



Master 2 : Biodiversité, Ecologie et Evolution  
Parcours Ecosystèmes et Anthropisation

**Auteur : Virgile Ballandras**

Délimitation d'espèces par différentes approches  
d'analyses moléculaires au sein du complexe  
*Cacopsylla pruni*.

Encadrant : Nicolas Sauvion  
Tuteur académique : Erick Campan  
Examinatrice : Cécile Ben

Stage effectué au sein de l'Institut de la Santé des plantes de Montpellier,  
dans les locaux du CIRAD de Baillarguet  
Du 1<sup>er</sup> mars au 31 août 2021

## Remerciements

Je souhaite particulièrement remercier Nicolas SAUVION pour son implication et sa disponibilité en tant que tuteur jusqu'à la fin de mon stage ainsi que pour les nombreux échanges aussi instructifs que constructif que nous avons eu durant cette période.

Je souhaite également remercier mon co-stagiaire Walid KANDOUCI pour son aide et le soutien mutuel que nous nous sommes apportés pendant ce stage.

Je désire aussi remercier Frédéric DEDIEU ainsi que toute l'équipe FORISK, Fabienne RIBEYRE, Bernard DUFOUR, Leïla BAGNY-BEILHE, Lorelei GUERY, Gerben Martijn TEN HOOPEN et Jean-Benoit MOREL, directeur de l'unité PHIM pour leur accueil chaleureux et bienveillant.

Enfin je remercie Erick CAMPAN et mes enseignants de l'université Paul Sabatier et de l'ENSAT pour leur encadrement durant cette année.

## Résumé

Le psylle *Cacopsylla pruni* (Hemiptera : Psyllidae) est un insecte vecteur de l'agent pathogène '*Candidatus Phytoplasma prunorum*' responsable d'une maladie des abricotiers et des pruniers. Des études antérieures avaient émis l'hypothèse que *C. pruni* était en fait un complexe de deux espèces, provisoirement nommé A et B. Cependant, cette conclusion était basée sur l'analyse d'un nombre limité de populations du sud de la France. Or, des travaux publiés cette année ont permis d'avoir une vision beaucoup plus précise des aires de distribution des deux groupes à l'échelle du paléarctique. Les individus du groupe A sont prédominants dans la région méditerranéenne mais sont aussi retrouvés dans le centre ouest de la France, voire plus au nord dans la région parisienne. Les individus du groupe B semblent plus adaptés au climat continental. Ils sont majoritaires voire en allopatrie stricte dans le centre et l'est de la France, en Allemagne et jusqu'en Turquie. En France, malgré une large zone de chevauchement des aires de répartition, les deux espèces ne s'hybrident quasiment pas. Cette barrière reproductive très forte n'exclurait cependant pas l'existence de flux de gènes au sein de chaque groupe, et donc l'existence de sous-groupes ou sous-espèces.

Les travaux présentés dans cette étude s'appuient sur l'exploitation d'un jeu de données couvrant une aire géographique plus étendue et plus proche de l'aire de répartition des psylles connue à ce jour. Notre étude repose sur l'exploitation de différents types de marqueurs (microsatellites, COI, ITS2) et la comparaison de différentes méthodes d'analyses (analyse multivariée, clustering, réseaux d'haplotypes). Nous confirmons qu'il existe deux espèces au sein du complexe *C. pruni*. Toutefois, nos résultats montrent que chaque espèce présente une structure des populations différente. L'espèce A semble se répartir de manière homogène à l'échelle d'un large territoire couvrant le sud-est de l'Espagne et la France, à l'exception du massif central, où aucun individu A n'a été observé. L'espèce B se répartit sur toute l'aire d'étude (de l'ouest de la France à la Turquie) mais semble se subdiviser en deux ou trois groupes génétiques, entre lesquels des flux de gènes sont encore visibles (y compris entre populations très éloignées) ce qui pourrait être le signe d'une partition en cours mais trop récente pour conduire à une ou de nouvelles espèces. Nous discutons l'intérêt de ces résultats dans le cadre d'une approche dite de taxonomie intégrative.

## Abstract

The psyllid *Cacopsylla pruni* (Hemiptera: Psyllidae) is the only known vector of the pathogen ‘*Candidatus* Phytoplasma prunorum’, responsible for apricots and plum trees disease. Earlier studies hypothesised that *C. pruni* was a complex of two cryptic species, tentatively named A and B. However, this conclusion was based on the analyses on a limited number of populations from the south of France. A work published this year has provided a much more accurate picture of the ranges of the two groups across the Palearctic. Group A individuals are dominant in the Mediterranean region but they are also found central-western France and even further north near Paris. Group B individuals seems to be better adapted to the continental climate and are dominant in the east of France. They are in the majority or even in strict allopatry in central and eastern France, in Germany and as far as Turkey. In France, despite a large area of overlap in distribution, the two species hardly hybridise. This very strong reproductive barrier does not, however, rule out the existence of gene flow within each group, and thus the existence of subgroups or subspecies.

The work presented in this study is based on the exploitation of a dataset covering a wider geographical area and closer to the distribution of psyllids known to date. Our study is based on the use of different types of markers (microsatellites, COI, ITS2) and the comparison of different analysis methods (multivariate analysis, clustering, haplotype networks). We confirm that there are two species within the *C. pruni* complex. However, our results show that each species has a different population structure. Species A seems to be homogeneously distributed over a large area covering south-eastern Spain and France, with the exception of the Massif Central, where no A individuals have been observed. Species B is distributed over the whole study area (from western France to Turkey) but seems to be subdivided into two or three genetic groups, between which gene flow is still visible (including between very distant populations), which could be a sign of an ongoing partition but too recent to lead to new species. We discuss the interest of these results in the context of an integrative taxonomy framework.



## Table des matières

Introduction :	1
I. Structure d'accueil et organisme finançant le stage	1
I.A. PHIM, PRISM et FORISK	1
I.B L'INRAE	2
I.C. Le GIS Fruits	2
II. Histoire, contexte et spécificités du pathosystème	2
II.A. L'ESFY	2
II.B Biologie du vecteur et statut taxonomique	3
II.C Génétique des populations et frontières spécifiques à large échelle spatiale	7
III. Objectifs du stage	8
Matériels et méthodes	9
I Analyses à partir des marqueurs microsatellites	10
I A. Présentation du jeu de données	10
I B. Clustering sans information spatiale	11
I C. clustering avec information spatiale	13
I D. Indicateurs statistiques	15
II Analyses à partir des séquences des gènes COI et ITS	16
III Comparaison des modèles et représentation des groupes	17
Résultats	18
I Analyses sur les données des marqueurs microsatellites	18
II Analyses sur les données des séquences ITS et COI	24
Discussion	26
Deux groupes génétiques fortement distincts	26
Des flux de gènes intraspécifiques plus ou moins importants	27
Conclusion	29
Perspectives	30
Bibliographie	31

Annexe 1 : Glossaire .....	35
Annexe 2 : Plantes hotes .....	38
Annexe 3 : Insectes vecteurs .....	39
Annexe 4 : Cartes d'occurences .....	47
Annexe 5 : <i>adeget</i> .....	35
Annexe 6 : STRUCTURE .....	53
Annexe 7 : TESS .....	56
Annexe 8 : GENELAND .....	62
Annexe 9 : <i>genepop</i> .....	66
Annexe 10 : Séquences ITS et COI .....	68
Annexe 11 : Morphologie .....	71

## Introduction :

### I. Structure d'accueil et organisme finançant le stage

Plusieurs organismes de recherche entrent dans l'encadrement de mon stage. L'unité PHIM (Plant Health Institut of Montpellier) est l'unité au sein de laquelle s'est déroulé mon stage. L'INRAE, (Institut National de la Recherche pour l'Agriculture, l'alimentation et l'Environnement) est l'organisme dont je dépendais administrativement. Ma bourse de stage a été attribuée par Groupement d'Intérêt Scientifique de la filière Fruits (GIS Fruits).

Enfin, j'ai été accueilli dans les locaux du centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD) sur le campus international de Baillarguet.

#### I.A. PHIM, PRISM et FORISK

L'UMR PHIM existe depuis le premier janvier 2021, suite à la fusion des unités BGPI (Biology and Genetics of Plant-Parasite Interactions), IPME (Plant Microorganism Interactions and Environment) et Bioagresseurs (Bioagressors: risk assessment and control). PHIM accueillait jusqu'à 200 personnes, temporaires et permanents, durant le mois de mars 2021 ce qui en fait un pôle majeur de recherche en pathologie végétale en France et à l'international.

PHIM regroupe des agents du CIRAD (46%), de l'INRAE (25%) de l'IRD (Institut de la recherche pour le développement) (29%) ainsi que deux agents de l'université de Montpellier et deux agents de Montpellier SupAgro. L'unité est sous la tutelle de l'Université de Montpellier 2 (Montpellier Université d'excellence, MUSE) et de l'Institut National d'Etudes Supérieures Agronomiques (Supagro) de Montpellier.

PHIM est divisé en 4 pôles<sup>1</sup> de recherche qui visent à analyser les interactions plantes–pathogènes, les interactions virus – vecteurs – plantes et à comprendre le phytobiome et les épidémies.

Mon stage s'est déroulé au sein du pôle PrisM (Plant pathogen and pest : dynamics and RISk Managment), dans l'équipe FORISK (FORecasting epidemic rISK). Cette équipe étudie comment les facteurs environnementaux affectent la démographie et les performances des bioagresseurs des plantes (virus, champignons, insectes ravageurs et vecteurs). J'ai réalisé mon stage sous la tutelle de Nicolas Sauvion, Entomologiste-Epidémiologiste, Ingénieur de Recherche et agent INRAE. À ce titre, je dépendais également de l'INRAE.

---

<sup>1</sup> <https://umr-phim.cirad.fr/>

## I.B L'INRAE

Le 1<sup>er</sup> janvier 2020, l'INRA (Institut National de la Recherche en Agronomie) a fusionné avec l'IRSTEA (Institut National de Recherche en Sciences et Technologies pour l'Environnement et l'Agriculture) pour former l'INRAE, institut de recherche sous la cotutelle du Ministère de l'Agriculture et de l'Alimentation et du Ministère de l'Enseignement Supérieur, la Recherche et l'Innovation.

À l'heure actuelle, il emploie plus de 10 000 agents dans 18 centres répartis sur toute la France ainsi qu'aux Antilles et en Guyane. L'INRAE se positionne sur le plan international comme un acteur majeur sur des enjeux tels que la sécurité alimentaire et nutritionnelle, l'agro-écologie, la gestion des ressources naturelles et des écosystèmes, l'érosion de la biodiversité, et les risques naturels.

### I.C. Le GIS Fruits

Afin de répondre aux questions environnementales tout en relevant le défi de garder une filière compétitive et productive, la filière Fruits a mis en place le GIS Fruits en 2012. Cet organisme remplit plusieurs objectifs dont celui de regrouper les différentes actions et programmes de recherche développements et formations scientifiques au service de la filière Fruits. Cela se traduit par le financement de projets de recherche en coordination avec l'INRAE. A travers l'objectif de formation, le GIS Fruits propose tous les ans des bourses de financements de stages de Master 2. En 2020, le GIS Fruits a accordé 14 bourses de stage de Master 2<sup>2</sup>. Il s'agit du programme dont j'ai bénéficié durant mon stage.

## II. Histoire, contexte et spécificités du pathosystème

### II.A. L'ESFY

#### **II.A.1. La maladie et les symptômes**

L'European stone fruit yellows (ESFY) est une maladie présente en Europe et recensée dans 15 pays de l'Union Européenne (Steffek et al., 2012). Durant les années 1990-2000, le progrès des techniques d'analyses génétiques a permis d'identifier l'agent causal de l'ESFY. Il s'agit d'une bactérie sans paroi de la classe des Mollicutes, connue aujourd'hui sous le nom '*Candidatus Phytoplasma prunorum*' (CaPp) (Danet et al., 2011; Lee et al., 2000).

Les *Prunus* sauvages, en particulier le prunelier, *Prunus spinosa*, (Annexe 2), jouent un rôle de **réservoir**<sup>3</sup> pour le phytoplasme CaPp au sein des agrosystèmes (Marie-Jeanne et al., 2020; Popescu et Caudullo, 2016). Visuellement, la maladie est quasi indétectable dans les massifs ou les haies de prunus sauvages. En revanche, les abricotiers (*Prunus armeniaca*) et les pruniers (*Prunus domestica*) présentent des symptômes caractéristiques de la présence du

---

<sup>2</sup> <https://www.gis-fruits.org/Actions-du-GIS/Bourses-de-Master-Fruits/Bilan-des-stages-2020>

<sup>3</sup> Les termes en **gras et italique** sont expliqués dans le glossaire en annexe 1

phytoplasme tels que le raccourcissement des entre-nœuds, un feuillage précoce durant la floraison, un enroulement chlorotique des feuilles ou encore une chute prématurée des fruits (Sauvion et al., 2012; Sauvion, 2020a). Ces dérèglements entraînent la mort des arbres à plus ou moins long terme (Jarausch et al., 2013; Sauvion, 2020a; Thébaud et al., 2009)

### **II.A.2. Mode de transmission du phytoplasme CaPp**

Les phytoplasmes se transmettent uniquement par voie horizontale, c'est-à-dire qu'ils passent d'une plante à l'autre par l'intermédiaire d'un vecteur ou *via* la propagation végétative (Arnaud et al., 2013). Les insectes vecteurs de phytoplasmes appartiennent tous à l'ordre des Hemiptera, des insectes piqueurs-suceurs de sève (Annexe 3). Ils transmettent les phytoplasmes selon le mode propagatif et persistant (Brown, 2016). Cela signifie que le pathogène est acquis pendant la période d'alimentation de l'insecte sur sa plante hôte et qu'un certain temps de latence est nécessaire pour qu'il soit ensuite inoculé à une plante saine. Chez CaPp, ce temps de latence dure au minimum 3 semaines. Cette durée correspond au temps nécessaire au pathogène pour se multiplier dans son vecteur (Thébaud et al., 2009). Le phytoplasme est capable de migrer de la lumière intestinale jusque dans les glandes salivaires (Brown, 2016). La bactérie pourra alors être inoculée durant la phase d'alimentation du vecteur sur une plante saine. Dans le cas du phytoplasme du prunier, un insecte vecteur peut potentiellement infecter autant de plantes qu'il en visite pour se nourrir. Cette période durant laquelle le vecteur reste infectieux est appelée période de rétention (Herrbach et al., 2013).

## **II.B Biologie du vecteur et statut taxonomique**

### **II.B.1 La biologie de l'insecte et son rôle dans la transmission du phytoplasme**

En 1998, Carraro *et al.* ont montré que CaPp était transmis par le psylle *Cacopsylla pruni* (Scopoli 1763) (Hemiptera : Psyllidae) (Carraro et al., 1998). Les psylles appartiennent au sous-ordre des Sternorrhyncha qui comprend notamment les pucerons et les cochenilles (Annexe 3). Plusieurs espèces du genre *Cacopsylla* transmettent des phytoplasmes aux arbres fruitiers (pruniers, poiriers, pommiers) mais chaque espèce fruitière à son cortège spécifique de vecteurs (Jarausch et al., 2019; Ouvrard, 2021)(Annexe3).

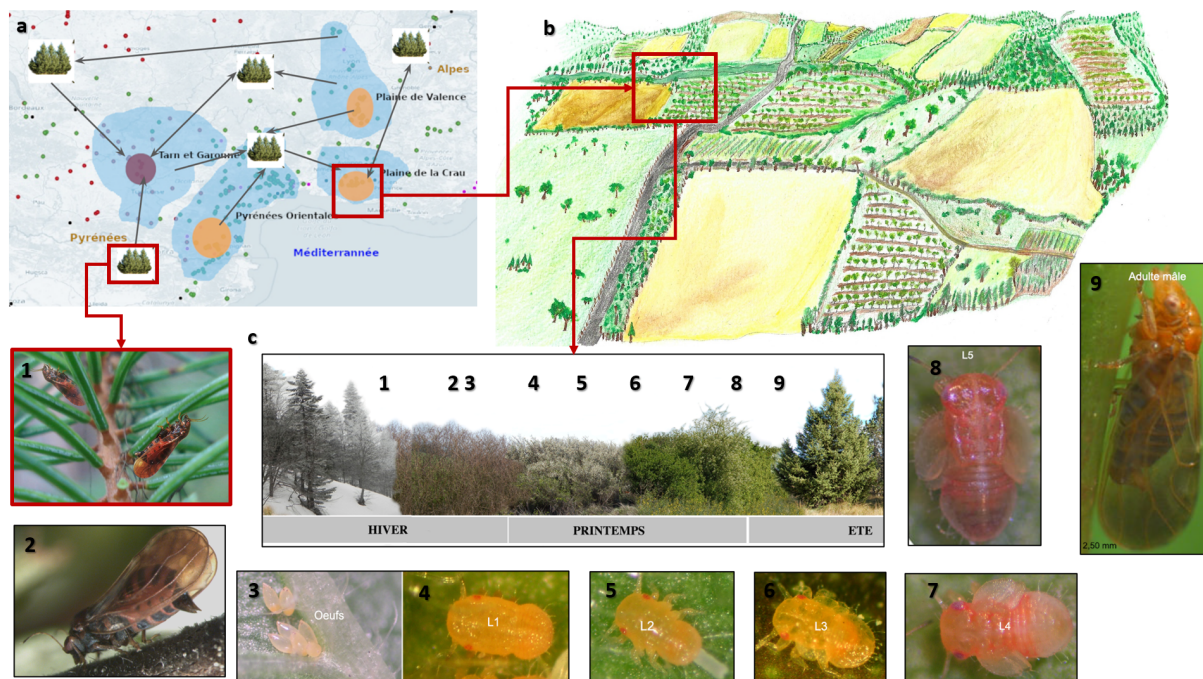


Figure 1: (Fig R6) Cycle biologique de *Cacopsylla pruni*. a : Migrations putatives à l'échelle régionale. Ces migrations entre plantes hôtes sont représentées par les flèches. Les points représentent les données d'occurrences des psylles (Sauvion, 2020b). Les cercles de couleur représentent les bassins de production de fruits (prunes en violet et abricots en orange) et les formes bleues représentent les populations hypothétiques des psylles. b : Représentation d'un agrosystème au sein duquel les psylles migrent et se développent ; c : développement des insectes à l'échelle de la plante hôte. Les numéros 1 à 9 en haut à gauche des photos d'insectes correspondent à ceux de la frise saisonnière. 1 : Psylle adultes sur conifères avant la première migration annuelle. 2 : Femelle en position de ponte. 3 à 8 : développement larvaire sur *Prunus*. 9 : Imagos avant la deuxième migration annuelle. Crédit photos, N. Sauvion, INRAE©.

*Cacopsylla pruni* est un insecte univoltin dont l'adulte meurt après la période de reproduction (Thébaud et al., 2009). A priori, cela empêche le chevauchement de génération, un élément caractérisant une population idéale selon le modèle **Hardy-Weinberg** (H-W) (Mayo, 2008).

Les individus de l'espèce *C. pruni* effectuent deux migrations par an (Fig 1). Une migration a lieu à la fin de l'hiver-début du printemps des conifères, où les psylles ont trouvé refuge depuis l'été de l'année précédente, vers les *Prunus*. Les accouplements se font rapidement et les premières pontes sont observables quelques jours plus tard. Après l'éclosion, les psylles connaissent un développement larvaire comprenant 5 stades nommés L1 à L5 (Fig 1, photos 2 à 8). A la fin du printemps, les psylles effectuent leur **mue imaginale** qui marque l'aboutissement des stades larvaires. Les imagos qui en résultent émergent, fin mai-juin, pour effectuer une migration en direction des conifères. Le départ des insectes vers ces plantes dites refuges se fait au plus tard quelques jours après l'émergence des imagos.

Le phytoplasme peut être acquis par les larves de derniers stades (Fig 1, photo 8) ou les imagos (photo 9) au cours de leur alimentation dans la sève phloémienne. Il est peu probable que des insectes porteurs de CaPp le transmettent à des arbres voisins avant leur départ vers les conifères (Marie-Jeanne et al., 2020; Thébaud et al., 2009). En effet, ils ne restent pas assez longtemps sur place pour multiplier la bactérie et donc pour être infectieux. C'est seulement

l'année suivante, à leur retour d'hivernage, qu'ils seront en capacité d'inoculer le phytoplasme à des plantes saines (Marie-Jeanne et al., 2020). Ces résultats montrent l'importance des déplacements de *C. pruni* dans la dissémination de CaPp. A l'heure actuelle demeure une question essentielle pour comprendre la dynamique de l'épidémie : à quelle(s) distance(s) et dans quelle(s) direction(s) migrent chaque année les psylles.

Ce problème a été toutefois complexifié par la découverte de deux groupes génétiquement distincts au sein du taxon *C. pruni* (Sauvion et al., 2007). Cette information laisse entrevoir un facteur de variation du processus de dissémination du phytoplasme.

### II.B.2 Un complexe spécifique

Le genre *Cacopsylla* comprend environ 465 espèces répertoriées et validées (Ouvrard, 2021). Toutefois, ce nombre est probablement sous-estimé. En effet, les premiers entomologistes ayant décrit ces espèces, comme Scopoli par exemple, se sont fondés sur la plante hôte puis sur des critères morphologiques tels que la forme ou la couleur des ailes. L'utilisation de séquences d'ADN permet aujourd'hui de mieux différencier les espèces, même très apparentées mais cette approche a ses limites, en particulier pour des espèces sœurs/cryptiques.

De manière tout à fait fortuite, N. Sauvion et son équipe ont découvert que *C. pruni* était un complexe d'au moins deux espèces (Sauvion et al., 2007). Dans le but d'analyser la structure des populations de *C. pruni*, neuf marqueurs microsatellites avaient été développés. Des analyses sur quelques populations du sud de la France ont alors révélé l'existence de deux groupes génétiquement très différenciés (appelés provisoirement A et B) en sympatrie à cette échelle géographique. A posteriori, des caractères morphologiques ont été recherchés pour différencier ces deux groupes. Seules les femelles sont distinguables d'après un critère très ténu au niveau des genitalia (Annexe 11). Les individus des groupes A et B peuvent s'accoupler *in natura* mais les hybrides sont très rares ( $<1/1000$ ) (Peccoud et al., 2018).

Des marqueurs mitochondriaux (COI : cytochrome oxydase I) et nucléaires (ITS2 : Internal Transcribed Spacer II) permettent aussi de parfaitement dissocier les deux groupes. Des amorces spécifiques ont été développées pour caractériser rapidement, en une PCR, les individus de chaque groupe (Peccoud et al., 2013). Cette démarche visant à différencier les espèces en utilisant plusieurs critères (écologiques, morphologique, nucléaires) s'inscrit dans une approche dite de **taxonomie intégrative**.

La taxonomie intégrative s'attache à prendre en compte un maximum de critères dans le but de départager des groupes d'individus et d'interpréter leur histoire évolutive (Padial et al., 2010). Cette discipline vise à intégrer les différences interspécifiques établies sur des

critères purement morphologiques dans une approche holistique de délimitation spécifique, en s'aidant d'autres critères (moléculaires, biologiques, physiologiques). Certains critères morphologiques, parfois sous-estimés, sont même réévalués comme les différences observées au niveau de l'édéage chez le genre *Aphrodes* (Hemiptera : Cicadellidae) (Bluemel et al., 2014). La précision des résultats et le progrès technique des analyses moléculaires incitent les chercheurs à se reposer davantage sur leurs résultats mais elles restent encore aujourd'hui coûteuses et non sans failles. L'association de ces techniques à d'autres critères, comportementaux, comme les appels des mâles *Cicadellidae* (Bluemel et al., 2014) ou encore écologiques mène à des délimitations spécifiques toujours plus pertinentes.

La standardisation des PCR diagnostiques ITS2 sur *C. pruni* a permis de typer des milliers d'individus collectés dans plus de 700 localités essentiellement en France mais aussi dans d'autres pays d'Europe (de l'Espagne à la Turquie). Ces informations ont été renseignées dans une base de données accessible en ligne (Sauvion, 2020b) et ont permis d'établir des cartes d'occurrence des groupes A et B à l'échelle du paléarctique occidental (Sauvion et al., 2021) (Annexe 4). Ces cartes montrent des zones d'allopatrie stricte et également des aires de sympatrie où les deux groupes cohabitent (Marie-Jeanne et al., 2020; Peccoud et al., 2018).

L'hypothèse qui semble la plus plausible concernant la formation de ces deux groupes est celle d'un processus de spéciation allopatrique. La barrière des Pyrénées aurait pu stopper les **flux de gènes** entre les deux populations situées de chaque côté de la chaîne de montagnes. Les cartes de distributions actuelles des deux groupes A et B (Annexe 4) indiqueraient que les individus de chaque groupe auraient pu ensuite se disperser vers l'Est. Les individus A se seraient limités au pourtour méditerranéen alors que les individus B, mieux adaptés au climat tempéré continental, se retrouvent plus au nord et dans les pays d'Europe centrale. Des études phylogéographiques sont en cours pour confirmer ce scénario.

La distribution sympatrique de ces deux groupes serait le fruit d'un contact secondaire récent. Les deux groupes ne sont pas en compétition et aucun ne supprime l'autre bien que les psylles des deux groupes exploitent la même plante hôte et se reproduisent au même moment (Sauvion N. com. pers). Les hybrides sont rares mais existent. Ces observations indiquent que les barrières d'espèces sont fortes mais pas complètes (Peccoud et al. 2018). Autrement dit, la frontière spécifique présenterait encore une certaine porosité. Cela pose la question d'une potentielle existence de groupes ou de sous-groupes jusque-là non détectés par les analyses. Rappelons que les premières analyses génétiques avec des marqueurs microsatellites (les plus représentatifs) ont été conduites sur des populations restreintes à une aire géographique relativement petite. Il faudrait donc pouvoir analyser de nouvelles populations plus



représentatives de la distribution connue à l'heure actuelle pour confirmer l'hypothèse de l'existence des deux groupes et uniquement deux groupes à ces échelles géographiques.

### II.C Génétique des populations et frontières spécifiques à large échelle spatiale

Au début de mon stage, l'hypothèse était que *C. pruni* était un complexe de deux espèces cryptiques (Marie-Jeanne et al., 2020; Peccoud et al., 2018, 2013). Cependant, des travaux publiés cette année montrent que les aires de distributions de ces psylles sont très grandes (Sauvion et al., 2021). Etant donné leurs capacités de dispersion (environ 50 km d'après Marie-Jeanne et al., 2020), il serait possible que des structures particulières puissent apparaissent sous l'effet d'isolements reproductifs par la distance notamment. Néanmoins, les capacités de vol des psylles, associées à leur transport probable par les vents lors des migrations, rendent certaines barrières géographiques franchissables. Notre modèle biologique tendrait alors à s'éloigner d'un modèle insecte terrestre à fortes contraintes topographiques (coléoptères, héteroptères ...) pour se rapprocher de modèles moins restreints géographiquement (oiseaux migrateurs, mammifères de savane, organismes océaniques ...).

Ces dernières années, de nombreuses études ont été menées sur ces modèles biologiques en utilisant souvent une combinaison de marqueurs mitochondriaux et nucléaires (COI et microsatellites) (Andrews et al., 2020; Keskin et Atar, 2011; Kholodova et al., 2006; Levy et al., 2016; Moreira et al., 2011; Ogden et al., 2015; Papadopoulos et al., 2005; Petzold et Hassanin, 2020; Sharma et al., 2020; Yannic et al., 2016). Ces études permettent de prendre du recul face au modèle *C. pruni* et formuler quelques hypothèses. Elles font ressortir que trois facteurs majeurs peuvent jouer sur la structuration des populations :

- La capacité intrinsèque des individus à se déplacer au sein d'un grand territoire ;
- Le comportement migratoire actif, répété et contrôlé dans le temps ;
- La niche écologique.

Le premier type de modèle concerne les taxons à grande capacité de dispersion sur un grand territoire. Ces taxons sont caractérisés par de faibles niveaux de différenciations génétiques car peu de barrières naturelles s'opposent à leurs flux de gènes. Ce modèle concerne les grands oiseaux comme les aigles royaux (Ogden et al., 2015), des poissons pélagiques et poissons des grands fonds (Andrews et al., 2020) ou encore des espèces de copépodes aux large des côtes européennes (Papadopoulos et al., 2005). Dans ces exemples, la subdivision des populations implique des distances de l'ordre 200 km, pour les aigles royaux, à plusieurs milliers de kilomètres, pour les poissons du genre *Etelis* dans l'indo-pacifique. Alors que les barrières géographiques que l'on pourrait qualifier de mineures comme les rivières ou les microclimats ne semblent pas affecter ces espèces, les barrières majeures jouent un rôle prépondérant dans leur structuration. En particulier, les alternances de milieux, terrestres/marins, marquent les

séparations entre clusters. Ce processus s'observe aussi bien au niveau du bras de mer entre les îles Hébrides extérieures et le continent Ecossais pour les aigles qu'aux détroits du Bosphore ou de Gibraltar pour les copépodes (Ogden et al., 2015 ; Papadopoulos et al., 2005).

Le second modèle concerne les espèces pour lesquels la structure est davantage un mécanisme actif dépendant d'un comportement migratoire comme chez les grands mammifères de plaine. Les girafes, gazelles, oryx (Petzold et Hassanin, 2020), caribou (Yannic et al., 2016), saïgas (Kholodova et al., 2006) ou encore les dromadaires (Sharma et al., 2020) suivent ce schéma. Chez ces taxons les différences génétiques entre clusters apparaissent pour des distances de l'ordre de 100 à 500 km. La diversité génétique au sein de leurs aires de distribution est plus importante que pour les taxons du premier type de modèle. Les taux d'admixture y sont également plus importants.

Enfin, le troisième type de modèle biologique se caractérise par une structuration influencée par des phénomènes environnementaux et climatiques. Ce schéma de structuration concerne des animaux côtiers comme des manchots (Levy et al., 2016), des pieuvres (Keskin et Atar, 2011; Moreira et al., 2011). Dans ce cas, la structure est déterminée par la capacité des populations locales à s'adapter aux habitats ou bien à en coloniser de nouveaux. Ce phénomène produit une importante diversité génétique. De ce point de vue, ce modèle se rapproche des structures observées chez les taxons à fortes contraintes géographiques et topographiques comme les tortues d'eau douce (Vecchioni et al., 2020) ou les rats (Richardson et al., 2017).

### III. Objectifs du stage

Initialement, mon stage visait à tester un jeu de nouveaux marqueurs microsatellites, mettre au point les protocoles PCR associés et finalement analyser plusieurs populations avec ces marqueurs pour répondre à la question de la distance de dissémination des psylles. Malheureusement, la situation sanitaire liée à la COVID et des problèmes techniques ont retardé la mise au point des PCR. C'est pourquoi il a été décidé assez rapidement avec N. Sauvion de réorienter partiellement le sujet de stage en me focalisant sur des résultats obtenus avant mon arrivée mais jusque-là partiellement analysés. Mon travail visait ainsi à répondre à plusieurs objectifs :

- Me familiariser avec les concepts associés à la problématique de délimitation des espèces ;
- Maîtriser au mieux différents logiciels utilisés en génétique des populations ;
- Confirmer l'hypothèse de l'existence deux groupes au sein d'une aire géographique couvrant l'aire connue de distribution de *C. pruni* ;

- Analyser la structure des populations au sein de chaque groupe, et représenter géographiquement les sous-groupes identifiés ;
- Estimer l'intensité des flux de gènes inter et intra-groupes.

## Matériels et méthodes

La part la plus importante de mon travail a été menée sur un jeu de données microsatellites. Ces dernières années, la multiplication des études de génétique des populations menées sur des données microsatellites multi-locus a conduit au développement de nombreux outils informatiques permettant de le traitement et l'analyse de ces données. Ces derniers peuvent être classés en plusieurs approches et pour chaque approche, les méthodes sont amenées à varier, les logiciels reposant sur des algorithmes différents. Chaque outil présente des caractéristiques qui le rend plus efficace dans une situation donnée. Ainsi, cette diversité d'approches et d'outils m'a incité à appliquer une démarche de validation croisée par comparaison des résultats obtenus avec les différents outils. Cette démarche m'a permis renforcer ou au contraire de nuancer les résultats obtenus mais également d'affiner certaines observations en utilisant de façon complémentaires les spécificités de ces outils. Le tableau 1 donne une vue synthétique des outils d'analyse que j'ai utilisés.

Dans un premier temps, j'ai utilisé une approche de partitionnement en K moyennes à travers une analyse discriminante en composantes principales (DAPC). Cette analyse exploratoire a comme principal intérêt d'évaluer un nombre de groupes sans a priori, contrairement aux approches de partitionnement bayésiens que j'ai utilisées ensuite. Les approches bayésiennes bénéficient néanmoins d'une grande puissance statistique par la possibilité d'effectuer des milliers d'itérations des modèles grâce la méthode de Monte Carlo par Chaîne de Markov (MCMC). De plus, ces méthodes permettent l'incorporation de paramètres comme l'admixture ou encore la spatialisation des données (tableau 1). STRUCTURE est le logiciel de référence pour les analyses de génétique des populations. Depuis son développement, la publication initiale (Pritchard et al., 2000) a été citée presque 32 000 fois selon Google Scholar.

TESS et STRUCTURE ont été comparés (Chen et al., 2007; François et Durand, 2010) et sont souvent utilisés conjointement (Durand et al., 2009b; Yannic et al., 2016). Ces deux logiciels reposent sur des algorithmes semblables (Durand et al., 2009b) mais la prise en compte des coordonnées géographique par TESS permet de détecter des structure biologiquement pertinentes qui pourraient passer inaperçues avec STRUCTURE (François et Durand, 2010).

GENELAND est un autre logiciel très utilisé en génétique des populations. Ses performances ont été comparées à celles de TESS et STRUCTURE (Chen et al., 2007; François

and Durand, 2010). GENELAND présente la particularité de représenter l'espace d'échantillonnage d'une autre manière que TESS. Alors que TESS représente l'espace d'échantillonnage par une tessellation (Annexe 7), GENELAND représente cet espace par un certain nombre de pixel. Cette particularité permet à l'utilisateur de régler la résolution de la figure géographique inférée.

J'ai confronté ces résultats à plusieurs indicateurs statistiques couramment utilisés en génétique des populations comme par exemple les valeurs de  $F_{st}$   $F_{is}$  ou  $F_{it}$  indicatrices de la subdivision des populations ainsi que de leur homogénéité génétique (tableau 1).

*Tableau 1: Résumé des méthodes et comparaison des spécificités des différents outils utilisés. Les mots en gras font référence aux critères qui ont le plus influencé mon choix.*

Approche	Méthode	N. groupes a priori	Spatialisation	Admixture	Outil utilisé
K-means clustering	Analyse multivariée => <b>DAPC</b>	<b>NON</b>	NON	NON	<i>adeget</i>
Bayesian clustering	<b>MCMC</b> , répartition des individus entre les cluster inférés	OUI	<b>NON</b>	OUI	STRUCTURE
			<b>OUI</b>	<b>OUI</b>	TESS
		OUI	OUI	<b>NON</b>	GENELAND
Indicateurs statistiques	<b>F-statistiques</b> <b>Déséquilibre de liaison</b> Isolement par la distance Taux de consanguinité	OUI	/	/	<i>genepop</i>

## I Analyses à partir des marqueurs microsatellites

### I A. Présentation du jeu de données

Les analyses de génétique des populations ont été conduites sur un jeu de données comprenant 1123 individus issus de 61 populations<sup>1</sup>. Ces populations ont été collectées entre 2002 et 2007 en France (53), en Espagne (4), en Serbie (2), en Italie (1) et en Allemagne (1). De tailles variables (entre 2 et 30 individus), elles étaient associées à des coordonnées spatiales, longitude (X) et latitude (Y). Ces données sont accessibles libre de droits depuis le DataServe INRAE (Sauvion et al., 2021).

Pour cette étude, j'ai utilisé 8 des 9 marqueurs microsatellites développés par l'équipe (Sauvion et al., 2009). Il s'agissait des marqueurs : Cp6–115; Cp5–45; Cp6–144; Cp4–127; Cp5–43; Cp6–15; Cp6–129; Cp4–108. Les séquences microsatellites sont des gènes caractérisés par la répétition en tandem de motifs d'une longueur variant entre 1 et 6 paires de bases. Ils ont été découverts chez tous les organismes examinés jusqu'à présent. Leur fonction en tant qu'éléments faisant partie intégrante du génome reste encore obscure mais leur important polymorphisme et taux de mutation en font d'excellents marqueurs pour les études

de génétique des populations (Goldstein et Schlötterer, 1999). Les mutations qui ont lieu sur ces séquences sont généralement des gains d'un motif (Goldstein et Schlötterer, 1999). Ainsi, les individus présentant des nombres de répétitions identiques sont considérés comme génétiquement proches. Les taux élevés de mutation (gènes hypervariables) de ces séquences permettent détecter les parentés entre les individus. Ces séquences sont également qualifiées de neutres car elles ne sont pas codantes pour des caractères morphologiques visibles. Ainsi, ces séquences ne seraient pas soumises aux sélections naturelles et sexuelles et s'associeraient selon un modèle panmictique. Cependant, il convient de rester prudent quant à cette dernière affirmation car mutations dans les séquences microsatellites induisent parfois des modifications fonctionnelles au niveau protéique ce qui peut impacter la physiologie et le développement des individus (Goldstein et Schlötterer, 1999). Ainsi, la sélection naturelle pourrait tout de même intervenir si ces individus présentent des handicaps écologiques, d'autant que ces variations pourraient jouer un rôle dans l'adaptation à de micro-niches écologiques (Goldstein et Schlötterer, 1999).

### I B. Clustering sans information spatiale

#### - **DAPC**

L'abréviation DAPC signifie « Discriminant Analysis of Principal Components ». J'ai effectué cette analyse grâce à la librairie R *adegenet*, (version 2.1.0)(Jombart, 2008; Jombart and Collins, 2017). La DAPC peut être divisée en deux étapes. La fonction *dapc* nécessite la définition préalable de groupes. Afin de déduire le nombre de groupes optimal, j'ai utilisé la procédure K.means implantée dans la fonction *find.clusters*. Cette fonction permet de résumer les données sous la forme de combinaisons linéaires de variables ou composantes principales. Le partitionnement est effectué sur ces variables synthétiques (Jombart et Collins, 2017). Pour éviter la perte d'information, toutes les composantes principales ont été retenues pour effectuer l'algorithme K.means. La fonction *dapc* m'a permis de représenter la distribution des données selon un nombre réduit de variables synthétiques (Fig 2 et 3).

*Tableau 2: Simulations effectuées avec la fonction DAPC*

Simulation	Groupe concerné	Fonction <i>find.clusters</i>		Fonction <i>dapc</i>	
		Nombre de composantes principales retenues	Nombre de cluster retenus	Nombre de composantes principales retenues	Nombre de fonctions discriminantes retenues
A vs B	Population totale	300	20	100	2
A vs B	Population totale	300	20	100	1
DAPC A 12	A uniquement	200	12	60	4
DAPC A 5	A uniquement	200	5	60	2
DAPC B 5	B uniquement	200	5	60	3
DAPC B 3	B uniquement	200	3	60	2

Les simulations A vs B correspondent aux assignations des individus aux groupes A et B. Les deux premières lignes du tableau 1 correspondent à la même simulation, seul le mode

de représentation en fonction du nombre de fonctions discriminantes varie. Cette variation permet d'afficher la figure 2a ou 2b. Les lignes suivantes du tableau correspondent aux analyses de partitionnement intra-groupes A et B. La simulation DAPC A 12 correspond à l'analyse de partitionnement du groupe A en 12 clusters et la simulation DAPC A 5 correspond à l'analyse de partitionnement du même groupe en 5 clusters (Fig 3a). Les simulations DAPC B sont nommées selon le même principe.

J'ai retenu le nombre de clusters en me basant sur le « Bayesian Information Criterion » (BIC). Plus la valeur du BIC est petite, plus le modèle est pertinent pour représenter la variabilité contenue dans le jeu de données (Jombart and Collins, 2017) (Annexe 5). Pour les simulations DAPC A 5 et DAPC B 3, j'ai fondé mon choix du nombre de clusters d'après la répartition des groupes selon les DAPC A 12 et DAPC B 5 respectivement (Annexe 5).

Le nombre de composantes principales retenues pour la fonction *dapc* représentait 90% de la variance totale. Chaque fonction discriminante utilisée pour former les cartes factorielles (Fig 2 et 3) correspond au ratio entre la variance inter-cluster par rapport à la variance intra-cluster.

#### - **STRUCTURE :**

Le modèle utilisé dans STRUCTURE (Pritchard et al., 2000) repose sur la méthode MCMC. Ce modèle stochastique décrit une séquence d'événements dont la probabilité d'occurrence ne dépend que de l'événement précédent. Il est également appelé « sans mémoire ». Ici un nombre de clusters maximal,  $K_{max}$ , doit être renseigné a priori. Chaque simulation comporte un nombre d'essais ou « runs ». Avant de lancer une simulation, un nombre d'itération du modèle bayésien doit être renseigné pour chaque essai (Nit par essai). Une phase d'initiation de la chaîne de Markov est également nécessaire pour chaque essai (Burnin). Il s'agit d'une phase durant laquelle le modèle tâtonne avant de **converger**.

J'ai effectué plusieurs types de simulations. Les simulations A vs B du tableau 3 correspondent aux assignations des individus aux groupes A et B. La première m'a servi de référence et les deux suivantes m'ont servi à vérifier si l'admixture et la corrélation des fréquences alléliques avaient un impact sur ces assignations.

Pour les simulations de clustering intragroupes A et B, j'ai utilisé le modèle avec admixture. En effet, ces modèles sont plus efficaces pour repérer des structures au sein de populations car ils sont robustes aux événements de divergences en sous populations alors que les modèles sans admixture ne sont pas robustes aux événements de fusion entre deux sous populations (François et Durand, 2010).

Les simulations delta K A et B comprennent des centaines d'essais en variant le nombre maximal de cluster ( $K_{max}$ ). Ce nombre important de runs m'a permis d'estimer les nombres

optimaux de clusters pour représenter les groupes A et B séparément par la méthode delta-K de Evanno (Evanno et al., 2005) grâce au programme CLUMPAK (Kopelman et al., 2015), disponible sur internet<sup>4</sup>.

Les simulations LOCPRIOR A et B ont été effectuées en tenant compte de l'appartenance des individus à leur population d'origine. Il ne s'agit pas ici d'une réelle spatialisation de l'information car les coordonnées géographiques n'interviennent pas dans l'analyse (Annexe 6).

*Tableau 3 : Simulations effectuées avec STRUCTURE, version 2.3.4 (juillet 2012)*

Simulation	Groupe concerné	Nombre d'essais	Nit par essai	Burnin	K <sub>max</sub>	Admixture	Fréquences alléliques	Localités à priori
A vs B	Population totale	30	14000	2000	2	Avec	Indépendantes	Non
A vs B	Population totale	10	20000	10000	2	Sans	Indépendantes	Non
A vs B	Population totale	10	20000	10000	2	Avec	Corrélées	Non
Delta K A	A uniquement	300	20000	10000	1 à 15	Avec	Corrélées	Non
LOCPRIOR A	A uniquement	50	20000	10000	1 à 10	Avec	Corrélées	Oui
Delta K B	B uniquement	500	20000	10000	1 à 25	Avec	Corrélées	Non
LOCPRIOR B	B uniquement	40	20000	10000	2 à 8	Avec	Corrélées	Oui

Pour les simulations A vs B, j'ai fixé les valeurs du nombre d'itérations et de la phase d'initialisation afin d'approcher un ordre de grandeur cohérent avec les valeurs de l'article de François et Durand, (2010). Les modèles convergeant rapidement, je n'ai pas jugé nécessaire d'augmenter ces valeurs. J'ai résumé chaque simulations grâce à la fonction *Main Pipeline* de CLUMPAK (Kopelman et al., 2015). Enfin, j'ai laissé le logiciel inférer le paramètre  $\alpha$  comme suggéré par Evanno et al, (2005). Il s'agit d'un indicateur permettant de connaître le niveau d'admixture au sein des groupes inférés par les modèles. Si  $\alpha$  est proche de 1, les groupes présentent un fort niveau d'admixture. En revanche, si  $\alpha$  s'approche de 0, les groupes sont génétiquement homogènes avec peu d'admixture.

### I C. clustering avec information spatiale

#### - TESS

TESS (Chen et al., 2007; Durand et al., 2009a) est un autre logiciel de partitionnement bayésien de type MCMC. Toutefois, TESS prend en compte les coordonnées spatiales des individus en créant une tessellation de Voronoï (Annexe 7). Pour toutes les simulations présentées dans le tableau 4, les tessellations ont été pondérées par les distances euclidiennes entre individus calculées par le logiciel à partir des coordonnées géographiques.

Dans le jeu de données initial, celles-ci étaient renseignées par localité. Tous les individus d'une même population partageaient donc les mêmes coordonnées. TESS peut fonctionner avec ce système de coordonnées mais cela entraine un biais dans la représentation

<sup>4</sup> <http://clumpak.tau.ac.il/>

et les assignations car les individus d'une même population se retrouvent dans une même cellule de la tessellation. J'ai donc créé un programme sous R pour faire varier la longitude d'échantillonnage des individus au sein de chaque population. La variation de cette coordonnée était suffisamment faible (de l'ordre d'un mètre) pour que les individus d'une population restent géographiquement plus proches entre eux que des individus d'autres populations. Le nouveau jeu de données a été employé pour les autres méthodes de clustering avec information spatiale.

La simulation A vs B correspond à l'assignation des individus aux groupes A et B en vue de comparer les résultats de TESS avec ceux de STRUCTURE. Les simulations de clustering A et B ont été effectuées en faisant varier les nombres d'itérations et  $K_{\max}$  afin d'évaluer leur impact sur le nombre de groupes inférés par le logiciel. J'ai également testé les 2 modèles d'admixture implémentés dans le logiciel, CAR (Conditionnal Auto Regressive) et BYM (Besag-York-Mollié) décrits par Durand et al (2009). Les deux modèles diffèrent dans la façon d'appréhender et d'estimer les niveaux d'admixture avant de lancer le modèle.

*Tableau 4 : Simulations effectuées avec TESS, version 2.3 (janvier 2010)*

Simulation	Groupe concerné	Nombre d'essais	N-it. par essai	Burnin	$K_{\max}$	Modèle Admixture
A vs B	Population totale	30	12000	2000	2	CAR
Clustering A	A uniquement	15	20000	4000	2 à 4	CAR
Clustering A	A uniquement	6	100000	25000	2 à 3	BYM
Clustering A	A uniquement	9	100000	25000	2 à 4	CAR
Clustering B	B uniquement	3	20000	4000	2 à 4	CAR
Clustering B	B uniquement	15	20000	4000	2 à 4	CAR
Clustering B	B uniquement	11	20000	10000	2 à 12	CAR
Clustering B	B uniquement	9	100000	25000	3 à 5	CAR
Clustering B	B uniquement	9	100000	25000	2 à 3	BYM

Les sorties du logiciel TESS ne pouvant pas être importées sur CLUMPAK comme avec STRUCTURE, j'ai sélectionné le meilleur des essais de la simulation A vs B afin de comparer les résultats d'assignation avec ceux de STRUCTURE et DAPC. J'ai choisi cet essai grâce au « Deviance Information Criterion » DIC (Durand et al., 2009a; François and Durand, 2010; Spiegelhalter et al., 2002). Ici, le DIC a la même fonction que BIC utilisé pour la DAPC. La valeur de DIC la plus basse indique l'essai le plus pertinent pour représenter le jeu de données.

#### - GENELAND :

J'ai utilisé un autre logiciel d'inférences bayésiennes, GENELAND, pour assigner les individus à différents groupes. A la différence de TESS et STRUCTURE, GENELAND ne comprend pas de modèles avec admixture, ce qui le rend moins robuste pour repérer les divergences au sein de populations (François et Durand, 2010). Les simulations ont été réalisées avec la librairie R GENELAND (version 4.9.2) (Guillot et al., 2005a, 2005b) et la fonction *mcmc*. L'originalité de cette méthode vient de la représentation de l'espace d'échantillonnage



selon un nombre de pixels fixés par l'utilisateur (tableau 5). Chaque pixel est assigné au groupe d'individus duquel il est le plus proche géographiquement.

La fonction *PostProcessChain* a servi à fixer les paramètres de représentation géographique et *Plotnpop* permettait d'afficher les graphiques d'inférences du modèle au cours de son exécution (Annexe 8).

Une interface utilisateur a été développée par les créateurs de la librairie. J'ai toutefois privilégié l'usage des lignes de commandes sur R. Le format des données recommandé, ainsi que le manque d'ergonomie et de visibilité sur l'exécution du programme m'ont poussé à ne pas utiliser cette interface (Annexe 8).

La simulation « A vs B » correspond à l'assignation des individus aux groupes A et B un seul essai a été réalisé en vue d'une comparaison avec les résultats de TESS et STRUCTURE. Les simulations « Boucles » A et B correspondent à 10 essais indépendant effectués à l'aide de boucles *for* sur R. Les simulations « Vérification » correspondent à des essais indépendants comprenant un grand nombre d'itérations afin de vérifier les résultats obtenus les simulations « Boucles ». J'ai effectué plusieurs tentatives afin d'atteindre maximum d'itérations supportées par mon équipement informatique,  $6.10^5$  itérations (tableau 5).

*Tableau 5 : Simulations effectuées avec GENELAND version 4.9.2 (2020)*

Fonction <i>mcmc</i>							Fonction <i>PostProcessChain</i>	
Simulation	Groupe concerné	Nombre d'essais	N-it. par essai	Précision	K <sub>max</sub>	Fréquences alléliques	Burnin	Nombre de pixels
A vs B	Population totale	1	75000	10	10	Indépendantes	1500	200 x 200
Boucle A	A uniquement	10	50000	10	5	Corrélées	200	300 x 300
Vérification A	A uniquement	1	500000	100	10	Corrélées	200	300 x 300
Boucle B	B uniquement	10	50000	10	5	Corrélées	200	300 x 300
Vérification B	B uniquement	1	600000	1000	6	Corrélées	200	300 x 300

Le paramètre précision, appelé « thinning » dans la littérature, définit le pas de sauvegarde des itérations. Dans GENELAND, seul un certain nombre d'itérations est pris en compte dans l'analyse. Fixer la précision à 10 pour la simulation A vs B signifie que 7 500 itérations ont été sauvegardées.

#### I D. Indicateurs statistiques

##### ***genepop* :**

J'ai utilisé la librairie R *genepop* (version 4.7.5) (Rousset, 2008; Rousset, 2020) pour estimer les valeurs des statistiques issues des différents tests listés dans le tableau 6.

Tableau 6 : Résumé des tests effectués avec *Genepop*, version 4.7.5 (2020)

Groupe concerné	Test de l'équilibre H-W	Rho-statistiques	Test de déséquilibre de liaison
	Paramètres		
Population totale	<i>which = Proba</i>	<i>pairs = TRUE</i>	<i>dememorization, batches et iterations par défaut</i>
Population totale	<i>which = Deficit</i>	<i>pairs = FALSE</i>	
A uniquement	<i>which = Proba</i>	<i>pairs = TRUE</i>	
A uniquement	<i>which = Proba</i>	<i>pairs = FALSE</i>	
B uniquement	<i>which = Proba</i>	<i>pairs = TRUE</i>	
B uniquement	<i>which = Proba</i>	<i>pairs = FALSE</i>	

La fonction *test\_HW* a servi à tester l'équilibre de Hardy-Weinberg. La méthode *Proba* a permis d'obtenir la probabilité de respecter l'équilibre H-W pour chaque ensemble testé : à chaque locus en prenant en compte toutes les populations, à chaque population en prenant en compte tous les locus et pour chaque population à chaque locus. La méthode *deficit* a servi à calculer le déficit d'hétérozygote en prenant en compte toute la population (tableau 6). Ce test a été effectué pour les mêmes ensembles que pour la méthode *Proba*.

J'ai calculé les Rho-statistiques ( $Rho_{is}$ ,  $Rho_{st}$ ,  $Rho_{it}$ ) grâce à la fonction *Fst*. Les Rho-statistiques s'interprètent de la même façon que les F-statistiques (Rousset, 2020). Ces valeurs s'obtiennent en renseignant le paramètre *sizes = TRUE*. Les Rho-statistiques sont les indicateurs adaptés aux variations alléliques basées sur le nombre de répétitions, format adapté pour les séquences microsatellites. Une valeur de  $Rho_{st}$  négative est indicatrice d'un excès d'hétérozygotes alors qu'une valeurs positive correspond à un déficit d'hétérozygotes.

Le paramètre *pairs = TRUE* permettait d'obtenir les valeurs pour chaque population 2 à 2 à chaque locus. Le paramètre *pairs = FALSE* permettait d'obtenir les valeurs à chaque locus indépendamment pour toutes les populations et également des valeurs globales, regroupant toutes les populations et tous les locus.

J'ai calculé les déséquilibres de liaison entre locus avec la fonction *test\_LD* (tableau 6).

Pour toutes ces analyses, j'ai laissé les paramètres *dememorization*, *batches* et *iterations* sont restés par défaut comme conseillé par Rousset, 2020. En effet, les résultats obtenus selon ces paramètres produisaient des erreurs standards petites (facteur entre 10 et 100) devant les probabilités calculées et les nombres de changements de configurations au cours des chaines de Markov étaient généralement supérieurs à 1000 (Rousset, 2020).

#### - *adegenet*

J'ai testé l'isolement par la distance au sein de la population totale, et au sein des deux groupes A et B, par des tests de Mantel avec la fonction *mantel.randtest*. J'ai calculé la probabilité de consanguinité par individu grâce à la fonction *inbreeding*. Pour comparer les taux de consanguinité entre les populations, j'ai sélectionné les individus dont la probabilité de consanguinité était supérieure à 50%. J'ai ensuite replacé ces individus dans leur population

respective afin d'obtenir une proportion d'individus probablement consanguins par localité (Annexe 5).

## II Analyses des séquences des gènes COI et ITS

Des séquences ITS2 et COI m'ont également été transmises par N. Sauvion. J'ai disposé ainsi de : 365 séquences COI (126 d'individus du groupe A et 239 du groupe B), et 785 séquences ITS2 (340 du groupe A et 445 du groupe B). Chaque séquence était associée à un individu identifié par le même système de code que pour les microsatellites, relié à la base de données globale. Ces séquences m'ont été fournies déjà alignées et nettoyées dans un format adéquat pour leur analyse.

Dans un premier temps, j'ai utilisé le logiciel DnaSP (version 6) (Rozas et al., 2017) pour calculer le nombre de site de ségrégation, le nombre de mutations, le nombre d'haplotypes, la diversité nucléotidique ainsi que, le D de Tajima, un indicateur de divergence des populations (Annexe 10). Ensuite, j'ai créé des réseaux d'haplotypes grâce au logiciel PopART (Leigh and Bryant, 2015). Ce logiciel m'a également fourni des statistiques comme le nombre de sites identiques ou la diversité nucléotidique (Annexe 10).

J'ai créé 6 réseaux d'haplotypes, trois par type de séquences (ITS ou COI). Pour chaque type de séquence, j'ai affiché un réseau global en intégrant les séquences des groupe A et B (Annexe 10) puis un réseau présentant chaque groupe séparément. J'ai créé des cartes affichant des clusters géographiques grâce aux coordonnées associées à chaque échantillon avec la fonctionnalité cartographique de PopART. J'ai utilisé le type de réseau *Median Joining Network* pour tous les types de séquences. J'ai retravaillé chaque réseau et carte sur le logiciel de retouche d'images GIMP<sup>5</sup> (version 2.10.22) afin que les couleurs des clusters géographiques et génétiques correspondent.

## III Comparaison des modèles et représentation des groupes

Pour les analyses de structure inter-groupes A et B, j'ai rassemblé les résultats d'assignations aux groupes A et B obtenus avec les différentes analyses dans un tableau Excel. À chaque identifiant d'individu, j'ai associé le pourcentage d'assignation dans les groupes pour chaque méthode d'analyse. J'ai ensuite ajouté une colonne dont chaque cellule contenait une fonction *si.condition* affichant 1 en face des individus pour lesquels la proportion d'assignation au groupe 1 était supérieure à 0.5, et 2 en face des autres. J'ai trié les données de chaque tableau par la colonne ID et j'ai comparé les colonnes sp pour chaque méthode 2 à 2. Pour chaque

---

<sup>5</sup> <https://gimp.cc/download-gimp/>

individu, si les deux méthodes comparées indiquaient la même valeur, le numéro 1 était indiqué dans une colonne supplémentaire et 0 si les valeurs étaient différentes. Par la suite, j'ai sommé les valeurs de cette nouvelle colonne pour réaliser un pourcentage d'assignations similaires par rapport aux 1123 individus.

J'ai utilisé le logiciel d'information géographique QGIS (version 3.20.0)(QGIS Development Team, 2021) pour créer des cartes permettant de commenter la distribution spatiale des individus en fonctions des groupes inférés par les différentes méthodes.

Pour les analyses de structure intragroupes A et B, je n'ai pas suivi cette démarche. En effet, chaque groupe comprenant plusieurs clusters, la tâche aurait nécessité la vérification de chacun des 1123 individus par rapport aux autres, pour chaque méthode. Ce travail aurait été trop fastidieux. Je me suis basé sur une comparaison visuelle des résultats graphiques pour faire ressortir les plus fortes tendances. Je me suis également référé aux indicateurs statistiques inférés par les différentes méthodes pour interpréter ces résultats.

## Résultats

### I Analyses sur les données des marqueurs microsatellites

#### - DAPC

Cette analyse fournit des cartes factorielles qui sont des plans formés des fonctions discriminantes sélectionnés par l'utilisateur lors de l'exécution de la fonction *dapc* (Fig 2 et 3).

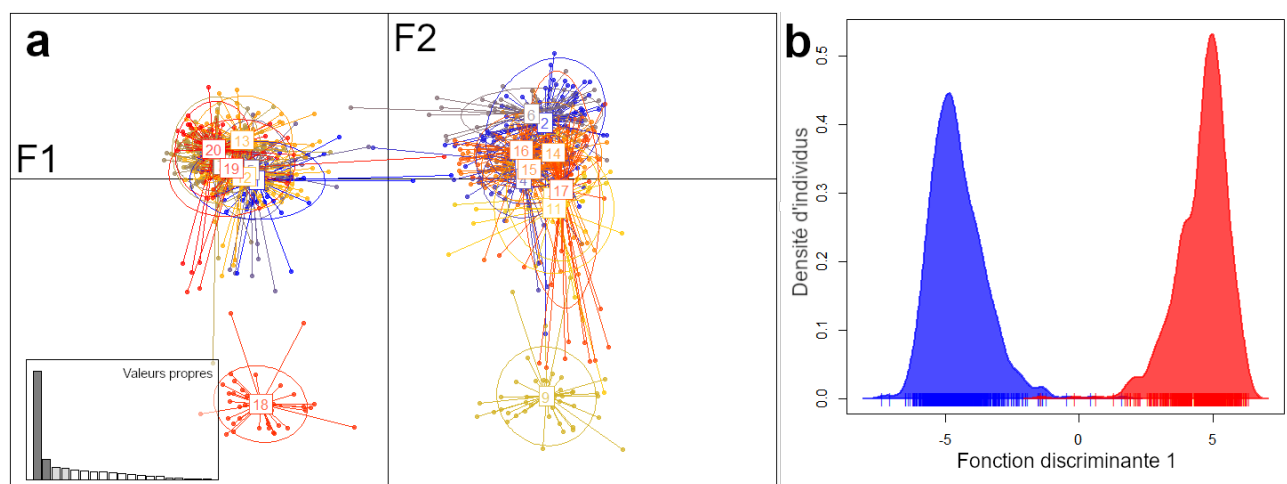


Figure 2 : Cartes factorielles formées par les fonctions discriminantes F1 et F2 des simulations DAPC A vs B (tableau 2). a) 20 clusters représentés dans le plan formé des deux premières fonctions discriminantes : 60% de l'information totale représentée. Valeurs propres F1 => 47.2% ; F2 => 12.2%. b) 2 clusters représentés selon la première fonction discriminante.

Sur les cartes factorielles, les points représentent les individus reliés aux centroïdes du cluster auquel ils se rapportent. Chaque couleur représente un cluster. Plus les clusters sont proches les uns des autres, plus ils sont proches génétiquement. Les clusters proches du centre d'origine apportent peu d'information selon ces fonctions discriminantes alors que les clusters

les plus éloignés du centre apportent le plus d'information. Sur la figure 2a, J'ai retenu 20. Ils se répartissent en deux groupes distincts sur l'axe 1, le plus informatif. Ces deux groupes sont éloignés l'un de l'autre et du centre de la carte factorielle. Deux clusters se différencient de ces groupes selon l'axe 2 mais restent inféodés à chacun de leurs groupes respectifs. Cette division m'a permis de ne tenir compte que de l'axe 1 et de représenter les individus selon une seule fonction discriminante (Fig 2b). Les assignations des individus selon la figure 2b m'ont permis de séparer les deux groupes A et B afin d'effectuer des analyses DAPC sur ces groupes séparément (tableau 2 ; Fig 3).

Trois groupes semblent se différencier sur la figure 3a. Le cluster 1 (bleu) se différencie des autres selon l'axe 1. Le cluster 2 (gris) se différencie clairement des autres selon l'axe 2, qui est presque aussi informatif que l'axe 1. Trois clusters (3, 4 et 5) se superposent et semblent former un groupe homogène. Cependant, tous ces clusters sont proches du centre d'origine de la carte. Cela indique que le ratio entre la variance inter-cluster et la variance intra-cluster n'est pas très important. Pour les individus B, (Fig 3b), un cluster se sépare nettement des autres selon l'axe 1 et les autres clusters se différencient selon l'axe 2. Ces clusters restent très proches du centre d'origine, indiquant une forte proximité génétique.

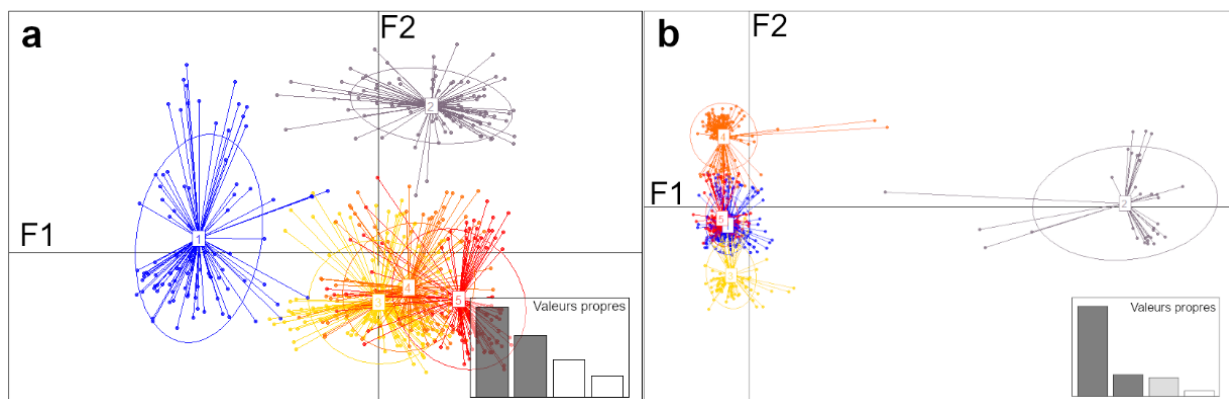


Figure 3 : Cartes factorielles formées par les fonctions discriminantes F1 et F2 des simulations DAPC A 5 et B 5 (Tableau 2). a) Groupe A pour lequel 5 clusters ont été retenus. b) Groupe B pour lequel 5 clusters ont été retenus.

## - STRUCTURE

Le logiciel STRUCTURE fournit des assignations qui se représentent par des diagrammes en barres dont chacune représente un individu. La probabilité d'appartenance à chaque cluster est indiquée par une couleur pour chaque individu (Fig 4a). Les individus dont la proportion d'assignation au groupe A était supérieur à 50% ont été assignés au group A et les autres au groupe B. Peu d'individus présentaient une ambiguïté d'assignation, y compris dans les zones de sympatrie.

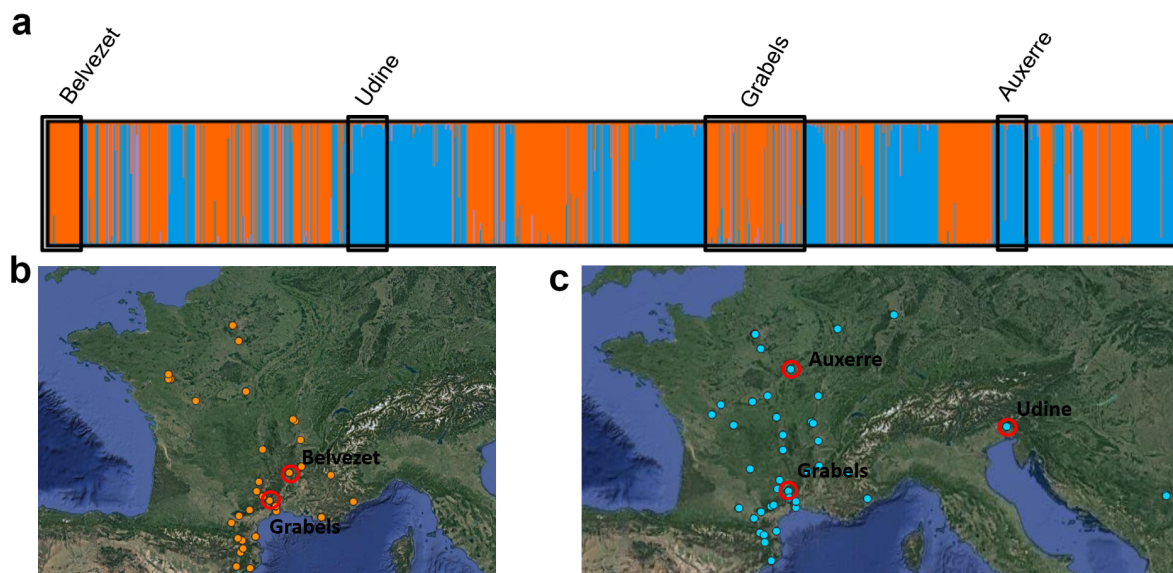


Figure 4 : Résultats de l'analyse A vs B (tableau3) effectuée avec STRUCTURE. a) Graph des assignations aux groupes A (rouge) et B (bleu) pour lequel certaines populations caractéristiques sont représentées. b) Répartition géographique des individus assignés au groupe A selon le graphique en a). c) Répartition géographique des individus assignés au groupe B selon le graphique en a)

Certaines populations comportaient uniquement des individus A (ex. population Belvezet, Fig.4a) et d'autres uniquement des individus B (ex. populations Udine et Auxerre) mais dans la majorité des localités, les deux groupes ont été échantillonnés (ex. population Grabels). Les individus du groupe A se concentrent principalement sur la côte méditerranéenne. Dans cette zone quand les deux groupes sont observés dans une même population, le groupe A est très majoritaire (proportion  $> 70\%$ ). Les individus A sont également retrouvés dans quelques localités du centre-ouest, nord-ouest de la France. Les individus du groupe B sont retrouvés quasiment partout en France en allopatrie plus ou moins stricte (très rare zone d'allopatrie stricte ne comportant que des individus A tel que le plateau du Larzac). Seuls des individus B sont observés dans les localités échantillonnées du massif central, du nord-est de la France, de l'ouest de l'Allemagne, du nord de l'Italie et de la Serbie.

Pour les simulations de clustering intragroupes A et B, les nombres optimaux de cluster inférés par la méthode delta K (tableau 3), étaient respectivement de 15 (groupe A, 300 runs) et 17 (groupe B, 500 runs). Pour les simulations LOCPRIOR A et B (tableau 3) le nombre optimal de clusters estimé par cette méthode était de 6 pour les deux groupes.

Entre 16 et 25 individus ont été assignés avec une proportion de plus de 40% à chaque cluster par la simulation delta K B (tableau 3) et 14 et 21 par la simulation delta K A. Toutefois, certains clusters du groupe A ne comprennent que 4 individus assignés à plus de 40 %. 400 individus ont été assignés à moins de 40% dans tous les clusters par simulation delta K A alors que ce nombre n'est que de 254 individus pour la simulation delta K B. Ce résultat indique une plus grande admixture individuelle au sein des individus A qu'au sein des individus B.

Les individus ont été répartis moins uniformément pour les simulations LOCPRIOR. Pour le groupe A, le logiciel a assigné plus de la moitié des individus à un seul cluster. 225 individus y sont assignés à plus de 40%. Le reste des individus été distribués dans 5 autres avec une plus grande importance pour deux d'entre eux. 30 et 25 individus y ont été assignés à plus de 40%. Le groupe B a été distribué entre 3 clusters majoritairement dont un principal (158 individus assignés à plus de 40%) et deux plus réduits (66 et 56 individus assignés à plus de 40%). Pour les deux groupes A et B, une grande partie des individus ne sont assignés à plus de 40% dans aucun cluster. La proportion de ces individus est, cette fois ci, similaire entre les deux groupes, 185 pour le groupe B et 197 pour le groupe A (Annexe 6).

#### - TESS

Pour les analyses intergroupes, TESS et STRUCTURE ont produit exactement les mêmes résultats. Pour les analyses intragroupes, le logiciel n'a pas détecté plus de 2 clusters, quel que soit le modèle utilisé, la valeur de  $K_{\max}$  ou le nombre d'itérations (Annexe 7). Le modèle BYM n'a pas fourni de résultats fondamentalement différents des résultats obtenus avec le modèle CAR. Les indices DIC associés au modèle BYM étaient en moyenne légèrement supérieurs à ceux du modèle CAR mais les inférences géographiques des deux modèles étaient semblables.

J'ai observé une certaine constance en comparant les tessellations *hard clustering* inférées pour chaque groupe ainsi qu'une ressemblance entre les 2 groupes. Les groupes A et B se diversifient progressivement suivant un gradient ouest-est. Par exemple, les individus de la région au nord d'Angers (ouest de la France) sont presque toujours classés dans un cluster différent de ceux de Coursegoules (sud-est de la France).

#### - GENELAND

Pour la distinction entre les groupes A et B, GENELAND a donné parfois des résultats différents de ceux de TESS et STRUCTURE. Cependant, la similarité d'assignations entre ces méthodes reste élevée (89%). Le modèle a convergé rapidement après environ 6 000 itérations. Le logiciel a rencontré des difficultés à reconnaître des individus appartenant à un groupe au sein de populations géographiquement dominées par des individus de l'autre groupe. Les individus "mal" identifiés ne correspondaient généralement pas à des individus diagnostiqués hybrides par PCR. Ce phénomène est probablement lié au mode de représentation de l'espace d'échantillonnage en pixel. En effet, plusieurs individus peuvent se trouver dans une même cellule (pixel). En revanche, les cellules d'une tessellation ne contiennent qu'un seul échantillon. L'astuce de déplacer géographiquement les individus d'une même population selon l'axe des abscisses pour éviter l'autocorrélation spatiales ne semble pas avoir aussi bien fonctionné avec GENELAND qu'avec TESS (Annexes 7 et 8).



J'ai observé certaines tendances dans la simulation de clustering intra-groupes ('Boucles A et B' et 'Vérifications A et B', tableau 5). Pour le groupe B, bien que beaucoup de modèles n'aient pas convergé, la majorité des itérations de la simulation 'Boucle' et de la simulation 'Vérification' (respectivement 54% et 63%) a inféré 3 clusters (Fig. 5). Les clusters étaient également structurés selon un gradient ouest-Est. J'ai observé le même type de structures géographiques pour  $5.10^4$  que pour  $5.10^5$  itérations, ce qui m'a conforté dans la vraisemblance de ces résultats.

La structure du groupe A était plus confuse, avec une importante variabilité entre les essais, même lorsque le modèle avait convergé. Le logiciel représentait des formes en mosaïque difficilement interprétables. (Tableaux en annexe 8)

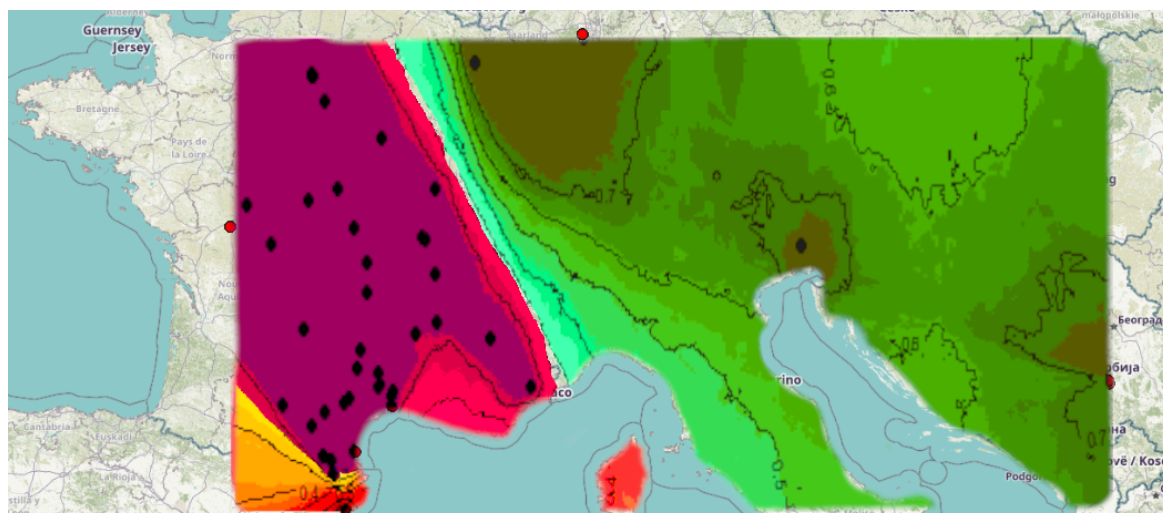


Figure 5 : Représentation graphique des clusters inférés par la simulation vérification B (tableau 5), chaque couleur représente un cluster différent. Les points représentent les localités d'échantillonnage et les lignes représentent les isoclines de proportion d'assignation de l'espace à un groupe

Sur la figure 5 les trois clusters se répartissent selon un gradient ouest-est avec un cluster associé à la région catalane espagnole limitée au nord par les Pyrénées. La majeure partie du territoire français est occupé par un cluster génétiquement homogène, limité à l'est par une frontière virtuelle allant de Nancy à Monaco.

#### - *genepop*

Les tests de d'équilibre de H-W ont montré que ni la population totale, ni les groupes A et B ne respectaient globalement l'équilibre. Pour la population totale, le test effectué avec la méthode *deficit* a montré que cette déviation de l'équilibre de H-W était due à un déficit en hétérozygotes dans la plupart des cas. Cependant, certaines populations à certains locus respectaient l'équilibre de H-W.

Selon les tests de déséquilibre de liaison, tous les locus étaient indépendants les uns des autres dans la population totale, à l'exception des locus 6\_115 et 5\_45, pour lesquels le déséquilibre de liaison était significatif dans la population totale. Toutefois, ce déséquilibre disparaît avec la subdivision en groupes A et B.



Les valeurs de  $Rho_{st}$  de la population totale indiquent une forte différenciation intergroupe (42,1 %) (Annexe 9). Les valeurs de  $Rho_{is}$  montrent également une forte diversité et différenciation génétique entre les individus eux-mêmes à l'intérieur des sous-populations (60,8 %). Les valeurs de  $Rho_{it}$  indiquent que ces différenciations génétiques entre individus sont encore plus fortes au sein de la population totale prise dans son ensemble (73,3%).

Les valeurs de  $Rho_{st}$  sont d'un ordre de grandeur 10 fois inférieur pour chaque groupe pris indépendamment. Mais les valeurs de  $Rho_{it}$  et  $Rho_{is}$  restent du même ordre de grandeur. La valeur de  $Rho_{is}$ , proche de 50% pour le groupe A, suggère une importante diversité génétique entre les individus.

Le  $Rho_{st}$  calculé pour le groupe A est 4 fois supérieur à celui calculé pour le groupe B. Toutefois, j'ai pu observer que la majorité de cette variation entre les 2 groupes était portée par un seul locus. Le logiciel a calculé une différenciation de 12% ( $Rho_{st}$ ) entre clusters du groupe A au locus 6\_15. Les autres locus tamponnant cette valeur, le  $Rho_{st}$  global passe à 0.04. Les valeurs de  $Rho_{st}$  étaient mieux réparties entre les locus pour le groupe B.

#### - *adegenet*

Les tests de Mantel ont montré des valeurs d'isolements par la distance significatives (corrélation entre distance génétique et distance géographique) pour la population totale et le groupe B (Annexe 5). En revanche, aucun isolement par la distance significatif n'a été trouvé pour le groupe A.

Pour la probabilité de consanguinité, le logiciel a considéré que 50 individus du groupe A avaient une probabilité supérieure à 50% d'être issus de parents partageant un ancêtre commun (environ 9% des individus). Pour le groupe B, ce nombre s'élève à 237 individus (environ 42%) (Annexe 5).

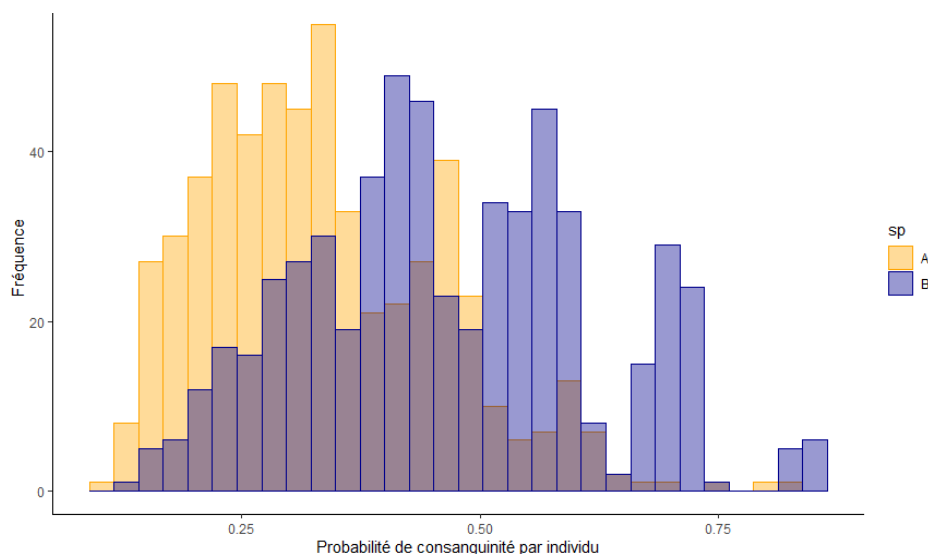


Figure 6 : Histogramme de probabilité de consanguinité pour chaque individu. L'axe de abscisses correspond à la probabilité pour chaque individu d'hériter deux allèles identiques issus d'un même ancêtre. L'axe des ordonnées représente le nombre cumulé d'individus

## II Analyses sur les données des séquences ITS et COI

Les analyses de séquences du gène ITS2 des individus assignés au groupe A par PCR diagnostique permettent de distinguer trois clusters (Fig 7).

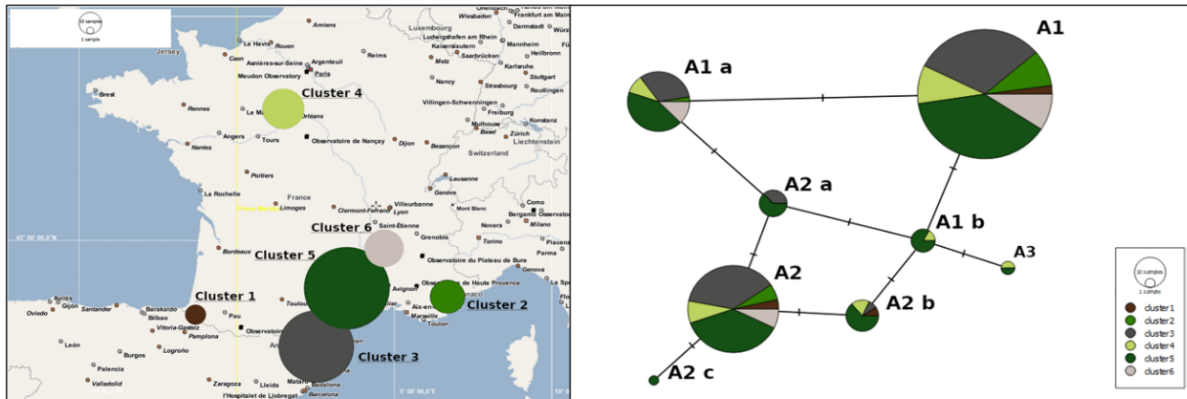


Figure 7 : Réseau d'haplotypes ITS groupe A. Les cercles représentent le regroupement d'individus haplotypes. Les barrettes perpendiculaires aux liens entre les cercles représentent les mutations qui différencient les clusters

Les cluster A1 et A2 sont composés d'un sous-cluster principal et plusieurs sous-clusters séparés du principal par une seule mutation. Le cluster A3, composé de 2 individus, est séparé des sous-clusters principaux par 2 mutations, c'est pourquoi je l'ai considéré comme différent. Les clusters 1 et 2 présentent la même répartition géographique. Certains sous-clusters sont cependant distribués dans des aires précises, comme les sous cluster A2a et A2b, répartis au sud-ouest. Ce réseau est globalement très connecté car au maximum 4 mutations séparent les clusters A1 et A2. Pour ces séquences, le D de Tajima a une valeur 1.308 (P-value >0.1). Cette valeur non significative indique que la population évoluerait de manière neutre, sans être soumise à une pression de sélection.

Les analyses de séquences du gène ITS2 des individus assignés au groupe B permettent de distinguer deux clusters principaux avec une forte différenciation car 7 mutations séparent les deux principaux sous clusters. (Fig.8)



Figure 8 : Réseau d'haplotypes ITS, groupe B

Les individus de ces clusters se distribuent aussi avec les mêmes proportions dans l'espace. Toutefois, certains clusters géographiques sont absents du cluster B 2 comme les

séquences de Serbie par exemple. Dans son ensemble, ce cluster comprend plus de séquences d'Europe de l'ouest que d'Europe de l'Est. Pour ces séquences, le D de Tajima a une valeur 1.452 (P-value >0.1) qui indique également que la population évoluerait de manière neutre.

Les analyses de séquences du gène COI des individus assignés au groupe A permettent de distinguer deux clusters principaux avec quelques individus différents d'une mutation qui s'en dégagent (Fig. 9).

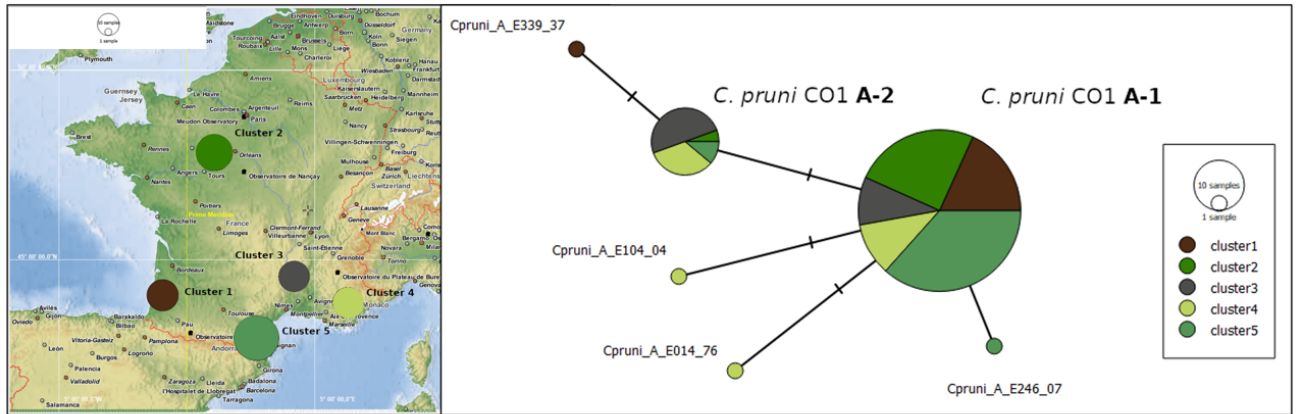


Figure 9 : Réseau d'haplotypes COI groupe A.

Le groupe A2 semble associé aux régions du sud, ne présentant aucune séquence de l'ouest de la France. En revanche le groupe A1 semble se répartir de manière plus homogène à l'échelle du territoire français. Le D de Tajima a une valeur -1.338 (P-value >0.1) signifiant également une évolution attendue en l'absence de sélection.

Les analyses de séquences du gène COI des individus assignés au groupe B ne permettent pas distinguer différents clusters. En effet, le groupe B présente un seul cluster central autour duquel gravitent des individus et des petits clusters composés de 2 à 5 individus (Fig. 10).

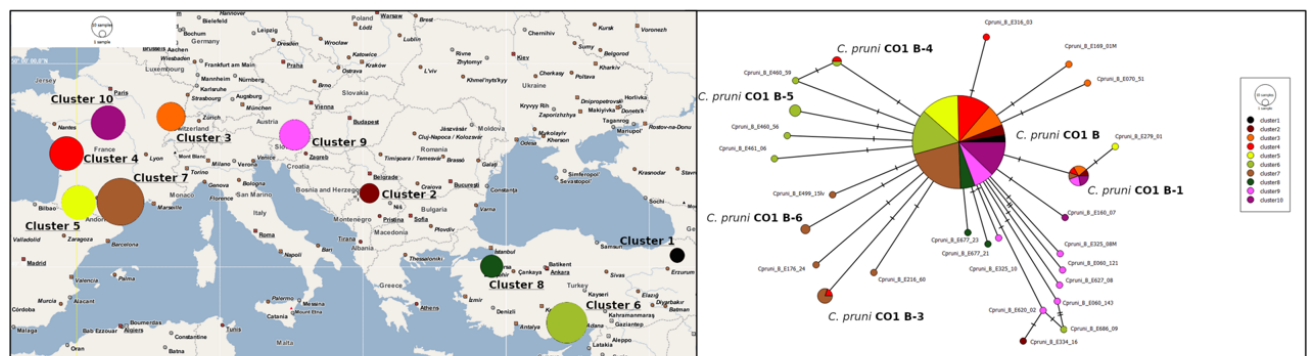


Figure 10 : Réseau d'haplotypes COI groupe B. D de Tajima = -2.512, P-value <0.001.

Certains clusters géographiques comme le cluster rose à l'est de l'Italie regroupe un nombre important d'individus différenciés les uns des autres et du cluster principal d'une ou deux mutations. D'autres, au contraire, comme ceux de l'ouest de la France sont quasi-entièrement compris dans le cluster principal. Ici la valeur du D de Tajima est de -2.512 (P-

value  $<0.001$ ), indiquant un excès d'allèles rares. Ce phénomène pourrait s'expliquer par une expansion de la population après un goulot d'étranglement génétique.

## Discussion

Des études antérieures menées au sein de mon équipe d'accueil avaient montré que *C. pruni* était très vraisemblablement un complexe de deux espèces cryptiques (Marie-Jeanne et al., 2020; Peccoud et al., 2013; Sauvion et al., 2009). Notre étude a porté sur un nombre important d'individus ( $> 1000$ ) représentatifs des aires de distributions connues des deux groupes (Sauvion et al., 2021) nous permet de conforter cette hypothèse.

### Deux groupes génétiques fortement distincts

Tout d'abord, nos résultats confirment sans ambiguïté que *C. pruni* est un taxon divisé en deux groupes très différents génétiquement. Plusieurs approches d'analyses (analyse multivarié, méthodes bayésiennes de clustering, incluant ou non de l'information spatiale) ont été utilisées pour analyser différents types de données moléculaires (marqueurs microsatellites, gènes COI et ITS2).

Tous les résultats obtenus ont montré une forte similitude. La simulation DAPC A vs B (tableau 2) a montré une seule différence d'assignation avec les inférences de STRUCTURE et TESS. Le succès d'une méthode basée sur l'identité des allèles (DAPC) à assigner correctement des individus grâce à des marqueurs microsatellites (codés par la taille des allèles) suggère un nombre important de mutations entre les deux groupes. Cela indiquerait une divergence ancienne et une absence de flux gènes pendant suffisamment longtemps pour amener au fort isolement observé aujourd'hui. En observant les réseaux d'haplotypes globaux (Annexe 10), nous dénombrons 22 (ITS) et 38 (COI) mutations portées sur les liens reliant les groupes A et B. Démontrant au passage une plus grande efficacité des séquences mitochondriales pour la distinction entre les deux groupes ces résultats statuent en faveur d'une (quasi) absence de flux gènes entre les deux groupes depuis une longue durée. Ces observations corroborent une étude antérieure de Peccoud et al. (2018) qui a démontré que les barrières reproductives sont très fortes entre les individus des deux groupes même si elles ne sont pas totales. Des accouplements inter-groupes sont possibles mais les hybrides sont très rares *in natura*. Notre étude le confirme, puisque nous avons trouvé un seul individu sur 1123 que nous pouvons formellement qualifier d'hybride.

Les simulations sans admixture et avec les fréquences alléliques corrélées (tableau 2) ont inféré des assignations identiques à celles de TESS et STRUCTURE avec admixture et fréquences indépendantes. Une conclusion est que la population totale ne présentait pas d'admixture A – B et que les fréquences alléliques étaient effectivement indépendantes entre ces deux groupes.

La conséquence de cet isolement reproductif se traduit parfaitement au travers différents indices génétiques. Globalement et pour chaque locus, la population totale déviait de l'équilibre de Hardy - Weinberg. Ce phénomène associé à un déficit en hétérozygotes démontre un *effet Wahlund*, indiquant une subdivision de la population totale en sous-populations (ici les deux groupes A et B). Les valeurs de  $Rho_{st}$  calculées pour chaque groupe séparément ont été divisées par un facteur 10 par rapport à la population totale. Cela montre que le déficit en hétérozygotes s'estompe lorsque l'on considère les groupes A et B séparément. Ces groupes ont des fréquences alléliques plus proches de celles attendues à l'équilibre de H-W que la population totale. La division en deux groupes est donc biologiquement plus satisfaisante que la prise en compte de la population dans son ensemble. De plus, les valeurs du D de Tajima, positives et significatives, et les réseaux d'haplotypes ITS et COI sur tous les individus (Annexe 10) indiquent une sélection diversifiante. Enfin, les taux de consanguinité et les valeurs d'isolement par la distance sont différents entre les deux groupes. En résumé, tout se passe comme si les psylles des deux groupes s'ignoraient et vivaient leur vie sans entrer en compétition pour la reproduction notamment.

#### Des flux de gènes intraspécifiques plus ou moins importants

Les forts indices génétiques de différenciation que nous avons trouvés confortent l'hypothèse que les groupes A et B seraient deux espèces différentes, mais peut-on statuer sur l'existence d'autres espèces ou sous-espèces ?

Les valeurs de  $Rho_{st}$  10 fois inférieures pour chacun des deux groupes par rapport aux valeurs de la population totale indiquent qu'il n'y aurait pas de troisième espèce. Un doute pourrait subsister au sein de l'espèce A. En effet, la valeur du  $Rho_{st}$  pour le groupe A est 4 fois supérieure à celle de l'espèce B. Cependant, cette valeur étant influencée par un seul locus, il nous semble peu probable qu'elle reflète l'existence d'une troisième espèce.

Les individus de l'espèce A présentent une structuration géographique différente de celle des individus de l'espèce B. En effet, ces individus semblent entretenir d'important flux de gènes comme le montre le peu d'individus consanguins évalués ainsi que l'absence d'isolement par la distance. La représentation géographique des clusters évalués par les méthodes ne tenant pas compte des coordonnées spatiales n'a pas reflété de structure particulière mais plutôt une distribution aléatoire de la diversité génétique. Il ne semble donc pas y avoir d'obstacles aux flux de gènes à l'échelle géographique de notre étude (France et nord de l'Espagne pour l'espèce A). Cela explique en partie la difficulté rencontrée par les logiciels bayésiens à converger vers des structures cohérentes entre les simulations.

Néanmoins, à l'échelle locale, j'ai pu observer quelques récurrences. Par exemple, les populations de la région de Millau et du Mont Aigoual, jusqu'à la région de Montpellier

semblent se démarquer génétiquement des populations de la région catalane espagnole. Ce phénomène suggère que la chaîne des Pyrénées ne constitue pas une barrière reproductive. Nous n'avons pas d'explication biologique à l'heure actuelle pouvant l'expliquer. De même, en région parisienne, les populations de Villepreux et Rennemoulin sont systématiquement classées par TESS dans des clusters différents malgré leur proximité géographique (la même tendance s'observe pour l'espèce B). Il s'agit de deux populations collectées le même jour (14/03/2007). Ces différences génétiques ne sont donc pas le fruit d'un décalage temporel. Ces structures s'observent également avec GENELAND, de manière toutefois moins flagrante.

Les réseaux d'haplotypes tendent à concilier les différentes méthodes bayésiennes. Le réseau COI présente une structure similaire à celle proposée par TESS avec la détection de 2 clusters. Le plus important des deux est réparti de manière homogène et le second est principalement composé d'individus du sud-est de la France (Fig. 9). Le nombre et la répartition des clusters représentés par le réseau ITS (Fig. 7) se rapprochent des structures inférées par GENELAND. La figure 7 montre une forte homogénéité dans la distribution spatiale des clusters génétiques inférés. La connectivité entre tous les clusters du réseau traduit un brassage génétique entre les populations.

La structuration géographique au sein des individus assignés à l'espèce B est beaucoup plus claire. L'interprétation conjointe de l'isolement par la distance et du nombre d'individus consanguins suggère un manque de contact entre les individus des populations les plus distantes entre elles. Le nombre de mutations entre les clusters du réseau ITS (Fig 8) m'a conforté dans cette idée. Les barrières reproductives ne semblent cependant pas totales car certains indicateurs montrent la subsistance de flux de gènes même à très longue distance. Par exemple, les figures 8 et 10 montrent les regroupements d'individus de Serbie et d'Occitanie au sein de mêmes clusters génétiques. Ainsi, l'isolement par la distance ne s'expliquerait que par la prise en compte d'une aire de distribution plus vaste pour l'espèce B que pour l'espèce A. Il existerait donc des mécanismes intrinsèques à chaque espèce qui expliqueraient pourquoi la barrière des Pyrénées, par exemple, constitue un obstacle aux flux de gènes pour une espèce et pas pour l'autre. La présence de séquences semblables ou identiques à travers tout le continent européen peut aussi s'interpréter comme une extension rapide de l'aire de répartition de l'espèce B, mieux adaptée aux conditions climatiques continentale que l'espèce A. Cette hypothèse peut être appuyée par la valeur du  $D$  de Tajima associée aux séquences ITS (Fig 10). Ainsi, certains clivages difficiles à interpréter comme la différenciation entre le cluster d'Europe de l'est et d'Europe de l'ouest (Fig 5) serait une limite arbitrairement inférée par les logiciels. Elle s'interpréterait plutôt par une transition plus lente d'un cluster à l'autre à travers des individus localement interconnectés. En effet, l'espèce B présente également des structures à échelles



géographiques très fines comme entre Villepreux et Rennemoulin, en région parisienne (quelques kilomètres de distance). De la même manière, un clivage est souvent repéré au cours des essais effectués avec TESS et également de manière plus diffuse avec STRUCTURE entre les deux populations serbes, pourtant distantes de seulement 5 km.

Ainsi, la structure des populations des psylles du complexe *C. pruni* semble se dessiner à deux échelles spatiales. A l'échelle du continent européen, la structure des populations, en particulier celle de l'espèce B, n'est pas sans rappeler celle des populations de grands mammifères (Kholodova et al., 2006; Petzold and Hassanin, 2020; Sharma et al., 2020). Les clusters couvrent de larges aires géographiques et ne sont contraints que par les obstacles majeurs (mer, grandes chaînes de montagnes). Des variations génétiques s'observent au sein de l'espèce B le long d'un gradient géographique. Ce phénomène pourrait annoncer de futures divergences entre les populations les plus éloignées. Néanmoins, les forts niveaux d'admixture observé au sein des deux espèces, en particulier pour l'espèce A indiquent un certain maintien des flux de gènes. Ces derniers favorisent l'homogénéité génétique observée chez les individus A et les individus B à l'échelle de la France par exemple. Cette homogénéité s'exprime également par une absence de déséquilibre de liaison entre les locus concernés par l'étude. Des clivages apparaissent cependant entre des populations échantillonnées à la même période distantes de 5 à 50 km. Cette observation à l'échelle des bassins de production de fruitiers est en adéquation avec les résultats de Marie-Jeanne et al. (Marie-Jeanne et al., 2020).

## Conclusion

Cette étude a révélé plusieurs phénomènes fondamentaux concernant le complexe spécifique *Cacopsylla pruni*. Tout d'abord, elle atteste l'existence de deux groupes génétiques distincts qui peuvent être considérés comme deux espèces à part entière. L'existence d'une barrière reproductive quasi-étanche, le constat d'un nombre important de mutations inter-groupes observées sur toutes les séquences analysées, l'observation de structures intra-groupes différentes sont autant d'éléments qui supportent cette affirmation. Cette observation conforte les études menées précédemment sur des jeux de données moins importants.

A travers l'analyse des structures intra-groupes, j'ai pu observer plusieurs schémas de distributions géographique et génétique des individus. Les deux espèces ont montré une structuration génétique, avec une espèce A présentant des groupes plus ou moins bien délimités en fonction des méthodes employées pour les mettre en évidence. Les groupes au sein de l'espèce A ont toutefois montré une grande connexion révélant des flux de gènes à plus ou moins longues distances à l'échelle géographique analysée (plusieurs centaines de kilomètres). Des phénomènes de spéciation au sein de cette espèce semblent peu probables étant donné le

brassage génétique et l'absence de sélection diversifiante qui la caractérisent. L'espèce B en revanche est distribuée sur une aire géographique plus large, ce qui se semble se traduire par un fort isolement par la distance entre les populations les plus éloignées géographiquement. Ainsi, au moins deux groupes génétiques se dessinent au sein de cette espèce selon un gradient ouest-est. Toutefois, des séquences génétiques identiques se retrouvent de la Turquie à la Bretagne. Ce phénomène peut s'interpréter comme le maintien de flux de gènes sur ces distances ou plus probablement attesterait d'une expansion rapide de l'aire de distribution de cette espèce.

À l'échelle des bassins de production, des structures plus fines s'observent pour les deux espèces. Certaines populations distantes de quelques dizaines de kilomètres appartiennent parfois à des groupes génétiques différents. Ces observations qui soutiennent l'étude de Marie-Jeanne et al. (2020) nécessiteraient d'être approfondies au cours de prochaines études.

Ainsi, nos analyses ont permis d'identifier les zones prioritaires à analyser dans une prochaine étude comme par exemple la région des Pyrénées ou des Alpes qui semble constituer une barrière pour l'espèce B mais pas pour l'espèce A ou encore la région Parisienne à plus fine échelle.

D'un point de vue pédagogique, ce stage est une réussite puisque j'ai acquis les outils et les connaissances nécessaires pour mener à bien d'autres études similaires, voire focalisées sur d'autres modèles biologiques.

## Perspectives

Afin de mieux comprendre les modes de transmission de CaPp, nous avons envisagé d'étudier les schémas géographiques de migration de *C. pruni*. En effet, à l'heure actuelle, les distances et directions de migrations des psylles ne sont pas réellement connues et il n'existe pas de carte ni de modèle représentant fidèlement la structure des populations de *C. pruni*. Les outils de génétique des populations permettent d'estimer indirectement les capacités de déplacement des insectes lors de leur migration.

Dans ce but l'équipe qui m'a accueilli en stage a développé 16 nouveaux marqueurs microsatellites. J'ai moi-même participé à la validation de ces marqueurs (tests sur quelques populations de références) avant leur utilisation ultérieure sur un plus grand nombre de populations. Il est ainsi envisagé d'en utiliser une trentaine parmi celles déjà disponibles (Sauvion, 2020b; Sauvion et al., 2021) pour tester différents scénarios de dispersion des psylles à différentes échelles spatio-temporelles. A terme, en combinant ces approches avec des modèles de dynamique des populations, l'objectif est de proposer des outils d'aide à la décision pour améliorer l'épidémiologie-surveillance de l'ESFY et ainsi proposer une meilleure prophylaxie plus respectueuse de l'environnement (BEYOND project, <https://www6.inrae.fr/beyond/>).



## Bibliographie :

- Andrews, K.R., Copus, J.M., Wilcox, C., Williams, A.J., Newman, S.J., Wakefield, C.B., Bowen, B.W., 2020. Range-Wide Population Structure of 3 Deepwater *Eteline* Snappers Across the Indo-Pacific Basin. *Journal of Heredity* 111, 471–485.  
<https://doi.org/10.1093/jhered/esaa029>
- Arnaud, A., Sauvion, N., Lavagne, C., Bonnet, H., Saint-Léger, M., Cavallazzi, Y., Langénieux, J.-P., 2013. Interactions insectes-plantes.
- Blumel, J.K., Derlink, M., Pavlovčič, P., Russo, I.-R.M., Andrew King, R., Corbett, E., Sherrard-Smith, E., Blejec, A., Wilson, M.R., Stewart, A.J.A., Symondson, W.O.C., Virant-Doberlet, M., 2014. Integrating vibrational signals, mitochondrial DNA and morphology for species determination in the genus *Aphrodes* (Hemiptera: Cicadellidae): Species determination in *Aphrodes*. *Syst Entomol* 39, 304–324.  
<https://doi.org/10.1111/syen.12056>
- Brown, J.K. (Ed.), 2016. Vector-Mediated Transmission of Plant Pathogens. The American Phytopathological Society. <https://doi.org/10.1094/9780890545355>
- Burckhardt, D., Hodkinson, I.D., 1986. A revision of the west Palaearctic pear psyllids (Hemiptera: Psyllidae). *Bull. Entomol. Res.* 76, 119–132.  
<https://doi.org/10.1017/S0007485300015340>
- Carraro, L., Osler, R., Loi, N., Ermacora, P., Refatti, E., 1998. Transmission of European stone fruit yellows phytoplasma by *Cacopsylla pruni*. *Journal of Plant Pathology* 80, 233–239.
- Chen, C., Durand, E., Forbes, F., François, O., 2007. Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Mol Ecol Notes* 7, 747–756. <https://doi.org/10.1111/j.1471-8286.2007.01769.x>
- Cho, G., Malenovský, I., Burckhardt, D., Inoue, H., Lee, S., 2020. DNA barcoding of pear psyllids (Hemiptera: Psylloidea: Psyllidae), a tale of continued misidentifications. *Bulletin of Entomological Research* 110, 521–534.  
<https://doi.org/10.1017/S0007485320000012>
- Danet, J.L., Balakishiyeva, G., Cimerman, A., Sauvion, N., Marie-Jeanne, V., Labonne, G., Laviña, A., Batlle, A., Križanac, I., Škorić, D., Ermacora, P., Serçe, Ç.U., Çağlayan, K., Jarausch, W., Foissac, X., 2011. Multilocus sequence analysis reveals the genetic diversity of European fruit tree phytoplasmas and supports the existence of inter-species recombination. *Microbiology* 157, 438–450. <https://doi.org/10.1099/mic.0.043547-0>
- Durand, E., Chen, C., François, O., 2009a. Tess version 2.3 - Reference Manual August 2009 30.
- Durand, E., Jay, F., Gaggiotti, O.E., François, O., 2009b. Spatial Inference of Admixture Proportions and Secondary Contact Zones. *Molecular Biology and Evolution* 26, 1963–1973. <https://doi.org/10.1093/molbev/msp106>
- François, O., Durand, E., 2010. Spatially explicit Bayesian clustering models in population genetics: SPATIAL CLUSTERING MODELS. *Molecular Ecology Resources* 10, 773–784. <https://doi.org/10.1111/j.1755-0998.2010.02868.x>
- Goldstein, D.B., Schlötterer, C. (Eds.), 1999. *Microsatellites: evolution and applications*. Oxford University Press, Oxford ; New York.

- Guillot, G., Estoup, A., Mortier, F., Cosson, J.F., 2005a. A Spatial Statistical Model for Landscape Genetics. *Genetics* 170, 1261–1280.  
<https://doi.org/10.1534/genetics.104.033803>
- Guillot, G., Mortier, F., Estoup, A., 2005b. Geneland: a computer package for landscape genetics. *Molecular Ecology Notes* 5, 712–715. <https://doi.org/10.1111/j.1471-8286.2005.01031.x>
- Herrbach, E., Sauvion, N., Boudon-Padieu, E., Lett, J.-M., Reynaud, B., Sforza, R., 2013. Chapitre 34. Une relation trophique originale : la vexion entomophile d’agents pathogènes, in: Calatayud, P.-A., Marion-Poll, F., Sauvion, N., Thiéry, D. (Eds.), *Interactions insectes-plantes*. IRD Éditions, pp. 511–548.  
<https://doi.org/10.4000/books.irdeditions.22614>
- Hodkinson, I.D., 2009. Life cycle variation and adaptation in jumping plant lice (Insecta: Hemiptera: Psylloidea): a global synthesis. *Journal of Natural History* 43, 65–179.  
<https://doi.org/10.1080/00222930802354167>
- Jarausch, B., Sauvion, N., Jarausch, W., 2013. Spread of European fruit tree phytoplasma diseases. *Phyt. Moll.* 3, 25. <https://doi.org/10.5958/j.2249-4677.3.1.006>
- Jarausch, B., Tedeschi, R., Sauvion, N., Gross, J., Jarausch, W., 2019. Psyllid Vectors, in: Bertaccini, A., Weintraub, P.G., Rao, G.P., Mori, N. (Eds.), *Phytoplasmas: Plant Pathogenic Bacteria - II: Transmission and Management of Phytoplasma - Associated Diseases*. Springer, Singapore, pp. 53–78. [https://doi.org/10.1007/978-981-13-2832-9\\_3](https://doi.org/10.1007/978-981-13-2832-9_3)
- Jombart, T., 2008. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24, 1403–1405. <https://doi.org/10.1093/bioinformatics/btn129>
- Jombart, T., Collins, C., 2017. A tutorial for Discriminant Analysis of Principal Components (DAPC) using adegenet 2.1.0 43.
- Keskin, E., Atar, H.H., 2011. Genetic divergence of *Octopus vulgaris* species in the eastern Mediterranean. *Biochemical Systematics and Ecology* 39, 277–282.  
<https://doi.org/10.1016/j.bse.2011.08.015>
- Kholodova, M.V., Milner-Gulland, E.J., Easton, A.J., Amgalan, L., Arylov, I.A., Bekenov, A., Grachev, I.A., Lushchekina, A.A., Ryder, O., 2006. Mitochondrial DNA variation and population structure of the Critically Endangered saiga antelope *Saiga tatarica*. *Oryx* 40, 103–107. <https://doi.org/10.1017/S0030605306000135>
- Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A., Mayrose, I., 2015. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources* 15, 1179–1191. <https://doi.org/10.1111/1755-0998.12387>
- Lee, I.-M., Davis, R.E., Gundersen-Rindal, D.E., 2000. Phytoplasma: Phytopathogenic Mollicutes. *Annu. Rev. Microbiol.* 54, 221–255.  
<https://doi.org/10.1146/annurev.micro.54.1.221>
- Leigh, J.W., Bryant, D., 2015. popart: full-feature software for haplotype network construction. *Methods in Ecology and Evolution* 6, 1110–1116.  
<https://doi.org/10.1111/2041-210X.12410>
- Levy, H., Clucas, G.V., Rogers, A.D., Leaché, A.D., Ciborowski, K.L., Polito, M.J., Lynch, H.J., Dunn, M.J., Hart, T., 2016. Population structure and phylogeography of the Gentoo Penguin (*Pygoscelis papua*) across the Scotia Arc. *Ecology and Evolution* 6, 1834–1853. <https://doi.org/10.1002/ece3.1929>

- Marie-Jeanne, V., Bonnot, F., Thébaud, G., Peccoud, J., Labonne, G., Sauvion, N., 2020. Multi-scale spatial genetic structure of the vector-borne pathogen '*Candidatus* Phytoplasma prunorum' in orchards and in wild habitats. *Scientific Reports* 10, 5002. <https://doi.org/10.1038/s41598-020-61908-0>
- Mayo, O., 2008. A Century of Hardy–Weinberg Equilibrium. *Twin Research and Human Genetics* 11, 249–256. <https://doi.org/10.1375/twin.11.3.249>
- Moreira, A.A., Tomás, A.R.G., Hilsdorf, A.W.S., 2011. Evidence for genetic differentiation of *Octopus vulgaris* (Mollusca, Cephalopoda) fishery populations from the southern coast of Brazil as revealed by microsatellites. *Journal of Experimental Marine Biology and Ecology* 407, 34–40. <https://doi.org/10.1016/j.jembe.2011.06.029>
- Ogden, R., Heap, E., McEwing, R., Tingay, R., Whitfield, D.P., 2015. Population structure and dispersal patterns in Scottish Golden Eagles *Aquila chrysaetos* revealed by molecular genetic analysis of territorial birds. *Ibis* 157, 837–848. <https://doi.org/10.1111/ibi.12282>
- Ouvrard, D., 2021. Psyl'list. <https://doi.org/10.5519/0029634>
- Padial, J.M., Miralles, A., De la Riva, I., Vences, M., 2010. The integrative future of taxonomy. *Front Zool* 7, 16. <https://doi.org/10.1186/1742-9994-7-16>
- Papadopoulos, L.N., Peijnenburg, K.T.C.A., Luttikhuisen, P.C., 2005. Phylogeography of the calanoid copepods *Calanus helgolandicus* and *C. euxinus* suggests Pleistocene divergences between Atlantic, Mediterranean, and Black Sea populations | SpringerLink [WWW Document]. URL <https://link.springer.com/article/10.1007/s00227-005-0038-x> (accessed 8.3.21).
- Peccoud, J., Labonne, G., Sauvion, N., 2013. Molecular Test to Assign Individuals within the *Cacopsylla pruni* Complex. *PLoS ONE* 8, e72454. <https://doi.org/10.1371/journal.pone.0072454>
- Peccoud, J., Pleydell, D.R.J., Sauvion, N., 2018. A framework for estimating the effects of sequential reproductive barriers: Implementation using Bayesian models with field data from cryptic species: BRIEF COMMUNICATION. *Evolution* 72, 2503–2512. <https://doi.org/10.1111/evo.13595>
- Petzold, A., Hassanin, A., 2020. A comparative approach for species delimitation based on multiple methods of multi-locus DNA sequence analysis: A case study of the genus *Giraffa* (Mammalia, Cetartiodactyla) 28.
- Popescu, I., Caudullo, G., 2016. *Prunus spinosa* in Europe: distribution, habitat, usage and threats 1.
- Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155, 945–959. <https://doi.org/10.1093/genetics/155.2.945>
- Richardson, J.L., Burak, M.K., Hernandez, C., Shirvell, J.M., Mariani, C., Carvalho-Pereira, T.S.A., Pertile, A.C., Panti-May, J.A., Pedra, G.G., Serrano, S., Taylor, J., Carvalho, M., Rodrigues, G., Costa, F., Childs, J.E., Ko, A.I., Caccone, A., 2017. Using fine-scale spatial genetics of Norway rats to improve control efforts and reduce leptospirosis risk in urban slum environments. *Evol Appl* 10, 323–337. <https://doi.org/10.1111/eva.12449>
- Rousset, F., 2020. 2020\_Rousset\_Genepop4.7-1.pdf.
- Rousset, F., 2008. genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources* 8, 103–106. <https://doi.org/10.1111/j.1471-8286.2007.01931.x>

- Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J.C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S.E., Sánchez-Gracia, A., 2017. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Molecular Biology and Evolution* 34, 3299–3302. <https://doi.org/10.1093/molbev/msx248>
- Sauvion, N., 2020a. Enroulement chlorotique de l'abricotier : situation en France - Phytoma [WWW Document]. URL <http://archives.phytoma-ldv.com/archivephytoma/article/enroulement-chlorotique-de-l-abricotier-situation-en-france-PH73500901.html> (accessed 1.17.21).
- Sauvion, N., 2020b. Compilation of occurrence data for two psyllid species of the *Cacopsylla pruni* complex (Hemiptera: Psylloidea). <https://doi.org/10.15454/VC9UR5>
- Sauvion, N., Lachenaud, O., Genson, G., Rasplus, J.-Y., Labonne, G., 2007. Are there several biotypes of *Cacopsylla pruni*? 2.
- Sauvion, N., Lachenaud, O., Mondor-Genson, G., Rasplus, J.-Y., Labonne, G., 2009. Nine polymorphic microsatellite loci from the psyllid *Cacopsylla pruni* (Scopoli), the vector of European stone fruit yellows. *Molecular Ecology Resources* 9, 1196–1199. <https://doi.org/10.1111/j.1755-0998.2009.02604.x>
- Sauvion, N., Peccoud, J., Meynard, C., Ouvrard, D., 2021. Occurrence data for the two cryptic species of *Cacopsylla pruni* (Hemiptera: Psylloidea). *BDJ* 9, e68860. <https://doi.org/10.3897/BDJ.9.e68860>
- Sauvion, N., THÉBAUT, G., Brun, L., 2012. Enroulement chlorotique de l'abricotier (ECA) 5.
- Sharma, R., Ahlawat, S., Sharma, H., Prakash, V., Shilpa, Khatak, S., Sawal, R.K., Tania, M.S., 2020. Identification of a new Indian camel germplasm by microsatellite markers based genetic diversity and population structure of three camel populations. *Saudi Journal of Biological Sciences* 27, 1699–1709. <https://doi.org/10.1016/j.sjbs.2020.04.046>
- Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Linde, A.V.D., 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64, 583–639. <https://doi.org/10.1111/1467-9868.00353>
- Steffek, R., Follak, S., Sauvion, N., Labonne, G., MacLeod, A., 2012. Distribution of ‘*Candidatus* Phytoplasma prunorum’ and its vector *Cacopsylla pruni* in European fruit-growing areas: a review. *EPPO Bull* 42, 191–202. <https://doi.org/10.1111/epp.2567>
- Thébaud, G., Yvon, M., Alary, R., Sauvion, N., Labonne, G., 2009. Efficient Transmission of ‘*Candidatus* Phytoplasma prunorum’ Is Delayed by Eight Months Due to a Long Latency in Its Host-Alternating Vector. *Phytopathology* 99, 265–273. <https://doi.org/10.1094/PHTO-99-3-0265>
- Vecchioni, L., Marrone, F., Arculeo, M., Fritz, U., Vamberger, M., 2020. Stand out from the Crowd: Small-Scale Genetic Structuring in the Endemic Sicilian Pond Turtle. *Diversity* 12, 343. <https://doi.org/10.3390/d12090343>
- Yannic, G., St-Laurent, M.-H., Ortego, J., Taillon, J., Beauchemin, A., Bernatchez, L., Dussault, C., Côté, S.D., 2016. Integrating ecological and genetic structure to define management units for caribou in Eastern Canada. *Conserv Genet* 17, 437–453. <https://doi.org/10.1007/s10592-015-0795-0>

## Annexe 1 :

### Glossaire

**Admixture** : résultat de la reproduction entre des individus issus de populations ancestrales reproductivement isolées. L'admixture peut être estimée à deux niveaux distincts, à l'échelle d'une population, le niveau d'admixture correspond à la proportion d'individus admixtés dans la population. A l'échelle des individus, le niveau d'admixture correspond à la proportion du génome hérité de chacune des populations parentales (Tang et al., 2005).

**Barcoding** : se définit comme la création et la lecture d'un code barre pour chaque espèce ou unité taxonomique. Il s'agit du développement de marqueurs dans le but d'identifier des espèces rapidement et avec une grande fiabilité. Plusieurs types de marqueurs sont utilisés :

- les marqueurs universels permettent une identification rapide des taxons présents dans un échantillon (ex. ARN 16S pour les bactéries) ;
- les marqueurs spécifiques permettent une identification plus précise des espèces au sein d'une famille ou d'un genre (ex. gène codant pour la cytochrome oxydase de type I, COI).

**Cline** : terme utilisé en génétique des populations pour faire référence à des tendances de fréquences alléliques ou de diversité génétique à larges échelles spatiales. Un cline dans les fréquences alléliques peut être le résultat d'une adaptation à un paramètre environnemental ou d'un phénomène d'admixture au cours d'un contact secondaire par exemple (François and Durand, 2010). Il peut représenter un gradient continu ou un changement abrupt pour le caractère étudié.

**'Candidatus Phytoplasma prunorum'** : Convention d'écriture des noms scientifiques pour les phytoplasmes : l'écriture en italique du nom "*Candidatus*" indique que « *Phytoplasma prunorum* » a été proposé pour nommer l'organisme responsable de la maladie « European stone fruit yellows » (ESFY) mais n'est pas encore validé par la communauté scientifique.

**Convergence des modèles** : Les modèles bayésiens fondés sur le principe des chaînes de Markov par méthode de Monte-Carlo (MCMC) reposent sur la réalisation d'un grand nombre d'itérations. Au cours du fonctionnement de ce type de modèle, le résultat de chaque itération dépend du résultat de l'itération précédente et uniquement de ce résultat. Ainsi ces modèles sont également appelés modèles sans mémoire car ils ne tiennent pas compte des autres itérations préalables. Au bout d'un certain nombre d'itérations, leurs résultats se ressemblent de plus en plus avant de devenir identiques jusqu'à la fin du fonctionnement du modèle. Ce dernier converge alors vers une solution unique ou un ensemble réduit de solutions inférées par la majorité des itérations. Cet état de convergence intervient généralement après une phase d'initialisation du modèle durant laquelle les solutions connaissent une variance maximale entre les itérations. Une fois la convergence atteinte, il est préférable de laisser tourner le modèle pendant un nombre important d'itérations afin de s'assurer qu'il n'entre pas dans une nouvelle phase chaotique après un semblant de convergence. Ainsi, la phase d'initialisation, comprenant un petit nombre d'itérations initiales, est détruite pour ne considérer que les résultats post-convergence.

**Délimitation spécifique** : concept utilisé en taxonomie pour définir sur quels critères et pour quels niveaux de différenciation il est possible de distinguer des espèces les unes des autres. La délimitation spécifique correspond également à la discipline de la biologie évolutive qui rassemble les outils et méthodes qui permettent ces distinctions (Yang and Rannala, 2014)

## Annexe 1 :

**Effet Whalund** : terme utilisé en génétique des populations signifiant une déviation du taux d'hétérozygotie au sein d'une population donnée par rapport au taux attendu pour une population à l'équilibre de Hardy – Weinberg. Cette déviation est due à la subdivision de la population en sous-populations de tailles différentes qui elles – mêmes peuvent être à l'équilibre de Hardy – Weinberg.

**Equilibre de Hardy-Weinberg (H-W)** : état d'équilibre défini pour une population de taille indéfiniment grande pour laquelle les forces évolutives comme la mutation ou la sélection naturelle sont négligées. Une telle population doit répondre aux critères suivants : diploïdie aux locus considérés, reproduction sexuée sous panmixie (association aléatoire des allèles), générations non chevauchantes, fréquences alléliques identiques chez les mâles et les femelles.

Soient deux allèles  $A$  et  $a$ , dont les fréquences au sein de cette population sont  $p$  et  $q$  ( $=1-p$ ), alors l'équilibre de Hardy-Weinberg est atteint lorsque les fréquences alléliques pour  $AA$ ,  $Aa$  et  $aa$  sont  $p^2$ ,  $2pq$  et  $q^2$  respectivement. Cet équilibre représente un point de repère permettant d'identifier des processus en génétiques des populations comme l'effet Whalund par exemple (Mayo, 2008).

**Flux de gènes** : représentation des mouvements et des déplacements de populations, de portion de populations ou d'individus, interprétés du point de vue génétique. Il s'agit de la façon dont les gènes sont transmis entre populations, par l'intermédiaire du déplacement et de la reproduction des individus. Par divers processus écologiques, environnementaux ou géographiques, en fonction des espèces considérées, ces flux de gènes peuvent être intensifiés, réduits, établis ou même rompus. Les conséquences de ces changements d'intensité de flux de gènes se répercutent à l'échelle des populations. La génétique des populations est la discipline qui cherche à identifier ces flux de gènes et leurs variations par des approches statistiques.

**Mue imaginale** : mue précédant le stade imago chez les insectes. Il s'agit de la dernière étape du développement larvaire durant laquelle l'insecte est doté des caractères qui le différencient d'un juvénile ou d'une larve.

**Organisme réservoir** : organisme au sein duquel un agent pathogène peut se maintenir, éventuellement se reproduire avant une nouvelle phase d'infection depuis cet organisme. Notion souvent abordée dans les relations plantes - pathogènes - insectes vecteurs dans lesquelles l'organisme réservoir peut être l'insecte, la plante ou les deux.

**Séquences microsatellites** : gènes caractérisés par la répétition en tandem de motifs d'une longueur variant entre 1 et 6 paires de bases (nucléotides). Leur fonction en tant qu'éléments faisant partie intégrante du génome reste encore obscure mais le fait qu'ils soient très abondants dans le génome des eucaryotes, en général très polymorphes, et qu'ils soient codominants en font d'excellents marqueurs pour les études de génétique des populations.

**Taxonomie intégrative** : approche en taxonomie qui consiste à définir une espèce et la différencier des autres non pas sur le seul critère morpho-anatomique mais vise à intégrer un maximum de composantes caractérisant l'espèce. Parmi ces caractéristiques, les critères constituant la niche écologique ainsi que les données génétiques sont les plus utilisées mais certains critères moins explorés peuvent également être pris en compte comme les sons émis par les individus (Blumel et al., 2014).

## Annexe 1 :

**Univoltin** : se dit d'un organisme qui ne produit qu'une seule génération par an, contrairement à un organisme multivoltin qui en produit plusieurs par an ou encore un organisme semi-voltin dont le développement nécessite plus d'une année.

## Références :

Bluemel, J.K., Derlink, M., Pavlovčič, P., Russo, I.-R.M., Andrew King, R., Corbett, E., Sherrard-Smith, E., Blejec, A., Wilson, M.R., Stewart, A.J.A., Symondson, W.O.C., Virant-Doberlet, M., 2014. Integrating vibrational signals, mitochondrial DNA and morphology for species determination in the genus *Aphrodes* (Hemiptera: Cicadellidae): Species determination in *Aphrodes*. Syst Entomol 39, 304–324. <https://doi.org/10.1111/syen.12056>

François, O., Durand, E., 2010. Spatially explicit Bayesian clustering models in population genetics: SPATIAL CLUSTERING MODELS. Molecular Ecology Resources 10, 773–784. <https://doi.org/10.1111/j.1755-0998.2010.02868.x>

Mayo, O., 2008. A Century of Hardy–Weinberg Equilibrium. Twin Research and Human Genetics 11, 249–256. <https://doi.org/10.1375/twin.11.3.249>

Tang, H., Peng, J., Wang, P., Risch, N.J., 2005. Estimation of individual admixture: Analytical and study design considerations. Genet. Epidemiol. 28, 289–301. <https://doi.org/10.1002/gepi.20064>

Yang, Z., Rannala, B., 2014. Unguided Species Delimitation Using DNA Sequence Data from Multiple Loci. Molecular Biology and Evolution 31, 3125–3135. <https://doi.org/10.1093/molbev/msu279>



## Annexe 2 :

### Précisions sur les plantes hôtes :

Le Prunellier ou Epine noire, *Prunus spinosa* (Linnaeus 1753), est l'espèce de *Prunus* sauvage la plus commune en France. Cet arbuste occupe une large gamme d'habitats et se rencontre en particulier dans les espaces semi-ouverts ou en cours de fermeture ainsi que dans les friches. Il est très commun dans les haies des paysages bocagés. Il est aujourd'hui encore un élément très important des corridors écologiques, favorisés par les actions de conservation écologiques dans les agroécosystèmes et sert de refuge à une importante diversité faunistique (Popescu and Caudullo, 2016).

D'autres espèces de prunus sauvages servent également de plante-hôte pour *C. pruni* : le myrobolan, *Prunus cerasifera* (Ehrhart 1784), assez commun dans la région de Perpignan ; le cerisier de Sainte Lucie, *Prunus mahaleb* (Linnaeus 1753), décrit comme plante-hôte de *C. pruni* en Espagne (Sabaté et al., 2016). )\_



Figure 1 : *P. spinosa* en fleur, INPN



Figure 3 : *P. cerasifera* en fleur, © Hugues Tinguy, INPN



Figure 2 : rameau de *P. mahaleb* Sébastien en fleur, © Sébastien Filoche



## Annexe 3 :

### Diversité des insectes vecteurs

*Cacopsylla pruni* appartient à l'ordre des hémiptères, le groupe d'insectes qui comprend la grande majorité des insectes vecteurs de pathogènes des plantes. Afin de comprendre l'importance de ces insectes en tant que vecteur, il est nécessaire de définir au préalable la vection *stricto sensu*.

La vection par les insectes est le processus actif par lequel un insecte transmet un agent pathogène d'un organisme infecté à un organisme sain (Herrbach et al., 2013). Les hémiptères ont des pièces buccales très particulières qui leur permettent d'enfoncer des stylets dans les tissus végétaux jusqu'aux vaisseaux conducteurs de sève du phloème ou du xylème, une sève dans laquelle ces insectes pourront trouver des éléments nutritifs (sucres, acides aminés essentiellement). Au cours de ce comportement alimentaire, les hémiptères rejettent de la salive ou aspirent de la sève. C'est à ce moment-là qu'ils peuvent respectivement inoculer ou acquérir un phytopathogène. En quelque sorte, l'agent pathogène « se sert » de son hôte-vecteur pour franchir les barrières physiques (ex. : parois cellulaires) ou physiologiques (mécanismes de défense) de son hôte-plante (Herrbach et al., 2013).

Il existe une ambiguïté dans l'utilisation du terme d'insecte vecteur dans la littérature. Certaines bactéries par exemple sont véhiculées par les insectes de manière passive d'une plante à l'autre. *Erwinia amylovora* (Proteobacteria : Enterobacteriaceae) par exemple est transportée par des insectes pollinisateurs (abeilles domestiques, papillons) et déposée avec le pollen sur le péristyle de la fleur visitée (Nadarasah et Stavrinides, 2011). Certains auteurs (Nadarasah et Stavrinides, 2011; Rutikanga et al., 2015) désignent ces insectes comme des vecteurs alors que l'internalisation de ces bactéries n'a pas été démontrée chez ces insectes. Ce processus de transport est plus généralement appelé dissémination afin de bien le différencier de la vection (Herrbach et al., 2013).

Chez les hémiptères, le rôle de transmission des phytoplasmes est principalement joué par les cicadelles (Membracoidae : Cicadomorpha ; Figure x)). Plusieurs espèces de psylles sont décrits comme vecteurs de phytoplasmes, à travers des relations souvent spécialisées et exclusives. Ainsi, les phytoplasmes du groupe 16Sr X sont transmis par les psylles du genre *Cacopsylla* (Jarausch et al. 2019).

'*Candidatus. Phytoplasma prunorum*' appartient au groupe 16Sr X-B (Seemüller et al., 1998). Cette bactérie est très spécifiquement associée au pathosystème *Cacopsylla pruni* – prunus, un système complexe dont le fonctionnement a fait l'objet d'un article récent (Marie-Jeanne et al. 2020)

## Annexe 3 :

Cependant, il existe d'autres pathosystèmes impactant la culture de Rosacées cultivés, impliquant des phytoplasmes de même groupe X. Ainsi, '*Candidatus Phytoplasma pyri*' (= groupe 16Sr X-C) est transmis aux poiriers par *Cacopsylla pyri* (Linnaeus, 1760) et *Cacopsylla pyricola* (Förster, 1848), et provoque une maladie appelée le Pear Decline. À l'instar de *C. pruni*, ces deux espèces de psylles sont vectrices du phytoplasme chez les plantes du genre *Pyrus* exclusivement mais sont compatibles avec plusieurs espèces de ce genre (Jarausch et al., 2013). Des taux d'acquisition et d'efficacité de transmission variables entre ces 2 espèces ont été observés en fonction de l'aire géographique examinée (Jarausch et al., 2013). Certains individus d'une troisième espèce de psylles des poiriers, *Cacopsylla pyrisuga* (Förster, 1848) ont également été échantillonnés porteurs de la bactérie mais la transmission aux poiriers par l'intermédiaire de cette espèce n'a pas été prouvée (Jarausch et al., 2013).

Un troisième phytoplasme du groupe X, '*Candidatus Phytoplasma mali*', provoque une maladie inféodée aux pommiers (*Malus* spp.), l'Apple Proliferation. Ce phytoplasme peut être transmis par les psylles, *Cacopsylla picta* (Förster, 1848) et *Cacopsylla melanoneura* (Förster, 1848). Toutefois, *C. melanoneura*, est polyphage et se nourrit sur plusieurs types de plantes de la famille des rosacées comme les aubépines, différentes espèces du genre de pommiers (*Malus* spp.), les néfliers ou encore les poiriers. En revanche, *C. picta* se nourrit et se reproduit uniquement sur les pommiers (Barthel et al., 2020). La prise en compte dans un pathosystème d'un si grand nombre d'organismes potentiellement acteurs au sein du pathosystème nécessite d'importants moyens d'analyses ne serait-ce que pour identifier le rôle de chaque organisme.

## Références :

- Barthel, D., Kerschbamer, C., Panassiti, B., Malenovský, I., Janik, K., 2020. Effect of Daytime and Tree Canopy Height on Sampling of *Cacopsylla melanoneura*, a '*Candidatus Phytoplasma mali*' Vector. *Plants* 9, 1168. <https://doi.org/10.3390/plants9091168>
- Herrbach, E., Sauvion, N., Boudon-Padieu, E., Lett, J.-M., Reynaud, B., Sforza, R., 2013. Chapitre 34. Une relation trophique originale : la vection entomophile d'agents pathogènes, in: Calatayud, P.-A., Marion-Poll, F., Sauvion, N., Thiéry, D. (Eds.), *Interactions insectes-plantes*. IRD Éditions, pp. 511–548. <https://doi.org/10.4000/books.irdeditions.22614>
- Jarausch, B., Sauvion, N., Jarausch, W., 2013. Spread of European fruit tree phytoplasma diseases. *Phyt. Moll.* 3, 25. <https://doi.org/10.5958/j.2249-4677.3.1.006>
- Nadarasah, G., Stavrinos, J., 2011. Insects as alternative hosts for phytopathogenic bacteria. *FEMS Microbiol Rev* 35, 555–575. <https://doi.org/10.1111/j.1574-6976.2011.00264>

### Annexe 3 :

Rutikanga, A., Night, G., Tusiime, G., Ocimati, W., Blomme, G., 2015. Spatial and Temporal Distribution of Insect Vectors of *Xanthomonas campestris* pv. *musacearum* and their Activity across Banana Cultivars Grown in Rwanda 15.

Seemüller, E., Marcone, C., Lauer, U., Ragozzino, A., Göschl, M., 1998. CURRENT STATUS OF MOLECULAR CLASSIFICATION OF THE PHYTOPLASMAS. Journal of Plant Pathology 80, 3–26.

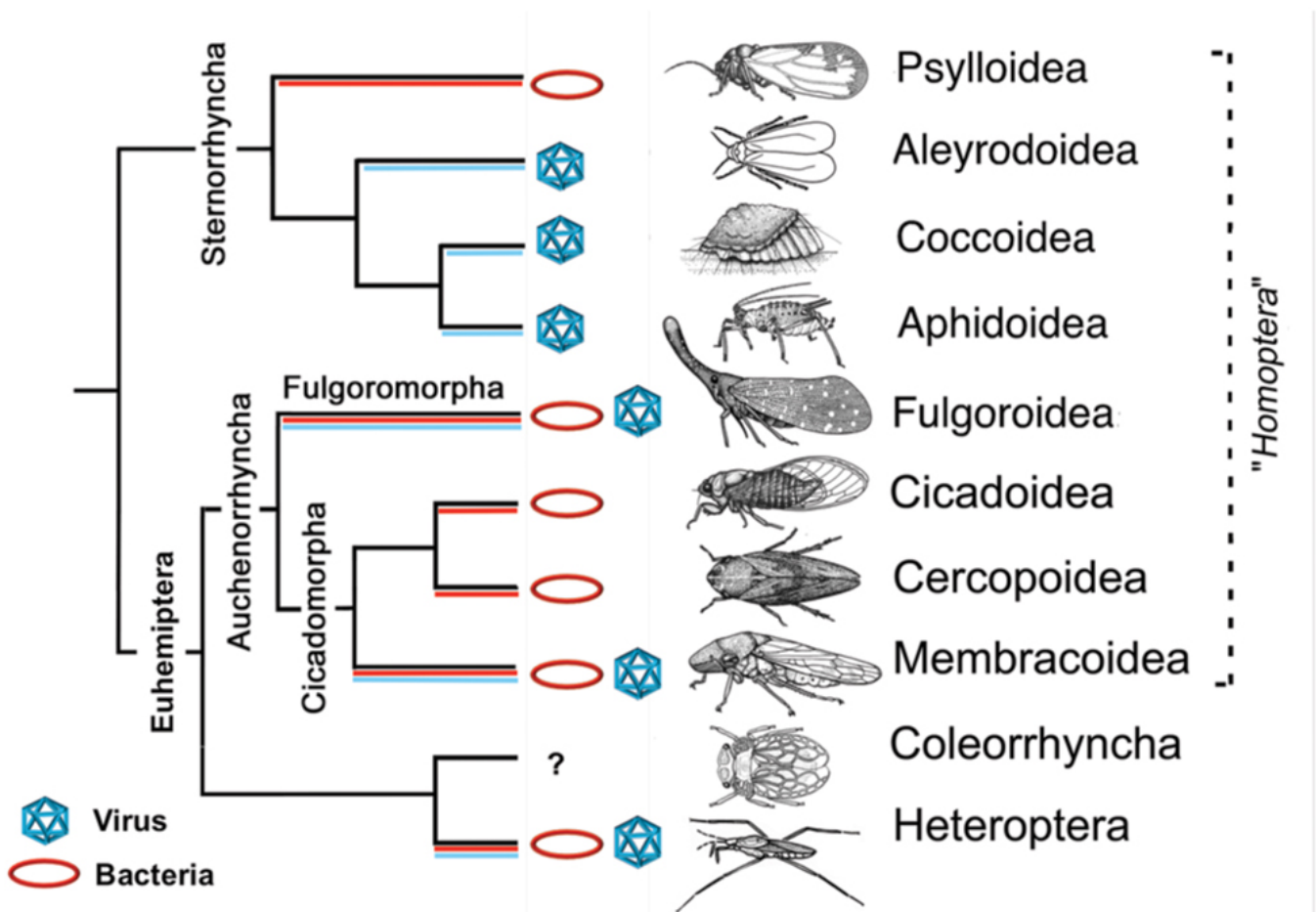


Figure 1 : Classification simplifiée de l'ordre des hémiptères et principaux organismes pathogènes vécés (Perilla-Henao and Casteel, 2016). D'un point de vue phylogénétique, cette classification a connu quelques modifications. Une version plus récente est présentée en figure 2

### Annexe 3 :

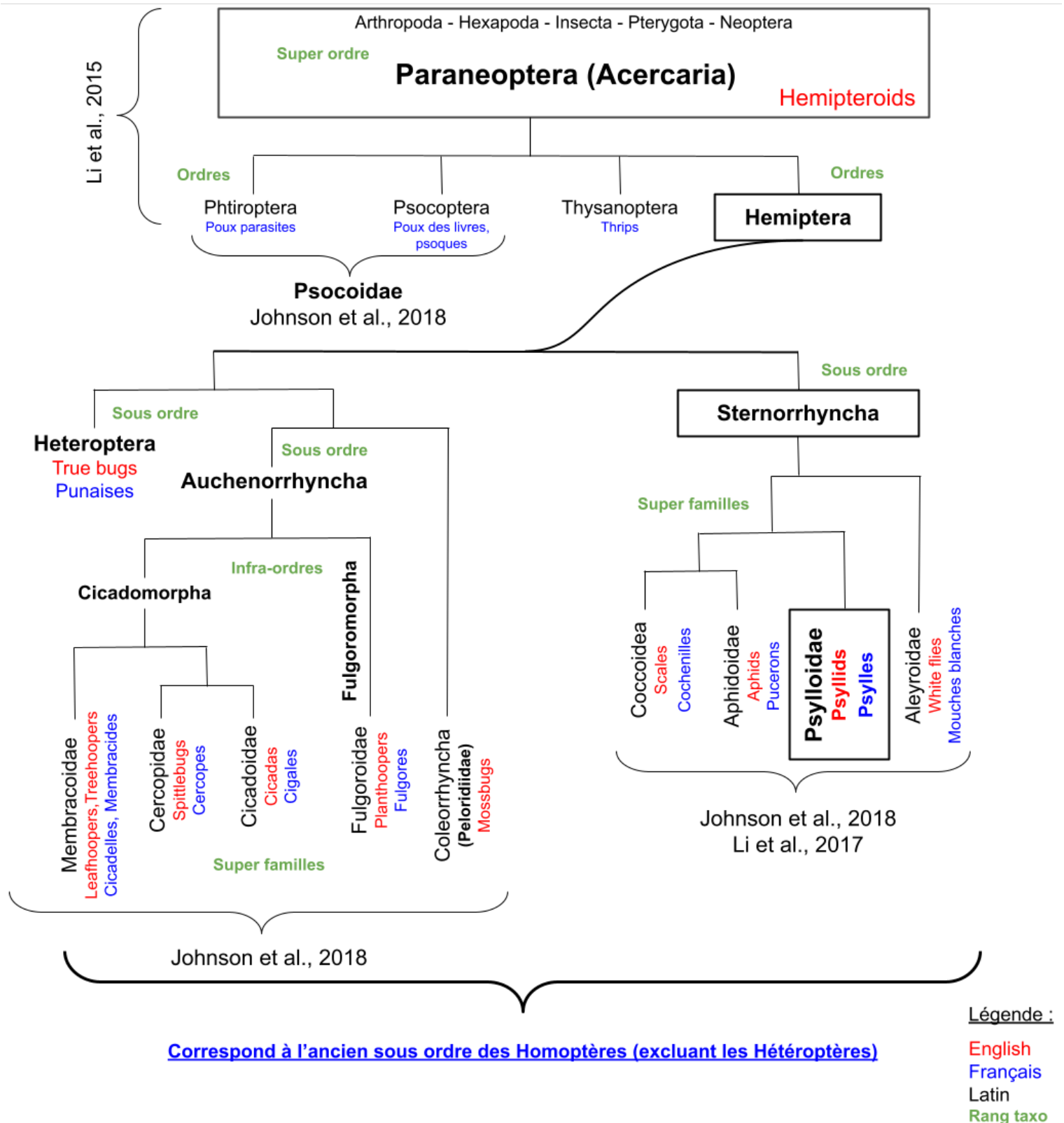


Figure 2 : Classification détaillée des différentes super familles de l'ordre des hémiptères. (Figure originale fondée sur Johnson et al., 2018 et Li et al., 2017).

## Annexe 3 :

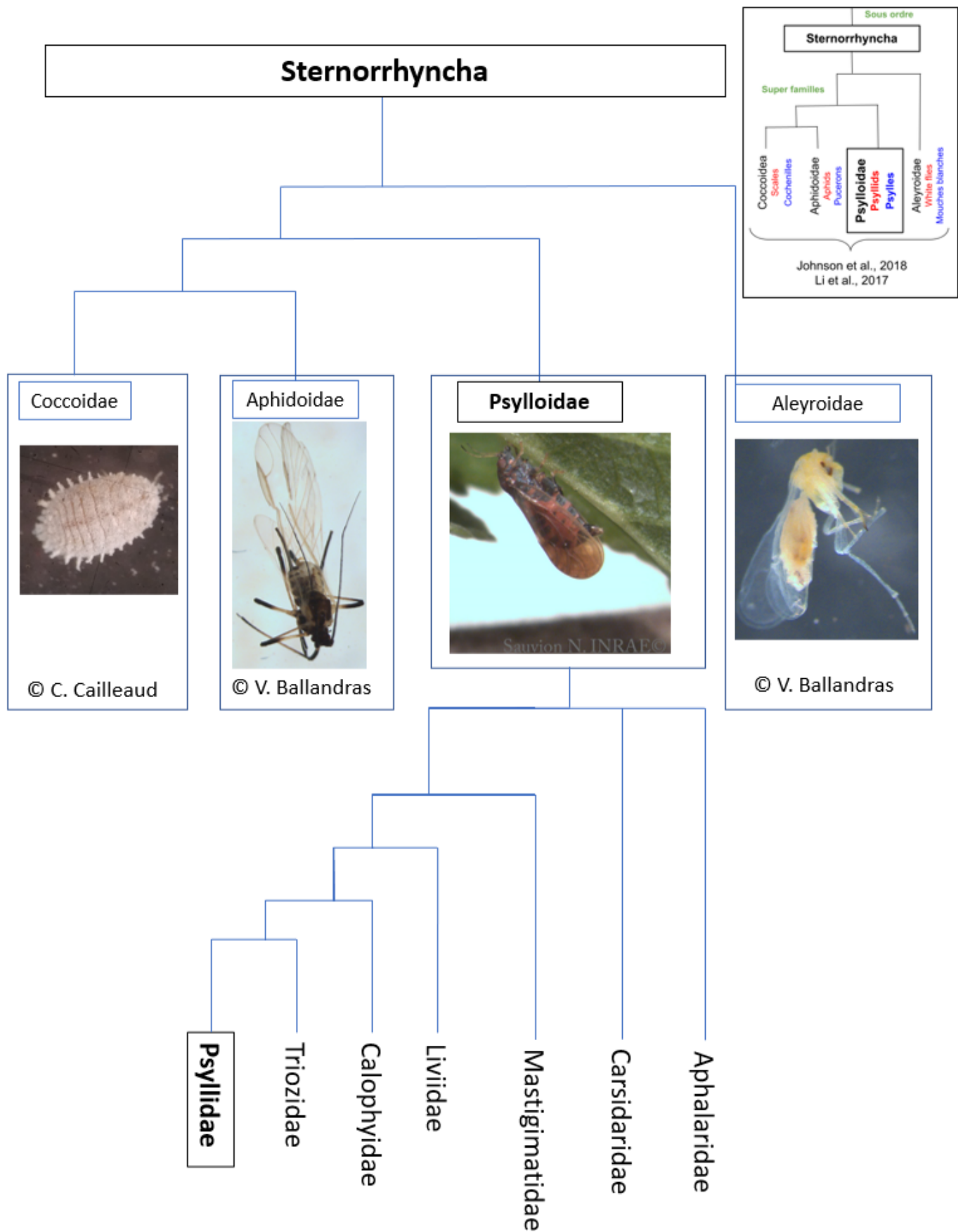


Figure 3 : Classification détaillée du sous ordre des Sternorrhyncha et des différentes familles de psylles fondée sur Cho et al., 2019 et Percy et al., 2018 . (Photos originales)



## Annexe 4 :

### Cartes d'occurrences de *Cacopsylla pruni*

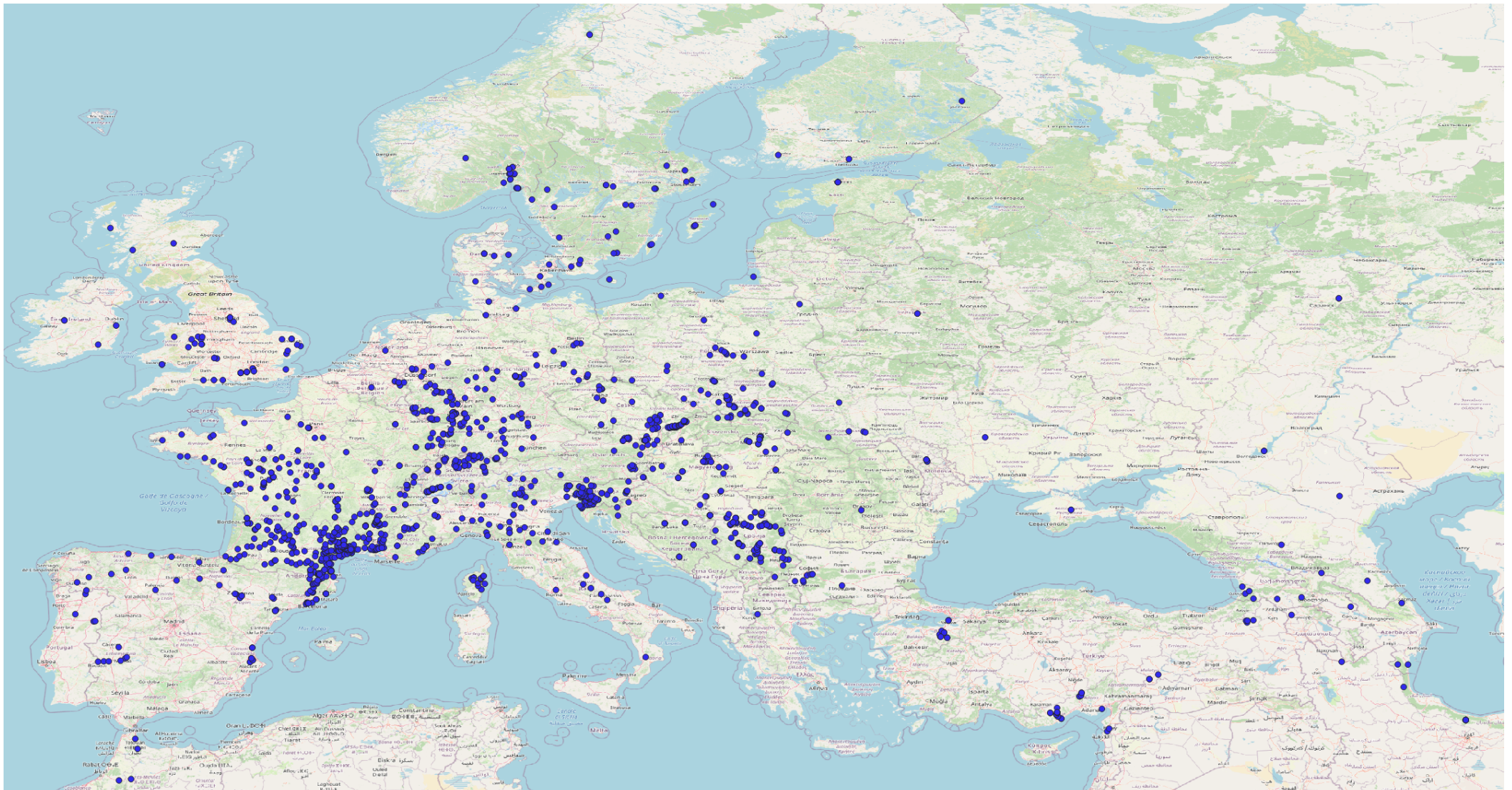


Figure 1 : Carte d'occurrences des 1975 localités échantillonnées par Nicolas Sauvion depuis 1998 dans toute l'Europe. Chaque point représente une population de psylles *Cacopsylla pruni* à une localité.



## Annexe 4 :

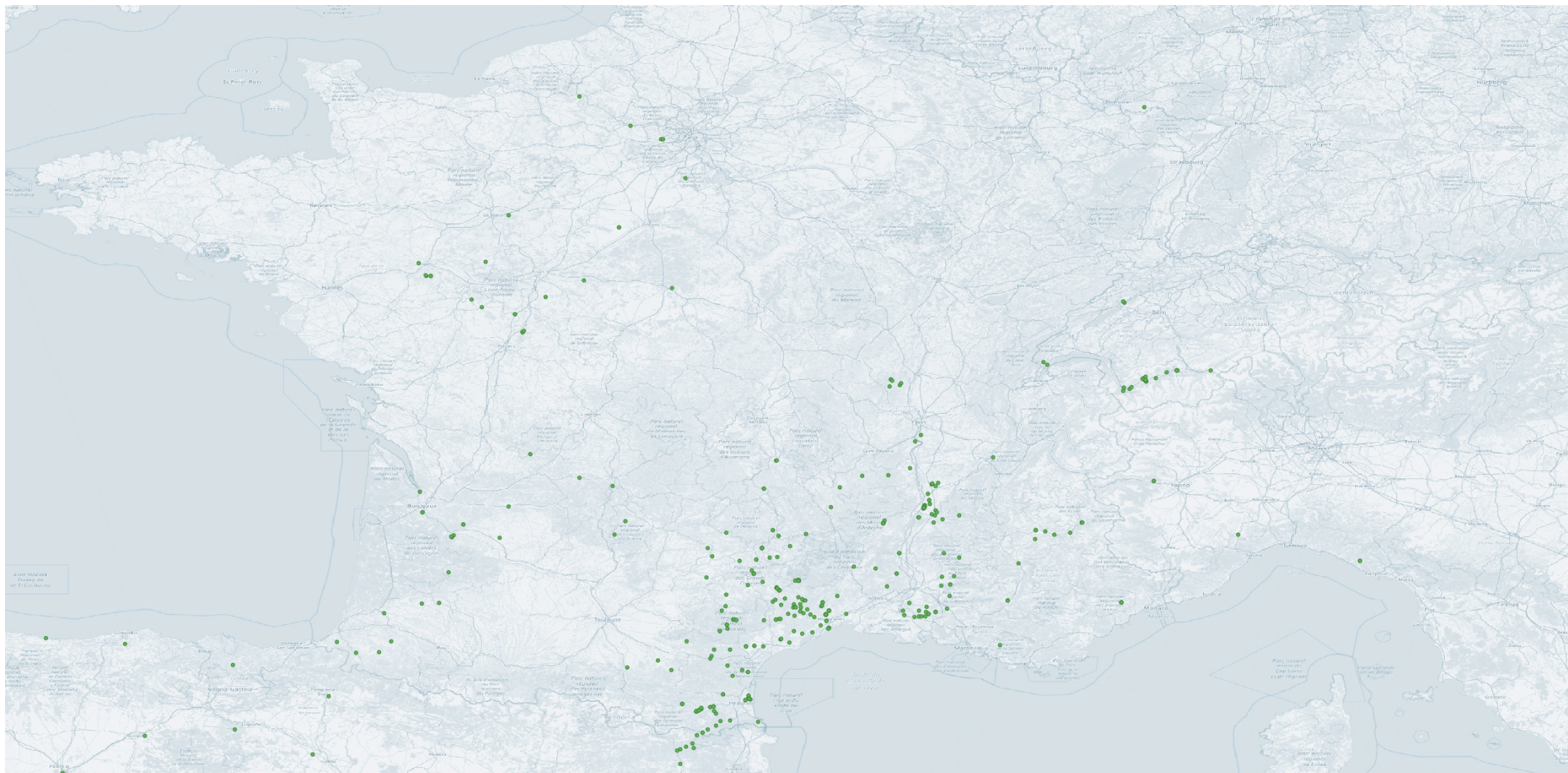


Figure 2 : Carte d'occurrences des psylles *Cacopsylla pruni* du groupe A échantillonnés par Nicolas Sauvion depuis 1998. Chaque point représente une population à une localité. La carte est centrée sur la région accumulant le plus de données.



## Annexe 4 :

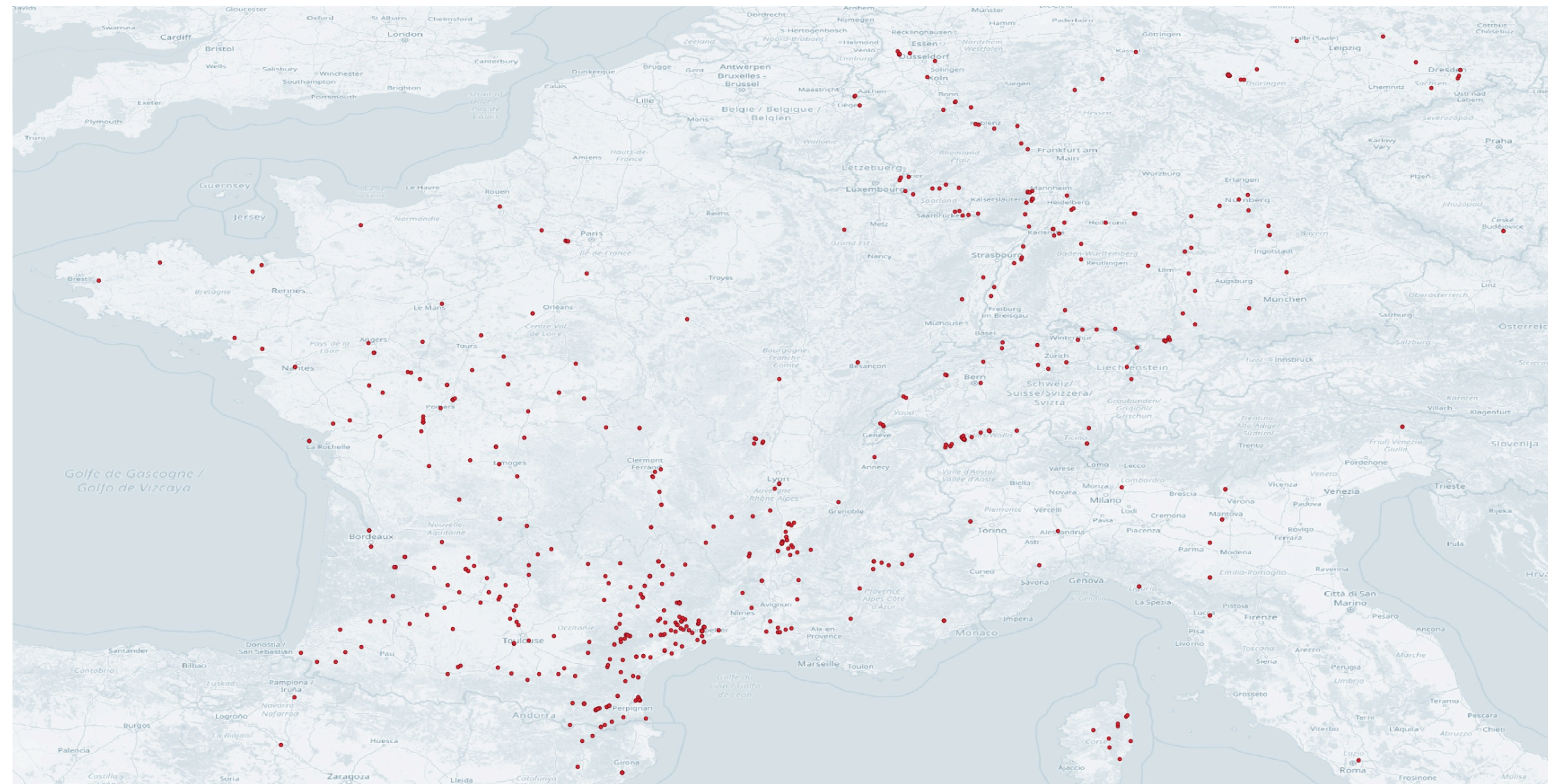


Figure 3 : Carte d'occurrences des psylles *Cacopsylla pruni* du groupe B échantillonnés par Nicolas Sauvion depuis 1998. Chaque point représente une population à une localité. La carte est centrée sur la région accumulant le plus de données.



## Annexe 5 :

### Librairie *adegenet*

#### Importation des données microsatellites :

Deux types d'objets sont pris en charge dans *adegenet*, les objets *genind* et les objets *genepop*. Les objets *genind* permettent de renseigner le génotype de chaque individu alors que les objets *genepop* permettent de regrouper ces individus par populations d'échantillonnage.

Un jeu de données peut être peut-être transformé en objet *genind* grâce à la fonction *df2genind*, à condition qu'il respecte un certain format.

Le jeu de données microsatellites était sous la forme suivante à son importation sur R :

Tableau 1: Format des données lors de leur importation sur R. La colonne ID correspond à l'identifiant de chaque individu, la colonne localité correspond à la localité d'échantillonnage, Les colonnes X et Y correspondent aux coordonnées géographiques de chaque population et les colonnes suivantes correspondent aux données microsatellites identifiées par locus.

	A	B	C	D	E	F	G	H	I	J	K
1	ID	localite	X	Y	X6_115.1	X6_115.2	X5_45.1	X5_45.2	X6_144.1	X6_144.2	X4_127.1
2	F207_17	Aigoual	3.52087	44.035855	129	132	119	119	72	75	363
3	F207_18	Aigoual	3.5208702	44.035855	132	132	119	119	81	81	363
4	F207_19	Aigoual	3.5208704	44.035855	129	132	119	119	75	105	367
5	F207_20	Aigoual	3.5208706	44.035855	129	132	119	119	75	75	359
6	F207_21	Aigoual	3.5208708	44.035855	129	132	119	119	78	84	375
7	F207_22	Aigoual	3.520871	44.035855	129	129	119	121	75	75	375
8	F207_23	Aigoual	3.5208712	44.035855	129	129	119	119	75	99	367
9	F207_24	Aigoual	3.5208714	44.035855	129	132	119	119	75	75 NA	
10	F207_25	Aigoual	3.5208716	44.035855	129	132	119	119	75	102 NA	

Le jeu de données étant composé d'individus diploïdes, chaque locus comportait 2 versions de la même séquence. Cependant, pour être importés sur R, chaque colonne devait être nommée indépendamment des autres. Sous ce format, l'objet *genind* créé considère le jeu de données comme composé de 16 locus haploïdes. J'ai alors dû concaténer les colonnes correspondant au même locus avec la fonction *paste* en utilisant un séparateur reconnu par la fonction *df2genind*. Cette fonction nécessite aussi de renseigner les caractères qui codent pour des données manquantes, ici, NA. L'utilisateur doit aussi renseigner la colonne du jeu de données qui contient les informations sur les localités d'échantillonnages.

Un objet *genind* peut être transformé en objet *genepop* grâce à la fonction *genind2genepop*. Des informations supplémentaires peuvent être incorporées dans ces objets comme par exemple les coordonnées géographiques par exemple. J'ai renseigné les coordonnées géographiques grâce à la fonction suivante : « *nom de l'objet genind* »@other\$xy = « *nom du jeu de données comportant les coordonnées* ». Celles-ci doivent absolument être dans l'ordre X puis Y. Ce procédé est indispensable pour effectuer les tests d'isolement par la distance car les distances géographiques entre les individus sont calculées grâce à ces coordonnées.

## Annexe 5 :

### DAPC

#### Fonction *find.clusters* :

Cette fonction permet de déterminer le nombre de cluster optimal pour représenter leur distribution sur les cartes factorielles créées par la fonction *dapc*. En premier lieu, *find.cluster* transforme les données par l'intermédiaire d'une analyse en composantes principales (ACP). A cet instant, l'utilisateur doit renseigner le nombre de composantes principales qu'il souhaite retenir dans l'analyse. En suite l'algorithme effectue le partitionnement en K moyennes en ajoutant un cluster supplémentaire à chaque étape. Chaque étape qui peut être identifiée comme une itération du K-means clustering est associée à une valeur de BIC. En retour, la fonction affiche le graphe des valeurs de BIC en fonction du nombre de cluster testés (Fig.1b). Le nombre de cluster le plus pertinent est celui auquel est associé une valeur de BIC minimale ou en pratique la valeur de BIC pour laquelle la courbe du graphe effectue un « coude » (Jombart and Collins, 2017).

Le BIC est un critère proche du critère d'Akaike (AIC). Il repose sur une fonction de vraisemblance qui, contrairement à ce dernier, prend en compte le nombre d'observations dans l'échantillon et le nombre de paramètres pris en compte dans le modèle. Le modèle sélectionné est celui associé à une vraisemblance maximale.

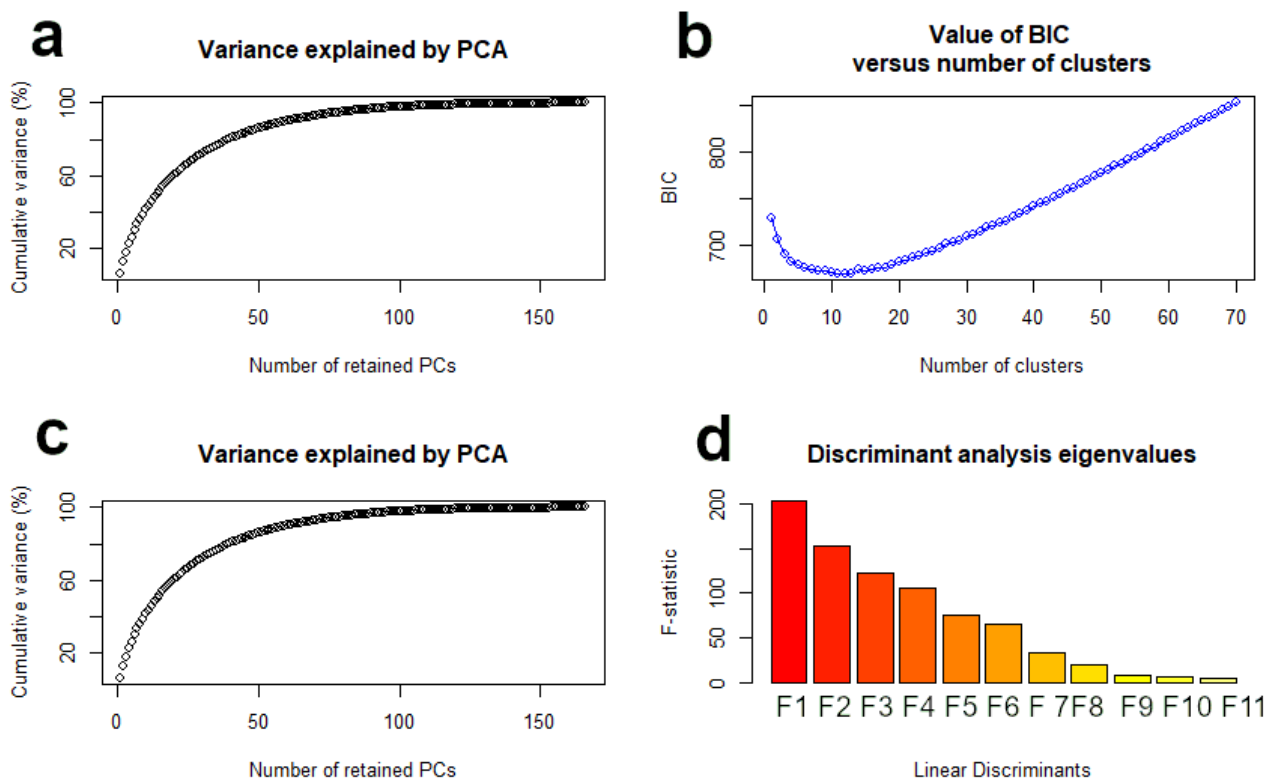


Figure 1 : Graphiques affichés interactivement par la fonction *find.cluster* a) et b) et *dapc* en c) et d) lors de la l'exécution de la simulation DAPC A 12.

## Annexe 5 :

Le graphique des valeurs de BIC en fonction du nombre de clusters (Fig.1b) montre que le nombre optimal de clusters à retenir pour cette simulation se situe entre 9 et 12.

### Fonction *dapc* :

Le graphique d de la figure 1 montre la part de variabilité portée par chaque fonction discriminante. Les fonctions qui représentent le mieux la variabilité inter-clusters (F1 et F2) ont été retenues pour construire la carte factorielle présentée en figure 2.

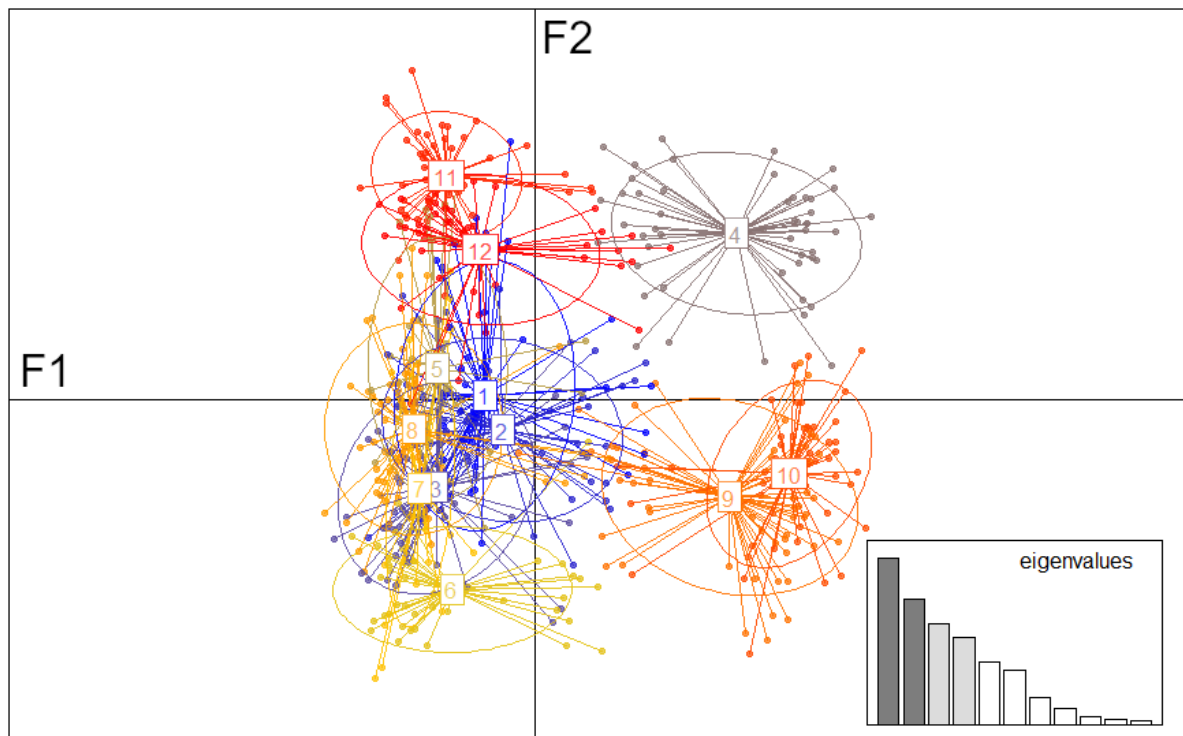


Figure 2 : Carte factorielle formée des fonctions discriminantes F1 et F2 de la simulation DAPC A 12

Dans un second temps, j'ai décidé de réduire le nombre de clusters représentés sur les cartes factorielles afin de mieux représenter les regroupements entre clusters. J'ai donc créé la figure 4a du rapport, correspondant à la simulation DAPC A 5 (tableau 2 du rapport). Le même procédé a été appliqué pour le groupe B.

### Isolement par la distance :

Les tests d'isolement par la distance ont été effectués avec la fonction *mantel.randtest*. Cette fonction compare des matrices de distances génétique et géographique, préalablement créées à partir des informations contenues dans les objets *genind*.

Pour chaque test d'isolement par la distance, un graphique peut être affiché avec la fonction *plot(mantel.randtest)*.

## Annexe 5 :

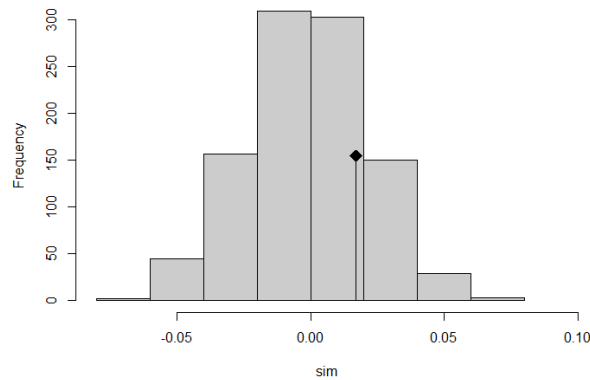


Figure 3 histogramme des simulations d'isolation par la distance pour le groupe A.

Sur la figure 3, Les barres grises montrent les simulations effectuées selon la méthode de Monte Carlo. La fonction effectue grand nombre de simulations en associant aléatoirement les distances génétiques aux distances géographiques trouvées dans les matrices. Ces itérations ont pour but de simuler des corrélations entre les deux types de distances sous l'effet du hasard. La gaussienne qui en résulte représente les valeurs de corrélations attendues sous l'hypothèse d'un isolement par la distance. Le point noir représente la valeur de la corrélation entre les distances génétique et distances géographiques entre les individus observée dans le jeu de données. La valeur observée pour le groupe A est de 0.017, associée à une P-value de 0.122. La corrélation observée pour le groupe A n'est pas significative, il n'y a donc pas d'isolement par la distance.

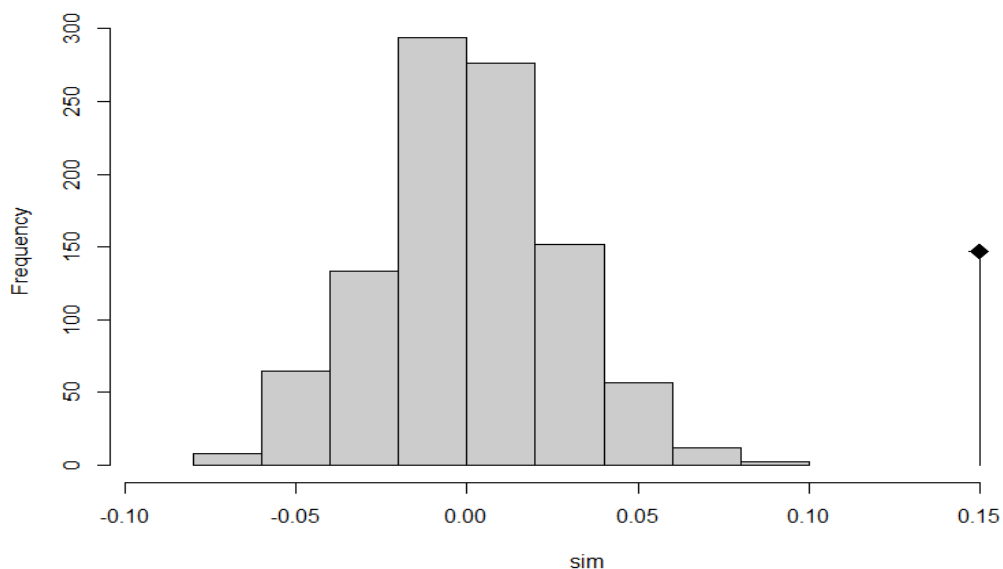


Figure 4 histogramme des simulations d'isolation par la distance pour le groupe B

Sur la figure 4, en revanche, on observe que cette corrélation est supérieure aux valeurs simulées sans isolement par la distance. Pour le groupe B, cette corrélation a une valeur de 0.150,

## Annexe 5 :

associée à une P-value de 0.001. Il y a donc bien un isolement par la distance significatif pour le groupe B.

### Taux de consanguinité :

Avec *adegenet*, les taux de consanguinités sont calculés avec la fonction *inbreeding*, elle associe à chaque individu une probabilité d'hériter deux allèles identiques issus d'un même ancêtre. La compilation de cette probabilité pour chaque individu m'a permis de créer le tableau suivant, répertoriant le nombre d'individu de chaque population ainsi que le nombres d'individus A et B pour lesquels la probabilité de consanguinité était supérieure à 0.5. La dernière colonne de ce tableau indique le pourcentage d'individus dont la probabilité de consanguinité était supérieure à 0.5, A et B réunis, par population.

Tableau 2 : Nombres d'individus consanguins par population

Localité	N. indiv	Consanguins A	Consanguins B	% d'individus consanguins
Aigoual	30	0	8	26,67
Angers Beaucouzé	5	0	0	0
Angers Chalbotières	5	0	0	0
Angers Plisson	2	0	0	0
Auxerre	30	0	14	46,67
Bellac	6	0	3	50
Belvezet1	14	1	0	7,14
Belvezet2	17	2	0	11,76
Belvezet3	29	1	0	3,49
Bouleternère	24	3	1	16,67
Cacak Rosci	8	0	3	37,5
Cacak Vracini	11	0	7	63,64
Chavenay	7	0	2	28,57
Col d'Ares	24	0	0	0
Col de Crie	22	0	7	31,82
Col de Jau	24	0	5	20,83
Communay	11	0	7	63,64
Coursegoules	30	5	6	36,67
La Couvertorade	24	2	0	8,33
Cuxac	11	0	1	9,09
Etoiles	24	4	2	25
Farcheville	23	1	3	17,39
Fesche	29	1	1	6,9
Fraisse-sur-Agout	5	0	4	80
GrabelsNG05	24	0	1	4,17
GrabelsNG07	24	4	4	33,33
GrabelsREIM06	24	2	2	16,67
GrabelsREIM07	23	0	1	4,35
Gramat	12	0	4	33,33

## Annexe 5 :

Hattonville	18	0	10	55,56
Les Monterniers	14	0	4	28,57
Les Natges	24	3	0	12,5
Levernois	24	0	7	29,17
Lorlanges	11	0	6	54,54
Lusignan1	18	0	7	38,89
Lusignan2	12	0	5	41,67
Millau Viaduc	23	0	1	4,35
Mirepoix	21	0	2	9,52
Moia	24	2	0	8,33
Montgamé	27	1	13	51,85
Montmarault	21	0	8	38,09
Neustadt	22	0	8	36,36
Pont du Châteaux	13	0	7	53,85
Prades	29	2	0	6,9
Prades le Lez	29	3	6	31,03
Rennemoulin	8	0	3	37,5
Romette	5	0	0	0
Séranne	29	0	14	48,28
Salvetat	13	1	6	53,85
San Pau	24	0	0	0
Seysses	10	0	5	50
Ste Baume	12	1	0	8,33
Tieule1	29	4	1	17,24
Tieule2	18	2	0	11,11
Tordera A	11	2	1	27,27
Tordera B	12	0	5	41,67
Torreilles	26	1	6	26,92
Udine	19	0	7	36,84
Villeneuve les Maguelonne	21	0	8	38,09
Villepreux	8	0	2	25
Vouillon	26	0	9	34,61

## Annexe 6 :

### STRUCTURE

#### Format des données

Le logiciel STRUCTURE ne supporte que des données enregistrées dans un fichier texte sous un certain format. Avant de démarrer un projet, qui rassemble toutes les simulations effectuées sur un même jeu de données, le logiciel affiche une fenêtre qui permet d'indiquer quel format de données prendre en compte. Chaque colonne du jeu de données doit être indiquée dans cette fenêtre au cours de l'importation des données. Le jeu de données de la présente étude comportait 8 marqueurs microsatellites diploïdes neutre, importé dans STRUCTURE sous la forme suivante :

popB structure.txt - Bloc-notes

Fichier	Edition	Format	Affichage	Aide												
6_115		5_45		6_144		4_127		5_43		6_15		6_129		4_108		
BJ403_14	-9	-9	-9	-9	-9	-9	363	363	136	136	122	165	191	191	117	136
BJ403_15	129	129	115	119	115	115	379	379	136	244	122	171	185	185	152	185
BJ403_16	129	129	119	119	78	78	239	239	132	136	185	185	191	191	136	136
BJ403_17	129	129	115	119	75	75	367	379	136	136	145	200	194	194	189	189
BJ403_18	129	129	119	119	75	75	379	379	136	136	162	185	194	194	148	161
BJ403_19	-9	-9	-9	-9	-9	-9	391	415	136	244	139	145	191	191	124	140
BJ403_20	132	132	119	119	78	78	359	359	132	244	151	153	194	194	136	177
BJ403_22	129	129	119	119	78	78	363	375	136	136	119	171	191	191	148	148
BJ403_23	-9	-9	-9	-9	-9	-9	375	375	136	136	142	153	191	191	148	148
BJ403_24	129	129	119	119	72	72	375	375	136	136	130	171	191	194	120	165
BJ403_25	129	129	119	119	75	75	-9	-9	136	139	101	101	191	191	136	136
BJ403_26	129	129	119	119	81	84	359	391	-9	-9	127	127	191	191	120	161
BJ403_27	-9	-9	-9	-9	-9	-9	263	287	136	136	148	188	191	191	136	136

Figure 1 : Format des données du groupe B pour leur importation sur STRUCTURE. La première ligne comporte 8 entités, il s'agit des marqueurs. Les lignes suivantes sont organisées en 17 colonnes, la première colonne correspond aux identifiants des individus. Les autres colonnes correspondent aux séquences microsatellites. Deux allèles sont renseignés pour chaque locus, chaque allèle représente une colonne, soit 16 colonnes au total. Les données manquantes sont représentées par le signe -9. Aucun autre élément ne doit apparaître dans ce jeu de données, autrement, il ne sera pas pris en charge par le logiciel.

Afin de rassembler les individus par population, une colonne supplémentaire doit être renseignée. Celle-ci doit contenir des nombres entiers dont chacun représente une population différente. Cette colonne doit être placée entre la colonne des individus et les données microsatellites afin que chaque individu soit associé à un entier référant à une population. Ce procédé a permis d'effectuer les simulations LOCPRIOR (tableau 3 du rapport)

#### Résultats graphiques

Plusieurs types de graphes sont générés par STRUCTURE à chaque run. Les plus informatifs sont ceux pour lesquels chaque couleur représente les proportions d'assignations pour les individus. Diverses options proposées par le logiciel permettent de modifier ces figures. L'utilisateur peut choisir d'afficher les résultats sur une seule ligne pour voir l'ensemble des clusters inférés ou plusieurs lignes afin d'observer plus en détail les assignations par individus. Il est également possible de classer les individus par cluster auxquels ils ont été assignés afin de comparer l'importance relative de ces clusters en terme de nombre d'individus. Pour les figures suivantes, j'ai choisi de classer les individus par cluster et de représenter les résultats



## Annexe 6 :

sur une seule ligne. Ces diagrammes permettent de bien visualiser les niveaux d'admixture individuels :

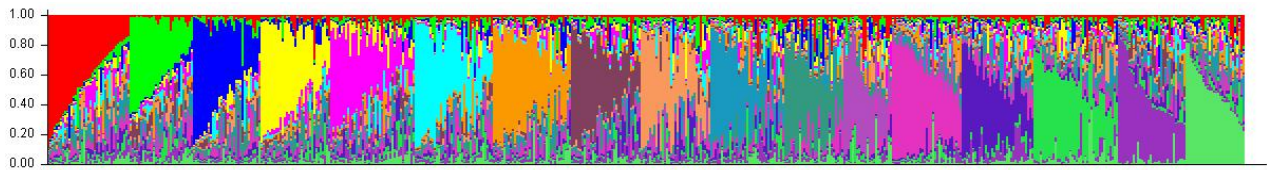


Figure 2 : Diagramme de la simulation effectuée sur les individus B sans prendre en compte les populations d'échantillonnage. Le meilleur des 20 runs de la simulation delta K B pour  $K_{\max} = 17$  est représenté (tableau 3 du rapport)

La figure 2 présente 17 clusters. Il s'agit du nombre optimal de cluster estimé par la méthode delta K de Evanno et al. (2005). Le choix parmi les 20 essais est basé sur le critère DIC affiché par STRUCTURE pour chaque run. Il s'agit d'un indicateur de vraisemblance du modèle proche du BIC. Contrairement à ce dernier, le DIC est pondéré par le nombre de clusters inférés par l'analyse et non par le nombre d'individus.

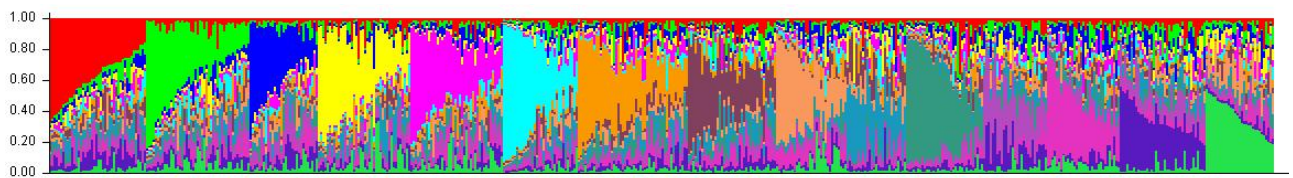


Figure 3: Diagramme de la simulation effectuée sur les individus A sans prendre en compte les populations d'échantillonnage. Le meilleur des 20 runs de la simulation delta K A pour  $K_{\max} = 15$  est représenté (tableau 3 du rapport), 15 clusters sont représentés

Les figures 2 et 3 présentent des nombres de cluster et des distributions des individus quasi identiques. Les individus semblent répartis équitablement entre les clusters. Beaucoup d'individus présentent une grande proportion d'assignation à plusieurs clusters ce qui dénote un haut niveau d'admixture.

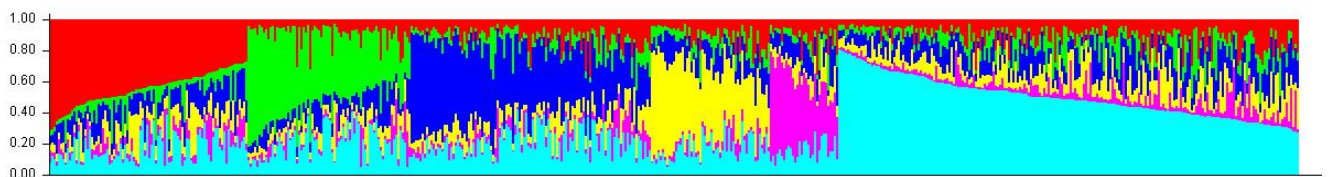


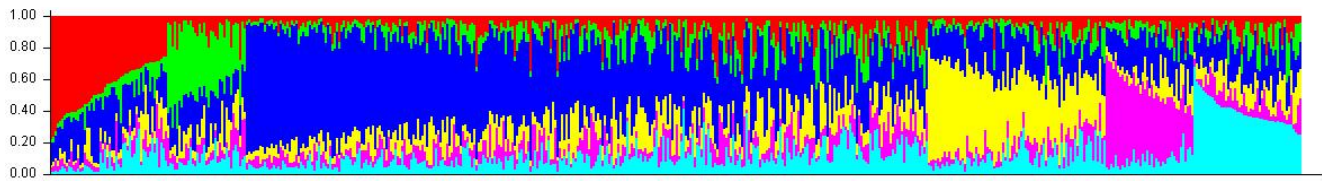
Figure 4 : Diagramme de la simulation effectuée sur les individus B en tenant compte des populations d'échantillonnage. (Simulation LOCPRIOR B, tableau 3 du rapport). Le meilleur des 5 runs de la simulation delta K A pour  $K_{\max} = 6$  est représenté

Sur les figures 4 et 5, les individus ne sont pas équitablement répartis dans les clusters. Pour la figure 4 (analyse sur individus B uniquement), le groupe bleu clair est le plus important en termes de nombre d'individus assignés. Les groupes rouge et bleu foncé semblent équivalents. Ces 3 clusters regroupent la majorité des individus. Toutefois, pour presque tous



## Annexe 6 :

les individus les fortes proportions d'assignation à plusieurs clusters montrent encore une fois un fort taux d'admixture.



*Figure 5* : Diagramme de la simulation effectuée sur les individus A en tenant compte des populations d'échantillonnage. (Simulation LOCPRIOR A, tableau 3 du rapport). Le meilleur des 5 runs de la simulation delta K A pour  $K_{\max} = 6$  est représenté

Pour la figure 5 (analyse sur individus A uniquement), le groupe bleu foncé se dégage des autres. Plus de la moitié des individus y sont assignés et tous les individus présentent une part non négligeable d'assignation à ce cluster. Ici également, ce phénomène dénote un fort taux d'admixture individuelle.

## Annexe 7

### TESS

#### Format et importation des données

Comme pour STRUCTURE, TESS nécessite un certain format de données pour fonctionner. Lors de la création d'un nouveau projet, une fenêtre permet également d'indiquer les caractéristiques du jeu de données :

- Le nombre d'individus
- Le type de ploïdie
- Le nombre de locus
- Le symbole indiquant une valeur manquante (par convention, les valeurs manquantes sont indiquées par la valeur -9 dans la majorité des études de génétique des populations et pour la plupart des outils informatiques utilisés pour ces études (Durand et al., 2009))

L'utilisateur peut également cocher une case indiquant au logiciel si chaque ligne représente un individu. Par défaut, le format pris en compte par TESS présente les deux allèles d'un même locus l'un en dessous de l'autre de sorte que pour des marqueurs diploïdes, chaque individu comprend deux lignes. Ce format demande de reprendre chaque ligne du jeu de données si celui-ci présente les deux allèles d'un même locus sur une seule ligne.

Une autre case peut être cochée, indiquant la présence d'une ligne d'allèles récessifs. Pour la présente étude, les marqueurs étant neutres cette case n'était pas à cocher.

Cette fenêtre nécessitait également de renseigner le fichier contenant les données ainsi que le dossier dans lequel enregistrer les résultats de sortie. Comme STRUCTURE, TESS ne supporte que des fichiers textes. Pour le dossier contenant le projet, il est conseillé d'utiliser le même dossier que celui contenant les données.

TESS permet également de prendre en compte les distances géographiques entre les individus. Ces données peuvent être générées par le logiciel par la fonction *Compute Geographic Distances*. Cependant, cette étape doit être effectuée avant de créer un nouveau projet pour que le fichier texte en résultant puisse y être incorporé.

#### Représentation spatiale

TESS prend en compte les données géographiques des individus pour créer une tessellation de Voronoï (Durand et al., 2009; François and Durand, 2010) (figure 1). Une tessellation de Voronoï ou diagramme de Voronoï, du nom du mathématicien russe Gueorgui Voronoï, est une représentation de l'espace composé d'un ensemble de cellules de Dirichlet. Une cellule de Dirichlet est une portion de l'espace incluant une unique observation. Et

## Annexe 7

comporte tous les points de l'espace qui sont géométriquement plus proche de cette observation que de n'importe quelle autre du diagramme.

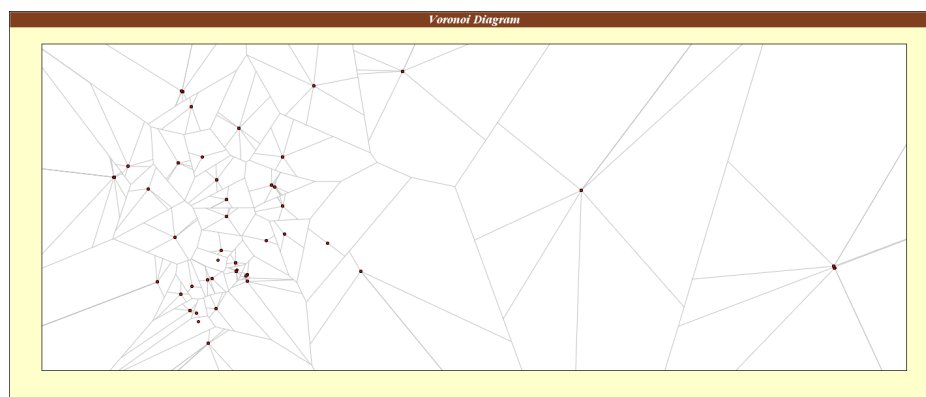


Figure 1 : Diagramme de Voronoï Inféré par TESS pour le groupe B. Chaque ligne représente la séparation entre deux cellules de Ditrichlet

La figure 1 semble présenter plus de cellules qu'il n'existe d'observations mais il s'agit d'un défaut de résolution. En effet, le jeu de données initial présentant le même couple de coordonnées géographiques pour tous les individus d'une même population (points de la figure 1), les résultats d'assignation étaient biaisés.

J'ai donc créé un programme sur R, avec l'aide d'un autre stagiaire M2 biostatisticien, Walid Kandouci, permettant de décaler chaque individu selon l'axe longitudinal afin d'assigner artificiellement un couple de coordonnées unique à chaque individu.

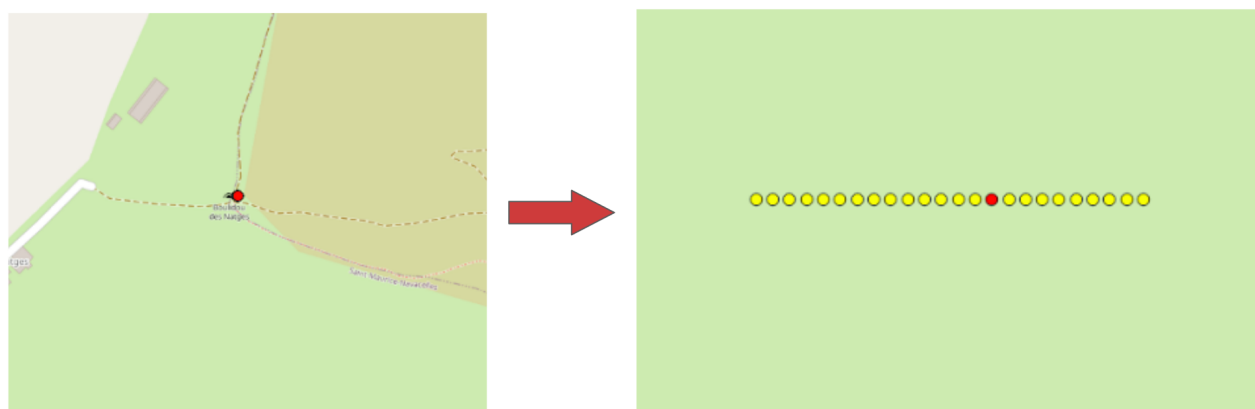


Figure 2 : Représentation géographique de la réassignation des coordonnées géographiques aux individus au sein d'une même population. Le point rouge représente un individu B et les points jaunes représentent des individus A. Ainsi, la majorité des individus de cette population appartiennent au groupe A mais ils auraient été classés en B sans le décalage géographique des individus car l'individu B était lu en premier par le logiciel.

## Annexe 7

Le résultat graphique présentant le clustering et l'assignation des cellules du diagramme à ces clusters (appelé *Hard Clustering*) peut être projeté sur une carte comme sur la figure 3.

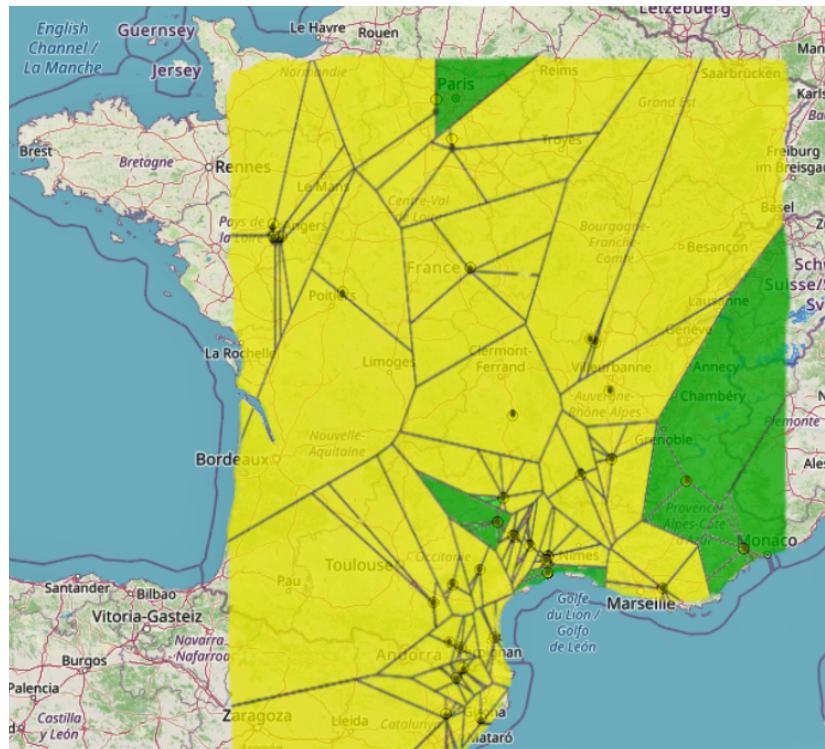


Figure 3 : Projection géographique des résultats de Hard Clustering, inféré sur la population A. Chaque couleur représente un cluster différent.

TESS produit également des diagrammes d'assignation en barre comme STRUCTURE.

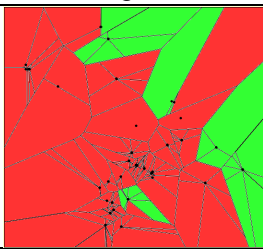
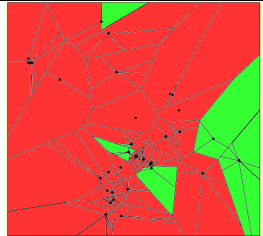
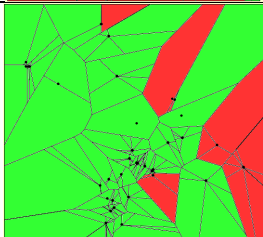
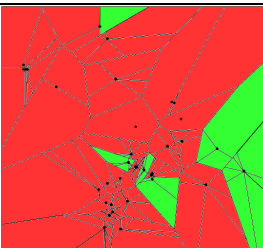
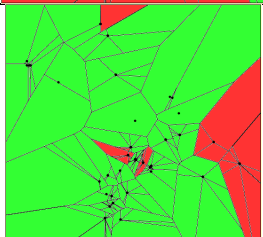
## Annexe 7

### Résultats

Les résultats des clustering avec TESS sur les groupes A et B synthétisés dans le tableau 4 du rapport sont présentés dans le détail ici :

*Tableau 1* : Résumé de tous les runs pour les simulations de clustering A (tableau 4 du rapport), classés en fonction du DIC par ordre croissant. Les numéros de runs sont classés dans l'ordre des simulations présentées dans le tableau 4 du rapport ; les runs numérotés de 1 à 15 correspondent à la simulation effectuée avec le modèle CAR et 20 000 itérations.

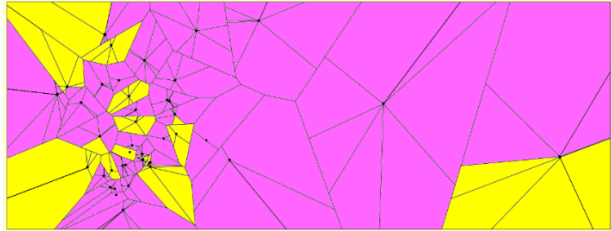
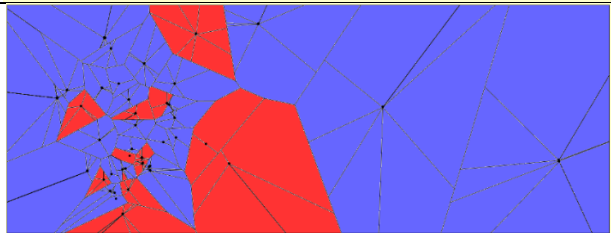
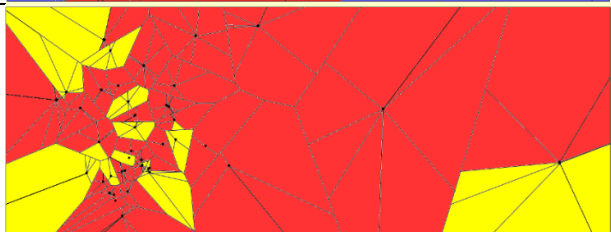
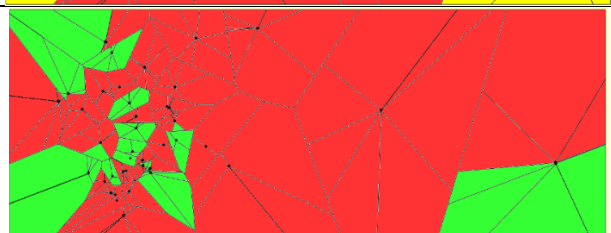
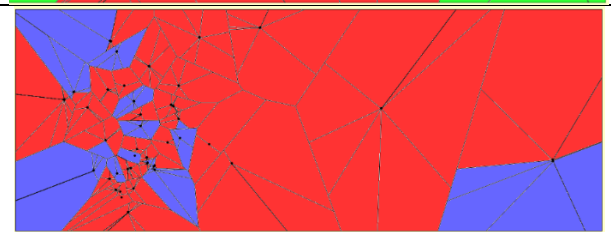
Les figures des 5 meilleurs runs ayant inférés 2 clusters sont présentées

Groupe	Run	K <sub>max</sub>	DIC	Modèle	K inférés	Figure
A	2	2	33289,5	CAR	2	
A	16	2	33354,4	BYM	2	
A	4	2	33378,9	CAR	2	
A	9	3	33387,3	CAR	1	
A	24	2	33425,6	CAR	2	
A	22	2	33427,9	CAR	2	
A	3	2	33428,5	CAR	2	
A	17	2	33440	BYM	2	
A	1	2	33488,3	CAR	2	
A	5	2	33586,3	CAR	2	
A	19	3	33621,3	BYM	1	
A	21	3	33655,6	BYM	1	
A	20	3	33655,7	BYM	1	
A	12	4	33907,3	CAR	1	

## Annexe 7

A	18	2	33908,6	BYM	1
A	15	4	33920,2	CAR	1
A	11	4	33921,1	CAR	1
A	29	4	33924,8	CAR	1
A	13	4	33925,2	CAR	1
A	28	4	33926,1	CAR	1
A	30	4	33926,2	CAR	1
A	14	4	33931,5	CAR	1
A	7	3	33981,7	CAR	1
A	26	3	33987,5	CAR	1
A	25	3	33990,7	CAR	1
A	10	3	33994	CAR	1
A	27	3	33996,7	CAR	1
A	8	3	33997,5	CAR	1
A	6	3	34001,9	CAR	1
A	23	2	34118,7	CAR	1

Tableau 2 : Résumé de tous les runs pour les simulations de clustering B (tableau 4 du rapport), classés en fonction du DIC par ordre croissant. Les runs numérotés de 1 à 15 correspondent à la simulation effectuée avec le modèle CAR et 20 000 itérations. Les figures des 5 meilleurs runs ayant inférés 2 clusters sont présentées

Groupe	Run	K <sub>max</sub>	DIC	Modèle	K inférés	Figure
B	22	5	26705,4	CAR	2	
B	16	4	26981,4	CAR	2	
B	15	4	27036,4	CAR	2	
B	3	4	27054,8	CAR	2	
B	33	4	27058,3	CAR	2	

## Annexe 7

B	12	3	27180,4	CAR	2
B	41	3	27243,9	BYM	2
B	9	3	27300,6	CAR	2
B	32	3	27321,8	CAR	2
B	11	3	27322,3	CAR	2
B	31	3	27329,7	CAR	2
B	20	3	27361,9	CAR	2
B	39	2	27835,8	BYM	2
B	40	2	27837	BYM	2
B	45	5	27862,7	BYM	1
B	5	2	27864	CAR	2
B	46	5	27906,6	BYM	1
B	27	10	27923	CAR	1
B	24	7	27945,4	CAR	1
B	8	2	27947,5	CAR	2
B	29	12	27951,3	CAR	1
B	43	4	27984	BYM	1
B	44	4	27986,8	BYM	1
B	1	2	27988,8	CAR	2
B	7	2	28027,8	CAR	1
B	25	8	28030,6	CAR	1
B	23	6	28039,4	CAR	1
B	26	9	28041,9	CAR	1
B	28	11	28048,3	CAR	1
B	38	5	28061,9	CAR	1
B	37	5	28063,4	CAR	1
B	36	5	28064,2	CAR	1
B	18	4	28083,9	CAR	1
B	21	4	28090,9	CAR	1
B	17	4	28097,9	CAR	1
B	34	4	28100,8	CAR	1
B	35	4	28107,3	CAR	1
B	14	4	28114,1	CAR	1
B	42	3	28140,7	BYM	1
B	30	3	28188	CAR	1
B	10	3	28223,3	CAR	1
B	2	3	28225,1	CAR	1
B	13	3	28229,7	CAR	1
B	19	2	28375	CAR	1
B	6	2	28432,9	CAR	1
B	4	2	28448,7	CAR	1

## Annexe 8

### Librairie Geneland

#### Installation de la librairie et format des données

La librairie *Geneland* ne peut pas être installée selon la procédure habituelle des packages sur R car elle ne figure pas dans les archives du réseau CRAN<sup>1</sup>. La librairie est à télécharger depuis la page GitHub de Giles Guillot<sup>2</sup>, son créateur. Sur R, l'installation s'effectue avec les lignes de code suivantes :

```
writeLines('PATH="${RTOOLS40_HOME}\\usr\\bin;${PATH}"',con=~/.Renviron")
Sys.which("make")
## "C:\\rtools40\\usr\\bin\\make.exe"
install.packages("jsonlite", type = "source")
devtools::install_github('gilles-guillot/Geneland', force = T)
devtools::install_github('gilles-guillot/Geneland', build_vignettes = TRUE, force = T)
library(Geneland)
```

Sous R, les fonctions prennent en charge les données sous la forme de tableau classique au format *data.frame*. Les données génétiques doivent toutefois être découplées des coordonnées géographiques dans des jeux de données différents. Les identifiants de chaque individu permettent de garder la liaison entre les deux jeux de données.

L'interface utilisateur nécessitait une mise en forme des jeux de données de manière similaire aux structures utilisées pour STRUCTURE ou TESS et de les enregistrer dans de nouveaux fichiers textes. Ce remaniement des données m'a dissuadé d'employer cette interface.

---

<sup>1</sup> <https://cran.r-project.org/>

<sup>2</sup> <https://github.com/gilles-guillot/>



## Annexe 8

### Résultats

La fonction *Plotnpop* permet d'afficher les graphiques résumant l'exécution du modèle. Ces graphiques informent de la convergence du modèle ainsi que du nombre de clusters inférés par itérations.

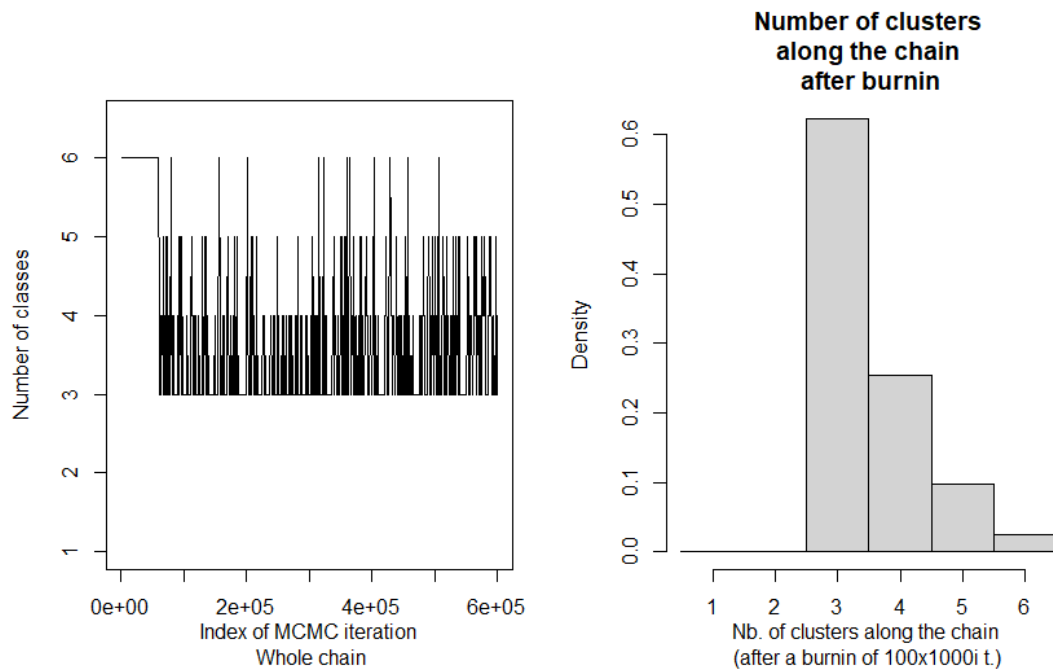
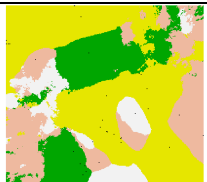
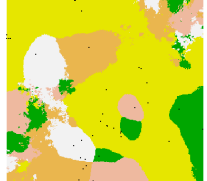
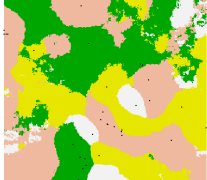
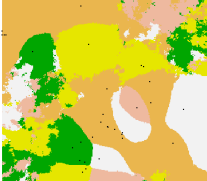

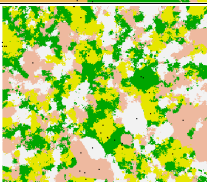
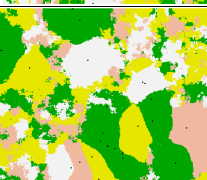
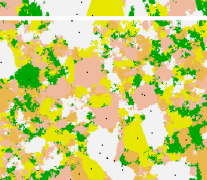


Figure 1 : Graphiques résumant la simulation vérification B, comprenant 600 000 itérations. A gauche, nombre de cluster inférés par itérations au cours de l'exécution de la simulation. A droite, histogramme des itérations classées en fonction du nombre de clusters inférés.

La figure 1 montre que le modèle n'a pas convergé vers une solution unique mais le nombre clusters le plus souvent détecté par les itérations est de 3. Ce nombre de cluster est récurant avec les simulations effectuées pour le groupe B.

## Annexe 8

Tableau 1 : Résumé des résultats de la simulation Boucle A (tableau 5 du rapport). Geneland affiche uniquement la figure pour le nombre de clusters détecté par le maximum d'itérations. La figure du premier essai n'est pas présentée car un seul cluster a été détecté majoritairement.

Simulation	Convergence	Pourcentage d'assignations détectant K clusters					Figure
		K=1	K=2	K=3	K=4	K=5	
Boucle A 1	non	40	20	30	8	2	
Boucle A 2	non	0	0	1	80	19	
Boucle A 3	non	0	0	14	16	70	
Boucle A 4	non	0	0	22	70	8	
Boucle A 5	oui	0	0	0	0	100	
Boucle A 6	oui	0	0	0	2	98	
Boucle A 7	oui	0	0	0	100	0	
Boucle A 8	oui	0	0	0	100	0	
Boucle A 9	oui	0	0	0	0	100	

## Annexe 8

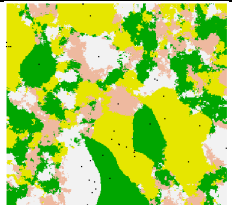
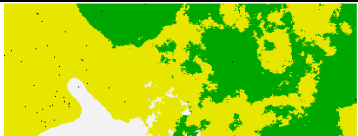
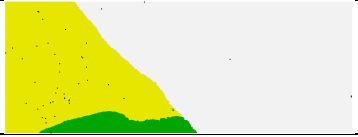
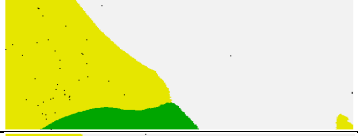
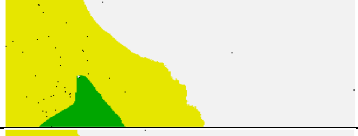


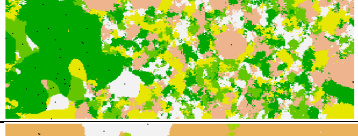
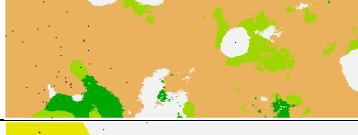
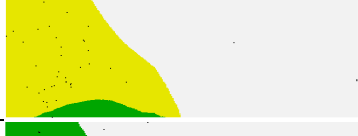
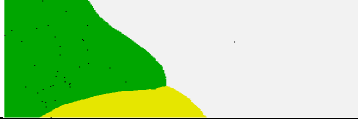
Boucle A 10	non	0	0	0	95	5	
-------------	-----	---	---	---	----	---	---

Tableau 2 : Résumé des résultats de la simulation Boucle B (tableau 5 du rapport).

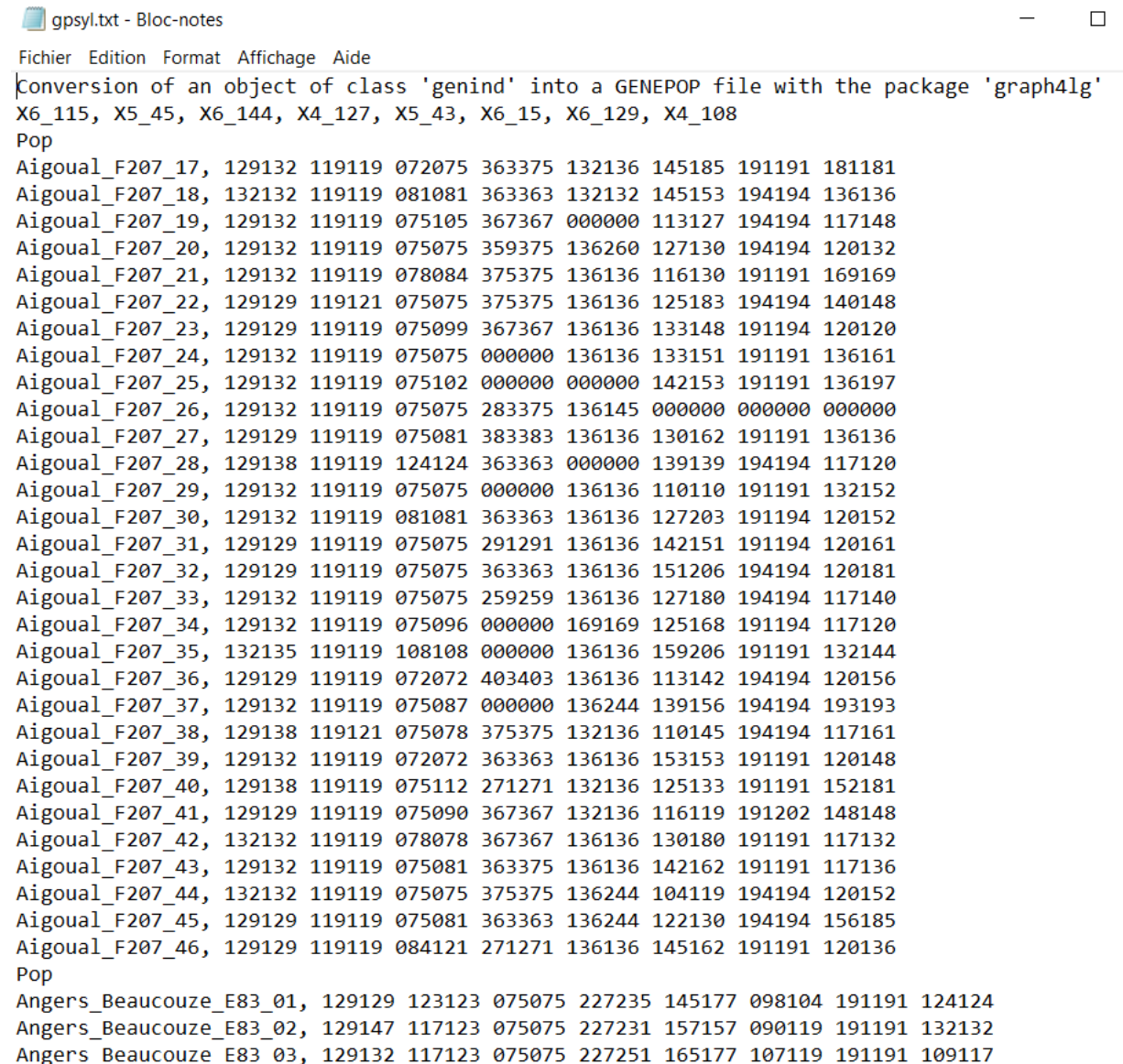
Simulation	Convergence	Pourcentage d'assignments détectant K clusters										Figure
		k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	
Boucle B 1	non	0	0	95	4	0,9	0,1	0	0	0	0	
Boucle B 2	non	0	0	45	35	15	3	1,9	0,1	0	0	
Boucle B 3	non	0	0	65	25	7	2	1	0	0	0	
Boucle B 4	non	0	0	87	10	3	0	0	0	0	0	
Boucle B 5	non	0	0	64	35	1	0	0	0	0	0	
Boucle B 6	non	0	0	62	28	8	1,5	0,5	0	0	0	
Boucle B 7	oui	0	0	0	0	100	0	0	0	0	0	
Boucle B 8	non	0	0	0	95	4	1	0	0	0	0	
Boucle B 9	non	0	0	57	29	10	3	1	0	0	0	
Boucle B 10	non	0	0	61	26	9	3,5	0,5	0	0	0	

## Annexe 9

### Librairie *genepop*

#### Format des données

La librairie *genepop* ne prend en charge qu'un format particulier de données (figure 1). La mise en forme supportée par *genepop* est spécifique à cette librairie, elle ne correspond pas aux formats supportés par STRUCTURE ni par TESS ni *Geneland*.



```
gpsyl.txt - Bloc-notes
Fichier Edition Format Affichage Aide
Conversion of an object of class 'genind' into a GENEPOP file with the package 'graph4lg'
X6_115, X5_45, X6_144, X4_127, X5_43, X6_15, X6_129, X4_108
Pop
Aigoual_F207_17, 129132 119119 072075 363375 132136 145185 191191 181181
Aigoual_F207_18, 132132 119119 081081 363363 132132 145153 194194 136136
Aigoual_F207_19, 129132 119119 075105 367367 000000 113127 194194 117148
Aigoual_F207_20, 129132 119119 075075 359375 136260 127130 194194 120132
Aigoual_F207_21, 129132 119119 078084 375375 136136 116130 191191 169169
Aigoual_F207_22, 129129 119121 075075 375375 136136 125183 194194 140148
Aigoual_F207_23, 129129 119119 075099 367367 136136 133148 191194 120120
Aigoual_F207_24, 129132 119119 075075 000000 136136 133151 191191 136161
Aigoual_F207_25, 129132 119119 075102 000000 000000 142153 191191 136197
Aigoual_F207_26, 129132 119119 075075 283375 136145 000000 000000 000000
Aigoual_F207_27, 129129 119119 075081 383383 136136 130162 191191 136136
Aigoual_F207_28, 129138 119119 124124 363363 000000 139139 194194 117120
Aigoual_F207_29, 129132 119119 075075 000000 136136 110110 191191 132152
Aigoual_F207_30, 129132 119119 081081 363363 136136 127203 191194 120152
Aigoual_F207_31, 129129 119119 075075 291291 136136 142151 191194 120161
Aigoual_F207_32, 129129 119119 075075 363363 136136 151206 194194 120181
Aigoual_F207_33, 129132 119119 075075 259259 136136 127180 194194 117140
Aigoual_F207_34, 129132 119119 075096 000000 169169 125168 191194 117120
Aigoual_F207_35, 132135 119119 108108 000000 136136 159206 191191 132144
Aigoual_F207_36, 129129 119119 072072 403403 136136 113142 194194 120156
Aigoual_F207_37, 129132 119119 075087 000000 136244 139156 194194 193193
Aigoual_F207_38, 129138 119121 075078 375375 132136 110145 194194 117161
Aigoual_F207_39, 129132 119119 072072 363363 136136 153153 191191 120148
Aigoual_F207_40, 129138 119119 075112 271271 132136 125133 191191 152181
Aigoual_F207_41, 129129 119119 075090 367367 132136 116119 191202 148148
Aigoual_F207_42, 132132 119119 078078 367367 136136 130180 191191 117132
Aigoual_F207_43, 129132 119119 075081 363375 136136 142162 191191 117136
Aigoual_F207_44, 132132 119119 075075 375375 136244 104119 194194 120152
Aigoual_F207_45, 129129 119119 075081 363363 136244 122130 194194 156185
Aigoual_F207_46, 129129 119119 084121 271271 136136 145162 191191 120136
Pop
Angers_Beaucouze_E83_01, 129129 123123 075075 227235 145177 098104 191191 124124
Angers_Beaucouze_E83_02, 129147 117123 075075 227231 157157 090119 191191 132132
Angers_Beaucouze_E83_03, 129132 117123 075075 227251 165177 107119 191191 109117
```

Figure 1 : format des données exigé par *genepop*. La première ligne indique les locus. Pour les lignes suivantes, les individus sont regroupés par population séparées par les caractères pop. Les données génétiques doivent être codées par 6 valeurs à chaque locus afin que le logiciel prenne en compte la diploïdie.

Ce format nécessite la modification complète du jeu de données pour respecter ces critères pour les 1123 individus de l'étude. Toutefois, il existe une fonction contenue dans une librairie appelée *graph4lg* qui permet de convertir un objet *genind* (le format supporté par

## Annexe 9

*adegenet*) en fichier *genepop*. Ce type de fichier texte (figure 1) est lisible avec la fonction de la librairie *genepop*.

### Résultats

#### - Rho statistiques :

Les Rho statistiques ont été calculées pour la population totale et les populations A et B indépendamment. Ces indicateurs de différenciation du jeu de données en sous populations ( $Rho_{st}$ ) et de la diversité génétique des individus au sein de ces sous populations ( $Rho_{is}$ ) et de la population totale ( $Rho_{it}$ ) ont d'abord été calculés par locus pour chaque population indépendamment puis moyennés par locus et ensuite résumés pour le jeu de données total.

Tableau 1 : Rho statistiques par locus et globales pour la population totale. Les valeurs significativement proches de 1 apparaissent en jaune et les valeurs significativement proches de 0 sont surlignées en bleu

Locus	$Rho_{is}$	$Rho_{st}$	$Rho_{it}$
6_115	0,5809	0,1496	0,6436
5_45	0,1557	0,0438	0,1927
6_144	0,8245	0,0911	0,8404
4_127	0,8433	0,5892	0,9356
5_43	0,1437	0,0642	0,1987
6_15	0,4474	0,4159	0,6772
6_129	0,3472	0,039	0,3727
4_108	0,4769	0,1012	0,5298
Total :	0,6078	0,4211	0,773

Tableau 2 Rho statistiques par locus et globales pour le groupe A. Les valeurs significativement proches de 1 apparaissent en jaune, les valeurs de  $Rho_{st}$  totales et du locus 6\_15 apparaissent en beige et les valeurs significativement proches de 0 sont surlignées en bleu

Locus	$Rho_{is}$	$Rho_{st}$	$Rho_{it}$
6_115	0,4732	-0,0005	0,4729
5_45	0,0757	0,0085	0,0836
6_144	0,8612	0,046	0,8676
4_127	-0,0621	0,0009	-0,0612
5_43	0,3179	0,0399	0,3451
6_15	0,0216	0,1221	0,1411
6_129	0,0499	0,0016	0,0514
4_108	0,6207	0,0255	0,6304
Total :	0,5075	0,0429	0,5287

Tableau 3 : Rho statistiques par locus et globales pour le groupe B. Les valeurs significativement proches de 1 apparaissent en jaune et les valeurs significativement proches de 0 sont surlignées en bleu

Locus	$Rho_{is}$	$Rho_{st}$	$Rho_{it}$
6_115	0,6719	-0,0106	0,6684
5_45	0,7498	0,0008	0,75
6_144	0,5634	0,0282	0,5757
4_127	0,8118	0,0029	0,8124
5_43	0,0751	0,0104	0,0848
6_15	0,3575	0,0316	0,3778
6_129	0,5817	-0,0036	0,5802
4_108	0,3944	0,01	0,4005
Total :	0,4264	0,0117	0,4331

## Annexe 10

### Analyses des séquences ITS et COI :

#### D de Tajima :

Le D de Tajima est un indicateur statistique nommé d'après son inventeur, le chercheur Fumio Tajima. Il s'agit d'un test qui permet de déterminer si une population ou une séquence d'ADN évolue de manière aléatoire, sous le simple effet de la dérive génétique ou si elle est soumise à une certaine forme de sélection. Il est calculé à partir de la différence entre deux autres indicateurs : la diversité nucléotidique et le nombre de sites de ségrégation.

Le nombre de sites de ségrégation représente de nombre de sites qui présentent des différences entre des gènes apparentés. La diversité nucléotidique correspond au nombre moyen de nucléotides différents par site. C'est-à-dire le la somme du nombre de nucléotides différents entre les paires de séquences divisé par le nombre de paires de séquences.

En fonction de la valeur de cet indice, D de Tajima, plusieurs interprétations peuvent être faites sur l'histoire évolutive de la population considérée :

- une valeur significativement négative peut indiquer une sélection directionnelle récente, une expansion rapide de la population après un goulot d'étranglement génétique ou encore la liaison du gène considéré à un gène qui subit une sélection directionnelle (auto-stop génétique) ;
- une valeur significativement positive indique plutôt une sélection disruptive ou une contraction de la population ;
- une valeur non significative ou nulle indique que la population évolue de manière neutre.

Tableau 1 : résumé des indicateurs permettant de calculer la valeur du D de Tajima pour les séquences ITS et COI concernées par l'étude

Séquences	Nombre de séquences	Nombre de sites de ségrégation (S)	Diversité nucléotidique (p)	Nombre total de sites	Valeur du D de Tajima	P-value
COI A	126	5	0.00052	621	-1.33761	NS, P > 0.10
COI B	239	25	0.00061	621	-2.51236	***, P<0.001
COI A + B	365	61	0.03043	621	2.55098	*, P < 0.05
IT'S A	340	5	0.00184	735	1.3084	NS, P > 0.10
ITS B	445	10	0.00368	744	1.45183	NS, P > 0.10
ITS A + B	785	35	0.01864	726	4.29287	***, P<0.001

## Annexe 10

### Réseaux d'haplotypes ITS et COI

Un haplotype est un groupe d'allèles de différents loci situés sur un même chromosome et habituellement transmis ensemble. Une manière de visualiser les différences entre haplotypes est de les représenter par des réseaux dans lesquels les haplotypes forment des clusters reliés par des liens qui rendent compte des mutations qui différencient les différents haplotypes.

Des réseaux ont été ici reconstruits à partir de séquences des gènes COI ou ITS2 à l'aide du logiciel Popart.

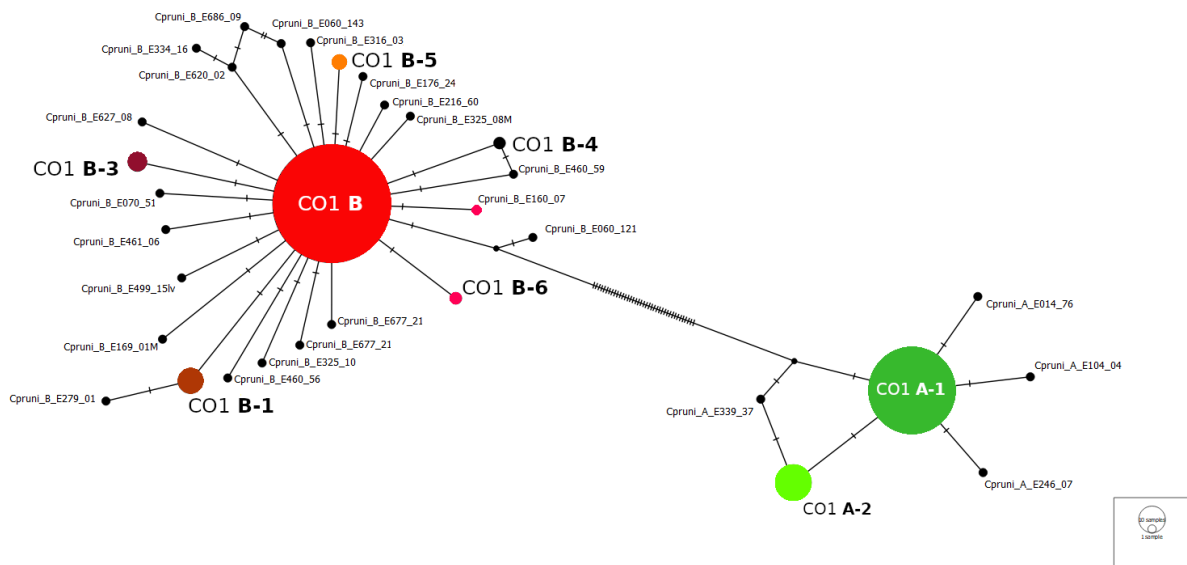


Figure 1 Réseau d'haplotypes prenant en compte les 365 séquences COI A et B.



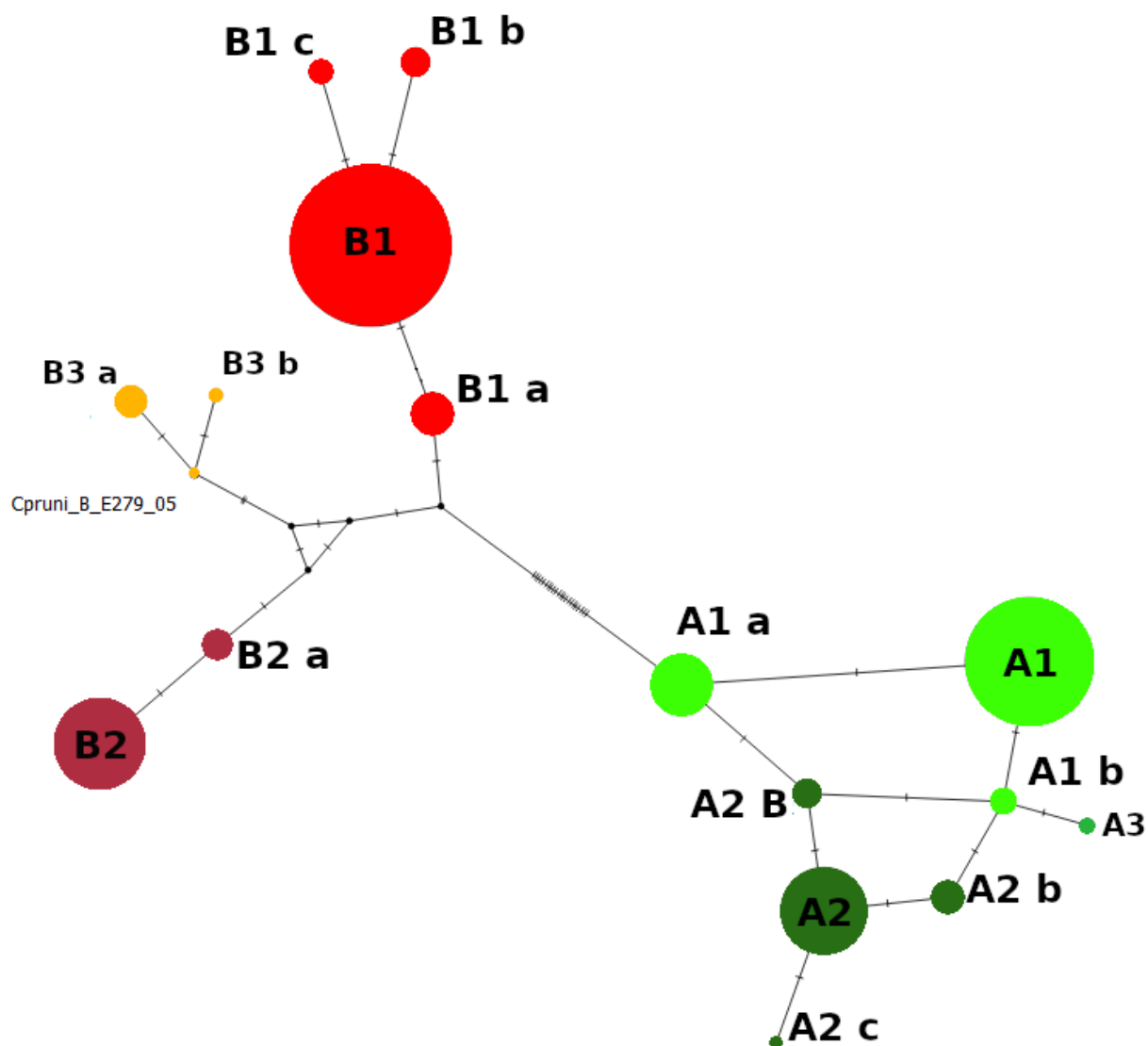

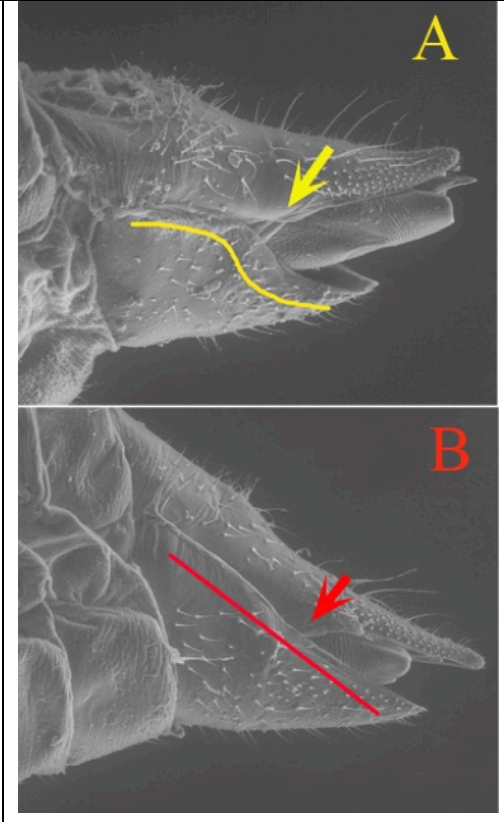


Figure 2 : Réseau d'haplotypes prenant en compte les 785 séquences ITS A et B.

Morphologie de *Cacopsylla pruni*

 <p><i>Cacopsylla pruni</i></p> <p>1,0 mm</p> <p>1,0 mm</p>	 <p>A</p> <p>B</p>
<p>Habitus de <i>Cacopsylla pruni</i> groupe A (vue latérale) : femelle en haut ; mâle en bas.</p>	<p>Vue latérale des genitalia d'une femelle du groupe A (en haut) et du groupe B (en bas). Les traits et les flèches indiquent les différences A vs. B.</p>
<p>Crédit photos : Sauvion N., INRAE©</p>	