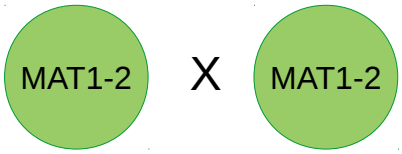


Système bipolaire : Un seul locus MAT code pour deux allèles et donc deux types sexuels

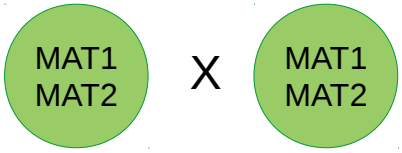


Système tétrapolaire : Deux locus MAT codent pour deux plus plusieurs allèles

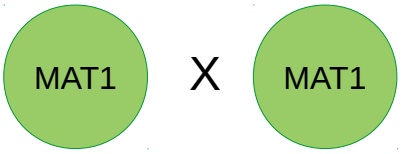
Homothallisme
La syngamie est possible entre deux isolats génétiquement identiques



Les deux allèles MAT sont présents dans le même noyau et ont fusionnés sur un seul locus



Les deux allèles MAT sont présents dans le même noyau et sur deux locus différents



Un seul allèle MAT

Figure 1, modifiée à partir de Ni et al., 2011 : Illustration des différents types d'accouplement les champignons

Encadré 1 – Modes de reproduction des champignons

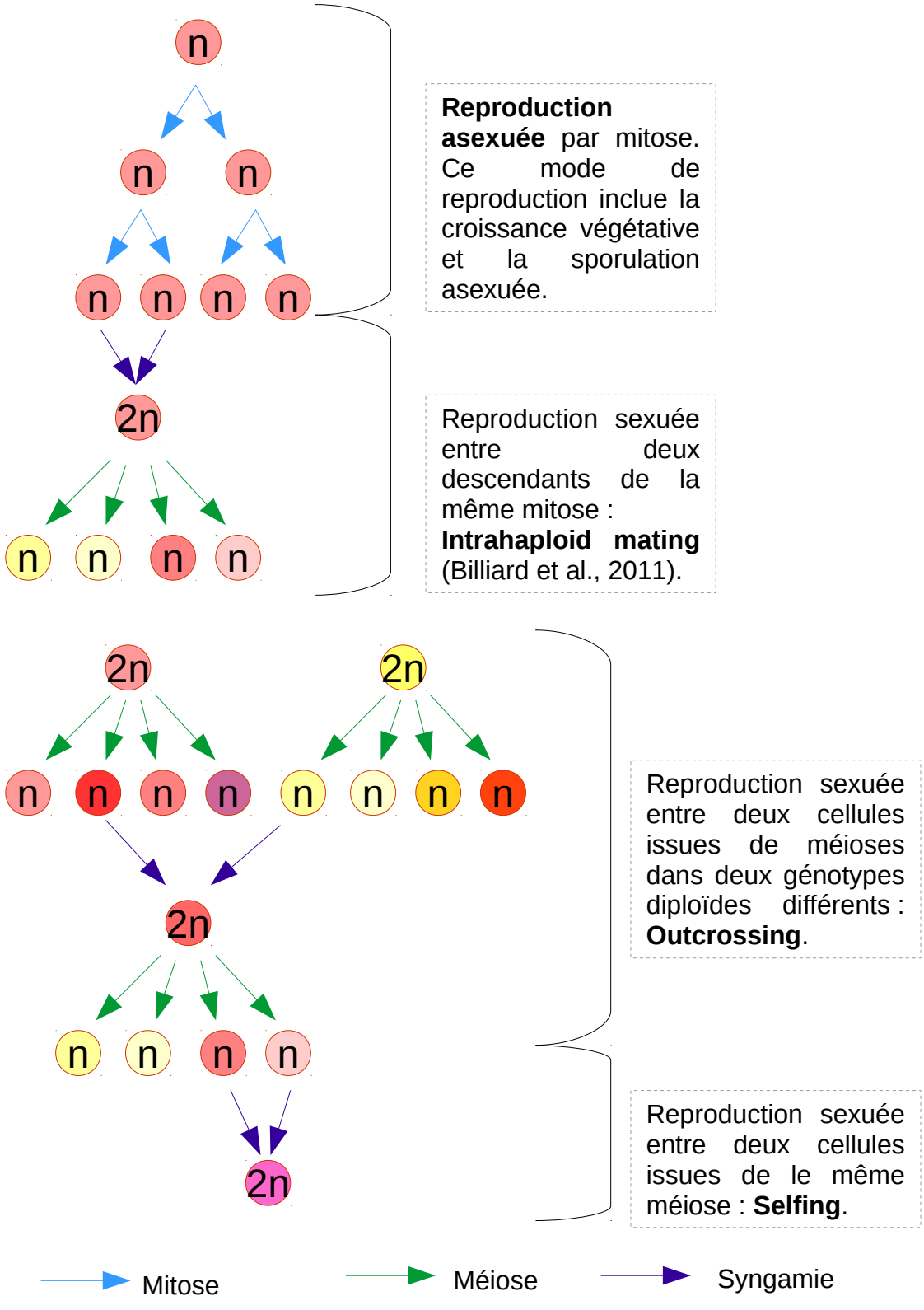


Figure 2 modifiée à partir de Billiard et al., 2012 : Illustration des différents modes de reproduction chez les champignons

Une des rares études en génomique évolutive des champignons forestiers concerne le complexe d'espèces *Heterobasidion annosum sensus lato.* (Sillo et al., 2015). Ce complexe compte cinq espèces de champignons phytopathogènes qui ont divergé en sympatrie ou en allopatrie depuis au moins 14 millions d'années (Dalman et al., 2013). L'espèce *H. irregulare* a été introduite depuis l'Amérique du Nord en Italie au milieu de XX^{ème} siècle où elle co-existe désormais en sympatrie avec l'espèce *H. annosum sensus stricto* (Gonthier et al., 2004). Malgré une divergence entre ces deux espèces datant de 34 à 41 millions d'années, les deux espèces ont conservé le genre *Pinus* comme hôte principal et demeurent hautement interfertiles (Garbelotto & Gonthier, 2013). Les deux espèces ont un pouvoir pathogène* comparable sur les pins européens et les pins nord-américains, pourtant l'espèce introduite d'Amérique du nord, *H. irregulare*, semble plus compétitrice en Europe et exerce une concurrence importante sur l'espèce *H. annosum s.s.*. Le succès invasif d'*H. irregulare* pourrait être lié à des différences dans le potentiel de propagation des deux espèces, car cette dernière est plus performante pour dégrader le bois et fructifier (Giordano et al., 2014). La comparaison de trois génomes de chacune de ces deux espèces a permis de montrer que la synténie* est globalement conservée le long du génome, mais qu'il y a une forte implication des variations structurales dans la création d'îlots génomiques fortement diversifiés entre les deux espèces (Sillo et al., 2015). Leurs résultats confirment que la variabilité génétique entre les deux espèces est plus importante dans les gènes associés à la fructification et la croissance saprophytique (dégradation du bois morts) que dans les gènes associés à l'agressivité (Giordano et al. 2014). Ce travail basé sur un séquençage utilisant des méthodes de deuxième génération* a permis d'identifier des régions du génome susceptibles d'être impliquées dans le succès invasif d'*H. Irregulare*, et suggère un rôle clef des éléments transposables dans la différentiation des espèces dans ces régions. Le séquençage de troisième génération* permettant de mieux assembler les génomes, en particulier dans les régions très variables riches en éléments transposables, permettrait de tester cette dernière hypothèse.

Ces exemples illustrent la diversité des dynamiques évolutives et l'intensité de flux de gènes au sein des populations de champignons pathogènes introduits en dehors de leur aire d'origine, et qui peuvent mener au succès d'une invasion biologique. Dans le cas d'un succès invasif, les champignons pathogènes s'adaptent parfois très rapidement à l'hôte et une admixture importante entre groupes génétiques différents ne semble pas être la clef de ce succès. Au contraire, il n'est pas rare que des barrières au flux de gènes s'installent en parallèle de l'adaptation au nouvel hôte. Au contraire, l'admixture entre les deux sous-espèces d'*Ophistoma novo-ulmi* en Europe semble mener au succès invasif de ces hybrides, mais l'origine et l'implication de ce phénomène restent floues. Dans le cas d'*Heterobasidion irregulare*, l'admixture avec les populations d'*Heterobasidion annosum s.s.* n'a pas été reliée au succès invasif et l'origine des facteurs adaptatifs ainsi que leur transmission dans les populations invasives restent non résolues.

Ces flux de gènes, et donc la possibilité d'admixture, sont aussi déterminés par le mode de reproduction des populations invasives. Ce dernier est sans doute un paramètre clef du succès invasif et peut varier en fonction du contexte environnemental chez de nombreux organismes (Barrett et al., 2008). Un changement de mode de reproduction est souvent constaté chez les populations de champignons pathogènes invasives par rapport à la zone native, la plupart du temps vers un taux de clonalité plus élevé dans l'aire d'introduction (Gladieux et al., 2015). Le système de reproduction, facteur essentiel de la transmission de la variabilité, doit donc être étudié précisément, ainsi que son évolution au cours des introductions pour comprendre les dynamiques évolutives de ces organismes pathogènes.

II.2 Variabilité du système de reproduction et conséquences sur la dynamique invasive et évolutive

Les champignons possèdent des stratégies et modes de reproductions diversifiés [encadré 1]. Beaucoup d'ascomycètes ont été décrits comme ayant un mode de reproduction uniquement asexuée. Mais depuis plusieurs dizaines d'années, de nombreuses études de génétique des populations rendent

compte que presque toutes ces espèces montrent un signal de recombinaison génétique, sans doute associé à une reproduction sexuée, alors même qu'elle n'est pas toujours décrite, ou dans certains cas par des mécanismes dits recombinaison parasexuée* (Nieuwenhuis et James, 2016). Ces études montrent par ailleurs qu'il existe aussi d'importantes variations dans la fréquence de la recombinaison génétique selon les espèces impliquant des différences marquées en terme de structure génétiques des populations. Ces variations peuvent s'expliquer par différentes stratégies écologiques associées à différentes contraintes évolutives comme le coût du sexe lié à la dépense énergétique nécessaire pour attirer et rechercher un partenaire sexuel, la possibilité de contracter des maladies sexuellement transmissibles ou encore le fait que le brassage génétique peut briser des associations de gènes adaptées à un environnement donné comme l'aire d'introduction (Otto 2009). Néanmoins, la reproduction sexuée peut conférer certains avantages évolutifs aux champignons pathogènes. Les spores issues de la méiose permettent la création de nouveaux génotypes grâce à la recombinaison génétique, caractéristique la plus importante de la reproduction sexuée. Ces spores peuvent aussi servir de structure de dispersion et de survie à ces champignons (Taylor et al., 1999 ; Stajich et al., 2009). Un autre bénéfice lié au sexe concerne les régions présentes en multiples copies et qui peuvent altérer l'intégrité du génome. Ces régions peuvent être ciblées et dégradées lors de la méiose par un mécanisme connu spécifiquement chez les champignons (repeat-induced point mutations, RIP, Irelan et al., 1994). Cette caractéristique sera développée en partie IV.2. A ces deux bénéfices de la reproduction sexuée plus spécifiques aux champignons, s'ajoute l'avantage de la recombinaison pour éliminer les mutations délétères, qui, sans recombinaison, peuvent s'accumuler plus rapidement dans les génomes des organismes à reproduction asexuée (cliquet de Muller, Muller 1964) et constituer à terme un cul de sac évolutif. A l'inverse, la reproduction asexuée peut aussi conférer des avantages évolutifs. Elle permet une multiplication rapide de certaines lignées clonales adaptées à un environnement donné et la dispersion de ces génotypes sans aucun partenaire sexuel (Sax et Brown, 2000 ; Barret et al., 2008). L'asexualité peut aussi protéger une population de la recombinaison génétique avec des

génotypes non adaptés à l'environnement (Bolnick et Nosil, 2007). En fait, ces deux modes de reproductions apportent différents avantages, et lorsqu'une espèce les combine dans son cycle de vie elle peut profiter d'un meilleur potentiel invasif. Bazin et ses collaborateurs (2014) ont ainsi suggéré dans une étude théorique basée sur des simulations que les espèces ayant le meilleur potentiel invasif étaient celles avec un taux d'asexualité de 95 %. Ils expliquent ce résultat par le fait qu'à des taux faibles de reproduction sexuée, la recombinaison génétique est trop rare pour casser le déséquilibre de liaison entre certains allèles favorables. Il en résulte que des mutations délétères sont entraînées par effet d'auto-stop* par les mutations qui sont bénéfiques et sous sélection dans l'environnement d'origine, ce qui augmente la variance génétique dans ces populations invasives. L'ensemble de cette variation (favorable ou non) permet aux populations introduites d'avoir une probabilité de contenir des génotypes adaptés à un nouvel environnement, et couplé à la multiplication clonale, d'avoir *in fine* une probabilité d'invasion accrue par rapport aux populations majoritairement sexuées (Bazin et al., 2014). Cette étude illustre clairement l'effet important du mode de reproduction sur la diversité génétique et les capacités adaptatives d'espèces à reproduction partiellement asexuée, comme c'est le cas de nombreux champignons phytopathogènes invasifs. Les différences environnementales entre milieu d'origine et aire d'introduction peuvent donc expliquer le changement de mode de reproduction souvent observé chez les espèces invasives (Gladieux et al. 2015). Néanmoins, au-delà de stratégie reproductive sélectionnée pouvant être à l'origine du succès invasif, des causes proximales associées à l'histoire de l'introduction sont parfois responsables de ces changements.

- Dans certains cas, ce changement de mode de reproduction est simplement dû à l'introduction d'un seul allèle sexuel. Par exemple, certaines populations invasives de *Magnaporthe oryzae* se reproduisent de manière asexuée en France, à Madagascar et en Colombie, alors que la population d'origine en Asie montre des signatures de reproduction sexuée (Saleh et al., 2012a). L'étude du locus sexuel a suggéré que ce changement de mode de reproduction entre l'aire d'origine et ces aires introductions serait dû à l'introduction d'un seul des

deux allèles nécessaires à la reproduction sexuée de cette espèce (Saleh et al., 2012a). Ce changement de mode de reproduction vers l'asexualité causé par l'introduction d'un seul type sexuel a été montré chez d'autres espèces pathogènes invasives : *Phytophthora infestans*, *P. cinnamomi*, *P. ramorum* et *Ophiostoma novo-ulmi* par exemple (Gladieux et al., 2015).

- Dans des populations possédant les deux allèles sexuels, l'adaptation génétique de différentes lignées invasives sur différents hôtes pourrait constituer en soit une barrière écologique suffisante à réduire le flux de gènes entre ces lignées et donc favoriser l'asexualité (Gladieux et al., 2018). En effet, une caractéristique liée au mode de reproduction des champignons pathogènes pourrait favoriser cette réduction du flux de gènes : de nombreux ascomycètes pathogènes réalisent leur cycle de reproduction sexué sur leur hôte. Pour que les croisements soient efficaces, il est donc nécessaire que les deux partenaires soient capables d'infecter le même hôte. Cela signifie qu'une adaptation à un nouvel hôte aura un effet pléiotropique* sur la capacité de s'accoupler sexuellement avec un autre partenaire (Giraud et al., 2010), et peut mener à une diminution du flux de gènes entre des lignées génétiques sympatriques d'une même population, jusqu'à un possible processus de spéciation (Giraud et al., 2008).

- Le passage à une reproduction asexuée exclusive et durable peut elle-même entraîner la diminution de la capacité de reproduction sexuée. La culture *in vitro* de quatre souches de *Magnaporthe oryzae* ayant subit jusqu'à 20 générations de reproduction asexuée (comptant chacune pour au moins 220 multiplications mitotiques) a permis de montrer une diminution de la production de structures sexuelles (les périthèces) par ces souches, au fil des générations (Saleh et al., 2012b).

- Enfin, si certains génotypes de la population invasive sont très adaptés au nouvel environnement grâce à des combinaisons d'allèles le long du génome, la recombinaison génétique causée par la reproduction sexuée entre ces génotypes pourrait casser ces combinaisons alléliques (outbreeding

depression ; Lynch, 1991). Il en résulterait un effondrement de la valeur adaptative moyenne des génotypes intermédiaires et donc un bénéfice à réaliser majoritairement une reproduction asexuée pour préserver ces combinaisons alléliques (Edmands, 1999). Ce phénomène est surtout avantageux dans des milieux présentant des populations d'hôtes génétiquement homogènes comme les cultures fortement anthropisées (Gladieux et al., 2015). Il n'est pas évident que ce phénomène soit favorisé dans des milieux plus hétérogènes dans lesquels évoluent les champignons pathogènes forestiers.

La diminution de la possibilité de reproduction sexuée peut avoir diverses conséquences sur l'évolution des populations invasives. L'asexualité ou l'autofécondation (*intra-haploid mating* chez les champignons haploïdes [encart 3]) peut réduire la taille efficace de la population et ainsi réduire la diversité génétique disponible pour la sélection (Charlesworth et Wright, 2001). Dans ce contexte, la dérive génétique serait un moteur majeur de l'évolution du génome, entraînant notamment l'accumulation de mutations délétères, de réarrangements chromosomiques ou d'éléments répétés (Whittle et al., 2011). L'étude du séquençage de génomes entiers de l'espèce *Zymoseptoria tritici* et de la sous-espèce *Z. tritici* S1 (sa plus proche espèce sœur) a permis de montrer l'absence de synténie et des polymorphismes de taille entre les petits chromosomes (< 1Mb) entre les deux sous-espèces (Stukenbrock et al., 2010). Les auteurs suggèrent que ces petits chromosomes ont une évolution structurale accélérée par rapport au reste du génome, et que ces réarrangements chromosomiques pourraient être liés à l'isolement reproducteur entre ces deux sous espèces. Il est difficile de savoir si ces modifications chromosomiques sont la cause ou la conséquence d'un isolement reproducteur, mais cet exemple suggère que ces réarrangements pourraient survenir rapidement à un niveau intra-populationnel dans des lignées génétiques au flux de gène limité, notamment par un changement de reproduction vers l'asexualité. De plus, les théories évolutionnistes montrent que l'adaptation à un nouvel hôte est plus efficace si le flux de gènes ancestraux vers la population en cours d'adaptation a été réduit ou interrompu

(Gavrilets, 2004), ce qui montre qu'une diminution de la reproduction sexuée dans une population peut avoir des conséquences bénéfiques dans le contexte invasif. A terme, il se peut que l'absence de flux de gènes mène à l'isolement reproducteur entre des lignées génétiques trop différenciées.

La caractérisation du mode de reproduction a fortement bénéficié des apports de l'étude du génome entier pour accéder à l'information haplotypique, permettant à la fois de détecter précisément les traces de recombinaisons et de déséquilibre de liaison*. Globalement, il ressort que l'admixture entre pools génétiques divergents ne semble pas être une solution fréquente pour résoudre le paradoxe génétique des invasions biologiques chez les champignons pathogènes. En outre, les populations invasives montrent souvent un changement de mode de reproduction vers l'asexualité qui ne favorise pas la diversification des génotypes dans les populations. Ceci maintient la question des causes de leur adaptation. La génomique permet de caractériser en profondeur le polymorphisme de ces génomes, dont les variations structurales, qui peuvent avoir des répercussions importantes et rapides dans l'évolution des populations.

II.3 Variations structurales de l'ADN

Les variations structurales, ou réarrangements chromosomiques, sont variées : fusion ou fission de chromosomes, duplication, délétion ou inversion d'une partie de chromosome, et translocation de séquences entre deux chromosomes non homologues. Ces évènements ont généralement lieu pendant la méiose suite à un mauvais appariement des chromosomes homologues, mais pourraient aussi avoir lieu lors de crossing-over mitotiques (Käfer, 1977). Si la plupart est contre-sélectionnée à cause d'un effet fortement délétère, elles peuvent être maintenues dans la descendance (Rieseberg, 2001). Depuis quelques années, les avancées technologiques du séquençage d'ADN ont permis d'obtenir des séquences de plusieurs dizaines de kilobases (*single molecule real-time technologies*), notamment grâce aux technologies proposées par *Pacific Bioscience* (Roberts et al., 2013) et par *Oxford Nanopore*

Technologies (Schneider et Dekker, 2012). Avec le développement des algorithmes intensifs de calcul et des méthodes en bio-informatique, ces séquences ont permis d'assembler et de ré-assembler de nombreux génomes à un niveau parfois chromosomique (Faino et al., 2015), et rendu possible la comparaison de plusieurs génomes entiers. Cette discipline, la génomique comparative, s'appuie sur l'étude de génomes de haute qualité pour détecter des variations intra- ou inter-spécifiques de répertoires de gènes, des variations structurales ou de contenus en éléments répétés.

Le cas de *Zymoseptoria tritici* (présenté en partie II.1.) est un excellent exemple qui illustre l'apport de la génomique comparative dans la détection de variants structuraux entre des génomes et l'impact que peuvent avoir ces variations sur les patrons de recombinaison génétique d'une espèce de champignon pathogène invasif. Une étude basée sur le séquençage d'ADN de seconde génération a permis de détecter la ségrégation d'un polymorphisme de délétion d'un gène de virulence (Zt_8_609) dans différents isolats de *Z. tritici* (Hartmann et al., 2017). Ce gène a été localisé à proximité d'une délétion mesurant entre 0,2 et 66,0 kb selon les isolats, affectant parfois la perte d'un ou de plusieurs exons. Ce polymorphisme suggère que le gène a été perdu plusieurs fois indépendamment dans différents isolats. Les auteurs ont constaté que la perte de ce gène entraînait une virulence accrue envers l'hôte, et proposent que ce réarrangement chromosomique pourrait être sélectionné. L'assemblage de cinq génomes complets d'isolats de *Z. tritici* à partir de données de séquençage de troisième génération (Plissonneau et al., 2018) a permis de montrer que le nombre de gènes perdus ou gagnés entre les différents isolats pourrait être beaucoup plus important (42% des gènes) que suggéré précédemment (Hartmann et al., 2017). De plus, ces gènes dits accessoires présentent un important polymorphisme de présence absence, et sont souvent localisés sur des chromosomes eux-même considérés comme accessoires* qui comptent de nombreux effecteurs susceptibles d'interagir avec la plante hôte, ce qui semble confirmer l'importance des modifications structurales du génome dans une adaptation rapide à l'hôte. Enfin l'établissement d'une carte de recombinaison génétique a montré que le taux de recombinaison génétique était plus élevé chez *Z. tritici* que chez son espèce

sœur infectant des herbacées sauvages, *Z. ardabiliae*, et que le taux de recombinaison dans les gènes effecteurs était significativement plus faible que dans les gènes non-effecteurs chez *Z. ardabiliae*, mais pas chez *Z. tritici* (Stukenbrock et Dutheil, 2018). Cette diversification des gènes en lien avec des réarrangements chromosomiques, pourrait ainsi permettre aux populations invasives de champignons pathogènes de s'adapter rapidement pour contourner les systèmes de défenses des plantes. Un autre exemple est donné par *Verticilium dahliae*, une espèce de champignon phytopathogène qui s'attaque aux racines des hôtes sensibles (Fradin et Thomma 2006). Bien que son cycle sexuel n'ait jamais été identifié, c'est un pathogène capable d'infecter des centaines d'hôtes différents (Inderbitzinand et Subbarao 2014). Plusieurs études ont mis en évidence que certaines régions du génome (appelées « lineage specific regions ») montraient des taux de réarrangements chromosomiques élevés (deJonge et al., 2013), qui seraient impliqué dans la duplication et la perte de gènes (Shi-Kunne et al., 2018). Certains de ces gènes sont des effecteurs, comme Ave1 qui détermine la capacité de *V. dahliae* à infecter de nombreuses plantes hôtes (Faino et al., 2016). Cependant, ces réarrangements chromosomiques ont aussi lieu dans des clades d'espèces proches qui ne sont pas pathogènes (Shi-Kunne et al.,2018). Il est donc peu probable que les réarrangements soient la cause unique de la pathogénicité de *V. dahliae*, mais il n'est pas exclu qu'ils puissent améliorer les capacités d'adaptation de l'espèce pathogène en accélérant l'évolution de certaines régions génomiques.

Les réarrangements chromosomiques peuvent aussi mener à la modification du mode de reproduction. Les espèces pathogènes du genre *Cryptococcus* présentent un système de reproduction bi-polaire (cad deux allèles à un locus de type sexuel MAT) alors que l'espèce *Cryptococcus amylolentus* a un système tetra-polaire, déterminé par deux locus MAT. L'assemblage du génome de deux isolats de *C. amylolentus* a permis de montrer que les deux locus étaient présent sur deux chromosomes différents, chacun étant lié au centromère de son chromosome (Sun et al., 2017). La comparaison de ces deux génomes avec des génomes d'espèces de *Cryptococcus* pathogènes a mis en évidence que des réarrangements

chromosomiques étaient la cause du changement du système tetra-polaire vers le système bipolaire. Une recombinaison ectopique* entre les deux centromères des chromosomes portant les deux locus MAT est permise par l'homologie entre les séquences répétées présentes dans les centromères, et a mené à la translocation des deux locus sur le même chromosome (Sun et al., 2017). Les chromosomes sexuels sont souvent sujets à des réarrangements chromosomiques et de tels réarrangements limitent les recombinaisons entre chromosomes homologues. L'assemblage du génome entier de deux isolats de types sexuels différents de *Microbotryum lychnidis-dioicae*, a permis de mettre en évidence un nombre très important de modifications chromosomiques entre les chromosomes portant les gènes de type sexuel d'isolats de type opposés (Badouin et al., 2015). Ces variations structurales s'accompagnent d'une recombinaison génétique supprimée sur près de 90 % du chromosome et de signes de dégénération avec la perte de centaines de gènes sur chacun des chromosomes.

Ces exemples de variations structurales dans les génomes de champignons pathogènes sont dans chaque cas liés à la présence de séquences répétées mobiles et mutagènes dans les génomes : les éléments transposables. Les éléments transposables sont vus comme des agents centraux de la structuration des génomes de champignons (Daboussi et Capy, 2003). Ils sont associés à des recombinaisons ectopiques, des explosions de prolifération de ces éléments menant à des variations de tailles de génomes, la dégénérescence de certaines régions du génome et la diversification de gènes ou encore des inactivations épigénétiques. Ces éléments transposables sont ainsi vu aujourd'hui comme un moteur de l'évolution chez de nombreux organismes, notamment les champignons pathogènes, et peuvent être particulièrement importants lors de processus d'invasions.

III Les éléments transposables comme moteurs de l'évolution du génome

III.1 Historique sur les études des éléments transposables

C'est dans les années 1940 que Barbara McClintock, chercheuse américaine qui travaillait sur la génétique du maïs, propose pour la première fois que des gènes puissent changer de position à l'intérieur du génome. Elle identifia deux locus génétiques qu'elle nomma Dissociator (Ds) et Activator (Ac), le premier capable de causer une cassure d'un chromosome et le deuxième régulant la transposition de Ds. Quelques années plus tard, elle observa que le changement de position de Ds impliquait un changement de couleur des grains des épis. Le déplacement de Ds modifiait l'expression de certains gènes codant pour la couleur des grains, causant un effet de mosaïque de couleurs sur les épis. Lorsqu'elle proposa que ces éléments de contrôle (ou éléments transposables comme elle les nomma par la suite ; McClintock, 1953) pouvait jouer un rôle important dans l'évolution, la communauté scientifique reçut cette découverte novatrice de façon « perplexe, voire hostile ». Jusque dans les années 1980, les éléments transposables furent considérés comme des déchets de l'ADN (*junk DNA*) et étaient souvent ignorés dans les études de biologie évolutive. En 1980, Doolittle et Sapienza proposaient que ces éléments transposables produisaient uniquement des effets délétères sur le génome, ce qui a conduit à l'idée que ces éléments auraient une évolution « égoïste », c'est à dire indépendante du reste du génome, terme initialement introduit par Richard Dawkin (*The selfish gene*, 1976). Les travaux sur l'élément P (un élément transposable similaire au Ds du maïs) chez les drosophiles ont aidé à populariser les éléments transposables dans la communauté scientifique. Notamment, Anxolabéhère et ses collaborateurs ont montré en 1988 par hybridation d'ADN que de nombreuses populations naturelles de drosophiles avaient été envahies par cet élément transposable très récemment et en seulement 30 ans. A partir des années 1990, grâce aux avancées des

techniques de séquençage d'ADN, les génomes humain, du maïs, de la drosophile et de nombreuses autres plantes ont pu être obtenus, ce qui a provoqué un regain d'intérêt pour ces éléments transposables. La grande quantité de ces éléments trouvée dans le génome de ces organismes a amené la communauté scientifique à ne plus voir les éléments transposables simplement comme des séquences d'ADN purement égoïstes qui envahissent les génomes, mais à également les placer dans un contexte plus large dans lequel ils pourraient avoir un impact sur le génome qu'ils colonisent, notamment en régulant l'expression ou modifiant la structure et la fonction de certains gènes. Environ 44 % du génome humain serait composé d'éléments transposables et ils représenteraient plus de 60 % du génome du maïs. Christian Biémont, dans une synthèse bibliographique écrite en 2010, exprime que la capacité des éléments transposables à être des séquences régulatrices (ce que suggérait déjà Barbara McClintock) a été difficile à accepter par la communauté scientifique, probablement car il est difficile d'accepter l'idée que l'expression de nos gènes puisse être modifiée par ces éléments (Biémont, 2010). Il a ensuite été montré que depuis la divergence entre les chimpanzés et l'humain il y a environ 6 millions d'années, plusieurs milliers de copies d'éléments transposables ont été insérées dans le génome humain (Lee et al., 2008 ; Baskaev & Buzdin, 2012), et environ 40 sous-familles d'éléments transposables seraient actifs dans notre génome (Mills et al., 2007). Depuis quelques dizaines d'années, un intérêt scientifique s'est développé sur la contribution positive que peuvent avoir ces éléments sur la régulation et la diversification des gènes (Sinzelle et al., 2009). Des événements d'intégrations de ces éléments transposables qui ne sont plus capables de se déplacer dans le génome ont été mis en évidence chez de nombreux eucaryotes (Volff, 2006). C'est le cas par exemple de la protéine Rag1, désormais régulée dans le génome des vertébrés à mâchoire (gnathostomes), et qui permet de générer une importante diversité des immunoglobulines impliquées dans le système immunitaire (Kapitonov et Jurka 2007). Les éléments transposables sont désormais pleinement considérés comme ayant un impact important sur le génome de leur hôte avec des conséquences directes sur la fitness des individus (Diwash et al., 2017).

III.2 Fonctionnement et impact des éléments transposables, et mécanismes de régulation

Les éléments transposables sont des séquences d'ADN capables de se multiplier dans le génome de leur hôte. Ces éléments sont autonomes car leur séquence contient toutes les informations nécessaires à leur transposition. Il existe deux grandes classes de transposons qui diffèrent dans leur mécanisme de transposition. La classe I concerne les transposons à ARN (rétrotransposons) qui se transposent grâce à un intermédiaire sous forme d'ARN. Ces éléments fonctionnent selon un schéma de copié-collé. Parmi la classe I sont trouvés les rétrotransposons à LTR (long terminal repeats), qui contiennent deux superfamilles d'éléments, appelés gypsy et copia, très représentés chez les champignons (Muszewska et al., 2011). La classe II concerne les transposons à ADN qui se transposent sous forme d'ADN selon un schéma de coupé-collé. Si ce mécanisme de coupé-collé ne permet pas en soit à ces éléments transposables de se multiplier dans les génomes, il peuvent toutefois augmenter leur nombre de copie par des mécanismes indirects (Daboussi 1997). Lors de la réplication de l'ADN durant la phase S de la méiose et de la mitose au cours de laquelle chaque chromosome est dupliqué, la transposition d'un élément transposable de classe II peut avoir lieu depuis un bras chromosomique non répliqué vers un bras chromosomique déjà répliqué ce qui résulte en un gain d'une copie. L'élément activator, premier élément transposable détecté par Barbara McClintock qui fait partie des transposons à ADN et de la super famille des hAT, est répliqué de cette façon dans le génome du maïs (Dash & Peterson, 1994). Ces éléments transposables peuvent induire des modification de l'expression de certains gènes chez de nombreux eucaryotes, incluant la levure *Candida albicans* (Atkinson, 2015). Une grande majorité des éléments transposables possède des régions répétées à leurs extrémités qui servent de sites de reconnaissances pour leur excision et leur intégration dans le génome de leur hôte (Wicker et al., 2007). Ces régions répétées peuvent induire des recombinaisons ectopiques dues à l'appariement entre ces séquences lorsqu'elles sont présentes sur différents chromosomes

lors de la méiose, comme cela a été montré entre différentes espèces de *Cryptococcus*, exemple présenté précédemment (Sun et al., 2017).

Au regard des effets délétères que peuvent avoir les éléments transposables sur le génome, les organismes ont évolué vers différents mécanismes moléculaires spécialisés dans leur inactivation ou la répression de leur expression (Daboussi et Capy, 2003). Les champignons ont notamment développé plusieurs mécanismes de défense pour enrayer l'invasion de leur génome par les éléments transposables : le RIP (Repeat-induced point mutation), le MIP (methylation induced premeiotically) et un mécanisme épigénétique impliquant des ARNs d'interférence (Irelan et Selker, 1996). Les deux premiers mécanismes nécessitent un appariement homologue des chromosomes qui a lieu pendant la méiose, et induisent des mutations des nucléotides cytosines en thymine (C vers T) entraînant des modifications de méthylations de la séquence ciblée, et à terme, l'inactivation de la transcription de l'élément transposable (Amselem et al., 2015). Le dernier est un mécanisme post-transcriptionnel et peut avoir lieu pendant la croissance végétative du champignon, et est similaire aux mécanismes de défense trouvés chez les plantes et les animaux.

Le RIP a lieu exclusivement pendant le cycle sexuel (Watters et al., 1999). Ce mécanisme reconnaît les duplications supérieures à 400bp avec une identité nucléotidique supérieure à 80 % et induit des mutations C:G vers T:A (Figure X extraite de Galagan et Selker 2004). Il est extrêmement efficace et peut en un seul passage modifier plus de 30 % des paires G:C contenues dans la séquence dupliquée (Cambareri, E.B. et al., 1989). Ces modifications génétiques peuvent avoir pour conséquence sur la méthylation de l'ADN et supprimer l'expression de certains gènes proches. De plus, le RIP cible les séquences dupliquées sans déterminer si ces séquences sont étrangères ou natives dans le génome. Ceci qui peut conduire à l'inactivation de gènes dupliqués, comme c'est le cas dans le génome de *Neurospora crassa* où seulement six des 10 000 gènes fonctionnels sont dupliqués (Galagan et Selker, 2004). A l'inverse, les mutations provoquées par le RIP peuvent permettre la diversification de certains gènes situés à proximté des éléments transposables. C'est le cas chez l'espèce *Leptosphaeria maculans*, champignon

hémibiotrophe qui s'attaque aux brassicacées, où le RIP provoque une diversification rapide des effecteurs situés dans des régions riches en TEs (Rouxel et al., 2011 ; Daverdin et al., 2012). Des mutations non-synonymes ou causant l'apparition de codons stop prématurés ont été produites par le RIP dans des effecteurs proches d'éléments transposables dégénérés et pourraient être sélectionnées positivement car elles permettraient aux génotypes de *L. maculans les portant* d'échapper aux mécanismes de reconnaissance des plantes hôtes.

Les modes de reproduction et les facteurs démographiques influent eux aussi la dynamique des éléments transposables. Il a été montré par simulation que lorsque l'autofécondation augmente dans une population, la probabilité d'invasion par un élément transposable diminue (Boutin, Le rouzic et Capy, 2012). La transition brusque d'une reproduction sexuée croisée vers l'autofécondation mène à l'homozygotie des insertions de TEs et les copies sont systématiquement perdues après quelques centaines de générations. Enfin, ces simulations ont montré qu'une taille réduite de population, et donc d'avantage soumises à la dérive génétique, mène souvent à une perte de l'activité des éléments transposables.

III.3 Les éléments transposables chez les champignons phytopathogènes

Les génomes des champignons pathogènes des plantes sauvages et cultivées incluent souvent des régions riches en éléments transposables (Raffaele & Kamoun, 2012 ; Anselem et al., 2015). L'émergence de concepts comme le « two-speed genome » de certains champignons pathogènes filamenteux et l'accumulation de données génomiques qui montrent des architectures de génomes différentes selon les espèces nous poussent à reconsidérer l'étude de l'évolution de ces pathogènes dans une vision plus générale qui intègre l'effet que peuvent avoir des éléments transposables sur l'évolution (Möller et Stukenbrock, 2017). En effet, si la reproduction sexuée conduit globalement à augmenter la diversité génotypique, les éléments transposables pourraient servir de moteur de diversification génétique chez

des champignons, notamment ceux se reproduisant majoritairement voir quasi-exclusivement par voie asexuée.

Il existe de nombreux exemples de diversification génétique associée aux éléments transposables chez les champignons pathogènes des plantes cultivées. Des études de génomique ont montré que les régions « lineage specific » de *V. dahliae* sont enrichies en éléments transposables supposés actifs et en effecteurs, et que ces éléments transposables jouent un rôle dans la plasticité, la recombinaison et l'expression des gènes de ces régions (deJonge et al., 2013). De même, une récente étude de génomique comparative a montré que les éléments transposables pourraient substituer la fonction de la recombinaison sexuée et contribuer au maintient de la diversité génétique chez *Magnaporthe oryzae* (Yoshida et al., 2016).

Il y a une focalisation très forte sur les champignons pathogènes dans le domaine de la génomique comparative et évolutive. A l'inverse, l'étude de l'impact des éléments transposables sur l'évolution des pathogènes d'arbres forestiers a peu été considérée jusqu'à présent, et seules quelques études très récentes traitent de ce sujet. Ainsi, une étude basée sur 19 génomes de *Dothistroma septosporum* récoltés à travers le monde a révélé un contenu en éléments transposables et un contenu en SNPs beaucoup plus important chez des souches d'Amérique centrale comparées aux autres (Ozturk et al., 2019). Ces souches ayant été isolées sur une espèce de pin différente, les auteurs suggèrent que ces éléments transposables pourraient être associés à la diversification de ces souches et à l'adaptation à l'hôte. Le séquençage et l'assemblage du génome de quatre espèces d'*Armillaria*, un champignon pathogène forestier infectant les racines et provoquant la mort de nombreux résineux à travers le monde, a permis de déterminer que leur contenu en éléments transposables n'était pas plus important par rapport à des espèces proches non pathogènes et que leur distribution n'est pas concentrée dans certaines régions riches du génome, et ainsi contraire aux hypothèses du « two-speed genome » (Sipos et al., 2017). De fait, nos connaissances sur le rôle des TEs dans l'évolution du génome des champignons pathogènes forestiers invasives demeurent clairsemées. La caractérisation de ces éléments dans de futures études de génomique comparative pourrait ainsi permettre

d’élucider leur impact dans les dynamiques d’invasions d’écosystèmes faiblement anthropisés.

IV Modèle d'étude : *Cryphonectria parasitica*

IV.1 Biologie de l'espèce et mode de reproduction

Cryphonectria parasitica est un Champignon filamenteux phtyopathogène de la classe des Sordariomycetes (Division des Ascomycetes), un clade regroupant de nombreux pathogènes. La forme imparfaite de C. parasitica, sa phase asexuée, est appelée *Endothia parasitica*. Sa gamme d'hôtes est majoritairement représentée par le genre *Castanea*, mais il est possible de détecter des infections sur *Quercus petraea* dans des parcelles fortement infectées (Bissegger et al., 1991). Son mycélium se développe dans le cambium de l'hôte, sous l'écorce, et provoque sur les troncs et les branches des lésions et des chancres. Il est capable de se reproduire asexuellement grâce à des spores asexuées (les conidies; Figure 3 a et c) qui sont disséminées sur de courtes distances, et des spores sexuées (les ascospores; Figure 3 b) disséminées sur de plus longues distances. La formation de ces spores sexuées nécessite la fécondation entre deux individus de types sexuels (MAT) opposés : l'hétérothalisme (encadré 1, figure 1). Pourtant, des études sur des populations naturelles de *C. parasitica* ont montré que de l'intra-haploid mating (définit dans l'encadré 1) est possible en milieu naturel, notamment car des isolats portent parfois les deux types sexuels (McGuire et al., 2004), mais que ce processus est difficilement reproductible en laboratoire (Marra & Milgroom, 2001). La présence des deux types sexuels dans un même isolat pourrait être due à la présence de deux noyaux dans chaque cellule, car la production d'une carte génétique a montré que le locus MAT est présent dans une région fortement diversifiée génétiquement, dans laquelle la recombinaison est fortement diminuée (Kubisiak & Milgroom, 2006). Ce champignon pathogène est donc capable de reproduction asexuée, de reproduction sexuée entre isolats possédant deux allèles différents sur le locus MAT et d'intra-haploid mating lorsque les deux allèles sont présents dans un même isolat. Enfin, la fusion de deux mycéliums, appelée anastomose, est possible lorsque deux souches ne sont pas incompatibles. Cette compatibilité végétative est

déterminée par au moins six locus (vic locus) bi-alléliques, et la fusion des cellules mycéliennes est possible lorsqu'elle portent les mêmes allèles sur ces locus (Cortesi & Milgroom, 1998).

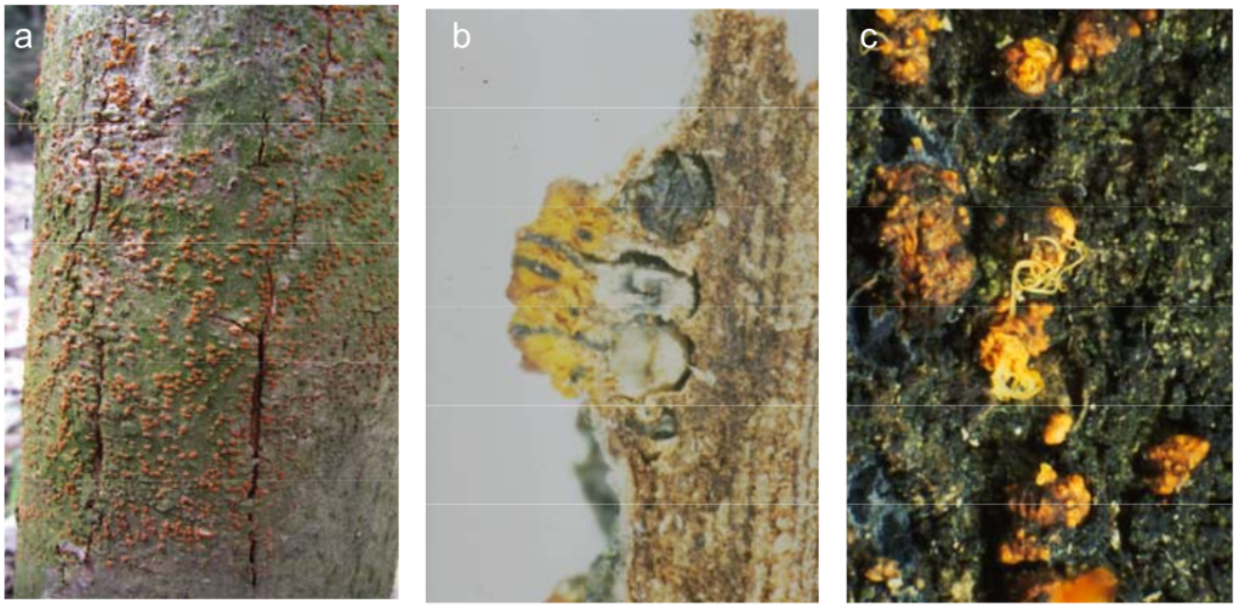


Figure 3 extraite de Rigling & Prospero (2018) : Photos montrant la sporulation sexuée et asexuée de Cryphonectria parasitica. a) Sur une écorce infectée, le champignon produit des pustules oranges qui abritent les fructifications sexuelles et asexuelles. b) Fructification sexuelle (périthèce). c) Fructifications asexuelles (pycnides) desquelles sont sécrétées les spores asexuées (conidies).

IV.2 Le contexte de double introduction en Europe

Originaire d'Asie, le champignon a été introduit à la fin du XIXème siècle en Amérique du Nord, probablement depuis le Japon (Milgroom et al., 1996). Il a causé une épidémie sur les châtaigniers nord-américain, *Castanea dentata*, et a causé la quasi-disparition des populations naturelles. En Europe, le pathogène a été signalé pour la première fois en 1938 en Italie (Biraghi, 1946) et infecte le châtaignier européen, *castanea sativa*. Il s'est ensuite répandu dans la majeure partie du sud de l'Europe (Rigling et Prospero, 2017), avec une colonisation récente plus au nord (Robin et al., 2017). Le génotypage de dix marqueurs moléculaires (locus microsatellites), il a été mis en évidence que les populations européennes étaient issues de plusieurs événements d'introduction différents. Les populations d'Amérique du nord seraient les populations d'origine d'une introduction ayant eu lieu en Italie et en Suisse (Dutech et al., 2012 ; Prospero & Rigling, 2012), tandis que les populations chinoises ont été la source d'isolats introduits dans le sud-ouest de la France (Dutech et al., 2012). Des premières mentions de chancres du châtaignier date de 1947 sur la côte nord de l'Espagne, et pourraient concorder avec cette introduction depuis les populations d'origine asiatique (Darpoux, 1949 ; Robin et al., 2009). *Cryphonectria parasitica* est désormais présent dans la majorité de l'aire de répartition de *Castanea sativa* en Europe (Figure 4). L'étude de la structure des populations européennes a mis en évidence une forte structure clonale dans les populations introduites (Dutech et al., 2010 ; Prospero & Rigling, 2012), contrairement au populations asiatiques et nord-américaines. Pourtant, les structures sexuelles sont régulièrement détectées sur les chancres dans les parcelles échantillonnées (Robin et al., 2009). Les deux allèles du locus MAT sont détectés en égales proportions dans ces parcelles, et même au sein d'une même lignée clonale (Dutech et al., 2010). Enfin, 32 % des 994 isolats échantillonnés en France et génotypés présentent un génotype rare, qui diffère souvent de seulement quelques allèles microsatellites (Dutech et al., 2010 ; Robin et al., 2017).

Contrairement aux épidémies dévastatrices de chancre qui ont eu lieu en Amérique du Nord, l'épidémie Européenne a été moins destructrices, même si *C. parasitica* demeure une des causes principales de maladie de *Castanea sativa* avec plusieurs espèces de *Phytophthora* (Vettraino et al., 2001). Les explications à ces épidémies moins importantes seraient que les châtaigniers européens sont moins sensibles qu les châtaigniers nord-américains, et qu'un virus (Cryphonectria hypovirus1, CHV-1) qui induit l'hypovirulence en diminuant la fitness des souches de *C. parasitica* infectées est présent en Europe, probablement dû à l'introduction de souches hypovirulentes directement depuis l'Asie (Nuss, 1992 ; Milgroom & Cortesi, 2004). Des souches de *C. parasititica* hypovirulentes sont notamment utilisées comme moyen en de lutte biologique car leur inoculation sur des châtaigniers infectés par *C. parasitica* peut mener à la transmission du virus vers la souche présente par fusion des deux mycéliums (Robin & Heiniger, 2001). Néanmoins, la reproduction sexuée, empêchant la transmission du virus à la descendance, et

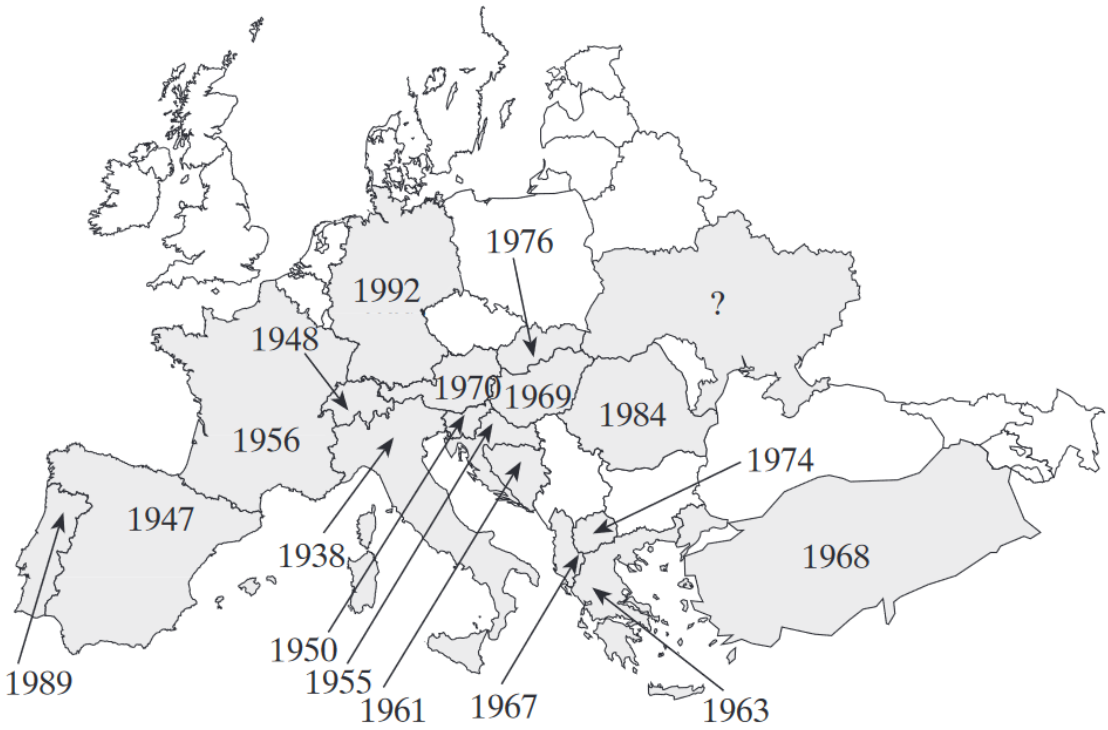


Figure 4 extraite de Robin & Heiniger (2001) : Carte de l'Europe montrant la présence de Cryphonectria parasitica et l'année de sa première observation par pays.

la diversité des groupes de compatibilité végétative, régulant la fusion des mycéliums entre deux génotypes, sont deux barrières à la transmission du virus dans les populations de *C. parasitica* (Montenegro et al., 2008).

IV.3 Importance de l'étude de *C. parasitica*

Au regard des concepts introduits, la double introduction de *C. parasitica* constitue un contexte d'étude adéquat pour comprendre les processus biologiques et mécanismes génétiques impliqués dans le succès invasif d'un champignon phytopathogène.

Premièrement, son introduction en Europe est estimée à moins d'une centaine d'année, ce qui représente un temps évolutif très court. Certaines invasions sont causées par des champignons pathogènes qui co-évoluent depuis plusieurs milliers d'années avec leurs espèces d'hôtes, comme c'est le cas lors de processus d'host tracking. L'introduction récente de *C. parasitica* et son succès invasif sur plusieurs espèces d'hôtes différentes de celles son aire d'origine en Asie permet d'étudier les mécanismes génétiques mis en place lors des phases initiales d'une invasion. De plus, les populations sont composées de deux groupes génétiques divergents, issus d'au moins deux processus d'introduction depuis l'aire d'origine et depuis l'aire d'introduction en Amérique du Nord. Ce contexte permet d'étudier en détail le processus d'admixture entre ces groupes génétiques, et de déterminer s'il constitue une réponse au paradoxe génétique des espèces invasives. Aussi, l'introduction successives de génotypes en Amérique du nord depuis l'Asie, puis en Europe depuis l'Amérique de nord, permet d'étudier l'effet de goulots d'étranglement démographiques et génétiques successifs et leur impact sur la diversité nucléotidique et l'architecture du génome de cette espèce.Deuxièmement, les études de génotypage suggèrent un changement de mode de reproduction vers l'asexualité en Europe, même si la possibilité de reproduction sexuée n'est pas exclue. Ce contexte permet d'étudier en détail l'ampleur et les causes d'un tel changement, qui semble mener à un isolement reproducteur au moins partiel entre les lignées clonales. Enfin, le two-speed génome est souvent invoqué chez des champignons pathogènes comme moteur d'évolution, notamment chez les champignons pathogènes de plantes cultivées. L'étude des populations invasives et des populations d'origine de Cryphonectria

parasitica permet de tester si ce cadre conceptuel peut s’adapter à un champignon pathogène forestier.

V Objectifs de la thèse

Nous avons vu que le mode de reproduction des champignons pathogènes est susceptible de changer au cours d'une introduction dans un nouvel environnement et peut entraîner une dynamique évolutive différente de celle observée en zone native. Lesétudes précédentes d'un petit nombre de marqueurs microsatellites ont permis de mettre en évidence que les populations invasives de *C. parasitica* en Europe sont structurées en lignées supposées clonales, mais que de rare croisements seraient parfois possible entre ce lignées (Dutech et al., 2010, 2012). Le premier objectif de la thèse a été de vérifier l'existence supposées d'échanges génétique entre les lignées clonales caractérisées par les SSR et d'en estimer l'ampleur. Par ailleurs, l'émergence apparente de nouvelles lignées clonales au cours de l'expansion de la maladie en Europe, notamment dans la région du massif central et ensuite plus au nord posait la question de leur origine : nouvelles introductions à partir de foyers non décrits jusqu'à présent ou recombinaison entre les principales lignées du sud de la France. Les données génomiques acquises dans ce travail ont ainsi permis de tester si la double introduction de *C. parasitica* en Europe s'était accompagnée d'une admixture entre les pools génétiques nord-américains et asiatiques, même limité, et on pu permettre l'émergence de ces nouvelles lignées génétiques. Dans le contexte général des invasions décrit ci-dessus, cette admixture aurait pu être susceptibles de générer des lignées mieux adaptées aux conditions environnementales de l'aire d'introduction européenne. Ces travaux sont présentés en chapitre I et ont fait l'objet d'une publication dans le journal *Fungal Genetics and Biology* (Demené et al., 2019).

Par ailleurs, l'observation des croisements limités entre les lignées de *C. parasitica* en Europe posent la question d'éventuelles barrières à la recombinaison génétique ; les lignées co-existant souvent dans une même parcelle forestière (Dutech et al., 2008, 2010). Une des hypothèses à ces barrières génétiques serait l'existence de variations structurales entre les groupes génétiques indépendamment introduits en Europe limitant les

croisements en perturbant l'appariement des chromosome homologue lors de la méiose. La première étape de ce travail a été d’assembler un nouveau génome de référence de haute qualité. Ce travail est présenté en annexe I, et fera l’objet d’une publication technique. Trois autres génomes ont ensuite été assemblés et comparés pour tester l’hypothèse de l’apparition de réarrangements chromosomiques ou de variations structurales impliquant certaines partie du génome et certains gènes entre ces génomes au cours du processus d’introduction depuis le centre d’origine en Asie. Ce travail est présenté en Chapitre II où une première version d’un article que je souhaite soumettre à l’issue de cette thèse est présentée.

Enfin, malgré la persistence de lignées génétiques majoritaires le long du gradient de colonisation de *C. parasitica*, un quart des populations françaises est représenté par des génotypesdistincts de ces lignées majoritaires, c'est à dire différents par un ou plusieurs allèles microsatellites avec une de ces lignées (Dutech et al., 2010 ; Robin et al., 2017). Des questions se posaient sur la contribution de ces génotypes plus rarement répétés dans la population française dans l’émergence de nouvelles lignées au centre et au nord de la France, complétant les résultats du chapitre 1 centré sur l'évolution des lignées clonales majoritaires. Vingt-neuf nouveaux génotypes génotypés préalablement à partir de dix locus microsatellites et distincts des lignées majoritaires ont donc été choisis et séquencés afin de comprendre quel lien de parenté ils avaient par rapport aux lignées principales et de quelle manière leur diversité génétique contribuait à l'émergence de nouvelles lignées dans le nord. Par ailleurs, l'existence de génotypes proches à un ou deux allèles microsatellites des lignées clonales majoritaires, observées dans les études antérieures, pourraient s’expliquer par une hybridation limitée entre les lignées principales, suivie de rétro-croisements successifs avec une des lignées parentales. Un projet de séquençage de plusieurs dizaines d' isolats échantillonnés dans une parcelle où plusieurs lignées génétiques majoritaires sont établies est présenté dans le dernier chapitre. Ce projet devrait permettre d'étudier plus précisément l'asymétrie de flux de gènes entre les lignées génétiques majoritaires et de tester cette hypothèse de rétro-croisement des hybrides avec une des lignées clonales majoritaires. Les attendus de cette

hypothèse associée au mode de reproduction partiellement asexuée de *C. parasitica* sont présentés et confrontés aux attendus d'un modèle où les hybrides seraient contre-sélectionnés.

Chapitre 1 : Whole-genome sequencing reveals recent and frequent genetic recombination between clonal lineages of Cryphonectria parasitica in western Europe

Whole-genome sequencing reveals recent and frequent genetic recombination between clonal lineages of *Cryphonectria parasitica* in western Europe

Running title: Recombinations between *C. parasitica* lineages

Arthur Demené[1], Ludovic Legrand[2], Jérôme Gouzy[2], Robert Debuchy[3], Gilles Saint-Jean[1], Olivier Fabreguettes[1], Cyril Dutech[1]

[1] BIOGECO, INRA, Université de Bordeaux, 69 route d'Arcachon, Cestas F-33610, France.

[2] LIPM, Université de Toulouse, INRA, CNRS, Castanet-Tolosan F-31326, France.

[3] Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198, Gif-sur-Yvette cedex, France

Correspondent author: D. Arthur, E-mail: arthur.demene@u-bordeaux.fr

Postal address: INRA - UMR 1202 BIOGECO - 69 route d'Arcachon - 33610 Cestas – France

Declarations of interest: none

Abstract

Changes in the mode of reproduction are frequently observed in invasive fungal populations. The ascomycete *Cryphonectria parasitica,* which causes Chestnut Blight, was introduced to Europe from North America and Asia in the 20th century. Previous genotyping studies based on ten microsatellite markers have identified several clonal lineages which have spread throughout western Europe, suggesting that asexuality was the main reproductive mode of this species during colonization, although occasional sexual reproduction is not excluded. Based on the whole-genome sequences alignment of 46 *C. parasitica* isolates from France, North America and Asia, genealogy and population structure analyses mostly confirmed these lineages as clonal. However, one of these clonal lineages showed a signal of strong recombination, suggesting different strategies of reproduction in western Europe. Signatures of several recent recombination events within all the French clonal lineages studied here were also identified, indicating that gene flow is regular between these lineages. In addition, haplotype identification of seven French clonal lineages revealed that emergences of new clonal lineages during colonization were the result of hybridization between the main expanding clonal lineages and minor haplotypes non-sequenced in the present study. This whole-genome sequencing study underlines the importance of recombination events in the invasive success of these clonal populations, and suggests that sexual reproduction may be more frequent within and between the western European clonal lineages of *C. parasitica* than previously assumed using few genetic markers.

Key-Words:

Bayesian inferences, clonal evolution, intra-haploid mating, recombination rates, whole genome sequencing

I Introduction

Worldwide, the expansion of a few clonal lineages (i.e., identical or closely related multi-locus genotypes) has often been described regarding populations of invasive pathogenic fungi (e.g., Steimel *et al.,* 2004; Raboin *et al.,* 2007; Goss *et al.,* 2014). Such population structure is usually associated with mainly asexual reproduction in the introduced area; unlike sexual reproduction, which is more frequently reported in the native area of the species (Gladieux *et al.,* 2015). Asexual reproduction is assumed to provide at least two benefits during colonization. First, it allows the rapid multiplication and dispersal of genotypes without any mating partner (Sax and Brown, 2000; Barrett *et al.,* 2008). Second, it protects the population from intensive recombinations with genotypes non-adapted to the sink environment (i.e., migration load; Travis *et al.,* 2005), and preserves the best allelic combinations adapted to some environments. However, the lack of genetic recombination is also known to lead to the accumulation of deleterious mutations (Muller, 1964), as well as a lower adaptability to a new environment through diminished genetic variance (Burt, 2000). Actually, species with a mixed reproductive mode during their life-cycle may efficiently combine the advantages of the two reproductive modes. A recent theoretical study has shown that the most invasive species are those with an asexuality rate close to 0.95 (Bazin *et al.,* 2014). Beyond sexual reproduction, the benefits of genetic recombination may also be a result of intra-genomic re-arrangement (Feschotte, 2008; Hua-Van *et al.,* 2011; Kaessmann *et al.,* 2009; Thon *et al.,* 2006) or non-homologous mitotic recombination (i.e., parasexuality) described for several fungal species (Chuma *et al.,* 2011; Huang, 2014, McGuire *et al*., 2004). All these genetic mechanisms may have a dramatic effect on the evolutionary trajectories of the species and should be more systematically investigated in introduced populations to understand their invasive success (Stuckenbrock and Dutheil, 2018). However, the detection of genetic recombinations can be challenging when they are rare and sparse in the genome, as expected in mainly asexual species. Although a few genetic markers, such as microsatellite loci, are efficient when describing

the overall clonal structure of populations (Bruford and Wayne, 1993, Steimel *et al.,* 2004; Raboin *et al.,* 2007), their limited number and scattered distribution in the genome make them unsuitable for the accurate detection of recombination signals. The analysis of single nucleotide polymorphism markers (SNPs) on the whole-genome provide a more efficient method to estimate the importance of genetic recombination along the genome, and to identify the main evolutionary mechanisms in invasive pathogenic fungi (Milgroom *et al.,* 2014, Gladieux *et al.,* 2018).

The Chestnut blight fungus, *Cryphonectria parasitica*, is a textbook example of the variation of reproductive modes among native and introduced areas (Milgroom *et al.,* 2008; Dutech *et al.,* 2010, 2012). Native to eastern Asia, *C. parasitica* was probably introduced to North America at the end of the 19th century most likely from Japan (Milgroom *et al.*, 1996), and it almost caused the extinction of the American chestnut (*Castanea dentata*). In Europe, the pathogen was first reported in 1938 in Italy (Biraghi, 1946), from where it expanded throughout most of southern Europe (Rigling & Prospero, 2017), with a recent colonization further north (Robin *et al*., 2017). Studies based on ten microsatellite loci found multiple introduction events of the pathogen in Europe: North America was the source for introductions into Italy and Switzerland (Dutech *et al.*, 2012, Prospero & Rigling 2012), while Asia was the source for introduction in south-western France (Dutech *et al.*, 2012). These south-western populations could derivate from strains mentioned in 1949 in the northern coast of Spain (Darpoux, 1949; Robin *et al.*, 2009). In contrast to North America, the European outbreak has been less destructive. This is most likely because European chestnuts (*Castanea sativa*) are less susceptible to *C. parasitica* than American ones, and due to the presence of *Cryphonectria* hypovirus 1 (CHV-1) that decreases the fungus fitness (Nuss 1992; Milgroom and Cortesi 2004).

Contrary to Asian and North American areas (Milgroom *et al.*, 1995; Liu *et al.,* 1996), a strong clonal structure was observed in most European populations (Milgroom *et al.,* 2008; Dutech *et al.,* 2010; Prospero and Rigling, 2012). Since two haploid strains of *C. parasitica* need the different idiomorphic alleles at the mating-type locus (*MAT1-1/MAT1-2*) to reproduce sexually (Mcguire *et al.,*

2001), this clonal structure may be explained by the detection of only one mating type in some eastern European populations (Milgroom *et al.,* 2008). In contrast, the frequent report of both mating types in the western populations challenges this hypothesis (Bragança *et al.,* 2007; Robin *et al.,* 2009; Dutech *et al.,* 2010). Using ten microsatellite loci, three genetic clusters were identified in western Europe, each including at least one multilocus genotype highly repeated in numerous sampled populations and defined as clonal lineages (Dutech *et al.,* 2008, 2010). Two of these clonal lineages were located in south-eastern France (RE019, RE092; Fig. 1), and are probably related to the migration of the Italian populations introduced from North America (cluster C1; Dutech *et al.,* 2010).The two other clusters located in south-western France are associated with the clonal lineages RE043 and RE053 for C2 and RE028 for C3 (Dutech *et al.*, 2010; Fig. 1), and were likely introduced from Asia. In addition, two other clonal lineages were clustered in C1, and related to North American genetic pool (RE079, RE103; Fig. 1). However, these lineages were sampled only in the south-central France, the only geographical area in southern France where Asian and North American genetic pools are in sympatry. Associated with this secondary contact, RE079 and RE103 might be the result of admixture between these two heterogeneous genetic pools. More recently, in northern France, six of these seven clonal lineages have been reported (RE028 has not been reported) with six additional emerging clonal lineages (H13, H53, H11, H39, H28 and H58; Robin *et al.,* 2017; Fig. 1).

If these previous results suggested that these clonal structures are due to asexual reproduction, other observations question the occurrence of sexual reproduction in French populations. First, the presence of sexual structures has been reported in all sampled locations (Robin *et al*, 2009), as the two mating-types, which are commonly found in equal proportions in the sampled chestnut stands and often identified within clonal lineages (Dutech *et al.,* 2010). Second, several rare multilocus genotypes, which sometimes differ at a few microsatellite loci from the clonal lineages, have been observed in 32% of the 994 French isolates genotyped previously (Dutech *et al.*, 2010, Robin *et al.*, 2017). Without ruling out somatic mutations, these rare genotypes could originate from crossings between the clonal lineages (Dutech *et al.,* 2008,

2010). In the same way, these crosses could explain the emergence of new clonal lineages during expansion of *C. parasitica* to the north (RE079 and RE103 in south-central France, the six new clonal lineages in the northern France). These observations suggest that genetic recombination possibly associated with sexual reproduction may be more frequent than estimated from the genetic structure obtained from the microsatellite analysis.

In order to better describe the genetic relatedness between nine of the most frequent French clonal lineages, especially between the earliest southern and the latest northern ones, and try to identify the main mode of evolution in this western European *C. parasitica* population, we conducted a genotyping by sequencing analysis. We hypothesized that genetic recombination has a more important role in the evolution of western European populations than assumed by the clonal structure described in the previous population genetic studies. Using the whole genome sequences of 46 *C. parasitica* isolates (32 isolates from France among the seven southern and two northern clonal lineages and some close genotypes, two isolates sampled at the beginning of the European colonization from the Pyrenees and Italy, ten additional North American and two Asian isolates), we addressed the following questions: 1) On the whole genome, what are the nucleotide diversity and the genetic relationships between isolates belonging to the most frequent multilocus genotypes which were initially defined by analyzing ten microsatellite loci (Dutech *et al.*, 2010; Robin *et al.*, 2017); can these repeated multilocus genotypes be considered as “clonal lineages”? 2) Are there evidences of recent genetic recombinations within the seven clonal lineages studied here, and what are the length, frequency and genetic origin of the detected genetic recombination events? 3) What is the genetic origin of the new clonal lineages in south-central (RE079 and RE103) and northern France (H13 and H53) and how are the Asian and North American genetic pools involved in these emergences? 4) Can we estimate the size of founding population introduced from North America and the timing of emergence of the clonal lineages?

II Materials and Methods

Sequenced isolates

We chose several isolates belonging to and close to the French clonal lineages already genotyped (using ten microsatellite loci) and analyzed in Dutech *et al.* (2010) and Robin *et al.* (2017). Forty-nine isolates were sequenced, including 36 sampled in northern and southern France, one in Italy, and 12 from the two main origins of European introductions (ten from North America and two from Asia; Fig. 1; Table S1 for details). Three French isolates with more than 30% of missing data (SNPs) were finally removed from the analysis. Among the French isolates, 24 have the same multilocus genotype as one of the seven clonal lineages in southern France, five others are different from one to two microsatellite alleles from these clonal lineages (Fig. 1; Table S1 for details). We also sequenced three isolates from two new emerging clonal lineages in northern France (H13 and H53; Robin *et al.,* 2017), and two historical isolates sampled in Italy in 1968 (VG1896, J. Grente unpublished) and in the French Pyrenees in 1975 (VG2106, J. Grente unpublished).

DNA extraction, genome sequencing and assembly

For each isolate, a monospore isolation was performed (Text S1 for details). DNA was extracted from mycelium grown on PDA overlaid with cellophane following Hoegger *et al.* (2000). Eighteen isolates were sequenced using the PGM/Proton Ion Torrent sequencer (Thermo Fisher scientific Inc.) at the platform Genome-transcriptome of Bordeaux (Inra-Université Bordeaux, Bordeaux, France). The 32 other isolates, including one isolate (ABI005) previously sequenced with the PGM technology, were sequenced using Illumina HiSeq2000 technology in paired-end at the Genome and transcriptome GenoToul facilities (INRA, Toulouse, France).

Reads obtained from both sequencing technologies were first mapped on the reference genome EP155 (available on http://genomeportal.jgi.doe.gov/Crypa2/Crypa2. home.html) using glint

software (Faraud and Courcelle, unpublished; lipm-bioinfo.toulouse.inra.fr/download/glint/). Several regions of the EP155 reference presented no mapped read from these 49 sequenced isolates, questioning the validity of this reference for SNP calling (J. Gouzy, unpublished results). A new reference genome was produced. High-molecular-weight genomic DNA of the French YVO003 monospore (i.e., H53) was extracted following Cheeseman *et al.* (2014) and sequenced using the Pacific Bioscience (PacBio, San Francisco, California, USA) long run sequencing technology (Institute for Genomic Medicine, UCSD, CA, USA). Genome was assembled using PbcR wgs-8.2 with the genome size specified to 45Mb, and polished using pbalign (available on https://github.com/PacificBiosciences/pbalign), Quiver and SMRT analysis v2.3.0 (https://github.com/PacificBiosciences/DevNet/wiki/SMRT-View) with default parameters. Contigs were ordered using Mauve v2.4.0 (Darling *et al.,* 2004), guided by the EP155 reference genome.

Annotation of protein-coding genes and transposable elements

Gene models were predicted with a fully automated and parallelized pipeline, egn-ep (http://eugene.toulouse.inra.fr/Downloads/egnep-Linux-x86_64.1.4.tar.gz, release 1.2, Text S2 for details). For each of the 46 isolates, a draft genome was constructed using CLCBio with default parameters in order to blast the two mating types alleles sequences on it (*MAT1-1* and *MAT1-2,* first described by McGuire *et al.,* In 2001). Transposable elements (TEs) detection was performed, using the REPET software (Jamilloux *et al.,* 2017). REPET uses two pipelines: first, TEdenovo (Flutre *et al.,* 2011), which build a TE consensus library, and second, TEannot (Quesneville *et al.,* 2005) which annotates TEs in the genome using the classification system proposed by Wicker *et al.* (2007).

Mapping and SNP calling

After mapping the 46 sequenced genomes on the new masked reference genome using glint software, SNP calling was performed using varscan v2.3.7

(Koboldt *et al.,* 2009, 2012) with parameters set as follows: minimum coverage 15 reads, minimum length of the high scoring pair (hsp) ≥90, number of mismatches ≤5, no gap allowed, and only best-scoring hits taken into account. Alternative variants were kept if they were present on a minimum of ten reads and at least once on each of the two strands. Sites with heterozygous SNPs (i.e., alternative variant frequency estimated between 0.25 and 0.75 for one isolate; *C. parasitica* monospore isolates sequenced here are haploid), insertion-deletion sites and non bi-allelic SNPs were removed from the dataset. Polymorphic sites with more than 10% of missing data and scaffolds sized to less than 100kb were removed from this analysis.

Levels of polymorphism, genealogical relationships, clustering and linkage disequilibrium between isolates of *C. parasitica in France*

Based on the filtered SNP dataset, VCFTOOLS (Danecek *et al.,* 2011) were used to estimate nucleotide diversity (π) and SNP density, on a non-overlaping sliding window counting 10,000 nucleotides along the genome. Nucleotide diversity was plotted using R-cran package ‘qqplot’ (Turner 2014). Pairwise linkage disequilibrium (LD) between SNPs was estimated using VCFTOOLS LD statistics after keeping one single isolate of each clonal lineage with the best coverage. Tajima’s D was tested on SNPs alignments for different sub-populations and seven clonal lineages using DnaSP v5 (Librado & Rozas, 2009).
We estimated the genealogical relationships between sequenced isolates using the neighbor-net algorithm implemented in SplitsTree4 (Huson and Bryant 2006) with the uncorrected P distance (i.e., in our case the proportion of sites at which two sequences differ compared to all the polymorphic sites of the filtered dataset). This phylogenetic network method is used to identify recombination events among haplotypes, incomplete lineages sorting, or homoplasic mutations (Bryant and Moulton 2003).
Population structure was estimated using BAPS Version 6 (Corander *et al.,* 2008), a Bayesian method for identifying the most likely number of sub-populations (k number of clusters) in a given population by comparing the allele frequencies among all the polymorphic sites among sub-populations and

the whole population. To increase the resolution and to get the substructure, which is hard to detect because of strongly divergent lineages, we used BAPS through a hierarchical clustering process (Cheng *et al.,* 2013). The largest cluster obtained from the first analysis of the whole dataset was used as input for the second analysis. We obtained the best K value describing these two datasets. From this, we used the fixed k option in BAPS with default parameters to perform four independent runs (K = 2,3,4,5) on the whole dataset and five runs (K = 2,3,4,5,6) on the largest cluster.

Genealogies used for the recombination detection and BEAST analysis were estimated using RAxML (Stamatakis 2014) with the New rapid hill-climbing algorithm [-f d]; the Generalized Time-Reversible (GTR) substitution model with gamma site heterogeneity model; and other parameters to the default values. Using jmodeltest2, the GTR substitution model was chosen as the most suitable among others for our data (Darriba *et al.,* 2012).

Detection of recombination between and within clonal lineages

We used the Pairwise Homoplasy Index (PHI) test implemented in SplitTree4 to detect the presence of recombination within six clonal lineages in which we had at least two isolates sequenced. Using three different methods, we identified recombination events within five clonal lineages (RE092 excluded because isolates are too genetically divergent; see results). First, we used a home-made method. This fast and well adapted approach to detect recombinations within clonal lineages uses the combined strategies of distance methods and phylogenetic methods (Posada and Crandall, 2001). We measured the nucleotide diversity within each French clonal lineage as described above. As the recombination between two divergent sequences tends to significantly increase nucleotide diversity in the recipient strain compared to the lineage genetic background, we associated the peaks of genetic diversity within the clonal lineages with a putative recombination event. Considering the low average variation within clonal lineages (see Results), a recombining region was assumed when at least two successive 10kb windows with more than one SNP were identified. Regions separated by less than 100kb were grouped as

one unique recombining region. Recipient strain and origin of each recombining region were determined using a genealogical tree inferred with RAxML. In addition to this home-made method, we used two other methods to confirm the detected recombinations: fastGEAR (Mostowy *et al.,* 2017), using a Bayesian clustering method to detect dissimilar regions within each clonal lineage, and ClonalFrameML (Didelot and Wilson 2015), using a phylogenetic approach to detect genomic fragments introducing novel polymorphism within each lineage. We used the relative rate of recombination to mutation (R/θ), the average length of DNA imports (δ), and the mean divergence between donor and recipient genotypes (ν) estimated on the four south-eastern clonal lineages by ClonalFrameML, to calculate the relative effect of recombination to mutation (r/m*).*

Identification of the different haplotypes among the French clonal lineages

We designed a pipeline combining BCFtools, VCFtools and R software (R Development Core Team, 2008) to identify and compare the haplotypic sequences along the genome of the seven southern French clonal lineages and the northern emergent genotype H13. For each of these lineages, we chose one reference isolate for which no recent recombination had been detected with the three methods described above. For the RE092 lineage, the most diverse lineage of this study (see results), we chose three reference isolates (VG_1896, STC36 and YVO006). Using VCFtools and per 10kb window along the genome alignment, we extracted a set of variants between each pair of ten reference isolates and all singeltons were removed. The historical isolate VG_1896, the oldest of this study (sampled in 1968), was used as the reference haplotype. For each 10kb window, a new haplotype was defined when more than one SNP differed from a previously identified haplotype.

Molecular parameter estimation of the south-eastern clonal lineages

Assuming that genomic regions with identical haplotypes shared between the four south-western and south-central clonal lineages (i.e., 22 isolates) have been inherited from a recent common ancestor, we estimated the divergence times and evolutionary rates of these genomic regions using a tips dated approach with BEAST 1.7 (Drummond and Rambaut 2007; Drummond *et al.,* 2012). From these regions, we removed the remaining recombination regions detected with ClonalFrameML to obtain a core alignment of 11Mb. For efficient posterior estimation of parameters, we chose to minimize the possibility of recombination by subdividing the 11Mb alignment according to the putative chromosomes obtained from the two reference genomes EP155 and YVO003. On the basis of the Mauve alignment (Fig. S1), we chose the four largest putative chromosomes lacking rearrangements between these two reference strains: S1 in EP155 (i.e., MS1-1, MS1-2 and MS1-3 in YVO003), S3 (i.e., MS3-1 and MS3-2 in YVO003), S4 (i.e., MS4-1, MS4-2, MS4-3 and MS4-4 in YVO003) and S8 (i.e., MS8-1 and MS8-2 in YVO003). We constructed four genealogical trees using RAxML and rooted the bestTree output with the option -f I. Rooted trees were used as the starting tree in BEAST for each alignment. GTR (Generalized Time-Reversible; Lanave *et al.,* 1984) substitution model and an empirical base frequencies model, as well as a strict molecular clock model (Zuckerkandl and Pauling 1965) were set up. As a *C. parasitica* colony can survive on its host for several years, tip dates for the 22 isolates were specified between the sampling date and twenty years ago. As no data are available in Europe on the demography of *C. parasitica*, the coalescent model was set to follow either an exponential population growth or a constant population size. The Markov Chain Monte Carlo (MCMC) length was 100,000,000 iterations, thinned every 10,000 to retain 10,000 final sampled trees. BEAST log files were checked using Tracer (Drummond *et al.,* 2012) to control the posterior distributions and the Effective Sample Size (ESS) was used to check the independence of parameters estimation through the chain. The highest posterior density (HPD) 95% given for the parameters is the shortest interval that contains 95% of the posterior probability. From these runs, we chose to discard the first 1,000 sampled trees as burn-in to generate a maximum clade

credibility tree with median node heights in treeannotator v. 1.7.5 (Drummon *et al.,* 2012).

III Results

Assembly of the new reference genome, gene and transposable element content

The PacBio sequencing of the strain YVO003 yielded 494,384 reads, N50=12,757 bp (L50=132,812 reads) with an average length of 9,064bp. The assembly produced 35 scaffolds for a total length of 39.3Mb (N50=2.7Mb; L50=6). This new reference genome was shorter than the EP155 reference genome v2 for 4.6Mb with several large genomic regions missing, such as on scaffold 2 (1.5Mb) or scaffold 6 (0.5Mb) and on putative scaffold rearrangements (Fig. S1). The number of predicted genes in this new *C. parasitica* reference genome is 12,146 (EP155 genome annotation v2: 11,609) comprising 52.4% of the assembly length with an average gene density of one gene per 3.2kb (EP155: one gene per 3.8kb). 276 complete plus 11 fragmented gene models out of a total of 290 (95.2% and 3.8% respectively) were detected. Using the two *MAT* sequences, we located the *MAT* locus on the RC05 scaffold between 52,138bp and 51,074bp. Except for the clonal lineage RE079 for which all sequenced isolates were *MAT1-2* (including YVO003), all lineages included the two *MAT* alleles (Table S1 for details).

The REPET package predicted 968 copies of transposable elements (TEs) from 28 TE families, covering 2.3% of the genome with an average density of one TE per 44.8 kb. The predicted TE copies were classified in 14 DNA transposons, 12 RNA transposons and 2 undefined families (Table S2 for details), the 28 families containing from two to 210 copies with an average copies number of 31.4 ($ci_{95\%}$ = 16.2 – 46.6). The 11 putatively active families represented 271 of the total number of copies (878) and 0.7% of the total genome size. The 17 other families showed incomplete consensus DNA sequence and may be the signature of ancient TE burst.

Mapping and SNP calling

All the 46 genome sequences from both sequencing technologies were mapped on the new PacBio reference genome with a coverage of 41.0 to 187.8X (mean = 100.5; Table S1 for details). We identified 118,182 SNPs using the Varscan method. Only three SNPs were detected between the Illumina and the Ion Torrent sequences of the ABI005 isolate after filtering, without considering missing data. This comparison suggests that the two sequencing technologies produced a similar set of SNPs after filtering. Finally, we kept a dataset of 46 isolates with 38,592 SNPs (i.e., average 1SNPs per kb), and focused on the 26 scaffolds longer than 100kb. Most of the discarded SNPs were present within TEs and incompletely covered regions, while only 44.0% of SNPs were discarded in the genes (some genes are in fact TEs) compared to 67.3% on the whole genome.

Levels of polymorphism, genealogical relationships, clustering and linkage disequilibrium between isolates of C. parasitica in France

Nearly half of this polymorphism was detected among the Japanese and Chinese isolates (18,263 SNPs, π=3.02E^{-4}; Table 1) while the ten American isolates showed 13,387 SNPs (π=1.3E^{-4}) with only 2,913 SNPs shared between the two geographical areas. Among the 34 European isolates, 26,090 SNPs were identified (π=1.8E^{-4}) mainly observed among the ten isolates introduced from Asia (18,201 SNPs, π=1.75E^{-4}). Only 12,807 SNPs (π=1.07E^{-4}) were observed among the 24 isolates introduced from North America and 4,918 SNPs were shared between the two introductions.

The neighbor-net network and the BAPS clustering generated from the 46 *C. parasitica* sequences of SNPs both showed two main genetic clusters containing respectively south-western and some northern French isolates related to Asian ones and south-eastern and other northern French isolates related to North American ones (Fig. 2.a and 2.c). The reticulations of the genetic network showed a greater proportion of shared markers within the

North American cluster than within the Asian one, suggesting a higher level of mating or a more recent divergence within the North American populations analyzed. The BAPS analysis indicated that the two south-eastern clonal lineages (RE019 and RE092) were genetically divergent relative to a second French cluster consisting of the two south-central lineages (RE103 and RE079), and to a third one associated with the northern clonal lineage H13 (Fig. 2.b).

Inside these clusters, French isolates were closely grouped following their clonal lineage relatedness, previously defined using the ten microsatellite locus analysis (Dutech *et al.,* 2010). The genetic variability within five clonal lineages (RE019, RE043, RE053, RE079 and RE103) was low with a P-distance comprising between 0.001 ($ci_{95\%}$ = 0-0.002; RE079) and 0.014 ($ci_{95\%}$ = 0.014-0.015; RE103) relative to the average P-distance of 0.237 ($ci_{95\%}$ = 0.224-0.250) estimated between two clonal lineages. Reticulations were sometimes detected within these lineages, but only due to few isolates (Fig. 2.c). The south-eastern RE092 clonal lineage was more variable than expected on the basis of microsatellite analysis, with a mean uncorrected P-distance of 0.064 ($ci_{95\%}$ = 0.056-0.073) and more reticulations than other French lineages. It can therefore not be strictly considered as a clonal lineage. However, all the RE092 isolates and close genotypes (RE093 and H68) remained clustered in the same clade in the network and in the same cluster in the BAPS best partitioning (Second analysis, k=6; Fig. 2.b). One of the least divergent pairs of isolates was surprisingly the historical isolate sampled in the Pyrenees in 1975 (VG2106) and a RE043 isolate (BAR002) sampled in 2006 in northern France (both introduced from Asia) with only 25 SNPs detected. Tajima's D estimates on the hole genome were non significant (P> 0.05) for nearly all the clonal lineages and sub-populations considered (Table 1), except for RE019 with D=-1.766 (P<0.001) indicating an excess of rare alleles.

A high linkage disequilibrium (LD) was estimated between pairs of SNPs separated by 1kb in the NA (North American isolates) and FNA (French isolates from North American genetic pool) subsets (r^2=0.84 $ci_{95\%}$=$9E^{-3}$ and 0.91 $ci_{95\%}$=$5E^{-3}$ respectively; Fig. 3). At this range of genomic distances, the estimate was only 0.54 ($ci_{95\%}$=$9e^{-3}$) for the Global subset. The distances at which the LD was half decayed - a useful indicator of the importance of recombination in a

population (Nieuwenhuis and James 2016) - were approximately ~3kb, ~28kb and ~400kb for the three datasets, Global, NA and FNA.

Detection of recombination between and within clonal lineages

The PHI test did not show significant evidence for recombination in the six French clonal lineages, except for RE092 (*p-value* = 0.0). However, the analysis of polymorphism within the five other lineages showed numerous peaks of diversity in the genome (Fig. S2), allowing us to identify several recombination events, confirmed by the two Bayesian methods FastGEAR and ClonalFrameML (Table 2). The diversity method detected 44 putative recombining regions observed on 21 of the 26 analyzed scaffolds, from which 31 were greater than or equal to 20kb. The average size of these 46 regions was 158kb ($ci_{95\%}$ = 96-220kb) and the largest region was 1,2Mb within the RE103 clonal lineage. Two regions on scaffolds MS8-2 (10kb) and C25 (30kb) each showed signals of recombination within two different clonal lineages: RE053 and RE103, and RE103 and RE079 respectively. One region on the RC05 scaffold showed a signal of recombination within four clonal lineages, which were estimated to be between 40 and 90 kb and to all encompass the mating-type locus. The FastGEAR method detected fewer recombination events (33), but confirmed 30 of the 31 largest regions detected with the first method; meanwhile, ClonalFrameML detected 42 recombinations (27 of the 31 large recombining regions detected with the first method) including seven smaller than 500bp and 28 large regions greater than 20kb (Table 2).

Among the 31 recombination events detected with the diversity method, we identified at least one - and often two - isolates within each clonal lineage having one or more signals of recombination (Fig. 4). Most of these signals of recombination (17/31) involved an exchange between lineages introduced from the same area of origin. Ten recombinations involved isolates from different introductions, mostly (8/10) detected in one isolate from the central RE103 lineage, which had received fragments from the south-western RE053 lineage. Four recombining fragments did not originate from a sequenced French isolate, but were related to the same genetic pool (i.e., Asian or NA) as the recipient

strain. Overall, regarding the four clonal lineages of the North American pool (excluding H13), we calculated the relative effect of recombination to mutation as r/m = 9.5. This means that recombination brought 9.5 times more substitutions than mutation within these lineages.

Identification of the different haplotypes among the French clonal lineages

Using our method of haplotype identification to estimate the proportion of 10kb sequences shared between the clonal lineages (Fig. S3), we estimated that the RE019 lineage and the closest RE092 isolate were genetically close with 79.2% of their genome sharing the same haplotype. The RE079 and RE103 lineages were highly similar with 87.2% of identical haplotypes, but 21.6% of their genome was found to be different from the other sequenced lineages. The H13 lineage included 17.9% of sequences that differed from the other lineages. However, these unknown haplotypes were phylogenetically related to the North American population (results not shown). RE019, RE092, RE079, RE103 and H13 were strongly associated with the North American introduction.

The genomic comparison of the five clonal lineages associated with the North American introduction showed that 99.5% of their genome may be reconstructed with only three different haplotypes (Fig. S3). Among these five lineages, 41.5% of the genomes was not variable. However, half of the identical regions (i.e., 20% of the whole genome) was variable in the North American isolates analyzed, revealing a lower level of genetic variability in the French population. In contrast, the three lineages related to the south-western introduction RE043, RE053 and RE028 were more diverse, with only 29.5% of their genome being identical and 43.5% having two haplotypes.

Molecular parameter estimation of the south-eastern clonal lineages

BEAST was used to reconstruct the history of the four French south-eastern and south-central clonal lineages from when they were introduced to Europe from North America. We chose four DNA fragments with no ancient

signal of recombination detected between lineages on scaffolds MS1 (2.4Mb, 29 SNPs), MS3 (1.4Mb, 155 SNPs), MS4 (1.Mb, 158 SNPs) and MS8 (0.7Mb, 182 SNPs). We did not include the H13 lineages in this analysis, because they diverged too much from the others analyzed in the different regions. The RAxML trees of the four fragments showed differing topologies (Fig. S4), and the molecular parameter estimations carried out on MS3, MS4 and MS8 did not converge. The assessment of alignments in these regions revealed several possible explanations for this non-convergence. First, the polymorphism is highly variable between lineages for the same scaffold, with no variation within some lineages and up to 37SNPs. Second, the most recently sampled isolates have nearly identical sequences to the ancestral state, whereas the isolates that were sampled the earliest showed several variations, thus challenging the estimation of the mutation rate. On the MS1 scaffold for which convergence was obtained, the posterior parameters which assumed a constant population size, or an exponential growth rate, gave values consistent with the population history of *C. parasitica* in western Europe. Under the hypothesis of constant population size (the growth rate estimated in the exponential model was close to zero; 0.05, 95% highest posterior density (HPD) = -0.02 – 0.13), the effective population size of the south-eastern introduced populations was estimated to be 44 individuals (95% HPD = 11 – 95), and the mutation rate per site and per year was $4.7E^{-8}$ (95% HPD = $1.5E^{-8}$ – $8.6E^{-8}$). The Date of the most recent common ancestor (MRCA) of this south-eastern population was 1931 (95% HPD: 1874 – 1966), which is concordant with the first European report of *C. parasitica* in 1938 in Italy (Biraghi 1946). The RE092 isolates were not clustered and it was not possible to infer a date of emergence for this lineage (Fig. 5). The emergence of the RE019 lineage was independent of the other lineages and dated to 1983 (95% HPD: 1959 – 2000). It was not possible to conclude which of the RE079 and RE103 lineages emerged first, due to the overlapping 95% HPD of the dates of emergence. However, it seems that they might have diverged as from around 1992 (95% HPD: 1980 – 2000) and separately from the RE092 lineage.

IV Discussion

Genealogical and clustering analyses based on the filtered dataset of 38,592 SNPs confirmed that isolates belonging to six multilocus genotypes (MLGs) highly repeated within French populations are highly similar on the whole genome (H13 and RE028 excluded of this analysis due to the sequencing of only one isolate). The clonal structure may be even stronger than described by the ten microsatellite loci (Dutech *et al.*, 2010; Robin *et al.*, 2017), since several MLGs differing by one or two microsatellite alleles were finally considered to belong to the same clonal lineage (for example the clonal lineage H53 identical to RE079). These observed differences in microsatellite alleles are likely to have resulted from recent mutations of the repeated microsatellite motif. Within these clonal lineages - excluding the recombinant regions identified (see below) - we detected a small number of variations (between 17 and 120 different SNPs within each lineage). For example, two isolates of the south-western RE043 clonal lineage, which was sampled between 1975 and 2012 (VG2106 and BAR002), differed by only 25 SNPs on the whole genome. The genome-wide linkage disequilibrium decay which was estimated for four lineages from North American genetic pool has a very similar profile to some clonal species. Nieuwenhuis and James (2016) described the most sexual species as having a LD decay lower than 1kb and highly clonal ones with half-decay points at greater than 100kb. Global and NA subsets have a LD decay similar to yeast species which suggest occasional sexual reproduction. By contrast, slow decay of LD of the FNA subset (French lineages and the historical isolate from North American genetic pool) suggested a mainly asexual or highly homogamic reproduction within each clonal lineage of this area. In contrast, for one lineage (RE092), we estimated high polymorphism (5,386 SNPs), a large number of reticulations in the neighbor-net network, and a PHI test indicating frequent genetic recombinations, challenging the assumption of clonal evolution. However, all RE092 isolates and some other MLGs (RE093 and H68) have been assigned to the same cluster by the BAPS structuring algorithm, suggesting that recombination events occur preferentially within this lineage. This regional genetic structure in south-eastern France is similar to the results

obtained in *Magnaporthe oryzae* by Gladieux *et al.,* (2018), who reported a central sexual lineage and several asexual lineages that have spread throughout the world. This could suggests different colonization strategies for these different lineages in invasive areas.

Although clonal evolution was confirmed at the genome scale for most of the French clonal lineages, the signature of recent recombination events between these lineages have been clearly identified using three different methods. Using ClonalFrameML, we estimated that the recombinations identified in the four south-eastern lineages introduced 9.5 times more substitutions than did mutation; thus indicating that even if they are limited in size, these events have important evolutionary consequences at the genome scale. In addition, since the detection of recombination events is based on the difference in genetic distance or phylogenetic branching between adjacent sequences, it is almost impossible to detect a recombination event between two sequences nearly identical. Thus, the estimates presented here are likely to be low, and the effect of recombination may be even greater than identified in this study. Most of these exchanges (87%) occurred between the seven clonal lineages studied here, while 13% involved haplotypes not identified in this study. The genetic mechanisms which cause this recent gene flow between these lineages remains unclear. McGuire *et al.,* (2004) previously reported such exchanges in *C.parasitica* isolates from the field in North America suggesting that they may be caused by non-meiotic crossing over between vegetatively incompatible genotypes (i.e., parasexual recombination; Pontecorvo, 1956, Milgroom *et al.*, 2009). Although we cannot rule out this possibility, other results suggest that regular sexual reproduction may be the main mechanism behind this limited genetic introgression within the lineages for two reasons. First, although a small number of isolates were analyzed in this study, we systematically detected one recombination event which encompassed the mating type locus within four clonal lineages and originated from one of the other French lineages. This genomic region would either be exchanged very regularly between these lineages by an undetermined genetic mechanism or would sometimes be exchanged and maintained in the lineages by regular sexual reproduction within the lineages during European colonization. Second, we expected

negative values of Tajima's D in asexually spread lineages (Gladieux *et al.*, 2017), and in the case of population expansion (Aris-Brosou & Excoffier, 1996) as for *C. parasitica* populations in Western Europe (Robin & Heiniger, 2001). But, Tajima's D estimates within each lineages were not significantly negative in all lineages but one (RE019). It is likely that the low polymorphism detected within the lineages and the small number of samples within these lineages decreases the power of the Tajima's D test (Ramos-Onsins & Rozas, 2002). More sequenced isolates are needed to understand how mutations are transmitted in these lineages. For now, the most parsimonious hypothesis is the regular occurrence of genetic recombination among isolates assigned to the same clonal lineage. These homogamic crossings would produce an inverse effect than asexuality and population expansion on the Tajima's D, and prevent us from obtaining a clear signal of rare alleles excess usually found in clonal lineages in other fungi species (Gladieux *et al.*, 2017). The occurrence of preferential crossings within lineages (also called intra-haploid mating, Giraud *et al.,* 2006) should be investigated in the future. It could be an important factor in the success of invasion of *C. parasitica* in Europe. This intra-haploid mating has the advantage of preserving the sequence of adapted haplotypes (with the exception of the mating types and some limited introgressions or mutations), as well as of limiting the accumulation of deleterious mutations, the transmission of virus via asexual reproduction (Day, 1977) and the invasion of transposable elements within genomes (Selker, 1990).

Genealogical, clustering and haplotypes analyses confirmed the two independent introductions of *C. parasitica* in France previously described in Dutech *et al.,* (2012). Our results showed that the Asian and North American genetic pools do not mix in the areas where they coexist (South-central and northern France). The two lineages RE079 and RE103 emerging in the south-central France are associated only with the North American gene pool. A similar result is found in the northern France with lineage H13. However, as a part of their genome was divergent from the other south-eastern clonal lineages, these emergences must have involved at least one other genotype not analyzed in this study. From our samples it is not possible to determine if these genotypes were recently introduced from North America, or if they have been present

since the early stages of the introduction in low frequencies in French *C. parasitica* populations. We first hypothesized that the new haplotypes came from the RE092 lineage, which is more diverse than others. However, a rarefaction curve carried out on the genetic diversity of this lineage suggests that most of its diversity has already been sampled in this study (Fig S5). We hypothesize that rare genotypes unrelated to the French clonal lineages, and identified in many southern populations (Dutech *et al.,* 2012), may sometimes cross with the clonal lineages to produce these new emerging lineages. The maintenance of rare genotypes is not consistent with several theoretical studies on expanding populations assumed to lead to rapid fixation of haplotypes along the colonization gradient (Excoffier 2004). However, another theoretical study has showed that the combination of the Allee effect (the decrease of effective reproduction due to the limitation of sexual partners) and unfavorable environmental factors in the early waves of the colonization may sometimes lead to a greater diversity in the colonization wave than in the colonization front (Roques *et al.,* 2012). Several factors in Europe, such as a temperature gradient (Robin *et al.,* 2017), or the presence of different lineages of the virus CHV-1 (Feau *et al.,* 2014), may be strong selective pressures for the fungus, thus explaining this pattern of genetic diversity in this invasive species.
Haplotype identification and estimates of molecular parameters using BEAST helped us to infer the evolutionary history of *C. parasitica* in Western Europe. We identified a maximum of three different haplotypes for reconstructing 99.5% of the whole-genome haplotype phase of the five lineages associated to the North American genetic pool. This result suggested that a very small founding population was introduced to Italy at the beginning of the 20th century. This result is also supported by an estimated effective population size of 44 individuals from the BEAST analysis. Surprisingly, we also estimated the same number of three haplotypes for reconstructing almost all the haplotype phase (98.7%) of the ten North American genomes analyzed in this study. These results suggest that following the first introduction of Asian strains in North America, a few invasive genotypes may have established and subsequently colonized new areas. Selection for the most invasive genotypes in the first steps of invasion, alongside the colonization of subsequent areas by their

progenies (known as the invasive Bridgehead effect; Lombaert *et al.,* 2010), may explain the invasive success of *C. parasitica,* which has such a limited genetic variation between the two invaded continents. We were not able to clearly reconstruct the timing of emergence of the clonal lineages originating from the North American introduction. One scaffold, on which no recombination events were detected between the clonal lineages, suggested dates of emergence between the different lineages that were consistent with the successive colonization of French regions. On other scaffolds, the lack of diversity associated with a recent divergence of the lineages may explain the lack of convergence of the model using BEAST. However, a surprising result from some genomic fragments was a lower number of mutations relative to the historical isolate for some recent isolates than for some of them sampled earlier. This result is inconsistent with an asexual evolution and an accumulation of mutations along the clonal branches. It again suggests the effect of sexual reproduction within these clonal lineages that may eliminate some new mutations, especially if they are slightly deleterious. This hypothesis should be investigated in the future.

This study shows that a clonal population structure does not necessarily imply that isolates reproduce only clonaly, as it has been described in other fungal species (Henk *et al.*, 2012; Milgroom *et al.*, 2014). First, *Cryphonectria parasitica* clonal lineages have regularly exchanged genomic regions. Second, a part of the genome of new emerging clonal lineages during the colonization are not related to the first emerging lineages. Associated with the presence of the two mating types within each clonal lineage, these results suggest that sexual reproduction may be more frequent than assumed from the simple description of the clonal structure in Europe. In this context, the stability of the lineages over time and through colonization in France, with a limited genetic introgression in their genome, raises the question about potential barriers to gene flow among the French clonal lineages. Pre- or post-zygotic barriers may be involved in these limited crossings between lineages. For example, when comparing the two assembled *C. parasitica* genomes (YVO003 and EP155), we found putative mis-assemblies or genomic rearrangements between them; the latter often being involved in reproductive isolation within species (Brown and

O’Neill 2010). We hypothesize that chromosomal rearrangements between clonal lineages may lead to partial reproductive isolation in French populations of C. parasitica. Without ruling out selective processes, a limited introgression after a genetic admixture may also be observed when there is asymmetric gene flow between parental sources (Verdu and Rosenberg 2011). Previous studies have shown the presence of spatial clonal patches of *C. parasitica* within European local chestnut stands (Dutech *et al.,* 2008, Hoegger *et al.,* 2000). If the hybrids resulting from crosses of two clonal lineages are mostly dispersed in one of these patches, conditions of unidirectional back-crosses to one of the parental genotypes can be created, leading to a limited introgression in few generations as observed in this study. Therefore, to disentangle selective or neutral factors explaining the genomic structure of this invasive species, we will now need to analyze allelic combinations from a larger number of different genotypes.

Acknowledgments

We thank Donald L. Nuss and the genome portal of the Department of Energy (DOE) Joint Genome Institute (JGI) for making the EP155 genome available; Pierre Gladieux, Christophe Lemaire, Aurélien Tellier, Simone Prospero and Jean-Paul Soularue for their numerous comments and corrections on previous versions of the manuscript; Sophie Siguenza for Mauve alignment of genomes; and Erika Sallet for gene annotations. Ion-torrent sequencing was performed at the Genome Transcriptome Facility of Bordeaux and Illumina at the GeT-PlaGe Facility of Toulouse. This work was supported by the ANR-12-ADAP-0009 (Gandalf project) and an innovative project INRA department EFPA. A. Demené was supported by a PhD fellowship from the Ministère français de l’enseignement supérieur – Université de Bordeaux.

Data Accessibility Statement

All DNA sequences are available on NCBI with the accession number SRP162210.

Genome, gene annotation and transposable elements: doi:10.25794/reference/UQ0T2ENU

Author contributions

A. D. carried out the data analyses and interpretation and wrote the manuscript. C. D. conceived and supervised the study and wrote the manuscript. L. L. and J. G. worked on the raw data. R. D. and O. F. performed the high molecular height DNA extraction of YVO003. O. F. did the molecular laboratory work. G. S.-J. carried out the monospore isolations.

V Tables and Figures

Tables

Table 1. Summary statistics of genomic variations through sliding and non-overlapping 100kb windows for different subset of isolates. † sample size; ‡ number of bi-allelic sites; § measure of nucleotide diversity; ¶ North America. In bold are Tajima's D values significantly different from zero. Nr define sets of isolate without recombination.

Lineage	n†	SNP‡	TajimaD	Pi§	Singleton Number
Asia	2	18263	-	3.02E-04	11204
Asian Introduction	10	18201	0.923	1.75E-04	3161
RE053	4	583	-0.870	4.24E-05	39
RE053 Nr[a]	2	23	-	6.67E-06	23
RE043	4	66	0.621	1.26E-05	30
RE043 Nr[a]	3	51	-	1.21E-05	17
NA¶	10	13387	0.958	1.30E-04	564
NA¶ Introduction	24	12807	1.192	1.07E-04	393
RE092	6	5386	0.280	8.71E-05	52
RE019	7	405	**-1.766**	1.36E-05	58
RE019 Nr[a]	5	120	-	8.80E-06	40
RE079	5	135	-1.268	8.33E-06	55
RE079 Nr[a]	4	51	-	4.17E-06	50
RE103	4	1135	-0.837	4.24E-05	116
RE103 Nr[a]	2	17	-	7.08E-06	17

Table 2. Recombination statistics of events detected within five clonal lineages with the three methods used: Nucleotide diversity: method using nucleotide diversity within clonal lineages and maximum likelihood genealogies, FastGEAR (Mostowy *et al.*, 2017) and ClonalFrameML (Didelot and Wilson, 2015). Largest recombinations are larger than 20kb. † kilo base pairs; ‡ 95% confidence interval.

Method	Nucleotide diversity	FastGEAR	ClonalFrameML
Recombination count	44	33	42
Largest recombinations count	31	30	28
Size of largest recombinations (kb†)	1190	5550	1190
Mean size of largest recombinations (kb†) ($CI_{95\%}$‡)	230 (76)	753 (483)	220 (93)
Mean size of smallest region (kb†)	10	3	0.011

Figure captions

Figure 1. Distribution of the multilocus genotypes (MLGs) of *Cryphonectria parasitica* isolates sampled in 6 French subpopulations (following Robin *et al.* 2017 classification : South-western France = S1 + S2, South-central France = S3, South-eastern France = S4, North-western France = N1, North-central France = N2 + N3, North-eastern France = N4). Colors represent the most frequent MLGs and white represents other MLGs. Pie charts were constructed using data from 583 isolates genotyped (Southern France, Dutech *et al.*, 2010) and 411 isolates genotyped (Northern France, Robin *et al.*, 2017) for ten microsatellite loci. Stars shows the affiliation to sub-populations and MLGs of each of the 32 French isolates sequenced and analyzed in this study. 1938 is the first official report of *C. parasitica* in Italy (Darpoux, 1949), 1949 is the first mention of *C. parasitica* on *Castanea crenata* trees in the northern coast of Spain (Darpoux, 1949). *C. parasitica* was reported in all but one sample sites in the survey carried out in 1996 and 1997 (de Villebonne 1998, Robin *et al.*, 2017)

Figure 2. Population subdivision defined by BAPS from the two hierarchical analysis and distance-based tree (Neighbor-joining) from SplitsTree4. Circles represent the partition obtained with different k values in the two BAPS analysis and color of each circle describes the cluster membership of each isolate. K fixed values are shown on the circles. Color patches define the clonal lineages: RE028 in yellow, RE043 in purple, RE053 in green, RE092 in grey, RE019 in blue, RE079 in orange, RE103 in red and H13 in turquoise. Names in dark blue are the ten North American isolates and the historical isolate introduced from North America is in light blue. Names in dark red are the two Asian isolates and the historical isolate introduced from Asia is in light purple.
a) BAPS analysis with all isolates (First analysis) and b) BAPS analysis with only the North American origin isolates (Second analysis).
c) Neighbor-net network of the 46 isolates of *C. parasitica*, based on 38,592 SNPs estimated from SplitsTree4 with uncorrected P distance. Potential

recombinations are shown by reticulations. Except for the internal reticulations, nodes showed bootstrap values greater than 0.95 (data not shown for clarity).

Figure 3. Log-linear plot of linkage disequilibrium (r^2) according to the distance between nucleotides from 1 to 500kb in three groups of *C. parasitica* isolates: global subset (two Asian isolates, ten North American, five French including the historical Italian isolates from the North American genetic pool) in green, NA subset (ten North American isolates) in purple and FNA (five French including the historical Italian isolates from the North American genetic pool) in blue. Change of the r^2 along the genome is represented using an estimated smoothed curve. Arrows indicate the half-decay values of LD of its maximum estimated value.

Figure 4. Genomic distribution of the 31 major recombination regions (>20kb) detected using the nucleotide diversity method of detection of recombinations within the five clonal lineages studied here. Colors of the regions of introgression correspond to the donor lineage of the fragment. Unknown origins are in white.

Figure 5. Maximum clade credibility trees of the 2.4Mb alignment located on the scaffold S1 of *C. parasitica*. Each internal node is labeled with the posterior probability of the robustness of the corresponding clade. The blue bars illustrate the extent of the 95% highest posterior density intervals for the node age. The x-axis scale is in years.

Figure 1.

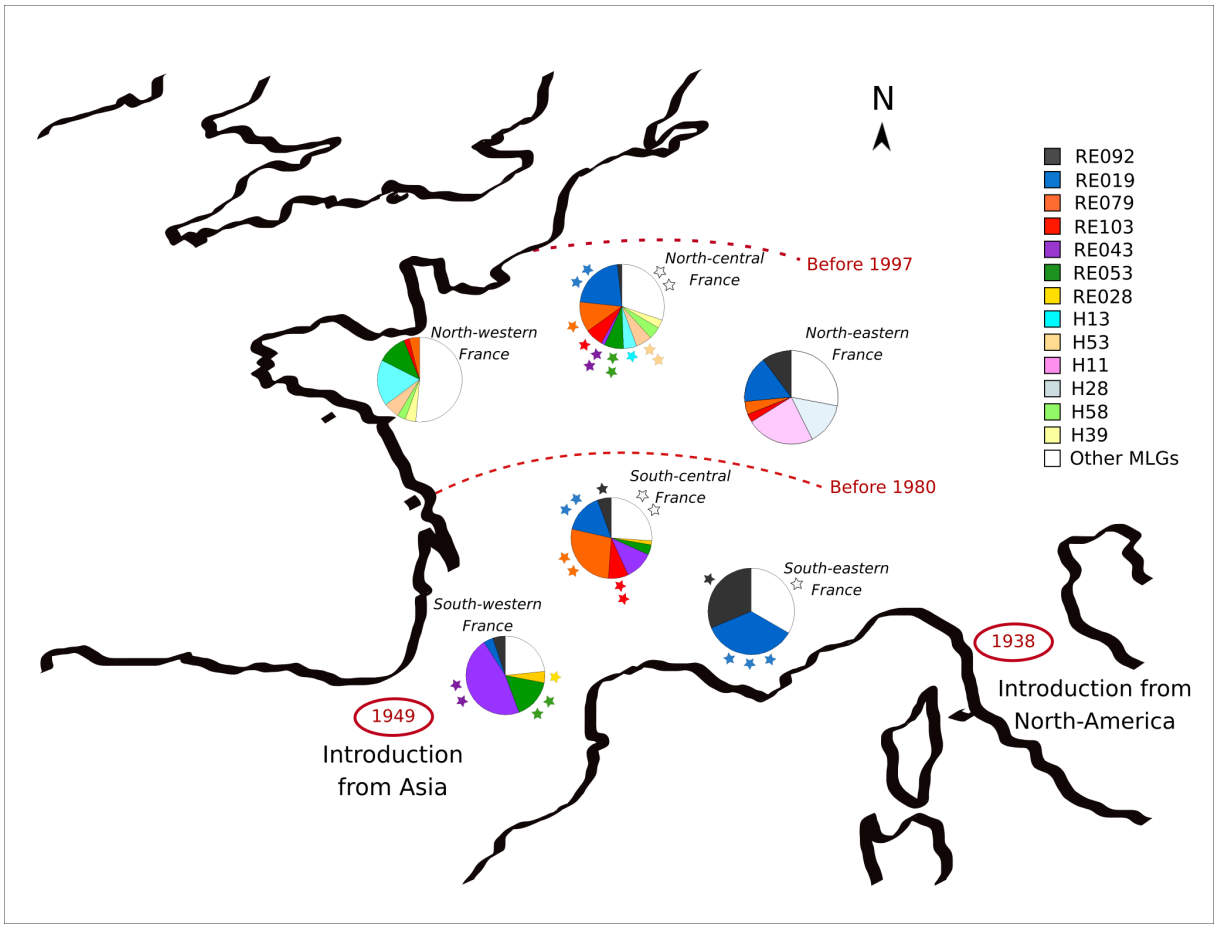

N
RE092
RE019
RE079
RE103
RE043
RE053
RE028
H13
H53
H11
H28
H58
H39
Other MLGs
North-central France
Before 1997
North-western France
North-eastern France
South-central France
Before 1980
South-eastern France
South-western France
1938
Introduction from North-America
1949
Introduction from Asia

Figure 2.

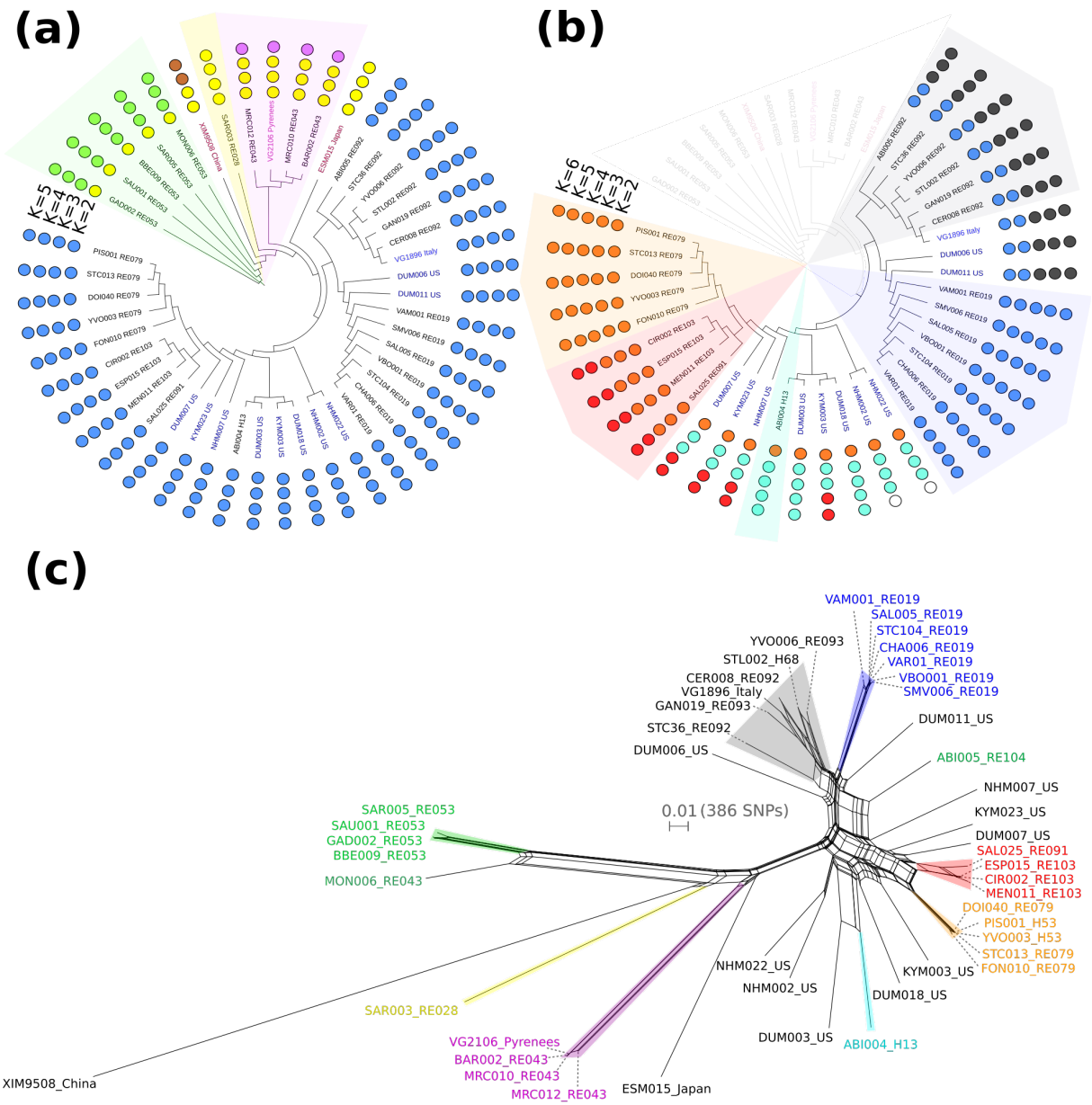

(a)
K=5
K=4
K=3
K=2
XIM9508 China
SAR003 RE028
MRC012 RE043
VG2106 Pyrenees
MRC010 RE043
BAR002 RE043
ESM015 Japan
ABI005 RE092
STC36 RE092
YVO006 RE092
STL002 RE092
GAN019 RE092
CER008 RE092
VG1896 Italy
DUM006 US
DUM011 US
VAM001 RE019
SMV006 RE019
SAL005 RE019
VBO001 RE019
STC104 RE019
CHA006 RE019
VAR01 RE019
NHM022 US
DUM018 US
KYM003 US
DUM009 US
ABI004 H13
NHM007 US
DUM007 US
KYM023 US
SAL025 RE091
MEN011 RE103
ESP015 RE103
CIR002 RE103
FON010 RE079
YVO003 RE079
DOI040 RE079
STC013 RE079
PIS001 RE079
MON006 RE053
SAR005 RE053
BBE009 RE053
SAU001 RE053
GAD002 RE053
(b)
K=6
K=5
K=4
K=3
K=2
(c)
VAM001_RE019
SAL005_RE019
STC104_RE019
CHA006_RE019
VAR01_RE019
VBO001_RE019
SMV006_RE019
YVO006_RE093
STL002_H68
CER008_RE092
VG1896_Italy
GAN019_RE093
STC36_RE092
DUM006_US
DUM011_US
ABI005_RE104
NHM007_US
KYM023_US
DUM007_US
SAL025_RE091
ESP015_RE103
CIR002_RE103
MEN011_RE103
DOI040_RE079
PIS001_H53
YVO003_H53
STC013_RE079
FON010_RE079
KYM003_US
DUM018_US
ABI004_H13
NHM022_US
NHM002_US
DUM003_US
ESM015_Japan
MRC012_RE043
MRC010_RE043
BAR002_RE043
VG2106_Pyrenees
SAR003_RE028
XIM9508_China
MON006_RE043
BBE009_RE053
GAD002_RE053
SAU001_RE053
SAR005_RE053
0.01 (386 SNPs)

Figure 3.

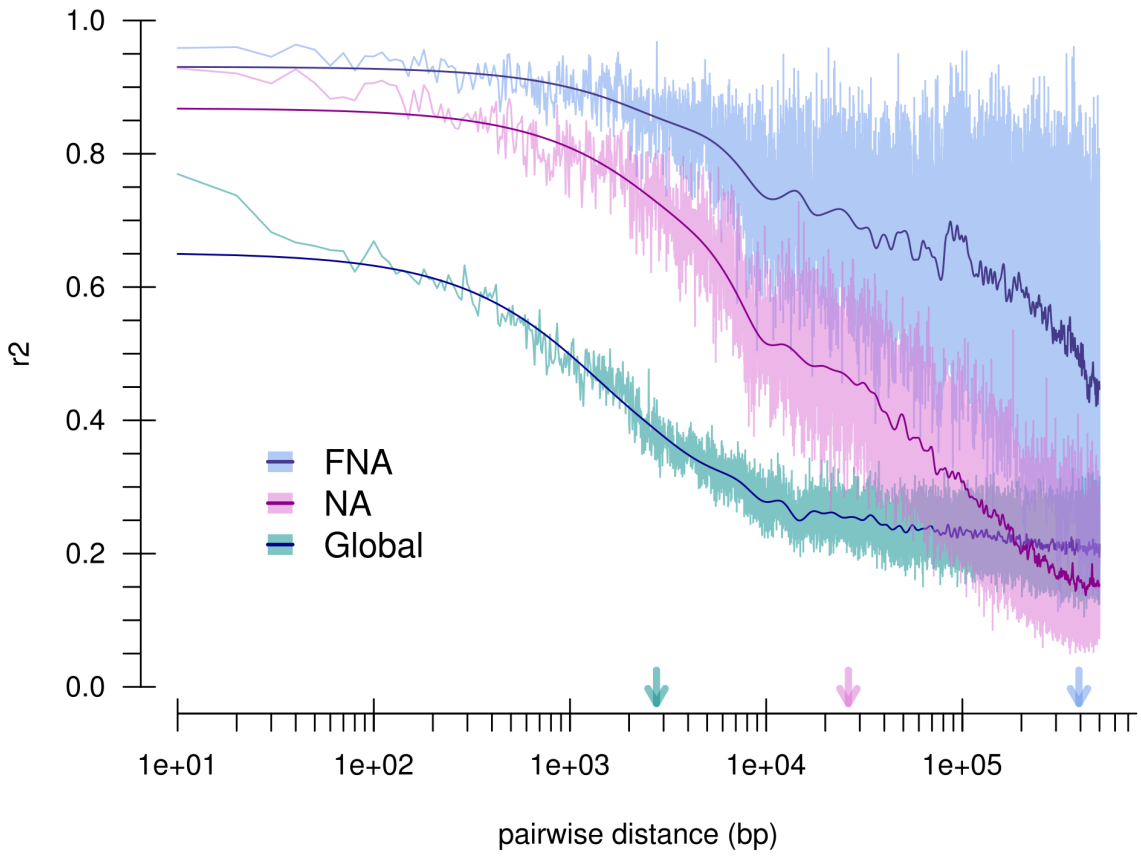

1.0
0.8
0.6
0.4
0.2
0.0
r2
FNA
NA
Global
1e+01
1e+02
1e+03
1e+04
1e+05
pairwise distance (bp)

Figure 4.

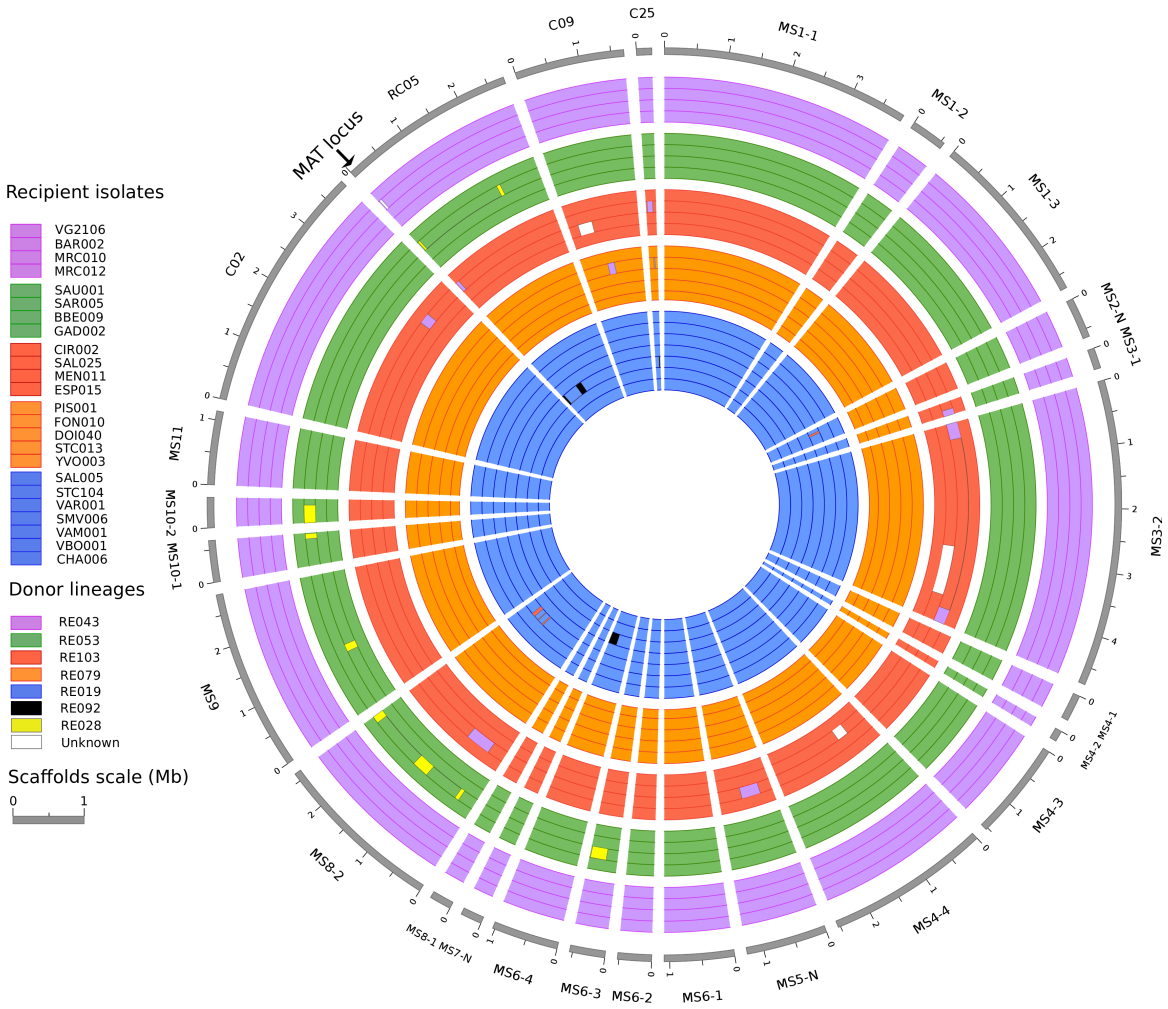
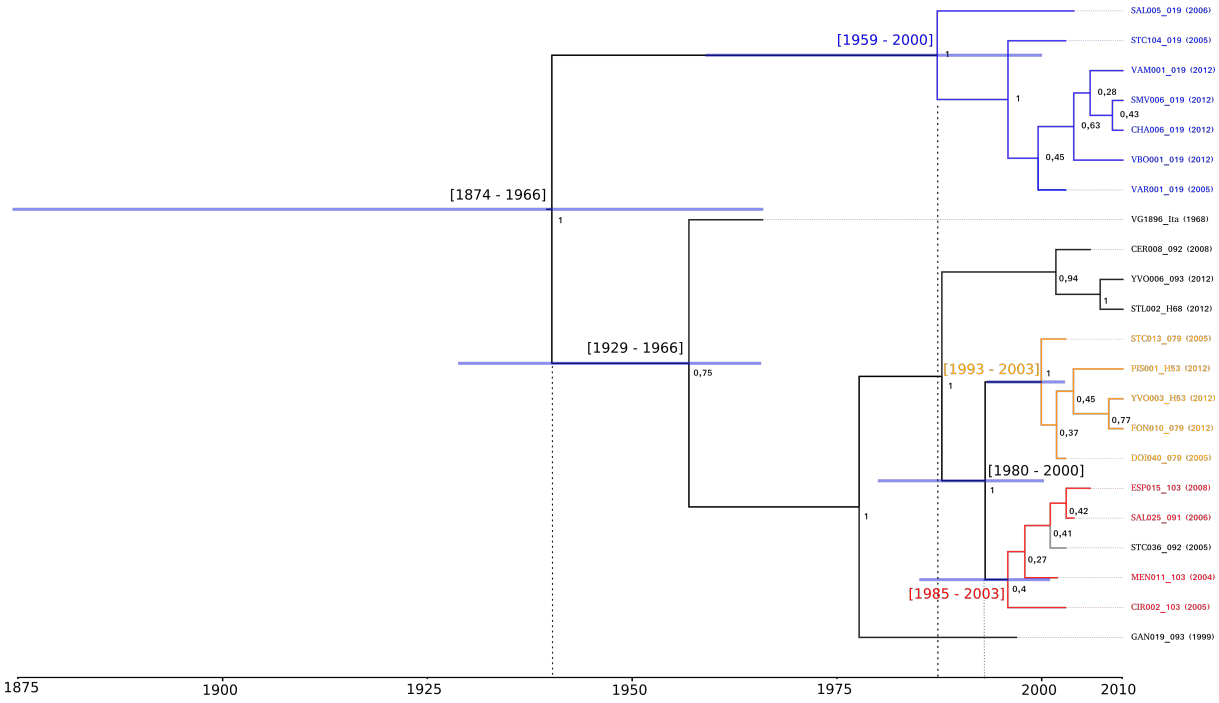

Recipient isolates
VG2106
BAR002
MRC010
MRC012
SAU001
SAR005
BBE009
GAD002
CIR002
SAL025
MEN011
ESP015
PIS001
FON010
DOI040
STC013
YVO003
SAL005
STC104
VAR001
SMV006
VAM001
VBO001
CHA006
Donor lineages
RE043
RE053
RE103
RE079
RE019
RE092
RE028
Unknown
Scaffolds scale (Mb)
0
1
MAT locus
C25
C09
RC05
C02
MS11
MS10-2
MS10-1
MS9
MS8-2
MS8-1
MS7-N
MS6-4
MS6-3
MS6-2
MS6-1
MS5-N
MS4-4
MS4-3
MS4-2
MS4-1
MS3-2
MS2-N
MS3-1
MS1-3
MS1-2
MS1-1

Figure 5.


[1959 - 2000]
[1874 - 1966]
[1929 - 1966]
[1993 - 2003]
[1980 - 2000]
[1985 - 2003]
SAL005_019 (2006)
STC104_019 (2005)
VAM001_019 (2012)
SMV006_019 (2012)
CHA006_019 (2012)
VBO001_019 (2012)
VAR001_019 (2005)
VG1896_Ita (1968)
CER008_092 (2008)
YVO006_093 (2012)
STL002_H68 (2012)
STC013_079 (2005)
PIS001_H53 (2012)
YVO003_H53 (2012)
PON010_079 (2012)
DOI040_079 (2005)
ESP015_103 (2008)
SAL025_091 (2006)
STC036_092 (2005)
MEN011_103 (2004)
CIR002_103 (2005)
GAN019_093 (1999)
0,28
0,43
0,63
0,45
0,94
0,75
0,45
0,77
0,37
0,42
0,41
0,27
0,4
1875
1900
1925
1950
1975
2000
2010

Chapitre 2 : Identification of structural variations in the genome of the chestnut blight fungus during successive world-wide introductions

Préambule

Dans le chapitre précédents, nous avons pu proposer ces deux conclusions : 1) en France, des signatures de recombinaisons génétiques sont détectées entre les lignées génétiques étudiées. 2) Les lignées émergentes détectées au nord de la France sont entièrement apparentées au groupe génétique nord-américain. La recombinaison génétique est donc possible entre les lignées génétiques présentes en France, comme cela était suggéré par la présence des deux types sexuels dans les populations (Dutech et al., 2010 ; Robin e al., 2017) et par la détection des structures sexuelles sur le terrain. Pourtant, les lignées génétiques se maintiennent au nord de la France et l'émergence des lignées n'est pas causée par l'admixture des deux groupes génétiques nord-américain et asiatique (Demené et al., 2019 ; Chapitre 1). L'observation des croisements limités entre les lignées de *C. parasitica* en Europe ont posé la question d'éventuelles barrières à la recombinaison génétique entre ces lignées ; les lignées co-existant souvent dans une même parcelle forestière (Dutech et al., 2008, 2010). Ce chapitre 2 a été construit pour répondre à une des hypothèses à ces barrières génétiques qui serait que l'existence de réarrangements chromosomiques majeurs entre les groupes génétiques introduits en Europe limitent les croisements en perturbant l'appariement des chromosome homologue lors de la méiose. Trois autres hypothèses permettant d'expliquer la détection de signatures de recombinaisons limitées dans les lignées clonales sont proposées dans le chapitre 3 du manuscrit.

Plusieurs méthodes permettent de mettre en évidence des réarrangements chromosomiques entre plusieurs génomes. En 2000, Pöggeler et ses collaborateurs ont utilisés des croisements contrôlés et des migrations d'ADN à l'aide d'électrophorèse afin de mettre en évidence que des différences de tailles de génome entre différentes souches de *Sordaria macrospora* pouvaient être liées à des inviabilité des hybrides. Plus récemment, De Jonge et ses collaborateurs (2013) ont produits des « draft » génomes de plusieurs souches de *Verticillium Dahliae*. En s'appuyant également sur une analyse de caryotypage, ils ont put mettre en évidence des réarrangements

chromosomiques (élongations) impliqués dans un isolement reproducteur entre certaines souches. Enfin, en 2015, Badouin et ses collaborateurs ont séquencé plusieurs souches de *Microbotryum lychnidis-sioicae* à l'aide de la technologie PacBio afin d'obtenir des génomes de haute qualité et mettre en évidence que des inversions et des translocations supprimaient la recombinaison dans certaines régions du génome. Afin de choisir quelle méthode utiliser, je me suis demandé quel type de données pouvait être le plus utile à la communauté scientifique, et quel type d'analyse pourrait être le plus pertinent dans le paysage scientifique actuel.

J'ai décidé d'assembler plusieurs génomes d'isolats de *C. parasitica* provenant de l'ensemble de son aire de répartition afin de détecter d'éventuels réarrangements chromosomiques. J'ai choisi d'utiliser la technologie Nanopore (Oxford Nanopore technologie) pour produire des longs reads, qui sont essentiels dans l'assemblage d'un génome *de novo* car ils permettent de traverser les régions de basse complexité, souvent riches en éléments répétés et éléments transposables. Les assemblages s'appuyant seulement sur l'utilisation de « short-reads » permettent d'assembler précisément les régions génomiques complexes, particulièrement utiles pour des études qui s'appuient sur le polymorphisme nucléotidique, mais résultent aussi en des assemblages très fragmentés ne permettant pas d'inférer précisément les variations structurales entre les génomes. Mon choix s'est porté vers la technologie Nanopore pour deux raisons. 1) La technologie m'a séduit car elle s'appuie sur l'utilisation d'une protéine transmembranaire possédant un pore d'un diamètre avoisinant le nanomètre (un nanopore) pour lire une séquence entière d'ADN à partir des variations de taille du pore lorsque les différents nucléotides la traversent. 2) La plateforme Génome-Transcriptome de Bordeaux s'est orientée vers cette technologie en 2017, et il m'a paru essentiel de participer à leur projet de recherche et développement afin de rendre possible, en local, le séquençage long brin.

La première étape de ce travail a été d'adapter un protocole d'extraction d'ADN correspondant aux critères du séquençage Nanopore. Les technologies de séquençage de long brin d'ADN (Nanopore et PacBio) nécessitent des extraits d'ADN génomique de haute quantité et qualité, et l'ADN doit être le

moins fragmenté possible pour espérer un séquençage de très longues séquences. Pour cela, il a fallut adapter un protocole d'extraction d'ADN de haut poids moléculaire, les kits d'extraction ne fournissant pas les quantités suffisantes d'ADN. J'ai décidé d'essayer d'adapter moi même un protocole d'extraction d'ADN pour *Cryphonectria parasitica* à partir d'un protocole d'extraction publié pour des espèces *Erisyphaceae*, des champignons biotrophes obligatoires (Feehan et al., 2017). Mes travaux ont permit d'obtenir un protocole d'extraction qui a mené à un premier séquençage Nanopore (Annexe 2). Cependant, la qualité du séquençage était très faible, et n'a pas permis d'obtenir un assemblage de qualité suffisante. J'ai ensuite pu obtenir un protocole d'extraction d'ADN de haut poids moléculaire développé sur Magnaporthe oryzae, de la part de Sandrine Cros-Arteil, ingénieure dans l'équipe Biologie Evolutive des champignons phytopathogènes de l'INRA de Montpellier. Après plusieurs essais et mises au points, ce protocole a permis d'obtenir des extraits d'ADN de qualité des quatre isolats sélectionnés pour notre étude, et ont fait l'objet de quatre séquençages Nanopore distincts.

La deuxième étape de ce travail a été d'assembler d'assembler un nouveau génome s'approchant du niveau chromosomique de la souche, afin de détecter des possibles ruptures de synténie entre : quatre génomes nouvellement assemblés, le génome de référence assemblé et présenté dans le chapitre 1 et le génome de référence non publié de l'isolat EP155 (disponible ici : https://genome.jgi.doe.gov/portal/Crypa2/Crypa2.download.ftp.html). Au moins trois assembleurs ont été comparés : Ra (https://github.com/lbcb-sci/ra), HybridSPAdes (Antipov et al., 2015) et Miniasm (Li, 2015). La synthèse des résultats de ces comparaisons est en cours de préparation et fera l'objet d'une publication technique avec Sandrine Cros-Arteil (Montpellier) et Christophe Boury (Plateforme Génome-Transcriptome de Bordeaux). Mon choix s'est porté sur l'assembleur HybridSPAdes, associé à plusieurs étapes de scaffolding à l'aide du logiciel npScarf (Cao et al., 2017). Cette méthode permet d'assembler les short-reads d'un isolat en contigs, qui sont ensuite assemblés en scaffolds à l'aide des long-reads. La publication permettra de présenter le nouveau génome de référence de l'isolat Japonais de *C. parasitica*, curé manuellement et présenté dans ce chapitre 2. L'étape de curation manuel d'un assemblage,

qui consiste à vérifier les erreurs d'assemblage menant à des scaffolds chimériques, est pour moi essentielle. A minima, il est essentiel de vérifier la couverture des reads de la souche mappés sur son propre assemblage, et si des reads pairé sont utilisés, d'identifier les régions contenant un fort taux de reads non pairés. Les régions présentant des pertes ou des gains importants de couverture, associés à un ensemble de reads non pairés devraient être coupées.

Ce chapitre 2 est présenté sous la forme d'un article scientifique unique. Cependant, après discussion avec des collaborateurs et les membres du jury de ma thèse, j'ai décidé que cet article serait découpé en deux articles distincts en raison de la densité trop importante d'information présenté ici. Le premier article permettra de présenter le protocole d'extraction d'ADN mis au point par Sandrine Cross-Arteil de l'UMR BGPI à l'INRA de Montpellier, les comparaisons d'assembleurs de génome, et le génome assemblé de l'isolat Japonais de *Cryphonectria parasitica* comme nouveau génome de référence. Le deuxième article pourra ainsi être concentré sur l'hypothèse de potentiels réarrangements chromosomiques entre les génomes de *C. parasitica* échantillonnés dans l'ensemble de l'aire de répartition.

Identification of structural variations in the genome of the chestnut blight fungus during successive world-wide introductions.

Arthur Demené[a,*], Sandrine Cros-Arteil[b], Benoît Laurent[a], Cyril Dutech[a]

[a] BIOGECO, INRA, Université de Bordeaux, 69 route d'Arcachon, Cestas F-33610, France

[b] BGPI, INRA / CIRAD – Campus International de Baillarguet, 34398 Montpellier Cedex 5, France

Correspondent author: D. Arthur, e-mail: arthur.demene@u-bordeaux.fr / arthurdemene@gmail.com

Postal address: INRA – UMR 1202 BIOGECO – 69 route d'Arcachon – 33610 Cestaas – France

Declaration of interest: none

Keywords: Cryphonectria parasitica, comparative genomics, structural variation, gene content, transposable elements,

I Introduction

Interactions between host and pathogens shapes the variation at the population and the genomic level in both plants and pathogens (Möller and Stukenbrock, 2017). In the context of biological introduction particularly, fungal pathogens are often confronted to a novel host and new biotic and abiotic conditions (Parker & Gilbert, 2004). The need of rapid adaptation may lead to changes in life-history traits between introduced and native populations, likely involving genetic changes to deserve adaptation (Gladieux et al., 2015). For example, pathogens of agricultural and often annual crops must rapidly adapt to artificially selected hosts and bypass host resistances or xenobiotic. Among all, structural variations (SVs) including chromosome rearrangement and fusion, large insertions-deletions, duplications, and genes copy number variations (CNVs), are known to be major drivers of evolution for animals (Weischendfeldt et al., 2013; Le Moan et al., 2019), plants (Anderson et al., 2014) and fungi (Raffaele & Kamoun, 2012 ; Galazka & Freitag, 2014). For the latter more particularly, modification of the chromosomal structure has been reported to allows adaptation and sometimes specialization to the host, for example, associated to high variation of lineage-specific (LS) genomic regions (Ma et al., 2010 ; de Jonge et al., 2013, Faino et al., 2016), gain or losses of genes (e.g. *Magnaporthe species* see Yoshida et al., 2016 ; in *Zymoseptoria tritica* see Hartmann et al., 2017), chromosome stretching (e.e. *Leptosphaeria maculans* see Grandaubert et al., 2014), interchromosomal translocations (e.g. *Colletotrichum higginsianum* see Tsushima et al., 2019) or whole genome duplication (e.g. *Phytophthora species* see van Hooff and al., 2014). In many cases, these highly variable regions are associated with transposable-elements (TEs) rich regions that act as important fuel for genomic and adaptative changes and has received an increasing consideration (Raffaele & Kamoun, 2012). Some fungal gene products that are involved in host-infection, including small-secreted protein (SSP) and secondary metabolites (SM), are often localized in or near these regions rich in TEs (Seidl & Thomma, 2017). This typical genome architecture brought the concept of the two-speed genome

model, in which filamentous pathogen species have a rapidly evolving part of genome promoting the adaptation and a slowly evolving one preserving the modification of house-keeping genes (Dong et al., 2015). Compared to the fungal crop pathogens, the study of the role of transposable elements for the evolution of forest tree pathogens has so far received little consideration, and only a few recent studies deal with this subject. For example, a study based on 19 genomes of worldwide *Dothistroma septosporum* isolates revealed much higher transposable element and SNPs content in Central American isolates than in other ones (Bradshaw et al., 2019). As these strains have been isolated on a different pine species, the authors suggested that these transposable elements could be associated with the diversification of these strains and adaptation to the host. Genomic comparison of four species of *Armillaria*, a forest pathogenic fungal species complex causing root rot disease in gymno- and angiosperms worldwide, highlighted lineage-specific genes and elongation of genomes between these species which have diverged since a very long time (Sipos et al., 2017). These authors also showed that first, the content of transposable elements within these genomes was no greater than that of 22 close non-pathogenic species and second, their distribution was not concentrated in TE-rich regions. As these pathogens evolve in less anthropized environments, with possibly various directional selective pressures for host-pathogen interactions, and a higher host genetic diversity, the evolutionary dynamics and molecular mechanisms underlying the adaptation of forest pathogens recently introduced in a new area could be different from those infecting crops. In such a context, understanding the interaction between SVs, genes involved in the molecular dialog with host or the response to abiotic stresses, and TEs is essential to understand the drivers of an invasive success.

Here, we focused on *Cryphonectria parasitica*, an ascomycete pathogen causing the chestnut blight disease on *Castanea sp*. which have been recently introduced multiple times in temperate areas around the world (Agnagnostakis, 1987; Robin & Heiniger,1992). Native to eastern Asia, *C. parasitica* was first introduced to North America at the end of the 19th century (Milgroom et al., 1996), and more recently in Europe from North American and Asian populations during the early 20th century, with a first report in Italy in 1938 (Biraghi, 1946).

Numerous studies have described the genetic diversity and the population structure in the world, showing a shift in reproductive mode between native and invasive populations (Liu et al., 1996; Milgroom et al. 1996; Hoegger et al., 2000; Dutech et al., 2010 ; Milgroom et al. 2008). Although a high clonal structure has been described in western Europe (Dutech et al., 2010; Prospero & Rigling, 2012), a recent genomic study highlighted that genetic recombinations could be frequent between and within these lineages (Demené et al., 2019). By contrast genomic features related to these populations subdivisions remain poorly understood. By comparing a French reference genome assembly to the North American reference genome (EP155, unpublished – https://genome.jgi.doe.gov/portal/Crypa2/Crypa2.download.ftp.html), Demené et al. (2019) showed that numerous losses of large genomic regions and many synteny ruptures may have occurred between the two isolates. We cannot rule out that these changes may be due to the use of different whole genome sequencing and assembly methods. But the fact that chromosome rearrangements occurred between these isolates is worthy of an investigation. Therefore, we tested in this study the hypothesis that genomic structural modifications is rapid in *C. parasitica*, especially in the context of multiple introductions such observed in Europe, and lead to rapid genomic changes, that may contribute to the invasive success of the species.

Compared to short read mapping based methods, that are limited by the size of the small size of the produced reads to detect structural variants (Sedlazeck et al., 2017), third generation sequencing (Pacific Biosciences, PacBio and Oxford Nanopore Technologies, ONT), allow now to easily produce DNA sequences of several tens of kilobases, which can be used to accurately detect these structural variants (Merker et al., 2018). In conjunction with the increase of our knowledge in genome assembly, it has therefore made possible to get near-chromosomes level assemblies allowing a better characterization of chromosomal structures, lengths and syntenies (Plissonneau et al., 2016; Jiao and Schneeberger, 2017; Frantzeskakis et al., 2018; Nattestad et al., 2018). In this study, we used such methods to assemble four new genomes of *C. parasitica* sampled in the native area in Japan and China, and in the two

introduced regions in North America and Europe in order to compare them, as well as with the previously assembled genome of the south-eastern French isolate introduced from North America (YVO003, Demené et al., 2019) and the reference genome EP155 assembly. More precisely, we were able to explore the variations in genome structure and features which could have lead to the success of *C. parasitica* populations in new areas. We present a near-chromosome-scale assembly of a *C. parasitica* monospore isolate sampled in Japan, which has been manually curated, and three other genome assemblies from a North American, a Chinese and a south-western French isolates. These genomes were chosen to represent the successive steps of the world chestnut blight invasion from Japan to North-America, then to Italy and the south-eastern France, and from China to directly south-western Europe (Dutech et al. 2012). This experimental design should allow us to compare the genomes of isolates from invasive populations with the genomes of isolates from native populations, and to identify possible genomic signatures of structural variations (SVs) possibly involved in response to the interaction with host, or abiotic and biotic stresses.

II Materials and Methods

Genome sequencing and assembly

The four monospore isolates chosen for the study were collected in the Japanese population (isolate hereafter called ESM015), the Chinese population (XIM), and the North-American population in Depot-Hill (DUM003) (Milgroom et al., 1996), and in the south-western French populations (MRC010) (Robin et al., 2017; see Table 1 for details). Strains were multiplied on patato dextrose agar culture as described in Hoegger et al. (2000), and DNA was extracted following the high molecular weight DNA extraction protocol described in (Cross Arteil et al., see Annexe 4 of this thesis). Long reads sequencing was performed using the GridION x5 (Oxford Nanopore technology, Cambridge UK) at the Genome Transcriptome Facility of Bordeaux (University of Bordeaux, France). For each DNA sample, DNA purification using 1.0X of homemade magnetic Beads (11 % PEG) and a DNA size selection (>1kb) and clean up using an optimized SPRI beads mixture according to Schalamun and Schwessinger's protocol (2017). Sequencing was performed with a R9.4.1 Flowcell on the GridION device with MinKNOW (version 2.0), using live basecalling and dedicated basecaller guppy v1.4.3 (Oxford Nanopore technology, Cambridge UK).

In order to produce a high confidence assembly, we tested for all isolate genomes multiple pipelines including either long read correction, polishing steps using the short reads produced previously for these four monospore isolates (Demené et al. 2019), and produced during this work for the North-American genome. In details, we tested four assembler which relied on the Overlap Layout Consensus (OLC) or the De-Bruijn Graph algorythm: Canu (Koren et al., 2017), Ra (https://github.com/lbcb-sci/ra), Miniasm (Li, 2015) and HybridSPAdes (Antipov et al., 2015) (See annexe 5 of this thesis). Best genome assemblies were finally obtained using HybridSPAdes (Antipav et al., 2016) which first builds a de Bruijn graph from Illumina short reads, and then uses long reads for gap closure and resolution of highly repeat regions. Kmer size for assembly has been tested for multiple values (33, 55, 77, 99, 111 and 127) on

the isolate originating from Japan and the final k-mer size was set at 127 as giving the best results. HybridSPAdes was ran with the short read error correction step (BayesHammer ; Nikolenko et al., 2013), the accurate mode on and the Mismatch correction step.

The Japanese isolate was chosen as the new reference genome because it is the origin of the two major introductions in North-America and Italy (Dutech et al. 2012). Scaffolding of contigs was performed using npScarf (Minh Duc Cao et al., 2017) in batch mode following the protocol described in https://github.com/mdcao/npScarf. Three cycles of scaffolding following manual curation were performed to obtain the final reference genome assembly. Manual curation first consisted in removing small contigs < 1000bp. Long reads were then mapped on its scaffolds using minimap2 (Li et al., 2018), and short reads using bwotie2 (Langmead and Salzberg, 2012). Reads with a map quality of zero were removed. A Bam Coverage (Deeptools, Ramirez et al., 2016) and samtools (Li et al., 2009) were used to detect regions with 1) more than five mis-paired short reads, 2) long reads coverage higher than twice the mean coverage of the genome and 3) less than 2 mapped long reads. All these regions were manually checked, and scaffolds were in some cases cut when chimeric assembly was detected (i.e incorrect mapping of paired-end reads). The mate pairs information was used to split together the extremities of contigs, and construct new correct scaffolds.

For the three other genome assemblies, scaffolds were only cut when an evidence of a chimeric connection was detected (i.e. mis-paired short reads). We used BUSCO v3 (Simão et al., 2015) with the ascomycota odb9 single-copy orthologs set (https://busco.ezlab.org/datasets/ascomycota_odb9.tar.gz) to assess genome assembly completeness. Telomeric regions were identified on the basis of -TTAGGG- repeated sequences described for fungi (Casas-Vila, 2015), and of -TTAGGGCTAGGG- and -CTAGGG- known to be specifically associated to such regions in *C. parasitica* genome (Xiaodong Qi, 2011).

Transposable elements prediction, annotation and curation

The REPET pipeline [https://urgi.versailles.inra.fr/Tools/REPET] was used on the Japanese isolate assembly for the detection of TEs, including TEdenovo for the consensus detection (Flutre et al., 2011, Hoede et al., 2014) and TEannot for their annotation (Quesneville et al., 2005). A manual curation step of the TEs consensus sequences has been done to supress all TE sequences without at least one full length copy (FLC) found in the genome, cut the part of the TE sequences which have been inferred from only high-scoring segment pair (HSP) and modify, suppress or validate the TE sequence on the basis of its repeated extremities and genes content using the Wicker's classification (Wicker et al., 2007). This curated TE sequences were used to annotate the assemblies of the three other genome assemblies and the previous PacBio assembly (Demené et al., 2019). The distribution of TEs aallong 100kb windows was compared to the random distribution hypothesis using a $\chi 2$ test. The alignment of the TE rich region spanning 1.7Mb near the mating type locus of the four Nanopore genomes on the scaffold2 of EP155 was conducted using the NUCmer (NUCleotide MUMmer) algorithm from MUMmer v3.0 (Delcher et al., 2002; Kurtz et al., 2004), and visualized using AliTV (Ankenbrand et al., 2017).

Identification of large scale rearrangements and structural variants (SVs)

To compare the genome length, structure and synteny of *C. parasitica* isolates from the different distribution areas, we used the four genomes assembled in this study, and included the genome of the eastern French isolate YVO003, previously assembled from PacBio and IonTorrent reads (Demené et al., 2019) and the previous reference genome (EP155). To identify genome large scale rearrangements (i.e. large gaps, translocation or inversion > 10kb) between genome assemblies, we used MUMmer v3.0 (Delcher et al., 2002; Kurtz et al., 2004) allowing to identify presence or absence of synteny between the different scaffolds of the four sequenced isolates. The alignments were conducted using NUCmer with the basic options. Output alignment blocks smaller than 5kb were removed with the command delta-filter. The synteny between genomes was visualized using circa (http://omgenomics.com/circa/).

Because short reads have a poor sensitivity for SVs shorter than 1kbp (Huddleston et al., 2016), we chose a method based on long-reads mapping to detect accurately the signature of SVs. Insertions-deletions, inversions, inter-spaced and tandem duplication, and translocations has been predicted among genomes using SVIM (Heller and Vingron, 2019) after alignments of long reads of each strain over the manually curated Japanese genome assembly. Long reads obtained from the Japanese isolate sequencing were also mapped on itself to detect and remove the false positive detection of SVs (i.e. mainly due to poly-A/T regions due to the frequent sequencing error of nanopore technology associated with homopolymers; Sárközy et al., 2017). Structural variations with a quality score below 15 were removed, as suggested for datasets with high coverage genome sequencing (i.e. >40x, Heller and Vingron, 2019). TEs rich regions were defined from the locations of the manually curated library of TEs of the Japanese isolate, merged when separated by less than 10kb. Since mapping quality within these region rich in repeated elements is low, we removed SVs associated with Transposable Elements (TEs) rich regions. SVs were separated in two categories: small SVs with a sequence size less than 1kb and large SVs larger than 1 kb. We compared the number of SVs detected by scaffolds to the number of SVs expected by scaffold under the random distribution hypothesis using a $\chi 2$ test.
We analyzed more precisely the content of large inserted sequences in the three genomes compared to the Japanese genome. We conducted a mummer alignment between the insertion locus detected on the Japanese isolate by SVIM with the corresponding isolate genome in which insertion has been detected, and extracted the sequences that have been inserted. A *de novo* prediction of transposable elements has been done on the inserted sequences from the North-American (the one showing the largest number of large insertion) using the REPET pipeline. We used these predicted consensus sequences to annotate the transposable element content of large inserted sequences within the three genomes. An annotation of the Japanse isolate has been done with these consensus sequences, and the comparison with the currated annotation allowed us to determine if these TEs were already found in the genome of the Japanese isolate.

Gene, effectors and small secreted proteins (SSPs) predictions

Gene prediction was performed on the four newly assembled genomes and the previous one (Demené et al., 2019) using the BRAKER v2.1.2 pipeline (Hoff et al., 2015). In this pipeline, gene prediction used AUGUSTUS (Stanke et al., 2006 ; Stanke et al., 2008), and the gene Catalog 20091217 of the reference genome EP155 for training. For each of the five assemblies, we followed these three steps with default parameters: 1) EP155 proteins aligned on the genome assembly using GenomeThreader (Gremme 2013), 2) Augustus training from the alignment file using NCBI BLAST (Altschul et al., 1990 ; Camacho et al., 2009) and 3) Augustus gene prediction. The predicted genes were then filtered by removing those with evidences transposable elements sequence identified as described below. Genes encoding secreted proteins were predicted using EffectorP version 2.0 (Sperschneider et al., 2018; only predictions with a likelihood of 0.6 or higher were kept) and using SignalP version 5.0 (Almagro Armenteros et al., 2019; only predictions with a likelihood of 0.5 or higher were kept). Only the common genes of the two prediction were kept after intersection of the two dataset using bedtools giving a high confidence gene set. Secondary metabolite biosynthesis gene clusters were identified and annotated using antiSMASH fungal version web tool (antiSMASH 5.0, Blin et al., 2019).

Ortholog gene inferences

To identify the putative variations in genes copies number (gCNVs) between the genomes, homologous amino-acid sequences between the predicted proteins (filtered for TE overlaps to avoid spurious gCNVs) were identified using OrthoFinder version 2.2.7, defining orthogroups (i.e. groups of ortholog genes; Emms & Kelly, 2015, 2018). OrthoFinder was also ran more specifically on the high-confidence SSPs genes described previously. Along the process, we realized that the Generalized Hidden Markov Model used by Braker to annotate genes could sometimes generated different annotation for similar nuclotidic

sequences, and generated artificial absence/presence of genes between isolates. So, we curated manualy a set of 241 genes which were predicted by Orthofinder to be absent in the North-American genome and present in the Japanese strain in order to report the importance of this prediction bias.

To go further in the study of absence/presence of genes, Illumina paired-end reads of the North-American, south-western French and Chinese isolates were mapped on the Japanese isolate filtered genes set in order to identify uncovered genic sequence associated with missing genes in these three strains. We also mapped the Japanese isolate paired-end reads on its predicted genes in order to detect poorly covered genes that we considered as false positives: all genes with a mean coverage below 2 were removed. Finally, we also mapped the paired-end reads from four other isolates associated with main French clonal lineages one another south-western French isolate (RE043) directly introduced from Asia , one south-eastern and two south-central French isolates (RE019, RE079 and RE103 respectively) associated with the introduced American gene pool (Dutech et al. 2012). We added for the comparison, an historic isolate from Italy and another North-American isolate. All these paired-end reads were sequenced in a previous study (Demené et al., 2019, table S1). Mean coverage was calculated through 200bp non overlaping sliding window using the software bamCoverage (Ramirez et al., 2016). Coverage was normalized by Counts Per Million mapped reads (--normalizeUsing CPM) to avoid a bias due to the different coverage depth between isolates. A minimum mapping quality of five has been set to remove poorly mapped reads (mapq < 5) and reads mapped to multiple locations (mapq = 0). All genes with a mean standardized coverage of mapped reads below 0.5 were considered as missing in the isolate genome.

III Results

I) Genome organization of five *C. parasitica* genomes

New reference genome from the Japanese isolate

The long-reads sequencing using Oxford Nanopore technology (Cambridge UK) of the ESM015 Japanese isolate produced 453,365 reads (4.3 GB) with a N50=15,460kb (Table I). These Nanopore long reads used in combination with Illumina paired-end reads library using HybridSPAdes allowed to obtain a first assembly with 113 scaffolds for a total length of 43.46Mb (L50=4, N50=3.66Mb). The kmer size of 127 has been used for the all assemblies, as it gave the best values of nodes and edges numbers, according to the criteria of a correct assembly described in Antipav et al. (2016). 93 small Scaffolds (< 10,000bp) were removed, 20 manually observed chimeric regions were cut, 15 of which have been bonds back to the corresponding part, and eight new bonds were determined. In this assembly, we identified the mitochondria genome on the basis of an extremely high coverage of ~ 1000X with Nanopore reads and ~4000X with Illumina paired-end reads for a length of 148kb (compared to the nuclear genome-wide average of 100X with both type of reads). One 680bp long gene was predicted on the mitochondria and correspond to a probable homing endonuclease involved in hydrolysis of genomic DNA. After this first polishing, the second assembly yielded 14 scaffolds for a total length of 43,31Mb (L50=4, N50=4.8Mb). The second manual curation consisted in identifying other chimeric association between sequences associated to small ployNs produced by the scaffolding step and cutting them. Three scaffolds of size 6.3Mb, 2.2Mb and 0.8Mb were respectively cut in 3, 2 and 2 new scaffolds, leading to 18 scaffolds. During this last checking step, we used again paired-end information of Illumina reads to bind three scaffolds of 1.2, 1.1 and 1.1 Mb in one scaffold of 3.4Mb (scaffold ESM_6 in the last assembly). After this final curation, we also cut the Scaffold 1 (9.5Mb) into two scaffolds measuring 4.5Mb and 5.0Mb (respectively the scaffold ESM_0 and scaffold ESM_1 in the last

assembly). This final assembly yielded 17 scaffolds (plus the mitochondria) for a total length of 43.31Mb (L50=5, N50=4,1Mb; Figure 1). This assembly was 4Mb larger than the previously assembled genome described in Demené et al. (2019).

The BUSCO verification using the ascomycota_odb9 conserved gene set lead to the recovery of 1297 complete genes out of the 1315 genes (98.7%) contained in the dataset. We found 14 telomeric regions at the end of scaffolds (Figure 1b). The 18S ribosomal RNA has been identified on the scaffold ESM_6 between 2,312kb and 2,322kb (region covered at 2,292X due to the tandem duplication of the element), within a locally enriched transposable elements (TEs) region of 100kb (20 TE copies per 100kb compared to 4.6 estimated along the genome). The mating-type locus (MAT1-2-1 allele, GenBank acsession: AF380364.1, McGuire et al., 2001) was located on the scaffold ESM_7 between 2,777,570bp and 2,781,852bp.

We aligned the curated Japanese genome assembly (ESM015) over the previous reference genome (EP155 v2), and found many synteny ruptures between the two strains (Sup. Figure 1). Alignment between ESM015 and EP155 assemblies using Mummer, lead to 43.34Mb alignment with a given mean nucleotidic identity of 95.1 % estimated by the nucmer's algorithm. When removing small alignment (length < 5kb) and poorly aligned (Identity < 95%), segments, alignment length is 38.65Mb with a mean identity of 99.6 %. This alignment revealed 11 major ruptures of synteny between the two genomes, concerning the EP155 scaffolds 1, 2, 5, 6, 7, 10 and 11 (Sup. Figure 1) as identified in Demené et al. (2019) with the south-eastern French isolates (YVO003) obtained with the Pacbio technology and assembled using the Celera assembler PbcR wgs-8 (Koren et al., 2013).

Other isolate assemblies statistics

Genome assemblies of the North-American, Chinese and south-western French isolates yielded 38, 32 and 24 scaffolds respectively, with genome length of 39Mb, 43.94Mb, and 43.25Mb respectively. The synteny between genomes showed that the missing regions in the shorter genomes were similar (Sup.

Figure 2.). Furthermore, the recover of ~98 % of Ascomycota single copies genes from BUSCO dataset in the four assemblies, including the North-American one, defend high completeness of those assemblies.

We tested if the 4.5Mb missing in the North-American genome assembly were due to a smaller depth of sequencing in short-reads and a fraction of Illumina short reads which have not been assembled, likely associated to low complexity regions (Faino et al. 2015). We first mapped short reads produced from this isolate on its own genome assembly using bowtie2 and get near 96% of reads aligned, suggesting that the majority of the reads were used for the assembly. Assembly of the remaining 4% unmapped reads using velvet (Zerbino & Birney, 2008) with basic parameters, produced the sequence of the mitochondria (length = 157kb). Checking the mapping coverage of the 96% mapped reads lead to the detection of 12 regions with a depth higher than 300X (mean depth of the assembly = 150X). We found that most of these regions were highly similar and mainly related to one sequence. These regions could be related to low complexity regions, probably containing TEs which are highly duplicated. We assembled a new genome of the North-American isolate using another assembly method (Ra, https://github.com/lbcb-sci/ra), which use first long reads to construct the backbone of the assembly and is therefore less sensitive to low short reads coverage. This new assembly lead to a 43.58Mb genome with 14 scaffolds, close to the other assembled native genomes and including most of the 4.5 missing regions of the first assembly. Except for these regions in the North-American isolate genome assembly, identical to the missing one in the south-eastern French PacBio assembly, no evidence of large scale rearrangement were found between the four assembled genomes and the PacBio assembly. The global synteny was conserved through the whole genomes, and the cuts between the scaffolds in all assemblies were mainly due to regions rich in TEs, known to be difficult to assemble.

II) Genomic features predictions

Transposable element content

De novo prediction and the first annotation of transposable elements in the Japanese isolate assembly lead to 29 consensus sequences with 3670 partial and complete copies accounting for 2.49Mb (5.75 %) of the genome. Two consensus sequences of 1.2kb were poorly annotated during the TEannot step (One no Categorized TE and one putative retrotransposon) but both matched with more than 1,000 copies in the genome assembly. Using de novo prediction and annotation in the three other genomes, we found a correspondence between these two sequences and a gypsy-like TE. This TE was well annotated in the Chinese and the south-western French isolate assemblies, and we replaced the two incomplete sequences in the Japanese isolate prediction by the complete sequence of the gypsy-like TE identified in the Chinese genome assembly. After manual curation (see sup. Table 1 for details), we finally kept 15 TE consensus sequences in the Japanese genome assembly. The second TEannot with these consensus sequences lead to 1,669 partial and complete copies accounting for 3.27Mb (7.46%) of the genome (Table 2), with 14 consensus having at least three full length copies. The mean TE number within this assembly is 4.6 per 100kb ($CI_{95\%}$=0.82), not randomly located along the genomes (Chi2=value, p-value << 0.01), with the highest density in small contigs (density > 25 per 100kb for the scaffolds ESM_11, ESM_12 and ESM_15; Sup. figure 3). TE contents between the five assembled genomes were different and ranged between 783 and 1934 copies (including incomplete copies) per genome (Table 5). However, new assembly of the North-American genome using Ra assembler has made it possible to detect 1,949 TEs copies accounting for 3.49Mb (8.00 %) of the genome. By contrast, the TE content within the same North-American genome, but assembled with HybridSPAdes, accounted for only 0.62Mb (1.58%) of the genome. Similarly, the TE content within the genome assembled by pacbio sequencing (Demené et al., 2019) accounted for 0.83Mb (2.12%), which could suggest an identical biased assembly and related to the unassembled regions described above.

The most widely represented TE is the gypsy like element outlined above, which contain the five protein coding domains and the long terminal repeats (LTR) sequences needed for its duplication. Its cumulative genome coverage represented more than 2.6Mb in the four genome assemblies (including the

North-American isolate assembly from Ra), and the full length copies (FLC) varied from 209 (North-American isolate), 212 (Chinese isolate), 215(Japanese isolate) and 248 (south-western French isolate)(Table 4). The assembly of the North-American strain with Ra allowed to annotate ~3Mb of TEs in addition to the HybridSPAdes assembly, a large part corresponding to the gypsy element (~2.5Mb), which underlines a poor assembly in regions rich in TEs using a de Bruijn graph assembler for the North-American with a lower coverage of paired-end Illumina reads (~50X) than the three other genomes (~100X). This gypsy TE was also not randomly distributed through the genome (χ2 value, p-value <<0.01) as other TEs, but more near the telomeric regions and within the putative centromeric regions.

The mating type locus was located on the scaffold ESM_7 in one of these regions rich in TEs of 270kb length. Ninety-three distinct TE complete and incomplete copies were identified in this region, including 50 copies of the gypsy like element. For the 14 remaining TE families, seven were DNA transposon (One Crypton like, one Helitron like and five Mariner like TEs), five were RNA transposons (another gypsy like, two copya like and two unclassified TEs). Two of these families were poorly annotated and described as one potential host gene and a RNA/DNA transposon which could be a chimeric annotation that we did not detect during the manual curation of the first TEannot. All described TE families had at least one FLC except for: the spurious RNA/DNA TE within the Japanese, the North-American the two French isolates assemblies, and the Crypt1 DNA transposon within the Chinese Assembly. The predicted consensus sequence of Crypt1 DNA transposon, an active Ac-like transposon, was 3,572 nucleotide length while Basso et al. (2001) found a consensus sequence of 3,563bp (Accesssion n° AF283502.1). Alignment of both sequences with Clustal Omega web tool (Sievers & Higgins, 2014) lead to 100 % of identity between them (Sup. Data1).

Gene predictions and gene copy number variations (gCNVs)

Using BRAKER and after removing predicted genes overlapping the predicted TEs, we found a number of gene between 11103 (Chinese isolate) to 11630

(Japan isolate), comparable to the 11,609 genes in EP155 genome annotation v2 (Table 3). Orthofinder allowed to detect 9,600 genes in single copy between the four assemblies, with 9,934 to 10,153 genes being in single copy between the Japanese, the Chinese and the south-western French isolate respectively. Beacause bias in gene predictions from BRAKER are possible, we manually checked the set of genes predicted to be absent by Orthofinder in the North-American genome (242) and in the south-western French (129) but present in the Japanese genome in one or more copies. Blast search of these genes against the North-American and the south-wester French genomes using megablast has resulted in the recovery of 129 (53 % of the genes) and 114 (88% of the genes) genes with a coverage and a nucleotide identity > 90%, respectively. Second, we checked the local coverage of the blast hit regions with coverage or identity <90%. We find similar pair-end read coverage than on the rest of the genome for 88 genes missing in the North-American genome. Last, 18 missing genes were located near TEs in the North-American genome and could be putative TEs not annotated by our TEannot. We finally kept seven and 15 genes showing deletion or insertion event altering their integrity in the North-American and the south-western French genome respectively. In the North-American genome, blast search on the NCBI nucleotide database and the EP155 gene catalogue v2 showed that four of the seven altered genes have similar profiles to known coding protein genes: one NAD(P)-binding, one polygalacturonase activity related to cell wall metabolism, one carbon-nitrogen ligase activity and the Vic1b allele involved in vegetative incompatibilities between genotypes (Zhang et al., 2014). Within the south-western French genome, six sequences had an homolog profile to one known gene: two fungal transcriptional regulatory protein, one cytochrome P450, one ankyrin repeat, one N-6 adenin-specific DNA methylase and one peptidase G1. The two last genes and one transcriptional regulatory protein coding gene were found in secondary metabolite clusters, and were altered because of insertions sized to 2 to 7kb.

Putative small secreted protein encoding gene content

EffectorP predicted ~1,000 genes likely to be small secreted proteins (SSPs) in the five genomes, and SignalP ~1,100 genes. We obtained between 77 (Chinese and North-American isolates) and 88 (Japanese isolate) high confidence predicted secreted protein coding genes, after intersection of EffectorP and SignalP predictions (Table 3). In order to determine whether the non predicted putative secreted protein genes are actually absent in the four genomes, we first predicted the orthology relationships between them. Orthofinder defined 81 groups of ortholog genes. One group contained three paralog genes within each genome, and 15 groups of single copy ortholog genes were lacking in at least one of the five genomes. After blast search of these genes in the genomes in which they were missing, we found that the set of putative secreted proteins encoding genes were completely conserved between the five genomes. Finally, only one of these putative secreted proteins encoding genes was fragmented in the Chinese genome, and no homology blast results found on the nucleotide collection from NCBI or on the EP155 gene catalogue v2.

Putative secondary metabolite genes clusters

A total of 47 predicted SM clusters, containing between 5 to 53 genes, were annotated from the Japanese reference genome. Near half (n=23) of these cluster were categorized in the Type 1 polyketide synthase group. Location of SM clusters according to TE rich regions was then investigated. These regions were predicted by merging all TEs separating by a distance less than 50kb between them, and keep all regions of size greater than 65kb. These regions contained 1,621 TE copies out of the 2,024 copies found on the Japanese isolate assembly, for a total length of 6.5Mb. The 47 secondary metabolite gene clusters were significantly closer to these putative heterochromatic regions (mean distance = 406kb, CI95%= 137kb) than a set of 1000 randomly chosen genes (mean distance = 649kb, CI95% = 34kb) (Figure 3.a.). Out of the 47 clusters, ten were located inside the TE-rich regions.

III) Identification of structural variations (SVs)

Long-reads from the Chinese, North-American and south-western French isolates were mapped on the Japanese curated assembly using minimap2, and we used the mapping depth to detect SVs with SVIM (Heller & Vingron, 2019). We removed 175 detected SVs (all were small SV < 1,000bp) considered as false positive according to SVs detection on the Japanese isolate mapped on itself using its long reads. These false positives were deletions (83%) or insertion (17%), mainly due to homopolymere of A/T nucleotides. A total of 482, 454 and 363 signatures of SVs were detected within North-American, Chinese and south-western French genome compared to the Japanese genome. Respectively, 153, 159 and 172 SVs were found in regions rich in TEs wich are highly variable between the four genomes, as suggested by the absence of synteny detected in these regions (Figure 2). Out of these regions rich in TEs in the Japanese genome, we detected more small SVs than large ones (Table 2). Structural variants were not randomly distributed for all three isolates, when considering all scaffolds (Chi2, $p\text{-}value << 0.01$). In particular, the small scaffolds ESM_11, ESM_12 and ESM_15 have a higher rate of SVs than expected for all the three isolates (respectively ~21, ~19, ~16 on ESM_11, ESM_12 and ESM_15 compared to a random expectation of ~6, ~5 and ~3).

Out of TEs rich regions, 333, 299 and 201 SVs were detected within the North-American, Chinese and south-western French genome. Their distribution was random when removing the scaffolds bellow 1Mb (six scaffolds and mitochondria). Most of these SVs were deletions (50.9%) or insertions (47.1%), and only 5 inversions and 18 duplications were detected (Table 2). Only one large deletion impacted an exon, and it was located on the scaffold ESM_6 between the position 3,391,000 and 3,393,000kb. This region was missing within the three genomes (Chinese and introduced isolates) while present in the Japanese genome, and a blast search on the EP155 gene model identified a protein homolog to a SPRT-like metalloprotease, a protein known play a role in DNA repair during replication of damaged DNA. On the basis of SVs detection, the most identical related isolate to the Japanese reference genome is the south-western French isolate introduced directly from Asia.

Twelve out of the 18 detected duplications contained the sequences of classI (Gypsy and copya) and classII (mariner) TEs. In 11 out of these 12 duplications, the TE sequence was complete. Six other interspaced duplications (meaning these duplications are not in tandem) were detected (for a total length of 11,918 to 11,920 bp): four in the south-eastern French isolate, one in the North-American and one in the Chinese isolate. These duplications all had the same sequence on the scaffold ESM_1, a length between 61kb and 73kb, and were not related to a TE or gene. Nucleotidic blast search of this 12kb length region on NCBI Nucleotide collection and on *C. parasitica* gene model gave no homology result. The origin and importance of this duplicated sequence within the three analyzed genomes remain unknown. These results suggest that at least three TEs families (including this unknown sequence) could be active within these genomes.

Characterization of large insertions within genomes

The total length of insertions was 183kb, 118kb and 172kb in the North-American, Chinese and the south-western French genomes respectively, while deletions total length was 36kb, 36kb and 16kb. The number of large insertions was higher within the genome of introduced isolates, 56 in the North-American isolate and 40 in the French isolate introduced from China, than in the Chinese isolate (35), and insertions was sometimes located in genes (Table 2). Because TEs seems to be largely implicated in the detected duplications, and that large insertions account for more than 100kb in the three studied genomes, we characterized the TE content in these insertions. Four complete sequences of transposable elements have been detected in the North American genome and used to annotate the insertion content in the two other genomes: crypt1, a class II transposon belonging to the hAT family of Activator transposable elements (Linder-Basso et al., 2001), crypt2 (Linder-Basso 2003), a mariner like DNA transposons, and one copya like RNA transposon. 43 out of the 56 insertions in the North-American isolate genome, 24 out of the 40 insertion in the French isolate genome, and 22 out of the 35 insertions in the Chinese isolate genome were caused by one of these TEs. The most frequent TE in

these inserted sequences was crypt2 (DNA transposon) and was found in 14 (North-American), 9 (French), and 16 (Chinese) full length copies. Only one copy of crypt1 was found in the Chinese isolate genome, with a crypt2 copy inserted within its sequence. This complex crypt1/crypt2 has already been detected in Chinese populations, and found to be ubiquitous in *C. parasitica* (Linder-Basso 2003). In contrast, seven full length copies of crypt1 were found within the inserted sequences of the North-American strain, plus one copy of the complex crypt1/crypt2. Five full length copy of crypt1 were found in the French isolate, and two copies of the complex crypt1/crypt2 suggesting a duplication of this complex in this genome.

Within the North-American isolate genome, the distance between the 56 large insertions and secondary metabolite clusters was significantly lower (mean distance = 316kb) than the distance between all the SVs and the secondary metabolite gene clusters (mean distance = 466kb; Kruskal-Wallis test $p<<0.05$; Figure 3b). This relation was not significant within the two other genomes, and six and five large insertions were found inside a secondary metabolite cluster in the two introduced isolates genomes (respectively the North-American and the south-western French), while only one was found inside a cluster in the Chinese isolate genome. Moreover, the copies of the four TEs found in these large insertions were all nearly identical (Mean pairwise identity = 95.6%; Figure 3.c). The crypt1 transposon was the most conserved of these four TEs (mean pairwise identity = 98.2%).

IV Discussion

Genomic structure is conserved throughout the geographical range of *C. parasitica*

We report a new reference genome of *C. parasitica* from both Nanopore (Oxford Nanopore technology, Cambridge UK) and Illumina sequencing of an isolate sampled in Japan. This genome is composed of 17 scaffolds and the mitochondria for a total size of 43.31Mb. This assembled genome is slightly

smaller and more fragmented than the nine chromosomes for a total size of 50Mb estimated by cytology (Eusebio-Cope et al., 2009). The scaffolds 2, 3, 4, and 6 owned two telomeric regions at there extremities and are likely to correspond to four complete chromosomes, out of the nine described previously (Eusebio-Cope et al., 2009). A deeper analysis in order to detect the nine centromeric regions should be performed to improve the quality of this assembly, but at least four regions could be centromeric canditates on this four scaffolds on the basis a low GC content (Smith et al., 2012). As discussed in Faino et al. (2015), it is clear that long run sequencing technology has greatly helped the assembly of this genome. However, our results also showed that region with low complexity remains difficult to assemble, even with the large depth obtained here (more than 100X), and still largely depend of the bio-informatic algorithm used. The Nanopore technology being still recent, we believe that methods of assembly will improve in a next future.

Contrary to our hypothesis and the observations made in other species (Plissonneau et a., 2016) we did not reveal important structural variation that could be easily linked to the invasion of *C. parasitica* into different geographical area. Structural variation are likely to be not involved in the adaptation of C. *parasitica* to new host species such as observed in Zymoseptoria species (Stuckenbrock et al. 2010), nor to the possible gene flow barriers suggested by the low genetic admixture between the different introduction observed in western Europe for the species (Dutech et al., 2012). However, we found at least 11 ruptures of synteny between the new reference genome assembled in this study and the previous EP155 reference genome assembly. At this time, it is not yet possible to disentangle if they are true variations or assembly errors but knowing that the EP155 isolate has been studied for at least 30 years (L'Hostis et al., 1985) and has been transplanted many times, the risk that *In vitro* culture has promoted the development of chromosomal modifications such as translocation and aneuploidy (Dunham et al., 2002; Huang et al., 2011) is worth to be considered. The re-sequencing of the EP155 isolates using similar technology and analysis pipeline will allow to identify the causes of these large ruptures of synteny observed in this study.

The structure of genes and more particularly effectors repertoire is conserved

We predicted a comparable number of genes (between 11,103 and 11,630 genes) in the five genomes. This estimation is close to the 11,609 genes annotated in EP155 genome. A first comparison between our new reference genome showed that most of genes present in single copy (86% of the predicted genes) were also present in the two genomes sampled in introduced areas. This result indicated that no strong variation of the gene repertoire likely occurred between the origin area of *C. parasitica* and the two independent introductions. This result does not support the hypothesis that such a genomic change may be at the origin of rapid adaptation to the host plants such as observed for some crop plant pathogens (Bao et al., 2017). This preliminary result should be improved by a better comparison including more isolates from the introduced areas and other genes detected in several copies in the different genomes. Furthermore, other source of polymorphism with potential role for adaptation (Laurent et al., 2017), as SNPs and InDel should be investigated as potential explanatory of the invasive success of *C. parasitica*. Noteworthy, we revealed that this comparison was limited by spurious losses and gains detected by Orthofinder and the certain level of stochasticity of the Braker gene annotation, as briefly mentioned in Frantzelkakis et al. (2018). Indeed, our manual verification showed that most of the absences of gene detected by Brakers were false positive. Nevertheless, we detected 8 and 15 true loss of genes in the North-American and the south-western French isolate introduced directly from Asia. The losses would be due to a small insertion disrupting the reading frame of these genes or deletions of these genes. These genes are involved in direct fonction such as polygalacturonase activity related to cell wall degration, or N-6 adenin-specific DNA methylase, peptidase G1 and transcriptionnal factors coding genes wich are located in SM clusters. A precise characterization of these genes in association with analyses of their transcriptional activity during infection could highlight the implication of their loss on the interaction between *C. parasitica* and *Castanea* species.

Overall, we identified 88 high confidence putative effector protein coding genes. The initial absences of these genes detected by Orthofinder could be due to a bias in the prediction of genes and their coding regions, altering the amino acid sequences used as input for EffectorP and SignalP. All these effectors was detected in the five genomes, except for one which was fragmented in the Chinese genome, and gave no homology with known protein. This set of 88 putative effector proteins is comparable to the effector repertoire that have been detected in *Botrytis cinerea* and *Sclerotinia sclerotiorum* using a similarmethod (*i.e.* 94 to 98 effector protein; Mousavi-Derazmahalleh et al., 2019). Both species are necrotrophic ascomycetes, similar to *C. parasitica*, with a broad host spectrum. *Seclortinia sclerotiorum* is able to infect more than 400 plant hosts (Boland & Hall, 2009), and *B. cinerea* is a devastating pathogen infecting more than 200 plant species (Ma & Michailides, 2005). *Cryphonectria parasitica* is able to infect different host families like many *Castanea* species, some *Quercus* species and even *Eucalyptus* species (Old & Kobayashi, 1988; Bissegger & Heiniger, 1991). A such limited repertoire of small secreted proteins sounds paradoxical to its apparent ability to jump to these different hosts, and to trigger out their defenses. Different arguments can be given to explain this point, either corresponding to methodological limitation, biological realty or both. First, the fact that SSP number and host range is strictly correlated is certainly not true, as some fungi can target common and basal mechanism of plant to process infection (Lo Presti et al., 2015), while a wide variety of other effectors other than SSP have been shown to be involved for pathogenesis as for example SM, CAZymes, transcription factors, small RNA, etc (Hua et al., 2018). The second point is that the stringency of our methodology, that was designed to limit the prediction of false positive, is likely producing false negatives. Indeed, another methods conducted to study the same *Botrytis cinerea* and *Sclerotinia sclerotiorum* referred to 499 and 432 secreted proteins respectively (Heard et al., 2015), compared to 94 and 98 with a similar approach than our. It highlights the limitation of bioinformatic prediction and the need to adapt the right method accordingly to the scope of the analysis.

Nonetheless, our findings support that variable number of genes and the presence/absence of SSP is maybe not the driving process of the *C. parasitica* host expansion during invasion both in North-America and western Europe. This result contrasts with the common findings on the evolution of invasive populations of crops pathogens, some of them being also necrotroph. These studies highlight significant variations in the SSP genes repertoire that are involved in adaptation to the host plant (Faino et al., 2016 ; Bao et al., 2017 ; Stukenbrock et Dutheil, 2018). Moreover, no relationship between SSP protein sequences location and transposable element rich regions has been found in *C. parasitica.* While the activity of transposons is often associated to one major driver of evolution in the frame of the “two speed genome” concept (Rouxel et al., 2011; Raffaele & Kamoun, 2012; Daverdin et al., 2012). Furthermore, it is also possible that host adaptation in less anthropogenic environment is less selective than in intensive agro-ecosystem (Ellis, 2015). Consequently invasive forest pathogen may adapt more to environmental factors than host defenses, favoring more universal- toolkit.

Similar transposable elements content among the four assembled genomes

We found comparable transposable elements content in the four studied genome*s*, with approximately 8% of the genome containing TEs. This rate is lower than several ascomycetes species harbouring up to 76.4 % of their genome having TEs (Amselem et al., 2015). Interestingly, the TE content of *C. parasitica* is comparable to those found in others necrotrophic pathogen, as *Botrytis cinerea* (0.7 to 2.2 % of genome covered by TEs) and *Sclerotinia sclerotiorum* (9.5%) (Amselem et al., 2015). The TE content of our assemblies with near four time higher than TE content found in the previously published genome of an isolate sampled in the south-eastern France (2.2%, PacBio reference genome YVO003; Demené et al., 2019). As for the North-American isolate genome assembled in this analysis, 1.6% of its total length was covered by TEs, that is comparable to the YVO003 genome (Demené et al., 2019). Alignment between these two genomes showed they both miss the same

genomic regions compared to the manually curated Japanese genome (ESM015), i.e. mainly telomeric regions. After re-assembling the North-American isolate gneome using a different assembler (Ra), a 4.8Mb larger genome was obtained, and lead to the detection of TE covering 8% of the genome. Combined, these results showed that regions highly concentrated in TEs were not correctly assembled in both the first assembly of the North-American genome and the PacBio genome published in Demené et al. (2019). Although the emergence of long-read technologies have considerably improved the assembly of these regions of low complexity (Faino et al., 2015; Badouin et al., 2015), we emphasize the importance of having sufficient sequencing coverage, and an assembler adapted to the type of sequences, to be able to accurately characterize TE-rich regions in genomes.

Transposable elements were mainly located in putative telomeric and centromeric region in *C. parasitica* genomes. We found 15 different famillies of TEs, which represent the expected composition of TEs in fungal genomes, with a high proportion of these TEs being long terminal repeats (LTR) retrotransposons (i.e. copya or gypsy superfamily; Muszewska et al., 2011). More precisely, we found that one gypsy-like element counted for up to 80% of the TE coverage in the genomes studied here. Preliminary results suggested that the multiplication of this element could cause elongation and translocations on the chromosome harboring the mating type locus (i.e. the scaffolds ESM_7, ESM_11, ESM_12 and ESM_15 in our reference genome). However, at this stage it is impossible to know if this is due to a poor assembly of these regions. The mating type locus is located inside this TE rich regions, this pseudo-autosomal region being located at the extremity of the putative chromosome and showing a higher rate of structural variations compared to the rest of the genome. Structural modifications on this chromosome between different isolates could result in improper matching between homologs chromosomes during meiosis, and induce decreased genetic recombination. This pattern of suppressed recombination has been found on 90% of the chromosome carrying the mating type locus of *Mycrobotryum lychnidis-dioicae*, which is linked to numerous loss of genes and a high TE content in this region (Badoin et al., 2015). It was further shown in this study that the absence of

recombination in this region resulted in a strong linkage disequilibrium between the alleles of genes involved in sexual reproduction such as pheromones and pheromone receptors, protecting the disruption of necessary combinations. Sexual reproduction is assumed to be less frequent in western Europe populations of *C. parasitica* compared to North-America and China, despite the presence of both mating type alleles in the populations (Demené et al., 2019). A better characterization of the gene content and linkage disequilibrium between alleles in this MAT region could help to understand the implication of a this TEs rich region on the apparent switch of reproduction mode in European populations of *C. parasitica*.

Insertions are mainly due to transposable elements

We detected between 363 and 482 structural variations between the four sequenced genomes using SVIM (Heller & Vingron, 2019). These SVs included mostly deletions and insertions that is a frequent result when such comparison is performed between species (Sillo et al., 2018). Since approximately one third of these SVs were located in TEs rich regions and mapping depth and quality in TEs rich regions being extremely low when mapping long reads, SVs number in these regions could indeed be highly underestimated. By limiting our comparison to well assembled genomic regions, the number of detected SVs out of these regions was higher between the North-American and the Japanese genomes than between the French and the Japanese genomes. This result is surprising because the two former ones are supposed to be more genetically related than the two latter ones. Indeed, the North-American introduction was assigned to Japanese genetic pool, and the northern French isolates sequenced here more assigned to the Chinese genetic pool (Dutech et al. 2012, Demené et al. 2019). This suggests that the occurrence of structural variations may be a frequent and random process, mainly neutral, leading to a genetic signature lowly related to genealogies of individuals in *C. parasitica*.

To characterize the impact of TEs movements and multiplication, we focused on large insertions outside the TEs rich regions. In *C. parasitica*, more than half of the large insertions were caused by TEs. We showed a significant relationship

between these insertions and secondary metabolites gene clusters, which were located closer to the telomeric and centromeric regions than other genes, a genomic structure commonly observed in fungi (Palmer & Keller, 2010). In particular, we found that the DNA transposable element Crypt1 previsouly identified in single copy in multiple Chinese isolates (Linder-Basso et al., 2001; Linder-Basso 2003), was found in multiple and highly conserved copies in the whole genomes studied in this study (>95% homology). Only one copy has been found in the Chinese isolate genome sequenced, but this TE has been found in four full length copies within the Japanese genome, and in seven and five copies in the insertions detected in the two introduced isolate genomes. These findings suggest that Crypt1 is active, with a recent duplication of this TE within the genomes of introduced isolates. Other genomes of the native and introduced area should be analyzed to give more credit to this hypothesis. Crypt1 belongs to the family of activator like transposable element (*hAT*) which as been first characterized in the maize and Drosophila, and other transposons of this group have been found in fungi (Kempken et al., 1998). Members of this family are known to cause horizontal transfers in Drosophila (Ottonelli Rossato et al., 2014) and to regulate gene expression in eukaryotes (Atkinson, 2015). Proximal insertion of a transposable element may alter the expression of close genes (Seidl & Thomma, 2017; Stuart et al., 2016). On the other hands, secondary metabolites have a wide range of roles in cellular processes, and some are involved in the molecular dialogue between pathogenic fungi and the regulation of their biosynthesis pathways is complex and involves several interconnecting networks (Brakhage, 2013). Here, we hypothesize that movement and replication of Crypt1, as well as other transposable elements that have not been characterized yet, may play a role in the regulation of secondary metabolite gene clusters. This could explain, for example, the non-Genetic adaptation patterns to temperature detected by Robin and her collaborators (2017). Indeed, they found that isolates belonging to the same clonal lineages, the latter having very little nucleotide diversity (Demené et al., 2019), showed different growth rates as a function of temperature depending on their sampling area (in northern or southern France, Robin et al., 2017).

Actually, secondary metabolite clusters also contain effectors (Lo Presti et al., 2015). We did not find any significant proximity of the set of 88 putative effector proteins coding genes with the TE rich regions, but a significant proximity between SM clusters and these regions. Redefining a new and broader set of effector genes, including both MS, and SSP with a less stringent method than that used in this study could allow a better estimate of the distribution of genes encoding proteins that play a role in the molecular dialogue between *C. parasitica* and its hosts. Such an analysis could provide more statistical power to estimate whether the genome of *C. parasitica* falls within the scope of the “two-speed genome”.

V Tables and Figures

Tables

Table 1. Statistics of the four genomes assembled with HybridSPAdes (China, Japan, North-America and western France), the genome assembled with Ra (North-America – Ra) and the previous assembly (south-eastern France; Demené et al., 2019). n represent the number of entities, bp is for base-pair.

	Japan	China	North-America	Western France	North-America – Ra	Eastern France
Total scaffolds (n)	18	32	38	24	14	35
Total length (bp)	43,305,322	43,250,891	38,746,792	43,938,120	43,579,384	39,261,148
Max scaffolds length (bp)	5381768	5601068	5215836	7591161	7,425,325	3861409
L50	5	6	6	4	4	6
N50	4,077,748	3,339,515	2655052	4,466,397	5,325,605	2,722,809
ONT Reads (n)	453,365	430,616	370,469	416705	370,47	X
Read length N50	15460	18819	30236	19245	30237	X
% GC	50.9	51.0	53.0	50.7	50.9	52.9
% Illumina reads unmapped	3.32	1.26	4.17	0.23	NA	NA
% ONT reads unmapped (n)	NA	NA	1,11 (4123)	10,4 (43347)	NA	NA
% BUSCO genes (Ascomycete)	98.7	98.6	97.0	98.7	NA	98.7

Table 2. Statistics of transposable elements annotations on the six genomes as presented in Table 1. bp is for base-pair.

Origin	Cumulative coverage (bp)	Genome coverage	No. of genome copies	No. of full-length genome copies	No. of gypsy like TE copies (full-length copies)	Cumulative coverage of gypsy like TE copies (bp)
Japan	3,269,996	7.46%	1,669	327 (19.59%)	935 (215)	2,630,698
China	3,424,842	7.82%	1,784	329 (18.44%)	968 (212)	2,787,202
North-America	618,529	1.58%	783	83 (10.60%)	242 (10)	179,068
Western France	3,770,125	8.48%	1,934	371 (19.18%)	1,037 (248)	2,972,940
North-America – Ra	3,485,996	8.00 %	1,949	346 (17.75%)	1,013 (209)	2,644216
Eastern France	832,19	2.12%	850	112 (13.18%)	291 (13)	300,571

Table 3. Statistics of the five gene prediction on the four HybridSPAdes genome assemblies and the previous assembly (south-eastern France; Demené et al., 2019) using BRAKER, the prediction of effector genes using EffectorP, the prediction of small secreted protein using SignalP and the intersect between the two last prediction (Putative effector genes).

	Japan	China	North-America	Western France	Eastern France
Genes predicted filtered number	11,630.	11,103	11,237	11,336	11,192
Genes total length	18,501,316	18,575,485	18,812,231	18,876,849	18,790,343
% of Assembly length	42,72	42,95	48,55	42,96	48,08
Genes mean size	1,590.8	1,673.0	1,674.1	1,672.0	1,678.9
EffectorP predicted gene number	1,119	980	939	964	953
SignalP predicted gene number	1,081	1,108	1,099	1,13	1,123
Putative effector genes	88	77	77	83	82
Putative effector genes total length	57,336	49,448	49,423	54,817	55,578
Putative effector genes mean size	651.5	642.2	641.9	660.4	677.8

Table 4. Summary statistics of structural variants (Svs) detected by mapping the North-American, the Chinese and the south-western French isolate nanopore reads on the Japanese isolate assembly. Only the SVs out of TEs are reported. DEL = deletions; INS = insertions; INV = inversions; DUP = interspaced and tandem duplications. Small SVs are < 1000bp; Large Svs are > 1000bp. N-Am = North-American isolate; S-W French = south-western French isolate.

Strain	Size	Feature	DEL	INS	INV	DUP	Total	Total	Total
N-Am	Small SVs	Exons	26	7	0	0	33	260	331
		Non-coding or introns	167	59	0	1	227		
	Large SVs	Exons	1	6	0	0	7	71	
		Non-coding or introns	6	50	2	6	64		
Chinese	Small SVs	Exons	27	10	0	0	37	246	296
		Non-coding or introns	100	108	1	0	209		
	Large SVs	Exons	1	4	0	0	5	50	
		Non-coding or introns	10	31	1	3	45		
S-W French	Small SVs	Exons	12	12	0	0	24	143	195
		Non-coding or introns	61	57	1	0	119		
	Large SVs	Exons	1	1	0	1	3	52	
		Non-coding or introns	3	39	0	7	49		
	Total		415	384	5	18	822	1105	822

Table 5. Graphical representation showing the composition of the identified ortholog gene groups (orthogroups) with Orthofinder. These results are not manually curated, and no re-assignment has been done to limit the bias induced by the different gene prediction. Grey cells show orthogroups having identical copy number in the two compared genomes, colors refer to the genome where an orthogroups have at least one more copy.

Native	Chinese									
Japanese	0	1	2	3	4	5	6	7	8	9
0	52	265	2	0	0	0	0	0	0	0
1	349	9934	13	1	0	0	0	0	0	0
2	0	38	225	3	0	0	0	0	0	0
3	1	1	5	52	1	0	0	0	0	0
4	0	0	1	0	14	1	0	0	0	0
5	1	0	0	0	1	3	0	0	0	0
6	0	0	0	0	0	1	3	0	0	0
7	0	0	0	0	0	0	0	1	0	0
8	0	0	0	0	0	1	0	0	1	0
9	0	0	0	0	0	0	0	0	0	1

South-western French										
	0	1	2	3	4	5	6	7	8	9
0	33	285	1	0	0	0	0	0	0	0
1	128	10153	16	0	0	0	0	0	0	0
2	1	20	243	2	0	0	0	0	0	0
3	0	1	3	56	0	0	0	0	0	0
4	0	0	1	0	14	1	0	0	0	0
5	0	0	0	0	1	4	0	0	0	0
6	0	0	0	0	0	0	4	0	0	0
7	0	0	0	0	0	0	0	1	0	0
8	0	0	0	0	0	1	0	1	0	0
9	0	0	0	0	0	0	0	0	0	1

Introduction	North-American									
Japanse	0	1	2	3	4	5	6	7	8	9
0	69	249	1	0	0	0	0	0	0	0
1	239	10045	13	0	0	0	0	0	0	0
2	2	21	242	1	0	0	0	0	0	0
3	0	2	6	50	2	0	0	0	0	0
4	0	0	1	1	13	1	0	0	0	0
5	0	0	0	0	0	5	0	0	0	0
6	0	0	0	0	0	0	4	0	0	0
7	0	0	0	0	0	0	0	1	0	0
8	0	0	0	1	1	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	1

South-eastern French										
	0	1	2	3	4	5	6	7	8	9
0	43	273	3	0	0	0	0	0	0	0
1	227	10057	13	0	0	0	0	0	0	0
2	2	26	237	1	0	0	0	0	0	0
3	1	1	5	53	0	0	0	0	0	0
4	0	0	0	2	14	0	0	0	0	0
5	0	0	0	0	0	5	0	0	0	0
6	0	0	0	0	0	0	4	0	0	0
7	0	0	0	0	0	0	0	1	0	0
8	1	0	0	0	0	0	1	0	0	0
9	0	0	0	0	0	0	0	0	0	1

Figure 1. Nanopore and Illumina reads hybrid assembly of the nuclear *C. parasitica* genome of the Japanese isolate ESM015. From outer to inner, circles represent : the localization of transposables elements, the localisation of genes predicted with BRAKER, the GC rate per 10kb (black line represent 0.5), blocs having > 95% of Identity with the Japanese genome in : the North-American isolate assembly, the south-eastern French isolate PacBio assembly, the Chinese isolate assemble, the south-western French isolate assembly and EP155 assembly. Scaffolds are numbered from 0 to 15 and M identified the mitochondria. Orange triangles show the location of telomeric motif. The purple triangle shows the location of the mating type locus. The squares show the location of nam-1 (blue), rDNA ITS (red) and β-tubulin (green) sequences (Eusebio-Cope et al., 2009)

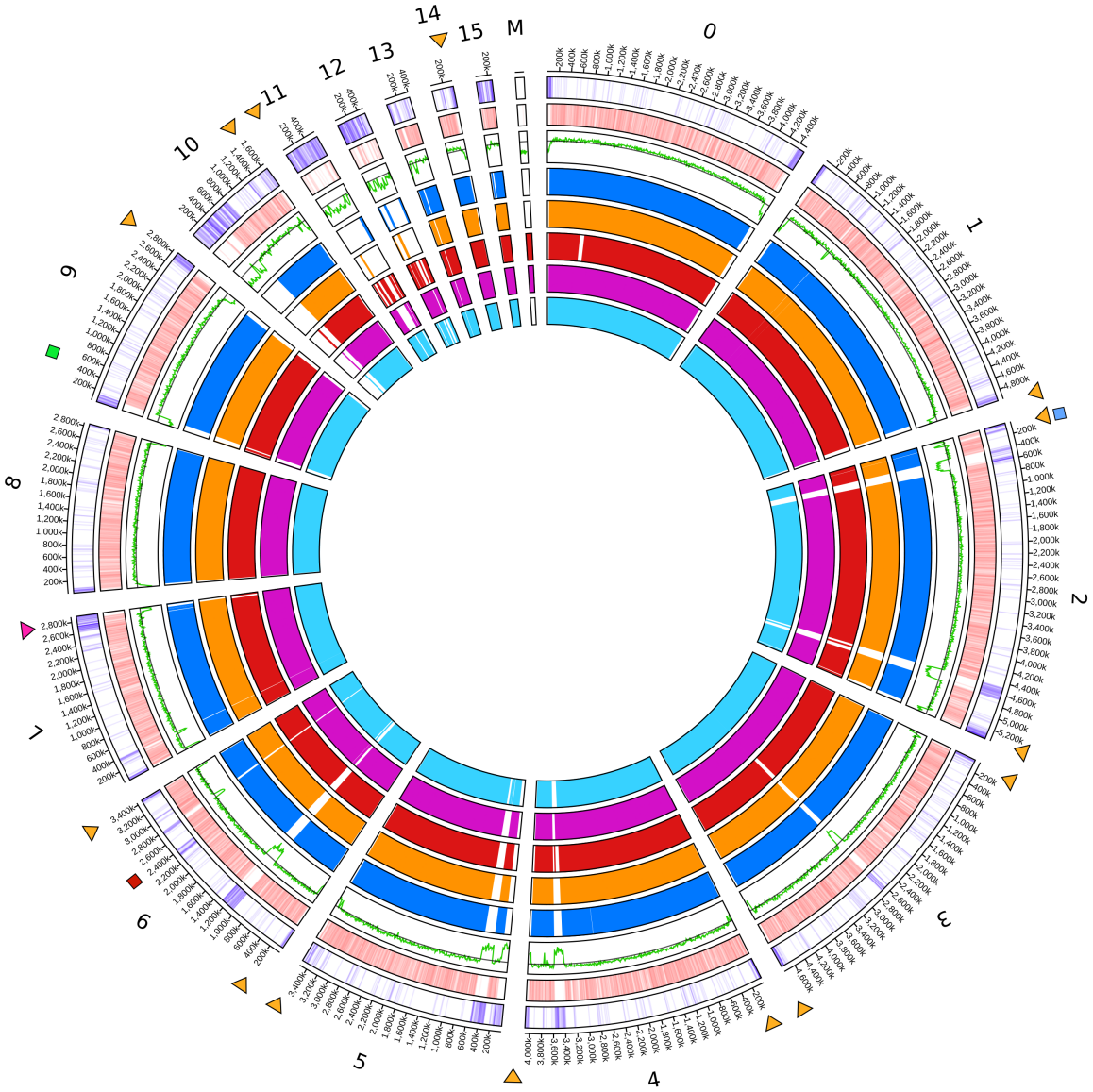


Figure 2. a) Distance of Secondary Metabolite genes cluster and of 1000 randomly chosen genes to the predicted TE-rich regions on the assembly of the *Cryphonectria parasitica* isolate from Japan. Difference between the mean of distances between the two data sets has been tested using a Kruskal-Wallis test.

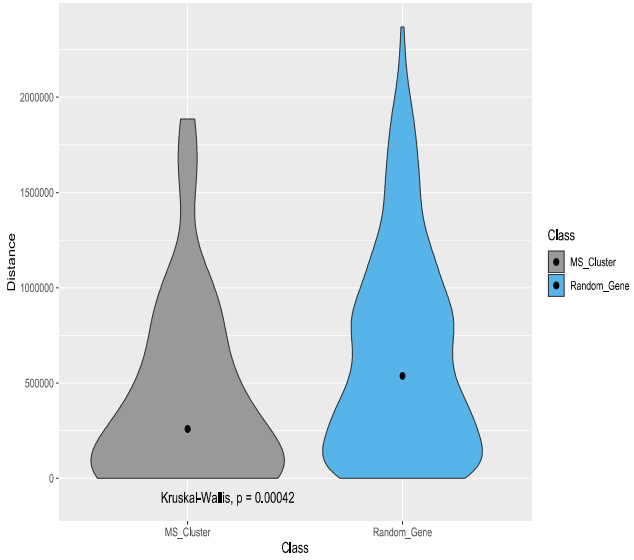


Figure 2. b) Distance of all structural variants (SVs) and large insertions to the Secondary Metabolite genes within the three genomes of Chinese, south-western French (French_SW) and North-American genomes. Difference between the mean of distances between the two data sets for each genome has been tested using a Kruskal-Wallis test. The colored points correspond to the mean distances for each data sets.

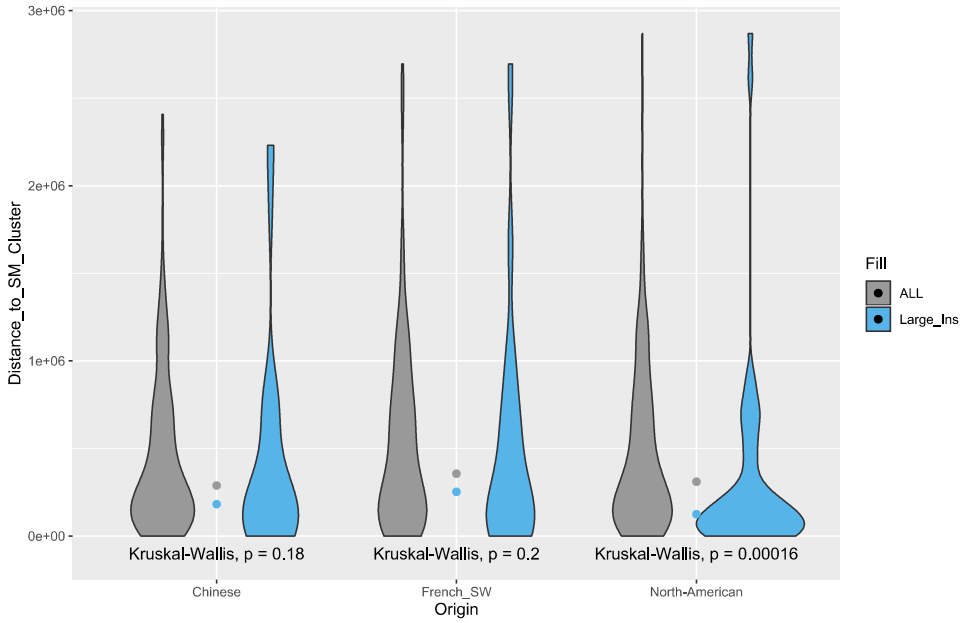
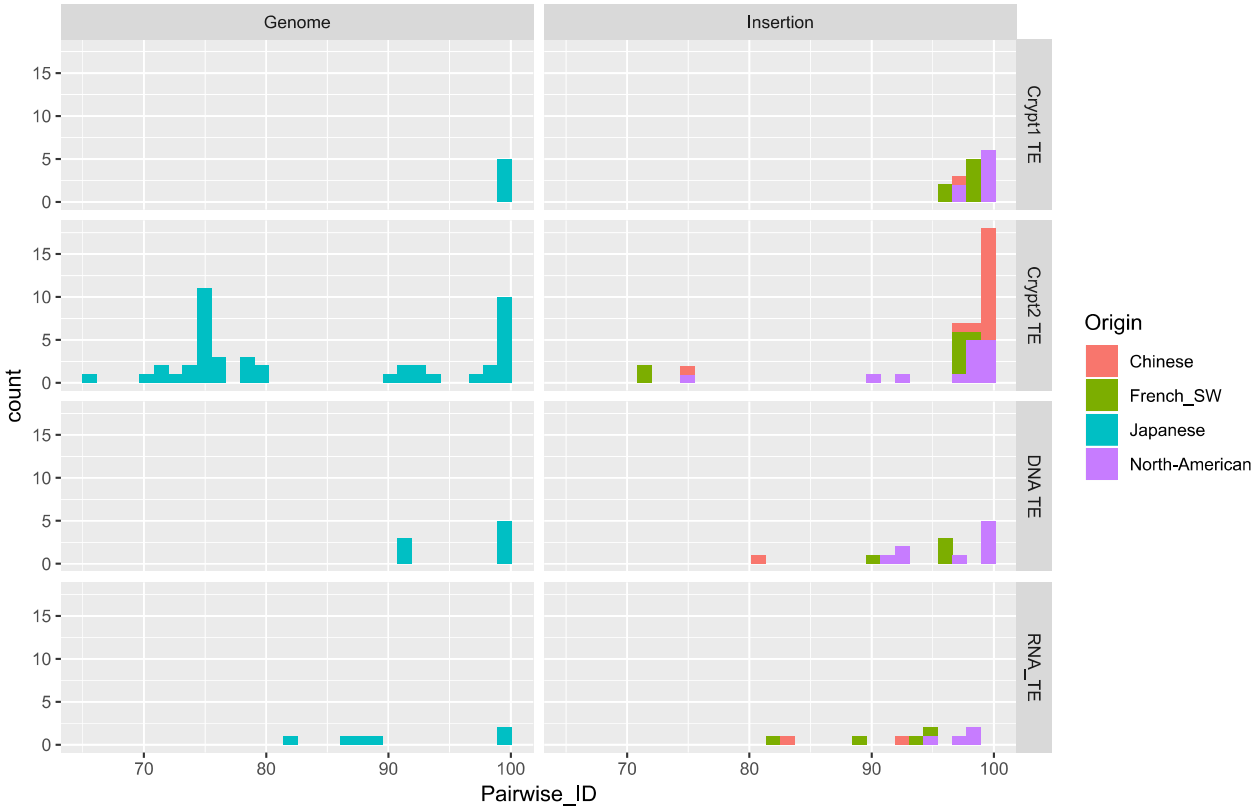


Figure 2. c) Plot of pairwise identity between transposable element (TEs) full length copies copies of the four consensus TEs detected in Large insertions estimated by the Clustal Omega multiple sequence aligner (ref). Pairwise identity between copies is given either for set of annotated sequence on the genome of the Japanese isolate (Genome) or on the inserted sequences which have moved in the Chinese, the south-western French or the North-American isolate (Insertion).



Chapitre 3 : Conclusions et perspectives

Dans le Chapitre I, le séquençage de larges parties du génome de 36 isolats de *C. parasitica* faisant partie de différentes lignées clonales identifiées précédemment à l'aide de dix marqueurs microsatellites (Dutech et al., 2010, 2012 ; Robin et al., 2017), a mis en évidence une diversité génétique intra-lignée très faible, essentiellement due à des recombinaisons génétiques entre lignées clonales, ou avec une source génétique non échantillonnée dans cette étude. Elles impliquent des régions limitées du génome dont le plus grand fragment échangé est d'environ 1Mb, soit 2,5 % du génome. Nous avons donc pu mettre en évidence que la structure clonale des populations françaises de *C. parasitica* n'implique pas seulement une reproduction purement asexuée.

La première hypothèse que nous avons proposé pour expliquer ces échanges génétiques limités entre lignées concerne des échanges parasexués, qui sont des échanges de matériel génétiques entre différents génotypes n'impliquant pas la méiose. La parasexualité a été évoquée chez *C. parasitica* par McGuire et ses collaborateurs (2005) pour expliquer la détection de variabilité de certains marqueurs (allèles sexuels et SCAR : sequence-characterized amplified regions) chez certains isolats d'une même lignée génétique. Ils font l'hypothèse que ces échanges ont été permis par la formation d'hétérocaryons (des cellules possédant plusieurs noyaux génétiquement différents) entre des isolats végétativement incompatibles. La compatibilité végétative entre deux génotypes est conditionnée par des gènes (vic : vegetative incompatibility) qui sont un moyen de différentier le sois du non-sois (Leslie, 1993). Ce mécanisme conditionne donc la possibilité de fusion cytoplasmique chez les champignons. Théoriquement, la fusion cytoplasmique (et donc la formation d'hétérocaryon) entre des isolats de *C. parasitica* possédant des allèles différents sur les six locus impliqués dans la compatibilité végétative n'est pas possible, et les isolats sont dits incompatibles (Cortesi et Milgroom, 1998). Pourtant, des croisement effectués en laboratoire ont permis d'obtenir des hétérocaryons entre des génotypes possédant certains allèles différents (Biella et al., 2002). En effet, certaines combinaisons d'allèles n'impliquent pas une réaction forte de mort cellulaire programmée lors de la fusion des mycéliums incompatibles (Biella et al., 2002). Cependant, McGuire et ses collaborateurs (2005) ne proposent pas de mécanisme qui

expliqueraient comment se font ces échanges via des recombinaisons parasexuées. Un des mécanismes qui pourraient expliquer de tels échanges génétiques n'impliquant que quelques parties du génome s'appelle le retour en croissance (return to growth). Le retour en croissance est un processus durant lequel une cellule initie la première phase de la méiose, puis quitte le cycle méiotique pour retourner en croissance mitotique. Chez la levure *Saccharomyces cerevisiae*, il a été montré que ce cycle peut être déclenché en modifiant le milieu nutritif des cellules, et qu'il entraîne un nombre de crossing over moins important que lors de la méiose (Dayani et al., 2011). Dans le cas de la formation d'hétérocaryons entre des isolats de *C. parasitica* végétativement incompatibles, on pourrait imaginer que suite à la caryogamie induisant un passage de la cellule vers un état diploïde, un retour en croissance pourrait être induit par la compatibilité incomplète des deux génomes et mener à ces échanges limités observés dans le chapitre I. Cependant, il est difficile de considérer que la fusion entre des isolats incompatibles soit si fréquente et que ce mécanisme soit la cause majeure de ces échanges qui ont été observés dans toutes les lignées étudiées.

J'ai proposé une deuxième hypothèse pouvant expliquer ces croisements limités entre les lignées clonales : les croisements sexués seraient rares du fait de réarrangements chromosomiques importants entre les lignées, et les rares croisements possibles impliquent peu de recombinaisons. Cette hypothèse a été testée dans le chapitre II et mes résultats montrent qu'a priori, cette hypothèse est à exclure car aucun réarrangement majeur n'a été détecté entre les génomes d'isolats échantillonnés dans l'ensemble de l'aire de répartition de *C. parasitica*.

Je propose enfin deux autres hypothèses dans lesquelles les croisements sont possibles et pourraient mener aux observations du chapitre I. La première hypothèse serait que les croisements sexués entre génotypes différents sont fréquents mais que seules quelques combinaisons alléliques sont favorables. Les échanges génétiques seraient donc limités à quelques zones du génome, car les recombinaisons génétiques qui brisent les combinaisons alléliques favorables induisent une épistasie négative menant à une perte de fitness des hybrides et donc, à une contre-sélection de ces derniers. La deuxième

hypothèse n'implique pas de contre-sélection des hybrides, mais s'appuie sur un processus neutre. Comme vérifié dans le chapitre I, les lignées clonales possèdent presque toutes les deux allèles sexuels permettant de réaliser de la reproduction sexuée (confirmant les résultats de Robin et al., 2009 ; et de Dutech et al., 2010). L'intra-haploid mating entre des isolats identiques tout le long de leur génome excepté sur leur locus sexuel est donc possible. Cette possibilité de croisement intra-lignée génétique, mimant l'homothallisme, couplée à une structure génétique clonale des populations pourrait largement favoriser les rétro-croisements des hybrides sur une des lignées parentale. En effet, une études d'isolats de *C. parasitica* sur une même parcelle a montré que les populations sont spatialement structurées en groupes clonaux (patch clonaux ; Dutech et al., 2008). De plus, les isolats sont pérennes et peuvent survivre plusieurs années sur le même hôte, ce qui accentue l'accumulation de clones dans les parcelles. Il est possible d'imaginer qu'un l'hybride issue de la reproduction sexuée entre deux lignées ai une forte probabilité de se développer dans un patch clonal préalablement établi. Cette structuration spatiale peut donc induire une forte disponibilité d'un même clone comme seul partenaire sexuel, et favoriser des rétro-croisements successifs sur ce clone. Ainsi, les traces d'hybridations s'effacent au fil des générations jusqu'à laisser seulement des traces d'introgressions comme observées dans le chapitre I. Ces deux dernières hypothèses sont discutées plus abondamment ci-dessous et un début de projet, malheureusement non abouti à la fin de la thèse, est proposé pour tenter d'identifier la plus probable des deux hypothèses.

Par ailleurs, au-delà de l'explication du mécanisme menant à ces échanges limités observés dans les lignées clonales, le chapitre I a permis de montrer que les lignées clonales émergentes au centre et au nord de la France présentent des mélanges d'haplotypes présents dans les lignées clonales de départ et d'haplotypes supplémentaires non retrouvés dans les lignées clonales. Nous nous sommes donc demandé quelles étaient les relations d'apparentement entre les génotypes minoritaires et les lignées clonales, et s'ils pouvaient constituer un réservoir de diversité génétique permettant l'émergence de nouvelles lignées clonales au cours de l'invasion en Europe. 29 isolats de *C. parasitica* ont été sélectionnés parmi les génotypes minoritaires

identifiés précédemment (Dutech et al., 2010 ; Robin et al., 2017). Plusieurs isolats présentent des allèles microsatellites rares non retrouvés dans les lignées principales et le séquençage de ces génotypes pourrait permettre d’estimer leur variabilité nucléotidique et haplotypique afin de déterminer s’ils constituent un réservoir de diversité génétique impliqué dans l’émergence de nouvelles lignées au cours de l’invasion Européenne.

Projet de caractérisation des génotypes minoritaires

Cette dernière partie du manuscrit est un projet d'analyses réalisable à partir de données produite au cours de cette thèse. Par faute de temps, du fait de l'énergie mise dans la production d'ADN de haute qualité, de l'assemblage et l'analyse des génomes produits pour le chapitre 2, ce projet n'a pu qu'être amorcé. J'ai dû aussi faire face à des difficultés de mise en culture de souches qui devaient servir à produire l'ADN pour cette analyse. Ces souches étaient conservées en laboratoire sur milieu PDA (potato dextrose agarose), au froid (-4°c) dans la mycothèque du laboratoire, mais la plupart d'entre elles datant de 2010 n'ont pas repoussé, l'inoculum étant mort ou contaminé par des bactéries. Environ 100 souches avaient été sélectionnées à partir de leur génotype définit par seize marqueurs microsatellites comme étant des hybrides entre les lignées génétiques, ou ayant au moins un allèle rare non retrouvé chez les lignées génétiques. Seulement 29 isolats ont pu être correctement cultivés et sont déjà séquencés.

1) Tester l'hypothèse de réservoir de diversité génétique dans les génotypes minoritaires

Hypothèse : Les génotypes minoritaires constituent un réservoir de diversité génétique permettant l'émergence de nouvelles lignées clonales au cours de l'invasion en Europe.

Les séquences des 29 isolats ont été mappées sur le nouveau génome de référence produit dans le chapitre II, et ont permis de produire un jeu de variants nucléotidiques (SNPs). Le jeu de SNPs produit devra tout d'abord être filtré afin d'obtenir un jeu de marqueur dépourvu de faux positifs. Cette méthodologie est décrite dans le chapitre I, et consiste principalement à exclure les SNPs hétérozygotes détectés dans les régions contenant des éléments transposables. En effet, même si le chapitre 2 présente un nouveau génome de référence de *C. parasitica* assemblé à un niveau proche de la formule chromosomique de l'espèce et qui pourra être utilisé pour ces

analyses, les régions répétées à l'échelle du génome peuvent toujours induire des erreurs de mapping des short-reads, ce qui mène à l'appel de SNPs faux positifs. Le jeu de SNPs filtrés pourra ensuite permettre d'estimer la variabilité génétique de cet échantillon d'hybride. Un réseau généalogique entre les différents génotypes pourra être produit grâce au logiciel SplitTree4, et permettra d'estimer les relations d'apparentement qu'ils partagent avec les lignées clonales principales. Une courbe de raréfaction de la diversité génétique de la lignée RE092 (la plus diverse génétiquement parmi celles étudiées dans le chapitre I) avait été réalisée et présentée en figure S5 de l'annexe 1 de cette thèse. Ma première hypothèse était que la lignée RE092 pouvait contenir une diversité génétique assez élevée pour être à l'origine de l'émergence de nouvelles lignées. Pourtant, la saturation de la courbe de raréfaction montrait que les six échantillons étudiés reflétaient l'essentiel de la diversité génétique de la lignée RE092. Ce résultat a infirmé mon hypothèse. Ma deuxième hypothèse est que les génotypes minoritaires détectés dans les populations seraient un réservoir de diversité génétique. Ce même type d'analyse pourra être utilisé pour estimer la diversité nucléotidique maximale théorique des génotypes minoritaires des populations Françaises. Si cette estimation s'avère être supérieure à la diversité nucléotidique estimée entre les lignées principales, cela signifiera qu'une part de la variation nucléotidique des populations de *C. parasitica* est contenue dans les génotypes minoritaires, et que ces derniers pourraient constituer un réservoir de diversité génétique latent. Notre attendu sur cette hypothèse est donc que certains génotypes minoritaires ne sont pas l'unique produit de l'hybridation entre les lignées principales. Si cette hypothèse est vérifiée, une perspective de travail essentielle sera de proposer des scénarios évolutifs pouvant expliquer le maintient de ces génotypes minoritaires dans les populations.

2) Tester l'hypothèse la plus probable concernant les mécanismes impliqués dans la détection de petites régions génomiques échangées entre les lignées

Hypothèse 1 : Les croisements entre lignées clonales sont possibles mais la faible part d'hybrides dans les populations est due à une contre sélection ou une faible viabilité de la plupart des hybrides.

Hypothèse 2 : Les croisements entre lignées clonales sont possibles mais la présence des lignées en larges taches clonales dans les parcelles crée une asymétrie dans les croisements induisant des rétro-croisements successifs des hybrides avec une seule lignée parentale.

Durant l'hiver 2019, je suis retourné sur le site d'étude initialement choisi pour une analyse géostatistique d'isolats de *C. parastica* dans une parcelle de châtaigniers par Dutech et ses collaborateurs (2008). Leurs analyses avaient permis de montrer que la population de cette parcelle était structurée en taches clonales. J'ai sélectionné une partie de la parcelle représentée par au moins deux taches clonales et présentant un gradient d'admixture entre ces deux taches. La figure 1 montre la localisation des nouveaux isolats que j'ai échantillonnés et les données de seize marqueurs microsatellites qui ont été obtenues récemment. Ces résultats permettront de choisir quels isolats seront séquencés, le but étant de sélectionner des isolats dont le génotype est minoritaire ou hybride entre les deux principales lignées clonales présentes sur cette parcelle. Certains isolats portant le génotype des lignées clonales seront sélectionnés, car il est possible qu'ils présentent des traces d'introgression réduites sur l'ensemble du génome comme cela a été montré dans le chapitre I. Ces derniers sont donc essentiels dans cette analyse car ils présentent un profile d'introgression très avancé.

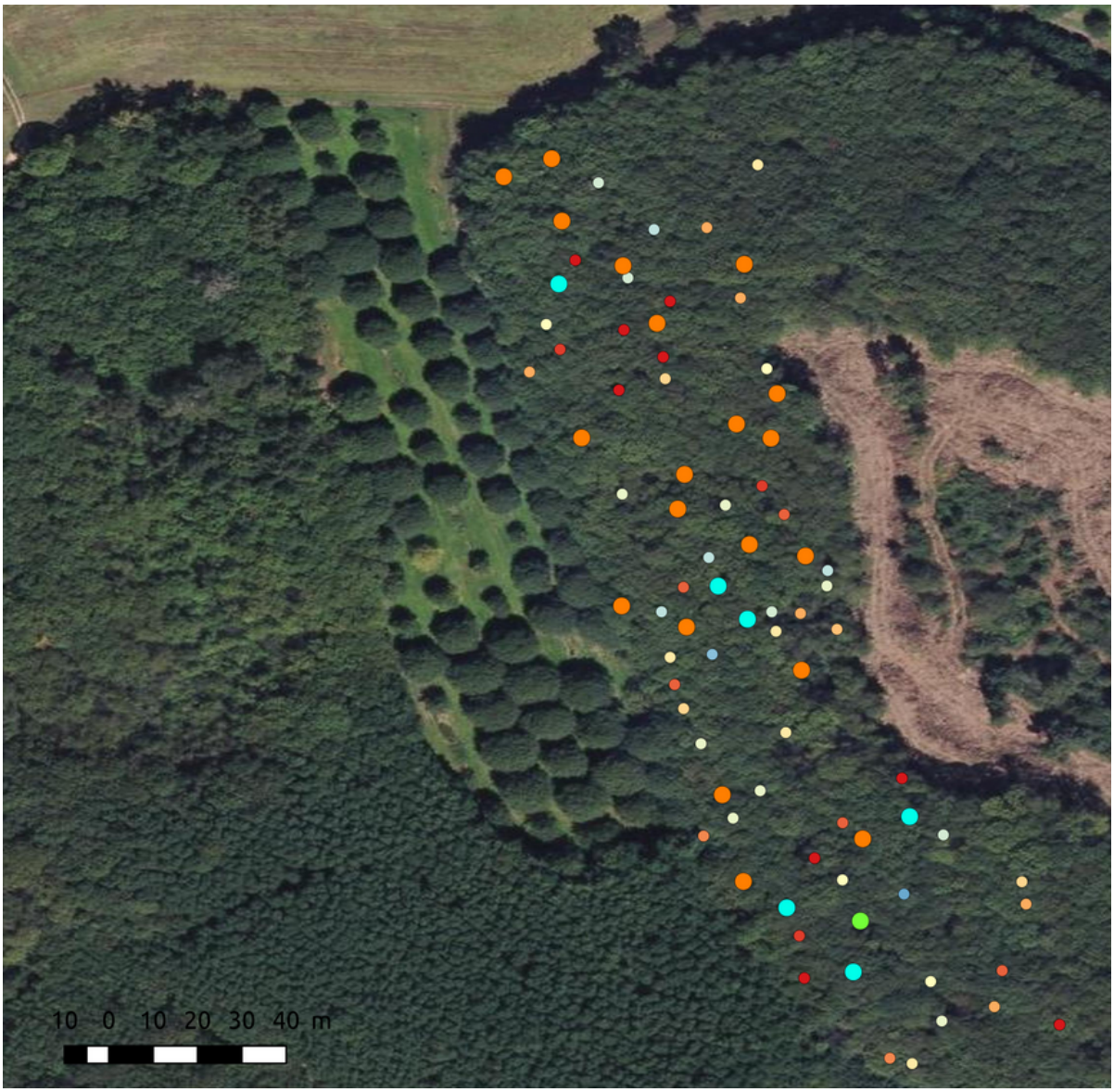


Figure 1 : Photo aérienne de la parcelle d'étude. Les points représentent les localisations des échantillons de chancre. Les couleurs représentent le génotype des isolats obtenu sur la base de seize marqueurs microsatellites. Les disques de grande taille représentent les génotypes apparentés aux lignées principales, tandis que les autres génotypes sont représentés par les disques de petite taille. La lignée RE079 est colorée en orange, la lignée RE019 est colorée en bleu turquoise et la lignée RE043 est colorée en vert. La couleur des disques de petite taille représente la proximité des génotypes à la lignée RE079 : de rouge (> 90 % des allèles identiques) à bleu foncé (0 % des allèles identiques).

Si certaines combinaisons d'allèles sont nécessaires à la viabilité des hybrides, les locus portant ces allèles seront donc retrouvés en fort déséquilibre de liaison. Pour mesurer le déséquilibre de liaison (DL), les SNPs détectés sur les scaffolds 2, 3, 4 et 6 du génome assemblé de l'isolat ESM015

pourront être utilisés pour obtenir un profile d'introgression à l'échelle chromosomique car, comme mentionné dans le chapitre 2 de cette thèse, ces scaffolds sont supposés être des chromosomes entiers. En effet, des signatures de séquences télomériques ont été détectées aux deux extrémités de chacun de ces scaffolds, et l'assemblage entier a été vérifié comme ne présentant aucun scaffold chimérique. L'utilisation de ces quatre scaffolds permettra d'établir un profile d'introgression et d'estimer un taux de transmission des SNPs des génotypes parentaux par chromosome. L'ensemble des isolats séquencés sera utilisé (les isolats récemment échantillonnés et les 29 génotypes minoritaires sélectionnés). Pour compléter ces profiles, l'ensemble des SNPs détecté le long du génome sera utilisé pour obtenir un profile d'introgression complet. Le déséquilibre de liaison entre ces SNPs sera estimé deux à deux (pairwise linkage disequilibrium) à l'aide des VCFtools (Danecek et al., 2011). Les valeurs de DL seront mesurées à partir du coefficient de corrélation au carré (r^2) estimé entre chaque paire de SNPs dans les populations considérées. Les taux de DL et la diminution de la valeur de DL seront tracées par scaffold en fonction de la distance nucléotidique en paire de bases à l'aide du package R LDheatmap (Shin *et al.*, 2006). Si certaines combinaisons d'allèles sont nécessaires à la viabilité des hybrides, les locus portant ces allèles devraient être retrouvés en fort déséquilibre de liaison. Il faut prendre en compte que la clonalité peut induire des patrons similaires de déséquilibre de liaison, et qu'il est donc difficile de détecter des signatures de sélections dans des populations largement clonales. Les profiles de déséquilibre de liaison pourraient donc être difficilement interprétables. Cependant, il sera possible de détecter des signaux forts de déséquilibre de liaisons entre des locus présents sur des chromosomes différents, qui pourraient s'ajouter au bruit de fond induit par la clonalité. Ces locus pourraient être des candidats de régions génomiques sous sélection dans les populations s'ils sont retrouvés en forts déséquilibre de liaison parmi les hybrides.

La mesure du déséquilibre de liaison dans les génotypes hybrides pourrait être difficile à interpréter, mais l'analyse des isolats de la parcelle présentée en figure 1 va surtout permettre de mieux identifier la fraction hybride dans cette population. La fraction hybride représente la proportion du

génome d’un hybride issue d’une lignée clonale. Chez les hybrides de première génération (F1) entre deux lignées clonales, on s’attend à retrouver 50 % du génome issue de chaque lignée parentale. Je m’attend à ce que les hybrides présentent une importante fraction génomique issue du génotype RE079, qui est la lignée clonale dominante dans la parcelle échantillonnée. Afin de déterminer la distribution de la fraction hybride dans cette population, je propose de calculer les paramètres de cette distribution et de les comparer aux attendus d’un modèle d’admixture entre des populations proposé par Verdu et Rosenberg (2011). Ce modèle a été développé pour estimer la probabilité de la fraction d’admixture d’un individu d’une population hybride issue du mélange de deux populations sources et généralisable à n populations. Les deux paramètres pris en compte dans ce modèle sont 1) La contribution initiale de chacune des populations sources à la population hybride à la génération 0 (g0), donnant une population hybride de première génération (F1/g1) et 2) La contribution de chacune des populations sources au fil des générations (de g2 à gn). Les attendus de probabilité de fraction d’admixture d’un individu de la population hybride à la génération 6 en fonction de différents paramètres initiaux sont présentés en figure 2. Pour adapter ce modèle à mon système d’étude, les lignées clonales seront considérées comme populations sources. A partir du jeu de SNPs obtenu à partir du séquençage de 96 isolats de la parcelle étudiée, je propose d’estimer la fraction d’admixture de chaque hybride par rapport au clone majoritaire RE079 de deux façons : 1) La proportion de SNPs partagés avec la lignée RE079 par rapport au nombre total de SNPs détectés par hybride, et 2) la proportion de fenêtres de 10kb partageant un haplotype identique à la lignée RE079 par rapport au nombre total de fenêtres de 10kb considérées. La première méthode s’avère plus facile à mettre en place, mais peut être biaisé par une forte concentration de SNPs dans certaines régions du génome. En effet, un même nombre de SNPs peut être détecté dans des régions de tailles variables, ce qui peut conduire à une mauvaise estimation de la fraction du génome partagée avec la lignée clonale. La deuxième méthode permet de s’affranchir de ce biais en calculant directement la portion de génome partagée. La distribution des fractions d’admixture de tous les hybrides permettra de prédire si les croisements dans cette parcelle sont

effectivement asymétriques, comme proposé dans l’hypothèse 2. Dans le cas d’une asymétrie de flux de gènes, due à des rétro-croisements successifs avec une lignée clonale, on s’attend à obtenir une distribution déséquilibrée vers une des lignées (scénarios C, D et E ; figure 2). D’autre part, la variance de la distribution peut donner une indication sur le nombre de générations (Verdu et Rosenberg, 2011). Si le flux de gènes est ancien et récurrent, la variance du signal d’admixture devrait être faible et la distribution régulière. A l’inverse, si le flux de gène est rare, peu de catégories de fraction d’admixture sont représentées. La variance du signal d’admixture est élevée et la distribution n’est pas homogène. Enfin, les fragments introgressés au fur et à mesure des générations de rétro-croisements pourraient ne pas être distribués aléatoirement le long du génome. Certaines régions pourraient être introgressées préférentiellement, ce qui supporterait l’hypothèse 1 proposant que certaines combinaisons alléliques sont bénéfiques.

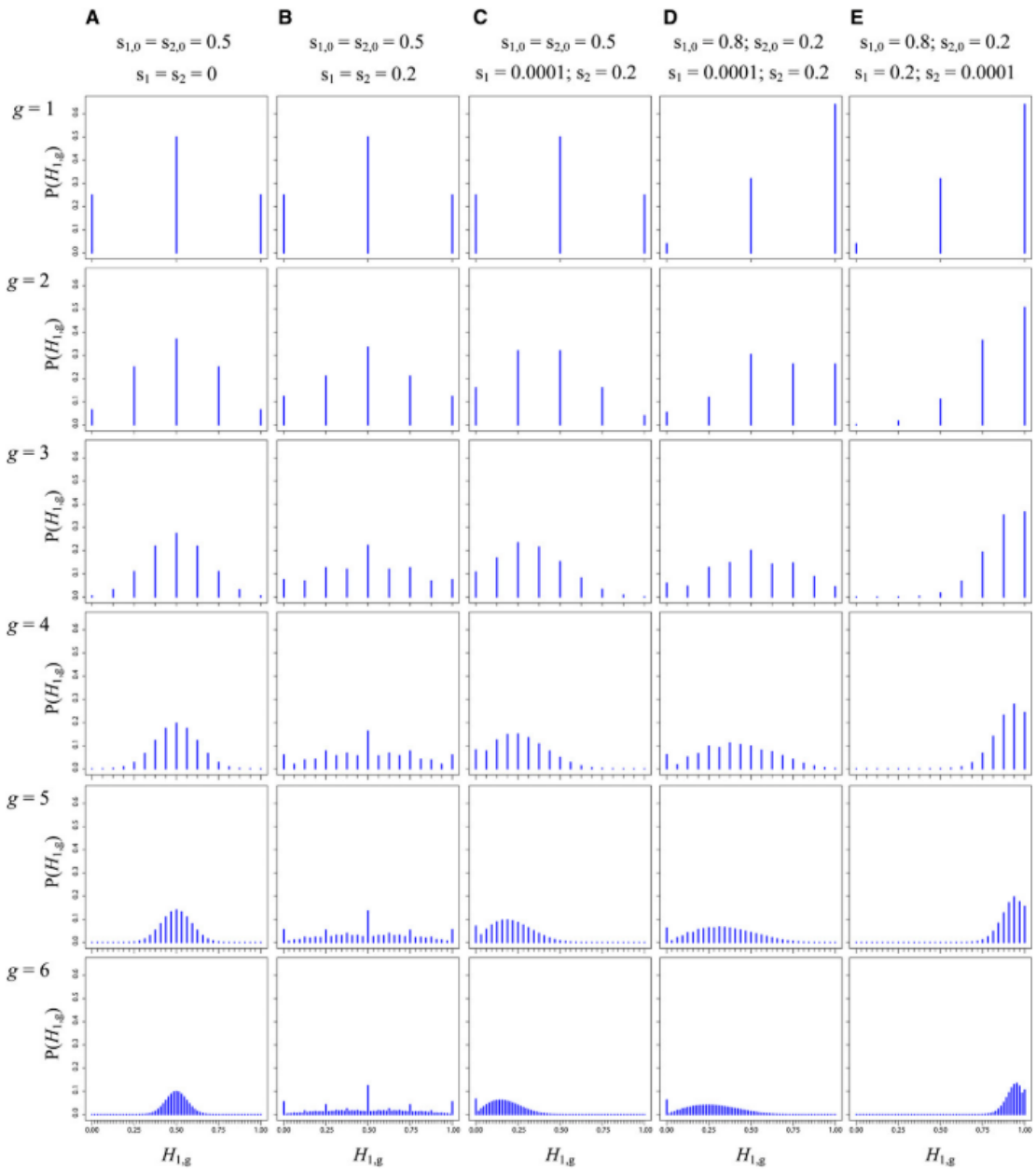


Figure 2 tirée de Verdu & Rosenberg, 2011 : Distribution des probabilités de fraction d'admixture de la population source S1 pour un individu sélectionné aléatoirement dans la population hybride (H) entre les populations sources S1 et S2 en fonction de différents nombre de générations. A) La population H est fondée à la génération 0 (g0) avec une égale contribution des deux populations sources. Aux générations suivantes, les populations sources ne contribuent plus à H. B) État initial (g0) identique à A mais les deux populations sources continuent de contribuer faiblement à H. C) État initial (g0) identique à A mais la population S2 continue de contribuer à H beaucoup plus fortement que S1. D) Contribution au fil des génération de S1 et S2 identique à C mais la population S1 a initialement (g0) contribué à 80 % à H. E) État initial (g0) identique à D mais la contribution de S1 et de S2 au fil des générations est inversée, la population S1 contribuant beaucoup plus fortement à H que S2.

Bibliographie

Agosta S. J. & Klemens J. A. (2008). Ecological fitting by phenotypically flexible genotypes: implications for species associations community assembly and evolution. Ecology Letters. 11, 1123-1134.

Almagro Armenteros J. J., Tsirigos K. D. Sonderby C. K., Petersen T. N., Winther O., Brunak S., Heijne G V & Nielsen H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. Nature biotechnology. 37(4). 420

Alonso A., Dallmeier, F., Granek, E., & Raven, P. (2001).Biodiversity: Connecting with the Tapestry of Life. Smithsonian Institution/Monitoring and Assessment of Biodiversity Program and President's Committee of Advisors on Science and Technology. Washington, D.C., U.S.A..

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.

Amselem J., Lebrun M-H. & Quesneville H. (2015). Whole genome comparative analysis of transposable elements provides new insight into mechanisms of their inactivation in funal genomes. BMC Genomics. 16:161.

Anagnostakis S. L. (1987). Chestnut Blight: The Classical Problem of an Introduced Pathogen. Mycologia. 79(1), 23-37

Anderson J. E., Kantar, M. B., Kono, T. Y., Fu, F., Stec, A. O., Song, Q.,Stupar, R. M. (2014). A roadmap for functional structural variants in the soybean genome. G3: Genes, Genomes, Genetics, 4(7), 1307–1318.

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 and Evolution, 19(10), 535–544.

Ankenbrand M. J., Hohlfeld S., Hackl T. & Forster F. (2017). AliTV-interactive visualization of whole genome comparisons. PeerJ Preprints 5:e2348v2.

Antipov D., Korobeynikov A., McLean J. S. & Pevzner P. A. (2015). HYBRIDSPADES: an algorithm for hybrid assembly of short and long reads. Bioinformatics. 32(7):btv688.

Anxolabéhère D., Kidwell M. G. & Periquet G. (1988). Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of Drosophila melanogaster by mobile P elements. 5(3); 252-259.

Aris-Brosou S., and Excoffier, L., 1996. The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. Molecular Bioliology and Evolution 13, 494–504.

Atkinson P. W. (2015). hAT Transposable Elements. Microbiology Spectrum. 3(4):MDNA3-0054- 2014

Badouin H., Hood, M. E., Gouzy, J., Aguileta, G., Siguenza, S., Perlin, M. H., Giraudtatiana, T. (2015). Chaos of rearrangements in the mating-type chromosomes of the anther-smut fungus Microbotryum lychnidis-dioicae. Genetics, 200(4), 1275–1284.

Bao J., Chen, M., Zhong, Z., Tang, W., Lin, L., Zhang, X., Wang, Z. (2017). PacBio Sequencing Reveals Transposable Elements as a Key Contributor to Genomic Plasticity and Virulence Variation in Magnaporthe oryzae. Molecular Plant, 10(11), 1465–1468.

Barrett S. C. H., Colautti, R. I., & Eckert, C. G. (2008). Plant reproductive systems and evolution during biological invasion. Molecular Ecology, 17(1), 373–383.

Baskaev K. K. & Buzdin A. A. (2012). Evolutionary recent insertions of mobile elements and their contribution to the structure of human genome. Zh Obshch Biol.73(1), 3-20.

Basso D. L. (2002). Viruses and transposons of the chestnut blight fungus, Cryphonectria parasitica: tools to examine virulence, fungal diversity, and evolution. PhD degree in AGRICULTURE, PLANT PATHOLOGY. Adviser : Hillman B. I. Rutgers The State University of New Jersey - New Brunswick.

Basso L. D., Foglia R., Zhu P. & Hillman B. I. (2001). Crypt1, an active Ac-like transposon from the chestnut blight fungus, Cryphonectria parasitica. Mol Genet Genomics. 265: 730-738.

Bazin É., Mathé-Hubert, H., Facon, B., Carlier, J., & Ravigné, V. (2014). The effect of mating system on invasiveness: Some genetic load may be advantageous when invading new environments. Biological Invasions, 16(4), 875–886.

Biemont C. A brief history of the status of transposable elements: from junk DNA to major players in evolution. Genetics. 186, 1085–93

Biraghi A (1946). Il cancro del castagno causato da Endothia parasitica. Ital Agric 7: 1–9.

Birch P., R., & Whisson S., C.(2001). Pathogen profile Phytophthora infestans enters the genomics era. (2001). 2, 257–263.

Bissegger M., & Heiniger, U. (1991). Chestnut blight (Cryphonectria parasitica) north of the Swiss alps. European Journal of Forest Pathology, 21(4), 250–252.

Blackburn T. M., Pyšek, P., Bacher, S., Carlton, J. T., Duncan, R. P., Jarošík, V., Richardson, D. M. (2011). A proposed unified framework for biological invasions. Trends in Ecology and Evolution, 26(7), 333–339.

Blin K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S. Y.,Weber, T. (2019). antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Research, 47(W1), W81–W87.

Boland G. J. & Hall R. (1994). Index of plants of of Sclerotinia sclerotiorum. Canadian Journal of plant pathology. 16, 93-108.

Bolnick D. I., & Nosil, P. (2007). Natural selection in populations subject to a migration load. Evolution, 61(9), 2229–2243.

Boutin T. S., Le rouzic A. & Capy P. (2012).How does selfing affect the dynamics of selfish transposable elements ? Mobile DNA 3, 5.

Bragança H., Simões, S., Onofre, N., Tenreiro, R., Rigling, D., 2007. Cryphonectria parasitica in Portugal: diversity of vegetative compatibility types, mating types, and occurrence of hypovirulence. Forest Pathology 37, 391–402.

Brakhage A. A. (2013). Regulation of fungal secondary metabolism. Nature Reviews Microbiology, 11(1), 21–32.

Brasier C. M. (1986). The population biology of Dutch elm disease: its principle features and some implications for other host-pathogen systems. Advances in Plant Pathology, 5. 53-118

Brasier C. M., & Buck, K. W. (2001). Rapid evolutionary changes in a globally invading fungal pathogen (Dutch elm disease). Biological Invasions, 3(3), 223–233.

Brasier C. M., & Kirk, S. A. (2010). Rapid emergence of hybrids between the two subspecies of Ophiostoma novo-ulmi with a high level of pathogenic fitness. Plant Pathology, 59(1), 186– 199.

Brasier C. M., Kirk, S. A., Pipe, N. D., & Buck, K. W. (1998). Rare interspecific hybrids in natural populations of the Dutch elm disease pathogens Ophiostoma ulmi and O. novo-ulmi. Mycological Research, 102(1), 45–57.

Brown J.D., O'Neill, R.J., 2010. Chromosomes, Conflict, and Epigenetics: Chromosomal Speciation Revisited. Annual Review of Genomics and Human Genetics 11, 291–316.

Brufor M.W., Wayne, R.K., 1993. Microsatellites and their application to population genetic studies. Current Opinion in Genetics & Development 3, 939–943.

Bryant D., Moulton, V., 2003. Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. Molecular Biology and Evolution 21, 255–265.

Burt A., 2000. Perspective: sex, recombination, and the efficacy of selection - Was Waismann right? Evolution 54, 337–351.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. (2009) BLAST+: architecture and applications. BMC bioinformatics 10: 421.

Cambareri E. B., B. C. Jensen, E. Schabtach and E. U. Selker, 1989 Repeat-induced G-C to A-T mutations in Neurospora. Science. 244; 1571–1575

Cao M. D., Nguyen S. H., Ganesamoorthy D., Elliott A. G., Cooper M. A. & Coin L. J. M. (2017). Scaffolding and completing genome assemblies in real-time with nanopore sequencing. Nature communications 8. 14515.

Casas-Vila N., Scheibe M., Freiwald A., Kappei D. & Butter F. (2015). Identification of TTAGGG- binding proteins in Neurospora crassa, a fungus with vertebrate-like telomere repeats. BMC Genomics. 16, 965.

Charlesworth D., & Wright, S. I. (2001). Breeding systems and genome evolution. Current Opinion in Genetics and Development, 11(6), 685–690.

Charlesworth D., Willis J. H. (2009). Fundamental concepts in genetics: the genetics of inbreeding depression. Nature Reviews Genetics. 10, 783–796.

Cheeseman K., Ropars, J., Renault, P., Dupont, J., Gouzy, J., Branca, A., Abraham, A.-L., Ceppi, M., Conseiller, E., Debuchy, R., Malagnac, F., Goarin, A., Silar, P., Lacoste, S., Sallet, E., Bensimon, A., Giraud, T., Brygoo, Y., 2014. Multiple recent horizontal transfers of a large genomic region in cheese making fungi. Nature Communications 5.

Cheng L., Connor, T.R., Siren, J., Aanensen, D.M., Corander, J., 2013. Hierarchical and Spatially Explicit Clustering of DNA Sequences with BAPS Software. Molecular Biology and Evolution 30, 1224–1228.

Chiapello H., Mallet, L., Guérin, C., Aguileta, G., Amselem, J., Kroj, T., Fournier, E. (2015). Deciphering genome content and evolutionary relationships of isolates from the fungus Magnaporthe oryzae attacking different host plants. Genome Biology and Evolution, 7(10), 2896–2912.

Chuma I., Isobe, C., Hotta, Y., Ibaragi, K., Futamata, N., Kusaba, M., Yoshida, K., Terauchi, R., Fujita, Y., Nakayashiki, H., Valent, B., Tosa, Y., 2011. Multiple Translocation of the AVR-Pita Effector Gene among Chromosomes of the Rice Blast Fungus Magnaporthe oryzae and Related Species. PLoS Pathogens 7, e1002147–e1002147.

Ciancio J. E., Rossi, C. R., Pascual, M., Anderson, E., & Garza, J. C. (2015). The invasion of an Atlantic Ocean river basin in Patagonia by Chinook salmon: new insights from SNPs. Biological Invasions, 17(10), 2989–2998.

Clemento A. J., Crandall, E. D., Garza, J. C., & Anderson, E. C. (2014). Evaluation of a single nucleotide polymorphism baseline for genetic stock identification of Chinook salmon (Oncorhynchus tshawytscha) in the California current large marine ecosystem. Fishery Bulletin, 112(2–3), 112–130.

Collemare J. & Seidl M. F. (2019) Chromatin-dependent regulation of secondary metabolite biosynthesis in fungi: is the picture complete? FEMS Microbiology Reviews. Fuz018

Corander J., Marttinen, P., Sirén, J., Tang, J., 2008. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics 9, 539–539.

Cornell M. J., Alam, I., Soanes, D. M., Han, M. W., Hedeler, C., Paton, N. W., Oliver, S. G. (2007). Comparative genome analysis across a kingdom of eukaryotic organisms: Specialization and diversification in the Fungi. Genome Research, 17(12), 1809–1822.

Couch B. C., Fudal, I., Lebrun, M. H., Tharreau, D., Valent, B., Van Kim, P., Kohn, L. M. (2005). Origins of host-specific populations of the blast pathogen Magnaporthe oryzae in crop domestication with subsequent expansion of pandemic clones on rice and weeds of rice. Genetics, 170(2), 613–630.

Cross Arteil et al., in prep

Crow J.F. (1993) How much do we know about spontaneous human mutation rates? Environ. Mol. Mutagen. 21, 122–129

Dalman K., Himmelstrand, K., Olson, Å., Lind, M., Brandström-Durling, M., & Stenlid, J. (2013). A Genome-Wide Association Study Identifies Genomic Regions for Virulence in the Non- Model Organism Heterobasidion annosum s.s. PLoS ONE, 8(1).

Danecek P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., 1000 Genomes Project Analysis Group, 1000 Genomes Project Analysis, 2011. The variant call format and VCFtools. Bioinformatics (Oxford, England) 27, 2156–8.

Darling A.C.E., Mau, B., Blattner, F.R., Perna, N.T., 2004. Mauve : Multiple Alignment of Conserved Genomic Sequence With Rearrangements Mauve : Multiple Alignment of Conserved Genomic Sequence With Rearrangements 1394–1403.

Darpoux H, 1949. Le chancre du châtaignier causé par l'Endothia parasitica. Document Phytosanitaire No 7. Paris, France: Ministère de l'Agriculture
Darriba D., Taboada, G.L., Doallo, R., Posada, D., 2012. jModelTest 2: more models, new heuristics and parallel computing. Nature Methods 9, 772–772.

Daverdin G., T. Rouxel, L. Gout, J.-N. Aubertot, I. Fudal et al. (2012) Genome structure and reproductive behaviour influence the evolutionary potential of a fungal phytopathogen. PLoS Pathog. 8: e1003020.

Davey J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics, 12(7), 499–510.

Dawkin R. The selfish gene. Oxford University Press. 1976. 224
Day P.R., 1977. Double-stranded RNA in Endothia parasitica. Phytopathology 77, 1393–1393.

Dayani, Y., Simchen, G., & Lichten, M. (2011). Meiotic Recombination Intermediates Are Resolved with Minimal Crossover Formation during Return-to-Growth , an Analogue of the Mitotic Cell Cycle, *7*(5).

de Jonge R., M. D. Bolton, A. Kombrink, G. C. M. van den Berg, K. A. Yadeta et al. (2013). Extensive chromosomal reshuffling drives evolution of virulence in an asexual pathogen. Genome Res. 23: 1271–1282.

de Vienne D. M., Hood M. E. & Giraud T. (2009). Phylogenetic determinants of potential host shifts in fungal pathogens. J. Evol. Biol. 22, 2532-2541.

Delcher A. L., Phillipy A., Carlton J. Salzberg S. L. (2002). Fast algorithms for large-scale genome alignment and comparison, Nucleic Acids Research. 30(11), 2478–2483

Demené A., Legrand L. Gouzy J. Debuchy R. Saint-Jean Gilles, Fabreguette O. & Dutech C. (2019). Whole-genome sequencing reveals recent and frequent genetic recombination between clonal lieanges of Cryophonectria parasitica in western Europe. Fungal Genet Biol. 130, 122-133.

Denver D.R., Morris, K., Lynch, M., Thomas, W.K., 2004. High mutation rate and predominance of insertions in the Caenorhabditis elegans nuclear genome. Nature 430, 679–682.

Desprez-Loustau M. L., Robin, C., Buée, M., Courtecuisse, R., Garbaye, J., Suffert, F., Rizzo, D. M. (2007). The fungal dimension of biological invasions. Trends in Ecology and Evolution, 22(9), 472–480.

Didelot X., Wilson, D.J., 2015. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. PLOS Computational Biology 11, e1004041–e1004041.
Diwash J., Feschotte C. & Betran E. (2017). Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. Trends in Genetics. 33(11). 817-831.

Dlugosch K. M., & Parker, I. M. (2008). Founding events in species invasions: Genetic variation, adaptive evolution, and the role of multiple introductions. Molecular Ecology, 17(1), 431– 449.

Dong S, Raffaele S, Kamoun S. (2015). The two-speed genomes of filamentous pathogens: waltz with plants. Curr Opin Genet Dev 35: 57–65.

Doolittle W., Sapienza, C. (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature 284, 601–603

Drummond A.J., Rambaut, A., (2007). BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology 7, 214–214.

Drummond A.J., Suchard, M.A., Xie, D., Rambaut, A. (2012). Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution 29, 1969–1973.

Duncan J. (1999). Phytophthora an Abiding Threat To Our Crops. Microbiology Today. 26, 114–116.

Dunham M. J., Badrane, H., Ferea, T., Adams, J., Brown, P. O., Rosenzweig, F., & Botstein, D. (2002). Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences of the United States of America, 99(25), 16144–16149.

Dutech C, Barres B, Bridier J et al. (2012) The chestnut blight fungus world tour: successive introduction events from diverse origins in an invasive plant fungal pathogen. Molecular Ecology., 21, 3931–3946

Dutech C., Fabreguettes O., Capdevielle X. & Robin C. (2010). Multiple introductions of divergent genetic lineages in an invasive fungal pathogen, Cryphonectria parasitica, in France. Heredity. 105, 220–228

Dutech C., Fabreguettes, O., Capdevielle, X., & Robin, C. (2010). Multiple introductions of divergent genetic lineages in an invasive fungal pathogen, Cryphonectria parasitica, in France. Heredity, 105(2), 220–228.

Dutech C., Rossi, J. P., Fabreguettes, O., & Robin, C. (2008). Geostatistical genetic analysis for inferring the dispersal pattern of a partially clonal species: Example of the chestnut blight fungus. Molecular Ecology, 17(21), 4597–4607.

Dutech, C., Barrès, B., Bridier, J., Robin, C., Milgroom, M.G., Ravigné, V., 2012. The chestnut blight fungus world tour: Successive introduction events from diverse origins in an invasive plant fungal pathogen. Molecular Ecology 21, 3931–3946.

Eckert C. G., Manicacci, D., & Barrett, S. C. H. (1996). Genetic Drift and Founder Effect in Native Versus Introduced Populations of an Invading Plant, Lythrum salicaria (Lythraceae). Evolution, 50(4), 1512.

Edmands S. (1999). Heterosis and Outbreeding Depression in Interpopulation Crosses Spanning a Wide Range of Divergence. Evolution, 53(6), 1757.

Eldredge N. & Gould, S. J. (1972). Punctuated equilibria: an alternative to phyletic gradualism.Schopf, T.J.M. (ed.), Models in Paleobiology. W.H. Freeman, New York. Methods in Paleobiology, pp. 82–115.

Ellegren H. (2000). Microsatellite mutations in the germline: Implications for evolutionary inference. Trends in Genetics, 16(12), 551–558.

Ellis E. C. (2015). Ecology in an anthropogenic biosphere. Ecological Monographs, 85(3), 287–331.

Emms D.M. & Kelly S. (2018). OrthoFinder2: fast and accurate phylogenomic orthology analysis from gene sequences. bioRxiv.

Estoup A., Ravigne, V., Hufbauer, R., Vitalis, R., Gautier, M., & Facon, B. (2016). Is there a Genetic Paradoc of Biological Invasion. Annual Review of Ecology, Evolution, and Systematics, 47, 51–72.

Estoup A., Wilson, I. J., Sullivan, C., Cornuet, J. M., & Moritz, C. (2001). Inferring population history from microsatellite and enzyme data in serially introduced cane toads, Bufo marinus. Genetics, 159(4), 1671–1687.

Eusebio-Cope A., Suzuki, N., Sadeghi-Garmaroodi, H., & Taga, M. (2009). Cytological and electrophoretic karyotyping of the chestnut blight fungus Cryphonectria parasitica. Fungal Genetics and Biology, 46(4), 342–351.

Excoffier L., 2004. Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. Molecular Ecology 13, 853–864.

Facon B., Genton, B. J., Shykoff, J., Jarne, P., Estoup, A., & David, P. (2006). A general eco- evolutionary framework for understanding bioinvasions. Trends in Ecology and Evolution, 21(3), 130–135.

Faino L., Seidl M. F., Datema E., van den Berg G. C., Janssen A., Wittenberg A. H., Thomma B. P. (2015). Single-Molecule Real-Time sequencing combined with optical mapping yields completely finished fungal genomes. MBio 6: pe00936–15.

Faino L., Seidl, M. F., Shi-Kunne, X., Pauper, M., van den, G. C., Wittenberg, A. H., Bart PHJ Thomma, D. (2016). Transposons passively and actively contribute to evolution of the two- speed genome 1 of a fungal pathogen The Netherlands Running title: Genome evolution by transposable elements. 1091–1100.

Feau N., Dutech, C., Brusini, J., Rigling, D., Robin, C., 2014. Multiple introductions and recombination in Cryphonectria hypovirus 1: Perspective for a sustainable biological control of chestnut blight. Evolutionary Applications 7, 580–596.

Feehan J.M., Katherine E.S., Salim B., William U., Beat K., Shauna C.S., 2017. Purification of High Molecular Weight Genomic DNA from Powdery Mildew for Long-Read Sequencing. *J. Vis. Exp.* (121), e55463

Feschotte C., & Pritham, E. J. (2007). DNA Transposons and the Evolution of Eukaryotic Genomes. Annual Review of Genetics, 41(1), 331–368.

Feschotte C., 2008. Transposable elements and the evolution of regulatory networks. NATURE REVIEWS | GENETICS 9, 397–405.

Finnegan D. J. (1989). Eukaryotic Transposable elements and genome evolution. Trends Genet. 5(4), 103-107.

Finnegan D. J. (1997). Transposable elements: How non-LTR retrotransposons do it. Current Biology, 7(4), 245–248.

Fisher R. A. (1930) The Genetical Theory of Natural Selection. Oxford Univ. Press, Oxford

Flutre T., Duprat E., Feuillet C., Quesneville H. (2011). Considering transposable element diversification in de novo annotation approaches. PLoS One. 6:e16526.

Forgetta V., Leveque, G., Dias, J., Grove, D., Lyons, R., Genik, S., Dewar, K. (2013). Sequencing of the Dutch elm disease fungus genome using the Roche/454 GS-FLX titanium system in a comparison of multiple genomics core facilities. Journal of Biomolecular Techniques, 24(1), 39–49.

Fournier E., & Giraud, T. (2008). Sympatric genetic differentiation of a generalist pathogenic fungus, Botrytis cinerea, on two different host plants, grapevine and bramble. Journal of Evolutionary Biology, 21(1), 122–132.

Fradin E.F., Thomma B. P. 2006. Physiology and molecular aspects of Verticillium wilt diseases caused by V. dahliae and V. albo-atrum. Mol Plant Pathol. 7, 71–86.

Frankham R. (2005). Genetics and extinction. Biological Conservation, 126(2), 131–140.

Frantzeskakis L., Kracher, B., Kusch, S., Yoshikawa-Maekawa, M., Bauer, S., Pedersen, C.,Panstruga, R. (2018). Signatures of host specialization and a recent transposable element burst in the dynamic one-speed genome of the fungal barley powdery mildew pathogen. BMC Genomics, 19(1), 1–23.

Galagan J. E., & Selker, E. U. (2004). RIP: The evolutionary cost of genome defense. Trends in Genetics, 20(9), 417–423.

Galazka J. M., & Freitag, M. (2014). Variability of chromosome structure in pathogenic fungi-of "ends and odds." Current Opinion in Microbiology, 20, 19–26.

Garbelotto M., & Gonthier, P. (2013). Biology, Epidemiology, and Control of Heterobasidion Species Worldwide . Annual Review of Phytopathology, 51(1), 39–59.

Gavrilets S. (2004). Fitness landscapes and the origin of species. Princeton University Press

George Allen & Uwin., Irelan J. T., & Selker, E. U. (1996). Gene silencing in filamentous fungi: RIP, MIP and quelling. Journal of Genetics, 75(3), 313–324.

Giordano L., Gonthier, P., Lione, G., Capretti, P., & Garbelotto, M. (2014). The saprobic and fruiting abilities of the exotic forest pathogen Heterobasidion irregulare may explain its invasiveness. Biological Invasions, 16(4), 803–814.

Giraud T., Gladieux P., Gavrilets S. (2010). Linking emergence of fungal plant diseases and ecological speciation. Trends Ecol Evol. 25, 387–395

Giraud T., Refrégier, G., Le Gac, M., de Vienne, D. M., & Hood, M. E. (2008). Speciation in fungi. Fungal Genetics and Biology, 45(6), 791–802.

Gladieux P., Condon B., Ravel S., Soanes D., Maciel J. L. N., Nhani A., Jr, Chen L., Terauchi R., Lebrun M-H., Tharreau D., Mitchell T., Pedley K.F., Valent B., Talbot N. J., Farman M., Fournier E. (2018). Gene flow between divergent

cereal- and grass-specific lineages of the rice blast fungus Magnaporthe oryzae. mBio 9:e01219-17.

Gladieux P., Feurtey A., Hood M. E., Snirc A., Clavel J., Dutech C., Roy M. & Giraud T. (2015). The population biology of fungal invasions. Molecular Ecology. 24, 1969-1986.

Gladieux P., Ravel, S., Rieux, A., Cros-Arteil, S., Adreit, H., 2017. Coexistence ofmultiple endemic and pandemic lineages of the rice blast pathogen. MBio9, e01806-17.

Gladieux P., Ropars, J., Badouin, H., Branca, A., Aguileta, G., De Vienne, D. M., Giraud, T. (2014). Fungal evolutionary genomics provides insight into the mechanisms of adaptive divergence in eukaryotes. Molecular Ecology, 23(4), 753–773.

Gladieux P., Wilson, B.A., Perraudeau, F., Montoya, L.A., Kowbel, D., Hann-Soden, C., Fischer, M., Sylvain, I., Jacobson, D.J., Taylor, J.W., 2015. Genomic sequencing reveals historical, demographic and selective factors associated with the diversification of the fire-associated fungus Neurospora discreta. Molecular Ecology 24, 5657–5675.

Gonthier P., Warner, R., Nicolotti, G., Mazzaglia, A., & Garbelotto, M. (2004). PATHOGEN INTRODUCTION AS A COLLATERAL EFFECT OF MILITARY ACTIVITY. Mycological Research, 108(5), 468-470.

Goss E. M., Tabima, J. F., Cooke, D. E. L., Restrepo, S., Frye, W. E., Forbes, G. A., Grünwald, N. J. (2014). The Irish potato famine pathogen Phytophthora infestans originated in central Mexico rather than the Andes. Proceedings of the National Academy of Sciences of the United States of America, 111(24), 8791–8796.

Grandaubert J., Bhattacharyya, A., & Stukenbrock, E. H. (2015). RNA-seq-Based gene annotation and comparative genomics of four fungal grass pathogens in the genus Zymoseptoria identify novel orphan genes and species-specific invasions of transposable elements. G3: Genes, Genomes, Genetics, 5(7), 1323–1333.

Gremme G., S. Steinbiss, and S. Kurtz, 2013 GenomeTools: a comprehensive software library for efficient processing of struc- tured genome annotations. IEEE/ACM Trans. Comput. Biol. Bi- oinform. 10: 645–656.

Haldane J. B. S., A mathematical theory of natural and artificial selection, 1924 – 1934

Hartmann F. E., Sánchez-Vallet, A., McDonald, B. A., & Croll, D. (2017). A fungal wheat pathogen evolved host specialization by extensive chromosomal rearrangements. ISME Journal, 11(5), 1189–1204.

Heard, S., Brown, N. A., & Hammond-kosack, K. (2015). An Interspecies Comparative Analysis of the Predicted Secretomes of the Necrotrophic Plant Pathogens Sclerotinia sclerotiorum and Botrytis cinerea, 1–27.

Heller D. & Vingron M. (2019). SVIM: structural variant identification using mapped long reads, Bioinformatics. 35(17), 2907–2915

Helyar S. J., Hemmer-Hansen J., Bekkevold D., Taylor M. I., Ogden R., Limborg M. T., Cariani A., Maes G. E., Diopere E., Carvalho G. R. & Nielsen E. E. (2011). Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. Molecular Ecology Resources (2011) 11 (Suppl. 1), 123–136

Henk D.A., Shahar-Golan R., Devi, K.R., Boyce, K.J., Zhan, N., Fedorova, N.D., Nierman, W.C., Hsueh, P.R., Yuen, K.Y., Sieu, T.P.M., Van Kinh, N., Wertheim, H., Baker, S.G., Day, J.N., Vanittanakom, N., Bignell, E.M., Andrianopoulos, A., Fisher, M.C., 2012. Clonality despite sex: the evolution of host-associated sexual neighborhoods in the pathogenic fungus Penicillium marneffei. PLoS Pathogens 8 (10), e1002851.

Hoede C., Arnoux S., Moisset M., Chaumier T., Inizan O., Jamilloux V., et al. (2014). PASTEC: an automatic transposable element classification tool. PLoS One. 9:e91929.

Hoegger P. J., Rigling, D., Holdenrieder, O., & Heiniger, U. (2000). Genetic structure of newly established populations of Cryphonectria parasitica. Mycological Research, 104(9), 1108– 1116.

Hoff K.J., Lange, S., Lomsadze, A., Borodovsky, M. and Stanke, M. (2015). BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics, 32(5):767-769.

Howlett B. J. (2006). Secondary metabolite toxins and nutrition of plant pathogenic fungi. Current Opinion in Plant Biology, 9(4), 371–375.

Hua, C., Zhao, J., & Guo, H. (2018). Trans-Kingdom RNA Silencing in Plant – Fungal Pathogen Interactions. *Molecular Plant*, *11*(2), 235–244.

Huang M., McClellan, M., Berman, J., & Kao, K. C. (2011). Evolutionary dynamics of candida albicans during in vitro evolution. Eukaryotic Cell, 10(11), 1413–1421.

Huang X. (2014). Horizontal transfer generates genetic variation in an asexual pathogen.

Hua-Van A., Le Rouzic, A., Boutin, T.S., Filée, J., Capy, P., 2011. The struggle for life of the genome's selfish architects. Biology Direct 6, 19–19.

Huddleston J. & Eichler E. E. (2016). An Incomplete Understanding of HumanGenetic Variation. Genetics. 202. 1251-1254.

Huson D.H., Bryant, D., 2006. Application of Phylogenetic Networks in Evolutionary Studies. Molecular Biology and Evolution 23, 254–267.

Huxley J. S., Pigliucci M & Muller G. B. (1942). The Modern Synthesis.

Irelan J. T., Hagemann, A. T., & Selker, E. U. (1994). High frequency repeat-induced point mutation (RIP) is not associated with efficient recombination in neurospora. Genetics, 138(4), 1093– 1103.

Jamilloux V., Daron, J., Choulet, F., Quesneville, H., 2017. De Novo Annotation of Transposable Elements: Tackling the Fat Genome Issue. Proceedings of the IEEE 105, 474–481.

Jensen J. D., Payseur, B. A., Stephan, W., Aquadro, C. F., Lynch, M., Charlesworth, D., & Charlesworth, B. (2019). The importance of the Neutral Theory in 1968 and 50 years on: A response to Kern and Hahn 2018. Evolution, 73(1), 111–114.

Jiao W. B., & Schneeberger, K. (2017). The impact of third generation genomic technologies on plant genome assembly. Current Opinion in Plant Biology, 36, 64–70.

Jørgensen L. N., Hovmøller, M. S., Hansen, J. G., Lassen, P., Clark, B., Bayles, R., Berg, G. (2014). IPM Strategies and Their Dilemmas Including an Introduction to www.eurowheat.org. Journal of Integrative Agriculture, 13(2), 265–281.

Ju G., & Skalka, A. M. (1980). Nucleotide sequence analysis of the Long Terminal Repeat (LTR) of avian retroviruses: Structural similarities with transposable elements. Cell, 22(2), 379–386.

Kaessmann H., Vinckenbosch, N., Long, M., 2009. RNA-based gene duplication: mechanistic and evolutionary insights. Nature Reviews Genetics 10, 19–31.

Käfer E., (1977). Meiotic and mitotic recombination in Aspergillus and its chromosomal aberrations. Adv. Genet. 19, 33–131.

Kapitonov VV, Jurka J. (2007). Helitrons on a roll: eukaryotic rolling-circle transposons. Trends Genet. 23, 521–529.

Kempken F., & Kück, U. (1998). Transposons in filamentous fungi-facts and perspectives. BioEssays, 20(8), 652–659.

Kern A. D. and M. W. Hahn. (2018). The neutral theory in light of natural selection. Mol. Biol. Evol. 35:1366–71.

Kimura M. (1968). Evolutionary rate at molecular level. Nature. 217, 624–626.

Kimura M. (1983) The Neutral Theory of Evolution, Cambridge University Press

Kimura M., Weiss G. H. (1964). The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics. 49(4), 561–576.

Koboldt D.C., Chen, K., Wylie, T., Larson, D.E., McLellan, M.D., Mardis, E.R., Weinstock, G.M., Wilson, R.K., Ding, L., 2009. VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25, 2283–2285.

Koboldt D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., Wilson, R.K., 2012. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research 22, 568–576.

Koren, S., Harhay, G. P., Smith, T. P. L., Bono, J. L., Harhay, D. M., Mcvey, S. D., ... Phillippy, A. M. (2013). Reducing assembly complexity of microbial genomes with single-molecule sequencing.

Koren S., Walenz B. P., Berlin K., Miller J. R., Bergman N. H. & Phillipy A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27. 722-736.

Koskinen M. T., Haugen, T. O., & Primmer, C. R. (2002). Contemporary fisherian life-history evolution in small salmonid populations. Nature, 419(6909), 826–830.
Koskinen M. T., Nilsson, J., Veselov, A. J., Potutkin, A. G., Ranta, E., & Primmer, C. R. (2002). Microsatellite data resolve phylogeographic patterns in European grayling, Thymallus thymallus, Salmonidae. Heredity, 88(5), 391–401.

Kulp D., Haussler D., Reese G. M., Eeckman F. H. (1996). A generalized Hidden Markov model for the recognition of Human genes in DNA. International Conference on Intelligent Systems for Molecular Biology. 4, 134-142.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large ge- nomes. Genome Biol 5: R12.

L'Hostis B., Hiremath, S. T., Rhoads, R. E., & Ghabrial, S. A. (1985). Lack of Sequence Homology between Double-stranded RNA from European and American Hypovirulent Strains of Endothia parasitica. Journal of General Virology, 66(2), 351–355.

Lande R. (2009). Adaptation to an extraordinary environment by evolution of phenotypic plasticity and genetic assimilation. Journal of Evolutionary Biology, 22(7), 1435–1446.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9:357–359.

Laurent, B., Moinard, M., Spataro, C., Ponts, N., Barreau, C., & Foulongne-oriol, M. (2017). Landscape of genomic diversity and host adaptation in Fusarium graminearum, 1–19.

Laurent B., Palaiokostas, C., Spataro, C., Moinard, M., Zehraoui, E., Houston, R. D., & Foulongne- Oriol, M. (2018). High-resolution mapping of the recombination landscape of the phytopathogen Fusarium graminearum suggests two-speed genome evolution. Molecular Plant Pathology, 19(2), 341–354

Lavergne S., & Molofsky, J. (2007). Increased genetic variation and evolutionary potential drive the success of an invasive grass. Proceedings of the National Academy of Sciences of the United States of America, 104(10), 3883–3888.

Lee J., Han K., Meyer T. J., Kim H-S & Batzer M. A. (2008). Chromosomal Inversions between Human and Chimpanzee Lineages Caused by Retrotransposons. PLoS ONE 3(12): e4047.

Leroy T., Lemaire, C., Dunemann, F., & Le Cam, B. (2013). The genetic structure of a Venturia inaequalis population in a heterogeneous host population composed of different Malus species. BMC Evolutionary Biology, 13(1).

Leslie F. John (1993). Fungal vegetative compatibility. Annual Review of Phytopathologist, 31,127-150.

Lewontin R. C. & Hubby J. L. (1966). A Molecular Approach to the Study of Genic Heterozygosity in Natural Populations. II. Amount of Variation and Degree of Heterozygosity in Natural Populations of DROSOPHILA PSEUDOOBSCURA. Genetics. 54(2), 595-609.

Li H. (2016). Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences. Bioinformatics, 32(14), 2103–2110.

Li H., Handsaker B.,Wysoker A., Fennell T., Ruan J., Homer G., Marth G., Abecasis G., & Durbin R. (2009). 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools, Bioinformatics. 25(16), 2078–2079

Liu Y. C., Cortesi, P., Double, M. L., MacDonald, W. L., & Milgroom, M. G. (1996). Diversity and multilocus genetic structure in populations of Cryphonectria parasitica. Phytopathology, Vol. 86, pp. 1344–1351.

Lo Presti L., Lanver, D., Schweizer, G., Tanaka, S., Liang, L., Tollot, M.,Kahmann, R. (2015). Fungal Effectors and Plant Susceptibility. Annual Review of Plant Biology, 66(1), 513–545.

Luikart G., England, P. R., Tallmon, D., Jordan, S., & Taberlet, P. (2003). The power and promise of population genomics: From genotyping to genome typing. Nature Reviews Genetics, 4(12), 981–994.

Ma Z., & Michailides, T. J. (2005). Genetic structure of Botrytis cinerea populations from different host plants in California. Plant Disease, 89(10), 1083–1089.

Daboussi M-J & Capy P. (2003). Transporable elements in filamentous fungi. Annual review of microbiology. 57, 99-275.

Ma, L. J., Van Der Does, H. C., Borkovich, K. A., Coleman, J. J., Daboussi, M. J., Di Pietro, A.,Rep, M. (2010). Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. Nature, 464(7287), 367–373.

Martin D. P., Lemey, P., & Posada, D. (2011). Analysing recombination in nucleotide sequences. Molecular Ecology Resources, 11(6), 943–955.

McClintock B. (1953). Induction of Instability at selected loci in Maize. Genetics. 38(6), 579-599.

McGuire I. C., Marra, R. E., Turgeon, B. G., & Milgroom, M. G. (2001). Analysis of mating-type genes in the chestnut blight fungus, Cryphonectria parasitica. Fungal Genetics and Biology, 34(2), 131–144.

McMullan M., Rafiqi, M., Kaithakottil, G., Clavijo, B. J., Bilham, L., Orton, E.,Clark, M. D. (2018). The ash dieback invasion of Europe was founded by two genetically divergent individuals. Nature Ecology and Evolution, 2(6), 1000–1008.

Merker J. D., Wenger, A. M., Sneddon, T., Grove, M., Zappala, Z., Fresard, L.,Ashley, E. A. (2018). Long-read genome sequencing identifies causal structural variation in a Mendelian disease. Genetics in Medicine, 20(1), 159–163.

Michael T. P. (2014). Plant genome size variation: Bloating and purging DNA. Briefings in Functional Genomics and Proteomics, 13(4), 308–317.

Michiels, A., Ende, W. Van Den, Tucker, M., & Riet, L. Van. (2003). Extraction of high-quality genomic DNA from latex-containing plants q, *315*, 85–89.

Milgroom M. G. 1996. Recombination and the multilocus structure of fungal populations. Annu. Rev. Phytopathol. 34:457–477.

Milgroom MG, Sotirovski K, Spica D et al. (2008) Clonal popu- lation structure of the chestnut blight fungus in expanding ranges in southeastern Europe. Molecular Ecology, 17, 4446– 4458.

Mills R. E., Bennett A. E., Iskow R. C. & Devine S. E. (2007). Which transposable elements are active in the human genome ? Trends in Genetics. 23(4), 183-191.

Moan, A. Le. (2019). Evolution at two-time frames shape structural variants and population structure of European plaice (Pleuronectes platessa).

Möller, E. M., Bahnweg, G., Sandermann, H., & Geiger, H. H. (1992). A simple and efficient protocol for isolation of high molecular weight DNA from filamentous fungi , fruit bodies , and infected plant tissues, *20*(22), 6115–6116.

Möller M., and E. H. Stukenbrock (2017). Evolution and genome architecture in fungal plant pathogens. Nat. Rev. Microbiol. 15: 756–771

Mousavi-Derazmahalleh M., Chang, S., Thomas, G., Derbyshire, M., Bayer, P. E., Edwards, D.,Hane, J. K. (2019). Prediction of pathogenicity genes involved in adaptation to a lupin host in the fungal pathogens Botrytis cinerea and Sclerotinia sclerotiorum via comparative genomics. BMC Genomics, 20(1), 1–11.

Muller H. J. (1964). The relation of recombination to mutational advance. Mutation Research 1:2–9

Muszewska A., Hoffman-Sommer, M., & Grynberg, M. (2011). LTR retrotransposons in fungi. PLoS ONE, 6(12).

Muszewska P., Feurtey A., Hood M. E;, Snirc A., Clavel J., Dutech C. Roy M. & Giraud T. (2011). The population biology of fungal invasions. Molecular ecology. 24, 1969-1986.

Naciri-Graven Y. Goudet, J. (2003). Number of Loci Involved in Epistatic Interactions. Evolution, 57(4), 706–716.

Nattestad M., Goodwin, S., Ng, K., Baslan, T., Sedlazeck, F. J., Rescheneder, P.,Schatz, M. C. (2018). Complex rearrangements and oncogene amplifications

revealed by long-read DNA and RNA sequencing of a breast cancer cell line. Genome Research, 28(8), 1126–1135.

Nieuwenhuis B. P. S., & James, T. Y. (2016). The frequency of sex in fungi. Philosophical Transactions of the Royal Society B: Biological Sciences, 371(1706).

Nikolenko S. I., Korobeynikov A. I. & Alekseyev M. A. (2013). BayesHammer: Bayesian clustering for error correction in single-cell sequencing. BMC Genomics 14. S7.

Nolte A. W., Freyhof., J., Stemshorn K. C., & Tautz D. (2005). An invasive lineage of sculpins, Cottus sp. (Pisces, Teleostei) in the Rhine with new habitat adaptations has originated from hybridization between old phylogeographic groups. Proceedings of the Royal society. 272, 2379–2387.

Old K. M & Kobayashi T. (1988). Eucalypts Are Susceptible to the Chestnut Blight Fungus, Cryphonectria Parasitica. Australian Journal of Botany. 36(5). 599-603.

Ordóñez V., Pascual, M., Rius, M., & Turon, X. (2013). Mixed but not admixed: A spatial analysis of genetic variation of an invasive ascidian on natural and artificial substrates. Marine Biology, 160(7), 1645–1660.

Otto S. P. (2009). The evolutionary enigma of sex. American Naturalist, 174(SUPPL. 1). S1–S14.

Ottonelli Rossato D., Ludwig A., Depra M. Loreto E. L. S., Ruiz A & Valente V. L. S. (2014). BuT2 Is a Member of the Third Major Group of hAT Transposons and Is Involved in Horizontal Transfer Events in the Genus Drosophila. Genome Biology and Evolution. 6(2), 352-365

Ou S., H. (1980). A look at worldwide Rice Blast disease control. American Phytopatholocal Society. 64(5), 439-445.

Ozsolak F., & Milos, P. M. (2011). RNA sequencing: Advances, challenges and opportunities. Nature Reviews Genetics, 12(2), 87–98.

Ozturk I. K., Dupont, P. Y., Chettri, P., McDougal, R., Böhl, O. J., Cox, R. J., & Bradshaw, R. E. (2019). Evolutionary relics dominate the small number of secondary metabolism genes in the hemibiotrophic fungus Dothistroma septosporum. Fungal Biology, 123(5), 397–407.

Paini D. R., Sheppard, A. W., Cook, D. C., De Barro, P. J., Worner, S. P., & Thomas, M. B. (2016). Global threat to agriculture from invasive species. Proceedings of the National Academy of Sciences of the United States of America, 113(27), 7575–7579.

Palmer J. M., & Keller, N. P. (2010). Secondary metabolism in fungi: Does chromosomal location matter? Current Opinion in Microbiology, 13(4), 431–436.

Paoletti M., Buck, K. W., & Brasier, C. M. (2006). Selective acquisition of novel mating type and vegetative incompatibility genes via interspecies gene transfer in the globally invading eukaryote Ophiostoma novo-ulmi. Molecular Ecology, 15(1), 249–262.

Park S. T., & Kim, J. (2016). Trends in next-generation sequencing and a new era for whole genome sequencing. International Neurourology Journal, 20, 76–83.

Parker I. M., & Gilbert, G. S. (2004). The evolutionary ecology of novel plant-pathogen interactions. Annual Review of Ecology, Evolution, and Systematics, 35, 675–700.

Pimentel D., McNair, S., Janecka, J., Wightman, J., Simmonds, C., O'Connell, C., Tsomondo, T. (2001). Economic and environmental threats of alien plant, animal, and microbe invasions. Agriculture, Ecosystems and Environment, 84(1), 1–20.

Pimentel D., Zuniga, R., & Morrison, D. (2005). Update on the environmental and economic costs associated with alien-invasive species in the United States. Ecological Economics, 52(3 SPEC. ISS.), 273–288.

Plissonneau C., Hartmann, F. E., & Croll, D. (2018). Pangenome analyses of the wheat pathogen Zymoseptoria tritici reveal the structural basis of a highly plastic eukaryotic genome. BMC Biology, 16(1), 1–16.

Plissonneau C., Stürchler, A., & Croll, D. (2016). The evolution of orphan regions in genomes of a fungal pathogen of wheat. MBio, 7(5).

Presti, L. Lo, Lanver, D., Schweizer, G., Tanaka, S., Liang, L., Tollot, M., … Kahmann, R. (2015). Fungal Effectors and Plant Susceptibility.

Prospero, S., & Rigling, D. (2012). Invasion Genetics of the Chestnut Blight Fungus Cryphonectria parasitica in Switzerland, *102*(1).

Rigling D. & Prospero S. (2017) Cryphonectria parasitica, the causal agent of chestnut blight: invasion history, population biology and disease control. Molecular Plant pathogolo. 19(1). 7-20.

Qi X. Structure, Function and Evolutionof Filamentous Fungal Telomerase RNA. PhD degree under the supervision of Dr. Julian J.L. Chen. Arizona state University. 2011.

Quesneville H., Bergman C. M., Andrieu O., Autard D., Nouaud D., Ashburner M., et al. (2015). Combined evidence annotation of transposable elements in genome sequences. PLoS Comput Biol. 1, 166–75.

Raffaele S., & Kamoun, S. (2012). Genome evolution in filamentous plant pathogens: Why bigger can be better. Nature Reviews Microbiology, 10(6), 417–430.

Ramírez F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S.,Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Research, 44(W1), W160–W165.

Reed D.H. & Frankham R. (2001). How closely correlated are molecular and quantitative measures of genetic variation? A meta-analysis. Evolution. 55,1095–103

Reed, D. H., & Frankham, R. (2003). Society for Conservation Biology Correlation between Fitness and Genetic Diversity. Conservation Biology, 17(1), 230–237.

Richardson D. M., Pysek P., Rejmanek M., Barbour M. G., Panetta F. D. & West C.J. (2000). Naturalization and invasion of alien plants: concepts and definitions. Diversity and Distributions. 6,93–107.

Rieseberg L. H. (2001). Chromosomal rearrangements and speciation. TRENDS in Ecology & Evolution, 16(7), 351–358.

Rius M., & Darling, J. A. (2014). How important is intraspecific genetic admixture to the success of colonising populations? Trends in Ecology and Evolution, 29(4), 233–242.

Riva Rossi C. M., Pascual, M. A., Aedo Marchant, E., Basso, N., Ciancio, J. E., Mezga, B., Ernst-Elizalde, B. (2012). The invasion of Patagonia by Chinook salmon (Oncorhynchus tshawytscha): Inferences from mitochondrial DNA patterns. Genetica, 140(10–12), 439–453.

Roberts R. J., Carneiro, M. O., & Schatz, M. C. (2013). The advantages of SMRT sequencing. Genome Biology. 405(14). 6–9.

Robin C., & Heiniger, U. (2001). Chestnut blight in Europe : Diversity of Cryphonectria parasitica , hypovirulence and biocontrol. Forest Snow and Landscape Research, 76(3), 361–367.

Robin C., Andanson, A., Saint-Jean, G., Fabreguettes, O., & Dutech, C. (2017). What was old is new again: thermal adaptation within clonal lineages during range expansion in a fungal pathogen. Molecular Ecology, 26(7), 1952–1963.

Roman J., Darling J. A. (2007). Paradox lost: genetic diversity and the success of aquatic invasions. Trends Ecol Evol. 22, 454-464.

Rouxel T., Grandaubert J. Howlett B. J. (2011). Effector diversification within compartments of leptosphaeria maculans genome affected by Repeat-Induced Point mutations. Nature

Sakai A.K., Allendorf J., With K.A., Baughman S., Cabin ey D.E., O'neil P., Parker I.M., Thom ılation biology of invasive spec Systematics. 32, 305-332.

Saleh D., Mila 012). Asexual reproduction indu reproduction capacity in the ric Results of in vitro experimental y, 12(1), 1–16.

Saleh D., Xu, P., Shen, D. (2012). Sex at the origin: An Asian po aporthe oryzae reproduces sexually. Molecu

Santini A., Ghelardini, L., De Pace, C., Desprez-Loustau, M. L., Capretti, P., Chandelier, A., Stenlid, J. (2013). Biogeographical patterns and determinants of invasion by forest pathogens in Europe. New Phytologist, 197(1), 238–250.

Sárközy P., Molnár, V., Fogl, D., Szalai, C. and Antal, P. (2017) "Beyond Homopolymer Errors: a Systematic Investigation of Nanopore-based DNA Sequencing Characteristics Using HLA- DQA2", Periodica Polytechnica Electrical Engineering and Computer Science. 61(3), 231- 237

Sax D.F. & Brown, J.H. (2000) The paradox ofinvasion. Global Ecol. Biogeogr. 9, 363–37

Schneider G. F., & Dekker, C. (2012). DNA sequencing with nanopores. Nature Biotechnology, 30(4), 326–328.

Sedlazeck F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. Nature Methods, 15(6), 461–468.

Seidl M. F., & Thomma, B. P. H. J. (2017). Transposable Elements Direct The Coevolution between Plants and Microbes. Trends in Genetics, 33(11), 842–851.

Selechnik D., Richardson, M. F., Shine, R., DeVore, J., Ducatez, S., & Rollins, L. A. (2019). Bottleneck revisited: increased adaptive variation despite reduced overall genetic diversity in a rapidly adapting invader. BioRxiv, 557868.

Shi-Kunne X., Faino, L., van den Berg, G. C. M., Thomma, B. P. H. J., & Seidl, M. F. (2018). Evolution within the fungal genus Verticillium is characterized by chromosomal rearrangement and gene loss. Environmental Microbiology, 20(4), 1362–1373.
Sievers F., & Higgins, D. G. (2014). Clustal Omega. Current Protocols in Bioinformatics, 2014(December), 3.13.1-3.13.16.

Sillo F., Garbelotto, M., Friedman, M., & Gonthier, P. (2015). Comparative genomics of sibling fungal pathogenic taxa identifies adaptive Evolution without divergence in pathogenicity genes or genomic structure. Genome Biology and Evolution, 7(12), 3190–3206.

Simão F. A., Waterhouse R. M., Ioannidis P., Kriventseva E. V., Zdobnov E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics. 31(19), 3210–3212

Sinzelle L., Izvsak E. & Ivics Z. (2009). Molecular domestication of transposable elements: From detrimental parasites to useful host genes. Cellular and Molecular Life Sciences. 66(6), 1073-109307/11/2019

Sipos G., Prasanna A. N., Walter M. C., O' Connor E. Bàlint B., Krizsàn K., Kiss B., Hess J., Varga T., Slot J., Riley R., Boka B., Ringling D., Barry K., Lee J., Mihaltcheva S., LaButti K., Lipzen A., Waldron R., Moloney N. M., Sperisen C., Kredics L., Vagvolgyi C., Patrignani A., Fitzpatrick D., Nagy I. Doyle S., Anderson J. B., Grigoriev I. V., Guldener U., Munsterkotter M & Nagy L. G. (2017). Genome expansion and lineage-specific genetic innovations in the forest pathogenic fungi Armillaria. Nature Ecology and Evolution. 1(12), 1931-1941.

Sipos G., Prasanna, A. N., Walter, M. C., O'Connor, E., Bálint, B., Krizsán, K., Nagy, L. G. (2017). Genome expansion and lineage-specific genetic innovations in the forest pathogenic fungi Armillaria. Nature Ecology and Evolution, 1(12), 1931–1941.

Slade R. W., & Moritz, C. (1998). Phylogeography of Bufo marinus from its natural and introduced ranges. Proceedings of the Royal Society B: Biological Sciences, 265(1398), 769–777.
Smith K. M., Galazka, J. M., Phatale, P. A., Connolly, L. R., & Freitag, M. (2012). Centromeres of filamentous fungi. Chromosome Research, 20(5), 635–656.

Soyer J. L., M. El Ghalid, N. Glaser, B. Ollivier, J. Linglin et al., (2014). Epigenetic control of effector gene expression in the plant path- ogenic fungus Leptosphaeria maculans. PLoS Genet. 10: e1004227

Sperschneider J, Dodds PN, Gardiner DM, Singh KB, Taylor JM. (2018). Improved prediction of fungal effector proteins from secretomes with EffectorP 2.0. Mol Plant Pathol. 19(9):2094– 2110

Stajich J. E., Berbee M. L., Blackwell M., Hibbett D. S. , James T. Y, Spatafora JW, Taylor J. W. (2009). Primer: the fungi. Curr Biol. 19, 840–845.

Stanke M., Diekhans, M., Baertsch, R. and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics

Stanke M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: A b initio prediction of alternative transcripts. Nucleic Acids Research, 34(WEB. SERV. ISS.), 435– 439.

Stephan W. (2010). Genetic hitchhiking versus background selection: The controversy and its implications. Philosophical Transactions of the Royal Society B: Biological Sciences, 365(1544), 1245–1253.

Strange R. N., & Scott, P. R. (2005). Plant Disease: A Threat to Global Food Security. Annual Review of Phytopathology, 43(1), 83–116.

Stuart T., Eichten, S. R., Cahn, J., Karpievitch, Y. V., Borevitz, J. O., & Lister, R. (2016). Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. ELife, 5(DECEMBER2016), 1–27.

Stukenbrock E. H. & Dutheil J. Y. (2018). Fine-Scale Recombination Maps of Fungal Plant Pathogens Reveal Dynamic Recombination Landscapes and Intragenic Hotspots. Genetics Investigation. 2018(march). 1209-1229.

Stukenbrock E. H., Banke, S., Javan-Nikkhah, M., & McDonald, B. A. (2007). Origin and domestication of the fungal wheat pathogen Mycosphaerella graminicola via sympatric speciation. Molecular Biology and Evolution, 24(2), 398–411.

Stukenbrock E. H., Jørgensen, F. G., Zala, M., Hansen, T. T., McDonald, B. A., & Schierup, M. H. (2010). Whole-genome and chromosome evolution associated with host adaptation and speciation of the wheat pathogen mycosphaerella graminicola. PLoS Genetics, 6(12), 1–13.

Stukenbrock E. H., Quaedvlieg, W., Javan-Nikhah, M., Zala, M., Crous, P. W., & McDonald, B. A. (2012). Zymoseptoria ardabiliae and Z. pseudotritici, two progenitor species of the septoria tritici leaf blotch fungus Z. tritici (synonym: Mycosphaerella graminicola). Mycologia, 104(6), 1397–1407.

Sun S., Yadav, V., Billmyre, R. B., Cuomo, C. A., Nowrousian, M., Wang, L., Heitman, J. (2017). Fungal genome and mating system transitions facilitated by chromosomal translocations involving intercentromeric recombination. PLoS Biology. 15.

Tajima F. (1989). The effect of change in population size on DNA polymorphism. Genetics, 123(3), 597–601.

Takabayashi N., Tosa, Y., Oh, H. S., & Mayama, S. (2002). A gene-for-gene relationship underlying the species-specific parasitism of Avena/Triticum isolates of Magnaporthe grisea on wheat cultivars. Phytopathology, 92(11), 1182–1188.

Taylor J. W., Geiser, D. M., Burt, A., & Koufopanou, V. (1999). The evolutionary biology and population genetics underlying fungal strain typing. Clinical Microbiology Reviews, 12(1), 126–146.

Thornton K. R. & J. D. Jensen. (2007). Controlling the false-positive rate in multi-locus genome scans for selection. Genetics. 175, 737–50.

Tsushima A., Gan, P., Kumakura, N., Narusaka, M., Takano, Y., Narusaka, Y., & Shirasu, K. (2019). Genomic plasticity mediated by transposable elements in the plant pathogenic fungus Colletotrichum higginsianum. Genome Biology and Evolution, 11(5), 1487–1500.

Van Hooff, J. J. E., Snel, B., & Seidl, M. F. (2014). Small homologous blocks in phytophthora genomes do not oint to an ancient whole-genome duplication. Genome Biology and Evolution, 6(5), 1079–1085.

Vitousek P. M., D'Antonio, C. M., Loope, L. L., & Westbrooks, R. (1996). Biological invasions as global environmental change. American Scientist, 84(5), 468–478.

Volff JN. (2006). Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. BioEssays. 28, 913–922.

Watters M. K., Randall, T. A., Margolin, B. S., Selker, E. U. & Stadler, D. R. (199). Action of repeat- induced point mutation on both strands of a duplex and on tandem duplications of various sizes in Neurospora. Genetics 153, 705–714.

Weber J. L., & Wong, C. (1993). Mutation of human short tandem repeats. Human Molecular Genetics, 2(8), 1123–1128.

Weischenfeldt J., Symmons, O., Spitz, F., & Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: Insights from and for human disease. Nature Reviews Genetics, 14(2), 125–138.

Whitney K. D., & Gabler, C. A. (2008). Rapid evolution in introduced species, "invasive traits" and recipient communities: Challenges for predicting invasive potential. Diversity and Distributions, 14(4), 569–580.

Whittle C. A., Nygren, K., & Johannesson, H. (2011). Consequences of reproductive mode on genome evolution in fungi. Fungal Genetics and Biology, 48(7), 661–667.

Wicker T., Sabot F., Jua-Van A., Bennetzen J. L., Capy P., Chalhoub B., Flavell A., Leroy P. Morgante M., Panaud O., Paux E., SanMiguel P. & Schulman A. H. (2007). A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics. 8(12), 973-982.

Willi Y., Griffin, P. & Van Buskirk, J. (2013). Drift load in populations of small size and low density. Heredity 110, 296–302

Williamson M. & Fitter A. (1996). The varying success of Invaders. Ecology. 77(6)- 1661-1666

Wright, S. (1931) Evolution in mendelian populations. Genetics. 16, 97–159.

Yoshida K., Saunders D. G. O., Mitsuoka C. Natsume S., Kosugi S., Saitoh H., Inoue Y., Chuma I., Tosa Y., Cano L. M., Kamoun S. & Terauchi R. (2016). Host specialization of blast fungus Magnaporthe oryzae is associated with dynamic gain and loss of genes linked to transposable elements. Genomics. 17:370.

Zerbino D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Research, 18(5), 821–829.

Zhang, M., Cui, S. W., Cheung, P. C. K., & Wang, Q. (2007). Antitumor polysaccharides from mushrooms : a review on their isolation process , structural characteristics and antitumor activity, *18*, 4–19.

Zhang D.X., Spiering M.J., Dawe A.L., Nuss D.L. (2014). Vegetative Incompatibility Loci with Dedicated Roles in Allorecognition Restrict Mycovirus transmission in chestun blight fungus, *197*(June), 701–714.

Zhu, H., Qu, F., & Zhu, L. (1993). Isolation of genomic DNAs from plants , fungi and bacteria using benzyl chloride, *21*(22), 5279–5280.

Annexes

I Annexe 1 : Informations supplémentaires du chapitre 1.

Supporting information for online publication

Text S1. Protocol of *C. parasitica* monospore isolation.

Text S2. Annotation of protein coding genes

Table S1. Sequencing information and mating-type of the 46 *Cryphonectria parasitica* isolates used in this study. † Number of different microsatellite allele(s) from the closest French clonal lineage; ‡ Coverage of sequencing in X; § Mating-type allele MAT1-1 and MAT1-2; ¶ North America.

Table S2. Summary results of TEdenovo and TEannot pipeline (REPET Package). REPET notation refer to the Wicker classification (Wicker 2007): the first letter refers to the initial of the class (retrotransposons or DNA transposons), the second to the Order (L for long terminal repeat (LTR), T for terminal inverted repeat (TIR)) and the last to the Super-family. X signifies an undetermined classification.

Figure S1. Comparison of the YVO003 (top) and EP155 (bottom) genomes obtained using progressive Mauve software (Darling *et al.*, 2004). The colors represent the 35 scaffolds defined in YVO003 denovo assembly. The potential structural rearrangements are indicated by the features that link both genome representations. Black lines represent the scaffolding of the two reference strains. Sizes of the two reference genomes are shown on the right.

Figure S2. Plot of the nucleotide diversity (π) calculated by 10kb windows along the 26 major scaffolds within the five studied French clonal lineages: RE019 (in blue), RE103 (in red), RE053 (in green), RE079 (in orange) and RE043 (in purple). Gray line represents the global nucleotide diversity (46 isolates).

Figure S3. The haplotypic patterns of ten *C. parasitica* French isolates based on 23,240 SNP (singletons removed) and given by 10kb windows along the genome. Color of each vertical line defines the frequency of the haplotype for the 10 isolates analyzed, and length of the sequence is represented by horizontal rectangles. Frequency of each haplotype was defined distinctly for the two introductions separated by the red line. White haplotype represents the most frequent haplotype found either in the south-eastern introduction, or in the south-western introduction. Black ones represent the second most frequent haplotype, blue the third and red the rarest.

Figure S4. RAxML genealogical trees used as starting tree for the BEAST analysis of the four fragments with no detected recombination using ClonalFrameML. a) MS1 b) MS3 c) MS4 d) MS8

Figure S5. Non-linear regression of the number of SNPs discovered as a function of the number of isolates of line RE092 considered between 1 and 6. The red line is the asymptote estimated by R for x[number of RE092 isolates]=+inf; it represents the supposed maximum genetic diversity that can be found in this line if an infinite number of individuals are sequenced. The tick named H13 shows the number of SNPs reached when comparing the six RE092 isolates and the H13 isolate. The red dotted lines represent the confidence intervals at +2.5% and -2.5%.

Text S1. Protocole of *C. parasitica* monospore isolation.

For each isolate, a monospore isolation was performed from a single pycnidia (the asexual structure of the fungus) after a first mycelial culture on Potato dextrose Agar (PDA) in a controlled growth chamber under fluorescente light (PAR = 30µmol per m^2 per seconde) and 8-hr photoperiod at 23°C. One pycnidia for each culture was collected, put in a 2mL microtube with 1mL of deionized water and successive decimal dilutions were applied. These dilutions

were then spread on PDA medium, and germination of a single spore, identified with a binocular, was subcultured on a new PDA medium.

Text S2. Annotation of protein-coding genes

Gene models were predicted with a fully automated and parallelized pipeline, egn-ep (http://eugene.toulouse.inra.fr/Downloads/egnep-Linux-x86_64.1.4.tar.gz, release 1.2), that manages probabilistic sequence model training, repeat masking, transcript and protein alignments and integrative gene modeling in EuGene software (release 4.2, Foissac *et al.* 2008). Three protein databases were aligned with ncbi-blastx (Camacho *et al.* 2009) to contribute to translated regions detection i) The proteome of *Cryphonectria parasitica* strain EP155 (https://genome.jgi.doe.gov/Crypa2) ii) Swiss-Prot - March 2013 iii) a fungi subset of Uniprot proteins – March 2013. Proteins similar to REPBASE (Bao *et al.* 2015) were removed from datasets prior mapping. The mRNA sequences of the annotated strain EP155 were used as transcript evidences (https://genome.jgi.doe.gov/Crypa2). The gene modeling algorithm used the standard EuGene 4.2 parameters, except that non canonical GC/donor sites were allowed. ncRNA genes were predicted by tRNAScan-SE (Lowe T and Eddy S 1996), RNAMMER (Lagesen *et al.* 2007) and rfamscan (Rfam release 12, Nawrocki *et al.* 2015). The set of predicted peptides was run on the BUSCO (release 3, Simao *et al.* 2015) fungi_odb9 dataset.

References of Text S1 and S2

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA, 6(1).

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden,T. L. (2009). BLAST+: architecture and applications. BMC Bioinformatics, 10(1), 421.

Foissac, S., Gouzy, J., Rombauts, S., Mathé, C., Amselem, J., & Sterck, L. (2008). Genome Annotation in Plants and Fungi: EuGène as a Model Platform, 11.

Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H.-H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Research, 35(9), 3100–3108.

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, 10.

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., …Finn, R. D. (2015). Rfam 12.0: updates to the RNA families database. Nucleic Acids Research, 43(D1), D130–D137.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 31(19), 3210–3212.

Table S1.

Area	Origin	Isolate	Genotype	Al. Close†	Sequencer	Depth‡	MAT§
Europe	Introduction from NA¶ In Italy	VAM001	re019	0	Ion-torrent	47.7	2
		VBO001				78.8	1
		STC104				66.3	1
		SAL005				50.0	1
		SMV006			Hiseq3000	144.0	1
		VAR01				104.7	1
		CHA006				107.1	1
		CER008	re092	0	Hiseq3000	121.2	2
		STC36_3				148.2	1
		MEN011	re103	0	Hiseq3000	110.2	1
		CIR002A				145.0	2
		ESP015A			Ion-torrent	74.5	2
		STC013-5	re079	0	Ion-torrent	64.6	2
		FON010A			Hiseq3000	119.9	2
		DOI040B				187.8	2
		YVO006	re093	1	Ion-torrent	42.4	1
		GAN019				44.5	2
		STL002	H68	2	Hiseq3000	134.5	2
		SAL025	re091	1	Hiseq3000	141.0	1
		ABI005G	re104	1	Hiseq3000	106.5	1
		YVO003	H53	1	Ion-torrent	104.2	2
		PIS001E			Hiseq3000	124.8	2
		ABI004	H13	3	Hiseq3000	97.9	2
	Italy – 1968	VG1896	-	-	Hiseq3000	146.3	1

Table S1.

Area	Origin	Isolate	Genotype	Al. Close†	Sequencer	Depth‡	MAT§
NA¶	New-York	DUM011	-	-	Hiseq3000	147.8	2
		DUM003				119.0	2
		DUM007				123.2	1
		DUM018			Ion-torrent	67.5	1
		DUM006X				63.3	2
	Kentucky	KYM023X			Hiseq3000	123.7	2
		KYM003X				110.2	2
	NewHampshire	NHM007X			Hiseq3000	141.0	2
		NHM022				122.6	1
		NHM002X				138.9	2
Europe	Introduction From Asia	MRC012B	re043	0	Ion-torrent	66.3	2
		MON006				41.0	2
		MRC010			Hiseq3000	137.5	2
		BAR002				113.4	1
		SAU001A	re053	0	Hiseq3000	64.7	2
		GAD002				123.2	2
		BBE009				104.0	2
		SAR005A				112.9	1
		SAR003	re028	0	Ion-torrent	74.7	1
	Pyrénées – 197	VG2106	-	-	Hiseq3000	102.7	2
Asia	China	XIM9508	-	-	Hiseq3000	107.2	1
	Japan	ESM015X	-	-	Hiseq3000	80.4	2

Table S2.

Classification	Structure	Families count	REPET notation	Copies count	Consensus size (kb)
Class I (retrotransposons)	Complete	2	RLX	69	8.0
				68	12.3
	Incomplete	10	RLX	5	7.0
				14	7.5
				30	7.8
				17	6.8
				36	5.1
				44	4.5
				209	19.0
				41	0.8
			RXX	26	4.0
				88	18.0
Class II (DNA transposons)	Complete	9	DTX	4	1.3
				15	2.2
				28	1.9
				11	1.8
				26	3.6
				14	1.9
				15	1.9
				11	1.8
				10	1.8
	Incomplete	5	DTX	23	3.9
				2	2.1
				8	3.9
				32	2.6
			DXX	17	1.0
Others	-	1	XXX	4	0.5
	-	1	PotentialHostGene	9	1.6

Figure S1.

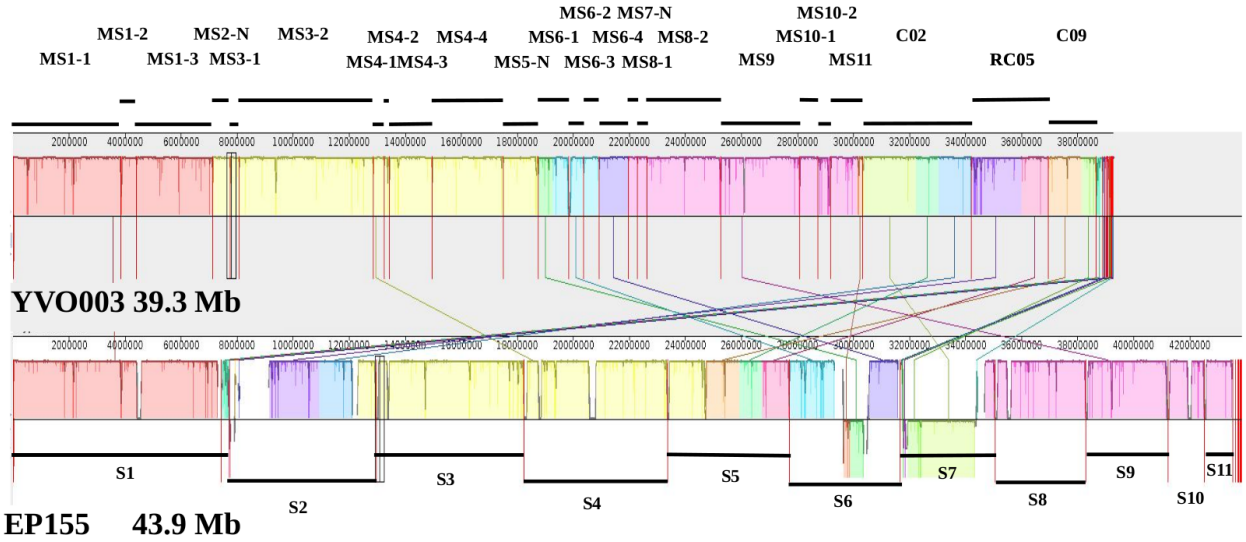

MS1-1
MS1-2
MS1-3
MS2-N
MS3-1
MS3-2
MS4-1
MS4-2
MS4-3
MS4-4
MS5-N
MS6-1
MS6-2
MS6-3
MS6-4
MS7-N
MS8-1
MS8-2
MS9
MS10-1
MS10-2
MS11
C02
RC05
C09
YVO003 39.3 Mb
EP155 43.9 Mb
S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
S11

Figure S2.

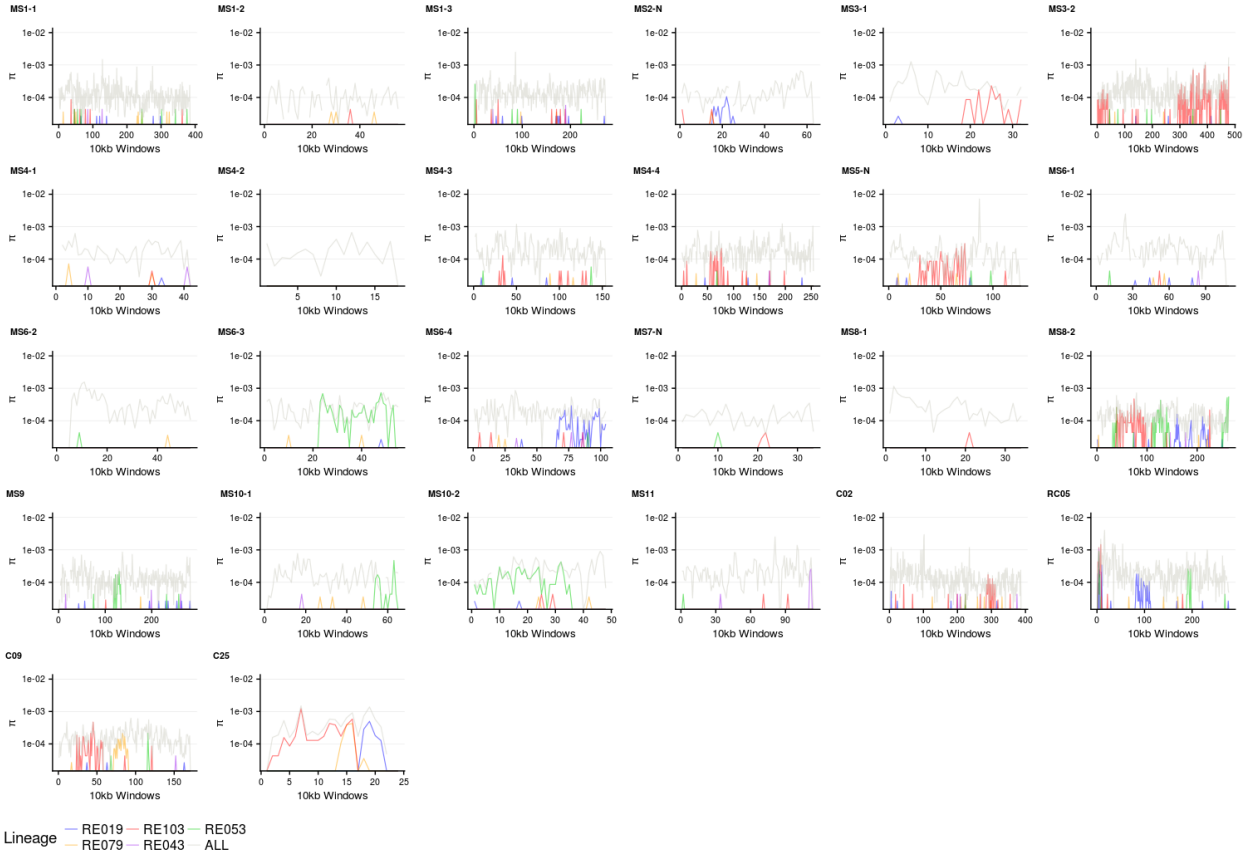

MS1-1
MS1-2
MS1-3
MS2-N
MS3-1
MS3-2
MS4-1
MS4-2
MS4-3
MS4-4
MS5-N
MS6-1
MS6-2
MS6-3
MS6-4
MS7-N
MS8-1
MS8-2
MS9
MS10-1
MS10-2
MS11
C02
RC05
C09
C25
π
1e-02
1e-03
1e-04
10kb Windows
Lineage
RE019
RE103
RE053
RE079
RE043
ALL

Figure S3.

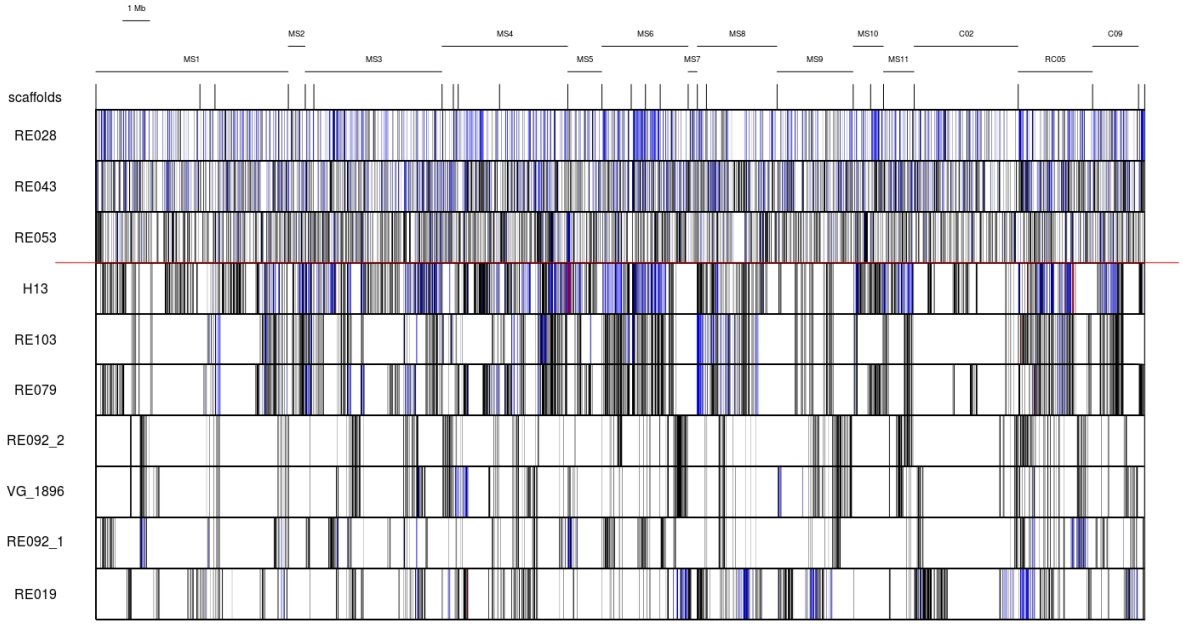

1 Mb
MS1
MS2
MS3
MS4
MS5
MS6
MS7
MS8
MS9
MS10
MS11
C02
RC05
C09
scaffolds
RE028
RE043
RE053
H13
RE103
RE079
RE092_2
VG_1896
RE092_1
RE019

Figure S4.

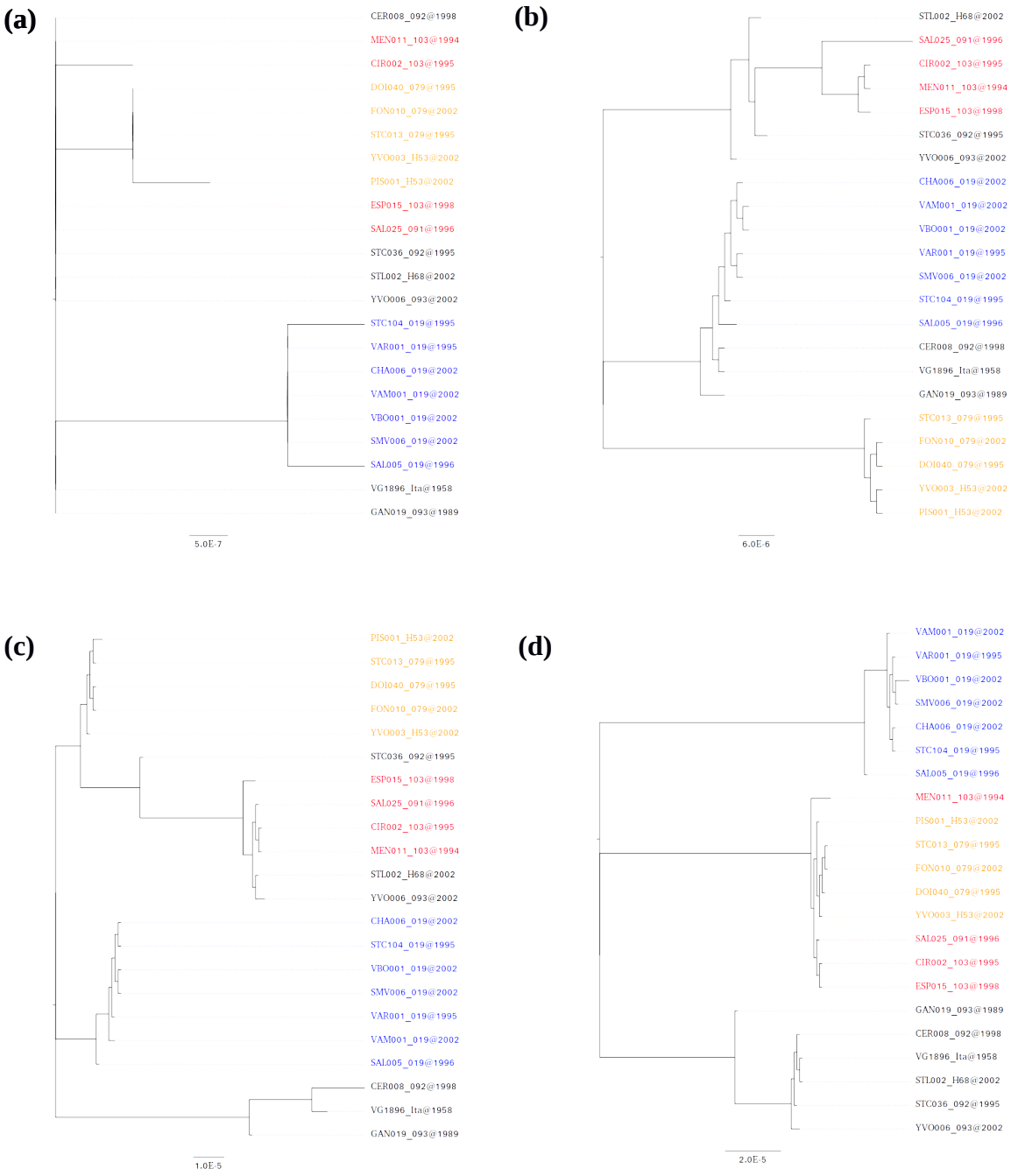
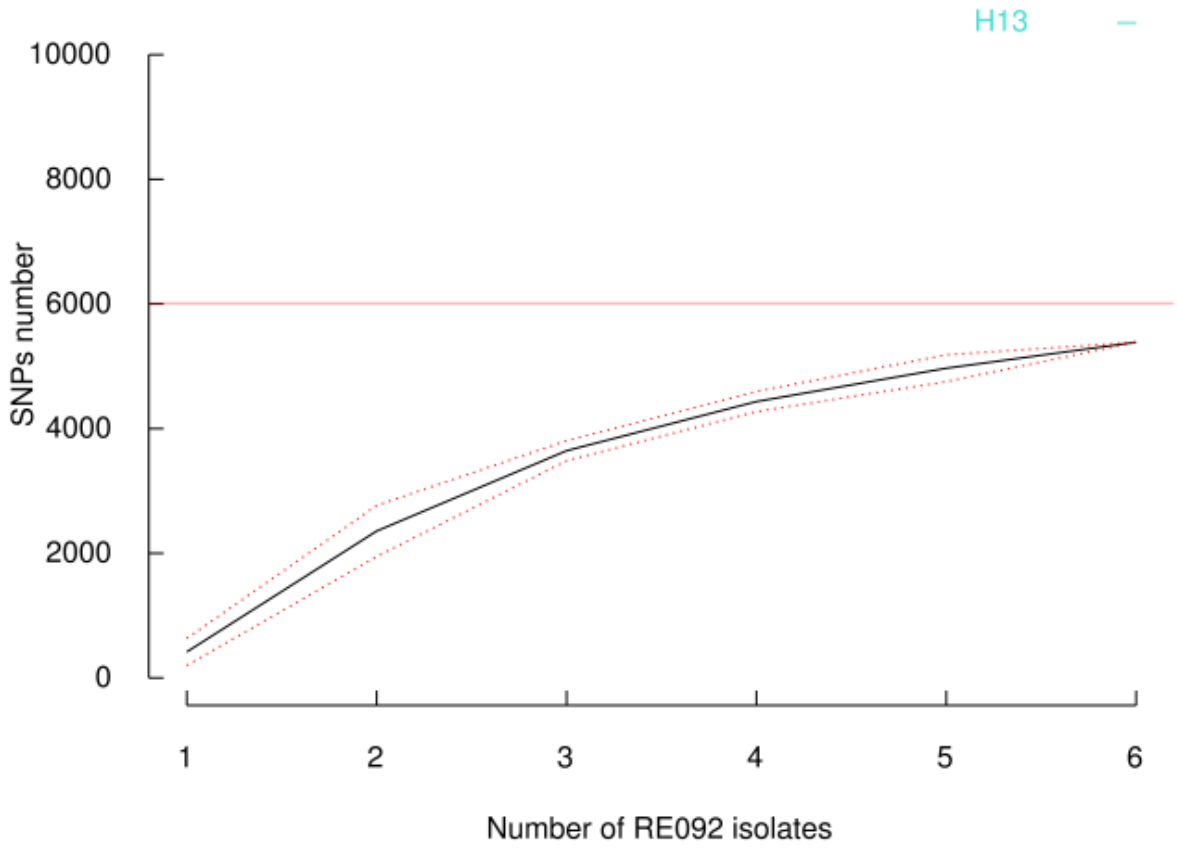

(a)
CER008_092@1998
MEN011_103@1994
CIR002_103@1995
DOI040_079@1995
FON010_079@2002
STC013_079@1995
YVO003_H53@2002
PIS001_H53@2002
ESP015_103@1998
SAL025_091@1996
STC036_092@1995
STL002_H68@2002
YVO006_093@2002
STC104_019@1995
VAR001_019@1995
CHA006_019@2002
VAM001_019@2002
VBO001_019@2002
SMV006_019@2002
SAL005_019@1996
VG1896_Ita@1958
GAN019_093@1989
5.0E-7
(b)
STL002_H68@2002
SAL025_091@1996
CIR002_103@1995
MEN011_103@1994
ESP015_103@1998
STC036_092@1995
YVO006_093@2002
CHA006_019@2002
VAM001_019@2002
VBO001_019@2002
VAR001_019@1995
SMV006_019@2002
STC104_019@1995
SAL005_019@1996
CER008_092@1998
VG1896_Ita@1958
GAN019_093@1989
STC013_079@1995
FON010_079@2002
DOI040_079@1995
YVO003_H53@2002
PIS001_H53@2002
6.0E-6
(c)
PIS001_H53@2002
STC013_079@1995
DOI040_079@1995
FON010_079@2002
YVO003_H53@2002
STC036_092@1995
ESP015_103@1998
SAL025_091@1996
CIR002_103@1995
MEN011_103@1994
STL002_H68@2002
YVO006_093@2002
CHA006_019@2002
STC104_019@1995
VBO001_019@2002
SMV006_019@2002
VAR001_019@1995
VAM001_019@2002
SAL005_019@1996
CER008_092@1998
VG1896_Ita@1958
GAN019_093@1989
1.0E-5
(d)
VAM001_019@2002
VAR001_019@1995
VBO001_019@2002
SMV006_019@2002
CHA006_019@2002
STC104_019@1995
SAL005_019@1996
MEN011_103@1994
PIS001_H53@2002
STC013_079@1995
FON010_079@2002
DOI040_079@1995
YVO003_H53@2002
SAL025_091@1996
CIR002_103@1995
ESP015_103@1998
GAN019_093@1989
CER008_092@1998
VG1896_Ita@1958
STL002_H68@2002
STC036_092@1995
YVO006_093@2002
2.0E-5

Figure S5.


H13
10000
8000
6000
4000
2000
0
SNPs number
1
2
3
4
5
6
Number of RE092 isolates

II Annexe 2 : Mise au point d'un protocole d'extraction d'ADN

1. Contexte

Les méthodes de séquençage de troisième génération nécessitent d'importantes quantité d'ADN de haut poids moléculaire (longues séquences d'ADN de plusieurs kilobases) et présenter peut d'inhibiteurs enzymatiques qui pourrait interférer avec le processus de séquençage (Arseneau et al., 2017). Il est donc essentiel d'avoir un protocole d'extraction qui permet de purifier un maximum l'ADN tout en préservant la longueur des brins. Le développement du séquençage de seconde génération a entraîné un tel engouement que de nombreux kit commerciaux ont été développés pour obtenir plus rapidement et facilement des échantillons d'ADN.

Cependant, les polysaccharides représentent un important pourcentage de la biomasse des champignons filamenteux, notamment dans les parois cellulaires de leurs hyphes dans lesquelles ont peut retrouver jusqu'à 75 % de polysaccharides (Gutiérrez et al., 1996). Les polysaccharides sont des polymères glucidiques qui peuvent servir de structures de soutient, de gaine protectrice du mycélium et certains seraient des molécules bio-actives (Zhang et al., 2007). Un des problèmes couramment rencontré dans les extraction d'ADN est la co-précipitation des polysaccharides avec les molécules d'ADN, ce qui peut conduire, dans les organismes produisant beaucoup de polysaccharides comme les champignons filamenteux et les plantes, à l'extraction de quantité insuffisante d'ADN, et la dégradation de la qualité de l'extraction (Möller et el., 1992).

2. Motivations

Un protocole d'extraction d'ADN de haut poids moléculaire a été mis en place par Robert Debuchy en 2013 (décrit dans Cheeseman et al., 2013). Ce

protocole a été utilisé pour produire l'extrait d'ADN de haut poids méloculaire qui a servit au séquençage PacBIO de l'isolat de *Cryphonectria parasitica* YVO003 décrit comme génome de référence dans le chapitre I (Demené et al., 2019. Cependant, ce protocole nécessite une étape d'ultra-centrifugation à 50,000 r.p.m. Il est rare que les laboratoires de biologie moléculaire aient un accès direct à une ultra-centrifugeuses. Aussi, une étape de dialyse de quatre jours dans une solution de 10mM TRIS et 1mM EDTA (pH 8.0) permet de purifier l'ADN. Et enfin la masse de matériel de départ doit être approximativement de 10 grammes, ce qui peut être difficile à obtenir dans certains cas. Au cours de ma thèse, j'ai voulu développer un protocole d'extraction d'ADN de haut poids moléculaire réalisable avec le matériel classique présent en laboratoire de biologie moléculaire.

3. Travail préliminaire

La qualité des extractions a été testé à l'aide de trois outils :

- Le Nanodrop (Thermo Scientific, Wilmington, MA) permet de mesurer l'absorbance d'une solution. Les nucléotides présents dans l'ARN et l'ADN ont une longueur d'onde d'absorption maximal à 260nm. Le ratio entre l'absorbance de la solution à 260nm et 280nm (260/280) donne une information sur la pureté de l'extrait d'ADN ou d'ARN par rapport aux protéines, aux phénoles et d'autres contaminants qui absorbent à 280nm. Les molécules d'ARN se diférentient des molécules d'ADN par la présence de bases nucléiques Uracil à la place des bases nucléiques Thymine, ces dernières ayant une capacité d'absorbtion plus faible (Nanodrop technical support bulletin). Un ratio 260/280 proche de 1,8 témoigne d'échantillons d'ADN pures alors que les solutions d'ARN pures auront un ratio proche de 2,2-2,3 (O'Neill et al., 2011). Le ratio entre l'absorbance à 260nm et l'absorbance à 230nm indiquent la pureté des acides nucléiques par rapport aux sucres (dont les polysaccharides) et certains phénols. Un échantillon d'ADN de bonne qualité devrait avoir un ratio 260/230 plus grand que 1,5 (O'Neill et al., 2011). Les concentrations estimées à partir de la conversion des valeurs d'absorbances seront données à

titre indicatif car de nombreuses molécules ont des absorbances maximales à 260nm et peuvent entraîner un biais.

- Le Qubit (Life Technologies, Carlsbad, CA, USA) permet de mesurer la fluorométrie d'une solution d'ADN à laquelle a été ajouté un réactif qui se lie à l'ADN double brin combiné à un fluorochrome. La conversion du signal de fluorescence en concentration d'ADN permet d'obtenir la quantité précise d'ADN double brin dans la solution.

- L'éléctrophorèse sur gel d'agarose permet de séparer les molécules d'ADN et d'ARN en fonction de leur masse moléculaire. Les molécule d'ADN obtenues par les extractions testées ici sont de grande taille et les gels ont été réalisé à 1 % d'agarose. Un marqueur de taille (« ladder ») 1kb a été utilisé.

3. a) Comparaison des tampons de lyse cellulaire

Méthode

Le premier essai d'extraction a été réalisé à partir du protocole « Purificationof High Molecular Weight Genomic DNA from Powdery Mildew for Long-Read Sequencing » (Feehan et al., 2017). Ce protocole utilise un tampon de lyse cellulaire avec du Cetyltrimethyl ammonium bromide (CTAB). Les protocoles à base de CTAB sont supposé éliminer les contaminations des extraits d'ADN par les polysaccharides (Michiels et al., 2002). Cependant, un protocole d'extraction spécifique à Cryphonectria parasitica utilisé en routine pour des extractions d'ADN à des fin de séquençage de seconde génération, spécifie une utilisation d'un tampon de lyse cellulaire avec du sodium dodecyl sulfate (SDS). Il a donc été réalisé deux extractions grâce au protocole de Feehen et ses collaborateurs (2017) en utilisant ces deux tampons de lyses (Tampon de lyse CTAB décrit dans Feehen et al., 2017 ; tampon de lyse SDS décrit dans l'annexe 1). Ces extractions ont été réalisées sur la souche ESM015 de *C. parasitica* échantillonnée au Japon, cultivée pendant 30 jours sur un milieu à base de potato dextrose agar (PDA) recouverts d'un papier cellophane stérile. Environ 0,3 grammes de mycelium ont été récoltés et broyés dans l'azote liquide à l'aide de mortiers et pilons stériles pendant trois minutes.

Résultats

Les résultats des extractions à l'aide du tampon de lyse au SDS montrent de meilleurs résultats que les extractions à l'aide du tampon à base de CTAB. L'analyse de l'absorbance (Nanodrop) des ces échantillons montre que les solutions d'ADN obtenues grâce au protocole SDS sont plus pures (260/280 moyen = 1,53 ; 260/230 moyen = 0,78; Tableau I) que les solutions obtenues grâce au protocole CTAB (260/280 moyen = 1,36; 260/230 moyen = 0,61 ; Tableau I). Les quantités d'ADN double brin obtenues sont beaucoup plus élevées grâce au protocole SDS (5 ,88µg et 2,75µg; Tableau I) que celles obtenues grâce au protocole CTAB (0,08µg et 0,26µg; Tableau I). Ces résultats sont cohérents avec la migration sur gel d'agarose (figure 1). Seuls les extraits d'ADN du protocole SDS montrent une bande nette d'ADN génomique de taille de fragments supérieure à 10kb. L'extrait C* du protocole CTAB ayant subit

deux traitements au chloroforme:isoamyl alcool montre une légère bande suite à la migration, et l'extrait d'ADN S* montre une bande d'intensité plus forte que S1, S2 et S3.

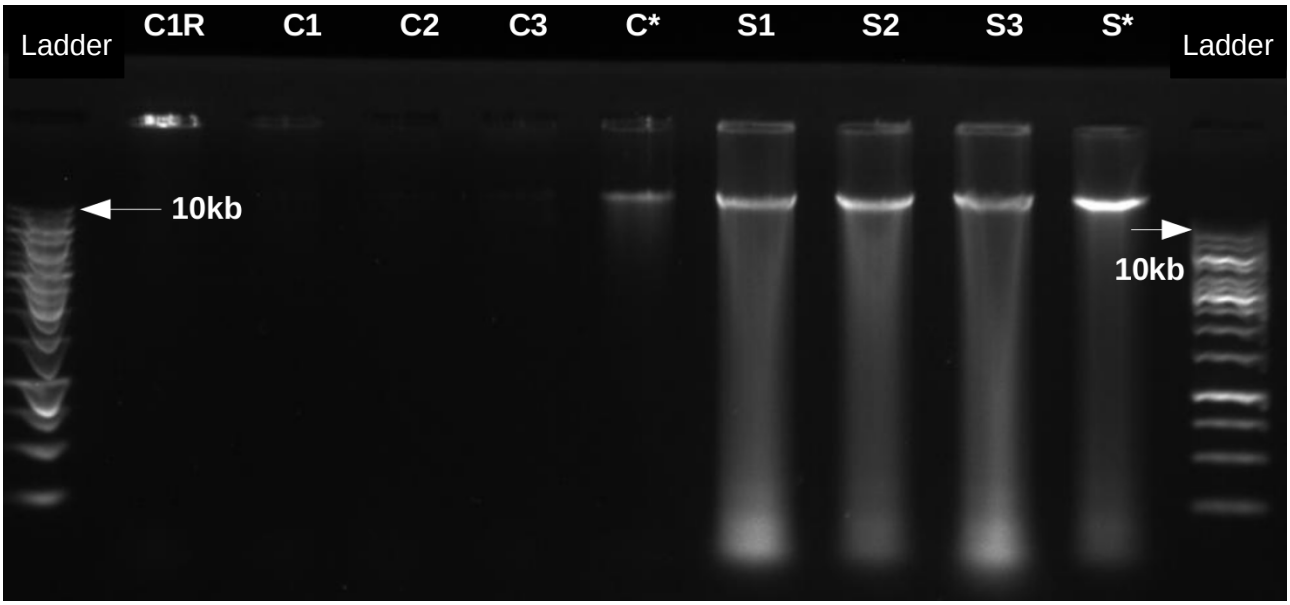


Figure 1 : Photo de la migration d'ADN par électrophorèse sur gel d'agarose 1 % (100 volts pendant 30minutes). Ladder correspond au marqueur de taille 1kb. L'identification des puits correspond à l'identification du tableau I. C : Tampon CTAB. S : Tampon SDS. * : Deux traitements au chloroforme:isoamyl alcool.

Conclusions

Le tampon à base de SDS semble plus adapté pour un protocole d'extraction d'ADN de *Cryphonectria parasitica*. Le gain d'intensité de la bande obtenue par migration de ADN génomique dans les deux cas où l'étape d'extraction d'ADN au chloroforme:isoamyl alcool a été répétée une deuxième fois suggère que la répétition de cette étape peut améliorer la qualité de l'extrait d'ADN.

b) Précipitation de l'ADN

Lors d'une discussion avec Michel Hernould, professeur à l'université de Bordeaux et chercheur dans l'UMR 1332 Biologie et Pathologie du Fruit à l'INRA Bordeaux-Aquitaine, j'ai pu soulever quelques points importants qui permettent d'éviter de précipiter les polysaccharides avec les molécules d'ADN. Deux points importants qui ont été retenus lors de ces échanges sont : 1) précipiter l'ADN avec de l'éthanol 100 % placé préalablement à -80°C et des faibles concentrations en sel (sodium acétate) permet d'éviter la précipitation des polysaccharides 2) Récupérer la pelote d'ADN qui se forme suite à la précipitation permet d'augmenter la pureté de l'extrait d'ADN. Ajouter de l'éthanol à -80°C permet de précipiter plus rapidement l'ADN, et de faciliter l'apparition d'une pelote d'ADN.

Méthode

J'ai donc modifié dans mon protocole l'étape de précipitation d'ADN en proposant d'ajouter 0.01 volume d'acétate de sodium (NaAc) 3M (pH5.2), puis de précipiter la solution d'ADN avec 2,5 volumes d'éthanol à 100 % préalablement placé à -80°C. Après 5 inversions du tube, j'ai récupéré la pelote d'ADN qui se formait pour ensuite procéder aux étapes de lavages dans l'éthanol à 70 %.

Résultats

Identification	Nanodrop		
	260/280	260/230	concentration (ng.uL)
1	2,05	2,1	268,2
2	1,99	2,13	181,4
3	2,04	2,01	226,1
4	1,92	1,48	81,88
5	2,04	2,02	198,4
6	1,8	1,28	38,27
7	1,99	1,69	202,8
8	1,76	0,59	11,76
		Moyennes	
	1,95	1,66	151,10

Tableau II : Tableau récapitulatif de huit extractions d'ADN à partir du protocole de Feehen et al. (2017) modifié avec un tampon SDS, et une précipitation de l'ADN avec 0,01 volume d'acétate de sodium NaAc 3M (pH5.2) et 2,5 volumes d'éthanol 100 %. Pour chaque extraction, la pelote d'ADN a été récupérée avec un cône préalablement coupé.

Les résultats des extractions réalisées avec la précipitation à faible concentration d'acétate de sodium et à large volume d'éthanol 100 % ont montré des résultats plus satisfaisants que ceux présentés en partie 3.a (Tableau II). Le ratio moyen du rapport 260/230 est de 1,66 comparé à 0,78 lors des précédentes extractions, et le ratio moyen du rapport 260/280 est de 1,95 comparé à 1,53. Les profiles de migration sur gel d'électrophorèse se sont révélés êtres très propres, sans aucune présence de « smear » qui témoigne habituellement d'une dégradation de l'ADN (figure 2). Cependant, les petites pelotes récupérées lors de ces extractions ont mené à des quantité d'ADN très faibles estimées à l'aide du Qubit (<0.1µg). Ces faibles quantités d'ADN sont confirmées par la faible intensité des bandes visibles sur le gel d'électrophorèse (figure 2).

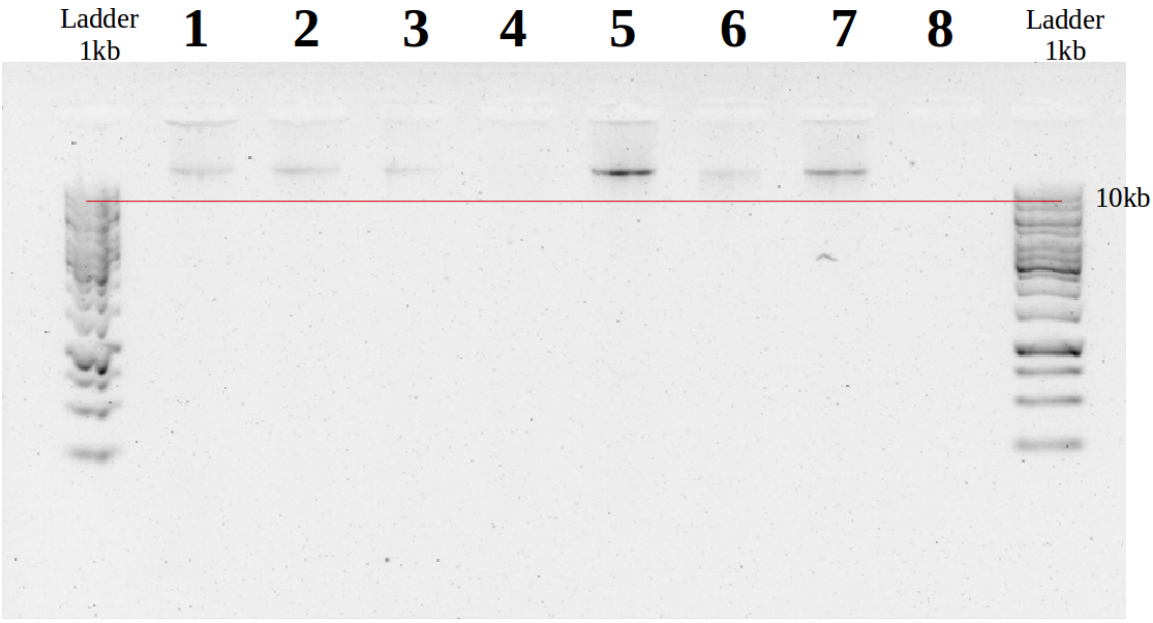


Figure 2 : Photo de la migration d'ADN par électrophorèse sur gel d'agarose 1 % (100 volts pendant 30minutes). Ladder correspond au marqueur de taille 1kb. L'identification des puits correspond à l'identification du tableau II.

Conclusions

La précipitation de l'ADN avec des faibles concentrations en acétate de sodium et un large volume d'éthanol 100 % ont permis d'améliorer grandement la qualité des extraits d'ADN génomiques. Cependant, les très faibles quantité d'ADN obtenues (probablement dû aux très petites pelotes d'ADN récupérées suite à la précipitation) suggèrent de laisser précipiter l'ADN génomique toute la nuit si la pelote n'est pas visible, ou de pooler plusieurs extraits d'ADN ensemble pour augmenter la quantité d'ADN final.

c) Benzil chloride

Méthode

Enfin, une dernière étape principale de la mise au point de ce protocole d'extraction d'ADN génomique a été de rajouter du Benzyl Chloride lors de la lyse cellulaire en début d'extraction. Cette idée a été tirée d'un protocole publié en 1993 pour isoler l'ADN génomique de plantes, champignons et bactéries et qui se concentre sur la dégradation des polysaccharides (Zhu et al., 1993). J'ai donc adapté les quantités décrites à mon protocole, et ajouté 300µL de Benzyl Chloride à la solution de lyse cellulaire à base de SDS. Afin de comparer l'effet de ce nouveau tampon d'extraction, six extractions d'ADN ont été réalisées : trois avec le tampon de lyse cellulaire classique et trois avec le tampon au benzyl chloride. Enfin, comme les quantités d'ADN obtenues précédemment étaient trop faible, j'ai finalement poolé les extraits d'ADN génomique issus des deux jeux de trois extractions.

Identification	Conditions d'extraction	Nanodrop			Qubit	
		260/280	260/230	concentration (ng.µL)	concentration (ng.µL)	quant tot (µg)
C1R	CTAB	1,15	0,31	66,29	NA	NA
C1	CTAB	1,43	0,86	113	3,08	0,08
C2	CTAB	1,46	0,42	56,68	NA	NA
C3	CTAB	1,33	0,58	26,56	NA	NA
C*	CTAB*	1,42	0,9	73,47	10,6	0,27
S1	SDS	1,35	0,8	554,4	NA	NA
S2	SDS	1,63	0,68	284,6	NA	NA
S3	SDS	1,7	0,92	601,8	235,2	5,88
S*	SDS*	1,45	0,71	110,3	110	2,75
Moyennes						
	CTAB	1,36	0,61	67,20	6,84	0,17
	SDS	1,53	0,78	387,78	172,60	4,32

Tableau I : Tableau récapitulatif de neuf extractions d'ADN à partir du protocole de Feehen et al. (2017). Les conditions nommées CTAB correspondent à l'utilisation du tampon de lyse cellulaire dans ce protocole. Les conditions nommées SDS correspondent à l'utilisation d'un tampon de lyse cellulaire à base de SDS (décrit dans l'annexe 1). Les étoiles conrrespondent aux extractions dans lesquelles l'étape 2.1. d'extraction d'ADN au chloroforme:isoamyle alcool a été réalisée deux fois car la solution de lyse était fortement colorée par des pigments.

Résulats

Identification	Conditions d'extraction	Nanodrop			Qbit	
		260/280	260/230	concentration (ng.uL)	concentration (ng.uL)	quant tot (ug)
1	Tampon lyse Classique	1,81	0,91	298,3	11,7	0,468
2		1,77	0,88	395,1	16,1	0,644
3		1,82	1	432,3	16,1	0,644
* (pool 1,2,3)		1,85	0,91	197,1	8,4	1,008
B1	Tampon lyse avec benzyl chloride à 30 % du volume	1,85	1,29	565,6	30,6	1,224
B2		1,88	1,23	523,1	30,8	1,232
B3		1,89	1,15	401,1	29,4	1,176
B* (pool B1,B2,B3)		1,89	1,22	475,8	29,4	3,528
Moyennes						
	Classique	1,80	0,93	375,23	14,63	0,59
	Benzyl Chloride	1,87	1,22	496,60	30,27	1,21

Tableau III : Tableau récapitulatif des six extractions d'ADN à partir du protocole de Feehen et al. (2017) modifié avec un tampon SDS, une précipitation de l'ADN avec 0,01 volume d'acétate de sodium NaAc 3M (pH5.2) et 2,5 volumes d'éthanol 100 %, et deux tampons de lyse cellulaires utilisés. 1, 2 et 3 représentent les résultats de l'extraction avec le tampon de lyse cellulaire classique, B1, B2 et B3 représentent les résultats de l'extraction avec le tampon de lyse cellulaire au benzyl chloride. Les étoiles représentent les résultats obtenus suite au pool des extraits d'ADN obtenus à partir des extractions 1,2 et 3 et des extractions B1, B2 et B3.

Les extractions avec le tampon de lyse cellulaire au benzyl chloride ont permis d'obtenir de meilleurs résultats que les extractions avec le tampon de lyse cellulaire classique (tableau III). En effet, le ratio 260/230 obtenu au nanodrop était en moyenne supérieur (1,22 en moyenne) avec le benzyl chloride, que sans (0,93 en moyenne). Dans les extractions classiques, les pelotes d'ADN n'étaient pas visibles, et l'ADN avait été placé toute une nuit à précipiter à -20°C. Ceci pourrait expliquer les résultats plus faibles obtenus pour les ratios 260/230 des extractions classiques par rapport à celles présentées en partie 3.b). Les profiles de migration sur gel d'électrophorèse se sont révélés êtres propres (figure 3), et les bandes d'intensité moyenne semble être corrélés avec les quantités obtenues d'ADN double brin déterminées à l'aide du Qubit (tableau III).

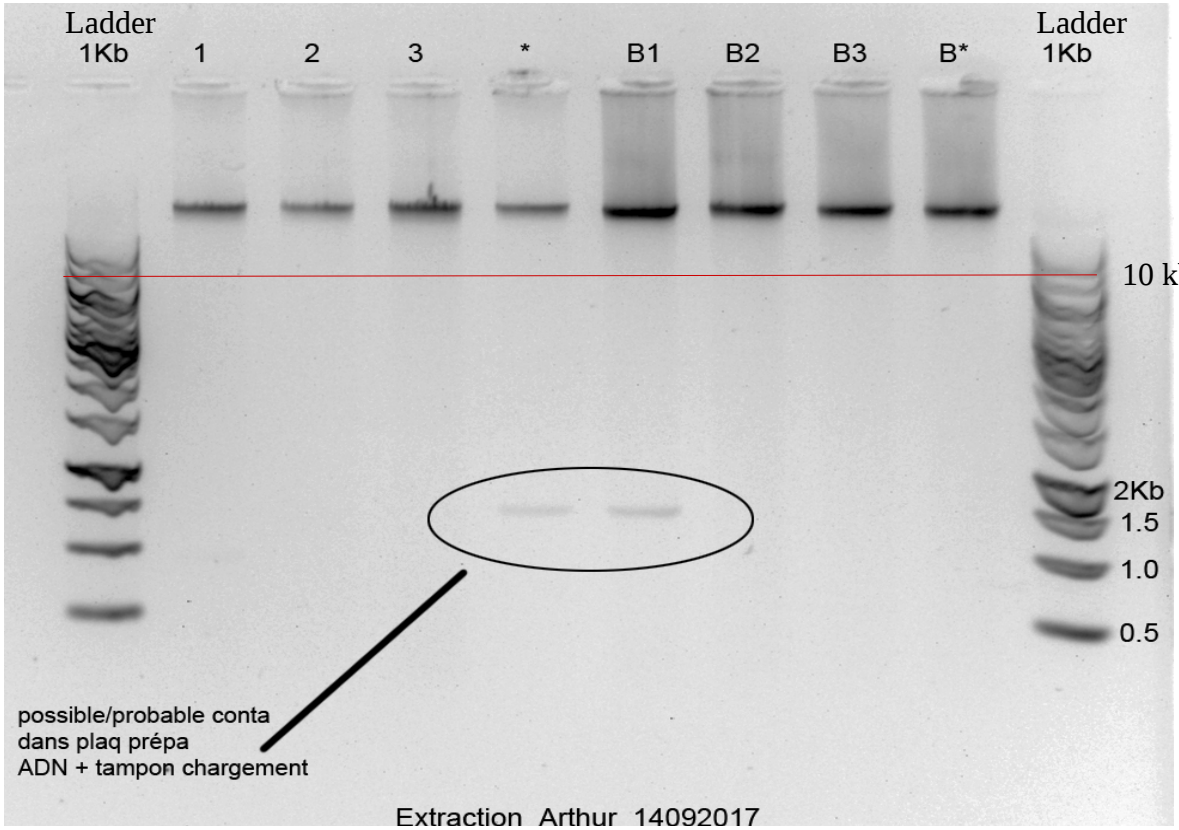


Figure 2 : Photo de la migration d'ADN par électrophorèse sur gel d'agarose 1 % (100 volts pendant 30minutes). Ladder correspond au marqueur de taille 1kb. L'identification des puits correspond à l'identification du tableau III.

Conclusions

L'extraction utilisant un tampon de lyse qui contient du benzyl chloride a été conservée dans le protocole. L'absence de pelotes visibles, ne permettant pas de récupérer l'ADN génomique juste après la précipitation a entraîné des puretés plus faibles. La précipitation de l'ADN pendant toute une nuit semble mener à la précipitation de plus de matériel génomique et d'autres composés moléculaires présents dans la solution et résulte donc en des extraits d'ADN plus concentrés et moins purs. Ces résultats illustrent que le choix du temps de précipitation d'ADN peut être cruciale si le but de l'extraction est d'obtenir à la fois une importante quantité d'ADN et une pureté élevée.

4) Protocole

Ci-dessous est présenté le protocole d'extraction d'ADN génomique mis au point au cours de ma thèse. Ce travail m'a permis de me rendre compte de la grande difficulté d'obtenir des extraits d'ADN de haut poids moléculaire d'une puretés nécessaire au séquençage long brin de troisième génération. Aussi, ces protocoles sont difficilement accessibles et la littérature scientifique est très éparse. Ces protocoles semblent aussi être très dépendant de l'espèce concernée, comme en témoignent les résultats obtenus à partir du protocole utilisé au départ de ce travail (figure 1, tableau I), initialement développé pour les espèces d'*Erysiphe* (Feehan et al., 2017).

Extraction of high-molecular-weight genomic DNA for long-read sequencing of *Cryphonectria parasitica*

Biological material

C. parasitica strains are cultured for six days on PDA medium overlaid with cellophane in a 22°C and 12 hours of light per day chamber. Mycellium must be young and white without stromae.

Material

Day 1 – Genomic DNA Extraction

Preliminary :

- Place SDS and Sarkozyl at 65°C
- Work quickly because samples should stay frozen.
- Place Ethanol 70 % at -20°C

1. Cell lysis

a) Broyer le mycelium de chaque boîte dans l'azote liquide dans un mortier autoclavé préalablement refroidit dans l'azote liquide. Le mycélium doit ressembler à une poudre blanche très fine.

b) Placer la poudre de mycélium dans un tube de 2mL pour micro-centrifugeuse et placer le tube fermer immédiatement dans l'azote liquide pour conservation.

c) Procéder ainsi de suite pour chaque échantillon avec un nouveau mortier pour chaque souche.

d) Sortir les tubes de l'azote et ajouter rapidement 500µL de tampon de lyse SDS à 65°C et vortexer 5-10sec jusqu'à homogénéisation de la solution.

e) Ajouter 200µL de Sarcosyl à 65°C et vortexer très gentiment 1 ou 2 sec. A partir de maintenant ne plus jamais vortexer les échantillons au risque

de briser les molécules d'ADN. Il faut inverser gentiment pour mélanger (1 inverstion max toutes les 3 secondes)

f) Ajouter 300µL de Benzyl Chloride et inverser 3 fois. (dégradation des composants des parois végétales/fungales notamment polysaccharides)

g) Incuber 30 min à 65°C en inversant 3 fois à 10, 20 et 30min.

Volume actuel de la solution 1mL + mycélium broyé

2. DNA Extraction

a) Ajouter environ 800µL de Chloroform:isoamyl alcohol (24:1 v/v) à la solution de lyse et inverser 5 fois. Incuber à température de la pièce pendant 10min en inversant 5 fois à 5 et 10min.

b) Centrifuger 15min à température de la pièce à 14,000G (rcf).

c) A l'aide d'un cone de pipette ou d'une pince autoclavée, retirer l'interface (galette).

d) Centrifuger 15min à température de la pièce à 14,000 G.

e) Prélever le surnageant à l'aide d'un cone de pipette dont le bout est coupé. Ne pas toucher à l'interface. Mieux vaut prendre moins de matériel que d'emporter des débris cellulaires de l'interface. Transférer dans un nouveau tube de 2mL.

f) Ajouter 800µL de Chloroform:isoamyl alcohol et inverser 5 fois.

g) Incuber pendant 5min en inversant 5 fois à 5min.

h) Centrifuger 15min à température de la pièce à 14,000 G.

i) Transférer le surnageant dans un nouveau tube de 2mL.

3. DNA precipitation

(700uL de surnageant est le max pour travailler en tube de 2mL)

a) Ajouter 1/10 volume de sodium acétate NaAc 3M (pH 5.2)

b) Ajouter 1/2 volume de Chlorure de sodium NaCl 5M.

c) Mesurer le volume de la solution actuelle et précipiter l'ADN en ajoutant environ 0.8 volume d'isopropanol 100 % à température de la pièce. (Ça devrait être entre 600µL et 850µL).

b) Inverser 6 fois et centrifuger 15min à température de la pièce à 14,000 G.

c) Lavage : Vider le surnageant, égoutter doucement les tubes et ajouter ~600µL d'éthanol 70 % à -20°C.

d) Lavage : Centrifuger 15min à température de la pièce à 14,000 G. Enlever le surnageant en faisant attention à ne pas perdre le culot et égoutter les tubes.

e) Renouveler le Lavage à l'éthanol 70 %

f) Centrifuger 5sec pour faire tomber l'éthanol des parois et aspirer à la pipette 200µL.

g) Laisser sécher à l'air libre jusqu'à ce que le culot devienne translucide (environ 15min)

h) Ajouter 400µL de TE (10mM TrisCl, 1mM éthylène diamine tetra-acetic, pH 8.0)

i) Placer les tubes à 4°C au frigo pendant la nuit

Day 2 – Genomic DNA Purification

Preliminary :

- Mettre le bain-marie à 37,5°C

1. Remoove RNA contamination

a) Ajouter 12µL de Rnase A (10mg/mL).

b) Inverser 3 fois et centrifuger 3 sec.

c) Placer à 37,5°C pour 2h

d) Toutes les 20 min, tapoter les tubes par le culot 10fois pour remettre en solution la pelotte d'ADN.

b) Ajouter 300µL de Phenol:Chloroforme:Isoamyl Alcohol (25:24:1)

c) Inverser 10 fois et centrifuger 10min à 14,000 G à température ambiante.

d) Récupérer le surnageant sans toucher la surface à l'aide d'un cône à la pointe coupée et placer dans un nouveau tube de 2mL.

e) Ajouter 0.01 volume de sodium acetate NaAc 3M (pH 5.2)

f) Ajouter 2.5 volumes d'ethanol à 100 % à -20°C. Vous devriez voir l'ADN génomic en transparence qui apparaît. Inverser 5 fois et placer à -20°C pour la nuit.

Day 3 – Genomic DNA Ellution

a) Préparer la centrifugeuse à -4°C et de l'éthanol 70 % à -20°C.

b) Centrifuger 30min à 4°C et 14,000 G.

c) Enlever cautionnement le surnageant et égoutter les tubes.

d) Ajouter 450µL D'ÉTHANOL 70 % À -20°C.

E) Centrifuger 5min à 4°C et 14,000 G.

f) enlever le surnageant et égouter

g) Centrifuger 5sec pour faire tomber l'éthanol des parois

h) Aspirer à la pipette 200µL le surplus d'éthanol

i) Laisser sécher entre 15min et 1h jusqu'à ce que la pelotte deviennt transparente

j) Ressuspendre dans 60µL d'H2O miliQ et placer à 4°C pour utilisation dans la semaine. Elluer dans le TE et placer à -20°C pour une conversation plus longue.

5) Séquençage

Un premier séquençage Nanopore (Oxford Nanopore Minion) a été réalisé à la plateforme Génome Transcriptome de Bordeaux (PGTB) avec la collaboration de Christophe Boury, à partir des extraits d'ADN poolés B* (tableau III). Les résultats de ce séquençage sont présentés dans le tableau IV et la figure 4. La quantité de reads exploitables (d'une qualité supérieure à Q9) était très faible (environ 168 000 reads) et correspondait à une couverture moyenne estimé inférieure à 5X. De plus, le pic de longueur des reads obtenus se situe entre 1 et 2kb alors qu'un pic de longueur à environ 20kb est habituellement attendu. Seulement 51 000 reads avaient une taille supérieure à 2kb (environ 3X de couverture). Ces résultats peuvent être expliqués par une quantité d'ADN trop faible. En effet, une étape de sélection de taille a été réalisée à l'aide de billes magnétiques afin de sélectionner seulement les plus grands fragments d'ADN dans l'extrait d'ADN. Suite à cette sélection de taille, environ 1µg d'ADN génomique a pu être séquencé. En combinaison avec une pureté de l'extrait d'ADN faible (tableau III), de nombreux pores de la cellule se sont bouchés au cours du séquençage.

L'affinement de ce protocole a été abandonné suite à la rencontre de Sandrine Cros-Arteil de l'équipe Biologie évolutive des champignons phytopathogènes de l'INRA de Montpellier, qui m'a donné l'opportunité de tester son protocole d'extraction d'ADN de haut poids moléculaire initiallement développé sur *Magnaporthe oryzae*. Ce nouveau protocole a fait l'objet de plusieurs adaptations selon notre matériel disponible au laboratoire et en fonction des résultats obtenus sur *Cryphonectria parasitica*. Ce protocole est présenté ci dessous et fera l'objet une publication en commun avec Sandrine Cros-Arteil, dans laquelle je présenterai les différents tests d'assemblages réalisés à partir des sorties de séquençage obtenues (Oxford Nanopore Minion) à la PGTB.

6) Protocole d'extraction d'ADN (Sandrine Cros-Arteil, en préparation)

High molecular weight DNA extraction from Sordariales mycelium

Author: Sandrine CROS-ARTEIL, INRA, Montpellier
PI: Pierre GLADIEUX, INRA, Montpellier
Adapted for fungi from Appl Environ Microbiol. 2013 Apr; 79(7): 2459–2462.
Contact: sandrine.cros-arteil@inra.fr; pierre.gladieux@inra.fr +33 499 624863

Requirements :
Use only materials and reagents DNase free.
Never vortex and store DNA extracts on ice from step 9.
Avoid DNA freezing-thawing
Notes apportées sur *Cryphonectria parasitica* par Arthur Demené
Détails importants suite aux discussion avec Sandinre Cros-Arteil

NE JAMAIS Vortexer ni Centrifuger à + de 8000rpm. NE JAMAIS pipeter/refouler l'ADN!

Day 1 :
1-Using liquid nitrogen and pre-chilled mortar and pestle, grind 0.2/0.3 g of fresh mycelium into powder, in less than 1 minute. Place the powder into a 15 mL Falcon tube and store in liquid nitrogen or at -80°C.
Note : Sur *C.parasitica,* la réalisation de trois extractions en parallèle du même échantillon *(*3 x ~0,3g de mycelium), on peut espérer en tout entre 15 et 100 ug d'ADN final.
2-Add 5 mL TE 1X pH8 and 300 µL SDS 10 %. Homogenize without vortexing until the complete thawing of the mixture.
3-Add 100 µL of proteinase K at 20 mg / mL. Homogenize by flicking the tube without vortexing and incubate at 37 °C over night in a dry oven.

Day 2 :
4-Centrifuge 3 min at 4000 rpm to eliminate tissues debris. Transfer the supernatant in a new 15 mL Falcon tube.
5-Add 200 µL of SDS 10 % and Homogenize by slow inversion.
6-Add 1,3 mL NaCl 5 M and mix vigorously until obtaining white foam.
7-Incubate 30 min in ice and mix again vigorously until obtaining white foam.
8-Centrifuge 25 min at 4000 rpm at 4 °C (with brake and acceleration on 4). Transfer the supernatant in a new 15 mL Falcon tube with caution.
9-Add 7 mL of phenol/chloroform/isoamyl alcohol 25:24:1. Mix by inversion 200 times and centrifuge 10 min at 5000 rpm at 4 °C. Transfer the upper aqueous phase into a new 15 mL Falcon tube taking care to avoid the aqueous/organic interface. Ne pas hésiter à laisser 1/2cm au dessus de la galette de débrits cellulaires. Warning: store tubes in ice during the aqueous phase transfer. (Empèche la galette de se resolubiliser dans la phase aqueuse)

9-b Ajouter 2% (ici souvent 140uL) de Rnase A (purelink utilisée) et placer 30min à 37°C dans une étuve ou un bain marie.
9-c Repeat the step 9. Warning: store tubes in ice during the aqueous phase transfer.
10-Add 1/10 volume of NaAc 3 M pH = 5.2 and 1 volume of isopropanol and mix gently by inversion. Place the tubes at -20 °C for 3 h.
11-Centrifuge 30 min at 8000 rpm at 4 °C. Eliminate the supernatant by spilling the tube. Briefly spin and eliminate the residual supernatant by pipetting.
12-Rinse the DNA pellet with 6 mL of ethanol 70 % (Note : Bien décoller la pelotte pour la laver sans pipeter/refouler. Tapoter avec le doigt – conseil de Michel Hernould Professeur à l'INRA dans l'UMR 1332 Biologie du Fruit & Pathologie) and centrifuge again 5 min at 8000 rpm at 4 °C. Eliminate the supernatant as above. (Note : Si possible, faire au moins trois lavages à l'ethanol 70% - conseil de Christophe Boury, assistant ingénieur à la plateforme Génome Transcriptome de Bordeaux – UMR 1202 Biogeco) Ici je repasse en tube de 2mL. Dry the DNA pellet at room temperature 5 min by opening the tube. (Note : Plus efficace de laisser la pelote d'ADN sécher toute une nuit pour enelever tous les résidus d'ethanol qui peuvent interférer avec les séquenceurs – conseil d'Olivier Fabreguette, technicien biologie moléculaire à l'INRA dans l'UMR 1065 Santé et Agroécologie du Vignoble)

13-Resuspend the DNA pellet with caution in 500 µL DNase-free water. (Note : Si la pelotte est difficile à resolubiliser même en tapotant avec le doigt, placer les tubes à agiter à 300rpm sur un agitateur à 37°C. Sinon mettre au frigo et attendre plusieurs jours.)

14-Possibility to store DNA extracts at -20 °C and avoid freeze-thaw cycles.

Dosages ici avant la prochaine étape.

Day 3 :

~~15-Add 2 µL of RNase A DNase-free at 10 mg/ mL (0.2 mg / mL final) and incubate 30 min at 37 °C in a dry oven.~~
16-To remove polysaccharides residues ~~(and RNase A)~~, add 500 µL of NaCl 2.4 M (1.2 M final) ~~and complete to 1 mL with DNase-free water and Homogenize.~~
17-Add 1 mL of diethyl ether saturated with DNase-free water (to prepare mix by vortexing 1 volume of diethyl ether and 1 volume of DNase-free water – let the two phases separating and recover the organic upper phase corresponding to diethyl ether saturated in water). (Note : Le diethyl ether saturé en eau est très difficile à pipeter car volatile, il faut aller vite ou s'y reprendre à plusieurs fois pour mettre 1mL.) Mix by inversion 200 times and centrifuge 20 min at 8000 rpm at 4 °C.

Transfer the lower aqueous phase into a new 15 mL Falcon tube. (Note importante : Ne pas récupérer ce qui est précipité au fond du tube (polysacharides). Bien penser à traverser la phase diethyl ether avec un peu d'air dans la pipette pour pouvoir chasser cet air dans la phase aqueuse et ne pas prélever de diethyl ether. Ne pas prélever du tout d'interface. Conseils de Sandrine Cros-Arteil qui a mis au point ce protocole)
18-Add 2.5 volumes of pure ethanol (prechilled at -80°C) and place the tubes at -20 °C over night. Si on voit très bien la pelotte d'ADN se former (inverser le tube doucement peu de fois pour bien la voir). Avec un cône de 1000uL préalablement coupé au cutter/scalpel, récupérer seulement cette pelotte et la placer dans 3mL d'ethanol à 70/75%.
18-b Laver 4 à 5 fois la pelottes dans 3 mL d'ethanol 70% sans jamais centrifuger. (Retirer les 3 mL d'ethanol lentement en laissant tomber la pelotte puis re-ajouter 3mL d'ethanol 70%). Lors du dernier lavage, récupérer la pelotte avec un cône coupé pour repasser en tube de 2mL. Pour les tribles extractions d'ADN a partir de 3 x 0,3 grammes de mycelium, placer toutes les pelottes dans le même tube de 2mL pour augmenter la quantité d'ADN.
19 Dry the DNA pellet at room temperature overnight by opening the tube. Laisser sécher toute une nuit les tubes ouverts sous un papier d'aluminium et sous la hotte allumée.
20 Resuspend the DNA pelette with caution in XXXuL (60uL) of Dnase-free water. Note : Pour les tests (nanodrop, Qbit, gels...) prendre 1uL de la solution et la diluer dans 9uL/99uL d'H2O DNAse free car la solution mère est souvent très concentré. Ne pas oublier de remultiplier les concentrations par 10 après les tests.

SI BESOIN :

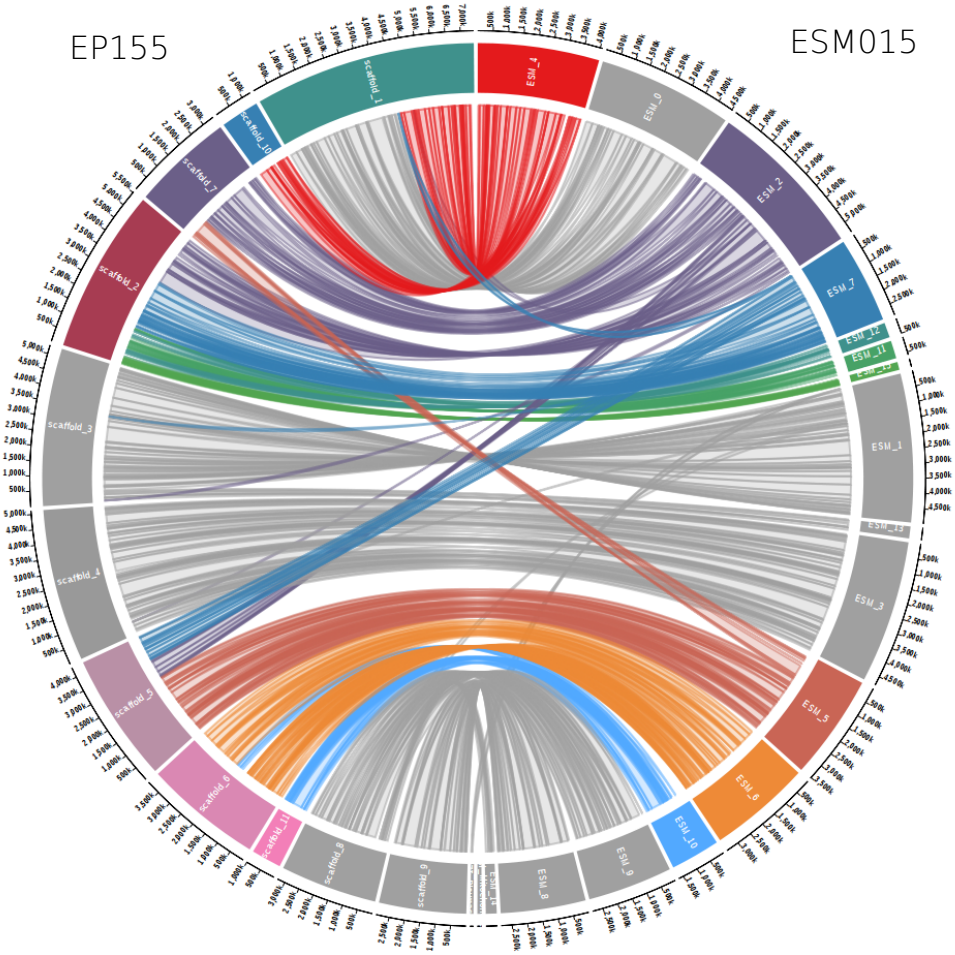
22-Purification on Genomic-tip 100/G - Blood & Cell Culture DNA Midi Kit (25) 13343 QIAGEN according to Special Applications protocol 'Purification of genomic DNA prepared by other methods*' , followed by QIAGEN Genomic-tip procedure - Part II: Genomic-tip Protocol Protocol for Isolation of Genomic DNA from Blood, Cultured Cells, Tissue, Yeast, or Bacteria - QIAGEN Genomic DNA Handbook 08/2001 with at step 5 an over night precipitation of the total DNA in isopropanol at – 20 °C, followed day 4 by a centrifugation 20 min at 8000 rpm to finish by step 6B and resuspension of the total DNA in 100 µL DNase-free water or buffer according to the sequencing company.
Note : At least 2/3 of DNA is lost throughout the final 100/G purification. At the end 8 µg of DNA must be obtained so the quantity of DNA extracted must be adapted to the prerequisite of the sequencing company !

23-*Ajouter une étape de (réparation sans traitement ExoVII + purification AMPure PB Beads) réalisée en plus au cours du workflow de préparation de librairies PacBio RSII ou Sequel. Cette étape permet de combler les

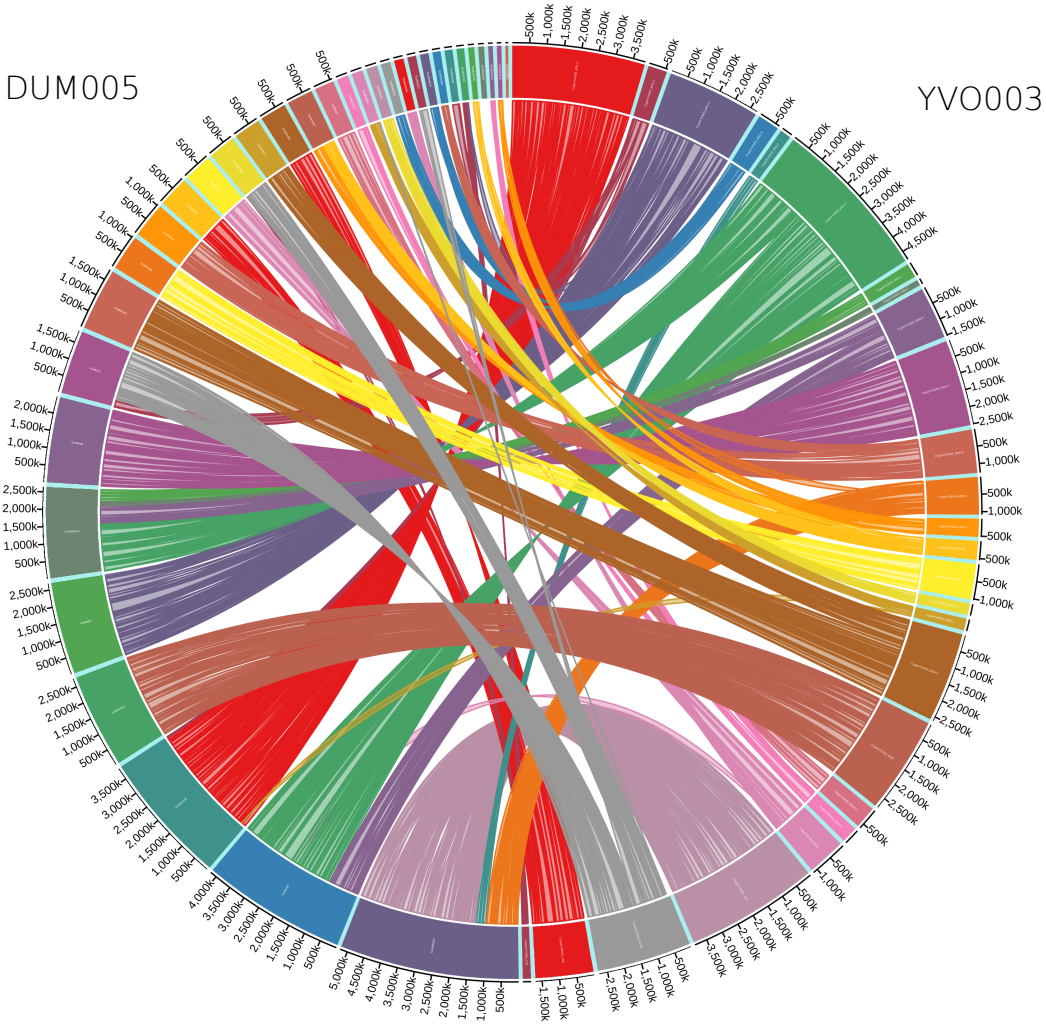
dommages générés sur l'ADN en lien avec le protocole d'extraction utilisé. En effet ce protocole d'extraction génère un ADN fragilisé qui par endroit est simple brin, ce qui entraine la perte de la quasi-totalité de la matrice ADN lors de l'étape 'exonucléases' ExoIII/VII si la double réparation n'est pas réalisée.
Cette adaptation a été développée par la plateforme Genotoul GeT-PlaGe INRA Castanet-Tolosan.

III Annexe 3 : Informations supplémentaires du chapitre 2.

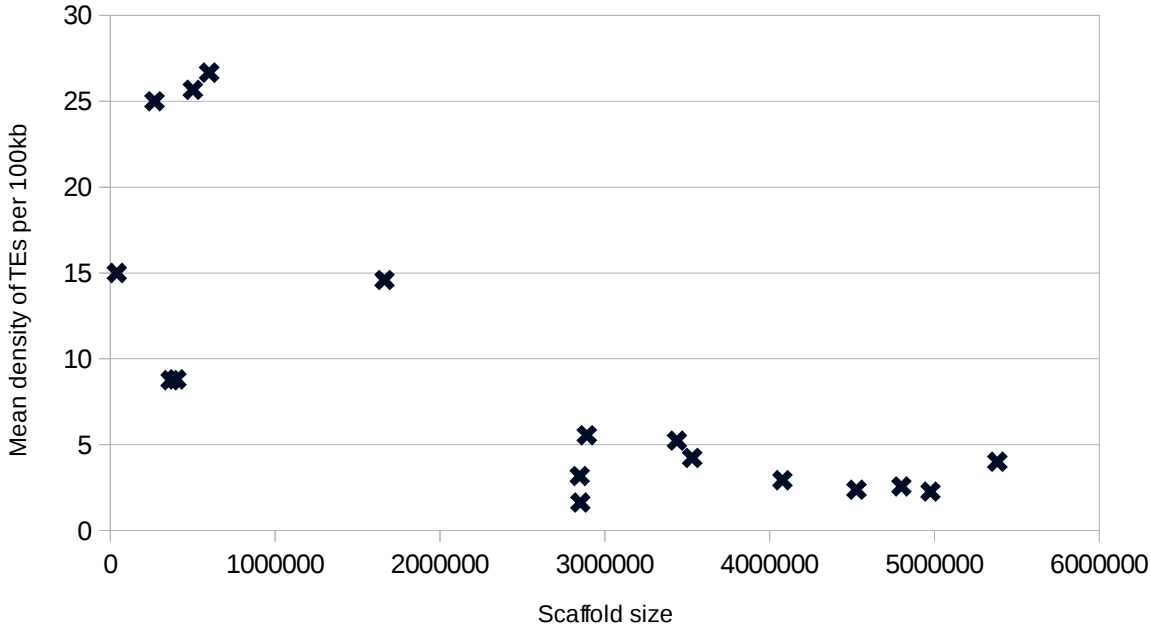
Supplementary figure 1. Visualisation of the alignment between the Japanese isolate ESM015 assembly manualy curated and the EP155 genome assembly using Mummer (available on https://genome.jgi.doe.gov/portal/Crypa2/Crypa2.download.ftp.html). Grey scaffolds and links between scaffolds highlights parts of the genome without rupture of synteny.



Supplementary figure 2. Visualisation of the alignment between the North-Aerican isolate DUM005 genome assembled using HybridSPAdes and the south-eastern French isolate YVO003 pacbio assembly using Mummer.



Supplementary figure 3. Plot of the mean density of TEs (second TEannot curated) per 100kb within the Japanese isolate genome assembly as a function of the scaffolds size.



Supplementary table 1. Summary of the manual curation of the first Teannot output ran on the Japanese isolate genome. Bold names in the new names column highlight the consensus sequences which have been keep, sometimes cut. Two consensus sequences have been replaced by the sequence of a gypsy element detected with a first Teannot on the Chinese isolate genome.

Teannot automatic Name	New Name	action
RXX-LARD_ESMdenovo-B-G10-Map20	**Doubt_LTR_TIR_RNA_DNA**	Keep – too complicated – chimeric
DTX-comp_ESMdenovo-B-G15-Map5_reversed	**DT_Supposed_Mariner_With_Endonuclease**	Keep – longuest Mariner like
DTX-comp_ESMdenovo-B-G22-Map4	*DT_Supposed_Mariner_With_Endonuclease*	supress
DTX-comp_ESMdenovo-B-G24-Map19	**DT_Supposed_Mariner_With_Endonuclease_CLEAN**	keep
DTX-comp_ESMdenovo-B-G26-Map17_reversed		supress no full length copy
RXX-LARD_ESMdenovo-B-G27-Map3	**RXX_LARD_onlyIntegrase**	keep
noCat_ESMdenovo-B-G2-Map20		supress no full length copy
RLX-incomp_ESMdenovo-B-G31-Map3		supress no full length copy
noCat_ESMdenovo-B-G5-Map20		supress no full length copy
PotentialHostGene_ESMdenovo-B-G64-Map3	**PotentialHostGene_ESMdenovo-B-G64-Map3**	keep
noCat_ESMdenovo-B-G6-Map15	Incomplete gypsy – Gypsy domains but no LTR	suppress
noCat_ESMdenovo-B-G7-Map20		supress no full length copy
RLX-incomp_ESMdenovo-B-P13.7-Map3	**RLC_COPYA_clean**	keep
DTX-comp_ESMdenovo-B-P15.5-Map7	*DT_Supposed_Mariner_With_Endonuclease*	supress
DTX-comp_ESMdenovo-B-P2.17-Map8_reversed	**DT_Supposed_Mariner_NoClearTransposase_With_Endonuclease**	keep
DHX-comp_ESMdenovo-B-P24.22-Map4	**DH_Helitron_like_remooved_repeat**	Cut : 0-7500
DTX-incomp_ESMdenovo-B-P3.16-Map4	*DT_Supposed_Mariner_With_Endonuclease*	supress
RLX-incomp_ESMdenovo-B-P6.13-Map4_reversed	*RL_COPYA_like_Missing_AP*	supress
DTX-comp_ESMdenovo-B-P8.11-Map5_reversed	**Crypt1_Transposon_Verif_Blast**	keep
RXX_ESMdenovo-B-R14-Map3_reversed	Uncomplete_1500copies_RichAT	**Replaced by GYPSY from XIM9508**
noCat_ESMdenovo-B-R25-Map12		**Replaced by GYPSY from XIM9508**
RLX-incomp_ESMdenovo-B-R26-Map6	**RL_GYPSY_like_missing_GAG_AP**	keep
RLX-incomp_ESMdenovo-B-R34-Map3_reversed	**RL_COPYA_like_Missing_AP**	keep
DTX-incomp_ESMdenovo-B-R3-Map6_reversed	*DT_Supposed_Mariner_With_Endonuclease*	supress
RXX-LARD_ESMdenovo-B-R51-Map10	**RXX-LARD_NomoreInformation**	keep
DTX-comp_ESMdenovo-B-R5-Map3_reversed	**DTX_comp_Mariner_Like**	keep
DTX-incomp_ESMdenovo-B-R72-Map9_reversed	*DT_Supposed_Mariner_With_Endonuclease*	supress
DTX-incomp_ESMdenovo-B-R7-Map3_reversed	**DTX_Mariner_Like_No_Transposase**	Cut 430 → end
DXX_ESMdenovo-B-R96-Map4	Only one copy	suppress

Grouper
Piller
Recon

Italic : Suppressed because multiple identical elements, detected by multiple tools