Développement et utilisation de marqueurs RADseq pour l'étude de l'impact de Wolbachia sur l'évolution des génomes mitochondriaux chez les Arthropodes

Marie Cariou

N° d'ordre : **92 - 2015** Année 2015

THÈSE de l'Université de Lyon

délivrée par l'Université Claude Bernard Lyon 1

École doctorale : Évolution, Écosystèmes, Microbiologie, Modélisation
Laboratoire de Biométrie et Biologie Évolutive, UMR CNRS 5558

Diplôme de DOCTORAT

(arrêté du 7 août 2006)

Présentée et soutenue publiquement le 8 juillet 2015 par

Marie CARIOU

Développement et utilisation de marqueurs RADseq pour l'étude de l'impact de *Wolbachia* sur l'évolution des génomes mitochondriaux chez les Arthropodes

---

JURY :

Frédéric FLEURY (Université Lyon 1)	Président
Sylvain CHARLAT (Université Lyon 1)	Directeur de thèse
Laurent DURET (Université Lyon 1)	Directeur de thèse
Nicolas BIERNE (Université de Montpellier 2)	Rapporteur
Xavier VEKEMANS (Université de Lille 1)	Rapporteur
Richard CORDAUX (Université de Poitier)	Examinateur

UNIVERSITE CLAUDE BERNARD - LYON 1

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

COMPOSANTES SANTE

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

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. F. DE MARCHI
Département Biologie	Directeur : M. le Professeur F. FLEURY
Département Chimie Biochimie	Directeur : Mme Caroline FELIX
Département GEP	Directeur : M. Hassan HAMMOURI
Département Informatique	Directeur : M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur : M. le Professeur Georges TOMANOV
Département Mécanique	Directeur : M. le Professeur H. BEN HADID
Département Physique	Directeur : M. Jean-Claude PLENET
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. Y.VANPOULLE
Observatoire des Sciences de l'Univers de Lyon	Directeur : M. B. GUIDERDONI
Polytech Lyon	Directeur : M. P. FOURNIER
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. le Professeur C. VITON
Ecole Supérieure du Professorat et de l'Education	Directeur : M. le Professeur A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Directeur : M. N. LEBOISNE

Remerciements

Je voudrais remercier tout d'abord et tout particulièrement mes directeurs de thèse Sylvain Charlat et Laurent Duret. Merci à tous les deux d'être une source d'inspiration par leur intelligence, leur gentillesse et leur enthousiasme. Je leur suis très reconnaissante bien sûr pour leur disponibilité, leur patience et leur bienveillance. Je ne pouvais pas imaginer un meilleur duo de chefs et je mesure avec émotion la chance que j'ai eue de les rencontrer et qu'ils acceptent de m'encadrer.

Je remercie les rapporteurs de cette thèse, Nicolas Bierne et Xavier Vekemans ainsi que les membres du jury, Frédéric Fleury et Richard Cordaux, qui ont bien voulu évaluer mon travail. Merci également aux autres membres de mon comité de pilotage, Nicolas Galtier, Tristan Lefébure, Vincent Daubin et Hélène Henri, pour leurs aides et conseils.

Mes remerciements vont également aux membres du LBBE pour avoir rendu ces 5 années si riches et passionnantes tant sur le plan scientifique qu'humain.

En particulier, merci à mes voisins de bureau, de l'époque de mes ainées à celle des concours de boulettes. Merci Emilie, Julien M, Anaël, Gabriel (pour les locus, les fonctions R et plein de choses pas si péraves), David (mon cher neveu), Mathilde, Julien C, Olivier, Taiisia, Caroline, Samuel, Manon… Merci pour les discussions et pour les nombreuses découvertes culturelles et scientifiques qui ont pu être partagées dans ce bureau. Continuez à vous la donner grave !

Merci à Fanny, Magali, Héloïse, Aline pour leur amitié et les moments passés ensembles au labo ou à l'extérieur.
Merci bien sûr aux autres collègues et amis du deuxième et d'un peu plus loin dans la couloir, Christophe, Clément, Michel, Rémi, Diamantis, Murray, Adil, Aurélie, Marie C, Laurent M. Je remercie particulièrement Yann d'avoir bien voulu partager son chef avec moi. Merci à Florent, à Mathieu, à Nicolas, à Eugénie... Merci à Joanna pour les brunchs et autres pub-quiz et à Erika pour son enthousiasme communicatif et ses conseils avisés. Merci Marie-Christine, Jos, Clothilde, Anna, Paf, Adrien.

Merci à Hélène, pour tellement de choses… bien sûr pour les RADseq du début à la fin ! Mais aussi pour sa gentillesse, son écoute, ses conseils et son exemple.

Merci aux autres grandes personnes, Fabrice, pour différentes discussions scientifiques et humaines, Natacha, pour ses sages conseils, Patricia, Julien, Laurence, Gladys (qui m'a fait faire mes presque premiers pas en bio mol). Merci à Dominique, à Bastien, à Céline, à Damien (pour m'avoir fait découvrir les Rotifères), à Laurent G (pour les mappings de dN/dS), Gabriel M, Sylvain M, Franck, Raquel, Cristina, Annabelle… Merci pour tous leurs conseils, les idées, les discussions. Merci à Marc BB, entre autres choses, pour m'avoir confié mes premiers enseignements, pour être toujours prêts à passer du temps à parler de tirages de lectures dans des urnes de locus (avec remise), et pour différents échanges de bon procédés en rapport avec l'immobilier. Merci à Simon, notamment pour avoir été mon compagnon de dessin cette dernière année et à Marie Sémon, pour être un modèle en toutes choses.

Merci à Nathalie, Aline, Laetitia et Odile ainsi qu'à Stéphane, Lionel et Bruno.

Merci à Thomas, pour les choses partagées.

Merci à Sara, pour tout ce qu'elle m'a apporté.

Grand merci également à Anne-Laure, Paul, Fabrice et Obey, mes amis depuis la cour du lycée.

Merci à Laurent, pour son soutien, sa tendresse, son inspiration et tout le reste. Je te remercie de tout cœur d'être là et d'être comme tu es.

Je remercie enfin à ma famille, mes parents, Monique et Bruno, ma petite sœur, Juliette, mes tantes et oncles, mes cousins, petits cousins et mes aïeules, Madeline et Mathilde, comme dit la chanson.

Et pour tout ceux auxquels ces mots ne rendent pas justice, je n'en pense pas moins.

Résumé

La propagation de bactéries intracellulaires invasives peut entrainer celle des génomes mitochondriaux qui leur sont liés génétiquement au sein du cytoplasme. Cette sélection par autostop peut conduire à une réduction de la taille efficace (N_e) pour le génome mitochondrial. Elle peut également favoriser l'introgression d'une mitochondrie introduite dans une espèce suite à une hybridation. Le principal objectif de ma thèse est de quantifier ces différents effets, de manière globale, au moyen d'un large échantillonnage d'Arthropodes.

Les événements d'introgressions mitochondriales sont à l'origine de discordances entre les histoires évolutives des génomes mitochondriaux et nucléaires. Afin de rechercher de telles discordances, nous avons développé des marqueurs génomiques nucléaires de type RADseq (*Restriction Associated DNA sequencing*), permettant de reconstruire l'histoire des populations étudiées. J'ai pu montrer au moyen de simulations que ce type de données pouvait être utilisé pour inférer des relations phylogénétiques entre espèces (Cariou et al. 2013). Des améliorations du protocole RADseq nous ont également permis de démontrer l'applicabilité de cette méthode à de nombreux spécimens au sein de librairies hautement multiplexées (Henri et al. 2015). A partir d'analyses *in silico*, j'ai par ailleurs évalué l'importance de différents biais liés à l'utilisation de marqueurs RADseq pour estimer les diversités génétiques et proposé une méthode ABC permettant de corriger certains d'entre eux.

A partir de ces développements, j'ai pu démontrer que sur 30 espèces de Diptères et de Lépidoptères testées à ce jour, la proximité génétique mitochondriale est systématiquement confirmée par les marqueurs nucléaires, rejetant ainsi l'hypothèse d'une introgression mitochondriale récente. Sur un plus large échantillon, nous avons en revanche mis en évidence une réduction significative du N_e mitochondrial dans les lignées infectées par *Wolbachia*, suffisante pour réduire le polymorphisme, mais insuffisante pour générer une réduction notable de l'efficacité de la sélection naturelle.

Abstract

The spread of endosymbiotic bacteria can drive that of the linked mitochondrial genomes within the cytoplasm. This hitchhiking selection can lead to a reduction of the effective population size of the mitochondrial genomes (Ne). It can also facilitate mitochondrial introgression, following the introduction of exogenous mitochondria in a species by hybridization. The main objective of my thesis is to quantify these different effects, on a global scale, using a large sample of Arthropods.

Mitochondrial introgressions can lead to discrepancies between the evolutionary histories of mitochondrial and nuclear genomes. To investigate such patterns, we used RADseq genomic markers, that allow reconstructing population histories, and developed improvements for the library preparation and data analysis. Using in silico experiments, I showed that RADseq data is suitable for phylogenetic inferences (Cariou et al. 2013). Adjustments in the RADseq protocol also allowed us to demonstrate the applicability of this method for highly multiplexed libraries (Henri et al. 2015). The impact of various biases related the estimation of population genetic diversity using RADseq was also investigated *in silico*, which lead me to propose an ABC method to correct some of them.

Following these developments, I showed on 30 species of Diptera and Lepidoptera that nuclear markers always confirmed the mitochondrial genetic relatedness, ruling out the hypothesis of recent mitochondrial introgressions. On a larger sample, we detected a reduction of the mitochondrial Ne in *Wolbachia* infected lineages. This reduction caused a significant decrease in the polymorphism of infected populations, but appeared insufficient to reduce the efficacy of natural selection.

Table résumée

Chapitre 1: Introduction **19**

1. Les mitochondries et les conséquences de leur transmission uni-parentale **20**
2. Les parasites de la reproduction **24**
3. Les mitochondries, les symbiotes et les histoires évolutives des populations **28**
4. Echantillonnage **31**
5. Objectifs de la thèse **32**

Chapitre 2: Développements autour du RADseq : Obtention, analyse et utilisation des marqueurs RAD **35**

1. Reconstruire les histoires évolutives des populations avec des marqueurs RAD **37**
2. L'utilisation des données RADseq en phylogénie **47**
3. Optimisation du protocole de RADseq **59**
4. Estimation de la diversité génétique avec des données de RADseq **99**
5. Conclusion **132**

Chapitre 3: *Wolbachia* et évolution des génomes mitochondriaux. **135**

1. Recherche d'introgressions mitochondriales liées à *Wolbachi*a **138**
2. Les *Wolbachia* affectent-elles la diversité mitochondriale des hôtes ? **156**
3. Les *Wolbachia* affectent-elles l'efficacité de la sélection sur les génomes mitochondriaux des espèces hôtes ? **169**

Chapitre 4: Discussion **185**

1. Le RADseq, différentes contraintes à différentes échelles **188**
2. Discordances entre divergences mitochondriales et nucléaires et fréquence des introgressions mitochondriales **192**
3. Quel impact de *Wolbachia* sur les Ne mitochondriaux ? **193**

Conclusion **199**

Références **203**

Table des matières

Chapitre 1: Introduction 19

1. Les mitochondries et les conséquences de leur transmission uni-parentale 20

2. Les parasites de la reproduction 24

2.1. Phénotypes et conséquences 24

2.2. Dynamique invasive des parasites de la reproduction 26

3. Les mitochondries, les symbiotes et les histoires évolutives des populations 28

3.1. Utilisation des marqueurs mitochondriaux en génétique des populations 28

3.2. Discordances entre histoires des populations et des génomes mitochondriaux : rôle des symbiotes cytoplasmiques 29

4. Echantillonnage 31

5. Objectifs de la thèse 32

Chapitre 2: Développements autour du RADseq : Obtention, analyse et utilisation des marqueurs RAD 35

1. Reconstruire les histoires évolutives des populations avec des marqueurs RAD 37

1.1. Réduction de la représentation (*reduced représentation, RR*) en génomique des populations 37

1.2. Caractères spécifiques du RADseq 38

1.3. De nombreuses variantes : flexibilité du protocole 41

1.4. De nombreux domaines d'application 45

2. L'utilisation des données RADseq en phylogénie 47

2.1. Peut-on résoudre des phylogénies avec des données de RADseq ? étude *in silico* (article) 47
2.2. Phylogénies RADseq dans la littérature : Confirmations in silico et empiriques 55
2.3. Discussions méthodologiques sur l'utilisation de données RAD en phylogénie 56
2.4. Conclusion 57
3. Optimisation du protocole de RADseq 59
3.1. Introduction 59
3.2. Une méthode de synthèse des adaptateurs permettant d'augmenter le niveau de multiplexage des librairies (article) 59
3.3. Expérience RADseq pilote : Variations de couverture dans les librairies hautement multiplexées 66
3.3.1. Introduction 66
3.3.2. Méthode 66
3.3.3. Résultats du séquençage de la librairie pilote 67
3.3.3.1. Variabilité du nombre de lectures obtenues par spécimens 68
3.3.3.2. Couverture incomplète et hétérogénéité de la profondeur de séquençage entre locus, au sein des individus 70
3.3.3.3. Qualité des données : alignement des lectures de *D. melanogaster* sur le génome de référence 74
3.3.4. Discussion : Apports de l'expérience pilote 76
3.4. Préparation de la librairie Symbiocode 77
3.4.1. Nombre le locus, profondeur et multiplexage 77
3.4.2. Description de la librairie Symbiocode, modifications apportées par rapport à la librairie pilote 81
3.5. Librairie Symbiocode : mise en place d'un protocole d'analyse des données RADseq .. 82
3.5.1. Méthodes 83
3.5.2. Analyse des lectures de la librairie Symbiocode 87
3.5.3. Evaluation de la proportion de locus couverts pour chaque individu 89
3.6. Discussion 97
4. Estimation de la diversité génétique avec des données de RADseq 99
4.1. Introduction 99
4.2. Correction du biais d'échantillonnage des coalescents par une méthode ABC 103
4.2.1. Principe des Approches Bayésiennes Approximées (*Approximate Bayesian Computation*) 103
4.2.2. Méthode RAD_abc : simulations, statistiques descriptives et estimations par ABC 106
4.2.2.1. Simulations, mesure du biais sur des données simulées 108
4.2.2.2. Statistiques descriptives 111
4.2.2.3. Estimation du polymorphisme 113
4.2.3. Résultats : Validation croisées (*Cross validation*), correction sur données simulées 114

4.3. Correction de π sur deux individus diploïdes : test sur des données réelles et impact d'une structuration en sous-populations 117
4.3.2. RADseq *in silico* : DPGP et Popphyl 117
4.3.3. Impact de la structuration sur le biais d'échantillonnage des coalescents 121
4.4. Estimation de la proportion de locus hétérozygotes sur un individu : Correction du biais d'hétérozygotie cachée 124
4.5. Discussion 131
5. Conclusion 132

Chapitre 3: *Wolbachia* et évolution des génomes mitochondriaux. 135
1. Recherche d'introgressions mitochondriales liées à *Wolbachi*a 138
1.1. Introduction : *Wolbachia*, introgressions mitochondriales et discordance nucléo-cytoplasmiques 138
1.2. Résultats 143
1.2.1. Estimation des divergences nucléaires à partir de données RADseq. 143
1.2.2. Répétabilité de l'estimation des divergences RADseq 146
1.2.3. Comparaison de l'hétérozygotie et des distances entre spécimens 147
1.3. Discussion 150
1.3.1. Fréquence des discordances causées par des introgressions mitochondriales 150
1.3.2. Comparaison des temps de coalescence mitochondriaux et nucléaires au sein des populations ? 153
1.4. Conclusion 155
2. Les *Wolbachia* affectent-elles la diversité mitochondriale des hôtes ? 156
2.1. Introduction : Dynamique d'invasion de Wolbachia et effets attendus sur le polymorphisme des mitochondries 156
2.2. Résultats 160
2.2.1. Diversité mitochondriale des espèces de SymbioCode 160
2.2.2. Impact des invasions récentes par *Wolbachia* sur les diversités mitochondriales au sein des espèces 161
2.2.3. Impact des infections par *Wolbachia* sur les diversités mitochondriales des espèces infectées par rapport aux espèces non-infectées 163
2.2.4. Impact global du statut d'infection sur diversité mitochondriale ? 165
2.3. Discussion et conclusion 167
3. Les *Wolbachia* affectent-elles l'efficacité de la sélection sur les génomes mitochondriaux des espèces hôtes ? 169
3.1. Introduction : *Wolbachia*, Ne et dN/dS 169

3.2. Inférence des statut d'infections par *Wolbachia* sur les branches de l'arbre des hôtes 171
3.3. Estimation des ω : Comptage (*mapping*) des substitutions synonymes et non-synonymes 172
3.4. Relation entre ω et la probabilité d'infection par *Wolbachia* 174
3.4.1. Méthode : comparaison entre ω des branches infectées et non-infectées 174
3.4.2. Résultat et discussion 176
3.5. *Drosophila recens* et *Drosophila subquinaria* 178
3.6. Discussion 182

Chapitre 4: Discussion 185
1. Le RADseq, différentes contraintes à différentes échelles 188
2. Discordances entre divergences mitochondriales et nucléaires et fréquence des introgressions mitochondriales 192
3. Quel impact de *Wolbachia* sur les Ne mitochondriaux ? 193

Conclusion 199

Références 203

Chapitre 1

Introduction

Les génomes des organismes eucaryotes se partagent entre deux compartiments aux modes de transmission distincts : le noyau et le cytoplasme. Contrairement aux génomes nucléaires, généralement hérités de deux parents, les mitochondries sont le plus souvent transmises de manière uni-parentale, de la mère aux descendants. Ce mode de transmission a différentes conséquences sur l'évolution des génomes mitochondriaux. En particulier, ces génomes peuvent être liés génétiquement avec ceux de symbiotes cytoplasmiques. Cette thèse a pour objet principal l'étude de l'impact de cette liaison génétique sur l'évolution des génomes mitochondriaux.

Cette étude a nécessité le développement de marqueurs permettant de reconstruire l'histoire évolutive des génomes nucléaires. J'ai utilisé pour cela une approche génomique par réduction (*Reduced Représentation*). Des marqueurs de *Restriction Associated DNA sequencing* (RADseq) ont ainsi été obtenus et analysés dans le cadre de l'étude d'une communauté d'Arthropodes de Polynésie. Ce système permet la comparaison des histoires évolutives des différents éléments génétiques composant chaque individu, dans le but de comprendre l'impact des infections symbiotiques sur l'évolution des génomes mitochondriaux.

1. Les mitochondries et les conséquences de leur transmission uni-parentale

Les mitochondries sont des organelles des cellules eucaryotes. Elles sont impliquées dans la respiration cellulaire et jouent ainsi un rôle fondamental dans le métabolisme énergétique de ces cellules. Par ailleurs, elles sont issues d'une endosymbiose entre une alpha-protéobactérie et un ancêtre des cellules eucaryotes (Sagan 1967, Margulis 1970). Cette origine endosymbiotique explique une grande partie de leurs caractéristiques morphologiques (double membrane) et moléculaires (composition des membranes, absence d'histone, similarité des ribosomes mitochondriaux et des ribosomes bactériens), mais également une grande partie des caractéristiques de leurs génomes (génomes circulaire, code génétique spécifique).

Une conséquence importante de cette symbiose originelle est le mode de transmission spécifique de ces organelles. Les mitochondries sont en effet transmises de façon uni-parentale. Chez les organismes diploïdes, à reproduction sexuée, le génome nucléaire de chaque individu lui a été transmis en deux copies, une par chacun de ses

parents, tandis que les mitochondries proviennent toutes des mitochondries d'un seul parent, généralement sa mère.

La transmission maternelle des mitochondries a quelques exceptions, par exemple chez les conifères où la transmission des organelles est paternelle (Neale et al. 1989). D'autre part, chez certains bivalves, les mitochondries sont transmises de façon dite « doublement uni-parentale » (Zouros 2013). Deux génomes mitochondriaux distincts sont en effet transmis au zygote, l'un par la mère et l'autre par le père. Mais, les mitochondries paternelles sont ensuite confinées à la lignée germinale des mâles, qui transmettent ensuite le génome mitochondrial hérité de leur père, tandis que les femelles transmettent celui hérité de leur mère. Cette transmission ne correspond donc pas à une recombinaison au sens de celle affectant les génomes nucléaires. Chacune des lignées maternelle et paternelle sont transmises parallèlement de façon uni-parentale. Notons que ce cas constitue une exception par rapport à une transmission très majoritairement maternelle des génomes mitochondriaux, chez les animaux.

Plusieurs conséquences importantes pour l'évolution des génomes mitochondriaux découlent de leur mode de transmission spécifique. Tout d'abord, les tailles efficaces des populations, Ne, de ces génomes haploïdes sont réduites par rapport à celles des locus nucléaires. L'importance de cette réduction dépend du sex-ratio des populations : le Ne mitochondriale est d'autant plus réduit que la proportion de mâles dans la population est élevée. Mais pour un sex-ratio de 50% de femelles, le Ne mitochondrial correspond théoriquement à 1/4 du Ne nucléaire. Cette différence de tailles de population se traduit théoriquement par des temps de coalescences plus courts des génomes mitochondriaux que des génomes nucléaires. Conformément à cette prédiction, les phylogénies établies au moyen de marqueurs mitochondriaux sont souvent considérées comme moins affectées par les tris de lignées incomplets que celles correspondant à des marqueurs nucléaires (Moore 1995).

D'autre part, un seul sexe contribue à la transmission des génomes mitochondriaux, ce qui génère des conflits génétiques particuliers entre compartiments cytoplasmiques et nucléaires. En effet, les mâles ne les transmettant pas, ils constituent des impasses évolutives pour les génomes cytoplasmiques. Les modèles de génétique des populations montrent qu'en population panmictique, les effets phénotypiques s'exprimant uniquement sur les mâles ne sont pas soumis à la sélection naturelle (Frank and Hurst 1996, Frank 2012). Cette absence de sélection explique vraisemblablement

que de nombreuses maladies génétiques liées au génome mitochondrial affectent plus fortement les mâles, ou la fertilité des mâles (par exemple, chez l'homme, la neuropathie de Leber ou le syndrome de Pearson sont des maladies génétiques associées à des variants mitochondriaux et généralement plus sévères chez les hommes porteurs que chez les femmes). Les mutations du génome mitochondrial causant des effets délétères uniquement pour les mâles sont ainsi plus difficilement éliminées par la sélection naturelle. Toutefois, ces mutations peuvent être soumises à une sélection de parentèle, dans le cas d'une population structurée. Dans ce cas, les femelles se reproduisant plus fréquemment avec leurs apparentés, leur fitness est indirectement affectées par celle des mâles.

Ce type de conflit s'illustre par exemple dans les phénomènes de stérilité mâle liés à un facteur cytoplasmique (CMS, *Cytoplasmic male sterility*), chez les plantes gynodioïques (Chase 2007). Comme son nom l'indique, la stérilité male cytoplasmique correspond à un défaut de production des anthères ou du pollen, causée par des locus mitochondriaux. Le même mécanisme peut également conduire à la sélection de locus mitochondriaux responsables de distorsions du sex-ratio en faveur des femelles. Ce phénomène est rarement observé chez les animaux, mais il a été décrit, par exemple, par Perlman et al. 2015 dans le cas d'une espèce de Psocoptères, *Liposcelis nr. bostrychophila*. Ces auteurs ont observé une distorsion du sex-ratio dans cette espèce, causée par un facteur transmis maternellement, probablement lié au génome mitochondrial (plutôt qu'à la présence de bactéries cytoplasmiques).

Toutefois, la sélection de mitochondries causant soit des effets délétères sur les mâles, soit une distorsion du sex-ratio en leur défaveur, peut être contrebalancée par celle de mécanismes compensatoires, codés par le génome nucléaire. Dans le cas de la stérilité mâle cytoplasmique, par exemple, de nombreux variants mitochondriaux responsables du CMS sont connus, mais également de nombreux locus « restaurateurs de fertilité », toujours situés sur le génome nucléaire. Chez *Drosophila melanogaster* également, Camus et al. 2012 ont montré l'existence de nombreux variants mitochondriaux délétères spécifiquement chez les mâles et contrebalancés par des locus nucléaires. Leur étude a consisté en l'introduction par introgression des génomes mitochondriaux issus de différentes populations des *D. melanogaster* au sein d'individus correspondant à une même lignée nucléaire. Au sein des lignées ainsi constituées, les auteurs ont observé une diminution de la fitness, spécifiquement chez les mâles. Ce

résultat suggère que des effets délétères uniquement pour les mâles étaient causés par ces génomes mitochondriaux. Le fait que ces effets ne soient pas visibles dans les populations d'origine suggère que dans ces populations, l'effet de ces mutations est contrebalancé par celui de locus « compensatoires » dans le génome nucléaire. C'est seulement associé à un nouveau fond génétique nucléaire, non coadapté, que l'effet phénotypique délétère de ces mutations mitochondriales a pu être observé.

Ces résultats montrent une importante coadaptation des génomes mitochondriaux et nucléaires. Plus généralement, on peut remarquer que le fonctionnement des mitochondries repose sur l'expression de nombreux gènes codés pour certains par le génome mitochondrial et pour d'autres par le génome nucléaire. Les mitochondries jouant un rôle central dans la production d'énergie par les cellules eucaryotes, on peut donc s'attendre à l'existence d'une coévolution entre génomes mitochondriaux et nucléaires au niveau des gènes impliqués conjointement dans ces fonctions.

Enfin, une autre conséquence de l'haploïdie et de la transmission uni-parentale de ces génomes concerne la liaison génétique entre locus. Les génomes nucléaires recombinent à chaque génération, c'est à dire que des allèles des différents locus sont réassociés à chaque génération, par des mécanismes de brassage inter et intra chromosomique lors de la méiose. A l'inverse, les différents locus mitochondriaux sont fortement liés génétiquement entre eux et plus largement avec tous les éléments cytoplasmiques transmis de la même façon.

Les mitochondries peuvent en particulier être liées génétiquement avec des bactéries symbiotiques présentes dans le cytoplasme. Les Arthropodes sont impliqués dans de nombreuses interactions avec des bactéries intracellulaires, qui illustrent l'impact de cette liaison. Ces symbioses peuvent être de natures différentes. Certaines sont obligatoires, et impliquent des relations de dépendance entre hôte et symbiotes. D'autres sont transitoires et peuvent correspondre à des associations plus ou moins avantageuses aux deux organismes. Certaines bactéries procurent de nouvelles fonctions au système constitué de l'hôte et du symbiote. C'est le cas par exemple des symbioses nutritionnelles dans lesquelles un micro-organisme permet à son hôte de synthétiser des métabolites ou des nutriments nécessaires, ce qui peut rendre possible l'utilisation de nouvelles ressources.

Dans cette thèse, je me suis intéressée plus particulièrement à un type de bactéries endosymbiotiques appelés « parasites de la reproduction ». Ces micro-organismes, transmis maternellement sont responsables de différents phénotypes ayant tous pour caractéristique de modifier les systèmes de reproduction de leurs hôtes de manière à augmenter leur transmission.

2. Les parasites de la reproduction

L'expression « parasites de la reproduction » désigne un ensemble de micro-organismes qui induisent des modifications des systèmes de reproduction de leurs hôtes. On rencontre ces micro-organismes dans des groupes taxonomiques variés (O'Neill et al. 1997). Toutefois, le groupe le plus étudié et probablement le plus répandu parmi ces symbiotes est celui des *Wolbachia* (Werren 1997, Werren et al. 2008). Selon les estimations, la proportion d'espèces infectées par *Wolbachia* varie entre 30 et 60% (Hilgenboecker et al. 2008). Pour cette raison, il sera principalement question de cette bactérie dans cette thèse, même si les mécanismes décrits peuvent théoriquement être causés par n'importe quel parasite de la reproduction.

2.1. Phénotypes et conséquences

Ces symbiotes ont pour caractéristique commune d'être principalement transmis verticalement de mère à descendants. Ce mode de transmission explique un certain nombre d'effets phénotypiques causés sur leurs hôtes, qui ont tous pour conséquence de favoriser les lignées cytoplasmiques infectées. Cela se traduit souvent par une augmentation de la fitness relative des femelles infectées, parfois au détriment des mâles.

Figure supprimée pour cause de droits d'auteur

Figure 1. Principaux phénotypes induits par *Wolbachia* (Werren et al. 2008)

La figure 1 illustre les différents types de parasitisme de la reproduction induits par *Wolbachia*. Certains de ces mécanismes ont pour conséquence directe de provoquer l'augmentation de la production de femelles infectées en augmentant la proportion de femelles dans la descendance des femelles infectées. Ainsi, certaines de ces bactéries amènent leurs hôtes à se reproduire par parthénogénèse thélytoque. Les femelles infectées produisent, sans s'accoupler, une descendance exclusivement femelle, ce qui favorise la transmission de la bactérie responsable. Ce phénotype est rencontré chez des espèces initialement arrhénotoques, c'est-à-dire où des mâles haploïdes sont produits par parthénogénèse à partir d'œufs non-fécondés (Stouthamer et al. 1990)

Des infections par *Wolbachia* peuvent également causer la féminisation de mâles génotypiques. Ce phénotype conduit au développement de femelles phénotypiques à partir des individus de génotype mâle dans la descendance les femelles infectées. Enfin, une mortalité embryonnaire mâle anormalement élevée (*male-killing*) peut également être induite par ce type d'infection. Ce dernier phénotype ne favorise pas de façon directe la production de femelles infectées. Mais la mort des mâles dans la descendance des femelles infectées peut permettre une réallocation des ressources vers les femelles, ce qui conduit donc indirectement à augmenter la valeur sélective des femelles infectées, par un phénomène dit "de compensation de fitness" (Hurst and Jiggins 2000).

Un dernier phénotype peut être causé par les parasites de la reproduction: l'incompatibilité cytoplasmique. Contrairement aux trois premiers effets décrits, ce phénomène ne conduit pas à un biais de sex-ratio. Dans ce cas, les croisements entre

mâles infectés et femelles non-infectées sont stériles. Ceci a pour effet de diminuer la fertilité moyenne des femelles non-infectées par rapport aux infectées ce qui favorise donc également les lignées cytoplasmiques infectées et conduit à la propagation de la bactérie responsable.

2.2. Dynamique invasive des parasites de la reproduction

Différents phénotypes affectant les systèmes de reproduction des hôtes sont donc induits par *Wolbachia*. Ces effets ont tous pour conséquence de favoriser la propagation de l'infection. D'autre part, on peut remarquer que ces bactéries peuvent avoir d'autres types d'effets, avantageux (Dobson et al. 2004) ou délétères (Fleury et al. 2000) sur leurs hôtes. Parmi les impacts positifs de *Wolbachia* sur la fitness de son hôte, on peut citer le cas des infections associées à une protection contre des virus (Hedges et al. 2008). *Wolbachia* peut également jouer le rôle de symbiote nutritionnel, par exemple chez les punaises des lits, où elle permet la synthèse de vitamine B (Hosokawa et al. 2010).

La propagation d'une infection par ce type de parasite dépend de l'intensité du phénotype lié à la reproduction et de son taux de transmission, ainsi que de l'impact de l'infection sur la valeur sélective de l'hôte, indépendamment des effets liés au parasitisme de la reproduction. Dans la plupart des cas, une forte pénétrance du parasitisme de la reproduction, un effet négatif faible ou un effet positif sur la fitness, et un taux de transmission efficace conduisent à l'augmentation en fréquence des bactéries dans les populations infectées jusqu'à une fréquence d'équilibre (Vavre and Charlat 2012). La figure 2 illustre cette dynamique dans le cas d'une infection causant une incompatibilité cytoplasmique. Au delà d'un seuil défini par l'intensité de l'incompatibilité cytoplasmique, du taux de transmission et de l'impact sur la fitness, la fréquence de l'infection augmente jusqu'à une valeur d'équilibre stable. En dessous de ce seuil, qui constitue en ce sens un équilibre instable, la fréquence diminue jusqu'à la perte de l'infection.

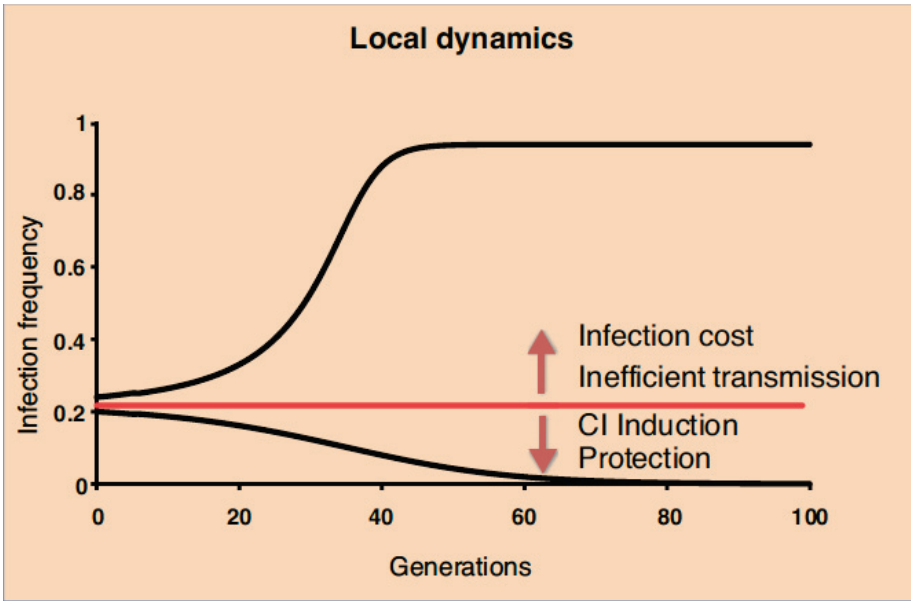

Figure 2. Dynamique d'invasion d'une infection causant une incompatibilité cytoplasmique. Au delà d'un certain seuil de fréquences (au dessus de la ligne rouge), le bénéfice lié à l'incompatibilité cytoplasmique permet de contrebalancer d'éventuels coûts de l'infection ainsi que des pertes dues aux transmissions imparfaites. Des effets bénéfiques en terme de fitness, comme par exemple, une protection contre des pathogènes, réduisent ce seuil et facilitent ainsi la propagation de l'infection (d'après Vavre and Charlat 2012).

Du fait de leur co-transmission, ces bactéries sont liées génétiquement avec les génomes mitochondriaux. La propagation d'une infection peut donc entrainer par auto-stop celle des mitochondries liées. L'impact de cette liaison sur les génomes mitochondriaux dépend alors fortement du pouvoir invasif de la bactérie, comme détaillé dans la partie suivante. A l'échelle des communautés, l'influence de *Wolbachia* sur l'évolution des génomes mitochondriaux va également dépendre de la dynamique des pertes et acquisition de *Wolbachia*, c'est-à-dire de la fréquence à laquelle les infections sont acquises et perdues dans les populations. Des phases d'invasion fréquentes causées par des remplacements rapides des *Wolbachia* le long des lignées auront probablement un impact plus fort que des invasions rares et stables.

3. Les mitochondries, les symbiotes et les histoires évolutives des populations

Les marqueurs mitochondriaux sont fréquemment utilisés pour reconstruire l'histoire des populations. Du fait de leurs temps de coalescences courts et de leurs taux de mutation élevés, ces marqueurs sont particulièrement appropriés dans le cadre de comparaisons entre populations ou entre espèces proches. Toutefois, l'évolution de ces génomes pourrait être influencée par des mécanismes spécifiques, mettant en question la pertinence de ces marqueurs pour inférer les histoires évolutives des populations correspondantes (Hurst and Jiggins 2005). Les symbiotes cytoplasmiques, en causant une sélection indirecte sur l'ensemble du cytoplasme au sein des populations infectées, peuvent notamment interférer avec l'évolution des mitochondries. Les invasions par ces symbiotes peuvent en effet entrainer des réductions d'effectifs efficaces des génomes mitochondriaux et favoriser des introgressions mitochondriales.

3.1. Utilisation des marqueurs mitochondriaux en génétique des populations

Les séquences mitochondriales constituent des marqueurs intéressants en génétique des populations. Tout d'abord, leurs temps de coalescence courts rend leurs phylogénies théoriquement moins sujettes aux tris de lignée incomplets que celles des marqueurs nucléaires, pour un même temps de divergence (Moore 1995). De plus, les taux de mutation élevés en font des marqueurs très variables à une échelle de temps relativement courte. Enfin, de nombreux marqueurs faciles à amplifier ont été développés pour ce génome.

Pour ces raisons, des marqueurs mitochondriaux sont notamment choisis pour les projets de *barcoding*, qui visent à l'usage de courtes séquences nucléotidiques pour la description de la biodiversité. En particulier, ces barcodes moléculaires peuvent être employés pour assigner des spécimens à des espèces connues, ou éventuellement pour révéler l'existence d'espèces cryptiques, non détectables par des approches morphologiques. Chez les animaux, ces projets utilisent généralement un fragment du gène mitochondrial CO1 (Hebert et al. 2003). Plusieurs études ont cherché à vérifier l'efficacité de cette approche, qui repose notamment sur la monophylie des lignées mitochondriales au sein des espèces. Par exemple, Huemer et al. 2014 ont étudié la performance du barcode CO1 pour l'assignation taxonomique de spécimens

correspondant à 1000 espèces de Lépidoptères échantillonnées chacune en Finlande d'une part et en Autriche d'autre part. Ces auteurs ont ainsi testé l'effet de la distance géographique sur la distance génétique et donc sur la qualité de l'assignation par *barcoding*. Les résultats obtenus ont montré que, dans ce groupe, les marqueurs mitochondriaux permettent de déterminer efficacement l'origine taxonomique des spécimens. Dans d'autres groupes taxonomiques, il avait été montré, à l'inverse, que la distance géographique affectait significativement les distances intra-espèce au point de limiter l'efficacité des barcodes CO1 pour la discrimination des espèces (Bergsten et al. 2012).

Plus généralement, les génomes mitochondriaux peuvent avoir des histoires évolutives discordantes de celles des génomes nucléaires, ce qui est à prendre en compte pour permettre l'emploi de ces marqueurs pour reconstruire l'histoire des organismes associés (Hurst and Jiggins 2005, Frézal and Leblois 2008).

3.2. Discordances entre histoires des populations et des génomes mitochondriaux : rôle des symbiotes cytoplasmiques

Des balayages sélectifs indirects affectant les génomes mitochondriaux peuvent être causés par la propagation de bactéries cytoplasmiques. Ces balayages sont associés à des réductions des effectifs efficaces des génomes mitochondriaux. L'effectif efficace d'une population, Ne, correspond à l'effectif d'une population idéale (de Wright-Fisher) dans laquelle l'impact de la dérive génétique serait le même que celui observé dans la population réelle. Cet effectif est donc relié au polymorphisme observé dans les populations. Toutefois, Bazin et al. 2006, ont montré que le polymorphisme mitochondrial moyen calculé sur huit grands taxons animaux n'était pas corrélé aux tailles de population réelles ni au polymorphisme nucléaire moyen. Cette observation pourrait s'expliquer par des balayages sélectifs récurrents pour la sélection de mutations avantageuses.

Des réductions de Ne mitochondriaux associées à des épisodes de sélection positive pourraient également, chez les Arthropodes, être causés par des invasions par des symbiotes cytoplasmiques. L'intensité des réductions de Ne mitochondriaux causées par les invasions par *Wolbachia* dépendent de leur effet phénotypique (Engelstädter

2010), ainsi que des modes de transmission de l'infection. En effet, l'impact de l'invasion du symbiote sur le Ne mitochondrial dépend de la liaison génétique entre leurs génomes. Ainsi, des transferts horizontaux de bactéries ou une transmission paternelle, en rompant la co-transmission des deux éléments, conduit à une réduction de l'impact de l'infection sur la diversité mitochondriale.

Notons que les infections par *Wolbachia* peuvent également avoir un effet opposé sur le polymorphisme mitochondrial au sein des populations infectées. Chez les guêpes du figuiers, *Ceratosolen solmsi* (Xiao et al. 2012) et chez *Drosophila quinaria* (Dyer et al. 2011), par exemple, les spécimens infectés ont des mitochondries fortement divergentes de celles des spécimens non-infectés (respectivement supérieurs à 6% et à 9% sur le gène CO1). Dans ces deux groupes, on peut supposer que les lignées mitochondriales des spécimens infectées ont été introduites dans ces espèces en même temps que l'infection par *Wolbachia*, suite à une hybridation. Ces évènements d'introgression mitochondriale peuvent en outre être à l'origine de discordances entre phylogénies des mitochondries et des populations (Hurst and Jiggins 2005).

L'introgression d'une mitochondrie dans une nouvelle population à la suite d'une hybridation peut théoriquement avoir lieu par hasard, dans le cas où la lignée mitochondriale exogène est fixée par dérive. L'haplotype mitochondrial introduit peut également être sélectionné en raison d'un avantage sélectif. Toutefois, au sein d'un organisme, les génomes mitochondriaux et nucléaires sont fortement liés fonctionnellement, dans la mesure où les mitochondries ont d'importantes fonctions cellulaires qui mettent en jeu des protéines codées par des gènes à la fois mitochondriaux et nucléaires. On peut donc supposer qu'un variant mitochondrial exogène est généralement moins adapté au nouveau contexte nucléaire que les variants coadaptés. En d'autre terme la coadaptation entre génomes mitochondriaux et nucléaires devrait représenter un frein aux introgressions mitochondriales. La présence d'un symbiote cytoplasmique peut néanmoins entrainer la propagation d'un haplotype mitochondrial exogène qui lui est lié génétiquement. Dans ce cas, l'introgression dépendra de la relation entre le pouvoir invasif du symbiote et de celle de la sélection contre le variant mitochondrial associé.

Au cours de cette thèse, j'ai cherché à évaluer la fréquence de ce type de discordances, causées par des introgressions mitochondriales entre marqueurs

mitochondriaux et nucléaires. Je me suis également intéressée à la mesure de l'impact des symbiotes cytoplasmiques sur les tailles efficaces des génomes mitochondriaux.

4. Echantillonnage

Les spécimens étudiés proviennent d'une communauté d'Arthropodes échantillonnés en Polynésie française dans le cadre du projet SymbioCode mené par Sylvain Charlat. Ce système comprend 11000 spécimens représentant plus de 1000 espèces d'Arthropode terrestres (Figure 3).

Figure 3. Quelques exemples d'échantillons du système Symbiocode

Ces communautés peuplent un archipel d'îles volcaniques éloignées des masses continentales (Figure 4), associées à un point chaud actif. Elles sont donc récentes (0.8 à 3 millions d'années) et leur peuplement est caractérisé par certaines spécificités liées à leur caractère insulaire. Tout d'abord, ces communautés sont peu diverses, ce qui signifie que les spécimens échantillonnés représentent probablement une proportion importante des espèces présentes. Ensuite, les évènements d'introduction d'espèce étant rares, les espèces présentes correspondent à des taxons isolés ou à des radiations récentes, datant au maximum de l'âge de l'archipel. Cet échantillonnage a été réalisé dans le cadre de l'étude des pertes et acquisitions de symbioses. L'organisation en îles maintient des structurations génétiques récentes, et permet donc d'observer des variations entre populations, même pour des phénomènes évolutifs rapides tels que la perte ou l'acquisition de symbiotes.

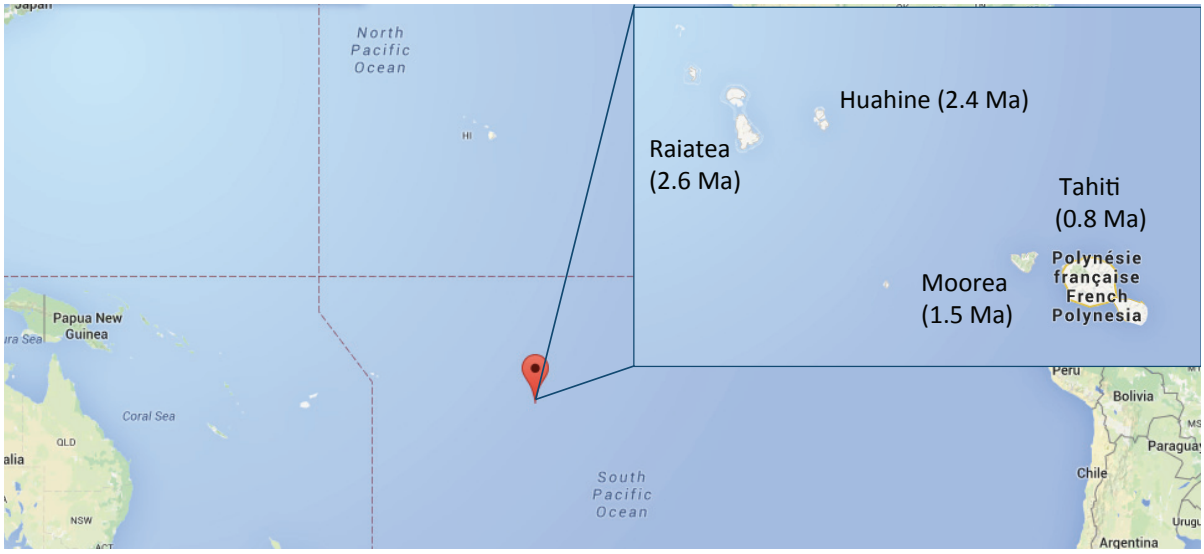

Figure 4. Localisation de l'Archipel des Sociétés.

Plusieurs études ont déjà été menées, ou sont en cours, sur les spécimens de ce système. Notamment, des projets visant à mesurer les transferts de *Wolbachia* entre taxons, ainsi que les transferts horizontaux d'éléments transposables, ont été conduites. Dans ce cadre, différents types de données ont déjà obtenues, dont j'ai pu utiliser une partie durant ma thèse. Notamment un barcode mitochondrial CO1 d'environ 660 paires de bases, a été séquencé pour 3600 spécimens. Des statuts d'infection par *Wolbachia* ont également été obtenus au moyen de PCR amplifiant un marqueur 16S et certaines de ces infections ont pu être caractérisées par le séquençage d'une autre marqueur, FbpA (Simões et al. 2011). Ces données sont utilisées pour la reconstruction des histoires des pertes et acquisitions de *Wolbachia* entre espèces de ce système.

5. Objectifs de la thèse

Ma thèse porte sur l'étude et la mesure de l'impact global des infections par *Wolbachia* sur l'évolution des génomes mitochondriaux des espèces hôtes. Tout d'abord, ces infections peuvent favoriser des événements d'introgressions mitochondriales à l'origine de discordances entre les histoires évolutives des génomes mitochondriaux et des génomes nucléaires. Il s'agira donc de quantifier ces discordances dans le but d'évaluer l'impact des introgressions mitochondriales sur l'évolution des génomes

mitochondriaux. Notamment, nous chercherons à évaluer si ces discordances sont trop fréquentes pour que les mitochondries constituent de bons marqueurs de l'histoire des populations.

Par ailleurs, les invasions par des symbiotes cytoplasmiques peuvent affecter les effectifs efficaces des populations mitochondriales (Ne). L'importance de cet effet dépend de l'intensité et de la fréquence des évènements de balayages sélectifs associés à ces symbiotes, c'est à dire de la dynamique invasive de ces infections. Ainsi, la mesure de l'impact global des infections sur les variations de Ne mitochondrial devra être interprétée en tenant compte ces dynamiques d'invasion. Enfin, cette étude permettra de mieux comprendre les déterminant des Ne mitochondriaux chez les Arthropodes, et le rôle de *Wolbachia* dans leurs variations.

Le chapitre 2 de cette thèse présente les travaux réalisés concernant l'obtention et l'analyse de données de RADseq. Nous avons tout d'abord étudié la question de l'utilisation des données de RADseq pour résoudre des relations phylogénétiques entre espèces. Puis, nous avons développé une méthode de préparation des librairies RADseq optimisée, notamment dans les cas impliquant le multiplexage de nombreux spécimens d'origines taxonomiques différentes. Nous avons ensuite cherché à mesurer et à corriger certains biais affectant la mesure du polymorphisme à partir de données de RADseq.

Ensuite, au cours du chapitre 3, j'aborderai la question de la détection et la quantification de différents effets potentiellement reliés aux infections par *Wolbachia* sur l'évolution des génomes mitochondriaux. Les questions suivantes constituent ainsi les différentes parties de ce chapitre: 1. Les discordances nucléo-cytoplasmiques liées à des introgressions sont-elles fréquentes? 2. Les réductions de Ne mitochondrial causées *Wolbachia* affectent-t-elles la diversité mitochondriale des populations infectées ? 3. Ces mêmes réductions de Ne mitochondrial affectent-t-elles l'efficacité de la sélection dans les lignées infectées ?

Chapitre 2

Développements autour du RADseq : Obtention, analyse et utilisation des marqueurs RAD

Dans le cadre de cette thèse, des marqueurs RADseq ont été obtenus pour de nombreux spécimens d'Arthropodes issus du système Symbiocode. Ce développement m'a permis d'étudier différents aspects de l'obtention et de l'utilisation de ce type de marqueurs. Tout d'abord, j'ai pu montrer que ces marqueurs pouvaient être utilisés pour résoudre des relations phylogénétiques entre spécimens à une échelle à laquelle cette méthode n'était pas utilisée précédemment. Nous avons également mis en place un protocole de préparation des librairies RADseq optimisé, adapté notamment au multiplexage de nombreux spécimens d'espèces différentes. Je présenterai ce protocole et l'analyse des données ainsi obtenues dans une troisième partie de ce chapitre. Il avait également été montré que l'utilisation de ce type de données pouvait introduire des biais dans l'estimation des diversités génétiques. Dans une quatrième partie, je présenterai ces biais et une étude concernant la mesure de leur impact et leur possible correction par une méthode bayésienne approximée (ABC).

1. Reconstruire les histoires évolutives des populations avec des marqueurs RAD

Un des principaux objectifs de ma thèse est l'étude de la concordance entre génomes mitochondriaux et nucléaires. Les introgressions mitochondriales causant des discordances entre ces génomes se traduisent en effet par le partage de mitochondries proches entre deux espèces distinctes. Une partie importante de ce travail a donc reposé sur l'obtention de marqueurs nucléaires permettant de comparer l'histoire de génomes nucléaires à celles des génomes mitochondriaux. Je présenterai tout d'abord la méthode RADseq utilisée, dans le contexte des approches par « réduction » utilisées en génomiques des populations.

1.1. Réduction de la représentation (*reduced représentation, RR*) en génomique des populations

Un ensemble de techniques de construction de librairies, dites par « réduction de la représentation », a été développé et utilisé dans le contexte de l'utilisation des NGS (*Next Generation Sequencing*) en génomique des populations. Ces approches visent au séquençage de sous-parties des génomes, homologues chez plusieurs spécimens. Cette définition est proche de celle des techniques de « génotypage par séquençage » (*Genotyping-by-sequencing*). Au sens large, ces méthodes regroupent l'ensemble des approches utilisant les techniques de séquençage haut débit dans le but d'identifier, de génotyper, des variations entre individus. Il peut ainsi s'agir d'expérience de reséquençage de génomes complets ou du séquençage de banques réduites. Toutefois, le terme *Genotyping-by-sequencing* est souvent utilisé comme synonyme de « séquençage de banques RADseq » (Narum et al. 2013, Schilling et al. 2014). En effet, parmi les approches par « représentation réduite » (RR), celles apparentées au RADseq, *Restriction Associated DNA sequencing*, c'est à dire celles utilisant des enzymes de restriction pour définir les régions du génome ciblées, représentent une part importante. Je détaillerai ces approches dans la suite de cette partie.

Mentionnons toutefois d'autres approches permettant également de construire des librairies réduites. Tout d'abord, le séquençage de transcriptomes permet d'obtenir les séquences des régions codantes des génomes. Les méthodes de RNAseq permettent

d'obtenir ce type de données. Elles sont souvent utilisées pour l'étude de l'expression des gènes, généralement dans le cadre d'études comparatives des transcriptomes et des niveaux d'expressions entre différentes conditions ou différents tissus. Mais elle peuvent également être utilisées en génomique des populations comme un moyen d'obtenir des séquences homologues de sous-parties des génomes chez plusieurs spécimens afin d'étudier la variabilité de ces marqueurs (Gayral et al. 2013). On peut remarquer que cette technique implique d'utilisation d'ARN frais, ce qui exclut son utilisation dans certains projets. Par ailleurs, la profondeur de séquençage des différents transcrits, c'est à dire la quantité de séquences obtenue pour chaque position, est reliée à leur niveaux d'expression. Ainsi, le séquençage des transcrits faiblement exprimés peut être difficile et se faire au prix d'une profondeur de séquençage moyenne élevée, ce qui peut limiter le nombre d'individus pouvant être intégré dans une librairie de RNAseq.

D'autres approches reposent sur le séquençage de produits de PCR multiplexés correspondant à de nombreux marqueurs. Par exemple, le séquençage ciblé d'amplicons, TAS, *targeted amplicon sequencing*, a permis à Bybee et al. 2011 de séquençer par 454 les amplicons de 6 locus chez 44 taxons de pancrustacés. Enfin, les méthodes de capture emploient de sondes s'hybridant à des régions ciblées du génome (Denonfoux et al. 2013) qui peuvent ensuite être séquencées à haut débit.

Ces trois approches sont généralement utilisées pour le séquençage de régions homologues codantes dans plusieurs génomes. L'obtention de ce type de séquences constitue le principe même du séquençage de transcriptomes. Dans les cas du séquençage d'amplicons et de produits de capture, ce sont les contraintes liées à l'utilisation de sondes ou d'amorces de PCR s'hybridant aux génomes qui expliquent que les régions ciblées correspondent le plus souvent à des régions codantes. A l'inverse dans le cas du RADseq, le principe de la sélection repose uniquement sur les occurrences de motifs courts dans les génomes, les régions obtenues sont donc théoriquement réparties aléatoirement le long des génomes.

1.2. Caractères spécifiques du RADseq

La figure 5, inspirée de Baird et al. 2008, représente le principe du protocole original de RADseq. Cette méthode permet le séquençage des régions flanquantes des

sites de restriction au sein d'un génome (Davey and Blaxter 2010). Les génomes sont tout d'abord digérés au moyen d'une enzyme de restriction. Des adaptateurs barcodés, c'est-à-dire comportant une partie spécifique pour chaque individu, sont liguées aux extrémités issues de la digestion. Ces barcodes servent d'identifiants moléculaires et permettent par la suite l'assignation des lectures à des individus.

Ces longs fragments, dont chaque extrémité correspond à un site de restriction associé à un adaptateur sont ensuite fragmentés, le plus souvent mécaniquement, par sonication, en plus petits fragments et un deuxième type d'adaptateurs est ligué à chaque extrémité. Les fragments comportant un adaptateur barcodé à une extrémité et un adaptateur non barcodé à l'autre sont ensuite amplifiés et séquencés, à partir de l'extrémité correspondant au site de restriction uniquement (*single-end*) ou des deux extrémités (*paired-end*). Cette méthode permet ainsi le séquençage des régions flanquantes de l'ensemble des sites de restriction d'une enzyme donnée.

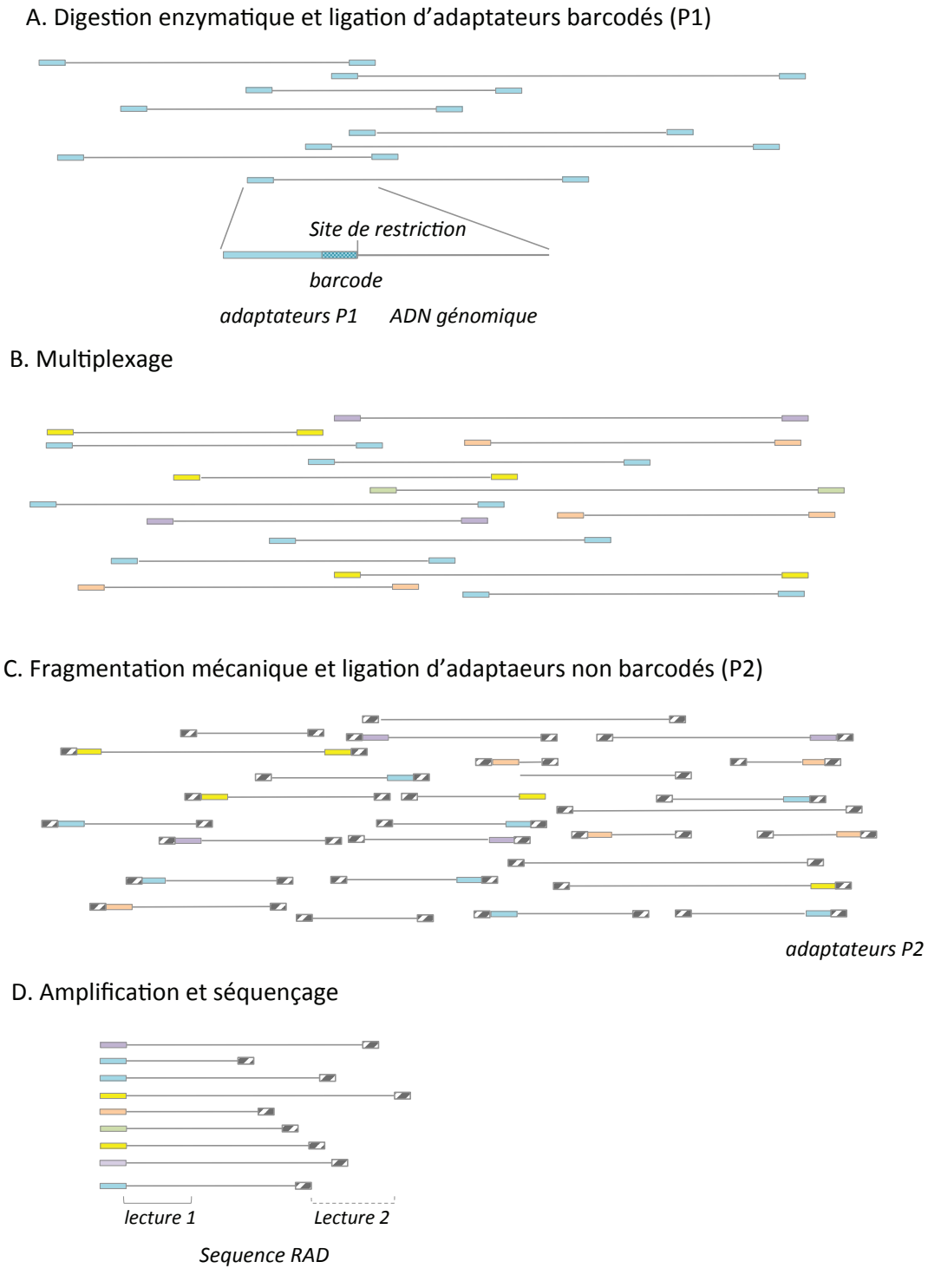

Figure 5. Protocole d'obtention de marqueurs RADseq, d'après Baird et al. 2008. A. Ligation d'adaptateurs barcodés (P1) aux extrémités issues de la digestion enzymatique pour chaque spécimen. B. Multiplexage des fragments correspondant aux différents spécimens. C. fragmentation mécanique et ligation d'adaptateurs non barcodés (P2) aux fragments. D. Amplification et séquençage Illumina des fragments associés à un site de restriction.

Ce protocole permet ainsi l'obtention de marqueurs répartis aléatoirement le long du génome. Plusieurs auteurs ont étudié l'impact de biais potentiels affectant la couverture des locus RAD et l'utilisation de ce type de données pour le génotypage. Par exemple, Davey et al. 2012 ont montré que les locus RAD ne sont pas tous couverts avec une même probabilité dans les librairies RADseq. Tout d'abord, dans les librairies correspondant à des enzymes dont les sites de coupure sont fréquents (*frequent cutter*), il peut exister une relation entre la longueur des fragments de restriction (écart entre deux sites de restriction consécutifs) et la profondeur de séquençage des sites correspondant. Ce biais peut conduire à une diminution de la couverture des locus correspondant aux extrémités de petits fragments. La réduction de la profondeur de séquençage observée affecte surtout des fragments de moins de 10 kBases. Il est donc probable que ce biais ait un impact plus faible dans le cas de la fragmentation mécanique de fragments plus longs correspondant à une enzyme de restriction aux sites de coupures plus rares. D'autre part, l'efficacité des amplifications par PCR peut être affectée par la composition en base des séquences amplifiées (Dabney and Meyer 2012). On peut donc également attendre un biais d'amplification en faveur des fragments riches en GC au sein des librairies RAD. Mais ce biais semble avoir un impact mineur sur la couverture des locus RAD (Davey et al. 2012).

Par ailleurs, le protocole de RADseq implique que seuls les fragments associés à des sites de restriction intacts seront séquencés. L'impact du polymorphisme sur les sites de restriction a été étudié notamment par Arnold et al. 2013 et Gautier et al. 2013. Ces biais feront l'objet de la quatrième partie de ce chapitre.

1.3. De nombreuses variantes : flexibilité du protocole

Une particularité du RADseq est d'avoir permis de nombreuses adaptations du protocole original rendant possible d'importantes variations de la densité et de la longueur des marqueurs obtenus. Cette flexibilité permet à cette méthode d'être adaptée à une grande variété de projets et d'objectifs correspondant à des nombres de marqueurs souhaités différents. Les expériences visant à la construction de cartes de liaison ou à la détection de sites sous sélection privilégieront ainsi l'obtention de marqueurs les plus denses possibles le long du génome d'un nombre modéré de

spécimens. En revanche, pour d'autres applications, par exemple pour la résolution de la biogéographie d'une espèce, l'obtention de centaines ou de milliers de marqueurs correspond déjà à une quantité de données très importante. Dans ce cas, la réduction du nombre de marqueurs obtenus par spécimen peut permettre l'obtention de données correspondant à un plus grand nombre de spécimens.

En premier lieu, le nombre de marqueurs obtenus pour un génome donné peut varier fortement en fonction de l'enzyme ou des enzymes de restriction utilisée(s). D'autre part, des méthodes de RADseq alternatives ont été développées, poursuivant deux objectifs distincts: tout d'abord, certaines de ces méthodes permettent une sélection plus fine du nombre de marqueurs obtenus pour chaque spécimen. Par ailleurs, certaines proposent également des simplifications techniques par rapport au protocole original. Par exemple, dans le protocole de ddRADseq (*double digest RAD sequencing*), une première digestion du génome est réalisée au moyen d'une enzyme à site de coupure rare. Puis, l'étape de fragmentation mécanique est remplacée par une deuxième digestion enzymatique, au moyen d'une enzyme à sites de coupures fréquents (Figure 6). Cette deuxième digestion génère des fragments de tailles variables entre locus, mais fixe pour un couple de sites de restriction adjacents donné. La sélection des fragments en fonction de leur taille permet ainsi ensuite de sélectionner une sous partie des locus RAD associés à la première enzyme (Peterson et al. 2012). Cette sélection des fragments peut être calibrée de façon à réduire plus ou moins drastiquement le nombre de fragments sélectionnés. Une limite possible de cette approche est que pour que les locus homologues soient séquencés chez deux individus, ils doivent être associés à des sites de restriction intacts pour les deux enzymes, correspondant au deux extrémités. On peut donc s'attendre à ce que la perte des locus homologues liés au polymorphisme sur les sites de restriction ait un impact plus fort sur ce type de données que sur celles issues du protocole classique (Gautier et al. 2013, Arnold et al. 2013).

Figure supprimée pour cause de droits d'auteur
Figure supprimée pour cause de droits d'auteur

Figure supprimée pour cause de droits d'auteur

Figure 6. Protocole de (A) RADseq classique et de (B) double digest RADseq (d'après Peterson et al. 2012). Le protocole ddRAD permet de sélectionner parmi les fragments associés à la première enzyme de restriction, ceux pour lesquels la deuxième digestion génére des fragments de taille comprises entre une longueur minimum et maximum. Dans le cas a, le site de restiction de l'enzyme à site de coupure fréquents (croix grise) est situé à proximité du site de restriction de l'enzyme à site de coupure rare (croix noir). Dans le cas b, le second site de coupure est très éloigné du premier. Dans les deux cas, le fragment généré ne sera pas présent dans la librairie ddRAD.

Wang et al. 2012 ont proposé un protocole différent impliquant l'utilisation d'enzymes de restriction des types IIB (Figure 7). Ces enzymes, comme les enzymes de restriction classique, reconnaissaient des motifs spécifiques dans les séquences d'ADN, mais celles-ci découpent l'ADN non à ces sites mais à une distance fixe en amont et en aval du motif. Ce protocole permet ainsi la génération de fragments de taille déterminée. Contrairement aux enzymes classiques, ces digestions génèrent des extrémités cohésives de séquence aléatoire. Les adaptateurs ligués à ces fragments peuvent ainsi avoir eux même des extrémités cohésives dégénérés, ou non. Dans le deuxième cas, ces bases cohésives permettent de sélectionner une sous partie des fragments de restriction possédant des extrémités complémentaires des adaptateurs.

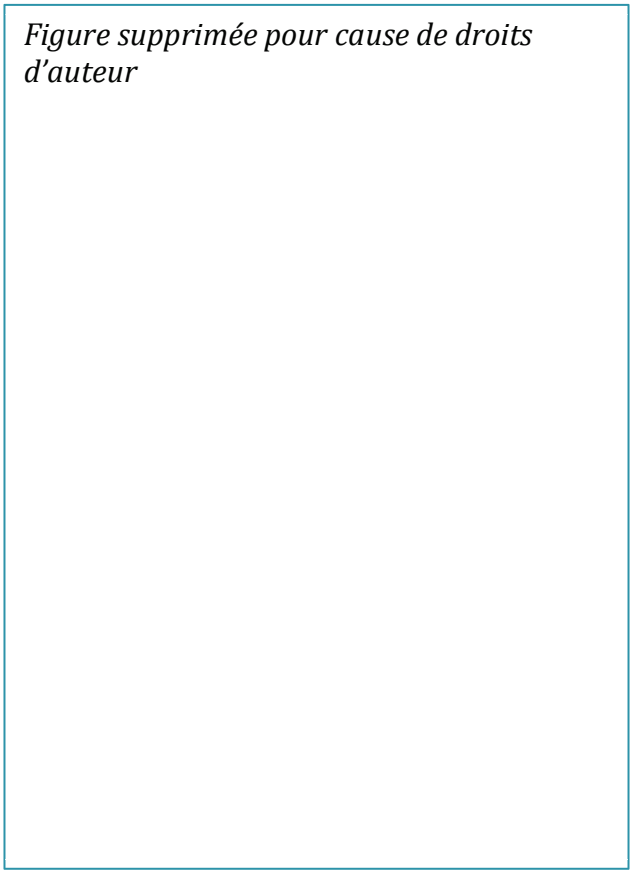

Figure 7. Protocole de 2b-RADseq, d'après (Wang et al. 2012). Les enzymes de restriction de type IIB digèrent les génomes en amont et en aval de leur site de reconnaissance.

D'autres protocoles alternatifs ne seront pas détaillés ici. Parmi ces approches, on peut citer le GBS (*genotyping-by-sequencing*) au sens restreint, utilisée par Elshire et al. (2011) et l'ezRAD (Toonen et al. 2013). La majorité de ces méthodes dérivées ont pour objectif de permettre la préparation de librairies RAD sans étape de fragmentation. En effet, cette étape, qui implique la sonication ou la nébulisation des fragments de restriction est techniquement délicate. Dans le protocole original, cette étape permet d'utiliser une digestion enzymatique correspondant à un nombre de site de restriction limité, et d'ensuite réduire mécaniquement la taille des fragments pour les rendre compatible avec les méthodes de séquençage NGS. Les protocoles alternatifs utilisent ainsi souvent des digestions enzymatiques générant un très grand nombre de fragments, et en sélectionnent ensuite une sous partie. Dans le protocole de GBS, une étape de PCR permet d'amplifier sélectivement des fragments de taille appropriée. En ezRAD (comme en ddRAD), une étape de sélection des fragments de la bonne taille, permet cette réduction de la taille et du nombre des fragments séquencés.

Les avantages et les inconvénients de ces différents approches ont été discutés dans plusieurs articles de synthèse (Andrews et al. 2014, Puritz et al. 2014). Brièvement, l'avantage principal des méthodes alternatives est une plus grande facilité de mise en œuvre. Néanmoins, seul le protocole principal permet l'obtention d'une deuxième lecture dont la position n'est pas unique. Le séquençage de cette deuxième extrémité permet d'assembler des contigs correspondant à l'ensemble de la région couverte par les fragments RAD, quand la quantité de lectures obtenues par locus est suffisante. L'assemblage de contigs à partir de lectures appariées permet d'obtenir des marqueurs plus longs (Etter et al. 2011, Willing et al. 2011). En outre, nous verrons au cours de ce chapitre que ces lectures peuvent être utiles à la réduction d'un biais causé par l'amplification par PCR des fragments RAD.

1.4. De nombreux domaines d'application

Les données de RADseq peuvent donc prendre des formes différentes en terme de nombre, de densité et de longueur des marqueurs obtenus. Le caractère modulable de cette approche explique la variété des études l'utilisant. Des données de RADseq ont en effet, été obtenues dans un grand nombre d'espèces, modèles ou non, issues de groupes taxonomiques variés (épinoche, Hohenlohe et al. 2010, aubergine, Barchi et al. 2011, teigne des crucifères, Baxter et al. 2011 etc.).

Parmi les études reposant sur l'analyse de données de RADseq, on peut tout d'abord citer celles employant ces marqueurs pour la détection de locus sous sélection entre différentes populations. Hohenlohe et al. 2010, par exemple, ont utilisé des données de RADseq pour identifier des régions associées à des patrons de diversité ou de différenciations spécifiques entre des populations d'épinoches marines et d'eau douce. Cette étude a ainsi permis l'identification de régions associées à des patrons de sélection balancées ou diversifiantes, dont certaines avaient été auparavant identifiées par des recherches de QTL. Plusieurs études ont également utilisé des données de RADseq pour identifier des locus impliqués dans la détermination de patron de couleur dans différentes espèces de cichlidés (Takahashi et al. 2013, Henning et al. 2014). Des données de RADseq sont en outre fréquemment utilisées pour réaliser des cartes de liaison génétiques pour des espèces dont les génomes n'ont pas été séquencés (Baxter et al. 2011).

Ces études impliquent l'utilisation de marqueurs répartis de façon dense sur les génomes pour explorer des variations entre régions génomiques. Une autre branche des utilisations des techniques de RADseq s'intéresse à la reconstruction des histoires évolutives des populations. Dans ce contexte, il est intéressant de disposer d'un grand nombre de marqueurs correspondant à des positions distinctes le long des génomes. En effet, ce type de données permet la résolution d'une histoire moyenne des génomes, en réduisant l'impact des locus individuels aux histoires potentiellement discordantes. Le nombre de marqueurs requis dans ce but est toutefois moins élevé, dans la mesure où il ne s'agit pas ici de couvrir l'ensemble du génome, mais seulement une portion suffisante pour constituer un « échantillon » représentatif. Dès les premiers développements de la méthode, des données de RADseq ont ainsi été employées pour reconstruire la biogéographie de populations au sein des espèces (Emerson et al. 2010) ou les phylogénies d'espèces correspondant à des radiations récentes (Wagner et al. 2013). Dans ces cas de diversifications rapides, l'approche RADseq permet l'obtention d'un grand nombre de marqueurs, ce qui augmente d'autant le nombre de positions variables utilisables. En outre, à ces échelles, de nombreux locus sont sujets à un tri de lignées incomplet. Dans ce contexte, l'obtention d'un grand nombre de marqueurs indépendants permet de réduire l'impact de ce polymorphisme ancestral sur l'histoire « moyenne » reconstruite.

Par ailleurs, j'ai étudié plus spécifiquement la question de l'utilisation des données de RADseq pour la résolution de relations phylogénétiques à une profondeur plus importante. Ces résultats seront détaillés dans la partie suivante.

2. L'utilisation des données RADseq en phylogénie

Au début de cette thèse, des données de ce type étaient utilisées depuis peu, principalement pour comparer des spécimens au sein des espèces. L'utilisation de ce type de marqueurs pour comparer des spécimens d'espèces différentes, dépend de l'existence d'un nombre de sites de restriction conservés suffisants entre spécimens. Le cas échéant, les marqueurs RADseq présentent de nombreux intérêt en phylogénie. En effet, cette méthode permet l'obtention de nombreux locus (10^3 à 10^5), et peut être appliquée à des espèces non-modèles.

2.1. Peut-on résoudre des phylogénies avec des données de RADseq ? étude *in silico* (article)

Afin de tester la pertinence de l'utilisation de ces marqueurs pour la résolution des relations phylogénétiques entre espèces, j'ai simulé l'obtention de données de RADseq à partir de génomes complets de 12 espèces de Drosophiles correspondant à des temps de divergences variant entre 4 et 63 Millions d'années. Ces données obtenues *in silico* nous ont permis d'estimer la relation entre le nombre de sites utilisables et le temps de divergence entre organismes, ainsi que de reconstruire les relations phylogénétiques attendues entre ces douze espèces. Ces résultats sont présentés dans l'article suivant.

Ecology and Evolution

Open Access

Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization

Marie Cariou, Laurent Duret & Sylvain Charlat

Université de Lyon, Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 43 boulevard du 11 novembre 1918, Villeurbanne F-69622, France

Keywords
Bioinfomatics/phyloinfomatics, molecular evolution, phylogenetic theory and methods, phylogeography.

Correspondence
Marie Cariou, Université de Lyon, Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 43 boulevard du 11 novembre 1918, Villeurbanne F-69622, France.
Tel: +33 4 72 43 29 08; Fax: +33 4 72 43 13 88; E-mail: marie.cariou@univ-lyon1.fr

Funding Information
The funding was provided by the CNRS-ATIP to SC.

Received: 14 December 2012; Revised: 10 January 2013; Accepted: 17 January 2013

Ecology and Evolution **2013; 3(4): 846–852**

doi: 10.1002/ece3.512

Abstract

Inferring phylogenetic relationships between closely related taxa can be hindered by three factors: (1) the lack of informative molecular variation at short evolutionary timescale; (2) the lack of established markers in poorly studied taxa; and (3) the potential phylogenetic conflicts among different genomic regions due to incomplete lineage sorting or introgression. In this context, Restriction site Associated DNA sequencing (RAD-seq) seems promising as this technique can generate sequence data from numerous DNA fragments scattered throughout the genome, from a large number of samples, and without preliminary knowledge on the taxa under study. However, divergence beyond the within-species level will necessarily reduce the number of conserved and non-duplicated restriction sites, and therefore the number of loci usable for phylogenetic inference. Here, we assess the suitability of RAD-seq for phylogeny using a simulated experiment on the 12 *Drosophila* genomes, with divergence times ranging from 5 to 63 million years. These simulations show that RAD-seq allows the recovery of the known *Drosophila* phylogeny with strong statistical support, even for relatively ancient nodes. Notably, this conclusion is robust to the potentially confounding effects of sequencing errors, heterozygosity, and low coverage. We further show that clustering RAD-seq data using the BLASTN and SiLiX programs significantly improves the recovery of orthologous RAD loci compared with previously proposed approaches, especially for distantly related species. This study therefore validates the view that RAD sequencing is a powerful tool for phylogenetic inference.

Introduction

Resolution of phylogenies between closely related species can be problematic for a number of reasons. First, due to incomplete lineage sorting and introgression, different loci might trace different evolutionary histories. In addition, most nuclear markers lack resolution at short evolutionary scales. And finally, in poorly studied taxa, molecular markers might not have been developed. In this context, Restriction site Associated DNA sequencing (RAD-seq) appears as a promising approach. This technique relies on the high throughput sequencing of genomic regions flanking restriction sites (Baird et al. 2008; Davey et al. 2010; McCormack et al. 2013; Rowe et al. 2011). It thus generates numerous homologous markers, scattered throughout genomes, potentially from hundreds of specimens in a single sequencing run. Moreover, if the precise number of loci is not critical, this approach is potentially universal, as it does not require preliminary knowledge on the taxa under study (Baxter et al. 2011).

The RAD-seq technique was initially designed to generate informative molecular variation within species, and has repeatedly proved its efficiency for this purpose (e.g., Emerson et al. 2010; Hohenlohe et al. 2010). However, the utilization of RAD-seq to compare genomes from different species can potentially be hindered by a number of caveats. Specifically, the following conditions must be fulfilled for RAD-seq to be suitable for phylogenetic inference: (1) enough restriction sites must be conserved between species; (2) the flanking regions must be sufficiently conserved for homology to be detectable by sequence similarity; and (3) the resulting alignments must contain enough phylogenetic signal.

Recent studies, published as this article was in preparation, concur to suggest that these conditions are often fulfilled (Rubin et al. 2012; Wagner et al. 2012). Here, we

use a simulated RAD-seq experiment on the 12 *Drosophila* genomes (Clark et al. 2007) to assess if this is the case at various degrees of molecular divergence, and to optimize the procedure of RAD data analysis for phylogeny. Consistently with Rubin et al. (2012), we were able to recover and align enough sequences from orthologous loci to reconstruct the known (whole-genome-based) phylogeny (Clark et al. 2007), with good statistical support. We further show here that sequence clustering based on the BLASTN and SiLiX programs significantly improves the recovery of orthologous RAD loci. This study therefore validates and reinforces RAD-seq as a powerful tool for phylogenetic inference.

Divergence time and RAD loci conservation

The number of RAD loci potentially usable to compare two specimens is the number of restriction sites conserved in their genomes, which is expected to decrease as divergence time increases. We used the 12 *Drosophila* genomes to establish the relationship between genome divergence and restriction sites conservation. Complete genomes and pairwise alignments of the *D. melanogaster* genome with each of the 11 others were downloaded from http://genome.ucsc.edu/. Conserved target sites of the restriction enzyme Sbf1 were counted in each pairwise alignment. This enzyme is one of the most commonly used in RAD-seq studies because its 8 bp, GC-rich recognition site is rare enough in genomes to maximize sequencing coverage for each locus.

Figure 1 shows the relationship between restriction site conservation and sequence divergence (measured at fourfold degenerated sites of codons), with divergence times ranging from 5.4 to 63 My (Tamura et al. 2004). The genome of *D. melanogaster* contains 2,948 sites, 49.1% and 50.5% of which are conserved in its two closest relatives *D. simulans* and *D. sechellia*. After 63 My of divergence, *D. melanogaster*, respectively, shares 4.9, 4.8, and 5.1% of its restriction sites with *D. grimshawi*, *D. mojavensis*, and *D. virilis*, which represent 145, 142, and 149 restriction sites. Notably, these estimates are conservative as the 2,948 restriction sites in *D. melanogaster* are counted from its complete genome, while conserved sites are only detected in the aligned genomic regions.

How many RAD tags are usable for phylogeny?

The above results suggest that more than 100 Sbf1 cut sites are conserved between species having diverged for 63 My (Fig. 1). However, for the genomic regions flanking these sites to be usable for phylogenetic inference,

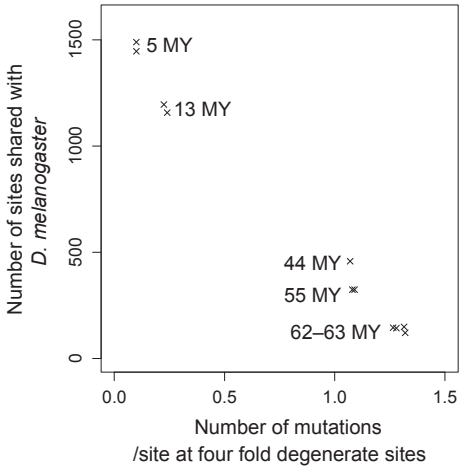

Figure 1. Relationship between molecular divergence at fourfold degenerate sites (*Drosophila* consortium, 2007) and Sbf1 restriction site conservation. Note: genomic regions aligned more than once against the *D. melanogaster* genome were excluded. Numbers next to crosses indicate divergence time.

sequences must be sufficiently conserved for homology to be detected. Moreover, paralogous loci may confound phylogenetic analysis and should therefore be excluded.

Illumina sequencing reactions yield 101 bp or 51 bp long reads. Here, we focus on the longest reads, which maximize the potential phylogenetic signal (see below for a discussion on this issue). Each read starts with a barcode sequence identifying a sample (up to 8 bp long) and the 8 bp restriction site followed by 85 bp of usable data. We thus extracted 85 bp in 5′ and 3′ of each restriction site to set up the list of RAD sequences existing in each *Drosophila* genome. To identify sets of homologous sequences among the 12 species, we used a two-step procedure: first, we performed all-against-all BLASTN comparisons (Altschul et al. 1990); and second, we analyzed these results with the SiLiX software (Miele et al. 2011), to cluster sequences sharing a minimum level of sequence identity over a minimal length (see below for a discussion on the optimal parameter values). As can be seen in Table 1, the proportion of orthologous RAD tag pairs retrieved by BLASTN and SiLiX (i.e., gathered in the same cluster) decreases with divergence time from 99% (between *D. melanogaster* and *D. simulans*, 5.4 My of divergence) to 49% (*D. melanogaster* and *D. wilistoni*, 63 My of divergence).

Restriction sites located in repeated regions may be problematic for phylogenetic analysis and should therefore be identified and discarded. For every pair of species, we thus identified clusters containing more than one locus from at least one of the two species. The exclusion of these clusters leads to a loss of less than 10% of

Table 1. Number of known and retrieved orthologous RAD tags in each species pair. "Orthologous tags": total number of orthologous RAD tags present in pairwise alignments (*D. melanogaster* vs. each of the 11 other species). "Retrieved orthologous tags": proportion of orthologous tags clustered by SiLiX. "In clusters including paralogs": proportion of the retrieved orthologous tags clustered in groups containing more than one locus. Loci are defined here based on genome sequences (see main text). Node depth from Tamura et al. (2004).

Species pair *D. melanogaster*	Node depth (My)	Orthologous tags	Retrieved orthologous tags (%)	In clusters including paralogs (%)
D.sechellia	5.4	2978	99	5
D.simulans	5.4	2892	99	4
D.erecta	12.6	2390	97	3
D.yakuba	12.8	2314	97	8
D.ananassae	44.2	916	68	9
D.persimilis	54.9	648	65	9
D.pseudoobscura	54.9	648	66	9
D.wilistoni	62.2	242	49	6
D.grimshawi	62.9	290	60	8
D.virilis	62.9	286	59	5
D.mojavensis	62.9	298	59	8

the RAD tags (Table 1). Among the clusters containing only single copy loci, we observed 2% of "false positives," that is, clusters containing loci that are not considered as orthologous based on whole-genome alignments.

Phylogenetic inference from simulated RAD-seq data

The above estimates rely on genome alignments for the identification of clusters containing paralogous RAD loci. In a true RAD-seq experiment, the "locus" definition would be solely based on an initial step of clustering of reads from each specimen. At this step, recently duplicated RAD tags may be mistakenly grouped into a single "locus" and too divergent alleles from the same locus can be mistakenly identified as two different loci. To estimate more realistically which proportion of the data would be usable for phylogenetic inference, we therefore simulated sequencing and "intra-specimen" clustering steps in our analysis. This simulation also allows us to test the impact of sequencing errors, heterozygosity, and coverage variation.

To assess the effect of heterozygosity on data analysis, a second haploid genome was simulated for each species using random mutations of the sequenced genome, to produce a 5% average distance between homologous alleles (the upper bound of realistic polymorphism values). RAD sequencing was then simulated by randomly sampling reads from the list of all possible RAD loci from diploid genomes, with a 10× mean coverage per locus, which was shown to cover 99% of all RAD loci at least once (see below). For each sampled read, sequencing errors were added with a uniform 1% error rate, that is, the upper bound of Illumina error rate estimates (Illumina technical support, Glenn 2011). Intra-individual clustering of the reads was performed using the ustacks program (Catchen et al. 2011). This program forms "stacks" (groups of strictly identical reads within individuals) and clusters similar stacks to form putative loci, with a user-defined maximum number of differences between stacks within a locus. Ustacks finally aggregates secondary reads (that were not initially placed in a stack) into existing stacks, to better estimate coverage. To maximize the number of retrieved heterozygous loci, we allowed up to 13 differences between stacks within a locus. We observed that this high value did not increase the number of inferred loci containing paralogous reads (3% of the loci). We also allowed up to nine mismatches to cluster secondary reads to putative loci. Using such parameters, 11.6% of the loci were mistakenly split by ustacks. This shows that a high level of heterozygosity in the data does not hamper the clustering of reads within loci (only few reads from the same locus are mistakenly clustered into different loci).

Orthologs were then searched with BLASTN and SiLiX as described above, using as input the consensus of each ustacks locus. We observe that 1.1% of the clusters that are known to contain paralogs (based on genome alignments) are not identified as such in the simulated experiment, as they were mistakenly merged by ustacks in a single locus. These correspond to very recent duplication events that should not confound the phylogenetic signal, unless many of these duplications are anterior to the latest speciation event.

For each cluster, sequences were aligned with Muscle v3.8.31 (Edgar 2004) using default parameters. We selected and concatenated the 2,275 topologically informative alignments (i.e., containing sequences from at least four different species, Fig. 2), adding gap sequences to represent missing orthologs. The proportion of gaps in the alignment varied from 19.2% for *D. simulans* to 94.9% for *D. wilistoni* (Table 2), which is expected considering the topology of the phylogenetic tree (the probability to recover at least three orthologs for a given locus is higher for species that have many close relatives in the set of sampled species). A maximum likelihood phylogeny was built using PhyML (Guindon et al. 2010). Bootstrap support was calculated with 100 replicates. The expected phylogenetic relationships among the 12 *Drosophila* species, established from whole-genome comparisons (Clark et al. 2007), were correctly recovered with strong bootstrap supports (Fig. 3). Because the number of RAD loci conserved between genomes decreases with divergence

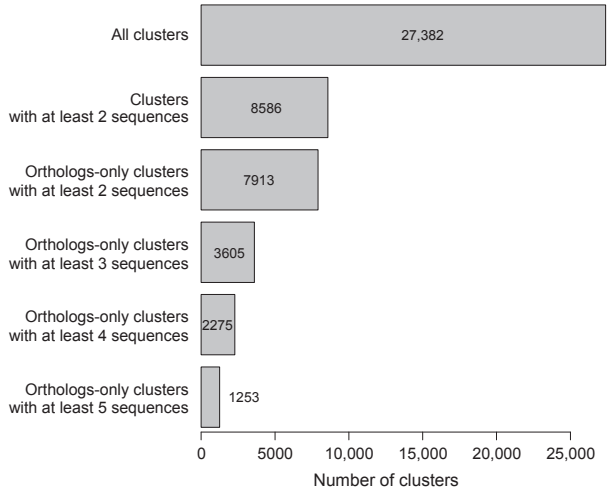

Figure 2. Results of SiLiX clustering of RAD sequences from the 12 *Drosophila* genomes.

Table 2. Percentage of gaps for each species in the concatenated alignment, after exclusion of loci present in less than four species.

Species	Percentage of gaps in the concatenated alignment (%)
D.melanogaster	24.6
D.sechellia	20.4
D.simulans	19.2
D.erecta	25.5
D.yakuba	23.4
D.ananassae	75.5
D.persimilis	80.4
D.pseudoobscura	80.0
D.wilistoni	94.9
D.grimshawi	91.8
D.virilis	91.5
D.mojavensis	91.1

time, the number of loci that can be used to infer phylogeny tends to decrease for the deepest branches of the tree (Fig. 3). However, there remains enough phylogenetic signal to infer the correct topology with high confidence, even for the most ancient branches.

Practical issues: sequencing coverage, number of specimens, and read length

In a typical RAD-seq experiment, DNA samples from several individuals are tagged with molecular identifiers and multiplexed in the same flow cell lane. The average sequencing coverage per locus per individual is given by the following formula:

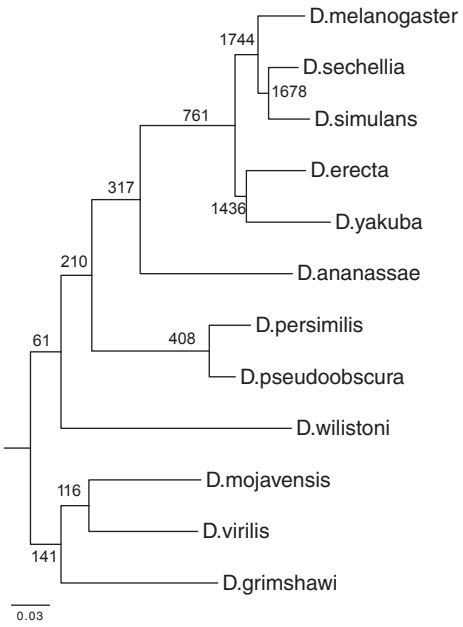

Figure 3. RAD-seq-based phylogeny of the 12 *Drosophila* species, based on 100-bp-long RAD-seq reads, inferred by maximum likelihood using PhyML 3.0. under a GTR+ G substitution model, using the concatenated alignments from orthologous-only clusters containing at least four sequences. Bootstrap values (100 replicates) were equal to 100% on every node. We indicate the number of informative loci at each node (shared by at least one species on each side of the bifurcation and at least one outgroup).

$$coverage = \frac{R}{\sum_{i=1}^{i=N} 2 \times S_i}$$

where R is the total number of reads, N the number of individuals included in the library preparation, and S_i the number of restriction sites in the genome of individual i. Increasing the pool size allows a decrease in sequencing cost per sample, but leads to a lower coverage per locus, which should affect the number of loci that can be recovered. The proportion of loci that was sequenced at least once in the in silico experiment follows a Poisson distribution: with 5× or higher coverage, more than 99% of the expected loci are sequenced at least once; this proportion drops to 95% with a 3× coverage. However, RAD loci represented by only one read are not identified as valid RAD loci by ustacks (Emerson et al. 2010; Morris et al. 2011). Using two as the minimum coverage to create a stack in ustacks parameters, the proportion of loci for which at least one allele was recovered after the intra-individual clustering step is 88.3% for a 10× mean coverage, which was sufficient in our simulations to recover the expected phylogeny. The fact that only 88% of the loci are recovered (although 99.95% are sequenced at least twice) is due to the high levels of polymorphism and sequencing

error used in our simulations. Although those values are conservative, we would recommend an increased coverage.

Notably, a significant proportion of restriction sites fall in recently duplicated regions, which reduces the number of RAD loci actually usable for phylogeny. For example, the reference sequence (haploid) of the *D. melanogaster* genome contains 2,948 restriction sites, that is, 5,896 potential RAD tags. Intra-individual clustering of reads from this genome using ustacks yields 4,296 clusters, 3% of which contain reads from more than one locus. This small proportion of clusters containing recently duplicated regions represents a substantial proportion of all reads (25.3%). Thus, 74.7% of reads actually fall in non-recently duplicated regions.

The 12 *Drosophila* species analyzed here contain on average 2,308 Sbf1 restriction sites per genome. Sbf1 is a rare-cutter enzyme because of its 8-bp-long and GC-rich restriction site. It is expected to yield among the lowest number of restriction sites per genome. This property is valuable for studies where a large number of specimens matters more than a large number of loci per genome. For example, with an Illumina Hiseq 2000 flow cell lane ($R = 1.5 \times 10^8$ reads), and for a targeted coverage of 10×, it would, in theory, be possible to pool DNA samples from up to 1000 specimens (for a genome size comparable to that of Drosophila). Of course, a number of issues would potentially lower the coverage for some loci or some specimens (e.g., genome size, GC content, DNA template concentration, and quality). Targeting a 50× mean coverage would allow the analysis of 200 specimens in one flow cell lane, and would be robust to a 10-fold variation in coverage between loci or specimens.

With existing sequencing technologies, it is also possible to increase coverage or reduce sequencing cost by decreasing the length of the reads. For example, Illumina sequencing runs can yield 101-bp- or 51-bp-long reads. We investigated whether 51 bp RAD-seq reads could be used for phylogenetic studies. We extracted 35 bp sequences on both side of each restriction sites and performed the in silico RAD-seq experiment and phylogenetic analysis as described above (see supplementary material for details). Overall, the topology shows only one error (the position of *D. wilistoni* is incorrect). Recent nodes (younger than 45 My) were resolved with high support within the *Melanogaster* group (*D. melanogaster, D. simulans, D. sechellia, D. yakuba, D. erecta,* and *D. ananassae*), but bootstrap values were low (<80%) for some deep nodes. This poor support for deep nodes suggests that short reads should only be used to resolve short-scale phylogenies.

Orthology inference: SiLiX versus uclust

As we were finishing this manuscript, Rubin et al. (2012) published a comparable analysis and also reached the conclusion that RAD-seq can be a useful tool for phylogenetic inference. One notable difference between the two studies is the use of different procedures to identify orthologous loci: uclust (Edgar 2010) versus SiLiX.

In an attempt to optimize this crucial step, we conducted a comparison of the two methods using the same data set and various parameter values. We estimated the efficiency of orthology detection by calculating the proportion of orthologous RAD loci (known from the whole-genome alignment) that were correctly recovered (i.e., known orthologs gathered in the same cluster, without inclusion of any paralog). It should be noticed that this definition of "efficiency" reflects the combined effect of sensitivity (the ability to gather orthologous sequences) and specificity (the ability to exclude paralogs). To assess the effect of evolutionary distance on clustering efficiency, we computed this measure for clusters shared by at least 4, 6, or 9 species (Fig. 4).

The first step of the SiLiX procedure consists in comparing all sequences against each other with BLASTN. To avoid the detection of spurious sequence similarities, we allowed the filtering of low complexity sequences (parameter F = T, which is set by default in BLASTN) and we set the *E*-value parameter to 10^{-4} (i.e., stringent enough to avoid the detection of similarities between non-homologous sequences). In the second step, SiLiX takes into account two parameters to cluster two sequences in a family: the fraction of their length covered by the BLASTN alignment (alignment overlap) and their sequence similarity (in the aligned region). The clustering of uclust is based on a global alignment, and requires one single parameter (the minimum percentage of identity over the entire sequence length). As expected, for both methods, the efficiency of orthology detection decreases with increasing evolutionary distance (compare in Fig. 4 the efficiency for RAD loci shared by 4, 6, or 9 taxa). As noted by Rubin et al. (2012), the efficiency of uclust decreases when the sequence similarity threshold is set to very high value ($\geq$80%), and we observed the same trend for SiLiX (Fig. 4). For SiLiX, the efficiency also tends to decrease with increasing alignment overlap threshold. We therefore recommend using relatively permissive thresholds (sequence identity $\geq$35%, and sequence overlap $\geq$35%). Interestingly, we observed that SiLiX was significantly more efficient than uclust, especially for clusters including some relatively distant homologs: for loci shared by at least nine species, SiLiX was two times more efficient than uclust (44.7% vs. 21.6% efficiency). In principle, the main interest of uclust is that it is extremely fast, because it does not require an exhaustive comparison of all sequences against each other's. However, given the number of sequences that have to be compared in a typical RAD-seq experiment (several thousands of loci for several hundreds of specimens),

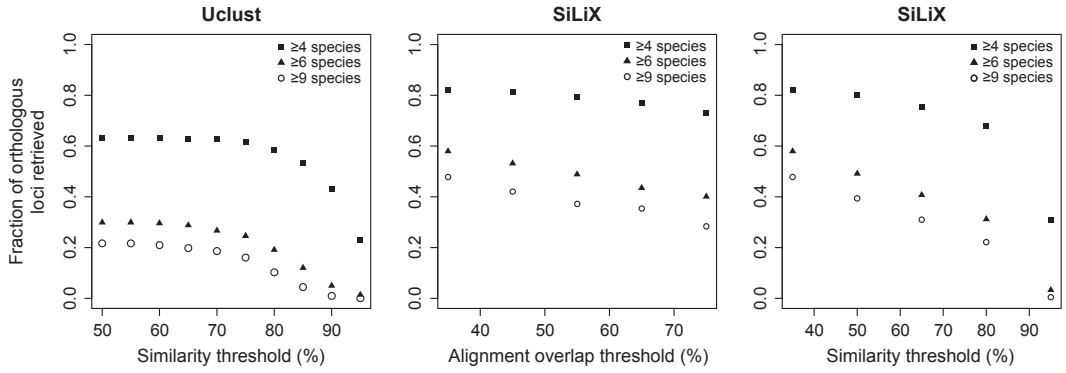

Figure 4. Proportion of known orthologous sequences retrieved by SiLiX or uclust, for different clustering parameters. Results are shown for clusters containing at least 4, 6, or 9 species. Note the x-axes are not the same in all figures. For SiLiX, the values of two parameters are tested: the minimum overlap threshold (on the second figure, where the minimum identity parameter is set at 0.35), and the identity threshold (on the third figure, where the minimum overlap is set at 0.35).

calculation time is not limiting. For example, in our analyses, the clustering of 55,324 sequences took 7 s with uclust versus 3 min for BLAST + SiLiX (on a single Macbook Pro with a 2.9 GHz processor and 8 Gb of memory). Moreover, one disadvantage of uclust is that its result depends on the sequence that is used as a seed to initiate the clustering. Hence, uclust clustering has to be replicated many times with different seeds to check results consistency. As the BLAST + SiLiX method is exhaustive, appears to be more efficient, and does not require to be replicated with different seeds, we argue that it should be preferred to uclust.

Conclusion

Our phylogenetic analysis using simulated RAD-seq data suggests that this method is suitable for interspecific comparisons, even for relatively large genetic divergences (1.3 substitutions per site, which corresponds to 63 My for the *Drosophila* clade). Enough non-duplicated restriction sites were conserved between species and sequence conservation between orthologous RAD tags was sufficient to detect homology for a large number of loci. Finally, the recovered alignments contained enough phylogenetic information to yield strongly supported phylogenies. It should be noted that in our simulations, we did not incorporate incomplete lineage sorting. However, empirical data suggest that RAD-seq produces data from enough loci to overcome this problem (Wagner et al. 2013).

This study further indicates that the SiLiX clustering program is more efficient than uclust to identify orthologous RAD sequences, especially for distantly related species, for which this feature is most critical. In addition, we observed that RAD-seq-based phylogenies are robust to sequencing errors and high polymorphism values. In practice, we recommend to target a 50× coverage, which is sufficient to sample 99% of all RAD alleles at least once, even with a 10-fold coverage variation between specimens, and to avoid the use of short reads (50 bp), which would lead to significant loss of phylogenetic signal for deep nodes. Overall, this study validates and reinforces RAD-seq as a powerful tool for phylogenetic inference.

Methods

We provide below program versions and parameter values used in this study, when not specified in the main text:

Ustacks (Catchen et al. 2011)
ustacks -t fasta –f file.fasta -r -m 2 –N 9 –M 13.
BLASTN: blastall 2.2.25 (Altschul et al. 1990)
blastall -i file.fasta -d file.fasta -p blastn -o file.blastn -W 11 -b 10000 -v 10000 -z 1000000 -e 1e-4 -m 8
SiLiX (Miele et al. 2011)
silix file.fasta file.blastn -r 0.35 –i 0.35
Muscle v3.8.31 (Edgar 2004)
Default parameters
PhyML 3.0. (Guindon et al. 2010)
phyml -i file -d nt -b 100 -m GTR.

Acknowledgments

We thank the CNRS INEE for funding (ATIP grant SymbioCode held by SC). This study was supported by the Centre National de la Recherche Scientifique, and by the Agence Nationale de la Recherche (Ancestrome: ANR-10-BINF-01-01).

Conflict of Interest

None declared.

References

Altschul, F., W. Gish, W. Miller, E. W. Myers, D. J. Lipman, and F. A. Altschul. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, et al. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3:e3376.

Baxter, S. W., J. W. Davey, J. S. Johnston, A. M. Shelton, D. G. Heckel, C. D. Jiggins, et al. 2011. Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. PLoS ONE 6:e19315.

Catchen, J. M., A. Amores, P. Hohenlohe, W. Cresko, and J. H. Postlethwait. 2011. Stacks: building and genotyping Loci de novo from short-read sequences. G3 (Bethesda, Md.) 1:171–82.

Clark, A. G., M. B. Eisen, D. R. Smith, C. M. Bergman, B. Oliver, T. A. Markow 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature 450:203–18.

Davey, J. W., J. L. Davey, M. L. Blaxter, and M. W. Blaxter. 2010. RADSeq: next-generation population genetics. Brief. Funct. Genomics 9:416–23.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–7.

Edgar, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics (Oxford, England) 26:2460–1.

Emerson, K. J., C. R. Merz, J. M. Catchen, P. A. Hohenlohe, W. A. Cresko, W. E. Bradshaw, et al. 2010. Resolving postglacial phylogeography using high-throughput sequencing.

Glenn, T. C. 2011. Field guide to next-generation DNA sequencers. Mol. Ecol. Resour. 11:759–69.

Guindon, S., J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59:307–21.

Hohenlohe, P. A., S. Bassham, P. D. Etter, N. Stiffler, E. A. Johnson, and W. A. Cresko. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet. 6:e1000862.

McCormack, J. E., S. M. Hird, A. J. Zellmer, B. C. Carstens, and R. T. Brumfield. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol. Phylogenet. Evol. 66:526–538.

Miele, V., S. Penel, and L. Duret. 2011. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics 12:116.

Morris, G. P., P. P. Grabowski, and J. O. Borevitz. 2011. Genomic diversity in switchgrass (Panicum virgatum): from the continental scale to a dune landscape. Mol. Ecol. 20:4938–52.

Rowe, H. C., S. Renaut, and A. Guggisberg. 2011. RAD in the realm of next-generation sequencing technologies. Mol. Ecol. 20:3499–502.

Rubin, B. E. R., R. H. Ree, and C. S. Moreau. 2012. Inferring Phylogenies from RAD Sequence Data. PLoS ONE 7:e33394.

Tamura, K., S. Subramanian, and S. Kumar. 2004. Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. Mol. Biol. Evol. 21:36–44.

Wagner, C. E., I. Keller, S. Wittwer, O. M. Selz, S. Mwaiko, L. Greuter, et al. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22:787–798.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Phylogeny of the 12 *Drosophila* species based on 50-bp-long RAD-seq reads, inferred by maximum likelihood using PhyML 3.0. under a GTR + G substitution model, using the concatenated alignments from orthologous-only clusters containing at least four sequences. Bootstrap values (100 replicates) are indicated in italics. Rad-seq data were simulated with 1% sequencing errors, but without polymorphism. Notice that bootstrap supports for several deep nodes are low and that the placement of *D. wilistoni* is incorrect. Non-italic numbers indicate the number of informative loci for each node (shared by at least one species on each side of the bifurcation and at least one outgroup).

2.2. Phylogénies RADseq dans la littérature : Confirmations in silico et empiriques

Nos résultats indiquent que des données de RADseq peuvent être utilisées pour résoudre des relations phylogénétiques entre espèces d'insectes divergeant depuis 4 à 63 millions d'années. D'autres études, publiées pendant la rédaction de cet article, ont confirmé ces résultats. Tout d'abord, Rubin et al. 2012 ont également réalisé des expériences de RADseq *in silico* sur des génomes complets afin d'étudier la pertinence de ce type de données pour la résolution de relations phylogénétiques. Trois groupes différents étaient considérés dans leur étude : 12 espèces de Drosophiles, ainsi que 8 espèces de levures associées à des divergences maximales de 300 Ma et 11 espèces de Mammifères dont la divergence maximale correspond à environ 100 Ma. Rubin et al ont pu retrouver la phylogénie attendue pour les Drosophiles, mais les nœuds les plus profonds des deux autres groupes, n'ont pas pu être inférés correctement, ce qui s'explique sans doute par une divergence trop importante de ces taxons. On peut remarquer que la méthode de *clustering* utilisée dans notre étude pour identifier les locus orthologues entre spécimens est plus sensible que celle utilisée par Rubin et al. Ainsi, Rubin et al. suggèrent que les données RADseq ne sont pas appropriées pour résoudre des phylogénies correspondant à des temps de divergence de l'ordre de 100 millions d'années, mais on peut se demander si un *clustering* plus sensible ne pourrait pas permettre de résoudre ces nœuds plus profonds. Toutefois, cette hypothèse n'a pas été testée dans notre étude.

Depuis ces premiers résultats obtenus *in silico*, plusieurs auteurs ont confirmé empiriquement l'intérêt de l'utilisation de données de RADseq pour reconstruire des relations phylogénétiques entre espèces. Parmi ces études, on peut citer notamment la résolution de relations phylogénétiques entre espèces de carabes (Cruaud et al. 2014) et de scarabées (Takahashi et al. 2014). D'une manière générale, des données de RADseq sont fréquemment utilisées pour répondre à des questions impliquant des comparaisons interspécifiques (Wang et al. 2013, Escudero et al. 2014, Hipp et al. 2014, Pante et al. 2014, Viricel et al. 2014). Les taxons étudiés dans ces études correspondent généralement à des temps de divergences courts. Les données de RADseq sont ainsi particulièrement utiles pour la résolution de relations phylogénétiques entre espèces correspondant à des radiations rapides et récentes, pour lesquels le nombre de

marqueurs variables disponibles étaient auparavant limitant, comme chez les cichlides (Wagner et al. 2013). Toutefois, des relations phylogénétiques entre espèces correspondant à des temps de divergences de plusieurs dizaines de millions d'années ont été résolues par l'utilisation de données RADseq par exemple chez le chêne (23-33 Ma, Hipp et al. 2014) et chez les poissons zèbres (17 Ma, McCluskey and Postlethwait 2014).

2.3. Discussions méthodologiques sur l'utilisation de données RAD en phylogénie

L'obtention de données de type RADseq, à partir d'individus d'espèces différentes permet de reconstruire les relations phylogénétiques, correspondant aux phylogénies moyennes de nombreux locus. Ce type de données permet également d'étudier la variabilité des histoires évolutives des différents locus. Les locus RADseq sont généralement courts, et chacun comporte peu d'information. Il est donc difficile d'évaluer si un locus en particulier est impliqué dans une introgression. Toutefois, ce type de données peut permettre de détecter de manière globale, l'existence d'un flux de gènes correspondant à un signal d'introgression d'une espèce vers une autre (Andrew et al. 2013).

Des données de RADseq obtenues pour des spécimens d'espèces différentes ont ainsi été étudiées par Eaton et al. pour détecter des introgressions entre espèces de *pedicularis* (une herbacée de la famille des Scrophulariaceae) (Eaton and Ree 2013). Cette première étude reposait sur l'utilisation d'une méthode ABBA BABA. Cette méthode permet de détecter des introgressions entre espèces par la comparaison des patrons de ségrégation d'allèles entre 4 espèces dont les relations sont de la forme (((P3,P2)P1)O). Un locus possédant un allèle ancestral A et un allèle dérivé B, retrouvé chacun chez 2 spécimens, soutient cette phylogénie si la forme B est retrouvées dans les espèces P2 et P3, ce qui correspond à un patron de ségrégation de type : BBAA. Les locus associés à des patron de ségrégation discordant (BABA ou ABBA), peuvent être dus à des tris de lignée incomplets, à des homoplasies ou à des introgression. Toutefois, les polymorphismes ancestraux et les homoplasies, devraient générer aussi fréquemment des patrons BABA que ABBA. La comparaison de la fréquence de ces deux patrons permet ainsi de détecter des cas ou la présence de locus ségrégant de façon discordante

n'est pas expliquée par le tri de lignée incomplet ou l'homoplasie mais probablement par l'existence d'une introgression.

Ces auteurs ont également développé des méthodes dédiées à la détection et à la quantification de ces flux asymétriques de matériel génétique. Ces méthodes permettent de mettre en évidence l'existence de topologies alternatives, soutenues par beaucoup de locus, mais masquées par le signal majoritaire (*partitioned RAD visualiation method* (Escudero et al. 2014, Hipp et al. 2014)). Cette méthode a été appliquée à la détection d'introgressions entre espèces de Carex d'une part et entre espèce de chênes, d'autre part.

Certaines analyses traitant de l'utilisation de données de RADseq en phylogénie se sont intéressées aux choix des paramètres optimaux pour ces inférences. Dans ce cadre, la proportion de données manquantes dans les alignements utilisés pour inférer des relations phylogénétiques est un paramètre important. On peut en effet choisir d'utiliser une proportion plus ou moins élevée des locus RAD en sélectionnant les locus partagés au minimum par un nombre plus ou moins grande de spécimens. Ce paramètre est d'autant plus critique que, quand la divergence entre organismes augmente, la proportion de données manquantes peut devenir très importante. Il s'agit donc de choisir le meilleur compromis entre l'utilisation d'un jeu de données relativement réduit mais comprenant peu de données manquantes, ou bien celle d'un très grand nombre de locus, associés à une proportion de données manquantes très élevée. Des études mesurant l'influence de ce paramètre suggèrent que l'utilisation de matrices comprenant une forte proportion de données manquantes est plus efficace que l'utilisation de filtres plus stringents (Rubin et al. 2012, Wagner et al. 2013, mais voir Takahashi et al. 2014).

2.4. Conclusion

Bien qu'initialement développées dans le cadre d'analyses intra-spécifiques, nous avons pu montrer au moyen de simulations que des données de RADseq pouvaient être utilisées pour résoudre des relations phylogénétiques entre espèces correspondant à des temps de divergence allant jusqu'à 63 Million d'année. Depuis, des marqueurs RADseq ont permis de résoudre de nombreuses phylogénies, notamment dans des groupes où des spéciations rapides conduisent à de forts tris de lignées incomplets, ou

bien où les hybridations et introgressions sont fréquentes. Dans ces contextes, les données de RADseq sont très intéressantes, d'une part du fait du grand nombre de marqueurs permettant de résoudre les relations entre taxons, d'autre part car le grand nombre de marqueurs permet de détecter de potentiels incongruences entre locus, associées par exemple à des cas d'introgressions.

3. Optimisation du protocole de RADseq

3.1. Introduction

Les méthodes RADseq permettent l'obtention de nombreux marqueurs homologues chez de nombreux spécimens. Un des objectifs principaux de cette thèse reposait sur l'obtention de données RADseq pour de nombreux spécimens, d'origines très hétérogènes et dont les extraits d'ADN possèdent également des caractéristiques très différentes (tailles de génomes, quantités d'ADN, composition en nucléotides…). L'obtention et l'analyse de ces données ont nécessité la mise au point de certaines adaptations qui seront détaillées dans cette partie.

Tout d'abord, la préparation des librairies RADseq implique l'utilisation d'adaptateurs nécessaires à la fois au séquençage et à l'assignation des lectures aux spécimens. Nous verrons que la synthèse et l'utilisation de ces adaptateurs sont critiques, à la fois financièrement et techniquement et nous présenterons une méthode de synthèse d'adaptateurs permettant d'optimiser ces deux aspects. Des adaptateurs synthétisés de cette manière ont été utilisés dans le cadre d'une expérience pilote. En séquençant cette librairie, nous avions également pour objectif de tester la robustesse du protocole RADseq, appliqué à la préparation de librairies hautement multiplexées. Nous souhaitions, en particulier, évaluer l'importance des variations de couverture entre spécimens et entre locus dans ce contexte spécifique. Enfin, cette expérience pilote a rendu possible la réalisation d'une seconde librairie RADseq correspondant à une sous-partie des échantillons étudiés dans le cadre de cette thèse. L'analyse des lectures issues du séquençage de cette librairie m'a permis de mettre au point un protocole d'analyse de données RADseq issues de ce type de librairies hautement multiplexées.

3.2. Une méthode de synthèse des adaptateurs permettant d'augmenter le niveau de multiplexage des librairies (article)

L'utilisation de librairies RADseq fortement multiplexées, c'est-à-dire comprenant de nombreux spécimens correspondant potentiellement à plusieurs espèces différentes, est limitée par des contraintes techniques. En particulier, le coût de synthèse des adaptateurs illumina contenant des identifiants moléculaires (MIDs)

devient excessif quand de nombreux spécimens doivent être multiplexés. Ensuite, la nécessité d'ajuster empiriquement le ratio de la concentration des adaptateurs et de l'ADN génomique constitue un obstacle à l'application du protocole de RADseq à des échantillons de nature hétérogène, en terme de concentration et de qualité d'ADN.

J'ai participé à l'élaboration d'un protocole de synthèse d'adaptateurs optimisés, permettant de résoudre ces difficultés. Tout d'abord, la partie commune des adaptateurs peut être synthétisée indépendamment de la partie unique (MID). Ces deux parties peuvent ensuite être liguées l'une à l'autre. Cette approche permet une réduction importante des coûts de synthèse des adaptateurs, ce qui permet ainsi le multiplexage de centaines de spécimens. Ensuite, les ligations des adaptateurs les uns avec les autres (formation de dimères d'adaptateurs) rendait la concentration d'adaptateurs critique lors de l'étape de ligation à l'ADN génomique. Nous montrerons que la formation de ces dimères peut être évitées par l'utilisation d'adaptateurs non-phosphorylés, ce qui améliore significativement le rendement de cette étape et la quantité de lectures utilisables produites.

Genetica (2015) 143:139–143
DOI 10.1007/s10709-015-9828-3

Optimization of multiplexed RADseq libraries using low-cost adaptors

Hélène Henri · Marie Cariou · Gabriel Terraz · Sonia Martinez · Adil El Filali · Marine Veyssiere · Laurent Duret · Sylvain Charlat

Received: 29 April 2014 / Accepted: 3 February 2015 / Published online: 11 February 2015

Abstract Reduced representation genomics approaches, of which RADseq is currently the most popular form, offer the possibility to produce genome wide data from potentially any species, without previous genomic information. The application of RADseq to highly multiplexed libraries (including numerous specimens, and potentially numerous different species) is however limited by technical constraints. First, the cost of synthesis of Illumina adaptors including molecular identifiers (MIDs) becomes excessive when numerous specimens are to be multiplexed. Second, the necessity to empirically adjust the ratio of adaptors to genomic DNA concentration impedes the high throughput application of RADseq to heterogeneous samples, of variable DNA concentration and quality. In an attempt to solve these problems, we propose here some adjustments regarding the adaptor synthesis. First, we show that the common and unique (MID) parts of adaptors can be synthesized separately and subsequently ligated, which drastically reduces the synthesis cost, and thus allows multiplexing hundreds of specimens. Second, we show that self-ligation of adaptors, which makes the adaptor concentration so critical, can be simply prevented by using unphosphorylated adaptors, which significantly improves the ligation and sequencing yield.

Keywords Reduced representation genomics · RADseq · Protocol · Multiplexing

Introduction

Reduced representation genomics, as opposed to full genome sequencing or random metagenomics, allow the sequencing of numerous homologous loci from numerous genomic extracts. In particular, methods involving the sequencing of regions flanking restriction sites (RADseq and affiliated approaches) attract growing interest, for they can produce powerful datasets in any organism to address any question related to genetic variability, from genetic mapping to deep phylogenies (Davey and Blaxter 2011; Cariou et al. 2013).

In this article, we propose some adjustments of standard RADseq protocols (Etter et al. 2011) to facilitate the application of this method to large numbers of heterogeneous genomic extracts in a single sequencing reaction. Our first objective is to reduce the cost of synthesizing large numbers of adaptors, which becomes excessive when several hundreds of genomic extracts are to be multiplexed. Our second objective is to design a method that would be more robust to variation and uncertainty in the initial DNA concentration and genome size.

RADseq library preparation starts with a digestion of genomic extracts, with one restriction enzyme, targeting rare or common restriction sites, depending on the required density of markers (Baird et al. 2008) (in the so-called "double digest" protocol, marker density can also be adjusted by combining two restriction enzymes; Peterson et al. 2012). Following digestion, adaptors (denoted P1) are specifically ligated to the free sticky ends. P1 adaptors include regions for PCR amplification and

Electronic supplementary material The online version of this article (doi:10.1007/s10709-015-9828-3) contains supplementary material, which is available to authorized users.

H. Henri (✉) · M. Cariou · G. Terraz · S. Martinez · A. El Filali · M. Veyssiere · L. Duret · S. Charlat
Laboratoire de Biométrie et Biologie Evolutive, CNRS, UMR 5558, Université Lyon 1, Université de Lyon, 43 boulevard du 11 novembre 1918, 69622 Villeurbanne, France
e-mail: helene.henri@univ-lyon1.fr

Illumina sequencing, as well as a Molecular Identifier (MID) that should be at least 8 bp long if several hundreds of genomic extracts are to be sequenced together. The full adaptors contain two strands of respectively 72 and 68 nucleotides, bringing the cost of synthesis to approximately 100 Euros (~120$) per adaptor, that is, 50,000 Euros (~62,000$) for an experiment where 500 genomic extracts would be multiplexed. It would obviously be possible to reduce the cost of such an experiment by pooling fewer specimens in several independent libraries and sequencing reactions, with the positive outcome of increasing the average sequencing depth per locus. But if a large sequencing depth is not required, or if money is limiting, increasing the number of libraries and sequencing reactions is not an optimal approach. Alternatively, paired end sequencing can be used, and a combination of different MIDs in the forward and reverse adaptors then allows one to envisage high levels of multiplexing with a limited number of distinct MIDs. However, this approach is also not economical since it requires the preparation of independent libraries (one for each combination of forward and reverse tag), that can only be pooled at a late stage (that is, after the P2 adaptors have been ligated). We propose instead an additional step in the preparation of P1 adaptors that drastically reduces the synthesis cost: the long region that is common to all adaptors is synthesized only once, while the short region containing the MIDs are synthesized independently. The common and unique parts are then ligated to produce the complete adaptors. With these modifications, the cost of synthesizing 500 distinct adaptors drops from 50,000 to 5600 Euros.

The second modification we introduce aims at making library preparation more robust to variation in the initial amount of genomic DNA. Indeed, current protocols indicate that a deficit but also an excess of P1 adaptor reduces ligation efficiency, so that P1 concentration must be appropriately adjusted to genomic DNA concentration (Peterson et al. 2012), which cannot be reasonably achieved for a large number of heterogeneous genomic extracts. We hypothesized that this problem stemmed from the formation of P1–P1 dimers, and could thus be solved by using P1 adaptors that are devoid of a 5′ phosphate and thus cannot ligate to themselves.

We assessed these protocol modifications on a test RADseq experiment including a diverse set of genomic extracts from 11 species, and several replicates of *Drosophila melanogaster* DNA used as positive controls, that is, samples where a reference genome could be used to assess the quality of the data. We show that our cheap P1 adaptors are efficient and that the use of unphosphorylated adaptors significantly improves the yield and robustness of RAD library preparation.

Adaptors design and synthesis

8 bp MIDs were designed using the barcrawl program (Frank 2009), which provided 755 potential MIDs with the following characteristics and provided in the supplementary material S1: at least three differences between all MIDs, no homopolymers of more than two nucleotides, no hairpins or heteroduplexes of more than four nucleotides, no possible convergence of MIDs to identical sequences through 1 bp deletion. For the test experiment described here, a combination of 100 MIDs (listed in supplementary material S2) was selected to ensure an approximately balanced base composition at each position (even ratios of A, C, G and T). Indeed, cluster detection during Illumina sequencing is facilitated by heterogeneity of the fragments to be sequenced, especially in the five first nucleotides.

Figure 1 provides a schematic view of the preparation of our P1 adaptors, containing a region common to all adaptors and a variable region unique to each adaptor. The common region includes the sequence required for PCR and Illumina sequencing. The variable region contains the MID and 2 sticky ends on 5′ and 3′. The 5′ end corresponds to an overhang of the bottom strand that is specific of the restriction enzyme (and will thus ligate to the genomic DNA) and the 3′ end to an overhang of the top strand that is cohesive to the 5′ end of the common part.

A full, step by step protocol for the preparation of adaptors is provided in supplementary material 3. In brief, the common and unique parts of the adaptors (respectively in blue and red in Fig. 1) are synthesized and hybridized separately and finally ligated to produce 30 µL of complete adaptors at 0.5 µM. The 5′ end of these full adaptors is not phosphorylated so that ligation to the genomic DNA relies on the 5′ phosphate carried by digested genomic DNA. The resulting nick is filled in with Bst 2.0 polymerase (NEB).

To verify that ligation of P1 adaptors to genomic DNA was effective, we randomly chose an Sbf1 cut site in the *D. melanogaster* genome, and designed PCR primers on the two flanking regions (mel1-CGA-CCA-GCA-GAC-CAA-TAA and mel2-GCT-CCA-CTA-CCA-GCT-ATA-A). Two PCRs were performed on each *D. melanogaster* sample, involving a forward primer targeting P1 (AAT-GAT-ACG-GCG-ACC-ACC-GA), and the mel1 and mel2 two reverse primers targeting the regions flanking the SbfI site. The amplification of these two fragments confirmed that P1 adaptors were properly ligated.

Design of the test experiment

Genomic DNA of 77 specimens from 11 species (see table in supplementary material 4 for details) was extracted either with an affinity column (DNeasy kit, Qiagen) or

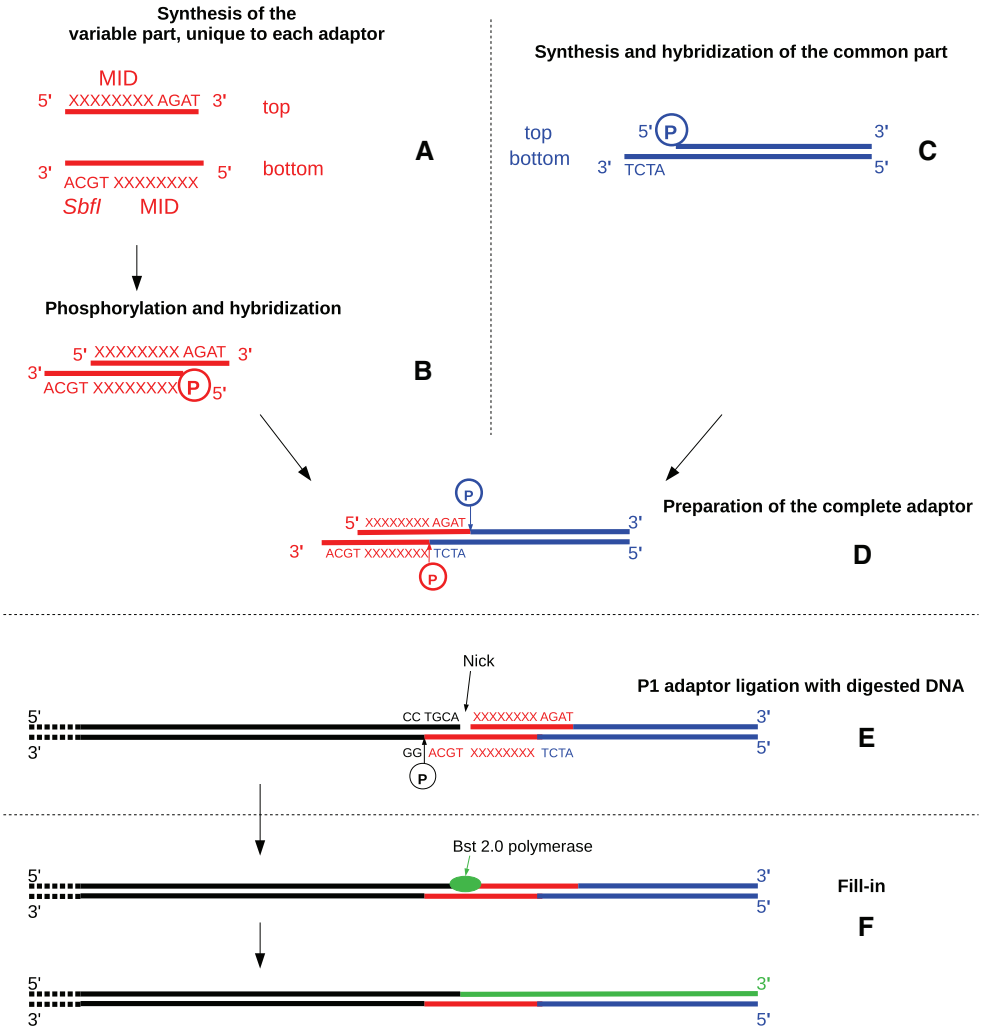

Fig. 1 Step by step synthesis of the P1 adaptors (*a–f*) and ligation to genomic DNA. *a* Synthesis of the two strands of the "MID + restriction site" part (*red*). *b* To reduce the cost of oligonucleotides synthesis, we order unphosphorylated oligos and add a phosphate to the 5′ end of the bottom strand only. The two strands are then hybridized. *c* Synthesis and hybridization of the common part (*blue*). The top strand is ordered with 5′ end phosphorylation. *d* The two parts of the adaptors are ligated with T4 DNA ligase to produce complete adaptors. *e* Ligation of the complete P1 adaptor with DNA digested with SbfI (*black*). Since the complete adaptor does not carry any 5′ phosphate, the ligation to the SbfI-digested DNA relies only on a single phosphate provided by the genomic DNA. *f* The nick on the top strand is filled-in with Bst2.0 polymerase through its 5′-3′ DNA polymerase and strand displacement activities. The newly synthesized region is shown in *green*. (Color figure online)

through phenol–chloroform precipitation (Sambrook and Russel 2001). Six specimens of *D. melanogaster* were included in this experiment to be used as positive controls, based on knowledge of their full genome. DNA concentration, as measured by fluorometry using Picogreen (Invitrogen) widely varied among specimens, ranging from 0.3 to 118 ng/μL, allowing us to assess the robustness of our protocol with regard to DNA concentration. Digestion and P1 ligation were performed as described in the step by step protocol provided in the supplementary material 3. Notably, the P1 adaptor was always in large excess (from 40 to 5000 X) relative to its genomic targets. Although some authors recommend a lower ratio of adaptor to genomic DNA (Peterson et al. 2012), presumably to avoid the excessive formation of P1–P1 dimers, we intentionally used a larger ratio to test the robustness of our approach to variation in genomic DNA concentration. To assess the efficiency of the two types of adaptors (classic, phosphorylated, versus new, unphosphorylated) we tested each of the *D. melanogaster* templates with the two types of adaptors (as well as some specimens of other species, see supplementary 4 for details). To assess the repeatability of the experiment, two *D. melanogaster* samples were also replicated, that is, tagged with two distinct MIDs. The experiment thus includes a total of 16 *D. melanogaster* samples: 2 adaptor types × (2 replicated templates + 4 non-replicated templates).

Library preparation and sequencing

Library preparation, starting from P1-ligated DNA, and library sequencing, were performed by the ProfileXpert facility in Lyon. The TruSeq library preparation was slightly modified from standard Illumina protocols. Namely, (1) samples were first pooled by species, purified using AmpureXP beads and DNA concentration was estimated. (2) One of our objectives was to test if the sequencing yield was robust to variation in the initial DNA concentration; we thus only partially standardized

concentrations in the pooled library, which resulted in a 55-fold variation between the most and least abundant genomic extract in the final pool. (3) Fragmentation by sonication, repair ends, size selection by gel excision, 3′-dA overhang addition and ligation to P2 adapters were performed following Illumina's recommendations. Finally (4), ligated fragments were amplified by PCR for 14 cycles from 10 μL of template and a single purification using AmpureXP beads was performed. 5 % of phiX DNA was added to the final library to facilitate cluster detection during the sequencing reaction.

Data processing

We used the Stacks software pipeline (version 1.11) to process the data (Catchen et al. 2013). The *process_radtags* program was used to assign reads to specimens (allowing no more than 1 mismatch in the MID) and to eliminate poor quality reads as well as reads devoid of the expected SbfI cut site after the MID sequence (options –barcode_dist 2 -q -e sbfI). For the *D. melanogaster* specimens, we used *Ustack* to produce consensus sequences of RAD tags (options -m 3 and -M 4). The consensus of each RAD tag (each stack) was then mapped to the *D. melanogaster* reference genome (dm3, Apr. 2006 assembly obtained from http://genome.ucsc.edu/) using BWA version 0.6.2 with default parameters (Li and Durbin 2009).

Results and discussion

One lane of HiSeq 2000 Illumina sequencing produced 192 million of single reads of 51 bp. 76 % of the reads passed the default quality filters and contained an identifiable MID (with no more than one mismatch). 58 % of those started with the expected SbfI overhang and were thus considered as valid for further analysis (see below for a solution to increase the proportion of valid reads). The *D. melanogaster* samples provide us with a mean to control the quality of the data. 92 % of the valid reads from these specimens mapped to RAD loci expected from the reference genome. Difference in the number of valid reads between two replicates varied from 0.1 to 26 %. Considering all samples, the average number of valid reads per specimen was 855,000, but varied from 20,120 to 13,930,998. As expected, a large part of this variation (48 %) was explained by the initial DNA concentration of each sample (ANOVA model: number of reads $\sim$concentration + type of adaptor, with square root of log transformation). Residual variation can be explained by variation among genomes in the density of cut sites. In particular, slight variation in GC content can have large impact on the density of restriction sites. For example a shift from 40 to 45 % GC induces a twofold increase in cut site density.

To compare the yield of our new adaptors (devoid of a 5′ phosphate and thus supposed to prevent the formation of P1 dimers) to that of classic adaptors, we used 11 DNA templates that were tagged with the two types of adaptors (two of which were replicated), making 13 possible comparisons between classic and new adaptors. In average, the templates ligated to the new adaptors produced 2.4 times more valid reads. Figure 2 provides the detailed data for the 13 comparisons, which clearly indicates that the avoidance of P1 dimers improves the yield of the reaction ($p = 0.001$, paired Wilcoxon signed rank test).

To further assess the benefit of using unphosphorylated P1 adaptors, we used data from the *D. melanogaster* controls to count the number of RAD loci that mapped to the

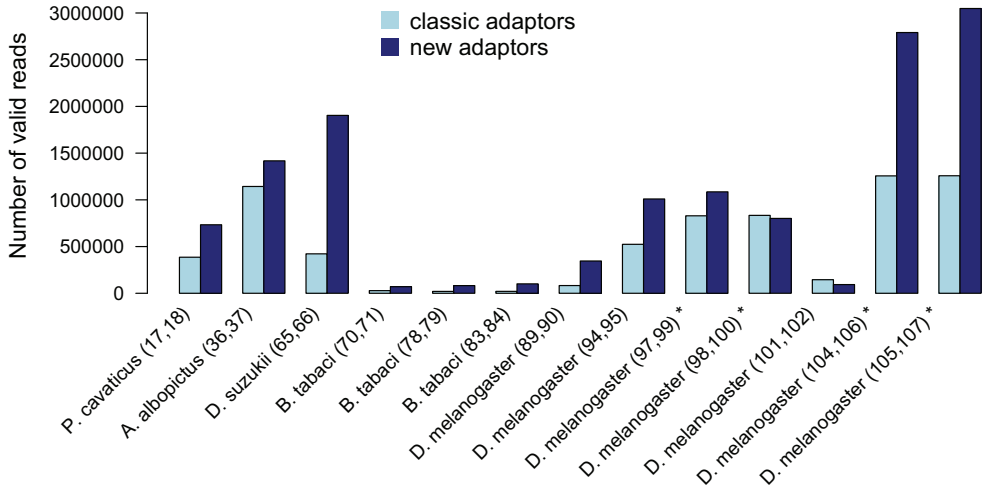

Fig. 2 Comparison of the classic and new adaptors, in number of valid reads per replicate. Sample ids are given in parenthesis after species name. *Experimental replicates

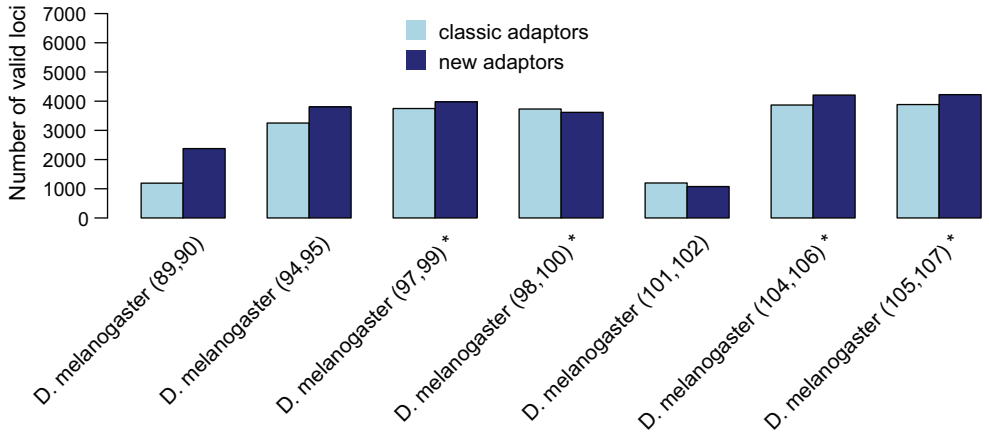

Fig. 3 Comparison of the classic and new adaptors, in number of loci per *D. melanogaster* replicate. Here only loci that mapped to the *D. melanogaster* reference genome were included. Sample ids are given in parenthesis after species name. *Experimental replicates

reference *D. melanogaster* genome (Fig. 3). The number of valid loci was significantly larger with unphosphorylated adaptors ($p = 0.04$, paired Wilcoxon signed rank test, 12 % more loci in average). (see table in supplementary material 5 for details). Notably, we did not observed P1–P1 dimers in our sequencing results, suggesting that such dimers are either eliminated by purification or do not get sequenced. In other words, the benefit of using unphosphorylated P1 adaptors appears to stem mainly from a more efficient ligation.

In contrast, we found that P1–P2 dimers represented a substantial proportion (52 %) of the "non valid reads" (devoid of the expected SbfI overhang), suggesting that free P1s remained in the solution and ligated non specifically to P2 adaptors. In an attempt to solve this problem, we added a second AMPure purification step after PCR amplification in a subsequent library preparation. This extra step increased the proportion of usable reads to much larger values (82 % on average, details not shown).

In conclusion, our experiment demonstrated that numerous distinct P1 adaptors can be obtained at reasonable cost if one synthesizes separately the constant and unique part of the adaptors, and that such "home made" adaptors are fully efficient. In addition, we showed that unphosphorylated P1 adaptors are more efficient than classic adaptors, providing more reads and more RAD tags. This likely results from the absence of P1–P1 dimers, making the ligation step more efficient, and also more robust to variation in the P1 to genomic DNA ratio. The modifications we propose allows one to envisage the preparation of highly multiplexed RADseq libraries, which will prove more and more useful as sequencing technologies increase the number of reads obtained in a single sequencing reaction.

Acknowledgments This work was funded by the Centre National de la Recherche Scientifique (ATIP Grant to SC) and the Agence Nationale de la Recherche (Grant ClimEvol). GT is the recipient of a Ph.D. studentship from the Rhône-Alpes region ("Program Cible" Grant). We would like to thank the two anonymous reviewers for their critical assessment of our work.

Conflict of interest The authors declare that they have no conflict of interest.

References

Baird N, Etter P, Atwood T et al (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3:e3376. doi:10.1371/journal.pone.0003376

Cariou M, Duret L, Charlat S (2013) Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. Ecol Evol 3(4):846–852. doi:10.1002/ece3.512

Catchen J, Hohenlohe PA, Bassham S et al (2013) Stacks: an analysis tool set for population genomics. Mol Ecol 22:3124–3140. doi:10.1111/mec.12354

Davey JW, Blaxter ML (2011) RADSeq: next generation population genetics. Brief Funct Genomics 9:416–423. doi:10.1093/bfgp/elq031

Etter PD, Bassham S, Hohenlohe PA et al (2011) SNP discovery and genotyping for evolutionary genetics using RAD sequencing. In: Orgogozo V, Rockman MV (eds) Molecular methods for evolutionary genetics. Humana Press, Totowa, pp 157–178

Frank DN (2009) BARCRAWL and BARTAB: software tools for the design and implementation of barcoded primers for highly multiplexed DNA sequencing. BMC Bioinformatics 10:362. doi:10.1186/1471-2105-10-362

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760. doi:10.1093/bioinformatics/btp324

Peterson BK, Weber JN, Kay EH et al (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One 7:e37135. doi:10.1371/journal.pone.0037135

Sambrook J, Russel D (2001) Commonly used techniques in molecular cloning. In: Sambrook J, Russel D (eds) Appendix 8, in molecular cloning, vol 3, 3rd edn. Cold Spring Harbor Laboratory Press, New York

3.3. Expérience RADseq pilote : Variations de couverture dans les librairies hautement multiplexées

3.3.1. Introduction

L'obtention de données de RADseq correspondant au multiplexage de nombreux spécimens d'espèces variées présente potentiellement certaines difficultés. Notamment, les variations de la concentration d'ADN génomique et du nombre de sites de restriction par génome pourraient avoir un impact critique sur la qualité des données obtenues. L'expérience de RADseq pilote présentée dans l'article précédent a également été réalisée dans le but de mieux comprendre le rôle de ces paramètres. Je présenterai dans cette partie les analyses réalisées dans cette perspective. Nous avons regroupé dans cette expérience des échantillons provenant de plusieurs projets indépendants et correspondant à des groupes d'Arthropodes variés. Les génomes des échantillons choisis pour cette expérience ont des tailles et des compositions variables.

Au delà du test d'une nouvelle méthode de synthèse d'adaptateurs, cette expérience pilote nous a donc permis (1) d'évaluer tester la robustesse du protocole RADseq appliqué à la réalisation de librairies RADseq hautement multiplexées et (2) de mieux comprendre les paramètres déterminant la quantité et la qualité des données obtenues pour chaque spécimen.

3.3.2. Méthode

100 individus appartenant à 11 espèces d'Arthropodes ont été inclus dans cette librairie. Le tableau 1 indique le nombre d'individus de chaque espèce ainsi que les quantités d'ADN moyennes utilisées pour les spécimens inclus dans cette librairie.

	Nombre de spécimens	Quantité d'ADN (pool final) (ng)
Leptopilina boulardi	15	2.53
Proasellus cavaticus	15	4.94
Aedes albopictus	11	4.93
Curculio elephas	2	6.40
Curculio glandium	3	4.10
*Culicidae sp**	5	0.94
*Paratrechina sp**	5	0.24
*Cnaphalocrocis poeyalis**	5	0.86
Drosophila suzukii	5	5.04
Bemisia tabaci	18	0.54
Drosophila melanogaster	16	4.625

Tableau 1. Spécimens de l'expérience RADseq pilote. (*) Spécimens du système SymbioCode.

Nous résumons ici le protocole utilisé pour la préparation la librairie, détaillé dans l'article précédent (partie 3.2.). En bref, l'enzyme Sbf1 a été utilisée pour l'étape de digestion enzymatique. Cette librairie comprend plusieurs réplicas permettant de tester, pour certains, l'utilisation d'adaptateurs optimisés et pour d'autres, la répétabilité du protocole et l'influence de la concentration d'ADN. L'ADN des mêmes spécimens ont pour cela été introduits dans la librairie associés à différents identifiants moléculaires et dans certains cas à plusieurs concentrations. Le séquençage de cette librairie RADseq sur une ligne illumina HiSeq 2000 en *single end*, 50 bp, a été effectuée par la plateforme de Génomique & Microgénomique de l'Université Lyon1 (ProfileXpert).

3.3.3. Résultats du séquençage de la librairie pilote

Les premières étapes de cette analyse ont également été décrites plus haut (partie 3.2.). Je me concentrerai ici sur la description des variations du nombre de

lectures obtenues à la fois entre espèces, entre spécimens au sein de chaque espèce et entre locus au sein de chaque spécimens.

3.3.3.1. Variabilité du nombre de lectures obtenues par spécimens

Nous avons tout d'abord voulu mesurer le nombre de lectures obtenues pour chaque spécimen, afin de mieux comprendre l'impact de l'organisme d'origine, et des concentrations d'ADN sur la variabilité de la quantité de données obtenues. La figure 8 indique la distribution des nombres de lectures obtenues pour chaque spécimen pour chaque modèle. Le nombre de lectures obtenues par spécimen varie entre 581 et 13930000 (médiane de 203100 lectures). On peut tout d'abord remarquer que des lectures ont été obtenues pour tous les spécimens. Toutefois, pour certains modèles, en particulier *Bemisia tabaci* et pour les spécimens issus du système Symbiocode, le nombre de lectures séquencées est plutôt faible pour tous les spécimens. Les médianes des nombres de lectures par individu sont respectivement 80640 et 89650 lectures. A l'inverse, pour certains modèles, un très grand nombre de lectures ont été obtenues, par exemple 12 et 13 millions ont été obtenus à partir des ADN de deux *Aedes albopictus*.

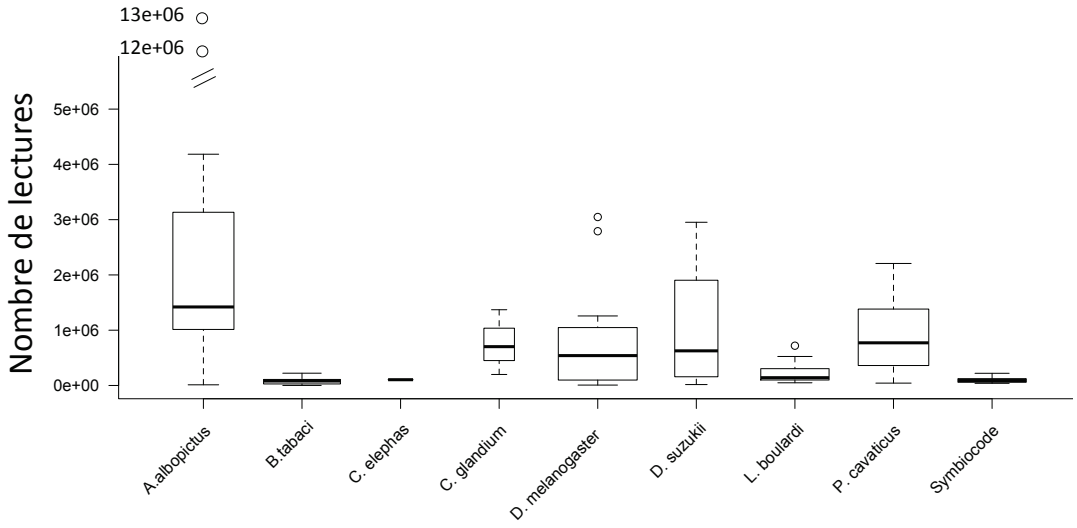

Figure 8. Distribution des nombres de lectures obtenues pour chaque spécimen de l'expérience pilote, par projet. La largeur des boites est proportionnelle au nombre de spécimens.

Par ailleurs, pour un génome donné, on pourrait s'attendre à ce que le nombre de lecture obtenu suive une relation linéaire avec la quantité d'ADN disponible. Afin de tester cette hypothèse, cinq spécimens ont été inclus dans la librairie à des concentrations différentes. La figure 9 représente le nombre de lectures obtenues en fonction du facteur de dilution. La relation obtenue est beaucoup moins forte que celle attendue pour trois spécimens, et inverse pour les deux autres.

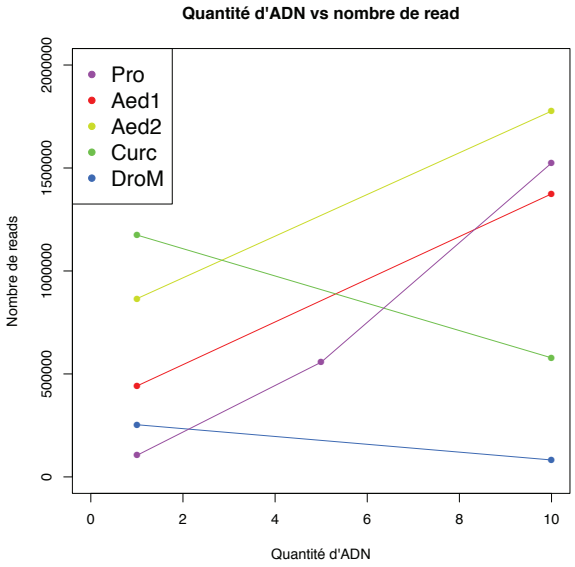

Figure 9. Relation entre le nombre de lectures et la quantité d'ADN (facteur de dilution), pour 5 spécimens inclus dans l'expérience pilotes à différentes concentrations (figure Gabriel Terraz).

Nous avons donc également mesuré la relation entre la concentration d'ADN et le nombre de lectures obtenues pour chaque individu au sein de chaque espèce. Certaines espèces ne sont représentées que par un très faible nombre de spécimens, mais parmi celles représentées par plus de 10 spécimens, une corrélation entre quantité d'ADN et le nombre de lectures obtenues ne semble exister que pour *Bemisia tabaci* (résultats non présentés). Notons que cette espèce est celle pour laquelle les concentrations sont les plus faibles. Cela suggère que la relation attendue entre quantité d'ADN et profondeur de séquençage est plus robuste pour de faibles concentrations d'ADN. Une conséquence de cette observation est que les spécimens associés à de fortes concentrations ne sont pas surreprésentés dans les données séquencées. Ainsi, la normalisation des concentrations

d'ADN semble avoir une importance moins cruciale que prévue, ce qui suggère que l'on devrait pouvoir multiplexer des ADNs de concentrations hétérogènes dans les prochaines librairies.

3.3.3.2. Couverture incomplète et hétérogénéité de la profondeur de séquençage entre locus, au sein des individus

La quantité de données obtenues varie donc fortement entre spécimens. On peut également s'interroger sur les variations du nombre de locus couverts entre individus. En effet, on peut supposer qu'au delà d'un seuil de nombre de lectures permettant de couvrir l'ensemble des locus d'un organisme avec une profondeur suffisante, les variations de nombre de lectures obtenues n'influent plus sur le nombre de marqueurs utilisables. J'ai donc mesuré la relation entre la profondeur moyenne, c'est-à-dire le nombre moyen de lectures par locus RAD, et le nombre de locus couverts par individus pour les différentes espèces (non montré). Dans cette partie, nous nous intéresserons plus particulièrement à l'analyse des lectures des 16 spécimens de *Drosophila melanogaster* inclus dans cette librairie. En effet, pour cette espèce, un génome de référence est disponible ce qui a permis d'estimer pour chaque spécimen la proportion de locus couverts parmi l'ensemble des locus présents.

J'ai identifié les locus obtenus pour chaque spécimen par un *clustering* intra-individu visant à regrouper les lectures correspondant à un même locus. Cette étape a été réalisée au moyen de la suite de programmes *pyRAD* (Eaton 2014) qui permet de réaliser ce *clustering* au moyen du programme *uclust* (Edgar 2010). *PyRAD* infère les locus présents au sein d'un individu en fonction de plusieurs critères, dont principalement, un seuil d'identité et un seuil de profondeur minimale. J'ai utilisé un seuil d'identité de 85% pour l'identification des locus de *D. melanogaster* et un seuil de profondeur minimale de 5 lectures. Le seuil de 85% d'identité utilisé est plutôt bas, ce qui permet de réunir des séquences correspondant à des allèles potentiellement très divergents d'un même locus. Par ailleurs, des études préliminaires sur des données simulées avaient montré que l'utilisation de seuils bas n'augmentait que très peu la proportion de séquences paralogues réunies à tord au sein d'un même locus.

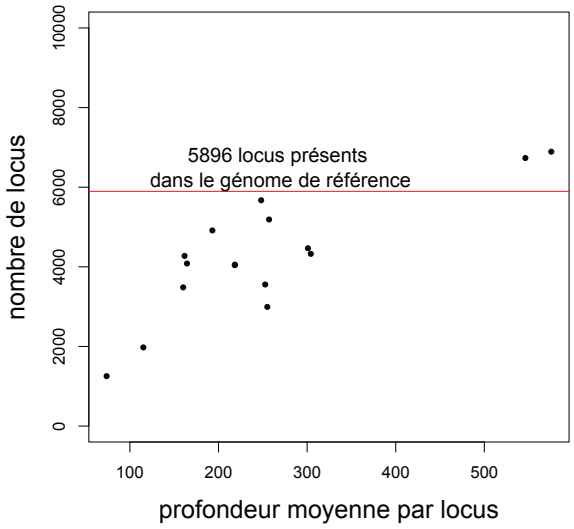

Figure 10: Relation entre la profondeur moyenne par locus et le nombre de locus couverts sur les 16 spécimens de *Drosophila melanogaster* (expérience pilote).

La figure 10 représente le nombre de locus identifiés par pyRAD, en fonction de la profondeur moyenne, c'est-à-dire du nombre de séquences moyen par locus pour chacun des 16 spécimens de *Drosophila melanogaster*. Le génome de référence de cette espèce contient 5 896 locus RAD. Toutefois, on peut remarquer que certains de ces locus se situent dans des régions répétées du génomes et génèrent des séquences paralogues généralement réunis lors de l'inférence des locus ; le clustering des locus RAD obtenu *in silico* (partie 2.2) conduit à l'inférence d'environ 4500 locus du fait de ces séquences paralogues, qui constitue donc le nombre de locus attendu dans cette expérience. Dans les échantillons les plus couverts, le nombre de locus obtenu est plus élevé (de l'ordre de 6000). Ceci pourrait s'expliquer en partie par le fait que l'hétérochromatine n'est pas intégralement représentée dans le génome de référence.

Par ailleurs, le nombre de locus couverts varie fortement entre individus, ce qui suggère qu'une forte proportion de locus n'est pas couverte chez certains d'entre eux. Pourtant, on peut remarquer que la profondeur moyenne de séquençage est généralement élevée ; une profondeur moyenne de 200X ne semble permettre de retrouver qu'environ 80% des locus. Or, si la distributions du nombre de lectures par locus suivait une loi de Poisson, en d'autre termes, sous l'hypothèse d'une distribution aléatoire des lectures entre tous les locus, une profondeur moyenne de 10X devrait

permettre de séquencer 99.95% des locus. Par ailleurs, comme le montre la figure 11, la profondeur du séquençage des locus au sein d'un même individu est beaucoup plus hétérogène qu'attendu d'après une loi de poisson. Ces résultats suggèrent que c'est cette hétérogénéité de profondeur entre locus qui explique que le nombre de locus couvert soit plus faible qu'attendu.

Nous nous sommes également demandés si cette hétérogénéité était répétable, c'est-à-dire si les même locus étaient couverts avec une profondeur importantes dans les différents individus. Dans les données obtenues, le nombre de locus séquencés en commun chez deux individus n'est pas supérieur à ce que l'on observerait si les locus étaient échantillonnés et couverts aléatoirement (non montré), ce qui suggère qu'une part importante de l'hétérogénéité de la profondeur n'est pas répétable. Ce résultat signifie qu'avec le protocole utilisé, une profondeur de séquençage anormalement élevée est nécessaire pour séquencer une grande proportion des locus RAD, suffisamment partagés entre individus.

.

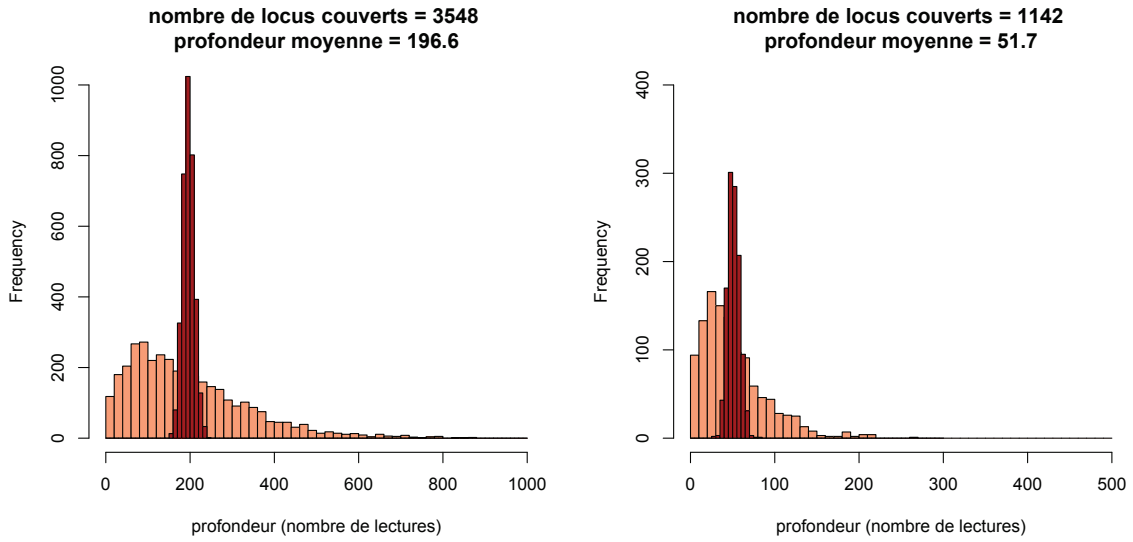

Figure 11. Distribution de la profondeur de la couverture des locus RADseq de 2 spécimens de *Drosophila melanogaster*. L'histogramme saumon représente la distribution observée des profondeurs de couvertures et l'histogramme bordeaux, celle attendue sous l'hypothèse d'une distribution aléatoire des lectures entre tous les locus (loi de poisson).

Une telle hétérogénéité dans la profondeur de séquençage des locus peut trouver son origine à l'étape de PCR qui est susceptible de provoquer une sur-amplification aléatoire de certains locus. La réalisation de plusieurs PCR indépendantes à cette étape d'amplification, dans les librairies ultérieures, devrait permettre de réduire l'hétérogénéité de la couverture entre locus. De plus, si cette hétérogénéité est bien expliquée par la présence de duplicats de PCR au sein des lectures analysées, elle pourrait également avoir des conséquences plus problématiques sur les analyses ultérieures. En effet, la présence de duplicats de PCR pourrait conduire à une surestimation du nombre de sites polymorphes dans ces données car l'identification des positions polymorphes pour chaque locus est basée sur la fréquence de chaque type de bases à chaque position. Deux nucléotides différents observés à une position donnée sur un locus peuvent correspondre à une position hétérozygote, ou à une erreur de PCR ou de séquençage. C'est la profondeur de séquençage des deux allèles potentiels qui permet de déterminer pour un taux d'erreur donné si la position correspond plus probablement au séquençage de deux allèles ou d'un allèle et d'une erreur. Or, la présence de duplicats de PCR, c'est-à-dire de séquences correspondant au séquençage multiple d'un même amplicon, peut conduire à observer des erreurs de séquençage avec une profondeur élevée pour un locus donné, et donc à inférer un faux SNP. Comme discuté plus loin, la multiplication des réactions de PCR devraient résoudre, au moins en partie ce problème. Notons par ailleurs qu'il est possible d'identifier les duplicats de PCR dans les données et donc de corriger ce problème *a posteriori*, en utilisant un séquençage en "*paired-end*". Lors d'un séquençage de ce type, des séquences sont obtenues à partir des deux extrémités de chaque fragment. Comme l'illustre la figure 12, ces deux lectures permettent de distinguer les lectures correspondant au même amplicon de PCR de celles correspondant à des amplicons différents pour un même locus.

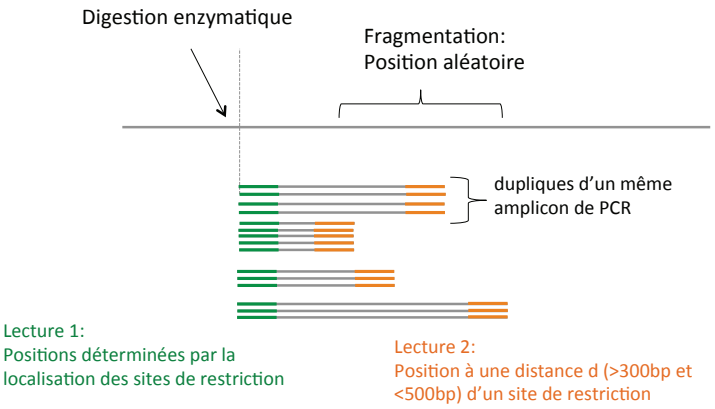

Figure 12. Localisation des lectures appariées (*paired-end*) correspondant au séquençage d'un locus RADseq.

3.3.3.3. Qualité des données : alignement des lectures de *D. melanogaster* sur le génome de référence

Nous avons ensuite voulu vérifier que les lectures obtenues correspondaient bien à des séquences du génome de *D. melanogaster*. J'ai donc aligné les séquences des lectures et des locus correspondant aux différents spécimens sur le génome de référence, par BWA (Li and Durbin 2010). La figure 13 montre la proportion de lecture alignables sur le génome de référence pour chaque spécimen. Parmi les lectures formant des locus, 90% correspondent bien à des séquences du génome de référence. Cette proportion est élevée ce qui indique que la grande majorité des lectures obtenues correspondent bien à des séquences RAD du génome des organismes étudiés. En outre, seules les lectures correspondant à des régions présentes dans le génome de référence de *D. melanogaster* peuvent être alignées. On peut donc supposer que certaines lectures que nous n'avons pas pu aligner sur le génome correspondent à de vrais locus RAD situés dans des régions non-séquencées du génome.

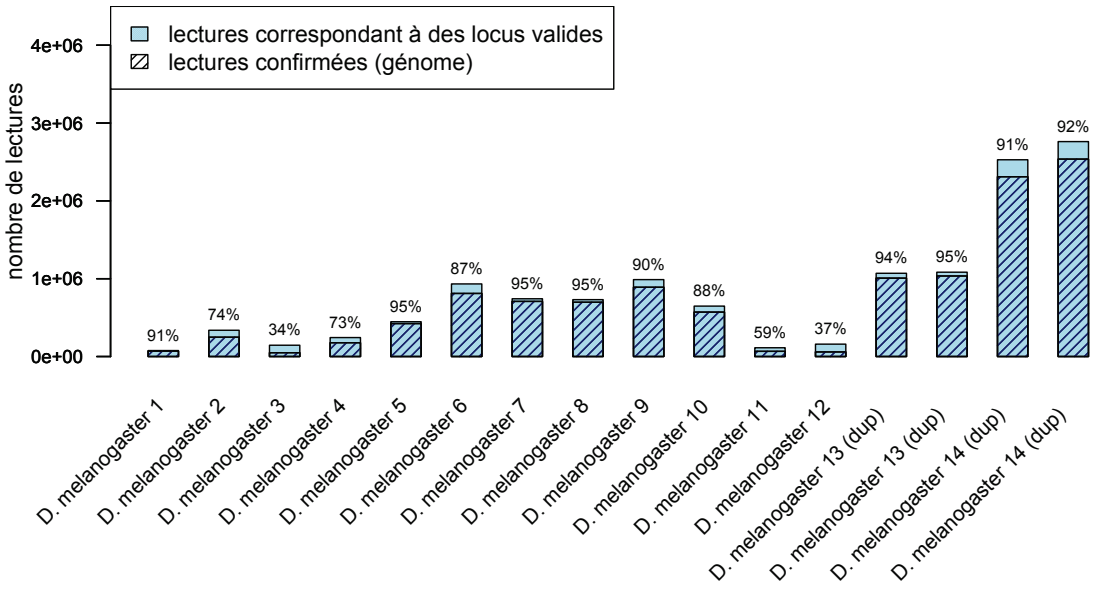

Figure 13. Nombre et proportion de lectures alignées sur le génome de *D. melanogaster* parmi les lectures formant des locus, pour les 16 *D. melanogaster* de l'expérience pilote.

Nous avons également mesuré la proportion de locus s'alignant sur le génome *D. melanogaster*. Celle-ci est plus faible (68%) que la proportion de lectures alignées. Cette différence s'explique par une plus faible profondeur de couverture des locus ne s'alignant pas sur le génome de (Figure 14). On observe ainsi des locus peu couverts ne correspondant pas à des séquences de l'ADN génomique de *D. melanogaster*. Ces locus pourraient correspondre à de l'ADN exogène effectivement présents dans l'échantillon de départ (symbiotes, bol alimentaire) ou à des contaminations expérimentales. Le fait que ces locus soient peu profondément couverts suggère qu'on pourrait en éliminer une partie en utilisant des seuils de profondeur plus élevés. Toutefois les profondeurs moyennes de ces locus sont quand même assez élevées (59 lectures en moyenne par locus), ce qui suggère que des seuils très élevées devraient être utilisés, conduisant à une perte de vraies données importante.

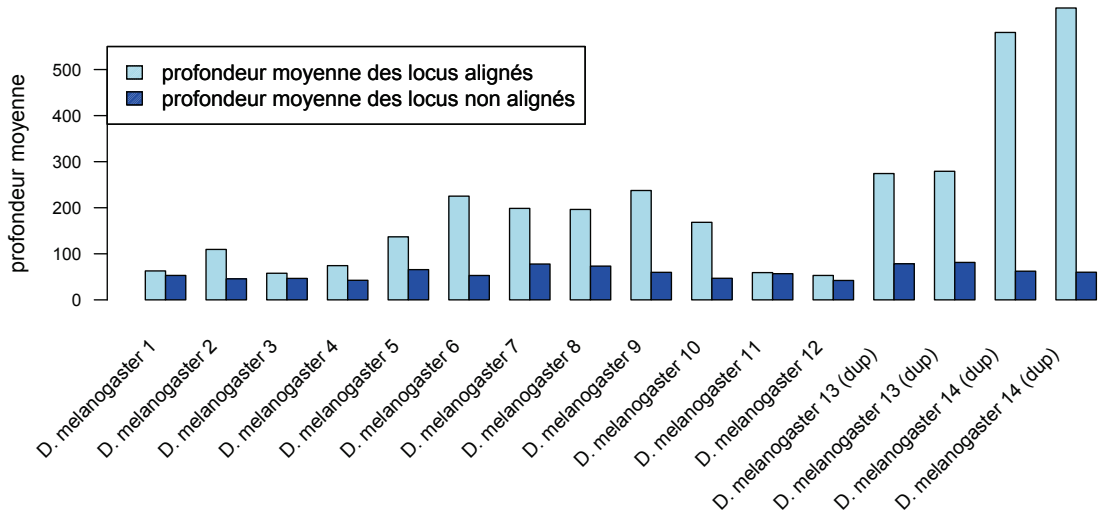

Figure 14. Profondeur moyenne (nombre de lecture moyen par locus) des locus alignés sur le génome de *D. melanogaster* ou non. Moyenne des lecture confirmées = 209 lecture par locus, moyenne des lectures non-confirmées= 59 lectures par locus.

3.3.4. Discussion : Apports de l'expérience pilote

Le séquençage de cette librairie a permis de vérifier la robustesse de cette méthode et de mettre au point un protocole de réalisation de librairies RADseq adapté à ce niveau de multiplexage. Tout d'abord, un protocole de synthèse d'adaptateurs optimisés a été développé au cours de cette étude (Henri et al. 2015). Ensuite, l'analyse du séquençage de cette librairie nous a également permis d'apporter quelques améliorations au protocole utilisé. En premier lieu, l'observation d'une faible relation entre la quantité d'ADN et le nombre de lectures obtenues incite à utiliser des ADNs de concentrations non-normalisées. De plus, une forte hétérogénéité de la profondeur de séquençage entre les locus a été observée dans cette expérience et semble expliquer la forte réduction de la couverture obtenue. Ce résultat implique qu'une profondeur moyenne par locus importante doit être obtenue pour permettre le séquençage d'une proportion de locus suffisante pour chaque individu. La réalisation de réplicats de PCR lors de la préparation de la librairie RADseq devrait permettre de résoudre partiellement ce problème. En outre, le séquençage en *paired-end* devrait permettre en

aval de supprimer ces duplicats, de manière à ce qu'ils ne perturbent pas les mesures de polymorphismes.

Cette expérience nous a également permis d'étudier différents aspects du protocole d'analyse des données RADseq. Notamment, nous avons remarqué que certains locus peu couverts, ne correspondaient pas à des séquences du génome de l'espèce étudié. La recherche d'un seuil de profondeur optimal ainsi que l'identification de l'origine et l'élimination de ces faux-locus seront donc importants pour l'analyse de ces données. Un protocole d'analyse permettant de résoudre ces questions a été développé dans le cadre de l'analyse d'une seconde librairie qui sera présentée dans les parties suivantes.

3.4. Préparation de la librairie Symbiocode

Nous avons réalisé une seconde librairie RADseq à partir des ADNs de spécimens issus de l'échantillonnage Symbiocode. Dans cette partie, je discuterai tout d'abord du nombre de locus attendu et de l'évaluation du nombre de spécimens pouvant être inclus dans une librairie de ce type. Puis, je présenterai la composition de la librairie réalisée et la méthode utilisée pour sa préparation.

3.4.1. Nombre le locus, profondeur et multiplexage

Dans le but de déterminer le nombre de spécimens pouvant être multiplexés dans une librairie RADseq, la première question qui se pose est celle du nombre de locus attendus par individu. De ce nombre de locus dépend en effet la relation entre la profondeur de séquençage moyenne par locus et le nombre de spécimens multiplexés.

$$N_{spécimens} = \frac{N_{lectures}}{ProfondeurMoyenne * N_{locusParIndividu}}$$

Le nombre de sites de coupure d'une enzyme de restriction donnée dépend tout d'abord de la taille et de la composition en bases du génome. La figure 15 représente les nombres de sites de restriction théoriques attendus par mégabase pour différentes enzymes, en

fonction de la composition en bases des séquences génomiques. Ces nombres de sites ont été calculés pour des séquences génomiques aléatoires ayant une composition en nucléotide donnée. Ils dépendent donc de la longueur et de la composition des motifs de restriction. On peut remarquer que l'utilisation d'une enzyme avec un site de 8bp (*rare cutter*) réduit d'un facteur environ 10 le nombre de site par méga base par rapport à une enzyme à site de coupure plus fréquent (*frequent cutter*). Le nombre de sites attendus dépend également du contenu en GC des séquences génomique, surtout quand le site de restriction est lui même biaisé. Ces résultats montrent que le choix de l'enzyme de restriction affecte fortement le nombre de locus RAD attendu par spécimen.

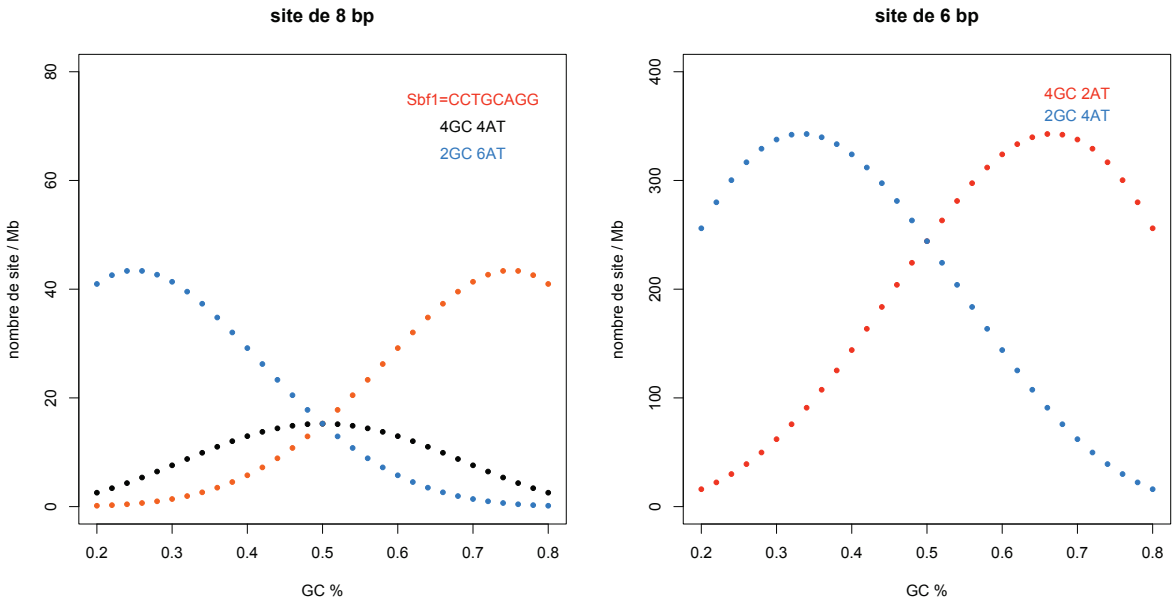

Figure 15. Nombre de sites de restriction attendus en fonction de la composition en bases GC du génome, sous l'hypothèse d'une composition en base aléatoire le long du génome, pour des sites de restriction de (A) 8 pb ou (B) 6 pb. En dehors du site Sbf1, les sites de coupures sont théoriques et visent simplement à illustrer l'effet de leur composition en base.

A l'aide de ces relations, nous avons cherché à évaluer combien de spécimens pouvaient être multiplexés dans une librairie RADseq Sbf1, séquencée par Illumina (HiSeq 2000). En supposant une taille de génome et une composition en GC relativement élevées, soit 500 Mbases et 45% de GC, le nombre de sites de restriction Sbf1 attendu par génome est d'environ 5000, ce qui correspond à 10000 locus RAD puisque chaque

site génère deux locus RAD. Si le séquençage illumina par HiSeq 2000 permet l'obtention de 100 millions de lectures utilisables, et que l'on souhaite obtenir une profondeur moyenne de 10 lectures par locus, on devrait pouvoir multiplexer en moyenne 1000 spécimens par librairie.

$$N_{spécimens} = \frac{100 * 10^6\ (lectures)}{10\ (lectures\ par\ locus) * 10000\ (locus\ par\ individu)} = 1000$$

Afin d'estimer dans quelle mesure cette relation simple, supposant une composition en base homogène le long du génome, permettait de prédire réellement la densité en site de restriction, nous avons calculé, en collaboration avec Tristan Lefébure (LEHNA), le nombre de sites présents dans 33 génomes d'arthropodes disponibles au moment de cette étude. Les tailles des génomes varient ici de 100 Mb (*Caenorhabditis elegans*) à 9 Gb (*Anopheles gambiae*), avec une moyenne de 606 Mb et une médiane de 191 Mb. Notons que cette distribution est sans doute biaisée vers les génomes de petites tailles, qui sont préférentiellement ciblés pour les projet de séquençage de génomes complets. Leurs contenus en bases varient quant à lui entre 27 et 45 % de GC (moyenne 39%). La figure 16 représente ainsi le nombre de sites de restriction présents dans ces génomes en fonction de leur composition. On peut remarquer que ces valeurs peuvent être très différents de celles attendues. En effet, l'écart entre valeurs observées et attendues peut correspondre à une fraction importante du nombre de sites. Par exemple, pour des génomes correspondant à des compositions en GC comprises entre 35 et 40%, l'écart entre le nombre de sites EcoR1 attendus et observés peut atteindre 300 sites par Mbases, ce qui est supérieur au double de la densité en sites attendue. De même on observe un écart de l'ordre de 30% entre le nombre de sites Sbf1 attendus et observés pour des génomes associés à des compositions en bases comprises entre 35 et 40%. Cette différence est probablement liée à l'hétérogénéité de la composition en base le long des génomes. Certaines régions peuvent en outre être enrichies en di-nucléotides, ou en certains motifs présents dans un site de restriction donnée. Ces régions seront donc enrichies en sites de coupures pour cet enzyme. Par exemple, McCluskey and Postlethwait (2014) ont observé dans le génome du poisson zèbre que le site de restriction Sbf1 était surreprésenté au sein des gènes de la famille des récepteurs *NOD-like*, ainsi qu'au sites accepteurs d'épissage. Ces enrichissement ou

appauvrissement locaux peuvent contribuer à expliquer l'écart entre les nombres de locus RAD attendus et observés.

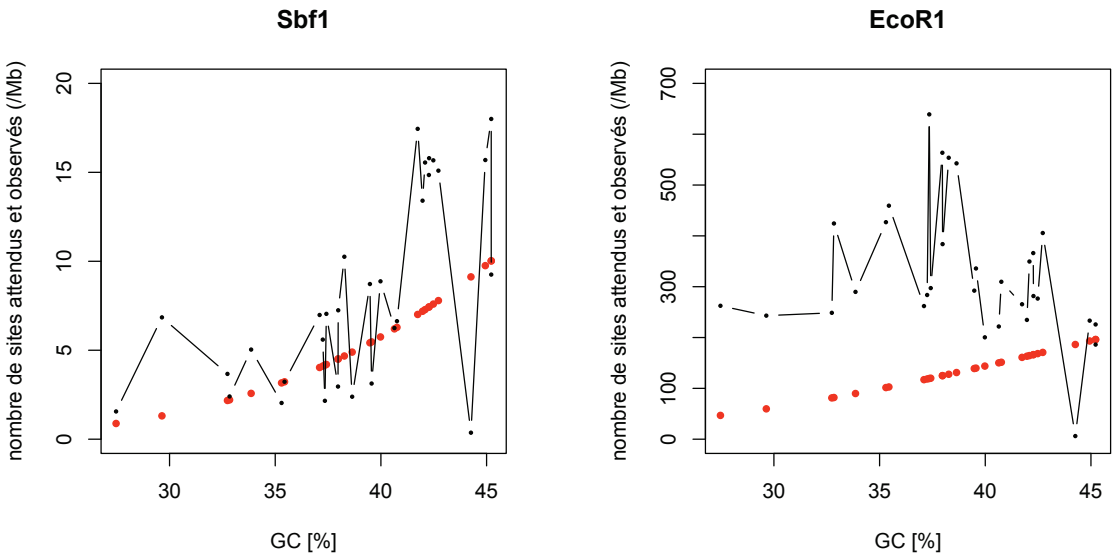

Figure 16. Nombres de site (A) Sbf1 et (B) EcoR1 attendus et observés par mégabase dans 31 génomes complets.

Ces résultats montrent qu'il peut-être est assez difficile de prédire le nombre de locus attendu dans une librairie RADseq, et ce d'autant plus que les caractéristiques des génomes des organismes étudiés sont variées et inconnues. Par ailleurs, on peut également s'attendre à ce que les profondeurs de séquençage varient entre spécimens. Notons que si le nombre de copies du génome est le même pour chaque spécimen au début de la préparation de la librairie, il ne devrait pas y avoir de différences de profondeur moyenne entre les locus des différents individus, quel que soit le nombre de locus par individu (en d'autres termes, un génomes 10 fois plus gros devraient fournir 10 fois plus de lectures, toutes choses égales par ailleurs). En revanche, une quantité d'ADN donnée peut correspondre à un nombre de « copies » des génomes différent selon la taille de ces derniers. Si la quantité d'ADN est la même pour chaque spécimen, on s'attend donc à ce que la profondeur de séquençage moyenne des locus soit inversement proportionnelle à leur taille. Enfin, si la quantité d'ADN disponible pour chaque individu varie, on s'attend également à ce que la profondeur moyenne des locus dépende de la quantité d'ADN disponible.

Afin de tenir compte des variations de couvertures et de profondeur entre individus et potentiellement entre locus, nous avons choisi de multiplexer 200 spécimens au sein de la librairie Symbiocode. Si les génomes inclus dans cette librairie ont en moyenne 10000 locus chacun, la profondeur moyenne visée par locus est ainsi de 50X.

3.4.2. Description de la librairie Symbiocode, modifications apportées par rapport à la librairie pilote

Une librairie RADseq a été construite, regroupant 180 individus représentant 90 espèces au sein des échantillons SymbioCode: 61 espèces de Diptères et 29 espèces de Lépidoptères. 10 spécimens ont en outre été répliqués, de façon à vérifier la répétabilité de l'expérience. Un spécimen de *Drosophila melanogaster* a également été introduit sous forme de deux réplicats, c'est-à-dire identifiés par deux barcodes différents, afin de constituer un contrôle de la qualité de cette librairie.

Pour la réalisation de cette nouvelle librairie, nous avons utilisé le protocole mis au point pour l'expérience pilote, avec quelques modifications notables. Tout d'abord, des adaptateurs comportant des barcodes de longueurs différentes ont été utilisés. Une librairie précédente (Emilie Delava, communication personnelle) avait montré que la présence de la même base pour toutes les lectures à certaines positions dans les séquences RADseq peut être à l'origine d'erreurs de séquençage à ces positions. En effet, le séquençage illumina implique la fixation et l'amplification des fragments d'ADN qui correspondant aux différentes lectures sur une cellule. L'identification des coordonnées de chaque lecture est ensuite permise par la détection des fluorescences émises par l'intégration des nucléotides lors du séquençage des premières bases de chaque lecture. Les quatre nucléotides émettent des fluorescences de longueurs d'ondes différentes, ce qui permet de distinguer les positions des différentes lectures proches les unes des autres. Si, à certaines positions lors de cette étape, la même fluorescence est émise par une grande proportion des séquences sur la cellule, cette identification des « positions de lecture » peut être compromise. Or, dans des séquences RADseq, toutes les lectures débutent par les séquences des barcodes, suivies de 6 bases correspondant aux sites de restriction. La présence de ces sites de restriction conduit donc au séquençage des

mêmes bases pour toutes les lectures entre les positions +9 et +15. Afin de réduire cette homogénéité, il est possible d'utiliser des barcodes de tailles différentes afin de décaler la position du site de restriction dans une certaine proportion des lectures. Des adaptateurs synthétisés par la méthode de Henri et al. (2015), correspondant à des barcode (MID) de 8 bp et 10bp (25%) ont donc été utilisés.

Nous avons également réalisé des purifications supplémentaires par rapport à la librairie pilote afin de diminuer le nombre de dimères P1-P2, qui avaient été retrouvés en grande quantité parmi les lectures non-utilisables de l'expérience pilote (Henri et al. 2015). Ces purifications ont permis de réduire la proportion de dimères des deux types d'adaptateurs dans les données séquencées qui était de 35% dans la librairie pilote à 2% dans cette librairie. L'étape d'amplification des fragments ligués aux deux types d'adaptateur a en outre été réalisée au moyen de 5 PCRs indépendantes, de manière à réduire les biais d'amplification aléatoires introduits à cette étape.

Cette librairie a ensuite été séquencée sur une ligne illumina HiSeq 2000 en *paired end*, de 100 bp, par la plateforme de Génomique & Microgénomique de l'Université Lyon1 (ProfileXpert), au moyen de deux réactions de séquençage.

3.5. Librairie Symbiocode : mise en place d'un protocole d'analyse des données RADseq

Dans cette partie, je présenterai le protocole mis en place pour l'analyse de ces données. Nous avons particulièrement mis l'accent dans cette étude sur la compréhension des artefacts potentiellement à l'origine de séquences ne correspondant pas à des RAD tags de l'organisme étudié au sein des données. Nous décrirons donc la méthode d'analyse développée dans ce cadre, ainsi que les résultats obtenus. Par ailleurs, afin d'évaluer la qualité des données obtenues, nous avons également cherché à estimer le nombre de locus présents au sein des génomes de chaque espèces, afin de pouvoir mesurer pour chaque spécimen, la proportion de locus couverts parmi l'ensemble des locus potentiels.

3.5.1. Méthodes

**Etapes classiques*

La figure 17 décrit le protocole utilisé pour l'analyse de ces données. Le démultiplexage des lectures a été réalisé par le programme *process_radtag* de *Stacks* (Catchen et al. 2011, 2013), en autorisant un mésappariement au niveau de l'ensemble des séquences des barcodes et des sites de restriction. L'identification des locus a quant à elle été réalisée par *pyRAD* (Eaton 2014), au moyen de paramètres de *clustering* correspondant à 85% d'identité minimale et à une profondeur minimale de 2 ou 5 lectures par locus, selon les analyses. Au moment de cette étude, les deux suites de programme les plus communément utilisées pour analyser des données de RADseq étaient *Stacks* et *pyRAD*. Ces deux programmes diffèrent notamment pour la méthode utilisée à cette étape d'identification des locus au sein des individus et des locus orthologues entre individus. Tandis que pyRAD utilise un *clustering* par *uclust* (Edgar 2010) qui regroupe les lectures correspondant à un même locus en fonction d'un seuil de similarité, *Stacks*, utilise un nombre maximum de différences autorisées entre séquences au sein d'un locus. Ainsi, l'inférence des locus par *Stacks* ne permet pas de regrouper au sein d'un même locus des séquences correspondant à deux allèles différenciées par des insertions ou des délétions. Nous avons donc choisi d'utiliser l'approche *pyRAD*, même si les résultats de Eaton (2014), indiquent que cette différence de méthode est plus importante pour l'étape d'identification des locus orthologues que pour l'étape d'identification des locus au sein des individus.

** suppression des duplicats de PCR*

L'analyse des résultats de l'expérience pilote avait suggéré que pour un locus donné, certaines lectures obtenues correspondent au séquençage de fragments d'ADN différents, tandis que d'autres correspondent au séquençage de plusieurs versions du même fragment, amplifié par PCR. La présence de ces duplicats peut introduire des biais dans les analyses ultérieures. C'est pourquoi nous avons voulu les identifier, ce qui est rendu possible par le séquençage des librairies RAD en *paired-end* (Figure 12). Dans le protocole de RADseq utilisé, la lecture correspondant à l'extrémité opposée au site de restriction, appelée « lecture 2 », est issue d'une fragmentation aléatoire, la position de cette lecture est donc théoriquement différente pour chaque fragment, avant l'étape d'amplification. L'identification des duplicats de PCR a été réalisé au moyen d'une

version modifiée du programme filterPCRdupl.pl, qui permet l'identification de "lectures 2" comportant le même point de départ (ConDeTri, Smeds and Künstner 2011). Ce programme repose sur l'identification d'ensembles de couples de lecture 1 et 2, correspondant aux extrémités d'un même fragment. Ce filtre est réalisé sur les 10 premières bases des lectures 1 d'une part et des lectures 2 d'autre part. Parmi les séquences regroupées, le programme sélectionne ensuite celle de meilleure qualité. L'identification des couples sur les 10 premières paires de bases permet de limiter l'impact des erreurs de séquençage sur ce filtre. En effet, si une lecture comporte une erreur entre la position 11 et la fin de la lecture, elle sera quand même identifiée comme duplicat. Pour diminuer encore l'impact des erreurs, nous avons ajouté une deuxième étape à cette méthode, au cours de laquelle des séquences issues de la première étape sont également filtrées par la recherche de couples identiques entre les positions 11 et 20. Ainsi, pour que deux couples correspondant au séquençage d'un même fragment ne soient pas identifiées, il faut supposer que ces séquences comportent une erreur entre les positions 1 et 10 et une autre entre les positions 11 et 20, ce qui doit être rare.

La recherche et l'élimination des duplicats de PCR parmi les lectures de la librairie Symbiocode a résulté en une réduction d'un facteur 10 du nombre de lectures. Ce résultat indique que chaque amplicon a été séquencé en moyenne 10 fois. Il est difficile de comparer ce niveau de duplication à celui obtenu lors de l'expérience pilote, mais ce résultat suggère que le problème d'hétérogénéité de la profondeur de séquençage des locus est toujours important dans cette librairie.

** filtres des faux locus*

L'analyse des résultats de l'expérience pilote, en particulier des données issues des spécimens de *Drosophila melanogaster*, avait mis en évidence l'existence de locus peu couverts ne correspondant pas à des séquences du génome nucléaire de l'organisme étudié. En outre, l'observation des séquences obtenues à partir de la librairie Symbiocode a permis d'identifier certaines constructions artéfactuelles. Notamment, une proportion significative des lectures contient une partie de la séquence des adaptateurs ou des constructions correspondant vraisemblablement à la ligation en tandem de certains barcodes. Nous avons donc choisi de filtrer ces locus contenant des séquences des adaptateurs ou les motifs correspondant au site de restriction ou au barcode caractérisant chaque individu. Par ailleurs, certaines lectures peuvent

correspondre au séquençage de locus RAD de différents organismes contaminants. Nous avons également voulu filtrer ces locus dans la mesure du possible. Pour cela, nous avons construit une banque de données contenant les séquences de tous les génomes complets de bactéries, d'archées, de virus et d'Eucaryotes unicellulaires contenus dans *Ensembl Genome* en octobre 2014. Nous avons ainsi pu comparer les lectures obtenues aux séquences de cette banque, au moyen d'une étape de BLASTs afin d'identifier d'éventuelles séquences exogènes homologues de génomes connus.

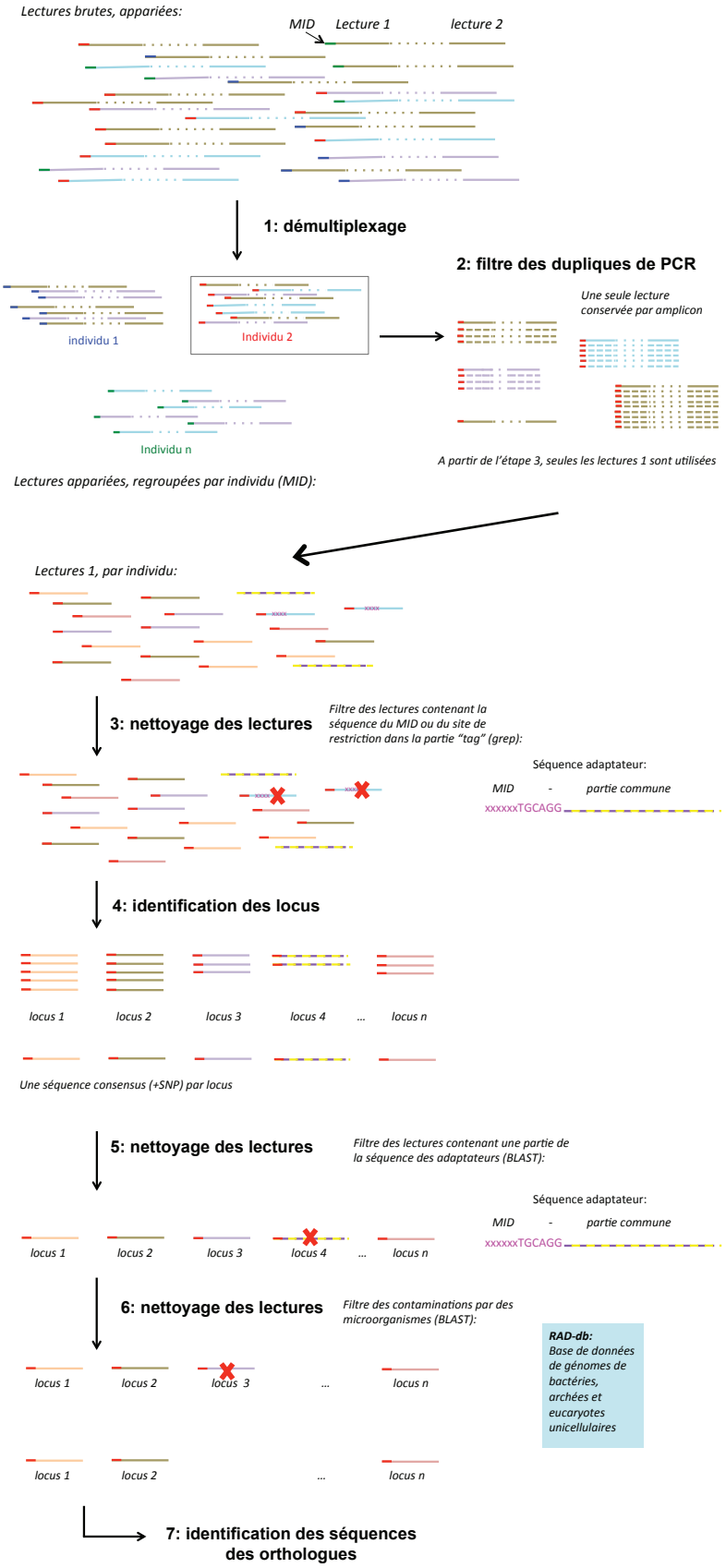

Figure 17. Schéma du pipeline d'analyse des données RADseq

3.5.2. Analyse des lectures de la librairie Symbiocode

Le séquençage par 2 réactions (*runs*) de cette librairie a permis d'obtenir 262 millions de lectures. Après démultiplexage, ces données correspondent à 119 millions de lectures de bonne qualité associées à un barcode et à un site de restriction correct à un mésappariement près. La proportion de lectures de bonnes qualités obtenues est donc similaire à celle correspondant aux lectures de l'expérience pilote (45%). Toutefois, la quantité de lectures obtenue pour chaque spécimen est beaucoup plus variable. Notamment, très peu de lectures ont été obtenues correspondant aux spécimens associés à des barcodes de 10 paires de base. Les 48 spécimens associés à des barcodes de 10 paires de bases sont représentés par un nombre de lectures compris entre 0 et 2575 (moyenne de 297 lectures, médiane de 67 lectures). Ce nombre de lecture très faible compromet fortement l'utilisation des données associées à ces spécimens pour la suite des analyses. La cause de cette efficacité anormalement faible des adaptateurs associés à des barcodes de 10 paires de bases est inconnue. On peut remarquer que ces mêmes adaptateurs ont été utilisés pour la préparation d'une librairie ultérieure, au sein de laquelle ils ont permis d'obtenir un nombre de séquences similaires aux adaptateurs de 8 paires de bases (communication personnelle Gabriel Terraz). En revanche, le séquençage des 144 spécimens associés à des barcodes de 8 paires de bases a permis d'obtenir des quantité de lectures beaucoup plus élevée, de 720400 lectures par individus en moyenne (compris entre 7279 et 10050000).

Nous avons réalisé les étapes d'analyses suivantes sur les lectures des spécimens associés à des barcodes de 8 paires de bases. Le tableau 2 indique les nombres de lectures et de locus moyens par spécimens obtenus après chaque étape.

étape	1er quartile	médiane	3eme quartile
lectures			
1 : démultiplexage	89582	356003	921266
2 : filtre des duplicats	10786	37391	89857
% perte	*entre 85 et 94% (médiane 90%)*		
3 : élimination des adaptateurs (partie barcode + site de restriction)	9687	36516	73785
% perte	*entre 1 et 59% (médiane 3%)*		
locus			
4 : identification des locus	183	1286	3676
5 : élimination des adaptateurs (partie commune)	183	1286	3676
% perte	*entre 0 et 5% (médiane 0%)*		
6 : nettoyage des micro-organismes	159	1192	3605
% perte	*Entre 0 et 73% (médiane 2%)*		

Tableau 2. Résultats du séquençage de la librairie RADseq Symbiocode. Les locus ont été identifiés par pyRAD, les paramètres de clustering correspondent à un seuil de similarité minimale de 85% et à une profondeur minimale de 5 lectures par locus.

Nous avons également voulu mesurer la qualité de ces données par l'alignement des lectures obtenues pour les deux spécimens de drosophiles inclus dans la librairie, sur le génome de *D. melanogaster*. La figure 18 montre le résultat de ces alignements. On peut observer, tout d'abord, que les proportions de locus confirmés par alignement sur le génome pour ces deux spécimens sont élevées (90 et 92%). De plus, pour les deux spécimens, l'application des différents filtres conduit à une augmentation de la proportion de lectures alignables sur le génome de *D. melanogaster*, ce qui suggère que les filtres utilisés permettent bien une augmentation de la qualité des données utilisées. Notons que leur application conduit également à une diminution du nombre total de lectures alignées, ce qui suggère que certaines lectures RAD « vraies » sont éliminées à tord. Toutefois, ce nombre de faux négatifs est faible. Il est donc probable que leur perte est moins problématique que l'utilisation de lectures contaminantes.

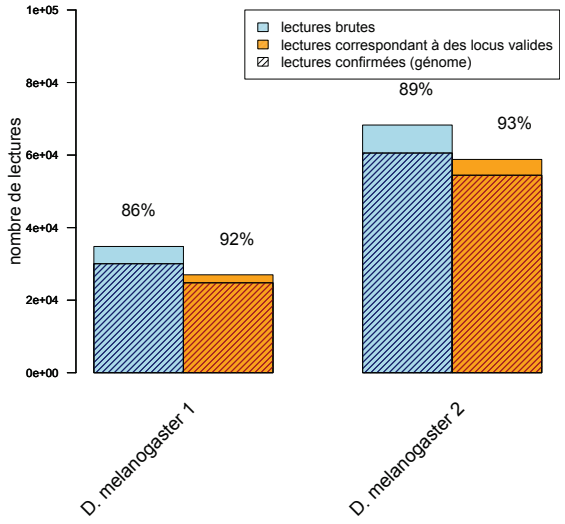

Figure 18. Nombre de lectures et proportion de lectures confirmées par l'alignement sur le génome avant et après les différents filtres (locus validés), pour les deux spécimens de *Drosophila melanogaster*.

Dans cette librairie, 90 espèces étaient représentées, chacune par l'échantillonnage de deux individus. En tenant compte des 48 spécimens associés à des barcodes de 10 paires de bases, non séquencés, j'ai pu étudier le partage de locus orthologues au sein de 51 paires complètes. Pour chacune de ces paires, la recherche des locus orthologues partagés a été réalisée au moyen d'un *clustering* par *SiLiX* des séquences consensus des locus identifiés pour chaque spécimen (Miele et al. 2011). Les 51 paires étudiées partagent en moyenne 946 locus couverts par au moins 5 lectures (valeur médiane: 279). De plus, 29 paires d'individus partagent plus de 100 locus orthologues. En d'autres termes, nous sommes en mesure d'estimer de façon robuste des distances nucléaires pour 60% des espèces considérées.

3.5.3. Evaluation de la proportion de locus couverts pour chaque individu

Nous nous sommes également demandés quelle proportion des locus RAD existants était couverte dans les données obtenues pour chaque spécimen. Cette estimation implique la connaissance du nombre de locus RAD total existant au sein des génomes de chaque espèce. Nous avons estimé ce nombre par deux méthodes différentes, l'une utilisant les proportions de locus partagés au sein de chaque paire de

spécimens, et l'autre, la forme de la distribution des profondeurs de séquençage des locus pour chaque spécimens.

**méthode 1 : Estimation du nombre de locus RAD au sein d'un génome en fonction des nombres de locus couverts chez deux individus*

Pour chaque paire, on peut en principe estimer le nombre total de locus RAD présents dans le génome de l'espèce, à partir de la relation reliant le nombre de locus total, le nombre de locus échantillonnés pour chaque individu et le nombre de locus échantillonnés dans les deux individus. Notons L_1 et L_2 les nombres de locus observés respectivement chez l'individu 1 et 2, et L, le nombre de locus total, que l'on suppose être le même dans les génomes des deux individus provenant de la même espèce. On appelle également p_1 et p_2, les probabilités d'avoir observé un locus chez les individus 1 et 2, et F_1 et F_2 le nombre de "faux locus", c'est-à-dire de séquences identifiées comme locus mais ne correspondant pas à une région RAD du génome étudié. Le nombre de locus observés chez un individu correspond à la somme du nombre de locus « vrais », $L*p_i$, et du nombre de locus « faux », F_i

$$L_1 = L * p_1 + F_1$$
$$L_2 = L * p_2 + F_2$$

Si l'on suppose que tous les locus observés proviennent bien du génome de l'organisme étudié, F_1 et F_2 sont nuls. Si, de plus, l'échantillonnage des locus est indépendant pour chacun des individus, L_{12}, le nombre de locus obtenu chez les deux individus vaut :

$$L_{12} = L * p_1 * p_2$$
$$\text{d'ou } L = \frac{L_1 * L_2}{L_{12}}$$

Nous avons utilisé cette relation pour évaluer le nombre de locus total pour chaque espèce, en utilisant les nombres de locus inférés pour des profondeurs minimales de 2 lectures ou de 5 lectures. Pour les deux drosophiles témoins, les nombres de locus total inférés par cette méthode sont de 3600 locus (seuil de profondeur 5 lectures), et 4000 locus (seuil de profondeur de 2 lectures). Le nombre total de locus attendu est d'environ 4500, compte tenu des locus paralogues observé

dans le génome de référence (Cariou et al. 2013), ce qui suggère que le nombre total de locus est légèrement sous-estimée par cette méthode.

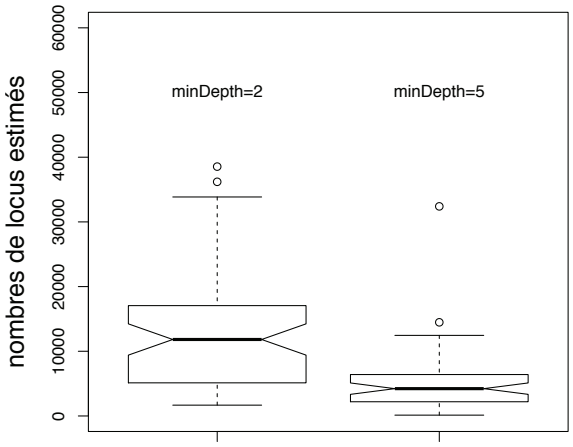

Figure 19. Distribution du nombre de locus total pour chaque paire, estimé à partir du nombre de locus séquencés dans les 2 individus (méthode 1), pour différents seuils de profondeur minimale utilisés par pyRAD (*minDepth*).

La figure 19 représente les distributions des nombres de locus total au sein des génomes de chaque espèce, estimés par la méthode décrite précédemment. On peut remarquer que les nombres de locus estimés sont plus élevés lorsqu'un seuil de profondeur minimal de 2 lectures est utilisé que quand ce seuil est de 5 lectures. Le nombre de lectures estimé médian est en effet supérieur à 10000 locus dans le premier cas et inférieur à 5000 locus dans le second.

Ces estimations reposent sur deux hypothèses qui pourraient conduire à une surestimation ou a une sous-estimation du nombre de locus total. Tout d'abord, nous avons supposé que les filtres effectués avaient permis de supprimer les locus ne correspondant pas à des RAD tags attendus, c'est à dire que le nombre de « faux » locus est proche de 0. Si ce nombre est en fait plus élevé et si, d'autres part, ces locus ne sont pas partagés entre les individu d'une paire, cela pourrait conduire à surestimer le nombre de locus total. Si, de plus, la proportion de locus faux augmente quand le seuil de profondeur minimal pour inférer un locus diminue (si les faux locus sont en moyenne moins couverts), cela pourrait expliquer que le nombre de locus total estimé soit plus

élevé avec un seuil de profondeur de 2 lectures. Le plus grand nombre de faux positifs quand on utilise un seuil de profondeur minimum de 2 lectures conduirait à une surestimation plus importante du nombre de locus total. D'autre part, ces estimations supposent également que l'échantillonnage des locus est indépendant dans les deux spécimens. Contrairement à cette attente, les locus montrant une grande profondeur de séquençage ont peut-être une probabilité supérieure d'être couverts. Ces locus seraient dans ce cas plus souvent couverts dans les deux spécimens que ce que l'on attendrait par hasard, ce qui pourrait conduire à une sous-estimation du nombre de locus total par cette méthode.

Ces deux estimations correspondent par ailleurs à des proportions de locus couvert médians (rapport du nombre de locus couvert sur le nombre de locus total) d'environ 20%, pour les locus couverts au moins par 5 locus et un peu supérieures à 45% pour les locus couverts par au moins 2 lectures (Figure 20).

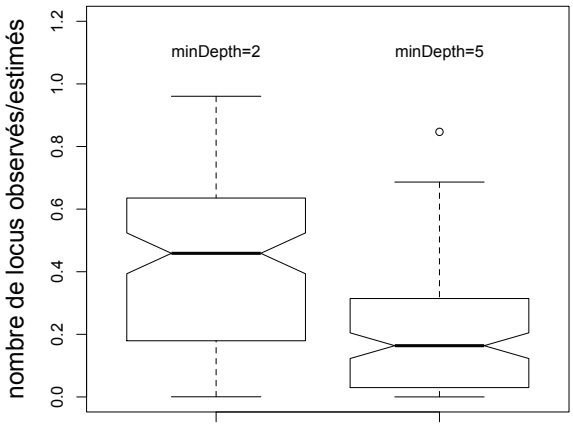

Figure 20. Distribution de la proportion de locus obtenus par rapport au nombre de locus total estimé pour chaque individu. Ces résultats correspondent à un seuil de profondeur minimale utilisé par pyRAD (*minDepth*) de 2 ou 5 lectures.

**méthode 2 : Estimation du nombre de locus RAD au sein d'un génome à partir de la profondeur de séquençage des locus couverts*

Le nombre de locus total attendu pour un individu peut également être estimé à partir de la distribution des profondeurs de séquençage par locus. On utilise pour cela une approche bayésienne approximée (voir présentation du principe de l'ABC *Approximate Bayesian Computation*, partie 4.2). J'ai ainsi réalisé des simulations

correspondant à l'échantillonnage de N lectures parmi L locus par des tirages aléatoires avec remise (script R). Pour chaque simulation, j'ai ensuite calculé les statistiques descriptives suivantes : le nombre de locus couverts (l_{obs}), la moyenne (m_{depth}), et la médiane (med_{depth}) des profondeurs de séquençage par locus et le nombre de lectures correspondant à des locus couverts par au moins 2 lectures (r_{obs}). 1,3 millions de simulations ont été réalisées pour un nombre de locus variant entre 500 et 100000 et pour des nombres de lectures correspondant à une profondeur moyenne variant entre 0.1 et 100, échantillonnés dans une distribution *a priori* uniforme sur le logarithme décimal.

Des estimations par ABC m'ont ensuite permis, pour les valeurs de statistiques calculées pour chaque individu, d'identifier les simulations correspondant aux statistiques les plus proches, et donc le nombre de locus, L, et le nombre de lectures totales, N, expliquant le mieux les données observées sous le modèle utilisé. L'efficacité de cette approche a été validée au moyen des méthodes classiques d'évaluation des procédures ABC. Ainsi, le test de validation croisée permet de vérifier que la méthode employée permet bien de retrouver les valeurs de paramètres correspondant aux statistiques descriptives de données simulées (dites pseudo-observée). Cette méthode sera détaillée plus longuement au cours de la partie 4.2, mais les résultats (non présentés) obtenus pour cette étude montrent que les nombres de locus et de lectures sont correctement estimés pour des données correspondant à des profondeurs de séquençage supérieures à une lecture en moyenne par locus. Un test de *goodness-of-fit* permet par ailleurs de vérifier l'adéquation du modèle aux données réelles. La statistique test de cette méthode est la médiane des distances entre statistiques descriptives acceptées et observées. La distribution nulle de cette statistique test est obtenue en calculant la même médiane pour des statistiques descriptives pseudo-observées. Ainsi, cette méthode permet de détecter des cas où les différentes statistiques descriptives ne seraient pas corrélées de la même manière aux valeurs de paramètre dans les données réelles et dans les données simulées. Cette différence peut conduire à observer une distance médiane entre statistiques observées et acceptées significativement supérieure à celle mesurée pour les statistiques « pseudo-observée », compatibles, par définition avec les variations des statistiques simulées, et donc à détecter des cas d'inadéquation entre le modèle utilisé et les données observées. Cette méthode n'a permis de mettre en évidence aucun cas d'inadéquation entre les données

étudiées et le modèle employé, ce qui suggère que cette procédure permet généralement une bonne estimation des valeurs de paramètres.

Par ailleurs, la précision de chaque estimation a également été estimée par l'étude de la forme des distributions postérieures des valeurs de L estimées. Nous avons pour cela mesuré l'intervalle entre le premier et le 3eme quartile de la distribution postérieure. Afin d'éliminer les estimations les moins précises, nous ne considérons ici que les résultats obtenus pour des distributions postérieures telles que 50% des valeurs encadrant la médiane ne s'écartent pas de plus ou moins de 5% de la valeur médiane.

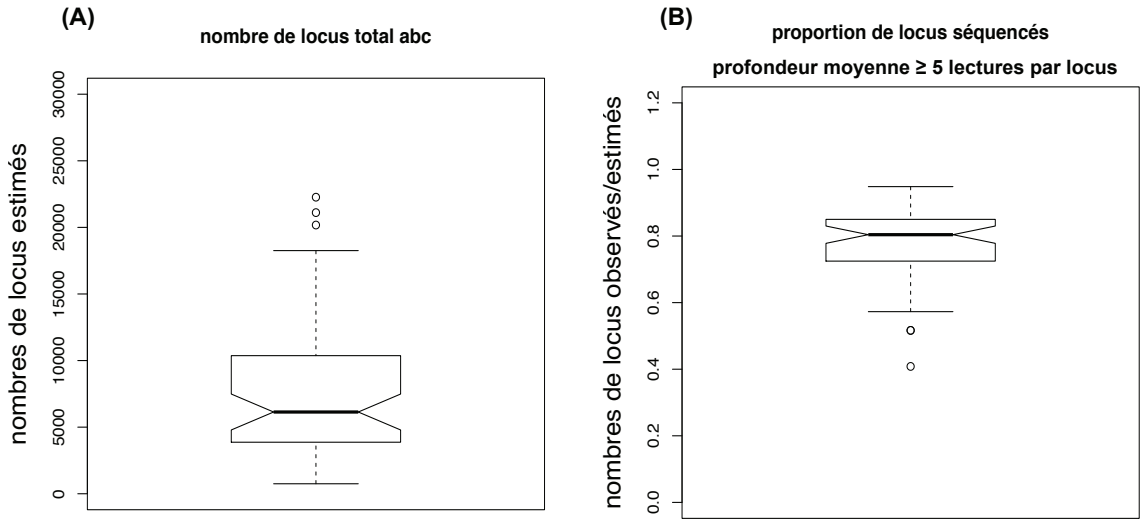

Figure 21. A. Distribution du nombre de locus total pour chaque paire, estimé en fonction de la distribution de la profondeur moyenne de séquençage des locus (méthode 2). B. Distribution de la proportion de locus obtenus par rapport au nombre de locus total estimé pour chaque individu. Les individus représentés sont ceux pour lesquels 50% des valeurs postérieures (intervalle entre le 1er et 3eme quartile) sont entre -5% et +5% de la valeur médiane (93/111) et correspondant à une profondeur moyenne par locus supérieure ou égale à 5 (47/111).

La figure 21 montre la distribution du nombre de locus total inférés par cette méthode. Les individus associés à des profondeurs moyennes inférieures à 5 lectures par locus ont été exclus de cette distribution. En effet, pour cinq de ces individus, le nombre de locus total inféré est inférieur au nombre de locus observé, ce qui suggère que la qualité de l'estimation pourrait diminuer avec la profondeur moyenne. Par ailleurs, notons que les estimations n'ont été réalisées que pour un seuil de profondeur

pour l'inférence des locus de 2 lectures, car les simulations ont été effectuées avec ce paramètre. Toutefois, la présence d'un nombre important de faux locus affecte peut-être également cette méthode d'estimation. En effet, il semble que ces faux locus correspondent à des distributions de profondeurs différentes des locus génomiques. On peut donc supposer que la présence de ces locus cause un écart entre les distributions de profondeurs observées et les simulations réalisées, qui modélisent un cas où tous les locus ont la même probabilité d'être couverts. Afin, d'approfondir cette question, il sera utile de réaliser des simulations correspondant à des seuils de profondeur minimale par locus plus élevés. Notons que la proportion de locus couverts, est ici de 80% en moyenne, car seuls sont considérés les spécimens pour lesquels l'estimation par ABC est suffisamment fiable, qui sont aussi les spécimens les mieux couverts. Sur les mêmes individus, la proportion médiane de locus couverts estimée par la première méthode est de 55%.

** Corrélation entre les 2 estimations*

La figure 22 montre la corrélation entre les nombres de locus estimés selon les deux méthodes. Pour la première méthode, j'ai utilisé les estimations correspondant à un seuil de profondeur minimal de 5 lectures, qui sont probablement moins affectées par la présence de « faux-locus » peu couverts. La corrélation observée est significative, ce qui suggère que ces deux méthodes permettent d'évaluer le nombre de locus présent dans un génome (test de corrélation de Pearson, p-value < 6.10^{-6}).

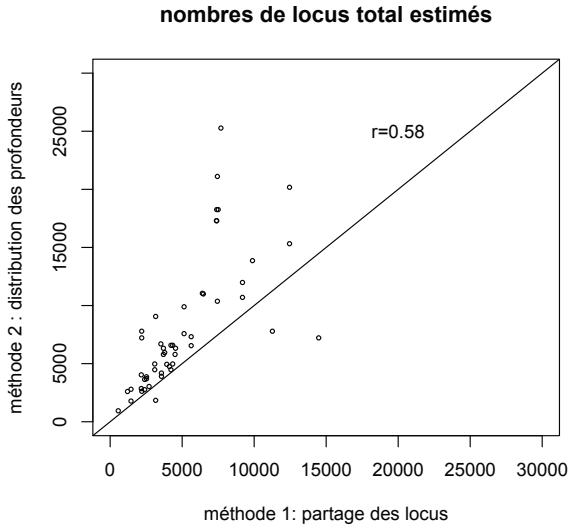

Figure 22. Corrélation entre les nombres de locus total estimé selon les deux méthodes d'estimation. L'estimation 1 correspond au nombre de locus total inféré à partir du partage des locus inférés par pyRAD avec les paramètres Wclust=85% et minDepth=5, l'estimation 2 correspond au nombre de locus inférés à partir de la profondeur de séquençage des locus inférés par pyRAD avec les paramètres Wclust=85% et minDepth=2.

La méthode utilisant la profondeur de couverture infère généralement un nombre de locus plus élevé que la méthode reposant sur les nombres de locus observés. On observe en médiane un écart d'environ 2093 locus entre ces deux estimations, ce qui représente en médiane, 42% de la valeur moyenne des deux estimations. Ce résultat suggère que l'une des méthodes surestime ou sous-estime systématiquement le nombre de locus. Par ailleurs, la même corrélation mesurée en utilisant les nombres de locus estimés par la méthode 1, en utilisant un seuil de profondeur minimale de 2 lectures (et non de 5 lectures) est également significativement corrélés (p-value < 7.10^{-6}, r=0.52). Mais pour ce paramètre les nombres de locus estimé par la première méthode sont cette fois plus élevé. On peut donc supposer que cette première méthode est fortement influencée par le choix des paramètres et par l'impact d'éventuels écarts par rapport aux hypothèses émises par cette approche. Il est également difficile d'estimer précisément la proportion de locus couverts dans les données observées par rapport au nombre de locus total. Celle-ci varie en effet entre une couverture médiane comprise entre 40 et 80% des locus (pour des paramètres similaires), selon la méthode utilisée. L'estimation des nombre de locus total, pour d'autres valeurs de paramètre avec la deuxième

méthode permettra sans doute de mieux comprendre l'influence des ces paramètre et de mieux estimer la proportion de locus couverts.

3.6. Discussion

Dans cette partie, nous avons présenté des améliorations du protocole de préparation des librairies RADseq, permettant le multiplexage de nombreux spécimens d'espèces différentes au sein d'une même librairie. Notamment, nous avons montré que l'utilisation d'adaptateurs optimisés permettait d'augmenter l'efficacité de la ligation de ces adaptateurs à l'ADN génomique tout en diminuant les coûts de préparation des librairies. Par ailleurs, nous avons réalisé une seconde librairie RADseq, correspondant aux ADN de 200 spécimens du système Symbiocode. Malgré plusieurs difficultés conduisant à une réduction de la couverture obtenues par rapport à celle attendue, le séquençage de cette librairie a permis d'obtenir des données RADseq pour 75% des spécimens. De plus 60% des paires de spécimens multiplexées correspondent à une couverture suffisante des deux spécimens pour permettre les analyses prévues qui seront présentées au chapitre 3.

Nous avons également cherché à comprendre l'origine de séquences ne correspondant pas à des RAD tags de l'organisme étudié. Nous avons ainsi pu identifier certaines sources de contaminations, soit par des artéfact techniques (constructions constituées de fragments d'adaptateurs), soit par d'autre organismes, notamment par des micro-organismes dont l'ADN est potentiellement présent dans les extraits d'ADN utilisés. Notons à ce sujet que ces données pourraient être utiles à l'identification de bactéries symbiotiques présentes dans l'ADN des individus. Par exemple, nous avons retrouver un nombre de locus de *Wolbachia* conséquent dans 30% des spécimens étudiés (26 individus possédant plus de 10 locus de *Wolbachia*). Ces infections sont confirmées par les données de PCR 16S obtenues précédemment dans 90% des cas (23/26). L'élimination de ces différentes sources de contamination a permis de réduire la proportion de « faux locus » au sein des données. On peut remarquer que ces sources de contaminations représentent globalement un nombre de locus limité, même si certaines peuvent ponctuellement représenter une forte proportion des locus identifiés chez certains individus. De plus, ces faux locus sont vraisemblablement rarement partagés par plusieurs spécimens. Ainsi, pour la plupart des analyses, pour lesquelles

seules les données partagées par un nombre minimum de spécimens sont utilisées, ces locus devraient avoir un impact faible sur les résultats obtenus.

4. Estimation de la diversité génétique avec des données de RADseq

Des données de RADseq peuvent être utilisées dans différents contextes. Notamment, elles permettent l'estimation des diversités génétiques au sein des populations. Toutefois, des études ont montré que l'usage de ce type de données pouvait introduire des biais dans les estimations du polymorphisme (Gautier et al. 2013, Arnold et al. 2013). Dans cette partie, je décrirai les différents biais affectant l'utilisation des marqueurs RAD pour la mesure du polymorphisme ainsi que les approches mises en œuvre pour mesurer leur impact et éventuellement les corriger.

4.1. Introduction

L'utilisation de données de RADseq pour mesurer la diversité génétique au sein d'une population peut conduire à une sous-estimation du polymorphisme. Un premier biais, lié à l'existence d'un polymorphisme sur les sites de restriction, a été décrit principalement par deux études reposant sur l'analyse de données RADseq simulées (Gautier et al. 2013, Arnold et al. 2013). Le polymorphisme sur les sites de restriction est responsable d'un biais en faveur des coalescents courts dans l'échantillonnage des locus par RADseq. En effet, au sein d'une population, les séquences associées à un site de restriction intact à un locus donné correspondent en moyenne à un coalescent plus court qu'un ensemble des séquences échantillonnées aléatoirement.

Arnold et al. 2013 ont ainsi montré que cet échantillonnage biaisé des coalescents conduisait à une sous-estimation de la diversité génétique, mesurée par π et $\theta_{watterson}$. Ces deux mesures correspondent respectivement à la fréquence moyenne de différences entre paires de séquences choisies au sein de la population et au nombre de sites ségrégeant corrigé par le nombre de séquences observées. En effet, aux locus associés à un site de restriction polymorphe, c'est-à-dire présent chez certains individus seulement, seuls les séquences des individus associés à un site fonctionnel peuvent être comparées. Or ces individus correspondent en moyenne à des coalescents plus courts que celui de l'ensemble de la population (Figure 23).On pourrait envisager d'éviter ce biais dans l'échantillonnage des séquences en restreignant l'analyse aux locus séquencés pour tous les individus. Toutefois, Arnold et al. ont également montré que ce tri des

marqueurs conduit également à un biais, cette fois-ci au niveau de l'échantillonnage des locus. En effet, parmi l'ensemble des locus RAD, on choisit ainsi ceux associés à un site de restriction partagé par un grand nombre d'individus; ces locus sont ceux associés aux régions les plus conservées, et donc aux coalescents les plus courts.

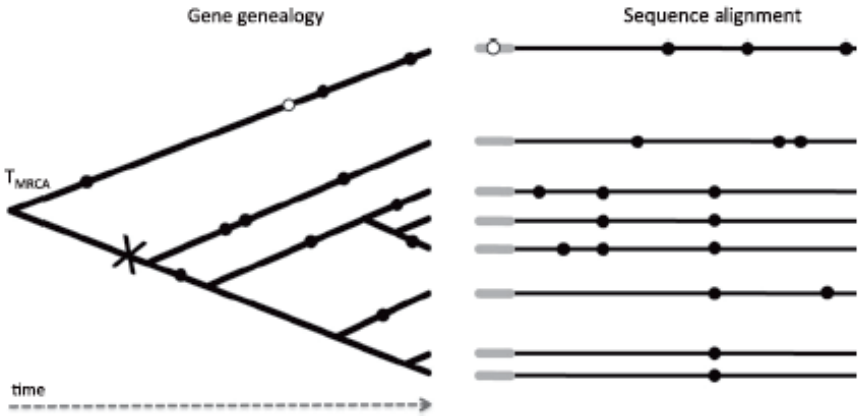

Figure 23. Exemple d'alignement (droite) correspondant à la généalogie d'un locus (gauche). Les points noirs représentent des mutations. Le point blanc représente une mutation altérant un site de restriction (partie grisées sur l'alignement). Dans cet exemple, le T_{MRCA}, correspondant à la coalescence des 8 individus est supérieur à celui correspondant aux 7 individus partageant un site de restriction intact. D'après (Arnold et al. 2013).

Gautier et al. 2013, ont également mesuré l'impact de cette perte de locus causée par le polymorphisme des sites de restriction dans l'estimation de la diversité génétique à partir de locus RAD. Ces auteurs relient ce biais à l'*allele-drop-out*, ADO, décrit par Luca et al. 2011, et qui correspond à la perte d'allèles associée au non-séquençage des allèles nuls, c'est-à-dire associés à des sites de restriction mutés. Dans cette étude, la relation attendue entre l'intensité de l'ADO et le polymorphisme a tout d'abord été calculée, ainsi que l'impact des longueurs des sites de restriction et des régions flanquantes séquencées. Puis, l'effet de l'ADO sur la mesure des fréquences alléliques, et sur l'estimation de la diversité génétique a également été mesuré sur des données simulées. La mesure de la diversité génétique utilisée dans cette étude est l'hétérozygotie attendue qui dépend des fréquences alléliques aux différents locus :

$$He = 2 * f_q * (1 - f_q)$$

où f_q correspond à la fréquence d'un des deux allèles à un locus di-allélique. Ces auteurs ont ainsi montré que l'ADO conduit à une surestimation de He, au sein d'une population. Cette observation s'explique par le fait que l'ADO conduit plus souvent à la perte d'allèles ancestraux majoritaires. Cette perte se traduit donc par des fréquences alléliques plus proches de 50%, valeur pour laquelle l'hétérozygotie théorique est maximale. Ce résultat est cohérent avec celui de Arnold et al. 2013, le même biais lié au polymorphisme sur les sites de restriction résultant en une surestimation de l'hétérozygotie théorique, et en une sous-estimation de la diversité nucléotidique mesurée par π.

En plus de ce biais lié à la longueur des coalescents, d'autres phénomènes peuvent conduire à une mesure biaisée de la diversité génétique à partir de données RADseq. Tout d'abord, la mesure de la proportion de locus homozygotes et hétérozygote pour un individu diploïde est également affectée par le polymorphisme des sites de restriction. En effet, les locus homozygotes « apparents », pour lesquels toutes les séquences obtenues sont identiques, peuvent correspondre à des locus pour lesquels un seul allèle a été séquencé. Ce « biais d'hétérozygotes cachés » est lié à la présence d'allèles associés à des sites de restrictions mutés, mais peut être aussi causé par une profondeur de séquençage limitante, ne permettant pas l'obtention de tous les allèles de chaque locus existant. Ces deux phénomènes conduisent à une surestimation de la proportion de locus homozygotes par hétérozygotie cachée. Soulignons qu'il est question ici de l'hétérozygotie observée, mesurée par la proportion de sites hétérozygotes au sein d'un génome diploïde, et non de l'hétérozygotie théorique déduite des fréquences alléliques, utilisée dans l'étude de Gautier et al. 2013.

Enfin, le protocole RADseq induit un troisième type d'erreur dans la mesure de la diversité génétique. Un biais dans la répartition des locus RAD sur le génome peut en effet conduire à une estimation biaisée du polymorphisme à partir de données RAD (DaCosta, Jeffrey M., Sorenson 2014). Ce bais est lié à la composition en base des sites de restriction, qui peuvent être enrichis en AT ou en GC, ce qui résulte en une hétérogénéité dans la localisation des sites de restriction sur le génome. La diversité génétique mesurée sur les locus RADseq correspond donc à une diversité mesurée sur une partie biaisée des génomes, qui ne correspond pas nécessairement à la diversité moyenne. En particulier, dans un génome de composition en bases hétérogène, les régions riches en

GC peuvent être également enrichies en gène. L'estimation de la diversité à partir de données de RADseq correspondant à un site de restriction CG riche, se fera ainsi sur une fraction du génome enrichie en gène, résultant en une sous-estimation du polymorphisme moyen.

Au cours de cette thèse, j'ai obtenu des données RADseq dans le but de comparer les distances entre génomes nucléaires, à des distances entre génomes mitochondriaux, obtenues par ailleurs, pour de nombreuses paires de spécimens. Il s'agissait donc de mesurer la distance nucléaire entre deux individus issus d'une même population ou de deux populations différentes, à partir de données RADseq. Cette estimation est donc potentiellement affectée par les biais décrits précédemment.

De plus, nous souhaitions également, le cas échéant, comparer la diversité mesurée entre individus à la diversité mesurée au sein de chaque individu (hétérozygotie). Or la mesure de l'hétérozygotie au sein de chaque individu est affectée par le biais « d'hétérozygotie cachée », qui surestime le nombre de locus homozygotes, du fait de l'existence de locus pour lesquels un seul allèle a été séquencé. Notons que l'on peut s'attendre à ce que l'impact de ce biais soit assez important dans les données obtenues, car elles correspondent au séquençage d'une librairie fortement multiplexée, avec des profondeurs parfois limitantes.

Dans cette partie, je me suis principalement intéressée à l'impact de deux de ces biais : la perte de certains locus RAD, du fait du polymorphisme des sites de restriction ou des limites expérimentales (biais d'échantillonnage des coalescent et biais d'hétérozygotie cachée). J'ai, tout d'abord cherché à évaluer la relation entre polymorphisme vrai et polymorphisme observé sur des données de RADseq à partir de données simulées correspondant, dans un premier temps, à des populations panmictiques évoluant de façon neutre. Ces simulations nous ont permis de modéliser le biais lié à l'échantillonnage des coalescents dans les données RADseq. Nous proposons en outre une approche par ABC permettant de corriger ce biais afin de rendre possible l'estimation de la diversité nucléotidique dans une population à partir de données RADseq.

J'ai ensuite étudié l'impact de ce biais sur des données issues de populations réelles au moyen de séquences génomiques. J'ai utilisé pour cela des génomes de spécimens de *Drosophila melanogaster*, provenant du *Drosophila Population Genomics Project* (Pool et al. 2012) ainsi que les transcriptomes de spécimens du projet *popphyl*

(Romiguier et al. 2014), qui regroupe des données associées à de nombreuses espèces animales représentées par plusieurs spécimens par espèce. Il s'agissait tout d'abord de comparer le biais observé dans des données de RADseq générées *in silico* à partir des génomes d'organismes provenant de populations réelles, à celui observé sur des données simulées. Puis, la comparaison du polymorphisme corrigé et du polymorphisme vrai, estimé à partir d'un échantillonnage non biaisé des locus des mêmes organismes nous a permis d'évaluer la qualité de l'estimation du polymorphisme par ABC. J'ai par la suite pu évaluer l'impact du biais d'échantillonnage des coalescents dans le cas de populations structurées, c'est à dire son effet sur la mesure de la distance entre deux spécimens provenant de populations différenciées. Enfin, j'ai également étudié la quantification et la correction par ABC du biais associée à l'hétérozygotie cachée, dans le cas de la mesure de l'hétérozygotie au sein d'un spécimen à partir de données de RADseq.

4.2. Correction du biais d'échantillonnage des coalescents par une méthode ABC

Cette partie présente une méthode permettant de mesurer la diversité génétique d'une population panmictique à partir de données de RADseq, en corrigeant le biais d'échantillonnage des coalescents associé à ce type de données, par une approche bayésienne approximée (ABC).

4.2.1. Principe des Approches Bayésiennes Approximées (*Approximate Bayesian Computation*)

Les méthodes d'inférences statistiques ont pour objet l'estimation de valeurs de paramètres dans une population à partir de données observées sur un échantillon de cette population. Dans ce cadre, les approches bayésiennes cherchent à déterminer la distribution *a posteriori* des valeurs des paramètres, à partir d'une distribution *a priori* et d'une fonction de vraisemblance qui décrit la probabilité associée à des données selon un modèle. L'ABC permet de réaliser ce type d'inférence sans calcul de fonction de vraisemblance (Beaumont et al. 2002, Sunnåker et al. 2013). Une des caractéristiques importantes de cette approche est, en effet, de remplacer l'estimation de la fonction de

vraisemblance par des simulations. Pour cela, de nombreuses simulations sont réalisées selon un modèle dont certains paramètres peuvent prendre différentes valeurs (Figure 24 étape 2), issues de distributions *a priori*. Ces simulations permettent d'obtenir des jeux de données simulées qui seront comparées aux données observées (étape 3 et 4) afin de déterminer une distribution postérieure des valeurs de paramètre (étape 5).

Cette approche implique donc de pouvoir déterminer pour chaque simulation si elle a généré un jeu de donnée suffisamment similaire aux données observées pour être acceptée. Pour cela, chaque jeu de donné (observé ou simulé) est « résumé » par un ensemble de statistiques descriptives (*summary statistics*, μ dans la figure 24) qui doivent être choisies de façon à représenter la variabilité des données (en fonction des valeurs de paramètres). C'est la distance entre ces statistiques, mesurées sur les données simulées et observées (étape 4), qui permet de déterminer quelles données simulées sont les plus proches des données observées. La procédure par laquelle les données simulées sont acceptées ou rejetées, en fonction de leur proximité avec les données réelles s'appelle un algorithme de rejet (*rejection algorithm*). Cet algorithme implique le choix d'un seuil ε, correspondant à la distance maximale autorisée entre les statistiques acceptées et observées, et qui peut être choisi de façon à accepter une fraction plus ou moins importante des simulations.

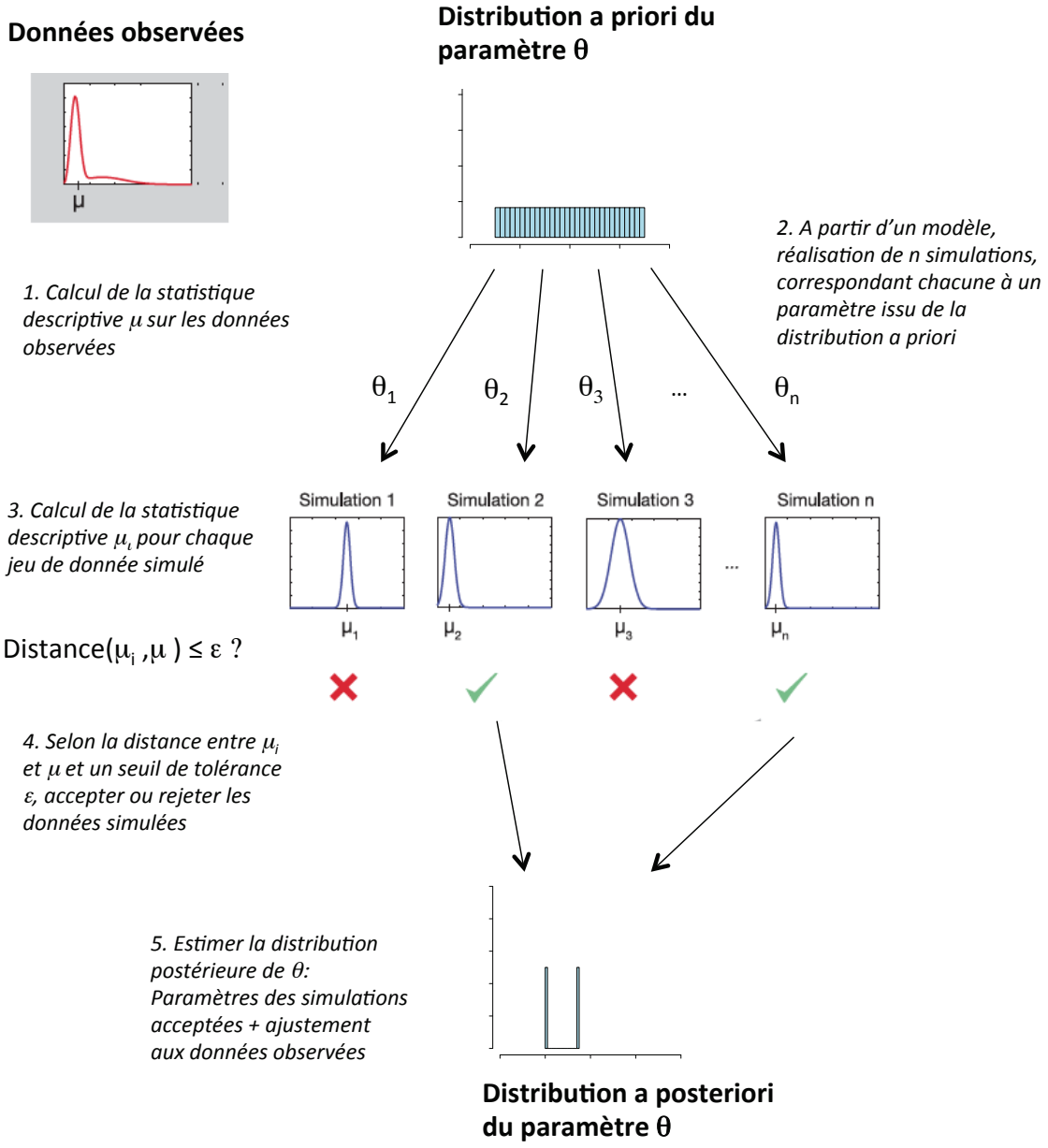

Figure 24. Représentation schématique de l'estimation d'un paramètre θ par ABC (d'après Sunaker et al. 2013)

La distribution des valeurs de paramètres correspondant aux simulations acceptées à l'issue du l' « algorithme de rejet » constitue une première approche de la distribution postérieure des valeurs des paramètres d'intérêt. Toutefois, il est possible d'améliorer l'estimation de la distribution postérieure en prenant en compte les écarts, ε, entre les statistiques descriptives des simulations et celles des données observées. On

peut par exemple modéliser la valeur du paramètre comme une fonction linéaire d' ε plus un terme résiduel (*Local linear regression*, Csilléry et al. 2010) (Figure 26). Estimer cet effet linéaire et le soustraire au paramètre de chaque simulation, c'est-à-dire remplacer chaque paramètre par son terme résiduel, permet de supprimer l'effet de l'écart autorisé entre la statistique des données observées et celles des simulations. On supprime alors de la distribution postérieure la part de variance causée par ces écarts.

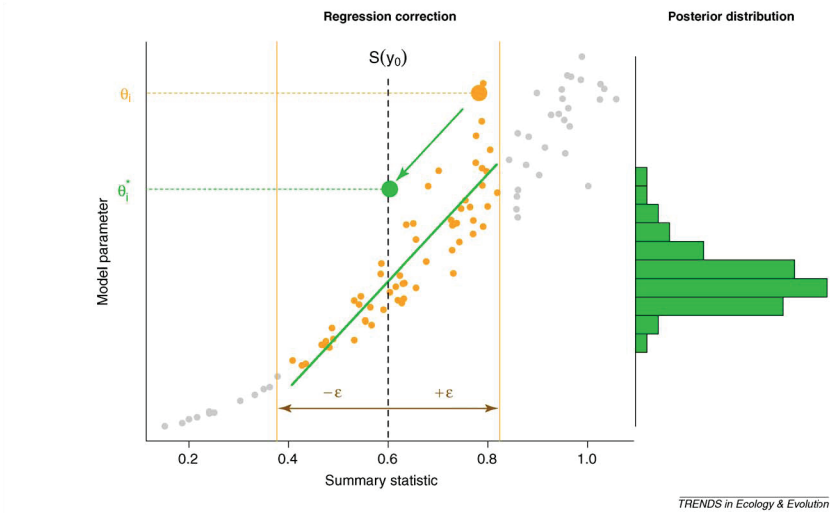

Figure 25. *Linear regression adjustment* dans l'algorithme de l'ABC. Les points oranges représentent les paramètres et statistiques associées aux simulations acceptées, qui sont ensuite « ajustés » par une transformation linéaire (représentée par la flèche verte) υi* = θi – b(S(y) – S(y0)) ou b est la pente de la régression linéaire. Les nouvelles valeurs de paramètre (verts) forment la distribution postérieure (Csilléry et al. 2010).

4.2.2. Méthode RAD_abc : simulations, statistiques descriptives et estimations par ABC

Nous avons appliqué cette approche, par *Approximate Bayesian Computation* à l'estimation de la diversité génétique à partir de données de RADseq. Le polymorphisme mesuré sur des données RADseq dépend à la fois de la diversité réelle dans les populations dont sont issues les données et de l'intensité du biais introduit par le RADseq qui dépend lui même du polymorphisme. L'intérêt de cette méthode est de permettre la modélisation conjointe de la part de l'histoire des population et celle du

protocole de RADseq dans le polymorphisme observés. Il est ainsi possible de corriger le biais introduit par le protocole de RADseq dans son estimation.

Des données de séquences ont été simulées sous un modèle de population neutre à partir de différentes valeurs de polymorphisme. Puis, nous avons simulé l'obtention de données de RADseq à partir de ces jeux de données. Le non-séquençage de certains allèles du fait d'une faible profondeur de séquençage d'une librairie RADseq peut introduire des biais dans la mesure du nombre de locus partagés entre individus. Pour cette raison, l'obtention de données RADseq a été simulée avec différents niveaux de couverture. Des statistiques descriptives correspondant à des mesures observées sur des données de RADseq, et qui seront détaillées plus loin, sont ensuite obtenues pour chaque jeu de paramètres et comparées à celles observées sur les données, afin de déterminer les valeurs de paramètres correspondant aux données les plus similaires à celles observées (Figure 26).

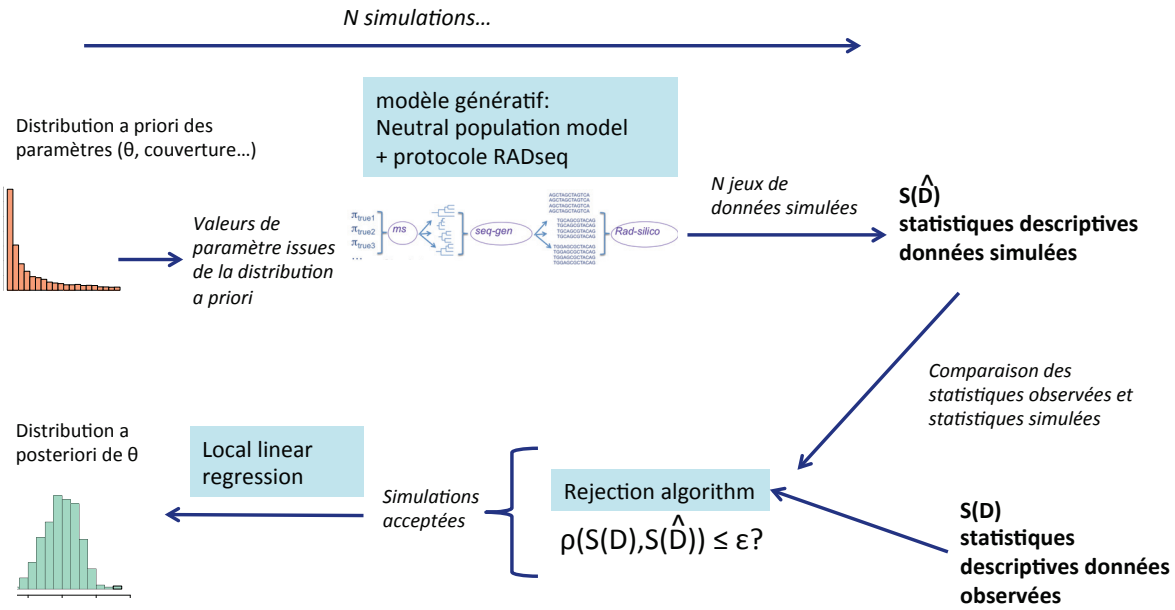

Figure 26: Représentation schématique de l'approche ABC appliquée à la correction du biais introduit par le RADseq dans l'estimation du polymorphisme (RAD_abc)

4.2.2.1. Simulations, mesure du biais sur des données simulées

Les simulations nécessaires à l'approche RAD_abc se déroulent en trois étapes. Il s'agit pour chaque simulation, de générer des séquences correspondant à plusieurs locus (ou chromosomes) échantillonnés chez deux spécimens diploïdes provenant d'une même population. Puis des données de RADseq sont générées *in silico* et les valeurs de statistiques associées, calculées, pour chaque simulation (Figure 27).

1. Le programme *ms* permet tout d'abord de générer des coalescents issus de populations évoluant selon un modèle neutre de Wright-Fisher (Hudson 2002). Dans notre étude, nous simulons les généalogies de locus issus de 2 individus diploïdes au sein de population panmictiques. Chaque simulation permet d'obtenir des généalogies aléatoires de 4 haplotypes (2 individus diploïdes) pour 1000 locus. Chacune de ces 1000 généalogies est représentée sous la forme d'un arbre dont les longueurs de branches, proportionnelles au temps, sont indiquées en unité de 4Ne générations.

Sous un modèle neutre, le polymorphisme est déterminé par Ne, la taille efficace de la population, et μ, le taux de mutation dans une population. Ainsi, le taux de mutation à l'échelle de la population, $\theta=4*Ne*\mu$, est le paramètre contrôlant la proportion de sites polymorphes dans une population. θ est le paramètre qui permet à *ms* de générer des coalescents correspondant à différents niveaux de polymorphisme. Il s'agit également du paramètre d'intérêt que l'on souhaite estimer sur les données réelles avec l'approche RAD_abc. Sous un modèle neutre, θ peut-être estimé par la mesure de la diversité nucléotidique π qui correspond au nombre de différences moyennes (par site) entre deux séquences (chromosomes) échantillonnées aléatoirement dans la population. Par la suite, « θ» désigne le paramètre utilisé pour réaliser les simulations et « π » la mesure de la diversité nucléotidique calculée sur des séquences (simulées ou observées).

2. *seq-gen* permet ensuite de générer des alignements de séquences pour chaque arbre *ms* (Rambaut and Grassly 1997). Sur chaque généalogie, et à partir de séquences ancestrales aléatoires, des mutations sont placées aléatoirement (suivant la valeur de θ). Pour chaque locus, des séquences de 10 000 bp, correspondant à des fréquences égales de chaque nucléotide, sont générées chez les 2 individus diploïdes.

3. Enfin, sur chaque jeu de données de séquences généré, une expérience de RADseq associée à une couverture donnée est simulée et les valeurs de différentes statistiques descriptives sont calculées. Cette étape est réalisée par un programme, *RAD_silico*, écrit dans le cadre de cette étude.

Les jeux de données ainsi simulées ont chacun une longueur de 1000*10000=10^7 paires de base par individu. Dans un jeu de données de cette taille, un motif de 8 paires de bases donné apparaît en moyenne 150 fois. Afin, d'augmenter le nombre de sites RADseq par individu sans augmenter la longueur de séquences, et donc le temps de calcul, 10 motifs de 8 paires de bases différents sont considérés comme sites de restriction possibles pour l'étape de RADseq, ce qui correspond donc à une moyenne d'environ 1500 locus par simulation.

Cette étape nécessite en outre l'utilisation de paramètres décrivant la couverture caractérisant l'expérience de RADseq simulée. Ainsi, les paramètres c1 et c2 (compris entre 0 et 1) correspondent aux probabilités d'échantillonner un allèle dans chacun des deux individus. On suppose que cette probabilité est la même pour tous les locus, et pour chacun des allèles, au sein d'un génome diploïdes. Ici, les paramètres qui déterminent le degré de couverture peuvent être considérés comme des paramètres « de nuisance », c'est-à-dire qu'ils sont inclus dans les simulations non pas parce qu'on cherche à les estimer, mais parce que leurs variations est associée à des variations dans les valeurs des statistiques descriptives. Leur prise en compte est donc nécessaire à une bonne caractérisation de la relation entre les paramètres d'intérêt et les statistiques descriptives.

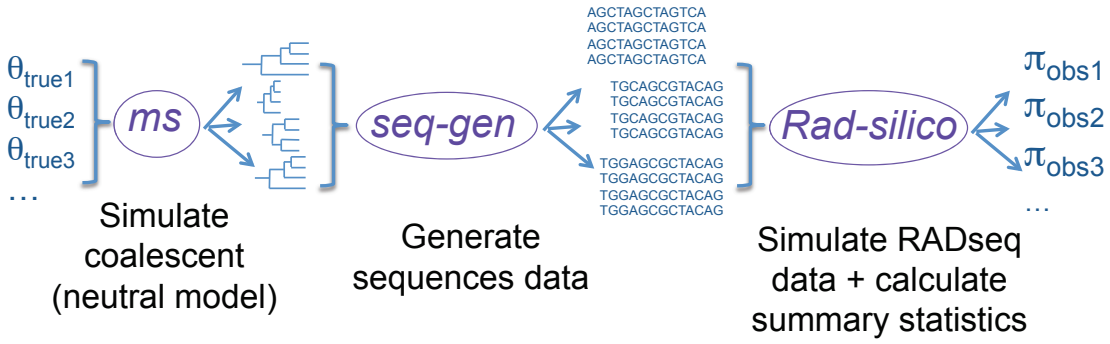

Figure 27: Schéma du protocole des simulations de données RADseq pour l'estimation de la diversité nucléotidique

200500 jeux de données ont été simulés suivant cette procédure pour 2005 valeurs de θ comprises entre 10^{-5} et 0.1, issues d'une distribution *a priori* uniforme de log(θ). Pour chaque valeur de θ, des simulations ont été réalisées pour 100 combinaisons de valeurs de couvertures pour les deux individus, chacune variant de 10% en 10 % entre 0 et 100% de locus échantillonnés.

En premier lieu, ces simulations ont permis de mesurer l'intensité du biais d'échantillonnage des coalescents en fonction de la valeur de θ. Ainsi, pour chaque jeu de données RAD simulé, on peut calculer une valeur de diversité nucléotidique π sur les données RAD simulées et comparer cette valeur à θ. On appelle π_{RAD_obs}, la valeur de π observée correspondant à des données RADseq, c'est-à-dire calculée en utilisant uniquement des séquences associées à un site de restriction intact. La figure 28 représente la relation entre θ et π_{RAD_obs}. On observe bien sur cette figure une sous-estimation de π_{RAD_obs} qui augmente quand la valeur de θ augmente.

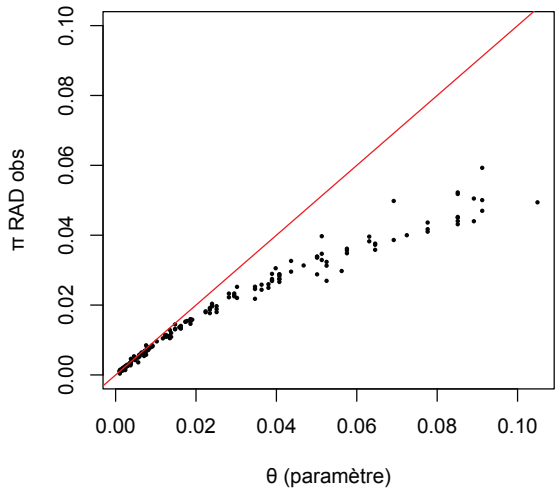

Figure 28. Biais RAD dans l'estimation du polymorphisme : relation entre θ et π_{RAD_OBS}.

Cette relation permet d'évaluer l'impact du biais RAD lié à l'échantillonnage des coalescents sur l'estimation de la diversité génétique à partir de données RAD. En effet, on peut observer que pour un polymorphisme de 10%, la sous-estimation est de l'ordre de 50%. Toutefois, ce type de polymorphisme est très exceptionnel. En revanche, une population associée à un polymorphisme de 2% verra son polymorphisme sous-estimé d'environ 20% par des données RAD. Ceci correspond à une diversité génétique forte, mais probablement pas rare dans les populations naturelles (Romiguier et al. 2014).

4.2.2.2. Statistiques descriptives

Les statistiques descriptives permettent la comparaison des données observées aux données simulées. Elles doivent donc être choisies de manière à « capturer » autant que possible la variabilité des valeurs de paramètres. Idéalement, une combinaison de valeurs de statistiques descriptives devrait correspondre à une seule combinaison de valeurs de paramètres. Les statistiques descriptives utilisées par RAD_abc pour l'estimation du paramètre d'intérêt (θ) sont les suivantes :

- π_{RAD_obs} correspond à la diversité nucléotidique π observée sur des données RADseq entre les deux individus diploïdes. Il s'agit ici de la distance entre individus (moyenne des distances entre individus pour tous les allèles observés) pour tous les locus partagés

par ces deux individus. π_{RAD_obs} est associée à θ de façon non-linéaire ; la sous-estimation de π_{RAD_obs} par rapport à θ augmente avec la valeur de θ (Figure 29 A).

- **p_shared_i1 et p_shared_i2,** sont les proportions de locus partagés entre les deux individus, c'est-à-dire le nombre de locus de l'individu 1 ou 2 qui sont aussi séquencés dans l'autre individu. Les figure 29 B et C, représentent la proportion de locus respectivement de l'individu 1 et de l'individu 2 partagés par les deux individus. La proportion de locus partagés diminue quand le polymorphisme augmente, du fait de la diminution du nombre de sites de restrictions en commun entre 2 individus, ainsi que quand la couverture (de l'autre individu) diminue. Par ailleurs, ces figures montrent la valeur de cette statistique pour les simulations correspondant à une couverture de 100% pour l'individu 1 et à des couvertures variables pour l'individu 2. Ainsi, la figure B montre le rapport entre le nombre de locus partagés et le nombre de locus observés pour l'individu 1. Ce rapport diminue avec le nombre de locus partagés du fait de la réduction de la couverture de l'individu 2. En revanche, la figure C montre le rapport entre le nombre de locus partagés et le nombre de locus observés pour l'individu 2. Ces deux valeurs diminuant avec la couverture de l'individu 2, leur rapport n'est pas affecté par le niveau de couverture.

La figure 29 montre que les statistiques descriptives choisies varient toutes avec la valeur du paramètre d'intérêt θ et que certaines sont également liées aux paramètres caractérisant le niveau de couverture des données de RADseq. Ainsi, les valeurs des paramètres du modèle semblent bien permettre de prédire les valeurs des statistiques descriptives associées. Si les données réelles suivent les mêmes relations entre paramètres et statistiques descriptives, la comparaison des statistiques correspondant à des données réelles à celles des simulations devrait permettre d'obtenir une distribution postérieure de valeurs de θ assez précise.

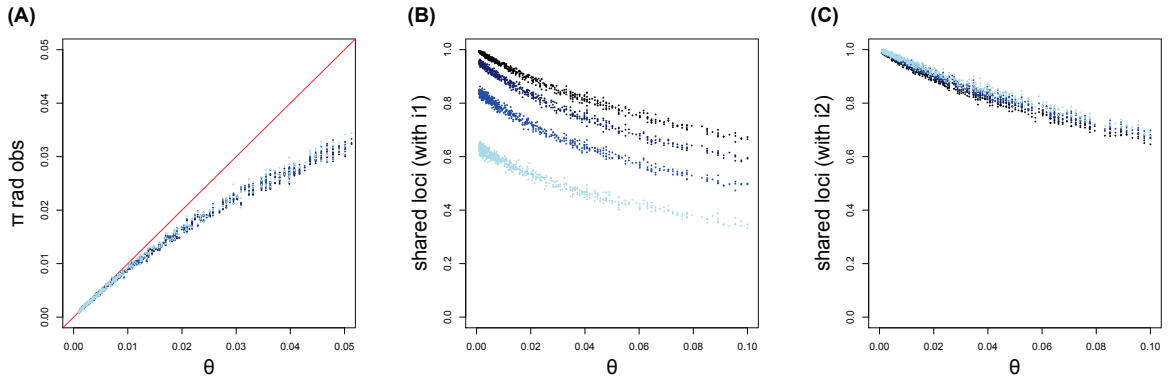

Figure 29: Relation entre statistiques descriptives utilisées pour estimer le polymorphisme à partir de données RAD obtenues pour deux individus d'une même population et θ, le paramètre mesurant le polymorphisme des populations, sur des données simulées. Les couleurs correspondent à différents niveaux de couverture dans l'individu 2. Noir : 100% de probabilité d'échantillonner chaque locus. Bleu foncé : 80%, bleu intermédiaire : 60%, bleu clair : 40%. Cette probabilité est fixée à 100 % dans le premier individu pour cette figure.

4.2.2.3. Estimation du polymorphisme

A partir de données RADseq réelles correspondant à deux spécimens provenant d'une même population, on peut mesurer les statistiques descriptives décrites à la partie 4.2.2.2. Ces statistiques sont comparées à celles associées à chacune des simulations réalisées à l'aide de la fonction abc() du package R *abc* (Csilléry et al. 2012).

Cet algorithme permet d'obtenir des distributions postérieures d'un ou plusieurs paramètres d'intérêt correspondant à un jeu de donné réel, à partir de jeux de données simulées correspondant à des paramètres issus d'une distribution *a priori*. Cette estimation nécessite le choix d'un « taux de tolérance », et d'une méthode d'ajustement. Plusieurs méthodes d'ajustement sont disponibles. La méthode « *loclinear* » utilisée dans cette étude permet d'obtenir des valeurs de paramètres ajustées par régressions linéaires locales, c'est à dire, corrigées pour le fait que les statistiques des simulations acceptées ne sont pas égales aux statistiques observées (voir Figure 25).

4.2.3. Résultats : Validation croisées (*Cross validation*), correction sur données simulées

Afin d'évaluer l'efficacité de cette méthode d'estimation du polymorphisme à partir de données de RADseq, nous avons tout d'abord voulu vérifier qu'elle permettait bien d'estimer le polymorphisme de séquences simulées à partir de données RAD simulées. Cette approche correspond à une validation croisée de la méthode développée. Les validations croisées permettent de tester la capacité d'une procédure ABC à estimer les paramètres du modèle, pour des données générés de la même façon que les données simulées utilisées pour l'estimation.

Les statistiques descriptives de la i-eme simulation, choisie au hasard, sont utilisées comme statistiques descriptives « pseudo-observées », et les paramètres sont ré-estimés en utilisant toutes les simulations sauf la i-ème. Cette méthode permet de mesurer l'erreur de prédiction pour une procédure ABC donnée, mais pas de tester la pertinence du modèle utilisé ou son adéquation par rapport aux processus qui expliquent les données réelles (pour cela, voir l'estimation de la *goodness-of-fit*, Csilléry et al. 2012). Cette approche permet en outre de quantifier l'erreur sur l'estimation d'un paramètre pour un ensemble de simulations et un ensemble de paramètres d'analyse donné. La mesure de l'erreur fournie par le programme utilisé (package R *abc,* Csilléry et al. 2012), correspond au rapport entre la somme des carrés de différences entre valeurs vraies et estimées et la variance des valeurs de paramètre. Cette mesure est donc une proportion de la variance du paramètre que représente l'écart entre valeur estimée et valeur réelle:

$$E_{pred} = (\sum_i (\theta_i estim - \theta_i)^2)/Var(\theta_i)$$

θ_i est la vraie valeur du paramètre correspondant à la i-eme simulation, $\theta_i$estim est la valeur estimée du paramètre, et Var(θ_i) la variance de la valeur du paramètre. L'erreur sur l'estimation de θ par ABC, est indiquée dans le tableau 3 pour différents seuils d'acceptation des simulations. Ce test permet de montrer que pour les 3 niveaux de tolérance testés, cette procédure permet d'obtenir une bonne estimation de θ. En effet, dans notre jeu de simulation, la variance de θ est de l'ordre de 10^{-3}, pour une erreur de l'ordre de 10^{-2}, on peut donc estimer grossièrement que la somme du carrée

des écarts est de l'ordre de 10^{-5}, ce qui correspond à un faible niveau d'erreur. Dans la suite des analyses, le taux de tolérance utilisé est de 0.05.

Tolerance rate	*E_{pred}, prediction error based on a cross-validation sample of 100*
0.01	0.0079
0.05	0.0078
0.1	0.0075

Tableau 3. Résultat du test de cross-validation. Erreur sur l'estimation de θ selon le taux de tolérance utilisé. $\boldsymbol{E_{pred} = \sum_i (\sigma_i estim - \sigma_i)^2 / Var(\sigma\ )}$

Le niveau de couverture caractérisant les données de RADseq a un impact sur la qualité de l'estimation de θ. La figure 30 (A), représente la relation entre valeur de θ et θ_{estim} pour des données pseudo observées, pour différents niveau de couverture. La valeur estimée est ainsi généralement plus proche de la valeur attendue pour des données pseudo-observées correspondant à des couvertures élevées. Cette influence du niveau de couverture est également illustrée par la relation entre le taux d'erreur et la couverture (Figure 30 B.). Pour une couverture inférieure à 50%, on observe une relation négative entre le taux d'erreur et la probabilité de séquencer un locus. En revanche, pour des couvertures correspondant à plus de 50% de locus couverts l'erreur sur θ_{estim} est toujours inférieure à 0.0032. De plus, on peut remarquer que même pour lorsque la couverture est très limitante (10% des locus couverts), l'erreur est assez faible (0.053).

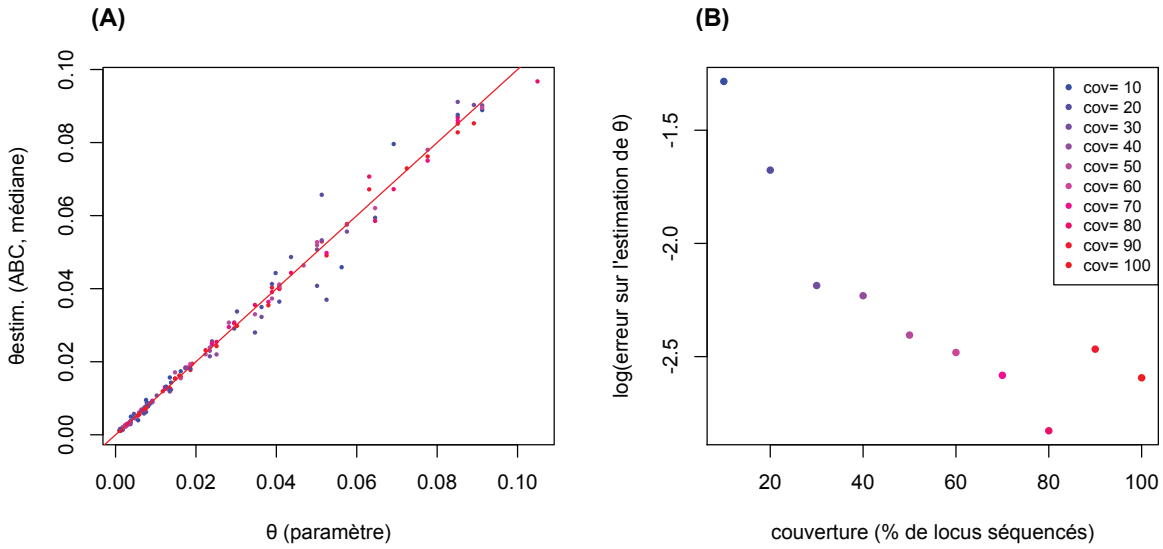

Figure 30. Contrôle positif de la correction du biais RAD dans l'estimation de π par validation croisée ; A. relation entre θ (paramètre) et θ estimé par ABC d'après $\pi_{\_RAD\_obs}$ correspondant aux données pseudo-observées pour différentes couvertures. B. Relation entre l'erreur sur l'estimation de θ et la couverture.

Ces résultats montrent que la méthode RAD_abc permet d'estimer de manière non-biaisée la diversité génétique d'une population à partir de données RADseq obtenues pour 2 spécimens diploïdes de cette population. Le test de validation croisée permet de montrer que le biais que l'on peut mesurer sur des données issues des simulations qui ont permis de décrire ce biais, est bien corrigé par ABC. Nous allons maintenant nous intéresser à l'efficacité de cette correction sur des données issues de populations réelles.

4.3. Correction de π sur deux individus diploïdes : test sur des données réelles et impact d'une structuration en sous-populations

4.3.2. RADseq *in silico* : DPGP et Popphyl

L'approche RAD_abc a été testée sur des données de RADseq générées *in silico* à partir des séquences de génomes et de transcriptomes de différents organismes. J'ai utilisé deux jeux de données dans cet objectif. Il s'agit tout d'abord des génomes complets de spécimens issus de différentes populations de *D. melanogaster* séquencés dans le cadre du projet *Drosophila Population Genomic Project* (Pool et al. 2012). D'autre part, les données du projet Popphyl, qui correspondent aux transcriptomes de plusieurs spécimens issus de nombreuses espèces de métazoaires (Romiguier et al. 2014), ont également été employées.

Ces jeux de données permettent la réalisation d'expériences *in silico* simulant l'obtention de marqueurs RAD pour plusieurs spécimens au sein de chaque espèce. Il est ensuite possible de mesurer sur ces séquences les valeurs de diversités nucléotidiques avec ou sans biais lié à l'échantillonnage des coalescents. π_{RAD_obs} correspond ainsi aux distances moyennes observées entre 2 individus diploïdes, mesuré sur des locus RAD, tandis que π_{RAD_vrai} est calculé sur les même locus (associés à un site de restriction sur un des chromosome au moins), mais sans tenir compte de l'état du site de restriction. Il s'agit ainsi d'une mesure non biaisée de π. Ces données permettent ensuite de comparer le polymorphisme RAD corrigé par ABC au polymorphisme non biaisé de chaque espèce.

Drosophila melanogaster : *Drosophila Population Genomic Project* (Pool et al. 2012).

Les génomes alignés de quadruplets de spécimens provenant de 4 populations de *D. melanogaster* ont été obtenus sur le site du DPGP (http://www.dpgp.org/). Ces quadruplets correspondent à des génomes haploïdes et peuvent être utilisés pour simuler l'obtention de données correspondant au séquençage RADseq de paires de spécimens diploïdes (Tableau 4).

population	Spécimens	$\pi_{génome}$ (%)	π_{RAD_vrai} (%)
ZI	(ZI91, ZI261), (ZI268, ZI468)	1.11	0.74
GU	(GU2, GU6), (GU7, GU10)	0.98	0.61
KR	(KR4N, KR7), (KR42, KR39)	0.96	0.62
RG	(RG2, RG10), (RG15, RG19)	0.99	0.66

Tableau 4. Spécimens des populations de *D. melanogaster* du *Drosophila Population Genomics Project* utilisés dans cette étude. $\pi_{génome}$ correspond à la diversité nucléotidique mesurée par Pool et al. sur l'ensemble des génomes de l'ensemble des spécimens pour une population. π_{RAD_vrai} correspond à la diversité non biaisée par l'échantillonnage des coalescents, mesurée sur les locus RAD sur les 4 spécimens étudiés. Les spécimens haploïdes associés pour constituer des génomes diploïdes sont indiqués entre parenthèse.

Tout d'abord, on peut remarquer que les valeurs π_{RAD_vrai} dans ces populations sont inférieures à 1% (Tableau 4). On peut remarquer en outre que ces valeurs sont inférieures aux valeurs de $\pi_{génome}$ calculées sur l'ensemble du génome par Pool et al. 2012. Cette différence est probablement due au fait que π_{RAD_vrai} correspond au polymorphisme mesuré uniquement aux locus RAD. π_{RAD_vrai} n'est pas biaisé par l'échantillonnage des coalescents, mais il est visiblement affecté par le biais de composition des locus RAD. En effet, si la répartition des locus RAD sur le génome n'est pas homogène (biais lié à la composition en base), la densité en sites de restriction et donc en locus RAD, peut être différentes dans des régions associés à des niveaux de diversité différents (DaCosta, Jeffrey M., Sorenson 2014). On peut remarquer que d'après ces valeurs, il semble que ces biais de composition aient un impact très fort, de l'ordre de 35%. Arnold et al. avaient également étudié l'effet de ce bais de composition sur l'estimation de la diversité nucléotidique par des données RADseq chez *D. melanogaste*r. Leur analyse montrait que le choix de l'enzyme de restriction influe fortement sur les types de régions génomiques échantillonnées, une enzyme GC riche échantillonnant une plus forte proportion de sites dans des exons qu'une enzyme AT riche.

La figure 31 montre la relation entre π_{RAD_obs} et π_{RAD_vrai} ainsi qu'entre π_{RAD_vrai} et $\theta_{estim.}$ pour les 4 populations de drosophiles. On peut remarquer que pour ces individus,

π_{RAD_obs} est très proche de π_{RAD_vrai}. Cette observation est cohérente avec les simulations qui prédisent un biais faible pour des valeurs de diversité nucléotidiques inférieures à 1%. Les valeurs de θ_{estim} sont très proches des valeurs attendues, ce qui correspond toutefois à une très faible correction de la diversité nucléotidique π_{RAD_obs}.

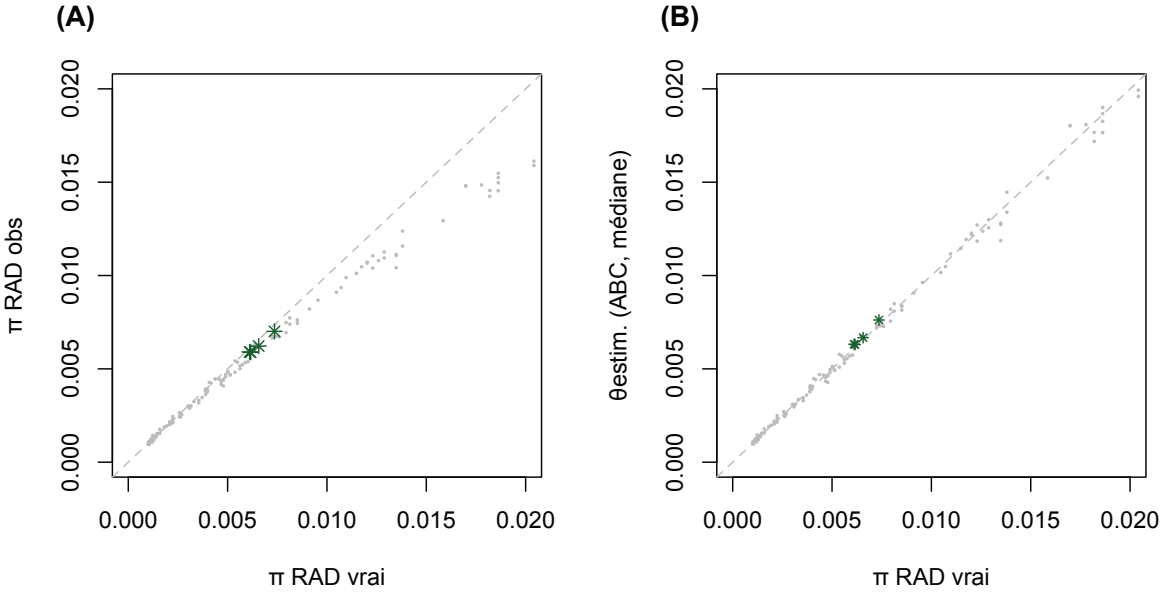

Figure 31. A: Relation entre π_{RAD_vrai} et π_{RAD_obs} et B. relation entre π_{RAD_vrai} et θ_{estim} obtenu par ABC, pour 4 populations de *D. melanogaster* (DPGP) (étoiles vertes). Les points gris correspondent à des données pseudo-observées (simulées). Couverture 100%

Popphyl (Romiguier et al. 2014)

Popphyl regroupe les transcriptomes de spécimens provenant de 90 espèces d'animaux représentées chacune par 2 à 8 spécimens. Pour chacun de ces spécimens, on dispose des séquences des deux allèles des transcrits séquencés par RNAseq. On peut donc, pour des paires de spécimens choisies dans chaque espèce, réaliser une expérience de RADseq *in silico*, et, de la même manière que pour les génomes de drosophile, mesurer π_{RAD_vrai} et π_{RAD_obs}.

Comme ces données correspondent au séquençage de régions codantes des génomes, la quantité de données disponibles par rapport aux séquences de génomes complets est réduite. Ainsi, parmi les 90 espèces de popphyl, seulement 47 paires de spécimens sont associées à plus de 50 locus RAD (Figure 32). D'autre part, tandis que

l'analyse de la diversité génétique à partir de ces données a montré que de nombreuses espèces sont associées à un polymorphisme synonyme très élevé (> 5%, suggérant un polymorphisme assez élevé sur l'ensemble du génome), la diversité calculée sur l'ensemble des sites des régions codantes est bien plus faible pour la plupart des espèces.

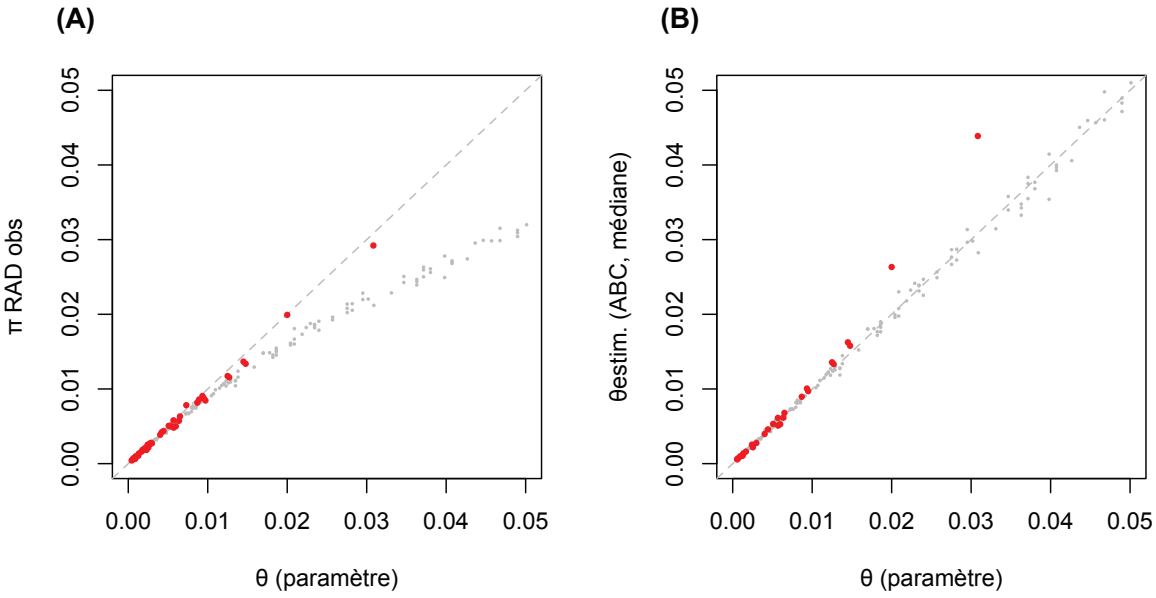

Figure 32 (A) Relation entre π_{RAD_vrai} et π_{RAD_obs} et (B) relation entre π_{RAD_vrai} et θ_{estim} obtenu par ABC pour 47 espèces de popphyl (points rouges). Les points gris correspondent à des données pseudo-observées (simulées). Couverture 100%.

La relation entre π_{RAD_obs} et π_{RAD_vrai} ne montre pas de sous-estimation de π_{RAD_obs} dans les données de popphyl (Figure 32 A). La plupart des populations étudiées sont caractérisées par un très faible polymorphisme, la sous-estimation attendue de π_{RAD_obs} était donc très faible. Néanmoins, on observe que même pour les populations associées à un π_{RAD_vrai} supérieur à 1%, π_{RAD_obs} n'est pas sous-estimé. Le biais observé pour ces espèces fortement polymorphes ne suit pas la relation attendue entre π_{RAD_vrai} et π_{RAD_obs}. En conséquence, dans ces espèces, les plus polymorphes, θ_{estim} est surestimé par l'approche ABC (Figure 32 B). Toutefois, ces données sont assez peu nombreuses et il est donc difficile de déterminer si cette différence tient à une propriété particulière de ces

populations ou au modèle utilisé. D'autres jeux de données pourraient permettre de tester de manière plus appropriée la méthode ABC. Il s'agirait de données de génotypes de plusieurs spécimens d'espèces caractérisées par un fort polymorphisme, parmi lesquels *Caenorhabditis brenneri* ou *Ciona savigni* et/ou *C. edulis* (Dey et al. 2013, Ilut et al. 2014).

Plusieurs hypothèses peuvent être envisagées pour expliquer cette absence de sous-estimation de π_{RAD_obs} pour les espèces les plus polymorphes de *popphyl*. En premier lieu, on peut supposer que les deux spécimens étudiés pour chaque espèce appartiennent à des populations différenciées. Cette structuration pourrait causer une diminution de l'effet du biais lié à l'échantillonnage des coalescents par RADseq sur le π_{RAD_obs}. Une autre hypothèse pourrait être liée au fait que les locus *popphyl* proviennent de régions codantes, tandis que les données simulées correspondent à des séquences évoluant de façon neutre. Ainsi, on peut imaginer que la fréquence des sites de restriction intacts ne correspond pas à la fréquence prédite par le polymorphisme mesuré sur les régions flanquantes de ces motifs, par exemple à cause de l'usage des codons. La modélisation du biais réalisée par les simulations pourrait de cette façon ne pas être appropriée pour ces données.

4.3.3. Impact de la structuration sur le biais d'échantillonnage des coalescents

Nous nous sommes penchés plus particulièrement sur la première de ces hypothèses. Nous avons cherché à mieux comprendre l'impact potentiel d'un écart à la panmixie dans les populations étudiées, et à déterminer si cet effet pourrait expliquer que pour certains spécimens de popphyl, la valeur de π_{RAD_obs} soit différente de celle prédite par les simulations. Cette différence pourrait peut-être s'expliquer par l'existence d'une structuration non prise en compte par le modèle au sein de la population étudiée. C'est–à-dire que les deux individus échantillonnés pourraient provenir de deux populations différenciées génétiquement et non d'une population panmictique (Figure 33). La figure 34 montre l'impact du biais RAD sur l'estimation de la distance entre deux spécimens issus de populations structurées, pour différents temps de divergences entre les populations. Ces données ont été simulées par une méthode similaire à celle décrite dans la partie précédente. Mais cette fois, le modèle

utilisé par *ms* ne correspond pas à l'échantillonnage de deux individus diploïdes dans une population panmictique, mais à l'échantillonnage de deux individus diploïdes provenant de deux populations divergeant depuis un temps t. Dans le modèle utilisé, les deux populations filles se caractérisent par un θ égal au θ de la population ancestrale (Figure 33). On observe sur la figure 34 que plus la divergence entre les deux populations est ancienne, moins la mesure de la distance RAD est biaisée par rapport à la distance vraie.

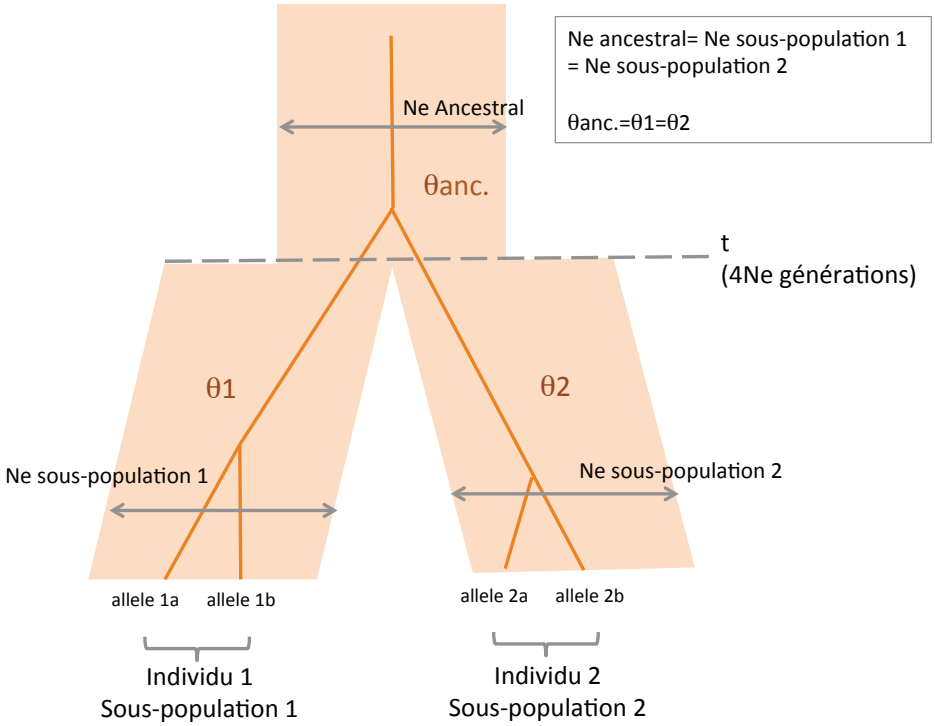

Figure 33. Echantillonnage de 2 individus diploïdes dans une population structurée. L'individu 1 et l'individu 2 proviennent de 2 sous-population divergeant depuis un temps t. Dans les simulations réalisées, les Ne des deux populations filles et de la population ancestrale sont les mêmes (θ=θ1=θ2).

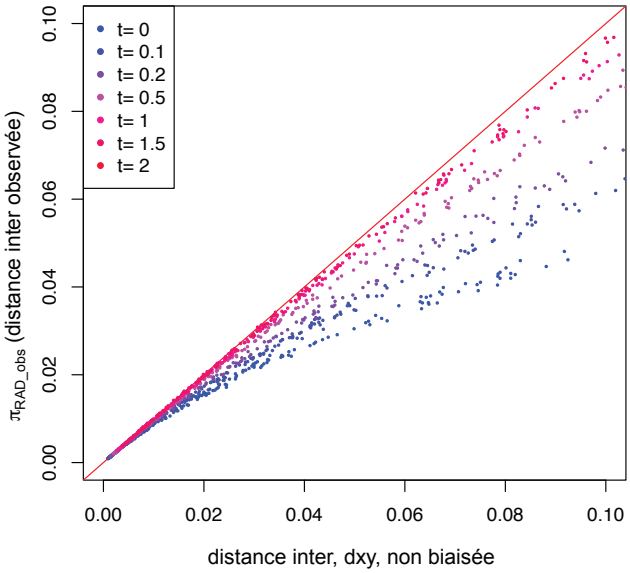

Figure 34. Relation entre la distance inter individu, et la distance inter RAD, mesurée sur les locus séquencés par RADseq. Les couleurs correspondent aux temps de divergence entre les sous-populations, mesurés en unité de 4Ne.

Ce résultat montre que l'intensité du biais affectant la mesure de la distance entre deux individus au sein d'une population, à partir de données de RADseq, est fortement influencée par des déviations par rapport à la panmixie dans ces populations. L'existence d'une structuration en sous-populations est donc un paramètre important à prendre en compte dans l'étude de la diversité génétique de populations potentiellement structurées à l'aide de données de RADseq.

Le biais lié à l'échantillonnage des coalescents a un impact faible sur la mesure de la distance entre spécimens provenant de populations différenciées, et ce biais est d'autant plus faible que les populations divergent depuis longtemps. L'existence d'une structuration non prise en compte lors de l'estimation de la distance moyenne entre individus, π_{RAD_obs}, conduit donc à l'observation d'une valeur de moins biaisée qu'elle ne le serait si les deux individus provenaient effectivement d'une population panmictique. Cet effet conduit donc à une sur-correction du biais lors de l'estimation de θ par ABC.

4.4. Estimation de la proportion de locus hétérozygotes sur un individu : Correction du biais d'hétérozygotie cachée

L'utilisation de données RAD conduit également à des mesures biaisées de l'hétérozygotie observée au sein des individus (Figure 35). En effet, cette mesure repose sur la quantification du nombre de locus homozygotes et hétérozygotes. Du fait du polymorphisme sur les sites de restriction, certains sites sont considérés comme homozygotes, car toutes les lectures associées sont identiques. Mais ces lectures peuvent en fait provenir du séquençage d'un seul allèle, l'autre allèle étant associé à un site de restriction modifié. On peut noter à cet égard que du fait de l'hétérogénéité de la profondeur de séquençage entre locus, il est généralement impossible d'utiliser la profondeur pour distinguer les locus pour lesquels un ou deux allèles sont séquencés. D'autre part, la proportion de locus « hémizygotes » ou « homozygotes apparents » augmente avec le polymorphisme et avec la baisse du niveau de couverture. En effet, quand la couverture permet de séquencer la totalité des séquences associées aux sites de restriction intacts, les locus hémizygotes ne sont dû qu'au polymorphisme sur les sites de restriction. En revanche, quand la couverture est faible, certains allèles ne sont pas séquencés faute de couverture et s'ajoutent aux allèles perdus causés par les sites de restrictions mutés. Nous nous sommes demandés si une approche similaire à celle mise en œuvre pour l'estimation de la diversité nucléotidique pourrait permettre d'estimer l'hétérozygotie mesurée chez un seul individu par une approche ABC, en corrigeant le biais « d'hétérozygotie cachée » introduit par le polymorphisme sur les sites de restriction et l'échantillonnage des locus lors du séquençage.

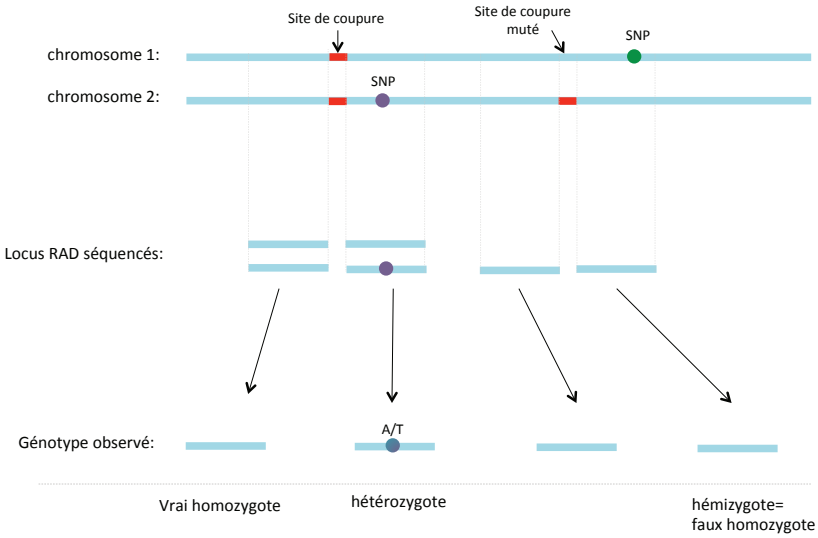

Figure 35. Biais lié à l'hétérozygotie cachée. La présence de locus hémizygotes, associés à un site de restriction polymorphe conduit à une sous-estimation de la proportion de locus homozygotes dans un individu diploïde.

Dons notre étude, l'hétérozygotie vraie, H_{vraie} est la moyenne des distances entre allèles observés au sein d'un individu. Dans une population panmictique, H_{vraie} correspond donc à θ. Dans le cadre de l'approche ABC développée, les statistiques descriptives permettant d'estimer H_{vraie} sont l'hétérozygotie RAD observée (nombre de SNP observés par site, sur les locus RADseq observé, $\mathbf{H_{RAD\_obs}}$) et l'hétérozygotie RAD observée sur les locus polymorphes uniquement ($\mathbf{H_{RAD\_obs\_polymorphe}}$). Ces deux statistiques varient avec H_{vraie} (Figure 36). Mais la première n'est corrélée à H_{vrai} que quand celle-ci est faible. Pour une H_{vraie} supérieure à 2%, la variation de l'hétérozygotie observée $\mathbf{H_{RAD\_obs}}$ est très faible. Ainsi, on peut imaginer que la mesure de l'hétérozygotie observée sur les locus polymorphe permette de mieux estimer H_{vrai}, dans la mesure où celui ci varie avec H_{vraie} sur toute la gamme de valeurs testées. En outre, pour une hétérozygotie vraie donnée, $\mathbf{H_{RAD\_obs}}$ est fortement influencée par le niveau de couverture. Une faible couverture des locus RADseq entraine en effet un forte diminution de $\mathbf{H_{RAD\_obs}}$, du fait de l'augmentation de la proportion de locus hémizygotes quand la couverture diminue. $\mathbf{H_{RAD\_obs\_polymorphe}}$, en revanche est calculé uniquement sur des locus pour lesquels les deux allèles sont séquencés. Cette statistique est donc peu influencée par le niveau de couverture et devrait permettre une meilleure estimation de l'hétérozygotie en cas de couverture imparfaite. Cette combinaison de statistiques

descriptives devrait donc permettre l'estimation de H_{vraie}, mais aussi du niveau de couverture (dont dépend $\mathbf{H_{RAD\_obs}}$), pour une expérience de RADseq donnée. Je me suis intéressée ici uniquement à l'estimation de l'hétérozygotie.

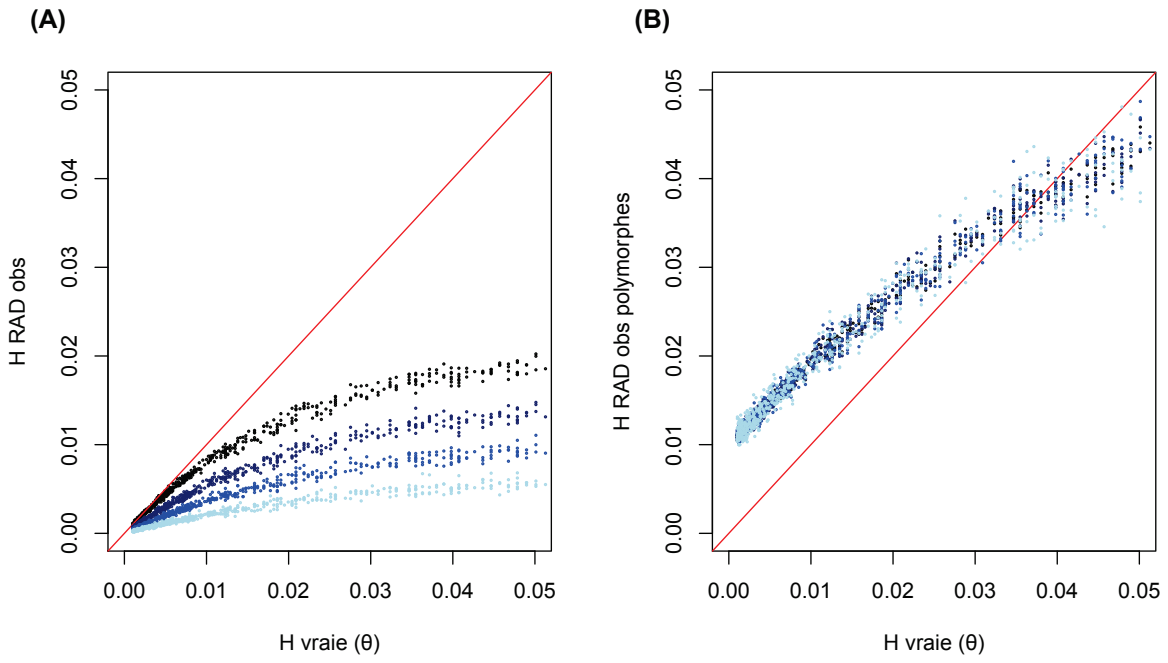

Figure 36. Relation entre polymorphisme et statistiques descriptives utilisées pour l'estimation du polymorphisme sur un individu diploïde. Les couleurs correspondent à différents niveaux de couverture. Noir : 100% de probabilité d'échantillonner chaque locus. Bleu foncé : 80%, bleu intermédiaire : 60%, bleu clair : 40%.

Une validation croisée permet de mesurer l'erreur que cette approche introduit sur la mesure de H_{vraie}. On peut observer que cette erreur est plus importante que celle affectant l'estimation de θ sur deux spécimens diploïdes (table 5), mais l'estimation de l'hétérozygotie à partir des marqueurs RAD d'un individu semble néanmoins assez précise, notamment quand la couverture est supérieure à 50% (Figure 37). D'autre part, de même que pour l'estimation de la diversité nucléotidique, la qualité de cette estimation dépend du niveau de couverture associé aux données RADseq utilisées. La figure 37 B montre ainsi une relation inverse entre l'erreur sur l'estimation de H estimée et le niveau de couverture.

Tolerance rate	*Prediction error based on a cross-validation sample of 100*
0.01	0.03
0.05	0.03
0.1	0.03

Tableau 5 Résultat des tests de validation croisée sur l'estimation de l'hétérozygotie

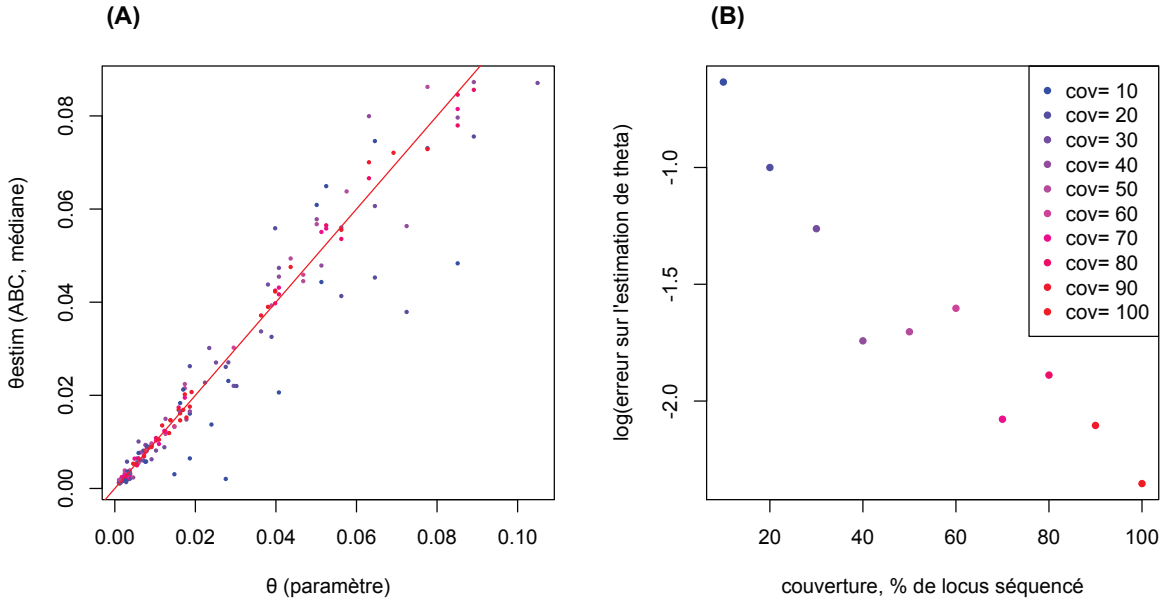

Figure 37. Validation croisée de la correction du biais RAD dans l'estimation de l'hétérozygotie (H) ; A. relation entre H_{RAD_vraie} et H estimée par ABC d'après H_{RAD_obs} correspondant aux données pseudo-observées pour différentes couvertures. B. Relation entre l'erreur sur l'estimation de H et la couverture (taux de tolérance de 0.05).

L'estimation de l'hétérozygotie à partir d'un génome diploïde a également été testée pour des paires de génomes haploïdes de *D. melanogaster* issues des quatre populations présentées précédemment. La figure 38 représente la relation entre H_{RAD_vraie} et les deux statistiques descriptives, H_{RAD_obs} et $H_{RAD\_obs\_polymorphe}$ ainsi que H estimé, pour des expérience de RADseq simulées avec différents niveaux de couverture. On peut ainsi observer que les relations entre H_{RAD_vraie} et H_{RAD_obs}, d'une part, et H_{RAD_vraie}

et $H_{RAD\_obs\_polymorphe}$, d'autre part, suivent celles prédites par les simulations. Les valeurs estimées, H_{estim}, correspondent également aux valeurs attendues. Par ailleurs, comme le suggéraient les validations croisées, tandis que l'estimation de H_{RAD_vraie} semble très bonne pour des niveaux de couverture correspondant à 100% ou 60% de locus couvert, elle semble moins précise quand seulement 40% des locus sont couverts.

Cette estimation a également été réalisée pour les spécimens de popphyl. On peut ainsi observer que, sur ces données, H_{RAD_obs}, H_{RAD_vraie} et $H_{RAD\_obs\_polymorphe}$ suivent la relation attendue et que les valeurs estimées de H_{vraie} (θ) par ABC sont proches des valeurs attendues. Ces résultats suggèrent que la méthode utilisée permet une assez bonne correction du biais d'hétérozygotie cachée dans l'estimation de l'hétérozygotie observée, chez un individu, à partir de données de RADseq. De la même façon que pour *D. melanogaster*, les estimations de H_{vraie} correspondant à des expériences de RADseq *in silico* associées à des couvertures respectivement de 60 et 40% montrent que la qualité de ces estimations décroit avec la proportion de locus couverts.

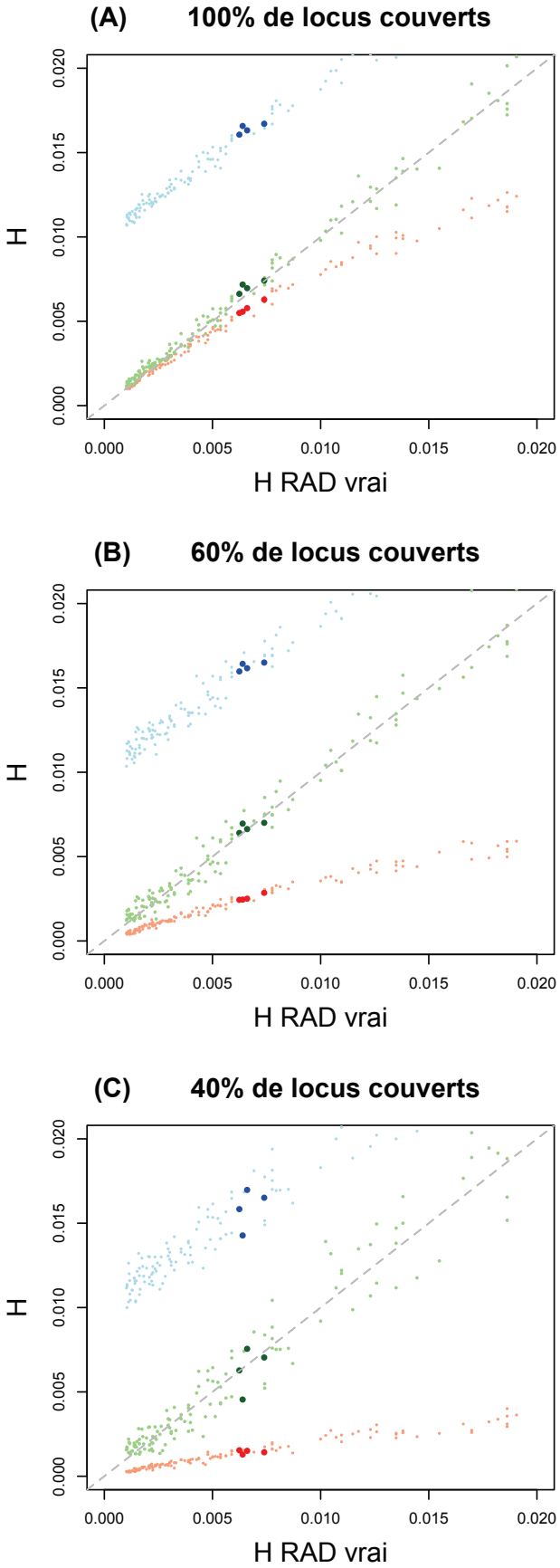

Figure 38 Relation entre $H_{RAD\ vraie}$ et H_{RAD_obs} (points rouges), $H_{RAD\_obs\_polymorphe}$ (points bleus) et H estimé par ABC (points verts) pour quatre spécimens de *D. melanogaster*, et selon différents niveaux de couverture : (A) 100%. (B) 60%. (C) 40%. Les points de couleur claire correspondent à des données pseudo-observées.

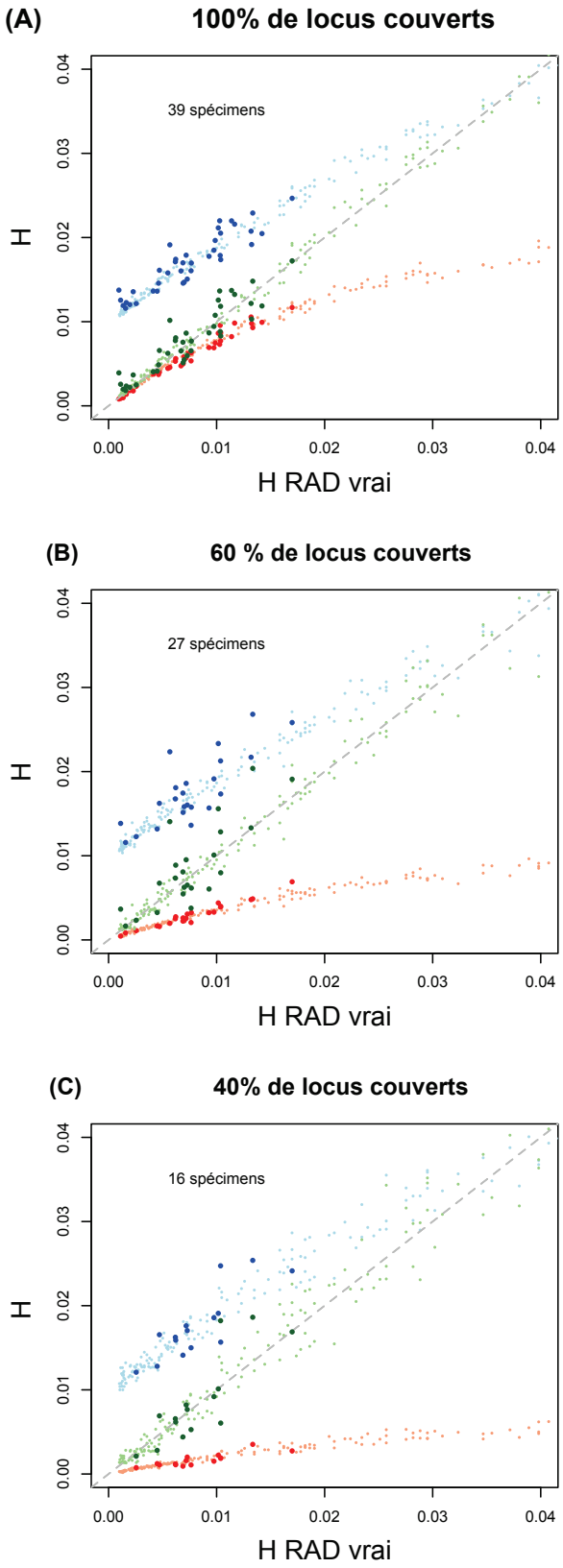

Figure 39 Relation entre $H_{RAD\ vraie}$ et H_{RAD_obs} (points rouges), $H_{RAD\_obs\_polymorphe}$ (points bleus) et H estimé par ABC (points verts) pour les spécimens de popphyl pour lesquels plus de 50 sites de restriction polymorphes ont été observés dans une expérience de RADseq *in silico* utilisant 10 motifs de restriction différents, et selon différents niveaux de couverture : (A) 100%. (B) 60%. (C) 40%. Pour chaque niveau de couverture, on considère ensuite les valeurs correspondant aux spécimens pour lesquels plus de 50 locus polymorphes sont obtenus. Les points de couleur claire correspondent à des données pseudo-observées.

4.5. Discussion

Trois sources d'erreurs affectant l'estimation de la diversité génétique dans les populations à partir de données de RADseq ont été identifiées. Tout d'abord, le biais d'échantillonnage des coalescents conduit à une sous estimation du polymorphisme. La mesure de l'hétérozygotie observée au sein des individus est également affectée par un biais lié à la présence de locus hétérozygotes « cachés » et conduisant également à une sous-estimation de la diversité. Enfin, l'hétérogénéité de la composition en bases du génome peut également être à l'origine d'un biais dans l'estimation du polymorphisme causée par une répartition non aléatoire des locus RAD le long du génome.

Mes simulations, ainsi que les données empiriques montrent que le biais lié à l'échantillonnage des coalescents, mis en évidence par Arnold et al. 2013 et Gautier et al. 2013, affecte peu l'estimation de la diversité nucléotidique dans des populations où la diversité est modérée. De plus, l'existence d'une structuration dans les populations étudiées semble réduire son impact ; la mesure de la distance moyenne entre spécimens est d'autant moins biaisée que la structuration est importante. Par ailleurs, j'ai développé une approche ABC permettant de corriger ce biais dans le cas de populations panmictiques évoluant de façon neutre. Toutefois, dans le cas d'échantillonnage d'individus provenant potentiellement de populations structurées, cette approche conduit à une sur-correction du biais d'échantillonnage des coalescents. La prise en compte de ces situations supposerait donc le développement d'une approche ABC permettant de modéliser cette possible structuration.

J'ai également mesuré l'impact du biais affectant l'estimation de l'hétérozygotie observée au sein des individus. Nous avons pu observer que ce biais, lié à l'existence d'une hétérozygotie cachée, peut être important, y compris pour des populations dont la diversité réelle est modérée. De plus, une couverture imparfaite des locus RADseq peut augmenter fortement la proportion de locus hétérozygotes « cachés », et donc la sous-estimation de l'hétérozygotie. Nous avons développé une approche RADseq permettant de corriger ce biais. Les expériences de validation croisée, mais également les tests effectués au moyen de données RAD obtenues *in silico* à partir de séquences génomiques suggèrent que cette méthode permet une correction efficace de l'estimation de l'hétérozygotie observée au moyen de données de RADseq.

Enfin, un biais est lié à la composition en nucléotides du site de restriction de l'enzyme utilisée pourrait également affecter l'estimation de la diversité génétique à partir de ce type de données. Les mesures de diversité nucléotidique effectuées sur les génomes de *Drosophila melanogaster* suggèrent que ce biais pourrait avoir généralement un impact plus fort que celui lié à l'échantillonnage des coalescents. Dans le cas d'une enzyme associée à un site de coupure GC riche, ce biais correspond à un enrichissement des données RAD en séquences codante, et donc à une sous-estimation de la diversité. Ainsi, dans certains cas, ce biais devrait pouvoir être réduit par l'identification des locus correspondant à des régions codantes ou non au sein des données, de manière à n'utiliser que des positions évoluant de façon neutre. Mais en l'absence de génome de référence annoté, ce filtre des locus RAD est impossible. Les données obtenues dans le cadre de cette thèse, dans le but d'identifier des paires d'individus associés à une divergence nucléaire incompatible avec le fort apparentement de leurs génomes mitochondriaux, seront donc probablement affectées par ce biais. Toutefois, notons que si la sous-estimation attendue de la divergence nucléaire mesurée est responsable d'une diminution de puissance de cette comparaison, elle devrait conduire à des estimations conservatives de la fréquence des discordances.

5. Conclusion

Différentes questions concernant la production et l'utilisation des données de RADseq ont été abordées au cours de cette thèse. Tout d'abord, les échantillons étudiés présentent certaines particularités, en terme de disparité et de concentration des ADN. L'utilisation de cette approche a donc nécessité le développement de méthodes de préparation des librairies et d'analyse des données spécifiques. De plus, d'une manière générale, les données générées par les méthodes de RADseq ont des caractéristiques particulières. Notamment, l'obtention des marqueurs homologues repose sur la présence de motifs déterminés qui sont susceptibles d'être polymorphes au sein des espèces et de ne pas être conservés quand les populations divergent. J'ai ainsi étudié au moyen d'expériences *in silico* l'utilisation de ce type de données pour la résolution de phylogénies. Je me suis également intéressée à la mesure de la diversité génétique au sein des espèces. Cette étude a permis de quantifier et de mieux comprendre les biais

causés par le polymorphisme des sites de restriction dans ce cadre et de proposer une méthode permettant l'estimation du polymorphisme à partir des données de RADseq en tenant compte de ces sources d'erreur.

Chapitre 3

Wolbachia et évolution des génomes mitochondriaux.

Les *Wolbachia*, comme de nombreux symbiotes cytoplasmiques, induisent différents phénotypes favorisant leur propagation. Ces symbiotes sont ainsi fréquemment invasifs. Or, de par leur mode de transmission, ils sont liés génétiquement aux autres éléments cytoplasmiques avec lesquels ils sont co-transmis d'une génération à l'autre. La propagation d'une infection peut ainsi entrainer celle des lignées mitochondriales liées. Nous allons nous intéresser à l'impact de cette sélection indirecte sur l'évolution des génomes mitochondriaux.

Ces balayages sélectifs induisent potentiellement une réduction de la taille efficace des populations mitochondriales, Ne, qui peut se traduire de différentes manières sur l'évolution des génomes mitochondriaux. Tout d'abord, elle peut conduire à une réduction de la diversité mitochondriale au sein des espèces infectées. Ainsi, des réductions du polymorphisme mitochondrial attribuées à la propagation d'une infection par *Wolbachia* ont déjà été observées dans différentes espèces. Par exemple, Turelli et al. (1992), ont étudié, chez *Drosophila simulans*, la propagation d'une infection par *Wolbachia* à travers la Californie. Cette propagation s'accompagnait d'un balayage sélectif sur les mitochondries de cette espèce, dont tous les spécimens infectés possédaient un même variant mitochondrial, alors que les individus non-infectés étaient polymorphes. D'autres études, théoriques, ont par ailleurs montré que la propagation de ce type de symbiotes induit des balayages sélectifs affectant à terme l'ensemble des mitochondries de la population infectée (Fine 1978, Johnstone et al. 1996).

D'autre part, dans le contexte d'hybridations entre espèces, et de flux de génomes mitochondriaux entre espèces, les balayages sélectifs causés par des symbiotes cytoplasmiques peuvent favoriser l'introgression de mitochondries exogènes dans une nouvelle espèce.

Enfin, la réduction de la taille efficace des populations de mitochondries peut également causer une diminution de l'efficacité de la sélection. Cet effet est associé à une augmentation de l'effet de la dérive qui conduit à une diminution des probabilités de fixation des mutations avantageuses et une augmentation de la probabilité de fixation des mutations délétères. Ainsi, si les mutations délétères sont plus fréquentes que les mutations avantageuses, la diminution de Ne se traduit par une augmentation des taux de substitution, en particulier non-synonymes, sur les lignées infectées. La diminution du Ne mitochondrial peut donc conduire à des différences de rapport dN/dS entre branches infectées et non-infectées dans les phylogénies des espèces hôtes.

Ainsi, les réductions de Ne causées par des symbiotes cytoplasmiques peuvent laisser des signatures moléculaires sur les génomes mitochondriaux, à la fois pendant et après leur propagation au sein d'une nouvelle population d'hôtes. Je me suis intéressée à la détection de trois types d'effets de *Wolbachia* sur l'évolution des génomes mitochondriaux : (1) facilitation des introgressions mitochondriales d'une part, (2) réduction du Ne mitochondrial d'autre part, conduisant à une réduction de la diversité mitochondriale ou (3) à une augmentation des rapport de dN/dS sur les lignées infectées.

L'impact de la réduction de Ne causée par l'invasion d'une bactérie cytoplasmique sur les génomes mitochondriaux dépend de son ancienneté. En effet, après un balayage sélectif causé par la propagation d'une infection, le polymorphisme des génomes mitochondriaux des hôtes peut être rétabli par accumulation de nouvelles mutations. Les diversités des génomes mitochondriaux associés à des infections anciennes sont ainsi de nouveau déterminées par l'équilibre entre mutation et dérive. Ces infections anciennes ne devraient donc pas être associées à des diminutions des diversités mitochondriales de leurs hôtes. En revanche, l'impact de la réduction du Ne mitochondrial sur l'accumulation de substitutions non-synonymes devrait pouvoir être détecté dans ces cas d'infections plus anciennes. Toutefois, notons que même dans ce cas, l'effet est dû uniquement aux substitutions fixées pendant que le Ne était réduit, c'est à dire pendant la phase d'invasion.

La détection des réductions de Ne mitochondrial liées à des invasions par *Wolbachia* dépend donc fortement des vitesses de remplacement (*turn-over*) des infections. Plus ce dernier est rapide, plus une proportion importante des infections observées à un temps donné sont récentes, et donc plus la relation entre diversité mitochondriale et statut d'infection devrait être forte. L'effet de la réduction de Ne sur les taux de substitutions sera également d'autant plus fort que les infections sont fréquemment en phase d'invasion. Toutefois, dans ce cas, un *turn-over* rapide des infections peut correspondre à des changements fréquents de statuts d'infection le long des branches de l'arbre des hôtes, ce qui pourrait entraver la comparaison des taux de substitutions sur les branches infectées et non-infectées. Ainsi, la mesure des différents effets des invasions par *Wolbachia* sur les Ne mitochondriaux est fortement liée à la dynamique de ces invasions, qu'il conviendra de discuter au cours de chacune des analyses présentées dans ce chapitre.

1. Recherche d'introgressions mitochondriales liées à *Wolbachi*a

Les hybridations, c'est à dire les croisements entre individus d'espèces non totalement isolées, permettent l'introduction d'allèles d'une espèce donneuse dans une espèce receveuse. On parle d'introgression lorsque cet évènement est suivi d'une augmentation de la fréquence de l'allèle « exogène » dans l'espèce receveuse. Cette augmentation en fréquence peut-être adaptative, quand la version introgressée du locus est sélectionnée positivement, ou se faire par dérive. Elle peut également être due à la sélection d'un élément avec lequel le locus introgressé est lié génétiquement. Par exemple, dans le cas d'une hybridation entre espèces dont une est infectée par *Wolbachia*, l'infection peut être introduite dans une nouvelle espèce, associée avec un haplotype mitochondrial exogène. Dans ce cas, la propagation de *Wolbachia* dans l'espèce receveuse peut favoriser l'augmentation en fréquence de cette mitochondrie introduite par une hybridation.

1.1. Introduction : *Wolbachia*, introgressions mitochondriales et discordance nucléo-cytoplasmiques

Des introgressions mitochondriales ont déjà été mises en évidence dans différents groupes d'insectes. Dans le cas du papillon *Hypolimnas bolina* (Charlat et al. 2009), par exemple, l'introgression mitochondriale a été mise en évidence par la comparaison des séquences mitochondriales et des souches de *Wolbachia* présentes chez différents individus. Ces auteurs ont ainsi observé que les individus infectés par une souche particulière de *Wolbachia* (wBol1a) étaient associés à un haplotype mitochondrial unique très divergent (>5%) des autres haplotypes de l'espèce. Le variant mitochondrial associé à wBol1a était en outre similaire à ceux d'espèces voisines ; les mitochondries de *H. bolina* formant ainsi un groupe paraphylétique. Ce patron peut s'expliquer par une introgression de cet haplotype associé à l'invasion de la bactérie wBol1a (Figure 40).

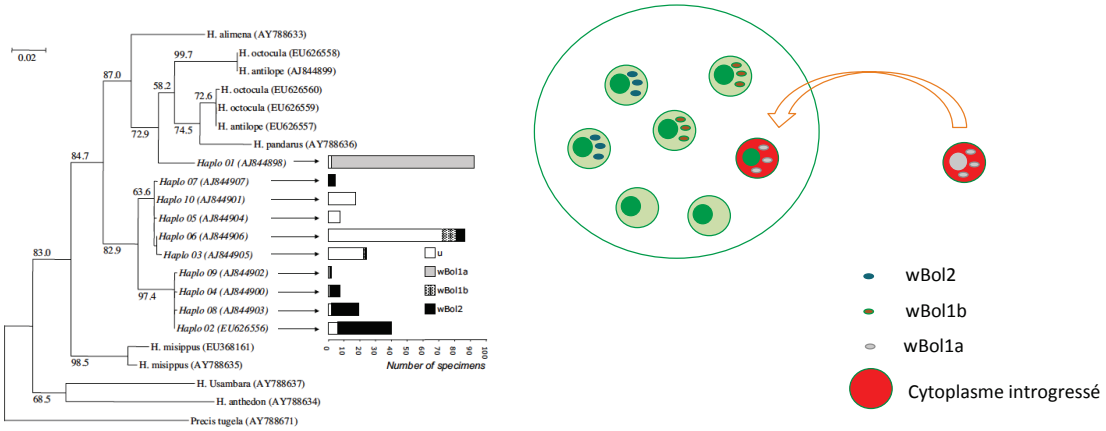

Figure 40. Introgression mitochondriale et infections par Wolbachia chez Hypolimnas *bolina*. D'après Charlat 2009, *BMC Evolutionary Biology*. La partie de droite schématise l'introduction d'une lignée cytoplasmique exogène par hybridation, qui peut être suivit d'une introgression. La partie de gauche présente la phylogénie des espèces du genre *Hypolimnas* basée sur un locus mitochondrial, et les fréquences des différents statuts d'infection par Wolbachia associées aux différents haplotype d'*H. bolina*. On observe en particulier que l'haplotype 01, strictement associé à la souche de *Wolbachia* wBol1a induisant du *male-killing*, est fortement divergent du reste de l'espèce, suggérant un patron d'introgression lié à l'invasion partielle de l'espèce par cette souche de *Wolbachia*.

De façon similaire, dans le cas d'*Acraea encedon* et *Acraea encedana*, Jiggins (2003) a observé une paraphylie des haplotypes mitochondriaux chez *A. encedon*, ce qui suggère une introgression d'*A. encedana* vers *A. encedon* (Figure 41). De plus, une *Wolbachia* induisant une mortalité anormalement élevée des embryons mâles (*male-killing*) infecte à la fois *A. encedana* et certains spécimens d'*A. encedon*, généralement associés à la lignée mitochondriale potentiellement introgressée. Enfin, la monophylie des marqueurs nucléaires de chaque espèce confirme que ce patron est spécifique des mitochondries et donc que la paraphylie des mitochondries de *A. encedon* s'explique probablement par une introgression mitochondriale causée par *Wolbachia*.

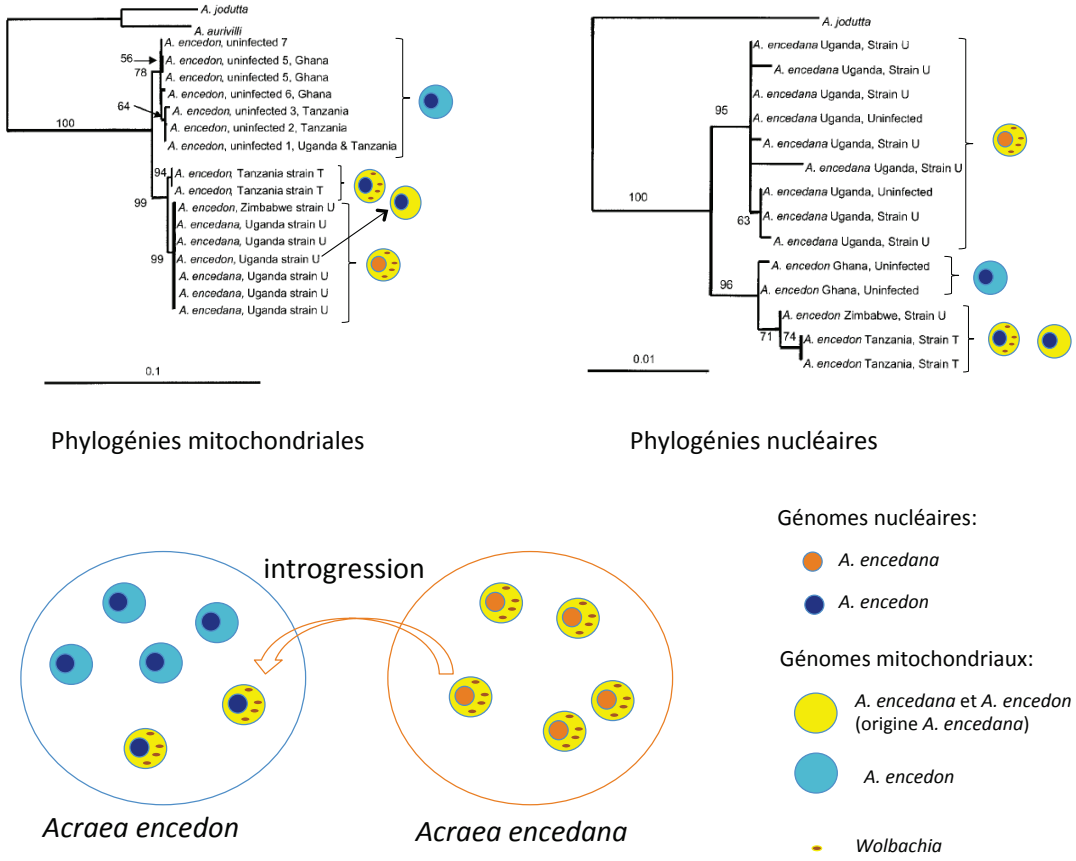

Figure 41. Introgression mitochondriale entre *Acraea encedana* et *A. encedon*. D'après Jiggins 2003, *Genetics*. La partie supérieure présente les phylogénies mitochondriales et nucléaires de ces deux espèces, ainsi que les statuts d'infections des individus étudiés. On observe ainsi que les mitochondries de *A. encedon* sont paraphylétiques et que les individus porteurs de mitochondries plus proches de celles de *A. encedana* sont en outre infectés par *Wolbachia*. La partie inférieure schématise l'introduction d'une lignée cytoplasmique d'*Acraea encedana* dans *A. encedon* par introgression.

Des discordances entre marqueurs nucléaires et cytoplasmiques ont également été observées entre spécimens *d'Eurema hecabe* type Y (*yellow*) et type B (*brown*) (Narita et al. 2006). L'étude de ces discordances a permis de détecter l'introgression d'un haplotype mitochondrial d'une population de type Y dans les populations de type B. De plus, la parfaite concordance entre la phylogénie des mitochondries et la répartition des infections suggère que cette introgression est liée à l'invasion d'une *Wolbachia*.

Les marqueurs mitochondriaux sont fréquemment utilisés pour reconstruire l'histoire des populations. Ce choix s'explique par plusieurs raisons. Tout d'abord, l'évolution des génomes mitochondriaux est généralement plus rapide que celle des

marqueurs nucléaires, ce qui leur permet d'être informatifs à une courte échelle évolutive. De plus, de nombreux marqueurs, d'obtention relativement facile, ont été développés. Cela explique par exemple le choix de séquences mitochondriales dans le cadre des projets de *barcoding* qui utilisent ces marqueurs pour assigner des spécimens à des espèces (Hebert et al. 2003). Toutefois, les cas d'introgression mitochondriale présentés précédemment suggèrent que dans certains cas, les marqueurs mitochondriaux constituent de mauvais marqueurs de l'histoire des populations (Hurst and Jiggins 2005). En effet, dans ces cas, l'histoire des mitochondries est en contradiction avec celles des génomes nucléaires et des populations. Il apparaît donc important d'estimer la fréquence de ces événements pour connaître l'impact de ce phénomène sur l'histoire des mitochondries, et plus généralement sur la pertinence de l'utilisation des mitochondries pour inférer l'histoire des populations. Un des objectifs de ma thèse était d'évaluer la fréquence de ces phénomènes d'introgressions mitochondriales ainsi que l'implication de *Wolbachia* dans ces évènements.

Les introgressions mitochondriales sont à l'origine de discordances entre génomes mitochondriaux et nucléaires. Une introgression peut notamment se traduire, comme dans le cas du papillon *Acraea* discuté précédemment, par des incongruences topologiques entre phylogénies mitochondriales et nucléaires. Ces différences de topologie peuvent donc permettre d'identifier certains évènements d'introgression mitochondriale. Toutefois, on peut remarquer que tous les événements d'introgression ne se traduisent pas par des incongruences topologiques. Une introgression aboutissant au remplacement de l'ensemble des mitochondries d'une espèce par celles d'une espèce sœur, par exemple, ne conduit pas à une incongruence topologique (Figure 42). De plus des incongruences phylogénétiques entre locus peuvent également être dues à des tris de lignées incomplets, liées au polymorphisme ancestral des locus considérés. Ainsi, dans le cas où l'on observe une incongruence entre locus mitochondriaux et nucléaires, il peut être difficile de distinguer une introgression d'un tri de lignées incomplet. Les incongruences topologiques ne constituent donc pas des critères infaillibles pour la détection des introgressions: ces incongruence ne sont pas nécessairement causées par des introgressions, et d'autre part, certaines introgressions ne génèrent pas d'incongruence topologique.

L'introgression mitochondriale cause également des discordances entre temps de divergences mitochondriaux et nucléaires (Figure 42). Dans le cas d'*Hypolimnas bolina*

par exemple, certains individus possèdent des mitochondries d'origine différentes. La divergence entre ces mitochondries correspond à la divergence entre l'espèce donneuse et l'espèce receveuse. A l'inverse, des paires de mitochondries échantillonnées au sein de l'espèce donneuse d'une part et de l'espèce receveuse d'autre part peuvent être plus proches qu'attendu compte tenu de la divergence nucléaire entre ces deux espèces.

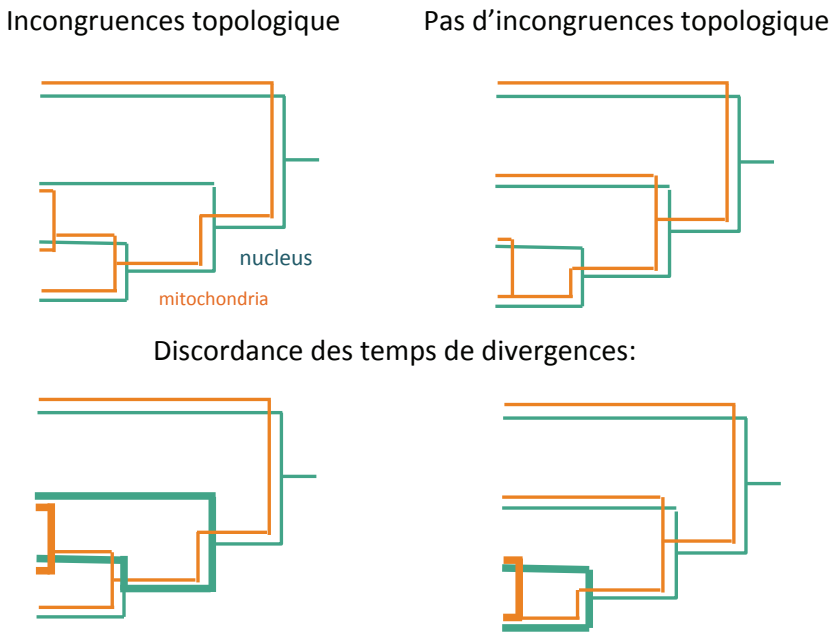

Figure 42 : Impact d'une introgression mitochondriale sur les discordances entre génomes mitochondriaux et nucléaires

Afin d'évaluer la fréquence des évènements d'introgressions mitochondriales, j'ai donc recherché des paires de spécimens affectées par ce deuxième type de discordance. Il s'agit d'identifier des paires de spécimens d'espèces différentes caractérisées par des mitochondries plus proches qu'attendues entre ces deux spécimens compte tenu de leur divergence nucléaire. En effet, si *Wolbachia* ou d'autres facteurs induisent fréquemment des introgressions mitochondriales, cela doit conduire fréquemment à l'observation de mitochondries excessivement proches dans des espèces différentes, et ce d'autant plus dans les espèces infectées. Au contraire, si l'introgression mitochondriale est anecdotique, le fort apparentement indiqué par les mitochondries devrait être confirmé par des marqueurs nucléaires.

Nous avons choisi de rechercher des discordances entre individus correspondant à des temps de coalescence mitochondriaux inférieurs aux temps de coalescence nucléaire et non l'inverse, tout d'abord car l'évolution des mitochondries est plus rapide que celle des génomes nucléaires. En effet, une plus grande divergence entre génomes mitochondriaux au sein d'une espèce peut, dans une certaine mesure, s'expliquer par un plus fort taux de mutation. A l'inverse, des mitochondries trop proches, ne peuvent correspondre qu'à une introgression. Par ailleurs, des séquences mitochondriales avaient été obtenues avant ma thèse, il était donc possible de choisir des paires de spécimens sur la base de leur distance mitochondriale.

Nous avons donc constitué des paires de spécimens de séquences CO1 identiques, et pour lesquelles nous avons estimé la divergence nucléaire. Il s'agit de déterminer si ces divergences nucléaires sont cohérentes avec le fort apparentement des mitochondries indiqué par la similarité des séquences CO1. Autrement dit, nous cherchons à déterminer si les génomes nucléaires de ces deux individus proviennent de groupes différenciés, entre lesquels l'observation de séquences mitochondriales aussi proches ne pourrait s'expliquer que par une introgression mitochondriale.

1.2. Résultats

1.2.1. Estimation des divergences nucléaires à partir de données RADseq.

Nous avons défini des paires de spécimens correspondant à des séquences mitochondriales identiques, afin de déterminer si certaines de ces paires correspondaient à des spécimens issus d'espèces différentes, dont la proximité des séquences mitochondriale serait due à une introgression. Dans le but d'estimer les divergences nucléaires, nous avons réalisé une librairie RADseq correspondant au multiplexage de 90 paires de spécimens de Lépidoptères et de Diptères. 10 spécimens ont également été répliqués au sein de cette librairie afin d'évaluer la répétabilité des mesures de distances effectuées. Des données ont été obtenues pour 51 paires de spécimens (cf. chapitre 2, la perte d'une partie des données s'explique notamment par des problèmes expérimentaux liés aux adaptateurs).

Les séquences RADseq ont été analysées en suivant le protocole décrit au chapitre 2. Pour cette analyse, j'ai utilisé un seuil à 5 lectures minimum pour

l'identification des locus (*clustering* des lectures au sein des individus). En effet, quand la couverture est faible, des erreurs de séquençage peuvent être interprétées comme des polymorphismes réels (SNP) au sein de la population considérée. Nous avions ainsi observé que les locus peu couverts tendent à surestimer les distances nucléaires entre individus (non montré). L'utilisation des locus bien couverts uniquement permet donc de limiter l'impact des erreurs de séquençage sur les calculs des distances entre allèles peu couverts, ce qui est important pour cette partie de mon étude qui porte sur l'estimation de la distance moyenne entre spécimens.

Pour les 51 paires étudiées, les deux individus partagent en moyenne 914 locus. La figure 43 représente la distribution du nombre de locus partagés au sein de chaque paire. Les 29 paires dont les deux membres partagent plus de 100 locus orthologues ont été utilisées pour la suite de l'analyse. Ces paires partagent en moyenne 1410 locus (la médiane correspond à 827 locus partagés et les premier et troisièmes quartiles à 372 et 2438 locus). La moyenne des distances génétiques observées, que nous noterons d_{xy}, a été mesurée pour ces 29 paires.

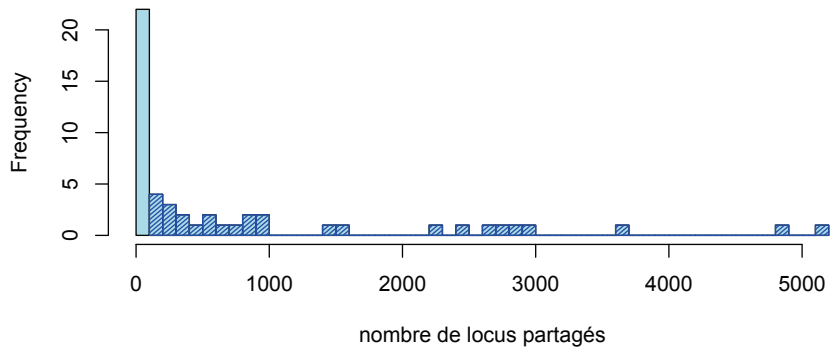

Figure 43. Distribution des nombres de locus partagés pour les 51 paires. La partie hachurée représente les 29 paires partageant plus de 100 locus.

La figure 44 montre la distribution des distances d_{xy} inférées. Ces distances varient entre 0.11% et 2% (moyenne 0.79%, médiane= 0.85%). Ce niveau de divergence n'est pas supérieur au polymorphisme observé classiquement au sein d'espèces d'insectes. Parmi les espèces du projet *popphyl* (Romiguier et al. 2014) par exemple, 26 espèces d'Arthropodes sont représentées. Le polymorphisme synonyme de

ces espèces varie entre 0.1% chez *Reticulitermes grassei* et 4.1% chez *Culex pipiens* (moyenne 1.41%, médiane= 0.97%). L'observation de ces distances ne permet donc pas de rejeter l'hypothèse selon laquelle ces paires correspondent à des individus issus de la même espèce. Ce résultat suggère qu'aucune de ces 29 paires ne correspond pas à des spécimens d'espèces différentes possédant des mitochondries proches acquises par introgression.

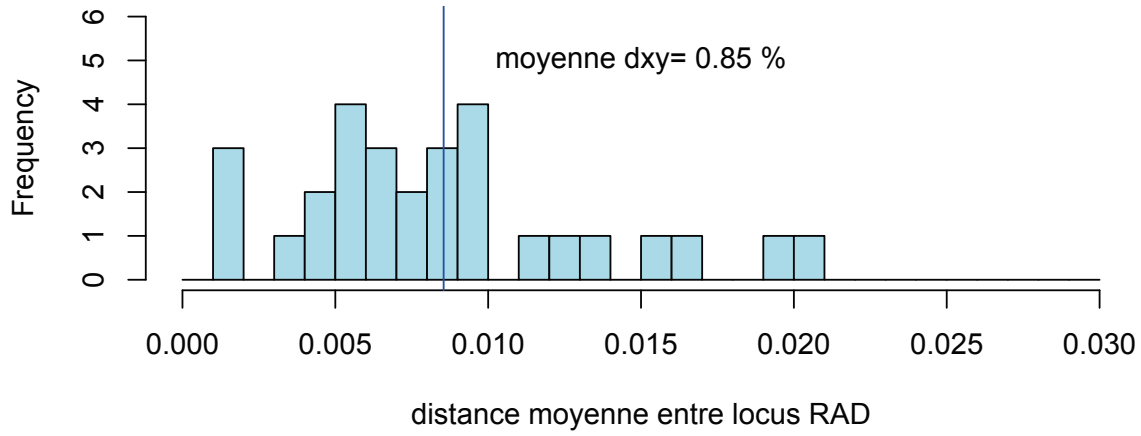

Figure 44. Distribution des diversités nucléotidiques estimées par RADseq pour ces 29 espèces, à partir de deux spécimens par espèce, présentant des séquences CO1 identiques.

Au sein d'une population panmictique, cette mesure des distances moyennes entre individus, d_{xy}, correspond à une diversité nucléotidique (π). Par ailleurs, nous avions montré au chapitre précédent que l'estimation de la distance entre deux spécimens pouvait être biaisée dans des données de type RADseq. Toutefois le biais lié à l'échantillonnage des coalescents est faible quand cette distance est faible. Or cette mesure est ici inférieure à 2% pour toutes les paires. La relation entre valeur de d_{xy} et π corrigée par ABC selon la méthode utilisée au chapitre 2 (Figure 45), confirme que le biais est probablement très faible sur ces données.

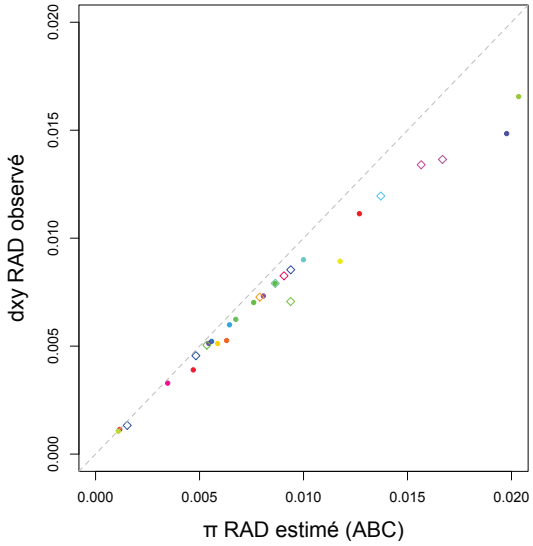

Figure 45. Relation entre distance inter individus corrigée par ABC et distance inter observée, d_{xy}, sur les 29 paires partageant plus de 100 locus. Les points carrés correspondent aux espèces de Lépidoptère et les points ronds aux espèces de Diptères

1.2.2. Répétabilité de l'estimation des divergences RADseq.

L'introduction de réplicats dans cette librairie m'a permis d'estimer la répétabilité de la mesure de la distance nucléaire entre spécimens, d_{xy}. Pour 5 paires de spécimens, un individu (que nous appellerons individu 2) a été introduit deux fois (échantillons 2a et 2b). Pour ces paires, on peut donc calculer la distance entre l'individu 1 et l'individu 2a, d'une part, et entre l'individu 1 et l'individu 2b, d'autre part. La figure 46 montre les valeurs obtenues pour chacun de ces réplicats. On observe un écart moyen de 6.10^{-4} (compris entre 5.10^{-5} et 0.0017) sur la mesure de d_{xy} entre réplicats. L'erreur sur l'estimation de d_{xy}, mesurée en proportion de la variance de d_{xy} sur l'ensemble des mesures est de 0.057 (écart de 5% de la variance totale):

$$erreur_{dxy} = (\sum (d_{xy1} - d_{xy2})^2)/(5 * var(d_{xy}))$$

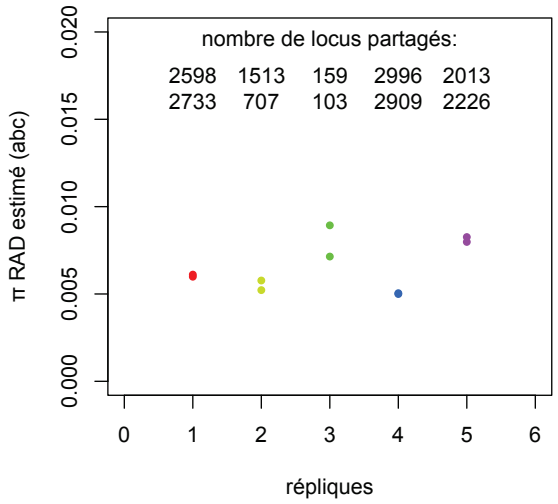

Figure 46. Mesure de distances entre spécimens pour les paires pour lesquels un spécimen est répliqué.

On peut remarquer que l'écart de d_{xy} entre paires répliquées est maximal pour la paire répliquée caractérisée par la plus faible couverture. On peut donc supposer que la répétabilité de la mesure de distance est affectée par la couverture des locus des spécimens impliqués. Toutefois, parmi les 29 paires associées à un nombre de locus partagé supérieur à 100, seules 4 partagent entre 100 et 200 locus. On peut donc supposer que l'incertitude affectant la mesure de d_{xy} pour ces paires est faible pour l'ensemble des données.

1.2.3. Comparaison de l'hétérozygotie et des distances entre spécimens

Les mesures de dxy ont montré que les 29 paires de spécimens étudiées étaient associées à des distances inférieurs à 2% entre spécimens ce qui ne correspond pas à une divergence entre spécimens anormalement élevée pour des spécimens issus de la même espèce. Il semble donc qu'aucune de ces paires ne corresponde à l'échantillonnage de deux spécimens d'espèces différentes impliquées dans une introgression mitochondriale. Des événement d'introgression peuvent également se produire entre populations différenciées, mais potentiellement faiblement divergentes.

Nous nous sommes donc demandés si certaines des paires de spécimens provenaient de populations différenciées.

Cette question peut être abordée en comparant la distance entre spécimens, dxy, à l'hétérozygotie mesurée au sein de chaque individu, H_{obs}. En effet, au sein d'une population panmictique, on s'attend à observer des distances génétiques identiques entre deux allèles choisis aléatoirement, qu'ils soient observés dans le même individu ou non. En d'autres termes, la distance moyenne entre allèles échantillonnés dans différents individus devrait être égale, en moyenne, à la distance moyenne entre allèles échantillonnés au sein d'un individu :

$$d_{xy} = H_{obs}$$

Afin de détecter une structuration entre spécimens, nous avons donc comparé la distance entre spécimens, dxy, à l'hétérozygotie mesurée au sein de chaque individu, H_{obs}. On peut rapprocher cette comparaison de la mesure de l'estimateur, *Fis*, utilisé en génétique des population pour mesurer un déficit en hétérozygotes. Le *Fis* mesure en effet l'écart entre l'hétérozygotie observée et l'hétérozygotie attendue en fonction des fréquences alléliques. La présence de structuration spatiale ou de tout autre processus induisant un déficit en hétérozygote par rapport à l'équilibre de Hardy-Weinberg est ainsi détectable par un *Fis* significativement supérieur à 0. De même, dans notre test, cet excès d'homozygotie devrait se traduire par une mesure de H_{obs} inférieures à dxy. Ce qui distingue la mesure du *Fis* de la comparaison que nous effectuons dans cette étude est que le *Fis* correspond à des mesures d'hétérozygoties observées moyennes entre plusieurs spécimens (généralement sur un nombre limité de locus) tandis que dans notre cas, la mesure est effectuée sur un seul individu, mais correspond à une moyenne sur un grand nombre de locus.

Nous avons mesuré l'hétérozygotie, H_{obs}, observée au sein de chaque spécimen. Il s'agit de la proportion de sites hétérozygotes au sein d'un individu. Cette mesure a été effectuée sur des données de RADseq, elle est donc vraisemblablement affectée par un biais d'hétérozygotie cachée, lié au polymorphisme sur les sites de restriction et à l'échantillonnage incomplets des locus lors du séquençage quand la couverture est faible (cf. Chapitre 2, partie 4.4.). Dans la librairie étudiée, les premiers résultats (cf. Chapitre 2, partie 3.5.) suggèrent que la couverture est faible au moins pour certains spécimens;

l'hétérozygotie observée pour ces spécimens est donc probablement sous estimée. Ces estimations de l'hétérozygotie ont donc été corrigées pour ce biais d'hétérozygotie cachée selon la méthode proposée au chapitre 2 (partie 3.5) (Figure 47A).

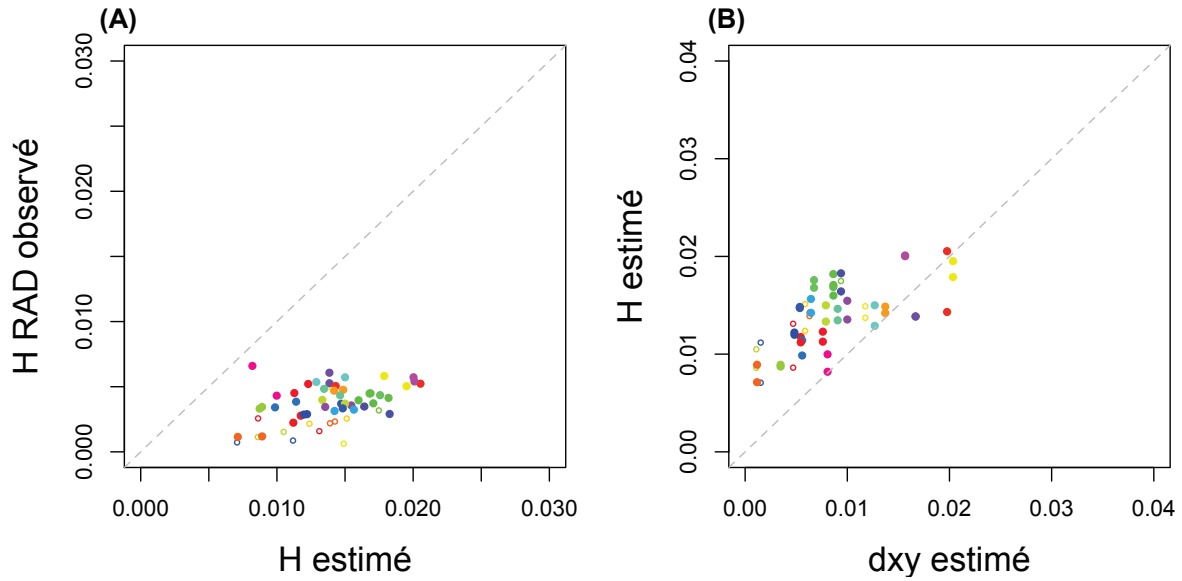

Figure 47. (A) Relation ente l'hétérozygotie RAD estimée par ABC et observée. (B) Relation entre distance entre spécimen (dxy) corrigée et hétérozygotie corrigée. Les points vides correspondent aux spécimens pour lesquels moins de 50 locus polymorphes sont disponibles.

Aucune paire de spécimen n'est associée à une distance entre spécimens, dxy, supérieure aux hétérozygoties mesurées au sein des spécimens (Figure 47B). Ce résultat semble indiquer qu'aucune paire ne correspond à l'échantillonnage de 2 spécimens issus de populations différentes. Toutefois, on peut remarquer que pour la majorité des paires, les valeurs d'hétérozygotie estimées par ABC sont supérieures aux mesures de distances entre individus. Cette observation pourrait s'expliquer de différentes façons. Une première hypothèse envisageable est que les individus comparés sont fortement apparentés au sein des paires. Les distances entre spécimens correspondraient ainsi à des distances entre haplotypes plus apparentés que la moyenne des haplotypes au sein de la population. Cette explication est assez peu vraisemblable, dans la mesure où cet effet semble affecter toutes les paires y compris celles échantillonnées dans des îles différentes. De plus, on n'observe pas de différence de dxy entre les paires en fonction de leurs distances géographiques, ce qui indique que l'apparentement ne biaise pas la

mesure des distances entre spécimens dans ces échantillons (communication personnelle Sylvain Charlat).

Une deuxième hypothèse, plus vraisemblable est que cette observation est due à un artefact de mesure. Notons par exemple que la correction de l'hétérozygotie par ABC repose sur la relation entre l'hétérozygotie vraie et l'hétérozygotie observée uniquement sur des locus RADseq polymorphes. Sur la figure 47, les paires correspondant à des nombres de locus polymorphes inférieurs à 50 ont été représentés par des points vides. On peut observer que ces paires, ne semblent pas associées à des estimations de H_{obs} particulièrement biaisées. Le faible nombre de locus ne semble donc pas expliquer une surestimation de l'hétérozygotie. Mais par ailleurs, cette relation entre H_{vraie} et $H_{obs\_polymorphe}$ est déduite de simulations correspondant à un modèle dans lequel le polymorphisme est homogène le long du génome. On peut donc imaginer que si cette hypothèse n'est pas vérifiée, la relation entre le polymorphisme mesuré sur les locus polymorphes et le polymorphisme global ne soit pas la même que dans les données simulées. Autrement dit, le modèle supposant une répartition homogène du polymorphisme pourrait introduire des biais dans l'estimation du polymorphisme quand il ne correspond pas aux données.

1.3. Discussion

1.3.1. Fréquence des discordances causées par des introgressions mitochondriales

Les introgressions mitochondriales sont à l'origine de discordances entre temps de divergences mitochondriaux et nucléaires. Notamment, un flux de mitochondries d'une espèce donneuse vers une espèce receveuse peut conduire au partage de mitochondries proches par des individus issus de deux espèces ou populations différenciées. Nous avons cherché à déterminer si certaines des paires de spécimens étudiées, possédant des séquences mitochondriales similaires, étaient associées à des divergences nucléaires discordantes par rapport à leurs séquences mitochondriales. Nous avons donc mesuré des distances nucléaires sur des marqueurs RADseq entre ces paires de spécimens. Les divergences obtenues ont montré que les 29 paires étudiées correspondaient à des divergences nucléaires faibles, compatibles avec les distances observées classiquement entre individus au sein d'une espèce.

Afin d'identifier des paires associées à des introgressions mitochondriales entre populations plus faiblement différenciées, nous avons ensuite souhaité déterminer si une structuration entre individus pouvait être mise en évidence, c'est à dire identifier des cas correspondant à l'échantillonnage de deux individus issus de populations divergentes. La comparaison de l'hétérozygotie mesurée au sein des individus et des distances entre individus, dxy, n'a pas permis d'observer de paires associées à une distance dxy, élevée et clairement supérieure aux hétérozygoties de ces individus. Ainsi, même si les conclusions de ces comparaisons sont limitées par le fait que l'estimation de l'hétérozygotie semble affectée par un biais à ce jour mal compris, il semble qu'aucune de ces paires ne montre de signe de discordance entre divergence nucléaires et mitochondriales.

La comparaison qualitative utilisée ici ne permet pas de tester formellement l'existence d'une différenciation au sein de chaque paire. Toutefois, une méthode ABC proche de celle présentée au chapitre précédent devrait permettre d'effectuer un tel test. En effet, nous avons jusque là pu utiliser cette approche pour mesurer le polymorphisme au sein de populations panmictiques, à partir de données de RADseq. D'autre part, nous avons montré que la méthode de simulation utilisée dans ce cadre permettait de modéliser l'obtention de données de RADseq à partir de spécimens issus de deux populations différenciées. Dans un avenir proche, une approche bayésienne approximée devrait ainsi pouvoir être développée, permettant de déterminer à partir des données de RADseq obtenues pour deux spécimens si le temps de divergence entre les populations dont ils sont issus est supérieur à 0, ou s'ils proviennent d'une population panmictique.

Aucune des 29 paires étudiées n'est associée à une discordance entre diversité mitochondriale et nucléaire. L'intervalle de confiance à 95% de la fréquence de ces discordances, compte tenu du nombre de paires testées (0 paires discordantes sur 29), est compris entre 0 et 0.15. L'absence d'introgression détectées dans cette étude ne permet donc pas d'exclure que les événements d'introgression mitochondriale soient suffisamment fréquents pour constituer un obstacle important à l'utilisation de marqueurs mitochondriaux pour retracer l'histoire des populations. L'ajout de nouvelles comparaisons à ces données sera nécessaire pour réduire cet intervalle de confiance. Notons d'autre part que les espèces étudiées appartiennent à des groupes taxonomiques

restreints (Lépidoptères et Diptères); on peut donc également se demander si cette observation est représentative de l'ensemble des Arthropodes.

Afin d'obtenir de nouvelles données permettant une meilleure estimation de la fréquence des évènements d'introgression, deux nouvelles librairies RADseq sont en cours de réalisation. Ces librairies comprennent 380 spécimens correspondant d'une part à 171 nouvelles espèces et d'autre part à 37 spécimens n'ayant pas données de résultat dans la première librairie du fait d'un problème technique lié aux barcodes utilisés pour ces spécimens. Ainsi, cette librairie permettra l'étude de 208 nouvelles espèces, s'ajoutant au 51 déjà disponibles. Ces nouvelles données représenteront l'échantillonnage de 54 espèces de Lépidoptères et 48 espèces de Diptères supplémentaires, ainsi que de 41 espèces d'Hyménoptères, 19 espèces d'Hémiptères et 9 espèces de Psocoptères. Cette nouvelle librairie devrait permettre de mieux estimer la fréquence des discordances entre divergences des génomes mitochondriaux et nucléaires. Nous pourrons ainsi mieux évaluer si les introgressions mitochondriale, sont suffisamment fréquentes pour limiter la pertinence des marqueurs mitochondriaux en taxonomie, génétique des populations et phylogénie.

Une autre question à considérer est celle de la représentativité de l'échantillon SymbioCode pour estimer la fréquence des évènements d'introgression entre espèces d'Arthropodes. En effet, ces échantillons proviennent tous de communautés insulaires fortement isolées des populations continentales. On peut se demander, par exemple, si la fréquence des introgressions dans ces îles pourrait être diminuée par la faible diversité de ces communautés. En effet, certaines de ces populations correspondent à des espèces invasives pour lesquelles aucune espèce proche n'existe localement. On peut donc imaginer que ces populations ont moins d'occasions d'hybridations que des populations continentales retrouvées fréquemment en sympatrie avec des espèces proches. Des estimations de la proportion d'espèces invasives et d'espèces endémiques dans ces communautés sont en cours. Elles devraient permettre de mieux comprendre les particularités de ces communautés par rapport aux communautés continentales.

1.3.2. Comparaison des temps de coalescence mitochondriaux et nucléaires au sein des populations ?

Nous avons pu estimer la fréquence des introgressions mitochondriales parmi les 29 paires de spécimens échantillonnées. Un autre type de discordances entre divergences mitochondriales et nucléaires pourrait être causé par des balayages sélectifs sur les génomes mitochondriaux au sein des populations. De tels balayages peuvent, par exemple, être causés par des infections par des symbiotes cytoplasmiques. Ils correspondraient ainsi également à des réductions des temps de coalescence mitochondriaux au sein des populations par rapport aux temps de coalescences moyens des locus nucléaires. On pourrait donc envisager de comparer les temps de coalescences mitochondriaux et nucléaires pour chacune des paires étudiées, afin d'estimer la fréquence de ces événements de balayages sélectifs affectant les génomes mitochondriaux.

Cette comparaison implique la prise en compte de différents paramètres responsables de différences de polymorphisme entre génomes mitochondriaux et nucléaires. Tout d'abord, le mode de transmission uniparental des mitochondries est associé à une réduction d'un facteur 4 de l'effectif efficace des populations mitochondriales par rapport au Ne nucléaire, dans le cas de populations au sex-ratio équilibré. De plus, indépendamment des différences de temps de coalescences, le polymorphisme mitochondrial est généralement plus élevé que le polymorphisme nucléaire, principalement en raison d'un taux de mutations plus élevé.

Nous pourrions néanmoins calculer la probabilité associée à l'échantillonnage d'un locus mitochondrial de séquence identique chez deux spécimens, en fonction de la proportion de locus nucléaires identiques observés, que nous appellerons $p0_{nuc}$, et en supposant que l'accumulation de mutation dans les génomes mitochondriaux et nucléaire suivent des lois de poisson. Ce calcul doit également tenir compte des longueurs respectives de ces locus mitochondriaux et nucléaires (respectivement l_{mito} et l_{nuc}) ainsi que des rapports des taux de mutation nucléaires et mitochondriaux, μ_{mito} et μ_{nuc}:

$$p0_m = (p0_{nuc})^{\left(\frac{\mu_{mito}}{4* \mu_{nuc}}\right)*\left(\frac{l_{mito}}{l_{nuc}}\right)}$$

Dans le cas présent, on ne connaît pas la valeur du rapport des taux de mutation mitochondriaux et nucléaires, μ_{mito}/μ_{nuc}. Nous pouvons toutefois évaluer la probabilité de l'échantillonnage de séquences mitochondriales identiques pour des valeurs hypothétiques plausibles de ce rapport. D'autre part, les locus nucléaires ont des longueurs de 89 paires de bases (l_{nuc}=89 pb), tandis que le locus mitochondrial utilisé à une longueur de 660 paires de bases, ce qui correspond à environ 220 positions synonymes (l_{mito}=220 pb). La figure 48 représente ainsi les probabilités associées à l'échantillonnage de séquences CO1 identiques, compte tenu de la proportion de locus RAD identiques et pour différentes valeurs des rapports des taux de mutation mitochondriaux et nucléaires.

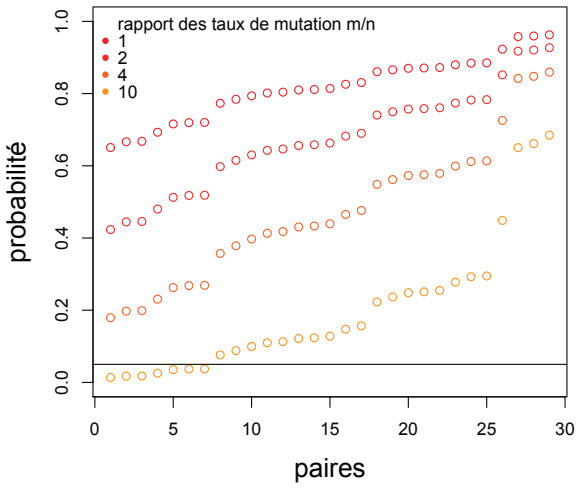

Figure 48. Pour chaque pair de spécimens, probabilités associées à l'échantillonnage de deux séquences CO1 identiques compte tenu des distributions des distances entre locus RAD, pour différents rapport de taux de mutations entre génomes mitochondriaux et nucléaires.

On peut remarquer que si les taux de mutation mitochondriaux sont 10 fois supérieurs aux taux de mutations nucléaires, la probabilité associée à l'échantillonnage de séquence CO1 identiques est inférieure à 5% pour 7 paires. Ces 7 paires sont associées au polymorphisme RAD le plus élevé, compris entre 0.9% et 1.7%. Toutefois, si nous utilisons une correction de Bonferroni pour tenir compte des tests multiples effectués dans cette expérience, le seuil de significativité devrait être de 0.0017. Aucune

paire ne correspond à une probabilité inférieure à ce seuil. De plus, ce test correspond à la probabilité associée à l'échantillonnage de deux séquences CO1 choisies aléatoirement au sein de ces espèces. Or, ces paires de spécimens ont été échantillonnées en fonction de la similarité de leur séquence CO1. Il s'agirait donc de comparer le polymorphisme mesuré sur les locus RAD non pas au polymorphisme observé sur ces deux spécimens choisis de façon biaisée, mais au polymorphisme moyen associé à ce locus, CO1, dans chaque espèce. Toutefois, dans la mesure où le test réalisé ne permet pas de mettre en évidence de paires associées à un polymorphisme mitochondrial significativement réduit avec le modèle utilisé, la comparaison du polymorphisme mitochondrial mesuré sur l'ensemble des spécimens de chaque espèce ne permettrait pas non plus, à plus forte raison, de détecter de réduction du polymorphisme mitochondrial dans ces espèces.

1.4. Conclusion

Des études de cas ont établi que les introgressions mitochondriales peuvent être favorisées par les infections par des symbiotes cytoplasmiques. En effet, le caractère invasif de ces symbiotes peut entrainer la propagation par auto-stop des mitochondries des spécimens infectées. Ainsi, l'introduction d'une infection au sein d'une espèce receveuse suite à une hybridation peut s'accompagner de l'introgression d'une mitochondrie exogène. Ces évènements d'introgression sont à l'origine de discordances entre génomes mitochondriaux et nucléaires. Si la fréquence de ces événements est importante, ils peuvent constituer un obstacle important à l'utilisation de marqueurs mitochondriaux pour retracer l'histoire des populations, et également perturber la coévolution des génomes nucléaires et cytoplasmiques. Nous avons évalué la fréquence des discordances entre divergences mitochondriales et nucléaires, causées par des évènements d'introgression. Sur 29 paires de spécimens étudiées, aucune discordance n'a pu être détectées entre divergence mitochondriale et nucléaire. Le nombre réduit de paires étudié ne permet cependant pas d'exclure la possibilité que cette fréquence soit relativement élevée (moins de 5% de chances d'excéder les 15%). Les nouveaux échantillons en cours de séquençage devraient nous permettre bientôt d'avoir une estimation plus précise de la fréquence d'évènements d'introgression.

2. Les *Wolbachia* affectent-elles la diversité mitochondriale des hôtes ?

La propagation d'une bactérie endosymbiotique peut entrainer par auto-stop un balayage sélectif sur les génomes mitochondriaux. Ces balayages sélectifs devraient se traduire par des réductions du polymorphisme mitochondrial, liées aux infections. Dans cette partie, nous avons cherché à mesurer cette réduction dans de nombreuses espèces d'Arthropodes de statuts d'infection différents.

2.1. Introduction : Dynamique d'invasion de Wolbachia et effets attendus sur le polymorphisme des mitochondries

L'impact d'une invasion par *Wolbachia* sur la diversité mitochondriale a été observé, par exemple, chez *Drosophila simulans* en Californie (Turelli and Hoffmann 1991), où l'invasion d'une souche de *Wolbachia* a conduit à la propagation du variant mitochondrial associé à l'infection. Cette sélection indirecte, par auto-stop, se traduit ainsi dans un premier temps par une réduction de la diversité mitochondriale des spécimens infectés par rapport à celle des spécimens non-infectés. Cela est dû au fait que toutes les mitochondries des spécimens infectés ont un ancêtre commun récent, qui correspond à la mitochondrie associée à l'infection initiale. La figure 49 (1) montre ainsi la propagation d'une lignée mitochondriale liée à un cytoplasme infecté, au sein d'une espèce.

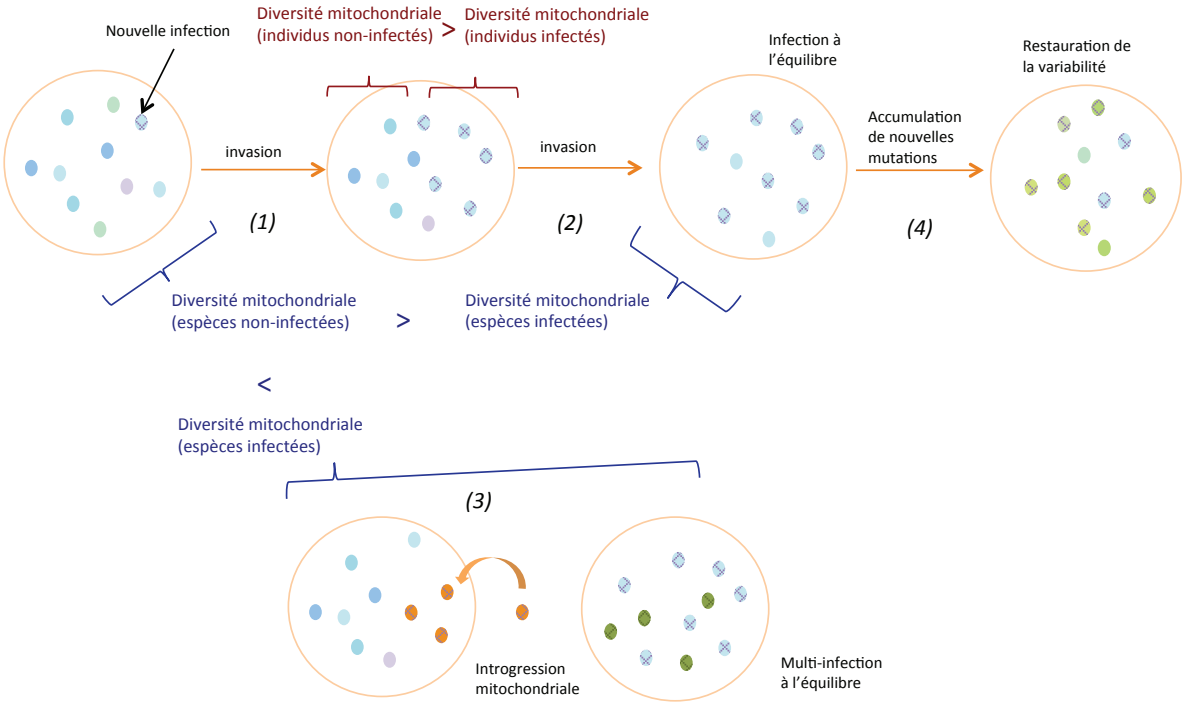

Figure 49. Dynamique d'invasion de *Wolbachia* et impact sur la diversité mitochondriale des hôtes.

La propagation d'une infection peut ensuite conduire à sa fixation et donc à celle de la lignée mitochondriale associée (Caspari et Watson 1959). L'infection peut également atteindre une fréquence d'équilibre qui est déterminée par le taux de transmission et de perte de la bactérie ainsi que par les valeurs sélectives (*fitness*) relatives des lignées infectées et non-infectées et par l'intensité, ou pénétrance, des phénotypes induits par *Wolbachia*. Cet état d'équilibre entre sélection et transmission imparfaite de *Wolbachia*, correspond à un équilibre dans lequel la fréquence d'individus infectés ne change plus de génération en génération. Des études théoriques ont montré que cet d'équilibre correspond à la fixation de l'haplotype mitochondrial initialement associé à l'infection, y compris chez les individus non infectés (Fine 1978). Brièvement, ce résultat s'explique par le fait que des individus non infectés sont produits à chaque génération par des mères infectées (transmission imparfaite). Ainsi, à l'état d'équilibre, tous les spécimens non-infectés ont des mitochondries issues d'une perte de l'infection. Cela signifie que toutes les mitochondries ont alors pour ancêtre commun l'haplotype mitochondrial initialement associé à l'infection, comme on peut l'observer sur la figure

49 (2). La diversité mitochondriale de la population est donc réduite par rapport à celle de la population ancestrale. Mais dans cet état d'équilibre de la fréquence de *Wolbachia*, l'infection n'est plus liée à une différence de diversité au sein des espèces entre les individus infectées et non infectées. En revanche, elle affecte l'ensemble des individus de la population qui a donc une diversité mitochondriale globalement réduite par rapport à une population non-infectée.

On peut noter que dans certains cas, la présence d'infections par *Wolbachia* et la propagation des mitochondries associées peut conduire à une augmentation et non a une diminution de la diversité mitochondriale au sein de l'espèce concernée. C'est par exemple le cas lorsque l'introduction de l'infection est due à une introgression incomplète. Comme nous l'avons vu au cours de la partie précédente, dans ce cas l'introduction de l'infection peut être concomitante de l'introduction d'un haplotype mitochondrial fortement divergent par rapport aux mitochondries préexistantes dans cette population. C'était le cas par exemple pour *Hypolimnas bolina* où l'introduction d'une nouvelle infection était liée à l'introgression d'une mitochondrie très divergente des mitochondries ancestrales, ce qui conduit à l'observation d'une diversité mitochondriale élevée dans cette espèce (Charlat et al. 2009). Une augmentation de la diversité peut également avoir lieu dans un cas où plusieurs infections coexistent dans une population spatialement structurée. Chaque infection peut alors être associé à une population de mitochondries homogène au sein de la sous-population, mais divergente de celles associées aux autres infections. Chez *D. simulans* par exemple, cinq infections par *Wolbachia* différentes sont connues, dont trois sont responsables d'incompatibilités cytoplasmiques et associées à des variants mitochondriaux fortement divergents. *Wolbachia* peut ainsi générer des incompatibilités entre populations qui maintiennent un grand polymorphisme mitochondrial. La figure 49 (3) illustre ces deux possibilités qui conduisent à des augmentations de la diversité mitochondriale au sein d'espèces infectées.

Enfin, dans le cas d'infections anciennes, de nouvelles mutations peuvent s'être accumulées au niveau des génomes mitochondriaux, jusqu'à restaurer la diversité des mitochondries à des niveaux comparable à celle des espèces non infectées. Ainsi, dans ces populations, l'ancêtre commun, qui correspond à l'haplotype mitochondrial associé à l'infection initiale, est ancien. Le temps de coalescence des génomes mitochondriaux actuels est donc égal ou inférieur à l'âge de cet ancêtre commun, c'est à dire à l'âge de

l'invasion par *Wolbachia*. Comme le montre la figure 49 (4), dans cette situation, on ne s'attend donc plus à observer de différences de diversité mitochondriale entre espèces infectées et non-infectées, à moins que la présence d'une bactérie ne cause des balayages sélectifs récurrents sur les cytoplasmes dans ces populations. Ainsi, l'observation d'un polymorphisme moyen similaire entre les espèces infectées et non-infectées pourrait signifier que la grande majorité des infections sont anciennes et que la diversité des génomes mitochondriaux est à l'équilibre entre mutations et dérive.

L'impact des infections par *Wolbachia* sur la diversité mitochondriale des espèces hôte dépend donc de la fréquence des évènements d'invasions, de pertes, et de remplacement. Si le taux de renouvellement des infections est faible, c'est-à-dire si les infections sont stables et rarement acquises par transfert horizontal, l'effet global du statut d'infection sur la diversité mitochondriale sera vraisemblablement modéré. En revanche, si les transferts horizontaux et les épisodes d'invasion sont fréquents, on observera un impact marqué de la présence de *Wolbachia* sur le polymorphisme.

La dynamique d'invasion de *Wolbachia* suggère, par ailleurs, que ces balayages sélectifs auront un impact différent sur les diversités mesurées au sein des espèces selon l'âge et le stade de l'infection. La comparaison du polymorphisme mesuré sur les individus infectés et non-infectés au sein des espèces, permet ainsi de détecter un effet de *Wolbachia*, si une proportion importante des espèces infectées sont à un stade précoce d'invasion. A ce stade, la diversité mitochondriale mesurée sur les individus non infectés est celle de la population initiale, avant infection, tandis que seuls les individus infectés sont affectés par le balayage sélectif. A un stade plus avancé de l'invasion, le balayage sélectif sur les mitochondries affecte généralement l'ensemble des mitochondries au sein de l'espèce infectées. La comparaison du polymorphisme des espèces infectées et non infectées devrait donc permettre de détecter un effet de *Wolbachia* même dans le cas où les infections sont à l'équilibre. Enfin, en comparant le polymorphisme des spécimens infectés au sein des espèces concernées, à celui des espèces non-infectées, il devrait être possible de détecter des réductions de polymorphisme dans les deux cas, à condition que les infections soient suffisamment récentes pour que la diversité mitochondriale n'ait pas été rétablie par l'accumulation de nouvelles mutations.

2.2. Résultats

2.2.1. Diversité mitochondriale des espèces de SymbioCode

Dans le but de mesurer l'impact de propagations d'infections par *Wolbachia* sur le polymorphisme mitochondrial, j'ai tout d'abord mesuré la valeur de la diversité nucléotidique mitochondriale au sein des différents taxons. J'ai utilisé pour cela des données de séquences mitochondriales obtenues pour 3600 spécimens représentant environ 1000 espèces d'Arthropodes. Ces données, correspondent aux séquences d'un fragment de 660 paires de bases du gène CO1 et ont été obtenues avant le début de ma thèse. Ces séquences ont permis de définir des unités taxonomiques opérationnelles (*OTU*) correspondant à une divergence maximale de 3% (sur toutes les positions) entre spécimens au niveau du marqueur CO1. Parmi les OTUs ainsi définies, 611 contiennent au moins 2 spécimens. La diversité nucléotidique mitochondriale de ces OTUs correspond à la distance moyenne entre paires de séquences. J'ai calculé ces diversités nucléotidiques sur les 3eme positions des codons afin d'obtenir une mesure de la diversité proche de celle correspondant aux sites synonymes uniquement. On peut remarquer que dans cet échantillon, π est calculé pour chaque espèce sur un faible nombre de spécimens (entre 2 et 48, médiane de 3 spécimens et moyenne de 5 spécimens). Certaines de ces mesures sont donc affectées par une forte incertitude due à ce faible échantillonnage. Toutefois, l'intérêt de ces comparaisons repose sur le nombre important de taxons étudiés. Même si l'incertitude sur chaque estimation est forte, le grand nombre d'estimations devrait permettre de mesurer des effets globaux de façon assez sensible.

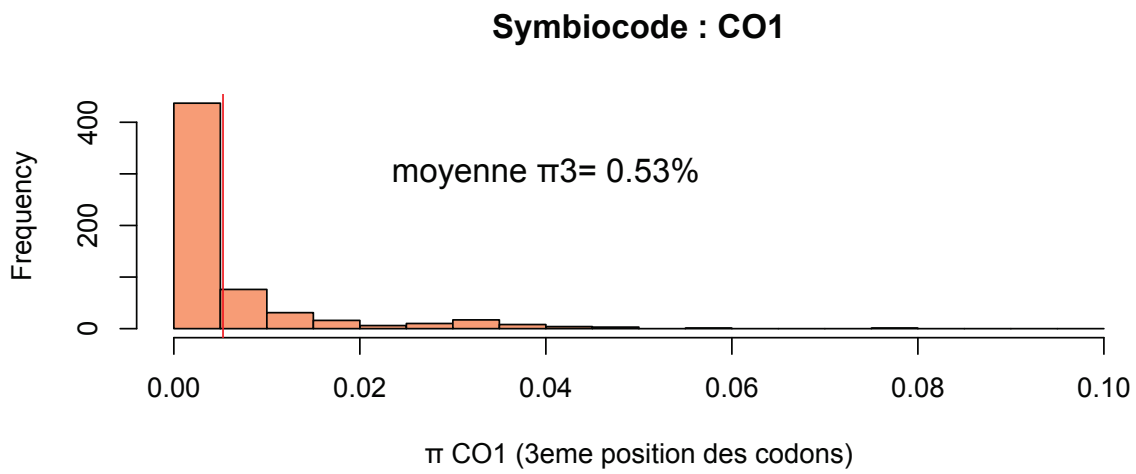

Figure 50. Distribution des diversités nucléotidiques, $\pi 3$, calculées sur les positions 3 des codons au sein des 611 OTUs SymbioCode pour lesquels au moins 2 spécimens ont été échantillonnés.

La figure 50 représente la distribution des diversités nucléotidiques mitochondriales synonymes pour chaque OTU. Cette diversité est en moyenne de 0.5%, ce qui correspond à une valeur plutôt faible par rapport aux mesures de diversité mitochondriales habituellement mesurées au sein des espèces (Bazin et al. 2006). Notons que la définition même de ces OTUs (groupes de spécimens dont aucun ne diverge d'un autre de plus de 3% sur l'ensemble de la séquence CO1) place une borne supérieure à cette diversité, même si la diversité nucléotidique aux 3èmes positions peut excéder 3%.

2.2.2. Impact des invasions récentes par *Wolbachia* sur les diversités mitochondriales au sein des espèces

Nous avons tout d'abord souhaité mesurer l'impact des invasions récentes de *Wolbachia* sur les polymorphismes mitochondriaux mesurés au sein des espèces. Les balayages sélectifs associés aux invasions affectent en premier lieu la diversité des spécimens infectés, affectés par le balayage, par rapport à la diversité des spécimens non-infectés. Il s'agit donc ici de comparer le polymorphisme mitochondrial des spécimens infectés par rapport à celui des spécimens non-infectés au sein des espèces

infectées. Sur 1089 OTUs, 78 comprennent un nombre suffisant de spécimens infectés et non-infectés pour permettre de mesurer des diversités nucléotidiques au sein de ces deux catégories d'individus.

Nous avons remarqué que 16 de ces OTUs sont infectés par plusieurs souches différentes de *Wolbachia*. Cela signifie que ces OTUs regroupent des spécimens infectés par des *Wolbachia* associées à des haplotypes de FbpA différents. Dans ces cas, on peut envisager que les spécimens associés à chacune des infections soient affectés par des balayages sélectifs différents. Mais la distance entre infections pourrait être importante, comme dans le cas de *D. simulans* ou des souches de *Wolbachia* distinctes sont présentes dans des régions différentes et associées à des lignées cytoplasmiques très divergentes. Dans cette étude, nous avons donc tenu compte de la nature des souches de *Wolbachia* responsables des infections. Au sein de chaque OTUs, j'ai mesuré la diversité nucléotidique parmi les spécimens infectés pour chaque souche de *Wolbachia* indépendamment. La diversité nucléotidique totale des spécimens infectés dans chaque OTU correspond ensuite à la moyenne de ces valeurs par infection.

Certains infections, identifiées par PCR 16S, n'ont pas pu être caractérisée par l'obtention d'une séquence FbpA. Cette échec du séquençage du marqueur FbpA, réalisée par la méthode de Sanger, est généralement due à la présence de multi-infections, c'est-à-dire de plusieurs souches de *Wolbachia* différentes au sein d'un individu. Dans ces situations, l'ensemble de ces spécimens associés à une infection « non identifiée » au sein d'une OTU, ont donc été considérés comme associés à une même infection. Toutefois, notons qu'une analyse écartant ces spécimens dont l'infection n'est pas clairement identifiée aboutit aux mêmes résultats que ceux présentés dans cette partie.

La figure 51 montre les distributions des diversités nucléotidiques mesurées de cette façon sur les individus infectés uniquement et sur les individus non infectés uniquement au sein des espèces pour lesquelles au moins 2 spécimens sont infectés. La comparaison des $\pi 3$ des individus infectés et non infectés au sein de chaque espèce ne permet pas d'observer de réduction significative du polymorphisme associée aux infections (test de wilcoxon apparié, p-value = 0.25).

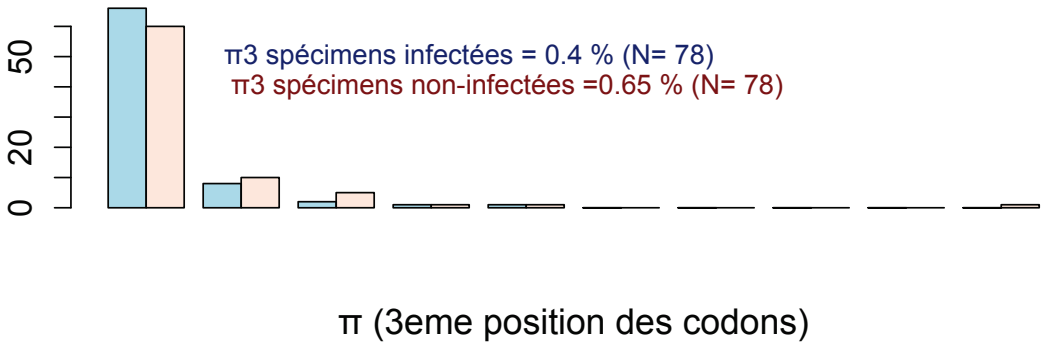

Figure 51. Distributions des diversités nucléotiques, $\pi 3$, des spécimens infectées (histogramme bleu) et des spécimens non-infectées (histogramme rose) au sein des OTUs infectées. La diversité nucléotidique est calculée sur les 3eme positions des codons d'un fragment du gène mitochondrial CO1.

Cette absence de différence entre le polymorphisme des spécimens infectés et celui des spécimens non-infectés au sein des OTUs pourrait s'expliquer par le fait qu'une faible proportion de ces infections soit en phase d'invasion. En effet, au sein des espèces où l'infection a atteint une fréquence d'équilibre, la réduction du Ne mitochondrial affecte l'ensemble des individus, on ne s'attend donc pas à observer de différence de polymorphisme sur ces espèces. Si la majorité des espèces sont dans cette situation, le polymorphisme des individus infectés peut ne pas être globalement différent de celui des individus non infectés. En revanche, dans ces cas le polymorphisme des espèces infectées devrait être réduit par rapport à celui des espèces non-infectées. Nous avons donc également recherché cette éventuelle différence de polymorphisme entre espèces infectées et non-infectées.

2.2.3. Impact des infections par *Wolbachia* sur les diversités mitochondriales des espèces infectées par rapport aux espèces non-infectées

Dans une espèce associée à une infection dont la fréquence est à l'équilibre, toutes les mitochondries sont affectées par le balayage sélectif lié à l'invasion. Toutefois, ces espèces devraient avoir une diversité réduite par rapport à leur diversité ancestrale,

tant que de nouvelles mutations ne se sont pas accumulées. La comparaison de la diversité nucléotidiques des espèces infectées et de celle des espèces non-infectées pourrait donc permettre de détecter une réduction de $\pi 3$ dans les espèces infectées.

Les relations phylogénétiques entre ces OTUs pourraient introduire un biais dans la comparaison des diversités des OTUs infectées et non-infectées. En effet, l'inertie phylogénétique affecte les mesures de corrélations entre traits mesurés sur plusieurs taxons car les relations de parenté entre ces taxons peuvent interférer avec la corrélation entre les traits. Par exemple dans le cas de la mesure de la relation entre la diversité nucléotidique et les infections par *Wolbachia*, on pourrait imaginer l'existence de larges groupes taxonomiques où une majorité d'espèces présenteraient la fois de faibles diversités nucléotidiques (pour des raisons indépendantes des infections cytoplasmiques) et une forte incidence de *Wolbachia*. L'existence de cette relation, liée à la parenté entre ces taxons et à l'inertie phylogénétique des traits étudiés, pourrait conduire à détecter une corrélation significative entre ces traits, pourtant sans relation causale. A l'inverse, l'effet opposé pourrait obscurcir la relation entre les deux traits: par exemple, de fortes variations des taux de mutations selon les taxons pourraient affecter les diversités nucléotidiques et masquer un éventuel effet des infections par *Wolbachia*. Une façon de tenir compte des relations phylogénétiques entre taxons est d'utiliser des comparaisons indépendantes (Felsenstein 1985). Dans le cas de l'étude de l'impact de *Wolbachia* sur la diversité mitochondriale, il s'agirait par exemple de mesurer l'écart entre la diversité de paires d'espèces sœurs, l'une étant infectée et l'autre pas. Nous pourrions par la suite déterminer si dans un plus grand nombre de comparaisons, l'espèce infectée est associée à une diversité plus faible. Cette méthode de comparaison appariée permet ainsi de tenir compte des effets de l'apparentement entre espèces sur les variations des traits étudiés.

Au lieu de comparer uniquement des espèces sœurs, et afin de maximiser la quantité de données utilisable, nous avons comparé les $\pi 3$ moyens des OTUs infectées aux $\pi 3$ moyens des OTUs non-infectées au sein de clades correspondant à 20% de divergence CO1 maximum entre spécimens, par un test de rang apparié. L'échantillon comporte 538 de ces taxons correspondant à une divergence CO1 maximale de 20%. 36 de ces taxons contiennent au moins une OTU infectée et une non-infectée, chacune comprenant plus de 1 spécimen (Figure 52). Un test de rang apparié n'a pas permis de

détecter de différences significative entre la diversité des OTUs infectées et non-infectées (*Wilcoxon signed rank test*, p-value = 0.158).

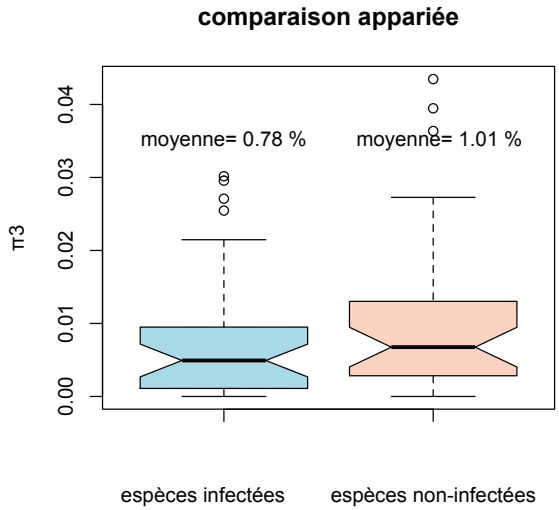

Figure 52. Distribution des diversités nucléotiques moyennes mesurées au sein des OTUs infectées (au moins un spécimen infecté) ou non-infectées, pour 35 clades correspondant à 20% de divergence mesurée au niveau du gène CO1, et comprenant au moins une OTU infectée et une OTU non-infectée.

La méthode utilisée pour tenir compte des relations phylogénétiques entre taxons réduit fortement le nombre d'OTUs utilisées par ce test. Toutefois, notons qu'une comparaison globale et non appariée (entre les 240 OTUs non-infectées du jeu de données et les 370 OTUs infectées par une souche unique de *Wolbachia*) ne permet pas non plus d'observer de différences significatives des diversités nucléotidiques (*Welch Two Sample t-test*, p-value = 0.81, *Wilcoxon rank sum test*, p-value = 0.38).

2.2.4. Impact global du statut d'infection sur diversité mitochondriale ?

On peut remarquer que selon le stade de l'invasion, l'effet sur la réduction de la diversité ne sera pas observable à la même échelle. On peut donc supposer que si une forte proportion de espèces est en phase d'invasion, pendant laquelle la diversité est maintenue au sein des populations, ces dernières vont réduire l'écart moyen de diversité entre espèces infectées et non infectées. A l'inverse si dans une grande proportion de

taxons *Wolbachia* a atteint une fréquence d'équilibre, stade auquel la réduction de la diversité affecte l'ensemble des spécimens, l'écart de diversité entre spécimens infectés et non infectés au sein des espèces pourrait être masqué. Ainsi, la coexistence de taxons dans les deux types de situations dans ce jeu de donnée, pourrait expliquer qu'on ne détecte aucun des deux effets.

Afin d'évaluer l'impact global de *Wolbachia* sur la diversité mitochondriale, nous avons donc envisagé un test moins sensible à cette hétérogénéité des stades d'infection, en comparant la diversité des individus infectés des espèces infectées à celle des individus non infectés des espèces non infectées. Il s'agit ici de comparer la diversité des espèces non infectées à celle des individus qui sont toujours affectés par le balayage sélectif, quel que soit le stade de l'invasion de *Wolbachia*. Comme pour la comparaison des diversités des espèces infectées et non-infectées, ces comparaisons ont été effectuées de manière appariée. De plus, comme pour la comparaison des $\pi 3$ des spécimens infectés par rapport à ceux des spécimens non infectés, les $\pi 3$ ont été mesurés sur chaque souche de *Wolbachia* indépendamment. La diversité nucléotidique des spécimens infectés au sein d'une espèce infectée par plusieurs souches de *Wolbachia* correspond ensuite à la moyenne des diversités correspondant associées à chaque souche.

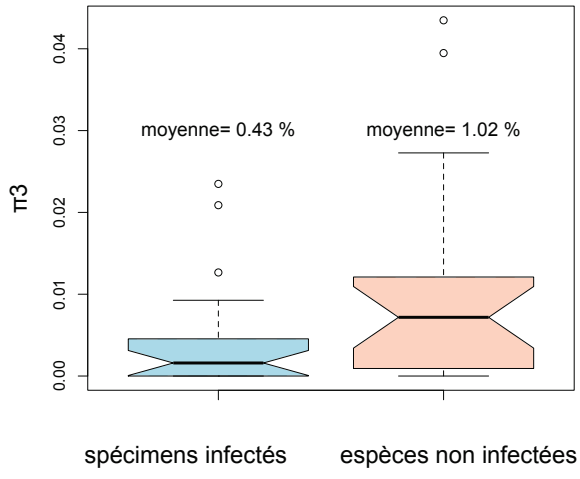

Figure 53. Distribution des diversités nucléotiques moyenne mesurées entre spécimens infectés au sein des OTUs infectées ou au sein des OTUs non-infectées, pour 22 clades correspondant à 20% de divergence mesurée au niveau du gène CO1.

La figure 53 montre la distribution des valeurs moyennes de π mesurées entre spécimens infectés des OTUs infectées et entre spécimens non-infectés des OTUs non-infectées, au sein de clades correspondant à 20% de divergence CO1. La comparaison appariée de ces mesure de diversité mitochondriale montre une diminution significative de la diversité mitochondriale des spécimens infectés par *Wolbachia* (*Wilcoxon signed rank test,* p-value = 0.008). Notons que cette p-value faible reste significative si l'on utilise un seuil tenant compte des 3 comparaisons effectuées.

Ce résultat suggère qu'il y a bien une relation entre la diversité mitochondriale et le statut d'infection par *Wolbachia*. Cette relation est compatible avec l'hypothèse selon laquelle les diversités mitochondriales sont affectées par des balayages sélectifs causés par *Wolbachia* dans les espèces infectées.

2.3. Discussion et conclusion

Des balayages sélectifs causés par la propagation d'infections par *Wolbachia* peuvent réduire la diversité génétique mitochondriale des espèces hôtes. Afin d'évaluer l'impact global de cette réduction, nous avons calculé et comparé les diversités nucléotidiques mitochondriales au sein de 611 espèces (OTUs) d'Arthropodes. Ces comparaisons ont permis de mesurer une réduction du polymorphisme d'un facteur 2 des spécimens infectés par rapport à celui des espèces non-infectées.

Nous avons également comparé, d'une part, le polymorphisme calculé sur les spécimens infectés ou non-infectés, au sein des espèces et d'autre part, le polymorphisme de l'ensemble des individus des espèces infectées à celui des espèces non-infectées. Dans les deux cas, les différences mesurées allaient dans le sens attendu, à savoir une diversité plus faible liée à la présence de *Wolbachia*, mais ces tendances n'était pas statistiquement significatives. Comme mentionné plus haut, ces deux résultats négatifs pourraient s'expliquer par l'hétérogénéité des stades d'invasion dans les différentes espèces infectées.

Pour explorer cette hypothèse, il devrait être possible, au moins pour certaines espèces au sein du jeu de données, de distinguer les différentes situations, c'est-à-dire de préciser les stades des infections (anciennes et à l'équilibre versus précoces et en phase d'invasion). Par exemple, une bactérie partagée par plusieurs espèces voisines est

plus probablement ancienne qu'une bactérie n'infectant que quelques spécimens au sein d'une OTU isolée. Ainsi, la connaissance des scénarios de pertes et d'acquisitions des infections pourrait permettre de mesurer l'effet des réductions de Ne causées par les infections par *Wolbachia* de façon plus fine. L'intégration des résultats de cophylogénies, établies entre la phylogénie des *Wolbachia* et des hôtes, qui seront présentés plus en détail au cours de la partie suivante de ce chapitre, devrait ainsi permettre une meilleure prise en compte des effets contrastés de *Wolbachia* sur le polymorphisme mitochondrial.

3. Les *Wolbachia* affectent-elles l'efficacité de la sélection sur les génomes mitochondriaux des espèces hôtes ?

Les infections par des symbiotes cytoplasmiques peuvent induire une sélection indirecte causant une réduction de la diversité génétique mitochondriale des populations infectées. Cette réduction correspond à une diminution des effectifs efficaces, Ne, des génomes mitochondriaux qui peut-être également responsable d'une réduction de l'efficacité de la sélection dans ces populations. Potentiellement, ces diminutions de Ne se traduisent, sur le long terme, par des modifications des patrons de substitution conduisant à une augmentation des rapports de taux de substitutions non-synonymes et synonymes, dN/dS dans les lignées infectées. C'est ce que nous allons chercher à mesurer dans cette partie.

3.1. Introduction : *Wolbachia*, Ne et dN/dS

Dans les séquences nucléotidiques codantes, le taux de substitution synonyme correspond au taux de substitution sur les sites dont la mutation ne change pas la séquence protéique, tandis que le taux de substitution non-synonyme correspond au taux de substitution des sites modifiant la séquence protéique. On considère généralement que les sites synonymes évoluent de façon principalement neutre, tandis que les sites non-synonymes peuvent être soumis à des pressions de sélections, positives ou négatives. Aussi, le rapport des taux de substitution non-synonyme, dN, et synonyme, dS, appelé $\omega = dN/dS$, est utilisé comme indicateur de la sélection naturelle. En absence de sélection, on s'attend à un ω de 1 ; c'est-à-dire à une évolution neutre des sites non-synonymes aussi bien que des sites synonymes. En revanche, les sites soumis à une sélection purifiante correspondent à un ω inférieur à 1, tandis qu'une sélection positive, diversifiante, se traduit par un ω supérieur à 1. Ainsi, à l'échelle d'une protéine, la mesure de ω sur l'ensemble des sites peut permettre d'estimer le niveau et le sens de la sélection à laquelle est, en moyenne, soumise la protéine, ou encore d'identifier des sites associés à un régime de sélection particulier au sein de la protéine.

Les tailles efficaces (Ne) des populations peuvent avoir un impact sur l'efficacité de la sélection, et par conséquent sur les taux de substitutions. Dans des populations de

tailles efficaces réduites, on s'attend en effet à ce qu'une plus grande proportion des mutations évolue de façon neutre. Cela signifie qu'une plus grande proportion des mutations légèrement délétères peut être fixée par dérive. A l'inverse, un plus grand nombre de mutations avantageuses peuvent être perdues par dérive. Etant donné que les mutations sont plus souvent délétères qu'avantageuses, la réduction de Ne doit donc se traduire par une augmentation des taux de substitutions non synonymes, sans effet sur les taux de substitution synonymes. En outre, on peut supposer que les mutations affectant les sites non-synonymes sont plus souvent délétères que celles affectant les sites synonymes. Ainsi, une réduction de Ne devrait généralement se traduire par une augmentation du taux de substitution non synonyme, et donc une augmentation du rapport ω (Woolfit 2009, Lanfear et al. 2014).

Nous avons vu précédemment que la propagation d'une infection par une bactérie endosymbiotique, en réduisant le Ne mitochondrial peut affecter le polymorphisme. Cette réduction de Ne peut donc également avoir un impact sur les taux de substitutions, qui conduirait à une augmentation du rapport ω sur les lignées infectées. Notons qu'à la différence des effets de *Wolbachia* sur le polymorphisme, explorés dans la section précédente, cet effet de la réduction du Ne mitochondrial sur les patrons de substitution serait mesurable sur les lignées infectées même après que les diversités génétiques aient été restaurées dans les populations actuelles.

Shoemaker et al. (2004) ont recherché un effet de *Wolbachia* sur la diversité mitochondriale et sur les taux de substitution mitochondriaux conformément à ces prédictions dans deux espèces de drosophiles. Ces auteurs ont ainsi observé une diminution de la diversité et une augmentation du rapport ω chez *Drosophila recens*, infectée par *Wolbachia*, par rapport à son espèce sœur non infectée, *Drosophila subquinaria*. Ces changements semblent affecter exclusivement le génome mitochondrial. En effet, trois marqueurs nucléaires, *period*, *adhr* et *tpi* analysés de la même façon ne permettent pas de mettre en évidence des différences similaires entre ces espèces. Ce résultat a donc été interprété comme un effet de l'infection par *Wolbachia* chez de *D. recens*.

On peut remarquer qu'une réduction de Ne mitochondrial dans cette lignée pourrait avoir d'autres causes que l'infection par *Wolbachia*. Les mitochondries de *D. recens* pourraient par exemple avoir subi un balayage sélectif récent indépendant de l'infection. D'autre part, si l'infection par *Wolbachia* est bien responsable de ce

changement de patron de substitutions au niveau du génome mitochondrial, on peut se demander si cet effet est fréquent ou si cette espèces constitue un cas isolé. Un patron similaire affecte-t-il suffisamment souvent les lignées infectées pour causer une augmentation globale des rapports ω sur les lignées infectés par *Wolbachia*?

Dans ce chapitre, j'ai cherché à mesurer l'existence de réductions de l'efficacité de la sélection, conséquences de diminutions des tailles efficaces mitochondriales par *Wolbachia*. J'ai pour cela comparé les rapports ω des lignées mitochondriales d'Arthropodes infectés ou non, au sein de la phylogénie des spécimens du système SymbioCode. Cette comparaison des lignées en fonction de leurs statuts d'infection implique la reconstruction de l'histoire des infections le long de l'arbre des hôtes et des statuts d'infection aux nœuds. Ces inférences ont été réalisées par Sylvain Charlat et Gergely Szöllosi par une réconciliation de l'arbre des hôtes et des *Wolbachia*. Ensuite, l'estimation d'un modèle d'évolution et le comptage (*mapping*) des substitutions sur l'arbre des hôtes a permis d'estimer des taux de substitution mitochondriaux synonymes et non-synonymes sur chaque branche. La comparaison des patrons de substitution entre lignées infectées et non infectées n'a pas permis mettre en évidence l'augmentation de ω attendue dans les lignées infectées. Les explications possibles de cette absence d'effet sont discutées, notamment en référence aux travaux antérieurs ayant suggéré une augmentation significative du dN/dS dans une espèce infectée.

3.2. Inférence des statut d'infections par *Wolbachia* sur les branches de l'arbre des hôtes

La comparaison des taux de substitution des lignées d'hôtes en fonction de leurs statuts d'infection implique la reconstruction de leurs états d'infection ancestraux. Tandis que des associations stables entre hôtes et symbiotes génèrent des phylogénies congruentes, les évènements de pertes et d'acquisitions par transfert horizontal génèrent des incongruences entre arbre des symbiotes et arbre des hôtes. Inversement, il est possible d'utiliser ces incongruences pour inférer des scénarios vraisemblables de pertes et acquisitions, et ainsi estimer les statuts d'infections aux noeuds internes de l'arbre des hôtes. Dans notre étude, la phylogénie des hôtes est basée sur le marqueur mitochondrial CO1 et celle des symbiotes sur une séquence du gène FbpA.

La réconciliation a été effectuée par Sylvain Charlat à l'aide du programme ALE (Szöllosi et al. 2013). Une des spécificités de ce programme est de tenir compte des âges relatifs des nœuds, ce qui facilite la reconstruction des scénarios de pertes, acquisitions et transferts, en particulier dans des cas comme celui-ci où les événements de pertes et acquisitions sont fréquents, multipliant le nombre de scénarios plausibles. D'autre part, ALE utilise non pas une phylogénie particulière pour l'arbre des symbiotes, mais recherche à la fois la meilleure phylogénie des symbiotes et le ou les meilleurs scénarios de réconciliation. Ainsi, ce programme intègre l'incertitude de la phylogénie des symbiotes dans l'estimation des scénarios de perte et acquisition. ALE permet de maximiser la vraisemblance des taux de transferts, de pertes et d'acquisition, qui correspond à la somme des vraisemblances des scénarios reconstruits pour un taux donné. Ce calcul est effectué pour 5000 arbres de *Wolbachia* échantillonnés en fonction de leur probabilité postérieure à l'issue d'une reconstruction Bayésienne. ALE ogramme permet ainsi d'obtenir une distribution des scénarios les plus vraisemblables en tenant compte de la vraisemblance de l'arbre et des taux correspondants. Un arbre CO1 ultramétrique, c'est à dire dont les longueurs de branches sont proportionnelles au temps et où toutes les séquences actuelles correspondent à un temps 0, des 3 600 spécimens du projet Symbiocode a été obtenu. Pour des raisons de limitations computationnelles, cet arbre a été coupé en 25 sous-arbres. Pour chacun de ces sous-arbres, ALE a permis de générer 10000 scénarios échantillonnés en fonction de leur vraisemblance (intégrant la vraisemblance des arbre de gènes, des taux de perte et acquisition, et du scénario). L'estimation des probabilités d'infections aux nœuds est ensuite obtenue à partir des moyennes des statuts d'infection des 10000 scénarios produits, pour chaque nœud de chaque sous-arbre.

3.3. Estimation des ω : Comptage (*mapping*) des substitutions synonymes et non-synonymes

Les infections par *Wolbachia* peuvent conduire à des réduction des Ne mitochondriaux qui pourraient, si elles sont suffisamment fortes et fréquents, se traduire par des augmentations du rapport ω des lignées infectées. Nous cherchons donc à comparer les ω des branches infectées à ceux des branches non-infectées. Pour cela,

les taux de substitutions synonymes (dS) et non-synonymes (dN) ont été estimés sur chaque branche de l'arbre CO1 par une méthode de comptage (*mapping*). Pour chacun des 25 sous-arbres utilisés pour la réconciliation de l'arbre des hôtes et des *Wolbachia* (partie précédente), un modèle d'évolution est déterminé par *bppml* (Bio++ *Maximum Likelihood*). *Bppml* permet d'estimer par maximum de vraisemblance les paramètres de modèles d'évolution de séquences associés à une topologie, imposée ou non (Dutheil and Boussau 2008). Pour chacun des sous-arbres, pour lesquels on dispose d'un alignement de séquence CO1 et d'une topologie, *bppml* a permis l'estimation des paramètres d'un modèle homogène. Ces paramètres correspondent aux compositions en bases des séquences à l'équilibre, au rapport des taux de transition et transversion à l'équilibre (κ=ts/tv) et à un ω global à l'équilibre, ainsi qu'aux compositions en bases à la racine.

Les substitutions synonymes et non-synonymes ont ensuite été « placées » et comptées sur les branches de l'arbre correspondant à ce modèle. *MapNH* (*TestNH*, Dutheil et al. 2012, Romiguier et al. 2012) permet d'effectuer ce comptage des différents types de substitutions sur un arbre phylogénétique à partir d'un alignement de séquences associé à un modèle d'évolution. Ce comptage permet donc d'obtenir pour chaque branche l'espérance du nombre de substitutions synonymes et non-synonymes. Contrairement à une approche où des taux de substitutions spécifiques sur chaque branche seraient estimés par maximum de vraisemblance, cette méthode permet donc à partir d'une estimation d'un taux unique correspondant à un modèle homogène, d'estimer *a posteriori* des taux de substitutions par branches. Cette estimation par comptage est ainsi plus rapide qu'une méthode nécessitant l'estimation d'un modèle non-homogène par maximum de vraisemblance (Figure 54).

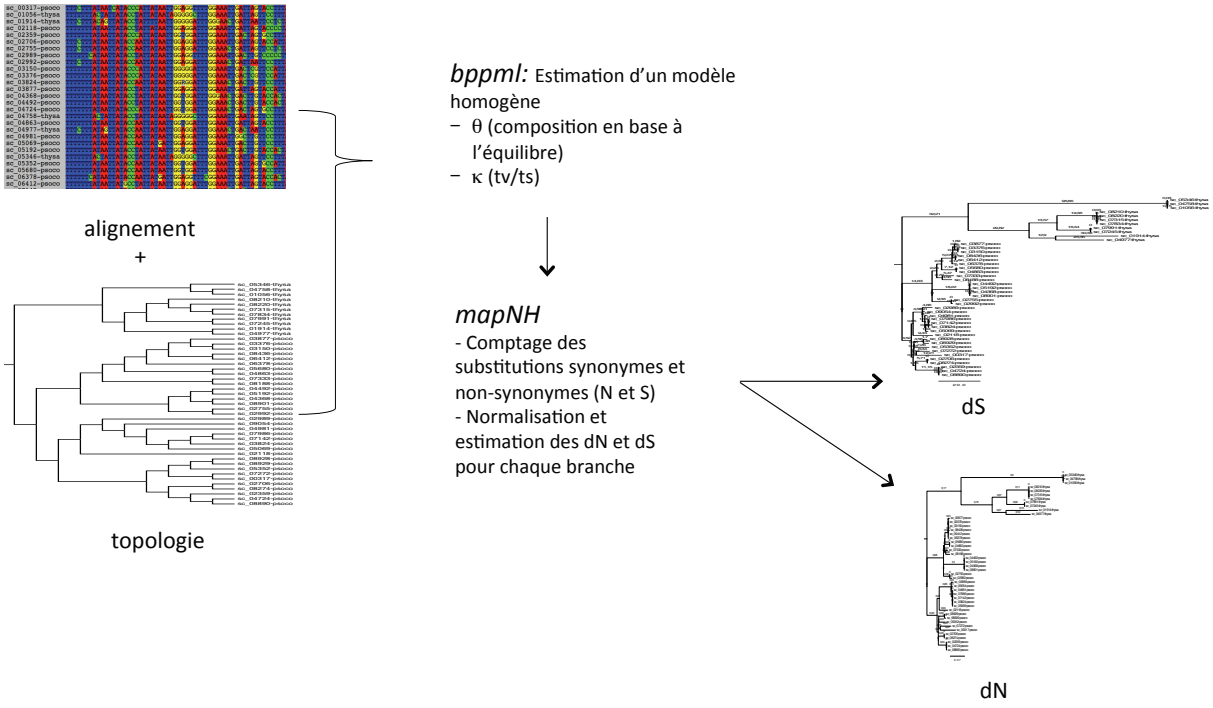

Figure 54. Estimation des taux de substitution synonymes et non-synonymes

Les taux de substitution synonymes et non-synonymes correspondent aux nombres de substitutions, mesurés par site synonyme ou non-synonyme. L'estimation des valeurs de dN et dS implique donc l'estimation non seulement des nombres de substitutions, mais également des nombres de sites synonymes et non-synonymes. Ces nombres de sites représentent en réalité les nombres de substitutions synonymes et non synonymes potentielles, en tenant compte de l'état d'une séquence à un moment donné le long d'une branche et d'un modèle de substitution. Cela permet une normalisation des nombres de substitutions synonymes et non synonymes en taux de substitutions synonymes et non synonymes pour chaque branche.

3.4. Relation entre ω et la probabilité d'infection par *Wolbachia*

3.4.1. Méthode : comparaison entre ω des branches infectées et non-infectées.

Ces estimations des taux de substitution par comptage permettent la mesure des rapports ω et la comparaison des rapports inférés sur les branches infectées à ceux des branches non infectées. Toutefois, les relations phylogénétiques entre les différents

taxons pourraient interférer avec la mesure de la corrélation entre statut d'infection et ω.

Par exemple, une variation des tailles efficaces dans la phylogénie pourrait masquer une corrélation entre patrons de substitutions et statut d'infection. Il est donc nécessaire ici encore de tenir compte de l'histoire commune plus où moins importante qui relie les différentes lignées étudiées. Comme nous l'avons déjà mentionné, différentes approches permettent d'étudier les corrélations entre traits mesurés sur des organismes en tenant compte de leurs relations de parenté, dont les méthodes de contraste. Une spécificité de notre étude, est que les traits étudiés sont mesurés sur les branches et non sur les individus aux feuilles de l'arbre. De plus, la plupart des paires de branches associées à des statuts d'infection différents sont courtes et donc associées à des taux de substitution très faibles et à des mesures de ω très imprécises.

La méthode que j'ai mise en oeuvre s'inspire de la méthode de contraste décrite précédemment. Il s'agit de mesurer des différences entre ω associés à des branches infectées et non infectées dans des sous-groupes indépendants. Mais au lieu de mesurer ces ω sur des branches sœurs uniquement, on utilise la somme des ω des branches infectées et non-infectées pour des clusters définis au sein de l'arbre global. Nous utilisons ici des clades dont la divergence maximale est de 20% sur le marqueur CO1. Ces clades sont suffisamment nombreux pour définir un grand nombre de points, et suffisamment grands pour contenir suffisamment de branches infectées et non infectées.

Pour chaque clade j, on mesure donc la somme des dN et la somme des dS des branches infectées et non-infectées et on en déduit les rapports $\omega_{inf}(j)$ et $\omega_{un}(j)$ (Figure 55). Les taux de substitutions correspondent aux nombres de substitutions par site de chaque catégorie. Ils dépendent donc de la longueur des branches associées. Les branches les plus courtes, dans lesquelles la mesure de ω est la plus imprécise, contribuent d'autant moins à la somme. D'autre part, chaque branche est associée à une probabilité d'infection aux deux extrémités. J'ai donc pondéré les sommes des taux de substitutions de chaque branche par leur probabilité d'être infectées. Ainsi, pour un sous-arbre j, on calcul les ω infectés et non infectés comme suit:

$$\omega_{inf}(j) = \frac{dN_{\mathrm{inf}(j)}}{dS_{\mathrm{inf}(j)}} = \frac{\sum_{k\epsilon j} dN_k * p_k}{\sum_{k\epsilon j} dS_k * p_k} \qquad (1)$$

$$\omega_{un}(j) = \frac{dN_{\mathrm{un}(j)}}{dS_{\mathrm{un}(j)}} = \frac{\sum_{k\epsilon j} dN_k * (1-p_k)}{\sum_{k\epsilon j} dS_k * (1-p_k)} \qquad (2)$$

Où P_k représente la probabilité de la branche k (appartenant au sous-arbre j) d'être infectée, et dN_k et dS_k représentent respectivement les taux de substitutions non synonymes et synonymes pour la branche k.

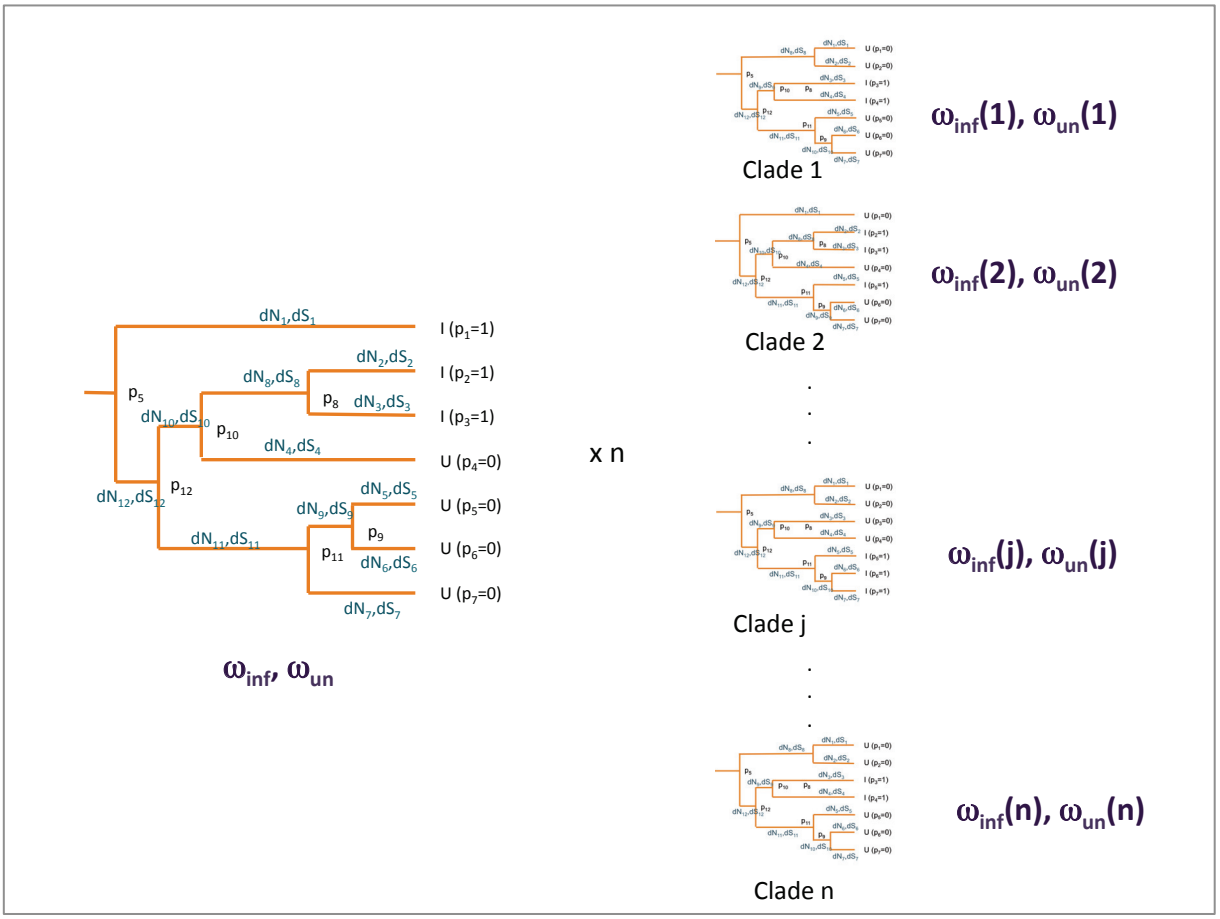

Figure 55. Calcul de ω_{inf} et ω_{un} pour chaque clade correspondant à une divergence CO1 maximale de 20%.

3.4.2. Résultat et discussion

Nous avons utilisé cette méthode pour comparer les valeurs de ω des lignées infectées par rapport à celles des lignées non-infectées dans l'arbre des spécimens de Symbiocode. Cet arbre comprend 514 clades de 20 % de divergence CO1 maximum. Parmi ces clades, nous avons sélectionné ceux comprenant des proportions

suffisamment équilibrées de lignées infectées et non-infectées pour autoriser une comparaison raisonnablement fiable. Ainsi, 120 clades contiennent au moins 10% de branches infectées et 10 % de branches non-infectées (10% de la longueur totale du sous-arbre), et ont été utilisés pour le suite de l'analyse. J'ai mesuré l'effet de la présence de *Wolbachia* sur ω en comparant de façon appariée des valeurs de ω_{inf} et ω_{un} dans ces différents clades. Le principe de ce test repose sur le fait que si les infections par *Wolbachia* résultent en une augmentation de ω sur les branches infectées, ω_{inf} sera significativement plus souvent supérieur à ω_{un}. Ce test permet donc de réduire l'influence de l'inertie phylogénétique en utilisant de nombreux groupes indépendants, tout en utilisant des valeurs de dN et dS mieux estimées que si l'on avait comparé uniquement des paires d'espèces, associées à des branches très courtes et donc à des dN et dS très faibles et donc des ω très imprécis. Comme indiqué sur la figure 56, les ω estimés pour les branches infectées et non infectées sont tous deux très faibles (médianes 0.6% et 0.7%) et leur différence non significative (test de wilcoxon apparié, p-val=0.138).

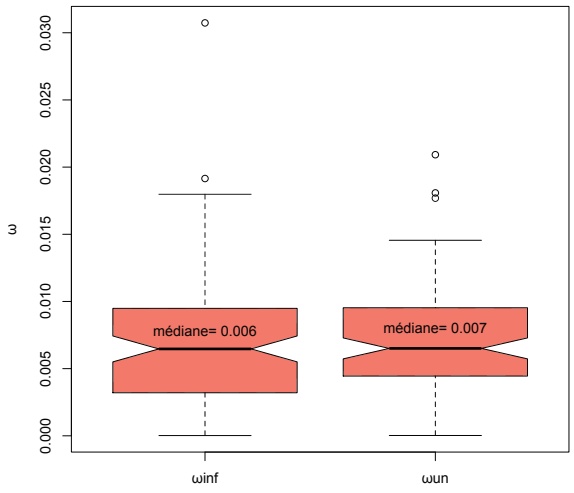

Figure 56: distribution des ω sur les branches associées ou non à une infection par *Wolbachia*, pour 106 clades de divergence CO1 de divergence maximale de 20 %. Test de Wilcoxon apparié, p-val=0.14.

Ce résultat semble indiquer que les infections par *Wolbachia* ne sont pas associées à une augmentation du ω spécifique des branches infectées au niveau du marqueur CO1 étudié. Il peut s'expliquer de différentes façons. Tout d'abord, il est possible que les infections par *Wolbachia* ne soient pas associées à des réductions de Ne suffisamment intenses ou suffisamment fréquentes pour induire des différences de rapports de taux de substitutions entre lignées infectées ou non. Une autre explication plausible serait que *Wolbachia* affecte bel et bien l'efficacité de la sélection, mais que cet effet soit masqué par des incertitudes concernant les statuts d'infection le long des branches. En effet, le test effectué repose sur l'identification des lignées infectées et non-infectées, qui comporte une forte incertitude puisqu'elle repose sur une réconciliation des phylogénies hôtes et symbiotes rendue difficile par des taux élevés de perte et acquisition des symbiotes. Enfin, le marqueur CO1 utilisé est vraisemblablement très contraint (ω très proche de 0). Une sélection très forte sur ce gène, empêchant la fixation de substitution non-synonymes même avec une réduction de Ne, pourrait expliquer qu'on ne détecte pas d'augmentation du ω, alors que cet effet aurait pu être détecté sur un autre marqueur. Le fait que ce marqueur soit court est peut-être également la cause d'une incertitude sur le calcul de ω, qui pourrait empêcher la détection des différences de ω entre lignées infectées et non-infectées.

Des études précédentes avaient permis de mettre en évidence des différences de patrons de substitutions entre deux lignées dont une était infectée par *Wolbachia* (Shoemaker 2004). Afin de mieux comprendre si l'absence de corrélation observée dans notre étude est due à un manque de puissance du marqueur utilisé ou à une absence d'effet dans notre jeu de données, nous avons ré-analysé les données présentées par Shoemaker et al 2004.

3.5. *Drosophila recens* et *Drosophila subquinaria*

Shoemaker et al. ont comparé la diversité mitochondriale et les taux de substitution entre deux espèces proches dont l'une est associée à une infection par *Wolbachia*. *D. recens* et *D. subquinaria* sont deux espèces de drosophiles mycophages, rencontrées respectivement au nord-est et au nord-ouest de l'Amérique du Nord. *D.*

recens est infectée par une *Wolbachia* responsable d'une incompatibilité cytoplasmique dans cette espèce, tandis qu'aucune infection n'est connue chez *D. subquinaria*.

L'étude de Shoemaker et al. 2004 avait porté sur la comparaison de valeurs de ω calculées sur 12 gènes mitochondriaux (7227 paires de bases), obtenus pour 12 *D. recens*, 1 D. *subquinaria* et 1 *D. quinaria*, qui constitue un groupe externe par rapport à celui formé par *D. recens* et D. *subquinaria*. Afin de reproduire les résultats de cette étude, j'ai tout d'abord estimé ω par *codeml* (PAML, Yang 1997, Yang 2007), sur l'alignement concaténé de 12 gènes mitochondriaux utilisé par Shoemaker et al. Trois modèles différents ont été utilisés pour ces estimations: (1) un ω unique pour les 3 espèces, (2) deux ω, un chez *D. recens* d'une part, et un chez *D. subquinaria* et *D. quinaria* d'autre part, et (3) un ω spécifique pour chacune des trois espèces. Comme dans l'étude de Shoemaker et al., la comparaison de ces résultats montre que la vraisemblance du modèle avec un ω spécifique pour *D. recens* (modèle 2) est significativement plus élevée que celle du modèle avec 1 seul oméga, alors que le modèle avec 3 ω n'est pas significativement meilleur que celui à 2 ω. Ce modèle 2 permet ainsi d'estimer une valeur de ω chez *D. recens* (ω =0.033), plus élevée que chez les deux autres espèces (ω =0.012).

Afin de mieux comprendre l'absence de différence entre les valeurs de ω associées aux branches infectées et non-infectées de l'arbre Symbiocode, nous avons étudié la contribution des différents marqueurs à l'effet observé entre *D. recens* et *D. subquinaria*. En effet, on peut supposer que si l'effet de l'infection par *Wolbachia* sur les taux de substitution non-synonymes est faible, celui-ci ne puisse-t-être détecté que sur un ensemble de marqueurs et non sur chaque marqueur indépendamment. Alternativement, on peut imaginer que les contraintes sélectives ne sont pas équivalentes sur tous les marqueurs et que seuls certains gènes sont affectés par l'effet observé sur l'ensemble des données.

Dans le but de comparer l'impact de la réduction de Ne sur chacun d'eux, nous avons voulu analyser séparément les différents marqueurs mitochondriaux constituant l'alignement concaténé utilisé par Shoemaker et al. J'ai donc identifié les séquences codantes de 13 gènes mitochondriaux au sein de cet alignement. Pour cela, les CDS des gènes mitochondriaux de *D. simulans* (numéro d'accession GeneBank AF200833) ont été comparé par BLAST à la séquence concaténée. Ces alignements ont permis d'identifier les parties du concaténât correspondant à chaque gène (Tableau 5). Chaque marqueur a

ensuite été analysé indépendamment par PAML, afin de comparer les vraisemblances de modèles correspondant à un ω unique dans l'ensemble de l'arbre ou un ω spécifique chez *D. recens*, différent de celui mesuré sur les autres branches. Seuls les séquences correspondant au gène *NADH deshydrogénase subunit 1* montrent une augmentation significative de ω chez *D. recens*.

L'effet de la réduction de Ne sur les taux de substitution mesurés sur les autres marqueurs est donc nul ou trop faible pour être détecté. Toutefois, on peut remarquer que pour presque tous les marqueurs de longueur suffisante (8 marqueurs de plus de 500 bp), PAML estime une valeur de ω plus élevée sur les lignées de *D. recens* que dans le reste de l'arbre, conformément à l'effet attendu de *Wolbachia*. Ce résultat suggèrerait que ces gènes sont bien affectés par une faible augmentation de ω, même si cet effet est trop faible pour être détecté sur chaque marqueur considéré indépendamment. Notons qu'un test de Wilcoxon, mesurant si la différence entre ω des branches infectées et non infectée est plus souvent négative ou positive, permet de détecter cette augmentation (p-value = 0.008). L'augmentation de ω dans les lignées infectées semble donc bien affecter tous les marqueurs mitochondriaux. Le marqueur CO1 utilisé semble être également associé à une faible augmentation de ω chez *D. recens*. On peut donc penser que si les lignées infectées de notre jeu de données étaient affectées par une réduction de Ne similaire à celle qui a affecté le génome mitochondrial de *D. recens*, l'étude de la différence entre ω_{inf} et ω_{un} du marqueur CO1 par le test de signe apparié que nous avons employé aurait bien permis de détecter une augmentation de ω sur les branches infectés. Autrement dit, même si le marqueur CO1, ne permet pas de mesurer une différence de ω significative dans chaque paire d'espèces infectées et non-infectées considérées indépendamment, le fait de comparer les ω de nombreuses espèces infectées et non infectées, devrait permettre de détecter des différences globales si elles existent.

	Longueur	**Longueur sans gaps**	**ω model 0**	**ω model 1 non-inf / inf**	**lnL0**	**lnL1**	**p-value***
ATPase_6	675	555	0.0188	0.0194 / 0.0178	-804.273	-804.271	0.94
ATPase_8	27	27	-	-	-	-	-
NADH dehydrogenase subunit 1	813	813	0.0044	*0.0001 / 0.0186*	-1119.458	-1116.841	0.022
NADH dehydrogenase subunit 2	852	732	0.0500	*0.0231 / 0.0918*	-1018.979	-1017.870	0.136
NADH dehydrogenase subunit 3	267	0	-	-	-	-	-
NADH dehydrogenase subunit 4	999	600	0.0061	*0.0001 / 0.0112*	-864.559	-863.470	0.140
NADH dehydrogenase subunit 4L	93	0	-	-	-	-	-
NADH dehydrogenase subunit 6	39	0	-	-	-	-	-
cytochrome_b	1140	1023	0.0148	*0.01245 / 0.0169*	-1513.916	-1513.867	0.753
Cytochrome c_oxidase subunit I	1530	1530	0.0089	*0.0050 / 0.0146*	-2425.560	-2424.787	0.214
COI-petit fragment	660	600	0.0027	*0.0016 / 0.0094*	-1376.821	-1376.142	1
Cytochrome c oxidase subunit II	675	675	0.0104	*0.0074 / 0.0187*	-961.431	-961.250	0.547
Cytochrome c oxidase subunit III	792	738	0.0177	*0.0150 / 0.02157*	-1099.751	-1099.695	0.738
concaténât	**9453**	**7227**	**0.0173**	***0.0123/ 0.0242***	**-10851.389**	**-10849.318**	**0.0419**

Tableau 5: Estimation par maximum de vraisemblance des ω (dN/dS) de différents gènes mitochondriaux chez *D. subquinaria* et *D. quinaria* (PAML, données de Shoemaker *et al.* 2004). Dans le modèles 0, toutes les lignées ont le même ω, dans le modèle 1, les lignées correspondant à *D. recens* ont un ω différent des autres lignées. *p-value du test LRT comparant la vraisemblance du modèle 1 à celle du modèle 0. d.f.=1. En italique : marqueurs pour lesquels ω_{inf} est supérieur à $\omega_{non\text{-}inf}$.

3.6. Discussion

Des réductions de Ne mitochondrial causées par des invasions par des symbiotes cytoplasmiques peuvent être associées à des réductions de l'efficacité de la sélection dans les populations concernées. Nous avons cherché à mesurer l'impact de cet effet sur les variations de ω entre lignées infectées ou non par *Wolbachia* dans un grand nombre de taxons. Contrairement à cette prédiction, nous avons observé des valeurs moyennes de ω similaires entre lignées infectées et non infectées de l'arbre des spécimens du système SymbioCode. Cette observation ne semble pas s'expliquer par un manque de sensibilité du marqueur mitochondrial utilisé.

Ce résultat suggère donc que l'impact des invasions par *Wolbachia* sur les rapports ω mitochondriaux pourrait être faible, ce qui pourrait s'expliquer soit par un impact faible ou nul des infections sur les diminutions de Ne mitochondriale, soit par l'influence d'autres effets affectant les tailles efficaces des génomes. Ainsi, la corrélation entre ω et le statut d'infection est peut-être masquée par d'autres facteurs affectant les variations ω. Quoi qu'il en soit, ces deux interprétations supposent que les infections par *Wolbachia* ne sont pas un déterminant majeur des différences d'efficacité de la sélection sur les génomes mitochondriaux.

D'autre part, on peut rappeler que cette étude repose sur les inférences des statuts d'infection ancestraux sur les branches des arbres des hôtes, qui sont nécessairement affectées par certaines d'incertitudes. Tout d'abord, leur qualité dépend de l'efficacité de la méthode utilisée pour inférer les scénarios de pertes et d'acquisition des infections. En effet, nous avons vu que ces inférences reposaient sur la reconstruction de cophylogénies, qui pouvaient être complexes dans les cas où les transferts, acquisitions et pertes sont fréquents, ce qui est le cas dans cette étude. D'autre part, même si les statuts inférés à chaque nœud sont corrects, les changements multiples de statuts d'infections le long des branches de l'arbre introduisent un autre type d'incertitude dans notre analyse. En effet, si le remplacement des infections est rapide, les branches dites « non-infectées » sont susceptibles d'avoir été infectées pendant une part de leur histoire par une bactérie perdue par la suite, et inversement. Ainsi, l'absence de différentes de ω entre lignées infectées et non-infectées pourrait

s'expliquer par des changements rapides de statuts le long des branches, qui obscurcissent la comparaison des branches « infectées » ou non.

Il devrait être possible de limiter l'impact de cette incertitude au moyen des scénarios de perte et d'acquisition reconstruits par co-phylogénie. Par exemple, on pourrait restreindre la mesure de ω « infectés » aux branches associées à l'arrivée et à l'invasion d'une infection, dans la mesure où l'augmentation de ω est vraisemblablement associée à cette phase de l'infection. Par ailleurs, l'incidence, c'est-à-dire la proportion d'espèces infectées, varie selon les taxons. Si l'on suppose que la « durée » moyenne d'une infection est la même dans ces différents groupes, cela suggère que les évènements d'invasions sont plus fréquents dans les groupes où l'incidence est élevée et donc que leur impact global sur l'efficacité de la sélection est plus élevé dans ces groupes. Il serait donc intéressant d'étudier la corrélation entre ω et incidence de *Wolbachia* dans ces différents groupes.

Chapitre 4

Discussion

Du fait de leur co-transmission au sein du compartiment cytoplasmique, les bactéries intracellulaires peuvent influencer l'évolution des génomes mitochondriaux de leurs hôtes. Leur propagation entraine celle des mitochondries qui leur sont liées, par autostop, ce qui peut conduire à une réduction des effectifs efficaces (N_e) de ces génomes. Cette sélection indirecte peut également faciliter l'introgression d'une mitochondrie introduite dans une espèce suite à une hybridation, si elle est liée à une de ces bactéries. Le principal objectif de ma thèse était de quantifier ces différents phénomènes, de manière globale, au moyen d'un large échantillonnage d'Arthropodes.

Les introgressions mitochondriales sont à l'origine de discordances entre les histoires évolutives des génomes mitochondriaux et nucléaires. Afin de quantifier ces discordances, nous avons obtenu des marqueurs génomiques nucléaires de type RADseq, permettant de reconstruire l'histoire de ces génomes. La production de ces données a été l'occasion de développements techniques concernant l'obtention et l'utilisation des marqueurs RAD. Tout d'abord, des améliorations du protocole de préparation des librairies RADseq nous ont permis de générer des données en multiplexant de nombreux spécimens appartenant à différents groupes d'Arthropodes potentiellement très divergents (Henri et al. 2015). Par ailleurs, l'utilisation des données de RADseq, qui reposent sur le partage de sites de restriction conservés entre spécimens, présente certaines difficultés que je me suis attachée à cerner et éventuellement à contourner. J'ai notamment étudié la question de l'utilisation de ce type de marqueurs pour inférer des relations phylogénétiques entre espèces au moyen d'expériences de RADseq *in silico* sur des génomes complets (Cariou et al. 2013). Ces analyses ont montré que des données de RADseq permettent de reconstruire des phylogénies entre des espèces de Drosophiles caractérisées par des temps de divergences allant jusqu'à 63 millions d'années. D'autre part, j'ai étudié l'impact de biais liés au polymorphisme des sites de restriction sur l'estimation des diversités génétiques au sein des populations. Des analyses *in silico* m'ont permis de mesurer l'importance de ce biais et de proposer une méthode pour le corriger dans certains cas.

Les données générées m'ont permis de démontrer que, sur 29 espèces de Diptères et de Lépidoptères testées à ce jour, la proximité génétique mitochondriale est systématiquement confirmée par les marqueurs nucléaires, rejetant ainsi l'hypothèse d'une introgression mitochondriale récente. Les marqueurs RAD ont dans ce cadre été utilisés pour la comparaison de spécimens proches, au sein de différents taxons. Je

discuterai dans cette partie de leur possible emploi dans le cadre de comparaisons entre espèces, par exemple pour la résolution de relations phylogénétiques entre certains des taxons étudiés.

Dans cette discussion, nous aborderons également la question de la quantification des évènements d'introgression. En effet, les résultats obtenus indiquent que les introgressions mitochondriales ne sont pas assez communes pour générer des discordances nombreuses entre divergences mitochondriales et nucléaires. Toutefois, nous verrons que la mesure de la fréquence de ces évènements requerrait la prise en compte de l'échantillonnage des spécimens au sein des espèces étudiées ainsi que de l'âge des événements que la méthode employée permettrait de détecter.

Par ailleurs, sur un plus large échantillon, nous avons mis en évidence une réduction significative de la diversité mitochondriale au sein des espèces infectées. Cette réduction confirme d'existence d'une réduction du Ne mitochondrial causée par les infections par *Wolbachia*. Toutefois, la comparaison des patrons de substitutions entre lignées mitochondriales montre que cette réduction de Ne ne semble pas générer de réductions notables de l'efficacité de la sélection naturelle dans les lignées infectées. Ces observations apparemment contradictoires peuvent s'expliquer de différentes façons. Tout d'abord, la comparaison des polymorphismes est effectuée sur les espèces actuelles tandis que celle des taux de substitutions implique des inférences indirectes, non seulement de ces taux mais également et des statuts d'infection de lignées ancestrales. Ces deux comparaisons mettent donc en jeu des estimations affectées par des types d'incertitudes différents. Cette disparité pourrait expliquer, au moins en partie, le fait que l'on détecte un effet et non l'autre. D'autre part, cette différence de temporalité pourrait également signifier que les deux effets ne sont pas affectés de la même façon par les réductions de Ne mitochondriales. Il se peut que des variations d'âge, de durée et d'intensité des balayages sélectifs entre lignées infectées affectent différemment les deux types de comparaisons.

1. Le RADseq, différentes contraintes à différentes échelles

Les méthodes de RADseq permettent l'obtention des séquences de locus nucléaires homologues de spécimens caractérisés par des temps de divergences variables. En effet, nous avons vu que ce type de données pouvait être utilisé aussi bien pour l'étude de la diversité génétique au sein des populations que pour la reconstruction de phylogénies entre espèces. L'utilisation des marqueurs RAD homologues repose sur le partage de sites de restriction conservés entre les différents génomes comparés ; les variations affectant la distribution et le partage de ces sites ont des impacts importants et différents selon le niveau de divergence des organismes étudiés.

Tout d'abord, au sein des populations, la proportion de sites de restriction conservés entre deux individus dépend du polymorphisme des sites de restriction, qui dépend lui-même de la diversité génétique. Ce polymorphisme peut être à l'origine d'un biais dans l'échantillonnage des coalescents, qui affecte la mesure de la diversité nucléotidique et de l'hétérozygotie observée (Gautier et al. 2013, Arnold et al. 2013). Nous avons confirmé et quantifié ce biais au moyen de simulations. L'intensité du biais observé dépend du niveau de polymorphisme ainsi que d'autre paramètres tels que des écarts à la panmixie au sein des populations étudiées. Par ailleurs nous avons pu montrer que des méthodes ABC pouvaient permettre de corriger ces biais.

Par ailleurs, une distribution non aléatoire des locus RAD le long du génome peut également causer un biais dans la mesure de la diversité génétique des populations. Des compositions en nucléotides différentes dans différentes régions du génome peuvent par exemple conduire à des densités hétérogènes en locus RAD. Ces variations peuvent introduire des biais si les compositions en base sont corrélées au niveau de polymorphisme. En particulier, les régions riches en GC peuvent être enrichies en régions codantes, souvent sous sélection purifiante, ce qui pourrait conduire à un enrichissement des données RAD en régions codante et donc à une sous-estimation du polymorphisme (DaCosta, Jeffrey M., Sorenson 2014, McCluskey and Postlethwait 2014). Si ce biais est principalement dû à la présence de régions codantes dans les données RAD, on peut envisager de détecter au moins en partie ces séquences codantes, ce qui permettrait de calculer des diversités correspondant aux positions synonymes.

Nous avons également étudié la question de l'impact de la divergence sur le nombre de locus conservés entre spécimens d'espèces différentes, ce qui nous a permis de montrer que ce type de données permettaient de reconstruire la phylogénie d'espèces du genre *Drosophila*. Les données de RADseq générées au cours de cette thèse ont principalement été utilisées pour la comparaison de spécimens caractérisés par des séquences mitochondriales très similaires, et donc étroitement apparentés. J'ai également cherché à évaluer si ces locus RAD pouvaient être utilisés pour des comparaisons entre espèces différentes, dans le but de résoudre leurs relations phylogénétiques. La résolution de ces phylogénies nucléaires pourrait par exemple permettre leur comparaison aux phylogénies mitochondriales, rendant ainsi possible la recherche d'introgressions mitochondriales plus anciennes que celles étudiées dans l'analyse précédente, potentiellement responsables d'incongruences topologiques entre phylogénies RAD et CO1.

Les individus de la librairie RAD SymbioCode séquencés appartiennent aux ordres des Diptères et des Lépidoptères. Au sein de chacun de ces ordres, nous avons recherché les locus orthologues sur l'ensemble des individus. Les résultats obtenus sont présentés dans la figure 57. On peut remarquer que le nombre de locus orthologues partagés au sein de paires de Diptères est très souvent inférieur à 10. En d'autres termes, le nombre de locus partagés entre espèces est ici insuffisant pour reconstruire des phylogénies. A l'inverse, on peut observer que le nombre de locus orthologues retrouvés entre Lépidoptères est variable et potentiellement suffisant pour résoudre certaines parties de la phylogénie. Certains spécimens ne partagent qu'un très faible nombre de locus avec l'ensemble des autres spécimens, mais d'autres ont plus de 100 locus orthologues en commun avec une forte proportion des individus. Cette différence peut s'expliquer en partie par les temps de divergence variables entre paires de spécimens. Toutefois, la plupart des spécimens partageant peu d'orthologues avec les autres se caractérisent également par des couvertures très faibles, comme on peut l'observer sur la diagonale de la figure 57A, suggérant que le nombre de locus couverts est ici plus critique que la conservation des locus.

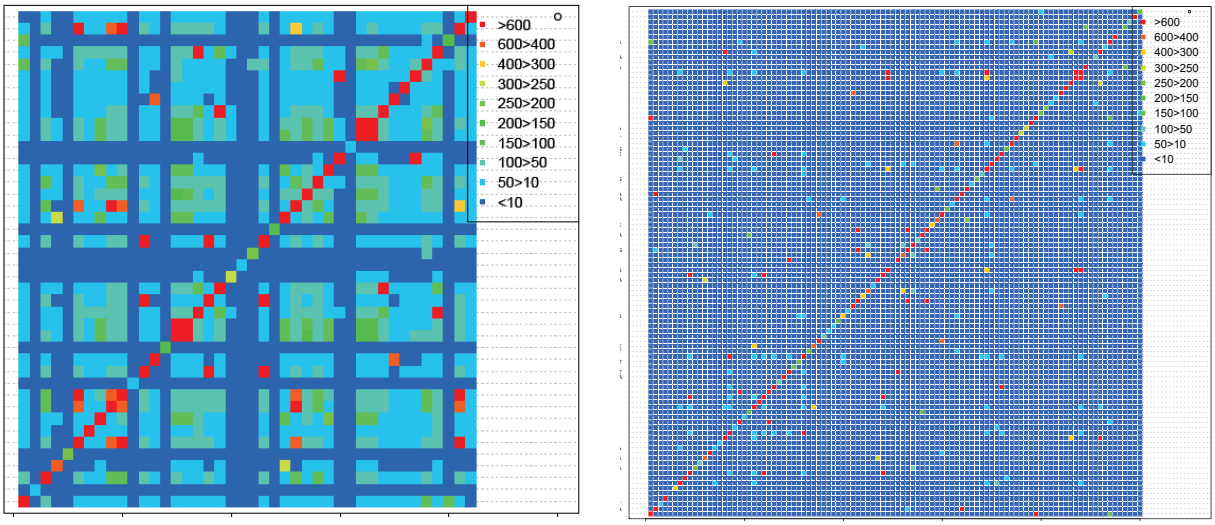

Figure 57. Nombres de locus partagés entre paires de spécimens A. de Lépidoptères B. de Diptères. Les couleurs indiquent le nombre de locus partagés (voir légende) entre l'individu en abscisse et celui en ordonné, pour toutes les paires d'individus.

En raison du plus grand nombre de locus orthologues identifiés entre espèces de Lépidoptères, nous avons concentré notre analyse sur ce groupe. Les phylogénies RAD ont été inférées pour les spécimens partageant au moins 100 locus avec au moins 3 autres spécimens (N=21). Le jeu de données ainsi construit regroupe les alignements de 1138 locus. La figure 58 représente les phylogénies obtenue par maximum de vraisemblance (phyML, Guindon et al. 2010) à partir de ce jeu de données RAD d'une part et des séquences CO1 des mêmes spécimens d'autre part. L'arbre obtenu au moyen des données RADseq comporte une proportion élevée de nœuds résolus de façon robuste (score aLRT supérieur à 95% pour tous les nœuds sauf 2). Notons que ces spécimens représentent des espèces de différentes familles de Lépidoptères (*Crambidae*, *Pyralidae*, *Erebidae*, *Noctuidae* et *Arctidae*), ce qui suggère que ces données permettent de reconstruire des relations phylogénétiques à des échelles plus profondes que celles déjà testées (inter familles).

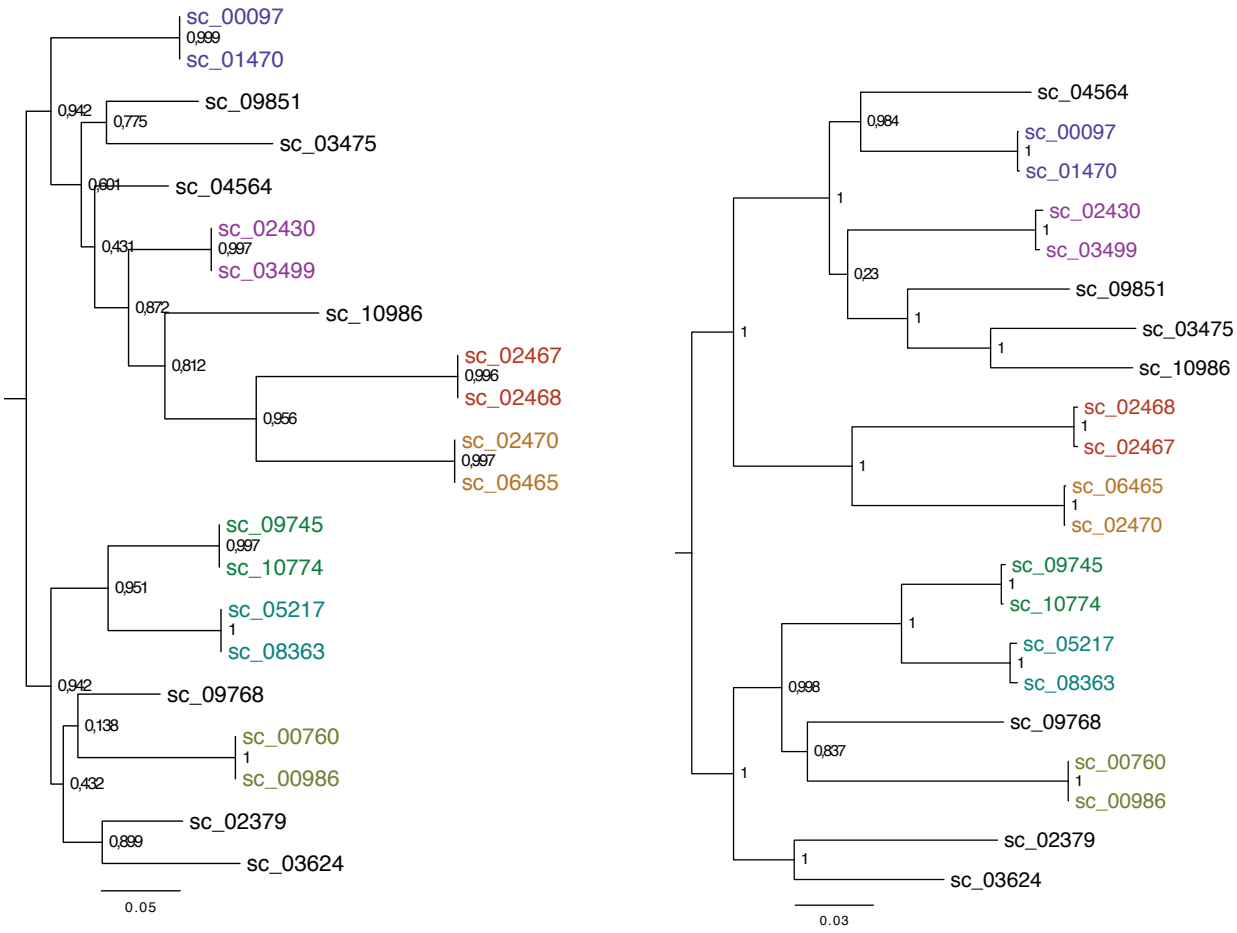

Figure 58. Phylogénie CO1 (gauche) et phylogénie RAD (droite) des 21 spécimens partageant au moins 100 locus avec au moins 3 autres spécimens de Lépidoptères. Les couleurs représentent les individus appartenant à une même paire. Les valeurs indiquées aux nœuds représentent leur soutien statistique (score aLRT).

Cette phylogénie a été comparée à celle des séquences CO1 des mêmes spécimens au moyen du test SH implémenté dans RAxML (Shimodaira and Hasegawa 1999, Stamatakis 2014). Ce test est fondé sur la comparaison des vraisemblances de la phylogénie CO1 générée par maximum de vraisemblance à celle d'une phylogénie CO1 dont la topologie est contrainte par celle de la phylogénie nucléaire. Cette comparaison a permis de montrer que la topologie imposée par la phylogénie RAD n'est pas moins vraisemblable que celle obtenue de façon non contrainte pour les séquences CO1. Ce résultat signifie qu'on ne détecte pas de signal d'incongruence entre les phylogénies RAD et CO1. Ainsi, parmi les 5 nœuds de la phylogénie CO1 résolus avec un score

supérieur à 0.9 (hors nœuds correspondant aux individus des paires), tous sont retrouvés et fortement soutenus dans l'arbre RAD.

Cette analyse préliminaire suggère que les données RAD rendent possible la résolution des relations phylogénétiques entre espèces à une échelle plus profonde que les marqueurs CO1. Ces données permettront donc l'étude de la congruences entre phylogénies mitochondriales et nucléaires au sein des taxons suffisamment récents pour correspondre à des phylogénies CO1 résolues de façon robuste. La librairie en cours de préparation comporte 55 espèces supplémentaires de Lépidoptères ainsi que pour plusieurs espèces d'autres ordres d'Arthropodes. Nous attendons en outre une meilleure couverture des locus RAD des spécimens inclus dans cette nouvelle librairie. En effet, de nouvelles améliorations du protocole de préparation de cette librairie RADseq devraient permettre d'accroitre la quantité et la qualité des données produites. Notamment, l'étape de fragmentation des produits de digestion sera réalisée par une nouvelle méthode enzymatique permettant de maximiser la proportion de locus RAD conservées dans la librairie RAD (communication personnelle Hélène Henri). Ces données devraient donc permettre la reconstruction des phylogénies d'un plus grand nombre de taxons.

2. Discordances entre divergences mitochondriales et nucléaires et fréquence des introgressions mitochondriales

Au cours de ma thèse, je me suis intéressée à la détection d'introgressions récentes par la rechercher d'éventuelles discordances entre les temps de divergences nucléaires et mitochondriaux. Plus spécifiquement, nous avons mesuré la fréquence des cas où deux spécimens possédant des génomes mitochondriaux suffisamment proches pour que leurs barcodes CO1 soient identiques correspondent à des individus d'espèces différentes. Une telle approche se concentre donc sur la détection d'introgressions récentes, susceptibles d'expliquer le partage de génomes mitochondriaux très similaires par des individus trop divergents au niveau nucléaire.

Parmi les 29 paires d'individus étudiées, aucune ne montre de discordance entre les divergences mitochondriales et nucléaires. Cela signifie que les introgressions mitochondriales ne sont pas assez fréquentes pour que ce type de discordances ait été

observé dans cette étude. Dans la mesure où les marqueurs mitochondriaux sont souvent utilisés pour reconstruire l'histoire des populations, il est intéressant de mesurer la fréquence des cas dans lesquels ces marqueurs sont en discordance avec l'histoire des génomes nucléaires. Toutefois, nous pourrions également nous interroger sur la mesure de la fréquence des évènements d'introgression mitochondriale d'une manière plus générale. Cette estimation nécessiterait tout d'abord la prise en compte de l'âge des évènements étudiés. Par exemple, dans l'étude que nous avons menée, seules des introgressions suffisamment récentes pour que les génomes introgressés n'aient accumulé aucune mutation sur le marqueur CO1 pouvaient être détectés. De plus, sur cet intervalle de temps court, il se peut que les introgressions détectées soient en cours et que des génomes mitochondriaux introgressés et non-introgressés coexistent au sein de l'espèce receveuse. Il est donc nécessaire de tenir compte de la proportion de spécimens affectés nécessaires pour détecter chaque événement, en fonction du nombre de spécimens échantillonnés.

Ainsi, une quantification de la fréquence des introgressions mitochondriales correspondrait à une mesure de la fréquence des espèces ayant reçu ou donné une mitochondrie par introgression depuis un nombre de générations donné. Il s'agirait de tenir compte, d'une part, de l'âge des évènements détectés et d'autre part de la proportion d'individus affectés nécessaire pour que la probabilité d'avoir détecté l'événement soit assez élevée.

3. Quel impact de *Wolbachia* sur les Ne mitochondriaux ?

Au cours de cette thèse, j'ai également étudié l'impact de la propagation de symbiotes cytoplasmiques invasifs sur les variations de tailles efficaces, Ne, des populations mitochondriales. Je me suis attachée à tester d'une part si les symbiotes pouvaient réduire significativement le polymorphisme, et d'autre part s'ils pouvaient diminuer l'efficacité de la sélection.

Les résultats obtenus montrent une réduction significative du polymorphisme mitochondrial des spécimens infectés au sein des espèces correspondantes, par rapport à celui des espèces non-infectées (1 : spécimens infectés (espèces infectées) *vs* espèces non-infectées). Nous avons toutefois également observé que cette diminution ne

s'accompagne pas d'une différence significative de polymorphisme entre espèces infectées ou non, quand celui-ci est mesuré sur l'ensemble des spécimens au sein de chaque espèce (2 : espèces infectées *vs* espèces non-infectées). En outre, au sein des espèces infectées, nous n'avons pas non plus mesuré de réduction significative du polymorphisme des individus infectés par rapport aux individus non-infectés (3 : individus infectés - espèces infectées *vs* individus non infectés - espèces infectées). Néanmoins pour les deux dernières comparaisons, les diversités génétiques moyennes de chaque catégorie varient dans le sens attendu, c'est-à-dire dans le sens d'une réduction causée par la présence de *Wolbachia*. Ces résultats traduisent peut-être l'existence d'effets plus faibles de *Wolbachia* dans ces contextes (test 2 et 3). En effet, on peut remarquer que le test comparant les polymorphismes des spécimens infectés à celui des espèces non-infectées (test 1) était celui permettant de détecter de la façon la plus sensible les variations de diversité causées par *Wolbachia*. L'observation d'une diminution globale du polymorphisme des espèces infectées par rapport aux espèces non-infectées (test 2) suppose, quant à lui, que les balayages sélectifs causés par la propagation de bactéries affectent l'ensemble des spécimens des espèces infectées. Ils doivent en outre être suffisamment récents pour que le polymorphisme n'ait pas été rétabli dans les populations concernées. Ainsi, le polymorphisme mitochondrial ne sera pas réduit dans les cas d'espèces où plusieurs infections coexistent et permettent le maintien d'une diversité globale élevée. De même, des infections introduites par introgression au sein d'une espèce peuvent être associées à un haplotype mitochondrial exogène divergent. Elles peuvent ainsi temporairement entrainer une augmentation du polymorphisme mitochondrial au sein de la population introgressée. Ainsi, plusieurs scénarios peuvent réduire la différence de polymorphisme global attendu entre espèces infectées et non-infectées (test 2). De façon similaire, la comparaison du polymorphisme des spécimens infectés à celui des spécimens non-infectés (test 3) suppose que dans de nombreux cas, une infection récente est en cours de propagation dans la population. En effet, dans les populations où les infections ont atteint une fréquence d'équilibre, la réduction du polymorphisme affecte l'ensemble de la population.

Ainsi, le fait que seule la comparaison du polymorphisme des spécimens infectés des espèces infectées à celui des espèces non infectées (test 1) indique une différence significative n'est pas en contradiction avec les résultats négatifs des autres tests. Ce résultat indique que les infections sont en général suffisamment récentes pour que le

polymorphisme mitochondrial n'ait pas été rétabli par l'accumulation de nouvelles mutations.

Par ailleurs, les modèles de génétique des populations montrent que des réductions d'effectif efficace sont associées à une diminution de l'efficacité de la sélection dans les populations concernées. Cette diminution peut conduire à une augmentation du rapport ω dans les lignées correspondantes. On pouvait donc s'attendre à observer des ω plus élevés dans les lignées infectées du système étudié. Contrairement à cette prédiction, nous n'avons pas observé de variations de ω en relation avec les statuts d'infection des différentes lignées. Plusieurs hypothèses pourraient expliquer ce résultat négatif.

Tout d'abord, cette étude repose sur l'inférence de statuts d'infection ancestraux qui peuvent être incertains, en particulier si les pertes et acquisitions d'infections s'enchainent rapidement au cours de l'histoire des lignées. Ainsi, l'impact des réductions de Ne mitochondrial sur l'efficacité de la sélection pourrait être moins facilement détecté que celui affectant le polymorphisme.

Par ailleurs, des différences de temporalité dans les effets de *Wolbachia* sur le polymorphisme et sur taux de substitution pourraient aussi expliquer ces observations contrastées. Par exemple, le polymorphisme mitochondrial est probablement fortement affecté par l'âge de l'invasion et l'intensité du balayage sélectif associé. En revanche, les augmentations de ω attendues sont en principe à des augmentations du taux de substitution non-synonymes pendant une certaine durée le long d'une branche infectée. Elles dépendent donc du nombre de substitutions « supplémentaires » fixées à cause de la réduction de Ne pendant le temps de l'invasion, plutôt que de l'âge de la dernière invasion dans une lignée.

La figure 59 illustre ainsi plusieurs évènements d'infection théoriques par *Wolbachia* et leurs effets attendus sur le polymorphisme mitochondrial d'une part et sur le rapport ω d'autre part. Le schéma de gauche illustre différents patrons de diminution du polymorphisme mitochondrial liés à des évènements d'infection. Après la phase d'invasion, le polymorphisme se restaure progressivement dans les populations concernées. Ainsi la diversité mesurée dans les populations actuelles dépend fortement de la date de la dernière phase d'invasion. Par exemple, dans la population B l'invasion est plus ancienne que dans la population D, si bien que le polymorphisme de cette population est plus élevé. La population C a été infectée par plusieurs souches

successives induisant des balayages récurrents. Dans ce cas, l'intensité de la réduction du polymorphisme ne dépend que de la dernière infection, c'est à dire du dernier goulot d'étranglement du Ne mitochondrial. A l'inverse, sur le schéma de droite, on peut observer que de rapport ω dépend non de l'âge des infections mais plutôt de la durée, la fréquence, et l'intensité de la diminution de Ne par rapport à la longueur totale de la branche. Par exemple, plusieurs invasions successives se succèdent sur la branche reliant l'espèce C au nœud le plus proche. Cette branche est ainsi affectée par des réductions de Ne pendant une durée plus longue que la branche sœur menant à l'espèce B. L'augmentation de ω est ainsi plus importante dans cette lignée que dans celle correspondant à la lignée B.

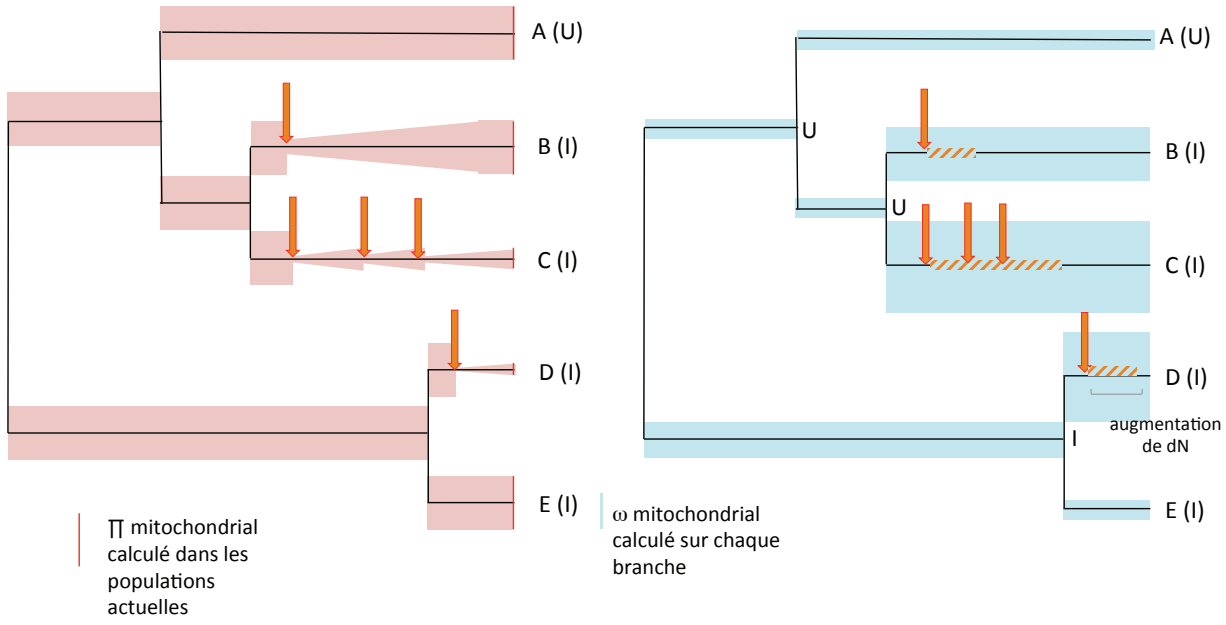

Figure 59. Impact de différentes infections théoriques par *Wolbachia* sur le polymorphisme mitochondrial (gauche) et sur ω (droite). Les flèches orange représentent l'arrivée d'une infection dans une lignée hôte. Les hachures orange représentent la période pendant laquelle le balayage sélectif associé à l'invasion affecte le Ne mitochondrial et donc l'efficacité de la sélection. On suppose que seules les infections sont responsables des variations de Ne entre ces lignées.

Afin de mieux comprendre la façon dont le polymorphisme mitochondrial et les rapports ω sont affectés par les infections par *Wolbachia*, il serait intéressant de restreindre les analyses à des sous-parties du jeu de données pour lesquelles les histoires des infections sont caractérisées de façon plus solide. Par exemple, nous avons vu que l'effet supposé de *Wolbachia* sur l'efficacité de la sélection sur les génomes mitochondriaux correspondrait à la fixation facilitée de mutations faiblement délétères lors de la phase d'invasion. Nous pourrions ainsi rechercher non pas une augmentation de ω sur l'ensemble des branches infectées, mais plus spécifiquement sur les branches associées à des évènements d'invasion. De plus, nous avons également vu que l'impact de ces évènements sur les variations de ω dépendaient de la fraction de la longueur de la lignée affectée par la diminution de Ne. En outre, le remplacement des infections le long des lignées, et l'alternance des phases infectées et non infectées introduit des incertitudes dans les comparaisons des ω en fonction des statuts inférés sur chaque branche. Il serait donc peut-être utile de tenir compte des longueurs des branches ainsi que du degré d'incertitude sur les statuts d'infections dans ces comparaisons.

Enfin, on peut remarquer que ces invasions mettent en jeu des mécanismes différents selon le type d'infection. Par exemple, les bactéries causant une mortalité embryonnaire des mâles se maintiennent souvent à une fréquence intermédiaire, ce qui correspond à une réduction de Ne sur le long terme mais d'intensité probablement plus faible que l'effet du balayage sélectif initial qui cause la fixation de la lignée mitochondriale infectée. On a vu par ailleurs que la rapidité de l'invasion et la fréquence d'équilibre des infections dépendent à la fois de l'intensité du phénotype affectant la reproduction et des effets positifs et négatifs induits sur la valeur sélective de l'hôte. On peut se demander quel est l'impact de ces variations sur l'effet sur ω. Des infections se fixant à des vitesses différentes ont-elles des impacts différents sur le ω mitochondrial de leur hôte ?

Conclusion

Cette thèse a eu pour objet principal l'étude des conséquences sur les génomes mitochondriaux des Arthropodes de leur liaison génétique avec des micro-organismes invasifs, capables de les entrainer dans leur propagation au sein des populations. Ces conséquences sont potentiellement multiples: facilitation des introgressions mitochondriales, diminution de la taille efficace des mitochondries se traduisant éventuellement par des réductions du polymorphisme et de l'efficacité de la sélection. Nous avons voulu mesurer ces effets de manière globale sur un ensemble d'espèces hétérogènes.

La comparaison des histoires des génomes mitochondriaux à celles des génomes nucléaires nécessitait l'obtention de marqueurs nucléaires. Nous avons ainsi mis au point des marqueurs RADseq, technique apparue peu avant le début de ce travail. L'étude des différentes particularités de cette méthode a pris une ampleur importante et constitue un deuxième axe majeur de ma thèse. Les développements méthodologiques présentés ici s'inscrivent dans l'ensemble des travaux qui ont permis à cette technique d'être aujourd'hui largement utilisée. J'ai employé des approches diverses (expériences *in silico*, biologie moléculaire, simulations, mise au point de pipeline d'analyse…), dont certaines ont permis l'obtention et l'analyse des données RAD présentées dans cette thèse. D'autre ont permis une meilleure définition des effets attendus et recherchés; par exemple, l'étude des biais RAD a permis de mieux formuler les effets attendus en terme de distance entre individus et d'hétérozygotie observée selon les cas envisagés…

En conclusion, nous avons pu mettre en évidence une réduction du polymorphisme mitochondrial liée à la présence d'une infection par *Wolbachia*. Mais nous avons également vu que cet effet global recouvre des situations plus hétérogènes. Différents processus, qu'il conviendrait de distinguer, sont à l'œuvre dans les différentes espèces infectées. D'autre part, aucune diminution de l'efficacité de la sélection, mesuré par des variations du rapport ω, n'a été détectée par les approches utilisées. Ici encore, cette absence d'effet global a permis de mieux cerner l'hétérogénéité des processus à l'oeuvre. Par exemple, il apparait maintenant clair que ce n'est pas le statut d'infection qui affecte le Ne mitochondrial, mais plutôt les occurrences des phénomènes d'invasion. De plus, des symbiotes associés à des mécanismes d'invasion différents affectent probablement de façon contrastée les populations hôtes. Ces différences pourraient être prises en compte et intégrées dans ces études globales des effets de la présence des symbiotes.

Ainsi, d'une manière générale, cette étude a permis de mieux comprendre les effets attendus de la liaison génétique entre mitochondries et symbiotes. Ceci permettra dans des études ultérieures de mieux caractériser, isoler et mesurer les conséquences évolutives des symbiotes cytoplasmiques.

Références

Andrew, R. L., N. C. Kane, G. J. Baute, C. J. Grassa, and L. H. Rieseberg. 2013. Recent nonhybrid origin of sunflower ecotypes in a novel habitat. Molecular Ecology 22:799–813.

Andrews, K. R., a Paul, M. R. Miller, and G. Luikart. 2014. Trade-offs and utility of alternative RADseq methods: reply to Puritz et al. 2014. Molecular Ecology:5943–5946.

Arnold, B., R. B. Corbett-Detig, D. Hartl, and K. Bomblies. 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Molecular ecology 22:3179–90.

Baird, N. a, P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. a Lewis, E. U. Selker, W. a Cresko, and E. a Johnson. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PloS one 3:e3376.

Barchi, L., S. Lanteri, E. Portis, A. Acquadro, G. Valè, L. Toppino, and G. L. Rotino. 2011. Identification of SNP and SSR markers in eggplant using RAD tag sequencing. BMC genomics 12:304.

Baxter, S. W., J. W. Davey, J. S. Johnston, A. M. Shelton, D. G. Heckel, C. D. Jiggins, and M. L. Blaxter. 2011. Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. PloS one 6:e19315.

Bazin, E., S. Glémin, and N. Galtier. 2006. Population size does not influence mitochondrial genetic diversity in animals. Science (New York, N.Y.) 312:570–2.

Beaumont, M. a., W. Zhang, and D. J. Balding. 2002. Approximate Bayesian computation in population genetics. Genetics 162:2025–2035.

Bergsten, J., D. T. Bilton, T. Fujisawa, M. Elliott, M. T. Monaghan, M. Balke, L. Hendrich, J. Geijer, J. Herrmann, G. N. Foster, I. Ribera, A. N. Nilsson, T. G. Barraclough, and A. P. Vogler. 2012. The effect of geographical scale of sampling on DNA barcoding. Systematic Biology 61:851–869.

Bybee, S. M., H. Bracken-Grissom, B. D. Haynes, R. a. Hermansen, R. L. Byers, M. J. Clement, J. a. Udall, E. R. Wilcox, and K. a. Crandall. 2011. Targeted amplicon sequencing (TAS): A scalable next-gen approach to multilocus, multitaxa phylogenetics. Genome Biology and Evolution 3:1312–1323.

Camus, M. F., D. J. Clancy, and D. K. Dowling. 2012. Mitochondria, maternal inheritance, and male aging. Current Biology 22:1717–1721.

Cariou, M., L. Duret, and S. Charlat. 2013. Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. Ecology and Evolution 3:846–852.

Catchen, J., P. a Hohenlohe, S. Bassham, A. Amores, and W. a Cresko. 2013. Stacks: an analysis tool set for population genomics. Molecular ecology 22:3124–40.

Catchen, J. M., A. Amores, P. Hohenlohe, W. Cresko, and J. H. Postlethwait. 2011. Stacks: building and genotyping Loci de novo from short-read sequences. G3 (Bethesda, Md.) 1:171–82.

Charlat, S., A. Duplouy, E. a Hornett, E. a Dyson, N. Davies, G. K. Roderick, N. Wedell, and G. D. Hurst. 2009. The joint evolutionary histories of Wolbachia and mitochondria in Hypolimnas bolina. BMC Evolutionary Biology 9:64.

Chase, C. D. 2007. Cytoplasmic male sterility: a window to the world of plant mitochondrial-nuclear interactions. Trends in Genetics 23:81–90.

Cruaud, A., M. Gautier, M. Galan, J. Foucaud, L. Sauné, G. Genson, E. Dubois, S. Nidelet, T. Deuve, and J. Y. Rasplus. 2014. Empirical assessment of rad sequencing for interspecific phylogeny. Molecular Biology and Evolution 31:1272–1274.

Csilléry, K., M. G. B. Blum, O. E. Gaggiotti, and O. François. 2010. Approximate Bayesian Computation (ABC) in practice. Trends in Ecology and Evolution 25:410–418.

Csilléry, K., O. François, and M. G. B. Blum. 2012. Abc: An R package for approximate Bayesian computation (ABC). Methods in Ecology and Evolution 3:475–479.

Dabney, J., and M. Meyer. 2012. Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. BioTechniques 52:87–94.

DaCosta, Jeffrey M., Sorenson, M. D. 2014. Amplifiation Biases and Consistent Recovery of Loci in a Double-digest RAD-seq Protocol 9.

Davey, J. L., and M. W. Blaxter. 2010. RADSeq: next-generation population genetics. Briefings in functional genomics 9:416–23.

Davey, J. W., T. Cezard, P. Fuentes-Utrilla, C. Eland, K. Gharbi, and M. L. Blaxter. 2012. Special features of RAD Sequencing data: implications for genotyping. Molecular ecology.

Denonfoux, J., N. Parisot, E. Dugat-Bony, C. Biderre-Petit, D. Boucher, D. P. Morgavi, D. Le Paslier, E. Peyretaillade, and P. Peyret. 2013. Gene capture coupled to high-throughput sequencing as a strategy for targeted metagenome exploration. DNA Research 20:185–196.

Dey, A., C. K. W. Chan, C. G. Thomas, and A. D. Cutter. 2013. Molecular hyperdiversity defines populations of the nematode Caenorhabditis brenneri. Proceedings of the National Academy of Sciences of the United States of America 110:11056–60.

Dobson, S. L., W. Rattanadechakul, and E. J. Marsland. 2004. Fitness advantage and cytoplasmic incompatibility in Wolbachia single- and superinfected Aedes albopictus. Heredity 93:135–142.

Dutheil, J., and B. Boussau. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. BMC evolutionary biology 8:255.

Dutheil, J. Y., N. Galtier, J. Romiguier, E. J. P. Douzery, V. Ranwez, and B. Boussau. 2012. Efficient selection of branch-specific models of sequence evolution. Molecular Biology and Evolution 29:1861–1874.

Dyer, K. a, C. Burke, and J. Jaenike. 2011. Wolbachia-mediated persistence of mtDNA from a potentially extinct species. Molecular ecology 20:2805–17.

Eaton, D. a R. 2014. PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics 30:1844–1849.

Eaton, D. a R., and R. H. Ree. 2013. Inferring Phylogeny and Introgression using RADseq Data: An Example from Flowering Plants (Pedicularis: Orobanchaceae). Systematic biology.

Edgar, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics (Oxford, England) 26:2460–1.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. a Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one 6:e19379.

Emerson, K. J., C. R. Merz, J. M. Catchen, P. A. Hohenlohe, W. A. Cresko, W. E. Bradshaw, and C. M. Holzapfel. 2010. Resolving postglacial phylogeography using high-throughput sequencing. Proceedings of the National Academy of Sciences of the United States of America 107:16196–16200.

Engelstädter, J. 2010. The Effective Size of Populations Infected With Cytoplasmic Sex-Ratio Distorters 320:309–320.

Escudero, M., D. a R. Eaton, M. Hahn, and A. L. Hipp. 2014. Genotyping-by-sequencing as a tool to infer phylogeny and ancestral hybridization: A case study in Carex (Cyperaceae). Molecular Phylogenetics and Evolution 79:359–367.

Etter, P. D., J. L. Preston, S. Bassham, W. a Cresko, and E. a Johnson. 2011. Local de novo assembly of RAD paired-end contigs using short sequencing reads. PloS one 6:e18561.

Felsenstein, J. 1985. Felsenstein_comparative_method_AmNat_1985.pdf. The American naturalist 125:1–15.

Fine, P. E. 1978. On the dynamics of symbiote-dependent cytoplasmic incompatibility in culicine mosquitoes. Journal of invertebrate pathology 31:10–18.

Fleury, F., F. Vavre, N. Ris, P. Fouillet, and M. Boulétreau. 2000. Physiological cost induced by the maternally-transmitted endosymbiont Wolbachia in the Drosophila parasitoid Leptopilina heterotoma. Parasitology 121 Pt 5:493–500.

Frank, S. a, and L. D. Hurst. 1996. Mitochondria and male disease.

Frank, S. a. 2012. Evolution: Mitochondrial burden on male health. Current Biology 22:R797–R799.

Frézal, L., and R. Leblois. 2008. Four years of DNA barcoding: Current advances and prospects. Infection, Genetics and Evolution 8:727–736.

Gautier, M., K. Gharbi, T. Cezard, J. Foucaud, C. Kerdelhué, P. Pudlo, J. M. Cornuet, and A. Estoup. 2013. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Molecular Ecology 22:3165–3178.

Gayral, P., J. Melo-Ferreira, S. Glémin, N. Bierne, M. Carneiro, B. Nabholz, J. M. Lourenco, P. C. Alves, M. Ballenghien, N. Faivre, K. Belkhir, V. Cahais, E. Loire, A. Bernard, and N. Galtier. 2013. Reference-Free Population Genomics from Next-Generation Transcriptome Data and the Vertebrate-Invertebrate Gap. PLoS Genetics 9.

Guindon, S., J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic biology 59:307–21.

Hebert, P. D. N., A. Cywinska, S. L. Ball, and J. R. deWaard. 2003. Biological identifications through DNA barcodes. Proceedings. Biological sciences / The Royal Society 270:313–321.

Hedges, L. M., J. C. Brownlie, S. L. O'Neill, and K. N. Johnson. 2008. Wolbachia and virus protection in insects. Science (New York, N.Y.) 322:702.

Henning, F., H. J. Lee, P. Franchini, and A. Meyer. 2014. Genetic mapping of horizontal stripes in Lake Victoria cichlid fishes: Benefits and pitfalls of using RAD markers for dense linkage mapping. Molecular Ecology:5224–5240.

Henri, H., M. Cariou, G. Terraz, S. Martinez, A. El Filali, M. Veyssiere, L. Duret, and S. Charlat. 2015. Optimization of multiplexed RADseq libraries using low-cost adaptors. Genetica:139–143.

Hilgenboecker, K., P. Hammerstein, P. Schlattmann, A. Telschow, and J. H. Werren. 2008. How many species are infected with Wolbachia? - A statistical analysis of current data. FEMS Microbiology Letters 281:215–220.

Hipp, A. L., D. a R. Eaton, J. Cavender-Bares, E. Fitzek, R. Nipper, and P. S. Manos. 2014. A framework phylogeny of the American oak clade based on sequenced RAD data. PLoS ONE 9.

Hohenlohe, P. a, S. Bassham, P. D. Etter, N. Stiffler, E. a Johnson, and W. a Cresko. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS genetics 6:e1000862.

Hosokawa, T., R. Koga, Y. Kikuchi, X.-Y. Meng, and T. Fukatsu. 2010. Wolbachia as a bacteriocyte-associated nutritional mutualist. Proceedings of the National Academy of Sciences of the United States of America 107:769–774.

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics (Oxford, England) 18:337–338.

Huemer, P., M. Mutanen, K. M. Sefc, and P. D. N. Hebert. 2014. Testing DNA Barcode Performance in 1000 Species of European Lepidoptera: Large Geographic Distances Have Small Genetic Impacts. PLoS ONE 9:e115774.

Hurst, G. D. D., and F. M. Jiggins. 2000. Male-killing bacteria in insects: Mechanisms, incidence, and implications. Emerging Infectious Diseases 6:329–336.

Hurst, G. D. D., and F. M. Jiggins. 2005. Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. Proceedings. Biological sciences / The Royal Society 272:1525–34.

Ilut, D. C., M. L. Nydam, and M. P. Hare. 2014. Defining loci in restriction-based reduced representation genomic data from nonmodel species: Sources of bias and diagnostics for optimal clustering. BioMed Research International 2014.

Jiggins, F. M. 2003. Male-killing Wolbachia and mitochondrial DNA: selective sweeps, hybrid introgression and parasite population dynamics. Genetics 164:5–12.

Johnstone, R. a, R. a Johnstone, G. D. D. Hurst, and G. D. D. Hurst. 1996. Maternally inherited male-killing microorganisms may confound interpretation of mitochondrial DNA variability. Biological Journal of the Linnean Society:453–470.

Lanfear, R., H. Kokko, and A. Eyre-Walker. 2014. Population size and the rate of evolution. Trends in Ecology and Evolution 29:33–41.

Li, H., and R. Durbin. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) 26:589–95.

Luca, F., R. R. Hudson, D. B. Witonsky, and A. Di Rienzo. 2011. A reduced representation approach to population genetic analyses and applications to human evolution. Genome research 21:1087–98.

Mackay, T. F. C., S. Richards, E. a. Stone, A. Barbadilla, J. F. Ayroles, D. Zhu, S. Casillas, Y. Han, M. M. Magwire, J. M. Cridland, M. F. Richardson, R. R. H. Anholt, M. Barrón, C. Bess, K. P. Blankenburg, M. A. Carbone, D. Castellano, L. Chaboub, L. Duncan, Z. Harris, M. Javaid, J. C. Jayaseelan, S. N. Jhangiani, K. W. Jordan, F. Lara, F. Lawrence, S. L. Lee, P. Librado, R. S. Linheiro, R. F. Lyman, A. J. Mackey, M. Munidasa, D. M. Muzny, L. Nazareth, I. Newsham, L. Perales, L.-L. Pu, C. Qu, M. Ràmia, J. G. Reid, S. M. Rollmann, J. Rozas, N. Saada, L. Turlapati, K. C. Worley, Y.-Q. Wu, A. Yamamoto, Y. Zhu, C. M. Bergman, K. R. Thornton, D. Mittelman, and R. a. Gibbs. 2012. The Drosophila melanogaster Genetic Reference Panel. Nature 482:173–178.

Margulis, L. 1970. Origin of Eukaryotic Cells. Yale University Press.

McCluskey, B. M., and J. H. Postlethwait. 2014. Phylogeny of Zebrafish, a "Model Species," within Danio, a "Model Genus." Molecular Biology and Evolution 32:635–652.

Miele, V., S. Penel, and L. Duret. 2011. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics 12:116.

Moore, W. 1995. Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees. Evolution 49:718–726.

Narita, S., M. Nomura, Y. Kato, and T. Fukatsu. 2006. Genetic structure of sibling butterfly species affected by Wolbachia infection sweep: evolutionary and biogeographical implications. Molecular ecology 15:1095–108.

Narum, S. R., C. A. Buerkle, J. W. Davey, M. R. Miller, and P. a. Hohenlohe. 2013. Genotyping-by-sequencing in ecological and conservation genomics. Molecular Ecology:n/a–n/a.

Neale, D. B., K. a Marshall, and R. R. Sederoff. 1989. Chloroplast and mitochondrial DNA are paternally inherited in Sequoia sempervirens D. Don Endl. Proceedings of the National Academy of Sciences of the United States of America 86:9347–9349.

O'Neill, S. L., A. A. Hoffman, J. H. Werren, and others. 1997. Influential passengers: inherited microorganisms and arthropod reproduction. Oxford University Press.

Pante, E., J. Abdelkrim, a Viricel, D. Gey, S. C. France, M. C. Boisselier, and S. Samadi. 2014. Use of RAD sequencing for delimiting species. Heredity:1–10.

Perlman, S. J., C. N. Hodson, P. T. Hamilton, G. P. Opit, and B. E. Gowen. 2015. Maternal transmission, sex ratio distortion, and mitochondria. Proceedings of the National Academy of Sciences:201421391.

Peterson, B. K., J. N. Weber, E. H. Kay, H. S. Fisher, and H. E. Hoekstra. 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PloS one 7:e37135.

Pool, J. E., R. B. Corbett-Detig, R. P. Sugino, K. a. Stevens, C. M. Cardeno, M. W. Crepeau, P. Duchen, J. J. Emerson, P. Saelao, D. J. Begun, and C. H. Langley. 2012. Population Genomics of Sub-Saharan Drosophila melanogaster: African Diversity and Non-African Admixture. PLoS Genetics 8.

Puritz, J. B., M. V Matz, R. J. Toonen, J. N. Weber, D. I. Bolnick, and C. E. Bird. 2014. Comment : Demystifying the RAD fad:1–18.

Rambaut, A., and N. C. Grassly. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. CABIOS 13:235–238.

Romiguier, J., E. Figuet, N. Galtier, E. J. P. Douzery, B. Boussau, J. Y. Dutheil, and V. Ranwez. 2012. Fast and robust characterization of time-heterogeneous sequence evolutionary processes using substitution mapping. PLoS ONE 7:1–10.

Romiguier, J., P. Gayral, M. Ballenghien, a. Bernard, V. Cahais, a. Chenuil, Y. Chiari, R. Dernat, L. Duret, N. Faivre, E. Loire, J. M. Lourenco, B. Nabholz, C. Roux, G. Tsagkogeorga, a. a.-T. Weber, L. a. Weinert, K. Belkhir, N. Bierne, S. Glémin, and N. Galtier. 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature.

Rubin, B. E. R., R. H. Ree, and C. S. Moreau. 2012. Inferring phylogenies from RAD sequence data. PloS one 7:e33394.

Sagan, L. 1967. On the origin of mitosing cells. Journal of theoretical biology 14:255–274.

Schilling, M. P., P. G. Wolf, A. M. Duffy, H. S. Rai, C. a. Rowe, B. a. Richardson, and K. E. Mock. 2014. Genotyping-by-sequencing for Populus population genomics: An assessment of genome sampling patterns and filtering approaches. PLoS ONE 9.

Shimodaira, H., and M. Hasegawa. 1999. Letter to the Editor Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. Molecular Biology and Evolution 16:1114–1116.

Shoemaker, D. D. 2004. Decreased Diversity but Increased Substitution Rate in Host mtDNA as a Consequence of Wolbachia Endosymbiont Infection. Genetics 168:2049–2058.

Simões, P. M., G. Mialdea, D. Reiss, M.-F. Sagot, and S. Charlat. 2011. Wolbachia detection: an assessment of standard PCR protocols. Molecular ecology resources 11:567–72.

Smeds, L., and A. Künstner. 2011. A next-generation approach to the characterization of a non-model plant transcriptome. PloS one 6:e26314.

Stamatakis, A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Stouthamer, R., R. F. Luck, and W. D. Hamilton. 1990. Antibiotics cause parthenogenetic Trichogramma (Hymenoptera/Trichogrammatidae) to revert to sex. Proceedings of the National Academy of Sciences of the United States of America 87:2424–2427.

Sunnåker, M., A. G. Busetto, E. Numminen, J. Corander, M. Foll, and C. Dessimoz. 2013. Approximate Bayesian Computation. PLoS Computational Biology 9.

Szöllosi, G. J., W. Rosikiewicz, B. Boussau, E. Tannier, and V. Daubin. 2013. Efficient exploration of the space of reconciled gene trees. Systematic Biology 62:901–912.

Takahashi, T., N. Nagata, and T. Sota. 2014. Application of RAD-based phylogenetics to complex relationships among variously related taxa in a species flock. Molecular phylogenetics and evolution 80:1–8.

Takahashi, T., T. Sota, and M. Hori. 2013. Genetic basis of male colour dimorphism in a Lake Tanganyika cichlid fish. Molecular ecology 22:3049–3060.

Toonen, R. J., J. B. Puritz, Z. H. Forsman, J. L. Whitney, I. Fernandez-Silva, K. R. Andrews, and C. E. Bird. 2013. ezRAD: a simplified method for genomic genotyping in non-model organisms. PeerJ 1:e203.

Turelli, M., and a a Hoffmann. 1991. Rapid spread of an inherited incompatibility factor in California Drosophila. Nature 353:440–442.

Turelli, M., a. a. Hoffmann, and S. W. McKechnie. 1992. Dynamics of cytoplasmic incompatibility and mtDNA variation in natural Drosophila simulans populations. Genetics 132:713–723.

Vavre, F., and S. Charlat. 2012. Making (good) use of Wolbachia: What the models say. Current Opinion in Microbiology 15:263–268.

Viricel, A., E. Pante, W. Dabin, and B. Simon-Bouhet. 2014. Applicability of RAD-tag genotyping for interfamilial comparisons: empirical data from two cetaceans. Molecular ecology resources 14:597–605.

Wagner, C. E., I. Keller, S. Wittwer, O. M. Selz, S. Mwaiko, L. Greuter, A. Sivasundar, and O. Seehausen. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Molecular ecology 22:787–98.

Wang, N., M. Thomson, W. J. a Bodles, R. M. M. Crawford, H. V. Hunt, A. W. Featherstone, J. Pellicer, and R. J. a Buggs. 2013. Genome sequence of dwarf birch (Betula nana) and cross-species RAD markers. Molecular Ecology 22:3098–3111.

Wang, S., E. Meyer, J. K. McKay, and M. V Matz. 2012. 2b-RAD: a simple and flexible method for genome-wide genotyping. Nature methods 9:808–10.

Werren, J. H. 1997. Biology of Wolbachia. Annual review of entomology 42:587–609.

Werren, J. H., L. Baldo, and M. E. Clark. 2008. Wolbachia: master manipulators of invertebrate biology. Nature reviews. Microbiology 6:741–51.

Willing, E.-M., M. Hoffmann, J. D. Klein, D. Weigel, and C. Dreyer. 2011. Paired-end RAD-seq for de novo assembly and marker design without available reference. Bioinformatics (Oxford, England) 27:2187–93.

Woolfit, M. 2009. Effective population size and the rate and pattern of nucleotide substitutions. Biology letters 5:417–420.

Xiao, J., N. Wang, R. W. Murphy, J. Cook, L. Jia, and D. Huang. 2012. WOLBACHIA INFECTION AND DRAMATIC INTRASPECIFIC MITOCHONDRIAL DNA DIVERGENCE IN A FIG WASP:1907–1916.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer applications in the biosciences : CABIOS 13:555–556.

Yang, Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24:1586–1591.

Zouros, E. 2013. Biparental Inheritance Through Uniparental Transmission: The Doubly Uniparental Inheritance (DUI) of Mitochondrial DNA. Evolutionary Biology 40:1–31.

Résumé

La propagation de bactéries intracellulaires invasives peut entrainer celle des génomes mitochondriaux qui leur sont liés génétiquement au sein du cytoplasme. Cette sélection par autostop peut conduire à une réduction de la taille efficace (N_e) pour le génome mitochondrial. Elle peut également favoriser l'introgression d'une mitochondrie introduite dans une espèce suite à une hybridation. Le principal objectif de ma thèse est de quantifier ces différents effets, de manière globale, au moyen d'un large échantillonnage d'Arthropodes de Polynésie française.

Les événements d'introgressions mitochondriales sont à l'origine de discordances entre les histoires évolutives des génomes mitochondriaux et nucléaires. Afin de rechercher de telles discordances, nous avons développé des marqueurs génomiques nucléaires de type RADseq, permettant de reconstruire l'histoire des populations étudiées. J'ai pu montrer au moyen de simulations que ce type de données pouvait être utilisé pour inférer des relations phylogénétiques entre espèces (Cariou et al. 2013). Des améliorations du protocole RADseq nous ont également permis de démontrer l'applicabilité de cette méthode à de nombreux spécimens au sein de librairies hautement multiplexées (Henri et al. 2015). A partir d'analyses *in silico*, j'ai par ailleurs évalué l'importance de différents biais liés à l'utilisation de marqueurs RADseq pour estimer les diversités génétiques et proposé une méthode permettant de corriger certains d'entre eux.

A partir de ces développements, j'ai pu démontrer que sur 30 espèces de Diptères et de Lépidoptères testées à ce jour, la proximité génétique mitochondriale est systématiquement confirmée par les marqueurs nucléaires, rejetant ainsi l'hypothèse d'une introgression mitochondriale récente. Sur un plus large échantillon, nous avons en revanche mis en évidence une réduction significative du N_e mitochondrial dans les lignées infectées par Wolbachia, suffisante pour réduire le polymorphisme, mais insuffisante pour générer une réduction notable de l'efficacité de la sélection naturelle.