Phylogéographie et évolution moléculaire chez les Procellariiformes : apport à la diversification des oiseaux marins

Lucas Torres

LA ROCHELLE UNIVERSITE

ÉCOLE DOCTORALE
Euclide

Laboratoire Centre d'Etudes Biologiques de Chizé UMR7372
& Littoral Environnement et Sociétés UMR 7266

THÈSE

présentée par :

Lucas TORRES

soutenue le 14 juin 2019
pour l'obtention du grade de Docteur de l'Université de La Rochelle
Discipline : Biologie de l'environnement, des populations, écologie

Phylogéographie et évolution moléculaire chez les Procellariiformes : Apport à la diversification des oiseaux marins

---

JURY :

Sarah SAMADI	Professeur, MNHN, Rapporteur
Pierre-André CROCHET	Directeur de recherche CNRS, Rapporteur
Didier FORCIOLI	Maître de conférences, Université de Nice Sophia Antipolis, Examinateur
Pascale GARCIA	Professeur, Université de la Rochelle, Examinatrice
Vincent BRETAGNOLLE	Directeur de recherche CNRS, Université de La Rochelle Directeur de thèse
Eric PANTE	Chargé de recherche CNRS, Université de La Rochelle, Directeur de thèse

Résumé

La génétique de la conservation a pour but de protéger à la fois la biodiversité et les processus qui l'ont forgée. Il est donc nécessaire de comprendre les processus qui régissent la diversification des espèces pour mieux les protéger. Le modèle de spéciation le plus répandu est le modèle de spéciation allopatrique, qui postule qu'une barrière physique aux flux de gènes est le catalyseur de la différenciation génétique des populations. La spéciation est donc étroitement liée aux niveaux de flux de gènes et aux barrières associées. Cependant les mécanismes qui entravent les flux de gènes sont multifactoriels et encore mal connus. C'est notamment le cas chez les organismes hautement dispersifs, comme les oiseaux marins chez qui des barrières géographiques seules ne peuvent pas empêcher les flux de gènes.

Parmi les oiseaux marins, les Procellariidae (puffins et pétrels) sont particulièrement intéressants, du fait de leur longue espérance de vie (de 20 à 30 ans), leurs tendances à vivre en large colonies sur des îles isolées et leur comportement hautement philopatrique. Ils présentent également un fort enjeu de conservation. Nous avons choisi de nous focaliser sur le complexe du Puffin d'Audubon, *Puffinus lherminieri*, du nord de l'Atlantique et ses deux lignées-sœurs en océan Indien. Le complexe Atlantique comprend trois lignées, *lherminieri* qui niche dans les Caraïbes, *boydi* au Cap Vert et *baroli* dans les Açores, les Canaries et Madère. Dans l'océan Indien on trouve *bailloni* sur l'île de la Réunion et *nicolae* aux Seychelles. La systématique de ces lignées est encore débattue et les précédentes études génétiques sur ce groupe ne se basaient que sur *cob* et une localité par lignée.

Pour étudier ce complexe, nous avons mené une étude multi-locus à l'échelle de la population, afin d'obtenir assez d'information pour montrer un signal de différenciation. Un total de 276 individus a été échantillonnés, couvrant toute la répartition géographique du groupe, pour lesquels nous avons séquencé trois marqueurs mitochondriaux et six marqueurs nucléaires. Nous avons montré une forte différentiation entre les populations séparées par l'Afrique qui reste une barrière infranchissable même pour les oiseaux marins. La caractérisation des cinq lignées est retrouvée par les marqueurs mitochondriaux mais pas nucléaires, ce que l'on explique en mettant en évidence une philopatrie plus forte chez les mâles et un tri de lignée incomplet. L'absence de structuration intra-lignée remet en cause le fort niveau de philopatrie supposé de ces oiseaux. Enfin nous avons reconstruit un scénario de colonisation et divergence corrélé à des événements climatiques et océaniques précis.

Pour chacun des trois marqueurs mitochondriaux, nous avons observé des doubles pics sur les chromatogrammes. Ces doubles pics créent de l'ambiguïté dans les séquences d'ADN et donc une perte d'information. Nous devons donc comprendre leur origine pour pouvoir utiliser au mieux nos données. Nous avons mis en avant deux phénomènes biologiques expliquant la présence de ces doubles pics. Le premier est l'existence de copies de gènes mitochondriaux dans le génome nucléaire (des numts). Nous avons empêché le séquençage de numts en digérant l'ADN nucléaire et ainsi obtenu des chromatogrammes exempts de doubles-pics pour *cob* et *co1* mais pas pour la région de contrôle.

Le séquençage involontaire de numts est probablement courant, en particulier dans les études d'oiseaux, car leur sang est pauvre en mitochondries. Pourtant ce phénomène est rarement mentionné et lorsque c'est le cas, différents traitements sont appliqués selon les auteurs pour corriger ces doubles pics. Afin de comprendre précisément l'impact de ces doubles-pics et des différentes méthodes de traitement de ceux-ci, nous avons réalisé une analyse comparative de ces traitements sur les différentes mesures de génétique de la conservation traditionnellement utilisées.

Le second phénomène expliquant les doubles pics est une duplication mitochondriale. Une région dupliquée a en effet été trouvée dans le génome mitochondrial de quatre espèces d'albatros (Procellariiformes: Diomedeidae). Cette duplication couvrait une partie de *cob*, *nad6* et la région de contrôle. Pour étudier ce phénomène plus en profondeur, nous avons séquencé le mitogénome complet d'un individu de *Puffinus lherminieri*, en utilisant une PCR long-range et la technique de séquençage MinION (ONT). Cette technique produit de longs fragments (jusqu'à 900 kb) mais peu précis (taux d'erreur de 1%). Nous avons utilisé des fragments Illumina (100 pb mais avec un taux d'erreur de 0.01%) en complément. Nous avons ainsi mis en évidence un phénomène d'hétéroplasmie mais aussi apporté une preuve directe de la présence d'une duplication dans le génome de *Puffinus lherminieri*. Cette dernière est différente de celle trouvée chez les Albatros, suggérant une évolution complexe au sein des Procellariiformes.

Table des matières

Remerciements

Je tiens à remercier mes directeurs de thèse, Eric et Vincent, pour m'avoir tout d'abord permis de me lancer dans cette grande aventure qu'est une thèse et pour m'avoir ensuite soutenu et guidé tout au long de ces trois années. Merci d'avoir cru que j'en étais capable. Merci pour nos échanges, vos idées, vos commentaires et votre appui.

Merci à toutes les personnes qui ont permis que ce travail de thèse soit réalisé. Merci à celles qui m'ont aidé en laboratoire : Cécile pour sa confiance, son sourire, sa bonne humeur, ses blagues, ses gâteaux et plus encore, Marie pour sa joie, sa gentillesse, ses cartons… Catherine pour m'avoir introduit aux NGS, Céline pour s'être investie dans différents projets, Maxime pour son appui essentiel, Vanessa pour ses conseils et son aide, Stéphanie pour son aide sur les extractions et les sexages, Colette pour sa gentillesse, Emmanuelle pour ses crêpes, Charline pour son humour et sa vice-direction et Fred pour m'avoir laissé faire des allers-retours dans le laboratoire librement.
Merci à Carine, Licia et à toutes les personnes qui ont effectué le terrain qui a permis de récolter les précieux échantillons : Francis Zino, Will MacKin, Julie Tourmetz et Hadoram Shirihai. Merci à Jacob Solis, Andreanna et Terry pour nous avoir fourni les échantillons manquants.
Merci à tous ceux qui m'ont apporté leur avis éclairé et de précieux conseils pour améliorer mon projet : Amélia, Joan, Pierre-André Crochet et Sarah Samadi.
Merci à l'université de La Rochelle, au ministère de l'enseignement supérieur, au Parc Naturel Régional de la Martinique et DEAL Martinique et à plusieurs autres organisations pour avoir fourni les financements nécessaires à cette thèse.
Tout ce travail a été effectué dans un institut qui repose sur plusieurs personnes. Merci donc à Xavier et Christophe pour l'avoir dirigé pendant mon séjour ici, à Delphine, Annie, Pascale, Marlène et Martine pour gérer parfaitement et avec beaucoup de gentillesse les tâches administratives, à Arnaud et André pour s'occuper de toute la partie informatique, à Simon et Patrice pour l'atelier. Bien sûr un énorme merci à Christophe et Arnaud pour leurs délicieux repas et leur tolérance envers mes jeux de mots.
Je remercie chaleureusement toutes les personnes qui ont eu l'incroyable bonheur de partager un bureau avec moi, en commençant par Juliette qui a eu à supporter le plus de mes jeux de mots et de mes humeurs et m'a offert beaucoup de joie en échange , Sophie qui a dû en profiter tout autant et s'est montrée très attentionnée en retour, Onde qui nous a fait profiter de sa joie de vivre, Edo avec qui j'ai adoré échangé, Landry qui m'a fait découvrir les joies de l'ornithologie, Léa qui est la meilleure successeure qu'on puisse rêver, Carine et Licia qui m'ont accueilli et protégé à mes débuts, Jessica qui a beaucoup égayé le bureau par ses anecdotes, Amandine qui a tenté de me sociabiliser du mieux qu'elle a pu, Adrien qui m'a rappelé le bonheur du début de thèse, Bruno qui a été le copain-geek que je n'attendais pas et Erika qui a toujours eu l'humour nécessaire pour rigoler avec moi. Merci à tous pour votre bonne humeur et pour tout ce que vous m'avez apporté, ce bureau me manquera.
La thèse a été enrichie de très beaux moments ludiques et je tiens à remercier également Baptiste pour m'avoir abreuvé de nourriture, de blagues, de points de vue et de jeux,

Christophe pour ses jets de dés et ses blagues, Pierre-Loup pour m'avoir replongé dans les jeux de rôles, Mme Jan pour d'autres types de jeux, Gaëtan pour toujours avoir sorti son sax au bon moment, Martin pour m'avoir laissé sa part de viande et Amélie pour avoir ri à vraiment chacune de mes blagues. J'espère qu'on aura l'occasion de faire d'autres parties.
Enfin il y a toutes les personnes qui ont fait qu'il a toujours été agréable pour moi de venir au CEBC. Je remercie donc Julien pour avoir toujours été là pour moi, Nathan pour les incroyables fous-rires qu'il m'a procuré, Candice pour ses blagues en accord avec les miennes, Eugénie avec qui on s'est très bien compris, Léa qui a toujours su garder le sourire, Anaïs pour sa gentillesse, Alexandre pour sa feinte méchanceté, Julie pour sa douceur, Emmanuelle pour son naturel, Joris pour sa bonne-humeur permanente, Karine pour sa sincérité, Tim pour son riche intérieur, Meumeu pour son riche extérieur, Gaëtane pour m'avoir sauvé la vie, Isabelle pour l'avoir égayée, Karine pour m'avoir appuyé, Solenne pour sa confiance et pour avoir été ma confidente de dernière minute, Fabien pour la fois où il a dit la vérité, Maxime pour son altruisme, Maxime 2ème du nom pour son sourire, Thierry pour son empathie, Marilyn pour son ouverture d'esprit, Florian pour son esprit, Joffrey pour son soutien, Yves pour son humour, Rémi pour son écoute, Pierre pour m'avoir accepté, Héloïse pour m'avoir bien cerné, Loriane pour sa générosité, Camille pour ses idées, Rui pour représenter les Portugais, Dimitri les Allemands, Thomas pour sa capacité à rire de tout, Paul pour son abnégation, Ryan pour son intégration, Yan pour ses blagues, Charlie pour sa gentillesse, Sylvie pour son attention, Matthias et Marion pour leur hospitalité, Marine pour avoir surmonté mes pires blagues, Sophie pour sa spontanéité, Bertille pour m'avoir démontré que j'avais tort sur les stagiaires, les monstres et les bretons, Adriana pour m'avoir prévenu de ce qu'était une fin de thèse, Romain pour ses paroles douces, Gildas pour m'avoir admis à Secondigné, Alexis pour ses conversations et son humour, Nicolas et Stan pour m'avoir fait rire de deux façons différentes et tous ceux que j'ai pu oublier pour ne pas m'en vouloir. Je n'aurais jamais imaginé avoir tant de monde à remercier, merci à tous.
Merci à Mme Morisset pour m'avoir loué un toit, un frigo, un lit, une prise télé et d'autres choses qui m'ont moins servies.
Merci à tout mon arbre phylogénétique, frère Ro, frère Da, Mère Man, Père Pa, Tante Tie, Grand-mère Mie 1, Grand-mère Mie 2, Belle-sœur Elo, Tonton Louis, Tante Jacquie, Fille Rousquille et Fille Pénélope, Neveu Oban, et merci les copains du pays, Franck et Betty, Gautier et Mélanie, Tom et Pauline, Gaultier et Personne-que-je-sache. Merci à tous pour m'avoir soutenu et appuyé sans savoir exactement ce que je faisais.

Chapitre I

Introduction générale

Depuis le début XIX^{ème} siècle, on constate une accélération du taux d'extinction des espèces animales et végétales à l'échelle de la planète. En effet, un minimum de 322 espèces de Vertébrés terrestres ont disparu depuis le XVI^{ème} siècle, tandis que les populations des espèces restantes, hormis l'Homme, auraient diminué de 25% (Dirzo et al. 2014) et de 60% pour les vertébrés (Living Plannet index, WWF 2018). Les invertébrés ne sont pas mieux lotis puisque 67% des populations suivies montrent une diminution moyenne d'abondance de 45% (Frankham et al. 2004). Le taux d'extinction actuel serait entre 8 et 100 fois plus grand qu'il y a 100 ans (Ceballos et al. 2015). La grande vitesse de disparition de ces espèces, qui n'est donc pas contrebalancée par l'émergence de nouvelles espèces, fait que ce phénomène est qualifié de 6ème extinction de masse (Jablonski & Raup 1995, Leakey & Lewin 1995, Thomas et al. 2004, Wake & Vredenburg 2008). L'extinction est donc semblable à celle qui a vu la disparition des dinosaures non-aviens il y a 65 Millions d'années. La particularité de cette 6ème extinction est qu'elle est due uniquement aux activités anthropiques.

La biologie de la conservation est une discipline qui a pour but la conservation de la biodiversité actuelle et de l'ensemble des processus qui ont façonné et qui maintiennent aujourd'hui cette diversité (Moritz 2002). La conservation a bien sûr un volet pratique qui se manifeste par exemple par des entreprises telles que celles de l'UICN qui entre autres évalue le risque d'extinction des espèces. L'UICN utilise différents critères tel que l'aire de distribution ou la taille de population pour établir le niveau de danger d'extinction d'une population. Cet inventaire est nécessaire afin de pouvoir déterminer les populations qui doivent faire l'objet de mesure de conservation particulières et sur lesquelles doivent se focaliser les politiques de conservation. La biologie de la conservation permet en effet de mettre en place des mesures de protection des espèces comme l'interdiction de chasse ou de prélèvement, l'arrêt de production de produits affectant la santé des espèces, la mise en place de dispositifs qui facilitent l'installation ou le mouvement de certaines espèces… La biologie de la conservation permet également une diffusion au grand public et un éveil des consciences sur les enjeux actuels de perte de diversité et de changement global.

La biologie de la conservation repose toutefois d'abord sur l'étude des organismes et des populations actuels. Ces études relèvent de différents champs disciplinaires. L'écologie permet de comprendre les diverses interactions entre les organismes et leurs environnements, ou entre les organismes eux-mêmes. La nature et le fonctionnement de ces interactions nous permet de déterminer l'impact que pourrait avoir une modification, même minime, de l'environnement dans lequel évoluent ces organismes. L'écophysiologie et l'écotoxicologie en sont des dérivés, étudiants respectivement les réponses comportementales et physiologiques des organismes à leur environnement et l'influence de différents produits toxicologiques sur les organismes. L'écologie des populations permet de connaître l'état actuel des tailles de populations et donc d'estimer la menace qui pèse sur ces populations. La biogéographie permet de connaître la répartition des différentes populations d'une espèce ainsi que les relations que ces populations entretiennent, la facilité et la vitesse auxquelles ces populations échangent des migrants par exemple. Enfin la génétique de la conservation constitue un outil puissant qui s'appuie sur la génétique évolutive afin de répondre aux enjeux majeurs de diminution drastique de la biodiversité.

I. L'évolution moléculaire

L'évolution moléculaire est la branche de la biologie évolutive qui étudie les macromolécules d'ADN, d'ARN et protéiques. Ces molécules portent l'information génétique et contrôlent donc tout ce qui constitue le phénotype des individus. C'est une discipline relativement récente puisque les premiers articles traitant d'évolution moléculaire remontent aux années 1960 et 1970, avec les premières phylogénies moléculaires (Fitch & Margoliash 1967), les premières études de l'horloge moléculaire (Zuckerkandl & Pauling 1965), les premiers débats entre neutralistes et sélectionnistes (revues dans Kimura 1979, Gillespie 1994). L'étude de l'échelle moléculaire a par exemple apporté des éléments nouveaux à la question « quelle part de variation génétique est expliquée par la sélection naturelle ». Les données de séquençage ont ainsi montré que de nombreuses régions du génome ne sont a priori pas exprimées et donc ne sont pas liées au phénotype. De plus les mutations qui apparaissent dans les régions fonctionnelles sont le plus souvent défavorables aux individus qui les portent. La majorité des différences observées entre deux génomes correspondent donc à des mutations neutres ou quasi-neutres. Depuis le débat qui a suivi (Gillespie 1994), l'évolution moléculaire étudie les forces sélectives qui agissent à l'échelle moléculaire. L'évolution à l'échelle moléculaire a ainsi permis l'observation de forces sélectives qui agissent spécifiquement à cette échelle. Par exemple les mécanismes à l'origine de la variation génétique, mutation ponctuelles, remaniements chromosomiques, transpositions, duplications et transferts horizontaux ont été ainsi caractérisés et leur impact sur l'évolution des génomes et des organismes identifiés. De plus l'évolution moléculaire a mis en évidence que les différents phénomènes qui agissent sur l'ADN, réplication, transcription, recombinaison, sont eux-mêmes contrôlés par des gènes et donc soumis à la sélection.

La démocratisation des méthodes de séquencage de 2^{e} et 3^{e} génération (Heather & Chain 2016) a permis un changement d'ordre de grandeur des jeux de données analysés permet de tester les hypothèses avec plus de puissance et de précision. Ainsi la base de données moléculaire Genbank regroupe en 2018 plus de 250 milliards de paires de bases pour plus de 200 millions de séquences enregistrées. Bénéficier du génome complet d'une espèce permet d'obtenir de nouvelles informations, telles que la présence ou l'absence de gènes, leur nombre, leur ordre dans le génome.

L'évolution moléculaire pose essentiellement deux types de questions, selon le point de vue adopté, « Qu'est-ce que les données moléculaires nous apprennent sur l'évolution ? » et « Qu'est-ce que l'approche comparative nous append sur les gènes et les génomes ? » Cette discipline permet donc de faire le lien entre les questions sur les déterminants de la diversification des taxons, via l'outil moléculaire, et les questions d'évolution à l'échelle génomique et génétique, permettant donc de mieux comprendre cet outil. L'évolution moléculaire est un domaine vaste qui peut être appliqué à tous les organismes, toutes les échelles temporelles et tous les types de séquences.

1. *La sélection à l'échelle moléculaire*

Les molécules d'ADN sont elles-mêmes soumises à des pressions de sélection, à différents niveaux. Les sites actifs des molécules doivent être fonctionnels et la structure tridimensionnelle des molécules doivent permettre à ces sites actifs d'être correctement placés dans l'espace. Ces contraintes influent sur la valeur sélective d'une mutation et donc sur sa probabilité de fixation. Ce ne sont cependant pas des contraintes absolues et elles laissent plusieurs degrés de liberté à l'évolution des séquences. La structure d'une molécule évolue ainsi nettement moins vite que sa séquence ADN. Des séquences de protéines au moins identiques à 50% possèdent une même structure (Chothia & Lesk 1986).

Le code génétique est dit dégénéré ou redondant puisque sur les 64 codons possibles plusieurs codent pour un même acide aminé. Certaines mutations de l'ADN ne modifient donc pas la séquence en acide aminé des protéines, on dit que ces sont des mutations synonymes. Ces mutations semblent neutres puisqu'elles ne modifient pas le phénotype moléculaire, mais en réalité les changements synonymes peuvent influer sur l'expression des gènes et sont parfois sous l'emprise de la sélection. L'intensité de cette force est toutefois négligeable et ces mutations sont donc considérées comme donnant accès aux effets attendus de l'évolution sous l'hypothèse nulle de neutralité sélective. Pour interpréter en termes d'effets sélectifs les patrons de variation des séquences codantes, il faut donc comparer les taux de substitutions synonymes et non-synonymes. Des biais peuvent en plus apparaitre si l'on compare des gènes qui ont des compositions en bases différentes (Bierne & Eyre-Walker 2003). Le ratio ω de mutations non-synonymes sur synonymes donne donc la direction générale de la sélection qui agit sur une protéine. En l'absence de sélection $\omega = 1$, si la sélection positive est prépondérante alors $\omega > 1$, puisque seule la fixation d'allèles avantageux peut expliquer une évolution plus rapide que l'attendu neutre, et à l'inverse si $\omega < 1$ alors la sélection purifiante des allèles délétères est majoritairement à l'œuvre.

Etudier la sélection permet de comprendre les liens génotype-environnement. Ainsi découvrir qu'un gène est sous sélection permet de se rendre compte que l'individu est soumis à un processus d'adaptation et de le lier à l'environnement (Kryazhimskiy & Plotkin 2008). Par ailleurs un marqueur sous sélection ne pourra pas être comparé dans des analyses de génétique des populations à un marqueur sélectivement neutre, puisque leur évolution au cours du temps aura été très différente. Enfin des différences de sélection peuvent permettre de comparer deux copies d'un même marqueur afin d'en déterminer l'origine, par exemple dans le cas de numts (Bensasson et al. 2001). La sélection à l'échelle moléculaire est donc une composante essentielle de la génétique de la conservation.

2. *Dynamique et évolution des génomes*

L'analyse des données génomiques a permis de mettre en évidence des processus évolutifs spécifiques au niveau d'organisation moléculaire.

On peut distinguer trois grandes classes d'éléments fonctionnels dans les séquences génomiques : les gènes protéiques et les gènes d'ARN non-codant (ARNnc dont le produit final est un ARN), les éléments régulateurs qui déterminent l'expression des gènes et les éléments nécessaires à la structure, réplication et ségrégation des chromosomes. Chez de nombreuses espèces, la plus grande fraction du génome est constituée de séquences non-codantes. Ainsi la variabilité de taille du génome dans le monde vivant, qui va de 10^5 paires

de bases (pb) chez une bactérie intracellulaire à 7.10^{12} pb chez l'amibe *Amoeba Dubia*, est due essentiellement à des différences de quantité d'ADN non-codant. Les régions codantes recouvrent 85% du génome des bactéries mais seulement 1,2% de celui des mammifères par exemple. Le nombre de gènes n'est donc pas corrélé à la complexité d'un organisme et l'Homme possède à peu près autant de gènes que le petit nématode *Caenorhabditis elegans*, constitué de seulement 1000 cellules contre 10^{14} chez l'Homme. De plus le nombre de gènes dans un génome peut doubler lors d'événements de polyploïdisation, qui aboutissent à la duplication du génome. La perte progressive de la quasi-totalité des gènes dupliqués qui s'en suit peut se dérouler sur plusieurs centaines de millions d'années. De telles duplications se sont produites dans de nombreux taxons eucaryotes, et chez ces organismes le nombre de gènes reflète sans doute essentiellement le temps qui s'est écoulé depuis le dernier événement de polyploïdisation (Aury et al. 2006).

Les génomes sont également composés de séquences répétées. La recombinaison inégale peut mener à la formation de répétitions en tandem, c'est-à-dire deux segments d'ADN identiques côte à côte (Richard & Pâques 2000). Un autre processus qui entraine la formation de séquences répétées est la formation de pseudogènes. Un pseudogène est une séquence qui dérive d'un gène ancestral ayant perdu sa capacité à coder son produit d'expression (ARNnc ou protéine). Le plus souvent le processus de formation d'un pseudogène fait suite à une duplication qui s'est produite directement au niveau de l'ADN, ou via un intermédiaire ARN, on parle alors de rétropseudogène. Les pseudogènes sont parfois aussi abondants que les gènes. Par exemple le génome humain contient environ 20 000 pseudogènes dérivant de gènes protéiques (Zheng et al. 2007). Les pseudogènes ont perdu leurs capacités à être transcrits du fait de l'absence de sélection sur la séquence. Leur fonction étant déjà accomplie par le gène original, les séquences ne sont pas contraintes par la sélection et l'apparition de codons-stop au sein de la séquence ne sera pas contre-sélectionnée. Certains gènes peuvent être délocalisés loin de leur position d'origine, c'est par exemple le cas des numts, copies nucléaires de gènes mitochondriaux. Dans ce cas en plus de l'absence de sélection, le code génétique qui permet de traduire les séquences nucléaires n'est pas le même que celui qui permet la traduction des séquences mitochondriales. La simple transposition d'une séquence mitochondriale vers le génome nucléaire peut donc entrainer l'apparition de codons-stops. Les numts sont fréquents chez les eucaryotes avec jusqu'à 500 copies au sein du génome humain (Richly & Leister 2004). Les numts peuvent être de différentes tailles (de quelques pb à 2000 kb, Hazkani-Covo et al. 2010) et certains englobent le génome mitochondrial en entier (Verscheure et al. 2015).

La duplication est le processus par lequel un gène, un fragment de génome ou un génome entier est copié à l'identique, mais dont les copies peuvent ensuite évoluer différemment. L'existence de duplications nombreuses et de longueurs variées est très visible dans les génomes. Ainsi la plupart des génomes contiennent des familles de gènes homologues qui peuvent compter jusqu'à plusieurs centaines de gènes. Certains organismes, comme la levure *Saccharomyces cerevisiae* ont subi une duplication de leur génome complet (Wolfe & Shields 1997). Chez les eucaryotes la duplication de gènes semble être aussi fréquente que la substitution d'un nucléotide par un autre (Lynch & Conery 2000). Par exemple le génome humain contient plusieurs gènes de récepteurs aux hormones stéroïdes, qui partagent une même organisation en domaines structuraux et une similarité de séquence importante qui indiquent une origine évolutive commune. Le génome humain comprend d'autres gènes

homologues à ceux-ci, récepteurs à d'autres hormones non-stéroïdes et d'autres ne liant aucune hormone. La diversification de cette famille s'est produite par duplication de gènes ou de génomes.

Les événements de duplication de génomes complets sont rares, les éléments dupliqués sont généralement relativement courts. Le processus le plus fréquent est la recombinaison ectopique qui concerne des répétitions relativement proches sur le même chromosome. Ceci produit alors deux copies inégales, l'une avec la région entre les deux répétitions en double, l'autre sans cette région. Ainsi si le gène A se trouve entre deux répétitions de R, après recombinaison ectopique de la région R-A-R, l'on peut obtenir un chromosome avec R-A-R-A-R, et donc une duplication du gène A. On parle de duplication en tandem.

La duplication peut permettre l'évolution de nouvelles fonctions, il a même été suggéré que ce soit la principale source de nouveauté en évolution (Ohno 1999). Sous ce modèle une paire de gènes dupliqués a deux destins possibles : soit l'un est perdu et l'autre conservé, on revient donc à la situation d'avant la duplication, soit l'un gagne une nouvelle fonction et est fixé par sélection positive, tandis que l'autre conserve la fonction ancestrale. La duplication du gène d'acétylcholinestérase chez le moustique *Culex pipiens* fournit un exemple de néo-fonctionnalisation (Labbé et al. 2007, Lenormand et al. 1998). Il est toutefois difficile d'imaginer que ce modèle puisse s'appliquer de manière générale : la probabilité d'occurrence d'une mutation avantageuse avant une mutation délétère sur l'une des copies parait très faible. Or on constate que de nombreux gènes dupliqués sont effectivement conservés sur le long terme. Un troisième destin possible pour les duplications a été envisagé : la sous-fonctionnalisation (Force et al. 1999). L'idée est que la plupart des gènes ne portent pas une unique fonction mais plutôt un ensemble de sous-fonctions. Une mutation peut donc aboutir à la perte de certaines sous-fonctions sans affecter les autres. Ce modèle explique le fort taux de conservation des copies après la duplication du génome et l'existence de copies à la fonction apparemment très similaires. Enfin il existe des cas où les deux copies d'un même marqueur sont toutes les deux conservées à l'identiques et présentent la même fonction. Ainsi la présence d'une région mitochondriale dupliquée a été détectée chez plusieurs familles d'oiseaux (Gibb et al. 2013). L'ancêtre de ces oiseaux était probablement pourvu de deux copies du gène *nad6* qui intervient dans la chaine de respiration cellulaire et de la région de contrôle mitochondriale qui intervient dans la réplication des autres gènes mitochondriaux (Urantowka et al. 2018). La présence de deux copies fonctionnelles de ce gène et la région de contrôle est avancée comme ayant eu des conséquences sur la biologie des oiseaux, notamment au niveau de leur métabolisme. En effet les gènes mitochondriaux sont impliqués dans la respiration cellulaire et la région de contrôle dans leur réplication. Deux copies de ces marqueurs pourraient augmenter la production d'énergie dans la cellule.

Les numts comme les régions dupliquées produisent une ou plusieurs copies d'un marqueur qui peuvent considérablement fausser les analyses génétiques lorsqu'ils sont co-amplifiés par erreur avec les marqueurs originels (e.g. Bensasson et al. 2001, Abbott et al. 2005). D'autant plus que ces copies ne peuvent pas toujours être détectées (Bertheau et al. 2011). Il est donc essentiel de prévenir cette co-amplification involontaire lorsque c'est possible (Calvignac et al. 2011). Ces copies supplémentaires peuvent donc être un obstacle à la réalisation d'analyses de génétique de la conservation fiables en bruitant ou biaisant les données. Ces multiples copies peuvent aussi être utiles si elles sont détectées. La détection d'évenements de

transposition permet de retracer leur évolution et de les dater afin de reconstruire l'histoire évolutive des marqueurs et de la lier à celle des espèces (e.g. Abbott et al. 2005, Arctander 1995). Etudier comment se sont formées et ont évoluées ces copies est donc essentiel pour comprendre leur impact sur les données génétiques et les espèces étudiées en conservation.
La présence de multiples copies de certains marqueurs peut avoir une influence directe sur les organismes eux-mêmes. Il a par exemple été montré que les numts sont responsables de maladies génétiques chez l'homme (voir Hazkani-Covo et al. 2010), de même que l'hétéroplasmie (voir Stewart & Chinnery 2015). Au contraire les multiples copies peuvent avoir des effets bénéfiques avec les phénomènes de surfonctionnalisation existent chez d'autres organismes (voir Kondrashov 2012 pour une revue).

II. Génétique des populations et spéciation

La génétique des populations est l'étude de la structure génétique et des changements évolutifs des populations. La population est définie comme un ensemble d'individus qui montrent une unité de reproduction. La génétique des populations se focalise sur la distribution et des changements de la fréquence des versions d'un gène dans une population sous l'influence de différents mécanismes évolutifs comme la sélection. Ces changements sont le support de l'adaptation et de l'évolution et sont à la base du phénomène de spéciation, qui correspond à une évolution divergente de deux populations.

1. La diversité génétique

Les gènes sont des séquences de nucléotides dans une région donnée, un locus, d'une molécule d'ADN. La diversité génétique d'une population correspond aux différences entre les séquences de différents individus d'une même population à un même locus. Les gènes sont codés en séquences d'acides aminés qui correspondent à des protéines. Il existe aussi des séquences non-codantes qui peuvent influencer l'expression des gènes et leur traduction en protéines. Ces protéines sont le support de la physiologie, la morphologie et le comportement d'un organisme vivant, que l'on appelle le phénotype. Les différences génétiques peuvent donc entrainer des dissimilarités dans la physiologie, la morphologie ou le comportement des individus. Ces différences génétiques sont le support de l'évolution des populations puisque ce sont elles qui sont responsables de l'adaptation ou la non-adaptation des organismes à leur environnement. En effet si un changement survient dans l'environnement, seuls les individus qui présenteront le phénotype adapté survivront à cet environnement et produiront des descendants, transmettant ainsi leurs gènes. C'est le principe de sélection naturelle (voir Dobzhansky 1937).
La diversité génétique est mesurée grâce à des techniques moléculaires qui permettent de lire les séquences de plusieurs individus sur un même locus ou une même position dans la séquence (Frankham et al. 2004). Il est alors possible de comparer l'information obtenue afin de calculer le nombre de séquences différentes, la fréquence de chacune de ces séquences et le nombre de différences qui les séparent. Ces différentes informations sont nécessaires pour analyser les mesures qui permettent de décrire la diversité génétique. Le polymorphisme représente la présence d'au moins deux séquences différentes, ou allèles (Cavalli-Sforza & Bodmer 1971). Les loci polymorphes sont habituellement définis comme montrant l'allèle le

plus fréquent à une fréquence de moins de 0.99, ou 0.95 pour minimiser les problèmes d'échantillonnages. A l'inverse un locus est monomorphe dans une population si un seul allèle est présent. Une autre mesure de la diversité génétique est l'hétérozygotie moyenne. Si un individu diploïde possède deux allèles différents à un même locus, il est hétérozygote. Chaque locus d'une population montre donc une proportion d'hétérozygote comprise entre 0 et 1. L'hétérozygotie moyenne est la somme des proportions d'hétérozygotes à tous les loci pondérée par le nombre total de loci échantillonnés. Enfin la diversité allélique est le nombre moyen d'allèles différents par locus.

Conserver la diversité génétique est l'une des trois priorités principales de l'UICN. La forte diminution de la diversité génétique d'une population réduit le potentiel de cette population à s'adapter aux changements environnementaux. D'autre part une baisse de la diversité génétique est généralement associée à une augmentation du taux de consanguinité dans la population et à une baisse générale de la survie et de la reproduction. Franklin & Frankham (1998) a ainsi estimé que la taille efficace (Ne) minimale d'une population devait être au minimum de 50 individus pour une conservation optimale à court terme et de 500 individus pour une conservation optimale à long terme. Cette règle des 50/500 a été longtemps appliquée (voir Jamieson & Allendorf 2012), mais a été récemment revue à 100/1000 (Frankham et al. 2014). Ainsi les mesures de diversité génétique sur des larges populations montrent un fort polymorphisme et une forte hétérozygotie moyenne et diversité allélique (Frankham et al. 2004). Ces grandes variations génétiques se manifestent par des fortes variations morphologiques, physiologiques et comportementales. Au contraire les espèces constituées de petites populations ou de populations qui ont subi une forte diminution de taille aujourd'hui enrayée, un goulot d'étranglement, montrent des mesures de diversité génétique faibles (Amos & Balmford 2001). Par exemple les éléphants de mer du nord ont été chassés presque jusqu'à l'extinction mais les populations ont aujourd'hui retrouvé un niveau stable. Cette espèce montre aujourd'hui des niveaux de diversité génétique faibles comparés aux espèces n'ayant pas subi de goulot d'étranglement (Stoffel et al. 2018). De même les lions asiatiques, considérés comme en danger d'extinction par l'UICN, montrent une faible diversité génétique par rapport aux populations de lions d'Afrique (Wildt et al. 1987). La plupart des espèces menacées ont une diversité génétique plus faible que les espèces proches qui ne sont pas en danger d'extinction. Ainsi Frankham et al. (2002) ont montré que sur 170 espèces menacées étudiées, 77% montraient une diversité génétique plus faible que des espèces proches non-menacées. En tout, la diversité génétique des espèces menacées représente environ 60% des espèces non-menacées. Cette baisse de la diversité génétique s'explique notamment par la diminution de taille efficace de population des espèces menacées, puisque les individus qui disparaissent sans laisser de descendants correspondent à des allèles qui disparaissent de la population.

2. La diversité génétique, support de la spéciation

La diversification du vivant résulte de la modification des populations sous l'effet des mutations, de la dérive génétique et de la sélection. La diversité génétique existante soumise à ces processus peut évoluer de manière divergente entre différentes populations.

Les taux de mutation varient d'une espèce à l'autre, d'un locus à l'autre et sans doute aussi d'un individu à l'autre (Baer et al. 2007). Ils varient aussi en fonction du milieu. Les taux de

mutation sont déterminés par l'efficacité avec laquelle les informations génétiques sont mal recopiées et réparées. Ces taux de mutation peuvent être estimés par des expériences particulières (Raquin et al. 2008 pour des marqueurs microsatellites) ou en comparant la divergence entre espèces sous hypothèse neutraliste. Les numts peuvent également être utilisés pour étudier l'émergence de mutations de novo (Bensasson et al. 2001).

Lors de la reproduction sexuée, sous l'effet du hasard des individus qui se reproduisent et des gamètes utilisés, on observe une évolution des caractères phénotypique de la population au cours du temps. Dans une population panmictique, chacune des informations génétiques d'une génération provient d'un tirage avec remise des informations génétique de la génération précédente. Cela suppose que la probabilité qu'un individu laisse un descendant ne soit pas influencée par le fait qu'il en ait laissé un auparavant. Ce processus entraine une variation aléatoire de la fréquence des allèles. La probabilité de tirer plusieurs fois un allèle est dépendante du nombre d'allèles et donc de la taille de la population. Plus la population est petite, plus la probabilité de perdre ou de fixer l'allèle est importante.

Dans L'Origine des espèces, Darwin (1859) considère que la sélection naturelle et la compétition sont les causes principales de la diversification biologique. L'adaptation à des environnements physiques et biotiques (proies, prédateurs, parasites, compétiteurs…) résulte dans la production de taxons divergents au moyen de la sélection naturelle. L'influence de la sélection naturelle sur la divergence des populations repose sur deux facteurs : la distribution des ressources et la compétition écologique, qui s'appliquent sur la valeur adaptative des individus. La valeur adaptative, ou sélective, d'un individu définit sa capacité à se reproduire. Un individu qui vivra vieux et aura de nombreux descendants aura ainsi une valeur sélective très forte. Un individu hybride aura une valeur adaptative plus faible et aura à gérer la compétition avec ses deux espèces parentes (Case & Taper 1986), bien que des cas de vigueur hybride existent où la rencontre d'une petite et d'une grande population avec différentes charges génétiques, mènent à des hybrides à valeur sélective supérieure à celle de leurs parents (e.g. Lohr & Haag 2015). De nombreux traits peuvent être soumis à des pressions écologiques directes, comme la taille du corps, l'habitat occupé ou la saison de reproduction.

La sélection sociale est définie comme la sélection sur des traits qui entrent en jeu dans la compétition pour une ressource (Price 2008; West-Eberhard 1983). La sélection sexuelle est un cas particulier où les ressources sont des partenaires pour la reproduction. La sélection sociale est différente de la sélection naturelle puisqu'elle intervient en l'absence de différences écologiques. Elles peuvent même être contradictoires. Considérons par exemple la queue du paon. Son envergure et ses couleurs vives en font un encombrement pour les mâles, et même un attrait pour les prédateurs. De telles contraintes auraient difficilement pu être favorisées par la sélection naturelle. Darwin (1871) suggéra que puisque cet organe était absent chez les femelles, il ne devait pas améliorer la survie ou la reproduction de son porteur. Il conclut donc qu'il conférait un avantage dans le choix du partenaire sexuel.

3. Applications de la génétique de la conservation

La génétique de la conservation permet de délimiter les unités de conservation. Le statut taxonomique de nombreuses populations est irrésolu. Des erreurs de délimitation peuvent entrainer des mesures de conservation erronées. Par exemple des espèces non-reconnues comme en danger peuvent s'éteindre, des espèces mal identifiées peuvent être hybridées avec

d'autres espèces, des populations en baisse de diversité peuvent être négligées. Par exemple le sphénodon ponctué (*Sphenodon punctatus*) en Nouvelle Zélande, s'est avéré être constitué de trois sous-espèces distinctes, dont l'une *S. p. guntheri* a été négligée et jusqu'à présenter un sérieux risque d'extinction (Daugherty et al. 1990). Il est donc nécessaire de classer les espèces et sous-espèces comme des unités séparées. Les populations qui présentent une divergence génétique significative et méritent donc des mesures de conservation indépendantes, sont généralement classées comme des Unités Evolutives Significatives (ESU, Ryder 1986). Ce concept pose néanmoins plusieurs problèmes. Des d'espèces à forte capacité de dispersion et ayant une large répartition géographique, comme de nombreux oiseaux et grands mammifères, ne seront ainsi pas caractérisées comme des ESU, alors qu'elles peuvent montrer des fortes différences adaptatives. Ces populations seront supposées reliées alors que les flux de gènes peuvent être restreints par d'autres facteurs que la capacité de dispersion et la géographie. De telles populations évolueraient indépendamment et seraient soumises aux problèmes de taille de populations évoqués plus hauts. De plus les ESU ne considèrent que deux états de divergence génétique, séparées ou non-séparées, alors que la divergence génétique est un processus long et continu. L'application des ESU est problématique dans le cas des premiers stades de divergence, ou zone grise (De Queiroz 2007). Crandall et al. (2000) ont proposé que les unités de management soient basées sur le fait que les populations soient écologiquement ou génétiquement échangeables ou remplaçables. Cette proposition permet de délimiter s'il y a différenciation adaptative, flux de gènes et si la différenciation est plus ou moins récente. Si deux populations sont adaptées à deux environnements différents alors elles ne sont pas écologiquement échangeables. Si elles sont génétiquement différenciées elles ne sont pas génétiquement échangeables. Comprendre les mécanismes qui façonnent la diversification des populations est donc essentiel pour mieux comprendre d'où vient cette diversité et comment optimiser sa conservation.

III. Diversification des espèces

La diversification des espèces, définie ici comme à la fois la divergence des espèces entre elles et la formation de la diversité intra-espèce, support de la divergence, est ce qui permet aux populations d'évoluer et de s'adapter indépendamment et joue un rôle majeur dans l'évolution et la conservation de ces espèces. Il est donc essentiel de bien comprendre les processus de diversification. Dans cette section, je commence par définir les concepts d'espèce et les mécanismes de spéciation.

1. Le concept d'espèce

Selon de nombreux auteurs, l'espèce est l'une des unités fondamentales en biologie, au même titre que le gène, la cellule ou l'organisme (Mayr 1982). Et de fait, la plupart des biologistes utilisent le concept d'espèce, comme référence taxinomique pour les naturalistes ou comme unité évolutive pour les systématiciens. Toutefois de nombreux biologistes défendent l'idée que l'espèce n'est qu'une division de la nature créée par l'Homme pour lui faciliter l'étude des organismes vivants. Ainsi Mayr (1982) a noté « Le problème des espèces peut être réduit à un simple choix entre deux alternatives: Les espèces sont-elles des réalités naturelles ou sont-elles simplement des constructions théoriques de l'esprit humain ? ». Il apparait légitime

de se demander si les assemblages de populations sont objectivement partitionnés en unités discrètes ou s'il ne s'agit que d'une vue de l'esprit.

On remarque que le partitionnement des populations en unités discrètes peut-être détecté par un œil non-averti. Ainsi plusieurs études ont montré qu'il y a une remarquable cohérence entre les espèces binominales et le groupement de populations utilisé par des populations humaines sans connaissances biologiques académiques. Diamond (1966) a par exemple montré que des habitants de Nouvelle-Guinée avaient nommé 120 espèces d'oiseaux dont 80% correspondaient à des espèces utilisées par les biologistes, certaines avec des différences morphologiques très subtiles. Des résultats similaires ont été trouvés par Majnep et Bulmer (1977), avec jusqu'à 95% de similarité entre les espèces de reptile définies par Linné et les espèces nommées par les indigènes. Les humains ne sont pas les seuls à pouvoir distinguer des espèces, les espèces elles-mêmes se reconnaissent entre elles. En effet un rouge-gorge mâle ne va tenter de se reproduire qu'avec une femelle rouge-gorge, et non pas avec un membre d'une autre espèce. De nombreux herbivores, pollinisateurs et parasites sont associés à des espèces qu'ils savent parfaitement distinguer des autres. Si cette reconnaissance par des non-scientifiques des espèces en tant qu'unités discrètes existe, elle est appuyée par des outils plus sophistiqués. L'outil statistique peut être appliqué à différentes catégories de traits-morphologiques, comportementaux, physiologiques ou génétiques- afin de déterminer si les populations forment des unités discrètes de façon objective. Ainsi les études taxonomiques ont permis de délimiter des espèces animales et végétales de façon objective, même si le niveau exact et la relation des différentes subdivisions est constamment remise en cause par le développement des outils. Il apparait donc que la nature est bien subdivisée en unités discontinues.

Si l'espèce existe bel et bien, il est difficile d'en avoir une définition claire car de nombreux concepts co-existent. Mayden (1997) a listé 24 concepts différents. Si ce débat peut paraitre philosophique, le problème est concret puisque nous ne pouvons pas étudier la formation et la composition des espèces si nous ne déterminons pas ce que sont les espèces. Des concepts d'espèces différents vont mener à des conclusions différentes sur le nombre d'espèces et leurs limites. Le premier et le plus utilisé des concepts d'espèce est le concept biologique d'espèce. Mayr (1942) l'énonce comme « Des espèces sont des groupes de populations naturelles réellement ou potentiellement interféconds, qui sont reproductivement isolés des autres groupes semblables ». Les espèces sont donc séparées par des barrières reproductives définies comme des paramètres qui empêchent l'échange de gènes avec les membres d'autres populations.

2. Le concept biologique d'espèce

Ce concept décrit les espèces comme des populations inter-fécondes. Etudier la spéciation revient donc à étudier l'évolution des mécanismes qui conduisent des sous-ensembles du réseau généalogique à ne plus échanger de matériel génétique. Ces mécanismes d'isolement reproductifs peuvent intervenir à différents stades du cycle de vie et peuvent être plus ou moins dépendants de l'environnement. Le concept d'espèce biologique considère donc les barrières aux flux de gènes comme principal mécanisme de spéciation. Ce concept est le plus utilisé car il montre plusieurs avantages. Le premier, énoncé par Mayr (1942) est que ce concept permet la diagnose d'espèces sœurs même si elles ne montrent pas ou peu de différences morphologiques. Un autre avantage est qu'il propose une explication à l'existence

des espèces. L'isolement reproducteur, en plus d'être le critère de distinction des espèces, est une cause de formation et de persistance des espèces. Toutefois le concept biologique d'espèce n'est pas exempt de problèmes, le plus évident est que ce concept n'est pas applicable aux organismes qui se reproduisent de façon asexuée. Un autre problème consiste en l'existence d'espèces allopatriques c'est à dire des populations qui ne se rencontrent jamais du fait de leur localisation géographique. S'il existe des différences morphologiques ou comportementales entre de telles populations, il est difficile de savoir si ces différences empêcheraient les deux populations d'échanger des gènes si elles se rencontraient. C'est par exemple le cas de douzaines d'espèces allopatriques de Cichlidés, dont on ne peut pas être sûr que les différences de coloration des mâles, utilisées pour la diagnose de ces espèces (Turner et al. 2001), empêcherait réellement la production d'hybrides. Il est possible que des migrations accidentelles entre populations, ou des expérimentations en élevage résolvent ce problème mais ce n'est pas applicable à toutes les espèces allopatriques. Le même argument s'applique aux chrono-espèces, il est impossible de savoir si deux populations fossiles pouvaient produire des hybrides viables. Un autre problème est que l'hybridation entre des espèces reconnues est relativement fréquente dans la nature. Par exemple 9,2% des 9672 des espèces décrites d'oiseaux, 10% des espèces animales et 25% des espèces végétales ont produit au moins un hybride avec une autre espèce dans la nature (Grant & Grant 1992, Mallet 2005). Si l'on considère que l'isolement reproductif doit être total pour distinguer deux espèces, alors la taxonomie de ces oiseaux devrait être revue. Le problème devient plus complexe lorsqu'on considère le cas des espèces en anneaux. Il s'agit de séries de populations voisines et connectées qui peuvent échanger des gènes et produire des hybrides viables, à l'exception des deux populations aux extrémités de la zone occupée. Ces dernières constituent donc des espèces, selon le concept biologique, mais sont connectées par des populations intermédiaires interfécondes. La question est donc de savoir où la limite entre les deux espèces est placée.

3. *Le concept d'espèce par reconnaissance*

Le concept d'espèce par reconnaissance stipule que les espèces sont « les populations d'organismes biparentaux individuels les plus inclusives, qui partagent un système de fertilisation commun » (Paterson 1985). Les espèces sont donc définies par des facteurs qui regroupent des populations plutôt que par des facteurs d'isolement, en l'occurrence l'ensemble des paramètres biologiques qui contribuent à la fertilisation. Ces paramètres incluent tout ce qui permet aux organismes de reconnaitre leurs partenaires potentiels, que ce soit de façon active (par la séduction) ou passive (reconnaissance des gamètes). Ce concept peut donc être considéré comme une version du concept biologique qui exclut les barrières post-zygotiques et écologiques comme facteur d'isolement reproductif et place en priorité le développement de barrières pré-zygotiques. Le principal moteur de la spéciation est alors l'allopatrie, ou spéciation par vicariance, énoncée par Mayr (1942). Ce mode de spéciation consiste en la séparation géographique d'une population en deux populations. Cette séparation peut apparaitre suite à divers événements tels que le changement climatique, la formation de montagnes, la dérive des continents… Si deux populations sont séparées par de telles barrières infranchissables pendant un temps suffisamment long, des mutations vont apparaitre de part et d'autre de ces barrières et certaines vont se fixer indépendamment dans les deux populations, provoquant leur divergence. Cette divergence peut être accélérée si les conditions

écologiques des deux côtés de la barrière sont différentes, ainsi les mutations sélectionnées ne seront pas les mêmes pour les deux populations. Si le temps de séparation est suffisant, les deux populations deviendront incapables de se reproduire entre elles même si elles se rencontraient à nouveau.

4. Le concept écologique

Le concept d'espèce écologique propose qu'une espèce soit définie comme « une lignée (ou un groupe de lignées proches), qui occupe une zone adaptative différente au minimum de toutes les autres lignées de sa zone et qui évolue séparément de toutes les autres lignées en dehors de cette zone. » (Valen 1976). Le concept écologique postule donc que le principal moteur de la spéciation est la sélection naturelle sur des facteurs écologique. L'influence de la sélection naturelle sur la divergence des populations repose sur deux facteurs : la distribution des ressources et la compétition écologique, qui s'appliquent sur la valeur adaptative des individus. La valeur adaptative, ou sélective, d'un individu définit sa capacité à se reproduire. Un individu qui vivra vieux et aura de nombreux descendants aura ainsi une valeur sélective très forte. Ce concept remédie aux problèmes d'entités écologiquement différenciées qui échangent toujours des gènes. La zone adaptative est donc le critère de distinction des espèces. Ceci permet de distinguer des espèces dans le cas d'espèces en anneau si des différences écologiques sont visibles, ou des espèces à reproduction asexuelle. Toutefois certains groupes peuvent coexister en tant qu'unités distinctes en sympatrie dans la même niche écologique sans échanger de gènes, c'est le cas par exemple d'espèces isolées dans le temps comme les cigales périodiques. De plus il est souvent difficile de déterminer si deux espèces sympatriques occupent deux zones adaptatives différentes. Par exemple les cichlidés haplochrominiens du lac Victoria ont souvent été considérés comme écologiquement identiques.

5. Le concept d'espèce cluster

Le concept d'espèce cluster génétique définit une espèce comme « un groupe d'individus génétiquement distinguable qui a peu ou pas d'intermédiaires quand il est en contact avec d'autres clusters similaires » (Mallet 1995). Ce concept est proposé davantage comme un moyen de reconnaitre les espèces, et de les séparer en unités opérationnelles taxinomiques, que de comprendre leur évolution. D'autres facteurs que l'isolement reproductif sont alors considérés comme causes du clustering, comme la sélection stabilisante, qui désavantage les individus présentant un phénotype intermédiaire entre deux populations. Ce concept est différent du concept biologique en ce qu'il permet de délimiter des espèces malgré la présence de flux de gènes, ainsi que des espèces largement ou complétement asexuées. Il répond également aux questions des espèces allopatriques. Le nombre des traits ou des gènes requis pour distinguer des espèces sympatriques selon ce concept n'est pas précisé, on peut donc considérer qu'un ou deux gènes peuvent suffire. Or il a été montré par exemple que la sélection densité- et fréquence-dépendante peut mener à la coexistence stable de génotypes distincts pour un ou deux loci (Wilson & Turelli 1986). Le concept d'espèce cluster génétique considérerait donc ici deux espèces alors qu'il ne s'agit que de variation intra-spécifique.

6. Le concept d'espèce par cohésion

Le concept d'espèce par cohésion dit qu'une espèce est « la population d'individus ayant le potentiel pour une cohésion phénotypique à travers des mécanismes de cohésion intrinsèque la plus inclusive » (Templeton 1992). Ce concept correspond donc à un concept d'espèce

cluster génétique étayé en incluant tous les facteurs qui préservent les clusters génétiques et morphologiques. Ces facteurs, appelés « mécanismes de cohésion » sont classés en deux catégories. D'abord les mécanismes « d'échangeabilité génétique » incluent tous les facteurs qui limitent la propagation de nouveaux variants génétiques par les flux de gènes, donc les barrières reproductives mais aussi les mécanismes facilitant les flux de gènes au sein de chaque lignée. L'aspect réellement nouveau du concept de cohésion correspond aux mécanismes « d'échangeabilité démographique », qui incluent tous les facteurs qui limitent la propagation de nouveaux variants génétiques par dérive génétique et sélection naturelle. Il s'agit de tous les facteurs qui déterminent la gamme d'environnements dans laquelle les individus sont potentiellement capables de survivre et de se reproduire. Ces facteurs permettent donc de délimiter des espèces par cohésion même s'il s'agit d'espèces asexuées.

7. Le concept évolutif

Le concept d'espèces évolutives décrit une espèce comme « une seule lignée de populations ou d'organismes ancêtres et descendants, qui maintient son identité d'autres lignées semblables et qui a ses propres tendances évolutives et destin historique » (Wiley 1978) d'après (Blackwelder & Simpson 1961). Ce concept diffère des précédents en donnant une plus grande importance à l'aspect évolutif des espèces. Les espèces sont les unités qui évoluent indépendamment les unes des autres. Les mécanismes de spécaiation considérés correspondent aux processus qui ont causés la séparation éolutive initiale, qu'ils soient intrinsèques ou extrinsèques aux espèces. Ce concept s'applique aussi bien aux espèces sexuées qu'aux espèces asexuées. Toutefois la quantité de divergence nécessaire pour distinguer deux populations comme des espèces n'est pas précisée. Il est donc difficile de déterminer si populations produisant des hybrides sont considérées ou non comme deux espèces.

8. Le concept de diagnose

Le concept de diagnose qui définit une espèce comme « un irréductible cluster d'organismes que l'on peut diagnostiquer d'autres clusters similaires, et au sein desquels il y a un patron parental d'ancestralité et de descendance » (Cracraft 1989). C'est donc un concept qui distingue les espèces sur des différences de traits fixées. Le principal moteur de la spéciation est ici la fixation de traits distincts chez les deux populations. Cette fixation peut être entrainée par de la sélection ou bien par de la dérive génétique, c'est-à-dire la fixation aléatoire de certains allèles au bout de plusieurs générations dans des populations de faible taille. Le nombre et la nature du trait fixé ne sont pas précisés, en théorie n'importe quel trait comme une différence de couleur ou d'un seul nucléotide dans les séquences ADN pourrait suffire à distinguer 2 espèces. L'application stricte de ce concept augmenterait donc de façon très significative le nombre d'espèces nommées. De plus ce concept ne résout pas le problème des espèces paraphylétiques puisque si, en reprenant l'exemple précédent, un trait est modifié entre les populations A2 et B, ce trait permettra de distinguer A1 et A2 de B mais pas A1 d'A2.

9. Le concept monophylétique

Le concept monophylétique définit une espèce comme « le plus petit groupe monophylétique à ancêtre commun » (Rosen 1979). Un groupe monophylétique comprend tous les individus qui descendent d'un ancêtre commun et uniquement ces individus. Il est opposé aux groupes paraphylétiques, évoqués précédemment et polyphylétiques, qui contiennent les descendants

de plusieurs ancêtres communs distincts. Les groupes monophylétiques, ou clades, sont définis par des caractères dérivés, par opposition aux caractères ancestraux, partagés par tous les membres du groupe. Ces caractères sont appelés synapomorphies. Le concept monophylétique diffère du concept de diagnose en basant la distinction d'espèces sur ces synapomorphies. Ce concept est donc parfaitement en accord avec l'histoire évolutive des taxons. Cependant il peut être difficile de déterminer si un groupe est monophylétique. L'outil le plus utilisé pour cela est la génétique. En effet les traits mesurés sur l'ADN ont deux avantages sur les traits morphologiques habituellement utilisés. D'abord les marqueurs génétiques ont une plus grande probabilité d'être sélectivement neutres et donc les changements se feront de manière constante au cours du temps, ce qui les rend utiles pour des reconstructions historiques. Ensuite les marqueurs génétiques sont très nombreux, offrant un nombre presque infini de traits potentiellement informatifs. Cependant ce nombre important peut aussi être un problème, en effet l'histoire des gènes ne correspond pas toujours à l'histoire des espèces. Comme noté par Avise et Ball (1990), après que 2 populations se soient séparées, leurs gènes passeront par des étapes successives de paraphylie et polyphylie avant de devenir réciproquement monophylétiques, l'étape où toutes les copies des gènes de chaque population sont plus proches entre elles que des copies de l'autre population. Ce problème ne peut pas être résolu simplement en échantillonnant d'avantage de gènes ou d'allèles car, tant que la monophylie réciproque n'est pas atteinte, différents gènes peuvent donner une histoire contradictoire, sans indice pour déterminer laquelle est celle qui correspond à l'histoire des espèces. Le concept monophylétique d'espèce ignore donc le problème de discordance entre monophylie des gènes et monophylie des espèces. C'est pourquoi une troisième version du concept a été introduite.

10. Le concept d'espèce généalogique

Le concept d'espèce généalogique définit une espèce comme « un groupe d'organisme basal et exclusif dont la coalescence de tous les gènes les uns avec les autres a eu lieu plus récemment qu'avec les gènes de n'importe quel groupe d'organisme en dehors du groupe et qui ne contient pas de groupes exclusif » (Baum & Donoghue 1995). Ce concept a été proposé pour diagnostiquer le statut phylogénétique des populations en utilisant les gènes. La différence avec le concept monophylétique est que le concept généalogique définit explicitement la monophylie d'un taxon comme la monophylie des gènes portés par ses membres. Bien qu'en principe définir une espèce en utilisant ce concept devrait impliquer de nombreux loci, en pratique la monophylie est déterminée en utilisant un ensemble limité de gènes. La tâche principale pour utiliser ce concept est de déterminer combien de marqueurs seront nécessaire pour diagnostiquer la monophylie d'une espèce. La formulation originelle requiert que tous les loci soient monophylétiques mais cette exigence est trop extrême car la sélection balancée, qui préserve plus de deux allèles dans une espèce, peut maintenir un ensemble de plusieurs allèles chez les descendants. Par exemple le locus MHC montre des allèles semblables chez les humains et les chimpanzés et chez les souris et les rats alors que les deux espèces de chaque paire ont divergées entre 5 et 10 millions d'années (Figueroa et al. 1988, Ayala & Escalante 1996). Sous le concept généalogique les espèces mettent donc un temps très long à se distinguer. Ainsi *Drosophila simulans* ne remplit pas les critères pour être définie comme une espèce par rapport à *D. mauritiana* même sous la règle des 50% alors qu'il existe plusieurs différences morphologiques entre ces deux taxons qui montrent un

isolement reproductif substantiel (Hudson et Coyne 2002). De plus ce concept considère comme important les mécanismes qui entrainent une divergence génétique, qui englobe donc la sélection naturelle et la dérive mais dépend aussi de paramètres intrinsèques aux gènes comme la vitesse de mutation ou le temps depuis la divergence, ce qui donne peu d'importance biologique à la scission de deux espèces. Il y a peu d'importance biologique à passer de 50% de monophylie génétique à 50,1%, contrairement aux concepts biologique ou écologique par exemple.

11. Le concept de compatibilité mitonucléaire

Le concept de compabilité mitonucléaire (Hill 2017) se focalise sur les interactions entre un ensemble de gènes mitochondriaux et nucléaires co-adaptés. L'hybridation entre deux populations dont ces ensembles de gènes ne seraient plus co-adaptés méneraient donc à des individus dont la valeur sélective serait réduite par rapport à leurs parents, ce qui permet la délimitation entre les deux espèces. Au contraire, observer de l'introgression pour ces gènes spécifiques est une preuve que les deux populations appartiennent à la même espèce. Ce concept se veut comme une actualisation du concept biologique, applicable à tous les organismes eucaryotes. Il est centré sur les processus d'isolements post-zygotiques qui semblent majoritaire chez les oiseaux chez qui ce concept a été développé (Hill 2017). Cependant ce concept ne présente pas non plus de seuil de divergence qui permette de distinguer des espèces et nécessite d'étduier spécifiquement les gènes impliqués dans le processus afin d'en attester la fonctionnalité chez des hybrides par exemple.

12. Le concept unifié

Tous ces différents concepts d'espèces, malgré les problèmes qu'ils soulèvent, ont des défenseurs de nos jours, et beaucoup de ces concepts sont au moins partiellement incompatibles. Par exemple de nombreux auteurs ont montré que l'adoption du concept biologique menait à la délimitation de moins d'espèces que l'adoption du concept de diagnose par exemple (e.g. Bremer & Wanntorp 1979, Cracraft 1989, Zink 1996). Ces incompatibilités entre les concepts sont dues aux différentes propriétés biologiques sur lesquelles sont basés les différents concepts. Chacun des différents concepts évoqués est associé à un mode de spéciation. Un concept n'a ainsi de sens que parce qu'il relie des patrons de divergences aux mécanimses qui lui sont associés. Ces mécanismes sont indépendants entre concepts. Par exemple l'isolement reproductif n'est pas nécessairement associé à l'occupation de deux niches écologiques différentes ou des différences fixées de certains traits. Ces différences sont attendues puisque les différentes propriétés utilisées présentent un plus grand intérêt selon les thématiques étudiées par les biologistes. Par exemple l'isolement reproductif sera d'une importance centrale pour les biologistes qui étudient les zones d'hybridation, alors que les différences de niches seront capitales pour les écologues et les différences morphologiques pour les paléontologues. Les biologistes adoptant une approche multidisciplinaire trouveront que tous les concepts ont des mérites et des inconvénients et sont basés sur des propriétés biologiques importantes. En effet malgré les différences qui les séparent ces concepts alternatifs ont un point commun sous-jacent qui permet de voir émerger un concept d'espèce unifié.

De Queiroz (2007) a souligné que tous les concepts d'espèce désignent les espèces comme « des segments de lignée de métapopulation évoluant séparément ». La lignée désigne ici une série d'ancêtres et descendants (Blackwelder & Simpson 1961, Hull 1980), ou plus

simplement une métapopulation étendue dans le temps. Le terme de métapopulation se réfère à une population incluant des sous-populations interconnectées (Levins 1970, Mattei et al. 2004). Le terme de segment permet de définir que des espèces donnent naissance à d'autres espèces en formant d'autres lignées. Le concept d'espèce internodale va plus loin en proposant des limites précises à l'espèce, la considérant comme un segment de branche entre deux nœuds ou entre un nœud et la fin de la branche dans l'arbre de la vie (Samadi & Barberousse 2009). Une fois que le point commun est observé et considéré comme la propriété primaire qui définit une espèce, on s'aperçoit que les propriétés sur lesquelles reposent beaucoup de concepts alternatifs sont en fait implicitement traités comme des propriétés secondaires. Ainsi si tous les concepts définissent deux espèces comme des lignées de métapopulation évoluant séparément (entre deux nœuds), pour le concept biologique ces lignées doivent en plus être isolées reproductivement, sous le concept écologique ils doivent occuper des niches écologiques différentes, et ainsi de suite. Ces propriétés secondaires différentes mènent à des concepts incompatibles en raison du temps qu'elles mettent à se mettre en place lors de la spéciation. La spéciation est un processus complexe qui découle de plusieurs mécanismes différents que sont la mutation, la sélection, la migration et la dérive. Ces mécanismes reposent sur des facteurs très divers, qu'ils soient génétiques, environnementaux ou démographiques et impliquent différents aspects de l'organisme comme la physiologie, le développement ou le comportement. Ainsi la mise en place de l'isolement reproductif entre deux populations ne mettra pas le même temps à se mettre en place que la fixation de caractères différents, et l'ordre dans lequel les différences apparaitront ne sera pas le même chez tous les taxons. C'est pourquoi différents concepts d'espèces peuvent mener à différentes conclusions. Le concept unifié et le concept internodal d'espèce permettent de retenir le point commun entre ces différents concepts et de considérer les propriétés secondaires comme des propriétés contingentes, qui peuvent être ou ne pas être acquises par les espèces durant leur existence. Ces propriétés peuvent toujours servir comme critère pour déterminer la séparation des lignées.

Le problème de définir une espèce trouve donc sa solution dans le concept unifié. L'unité d'étude et de délimitation des populations naturelles ainsi définie, la question qui se pose désormais est de savoir comment ces espèces se forment ? Qu'est ce qui fait que des lignées de métapopulation évoluent séparément ?

IV. Problématique de la thèse

Pour optimiser la conservation des espèces il est essentiel de comprendre comment la biodiversité s'est formée et comment elle s'est maintenue. La différenciation des espèces entraine des différences morphologiques ou comportementales visibles mais peut aussi entrainer des différences génomiques entre les populations. L'évolution moléculaire nous a montré que les génomes de différentes espèces montraient des propriétés différentes, que ce soit au niveau du contenu, comme avec des événements de duplication, ou au niveau des forces sélectives qui s'y appliquent. Ces différences génomiques peuvent avoir un effet direct sur la biologie des espèces. Certains d'entre eux peuvent même être à l'origine de la spéciation. Des phénomènes comme la duplication semblent avoir une évolution parallèle à

celle des espèces sans pour autant être exactement identique. Dans la première partie de cette thèse, nous allons nous demander comment les phénomènes d'évolution moléculaire évoluent en association avec la différenciation des espèces.
La formation de la biodiversité est également associée à des paramètres internes aux populations. Ainsi la taille efficace de la population ou le taux de mutation vont avoir un impact sur la vitesse d'évolution des populations et l'importance des processus à l'origine de leur divergence comme la sélection et la dérive génétique. Ces processus sont le moteur de la diversification et reposent sur la diversité génétique existante des populations. C'est pourquoi la diversité génétique est le support de l'étude de la diversification des populations. Certains processus d'évolution moléculaire peuvent également jouer sur cette diversité, soit en la faisant biologiquement évoluer, soit en la biaisant artificiellement. Pourtant l'impact de ces phénomènes sur la diversité génétique a été peu étudié. Dans la deuxième partie de cette thèse nous allons nous demander comment des phénomènes d'évolution moléculaire peuvent biaiser des analyses de diversité génétique.
Nous avons vu qu'il existe différents mécanismes d'isolement reproductif qui peuvent être dépendants de l'environnement : l'isolement écologique (spécialisation et reproduction dans des environnements spécifiques), le choix d'habitat (contraint par la compétition ou la capacité de dispersion par exemple), la sélection dans l'habitat. Il existe aussi des mécanismes d'isolement qui ne sont pas liés à l'habitat comme l'isolement temporel, l'isolement sexuel (par accouplement préférentiel par exemple), l'isolement gamétique ou la stérilité des hybrides. Ces différents mécanismes d'isolement sont associés à différents concepts d'espèce et de spéciation. Ainsi selon que sont considérés comme prioritaires les mécanismes d'isolement au niveau pré-zygotique ou post-zygotique, on emploiera plutôt le concept biologique de l'espèce ou le concept d'espèce par reconnaissance. Le concept général et le concept d'espèce internodale ont permis de rassembler les différents concepts d'espèce et donc de spéciation autour de leur point commun de lignées de métapopulations divergentes. La spéciation est donc vue maintenant comme un continuum de niveaux de divergences qui n'est pas le même et qui ne fait pas intervenir les mêmes mécanismes dans le même ordre pour les différentes espèces. Toutefois l'importance relative des différents mécanismes de différenciation dans le façonnement de la biodiversité, notamment dans les premières étapes de la spéciation, ou zone grise, n'a pas encore été élucidée. Dans la troisième partie de cette thèse nous allons nous demander quel est l'impact relatif de différents mécanismes de différenciation.

V. Les Procellariiformes

Les oiseaux constituent un modèle d'étude approprié pour étudier la spéciation. En effet les oiseaux sont relativement faciles à observer sur le terrain et ont été largement étudiés, en faisant l'un des groupes taxonomiques avec lesquels il est le plus facile d'étudier les facteurs de divergence et leurs conséquences génétiques, morphologiques et comportementales. Ainsi de nombreux biologistes s'intéressant à la spéciation ont fondé leur théories en étudiant particulièrement les oiseaux, c'est le cas d'Ernst Mayr (1942, 1963) et de David Lack (1976) par exemple. On notera aussi que dans *L'Origine des espèces* Charles Darwin (1859) reprend l'exemple des oiseaux domestiques (les pigeons) et sauvages (les pinsons) pour illustrer et

étayer son propos. Les oiseaux constituent également un groupe intéressant pour étudier le rôle du comportement sur la divergence, notamment sur le rôle de la sélection sexuelle. Le choix du partenaire est en effet un facteur important de divergence et les critères de sélection sont particulièrement étudiés chez les oiseaux (Fear & Price 1998). Cependant, la génétique de la spéciation des oiseaux est encore méconnue de nos jours, du fait notamment de temps de générations relativement longs et de difficultés pour élever un grand nombre d'oiseaux en captivité, ou d'en échantillonner un grand nombre dans la nature (Price 2008). L'information disponible est cependant suffisante pour être liée aux données écologiques et comportementales des oiseaux et aux résultats génétiques des autres groupes (revue dans Coyne & Orr 2004). Les oiseaux constituent donc un très bon modèle d'étude pour synthétiser les rôles de l'écologie, du comportement et de la génétique sur la spéciation. Dans le reste de cette section nous allons donc nous focaliser sur les facteurs de divergence chez les oiseaux, mais la plupart d'entre eux peuvent être appliqués aux autres groupes taxonomiques.

L'ordre des Procellariiformes comprend classiquement quatre familles (Brooke 2004) : les Diomedeidae (les albatros, 32 espèces dans quatre genres), les Hydrobatidae (les pétrels-tempête, 27 espèces dans six genres), les Pelecanoidae (les puffinures ou pétrels-plongeurs, quatre espèces dans un genre) et les Procellariidae (regroupant les pétrels, les puffins, les prions et les fulmarines, 108 espèces dans 14 genres). Le terme pétrel *sensu largo* incluant tous les Procellariiformes sauf les albatros (Brooke 2004, Warham 1990), ou même tous les Procellariiformes. L'ordre compte 141 espèces reconnues par BirdLife International (2018). La principale caractéristique des Procellariiformes les distinguant des autres oiseaux marins est leurs narines tubulaires externes présentes sur le bec et supposément utilisées pour l'olfaction (Lequette et al. 1989). Les Procellariiformes sont capables de boire de l'eau de mer et d'en extraire le sel grâce à des glandes qui sont présentes chez tous les oiseaux et actives chez les Procellariiformes. Une autre particularité est leur tube digestif qui ne comprend pas de jabot. Ces oiseaux se nourrissent principalement de poissons, céphalopodes, krill et d'autres organismes marins zooplanctoniques, à l'exception des deux espèces de pétrels-géants (*Macronectes*) qui se nourrissent de carcasses d'autres oiseaux.
Les Procellariiformes ont une écologie typique des oiseaux marins. Ce sont des oiseaux coloniaux et pélagiques. Ils ont un temps de génération long, ils sont longévifs, ne se reproduisent pas avant un âge de plusieurs années et ne pondent qu'un seul œuf par an ou tous les deux ans. Cet œuf ne peut être remplacé si la reproduction échoue, bien que des cas de pontes de remplacement aient été observés chez des pétrels plongeurs (Warham 1990). Les deux sexes partagent l'incubation et l'élevage du jeune. Toutes les espèces présentent une forte fidélité entre partenaires d'une année sur l'autre, une caractéristique probablement associée au grand degré de coordination nécessaire pour élever un poussin jusqu'à l'envol, la période de reproduction est longue voire très longue (le grand albatros se reproduit tous les deux ans, car il faut 11 mois pour l'incubation de l'œuf, Brooke 2004). Toutes les espèces présentent des combinaisons de blanc, noir, brun et gris. Les Procellariiformes présentent la plus grande gamme de variation de taille chez les oiseaux, de l'océanite minute (*Halocyptena microsoma*) et ses 20,5g à l'albatros royal du sud (*Diomedea epomophora*) et ses 10,3/7,7 kg (mâles/femelles) (Brooke 2004).

Les Procellariiformes représentent un enjeu majeur de conservation. Les oiseaux marins sont parmi les oiseaux déclinant le plus vite au monde depuis une trentaine d'années. Sur les 140 espèces existantes dénombrées par l'UICN, 29 (21%) sont classées comme Vulnérables, 21 (15%) comme En danger d'extinction, et 14 (10%) comme En danger critique d'extinction, soit près de la moitié des espèces considérées comme menacées. Au niveau des familles cela représente 68% des Diomedeidae, 30% des Hydrobatidae, 25% des Pelecanoidae et 51% des Procellaridae. Deux espèces sont considérées éteintes par l'IUCN, *Bulweria bifax* et *Pterodroma rupinarum* tous deux originaires de l'ile de Sainte Hélène dans l'Atlantique sud et qui se sont éteintes au XVI^{ème} siècle suite à l'arrivée des premiers colons sur l'ile (Olson 1975). Le pétrel de Jamaïque (*Pterodroma carribaea*), bien que n'étant pas considéré comme éteint, car difficile à observer, n'a pas été vu depuis 1879. Malgré tout, la récente redécouverte du pétrel de Beck (*Pseudobulweria becki*; Shirihai 2008) présumé éteint jusque-là, laisse espérer de revoir un jour le pétrel de Jamaïque. Par ailleurs, quatre espèces de puffins sont connues à l'état fossile et se sont éteintes peu après l'arrivée de l'homme, et avec lui de prédateurs tels que le rat *Rattus exulans*, sur leur île de reproduction. C'est le cas de *Puffius olsoni* et *P. holeae* aux Canaries (Rando & Alcover 2007, 2010) aux alentours de 1270, *P. spelaeus (Holdaway & Worthy 1994)* en Nouvelle-Zélande, et *P. parvus* aux Bermudes (Olson 2004) XVI^{ème} siècle. Si le déclin de populations de Procellariiformes s'est accéléré de nos jours, leur vulnérabilité a été fatale à des espèces dès leur premier contact avec l'homme.

1. Evolution du génome des Procellariiformes

De nombreuses études ont été réalisées sur l'histoire évolutive et la phylogéographie des Procellariiformes, et la plupart reposent sur l'étude de marqueurs mitochondriaux. Que ce soit pour de la phylogénie pure (Austin et al. 2004, Jesus et al. 2009, Welch et al. 2014), de la phylogéographie (Cagnon et al. 2004, Kerr & Dove 2013, Smith et al. 2007) ou de la génétique des populations (Alderman et al. 2005, Burg & Croxall 2001). Le choix des marqueurs mitochondriaux repose sur plusieurs propriétés qui les rendent théoriquement commodes pour ces types d'études (Avise et al. 1987, Moritz et al. 1987). D'abord ils sont présents en multiples copies dans une même mitochondrie et plusieurs mitochondries sont présentes dans une seule cellule. Les marqueurs mitochondriaux sont donc présents en très grand nombre dans un individu, ce qui les rend théoriquement faciles à séquencer. De plus l'ADN mitochondrial est fortement conservé au sein du règne animal, les gènes sont, sauf exceptions, présents en une seule copie, sans introns et avec de très courtes régions intergéniques (Gissi et al. 2008). Les marqueurs mitochondriaux sont considérés comme montrant un taux d'évolution optimal pour les études de biologie évolutive, assez rapide pour générer un signal sur les divergences des populations, mais assez lent pour montrer une structuration entre populations. Enfin l'ADN mitochondrial nous permet d'étudier la dispersion par les femelles puisque, chez la plupart des espèces, il est transmis uniquement de la mère aux enfants.

Pourtant on sait aujourd'hui que plusieurs marqueurs mitochondriaux des Procellariiformes sont présents en de multiples copies chez un seul individu. Les numts, ces copies nucléaires de gènes mitochondriaux, ont de grandes chances d'être trouvés chez les Procellariiformes. En effet, le sang d'oiseau est particulièrement pauvre en mitochondries (Sorenson & Quinn

1998). Si des numts sont présents, l'amplification d'ADN par PCR et le séquençage auront plus de chance de mener au séquençage des copies nucléaires, plus nombreuses, que des copies mitochondriales. Deux copies différentes des marqueurs mitochondriaux coexistant chez un même individu peuvent être le résultat du phénomène d'hétéroplasmie, où deux génomes mitochondriaux différents sont présents chez un seul individu, du fait de transmission accidentelle des mitochondries par les deux parents, par exemple. Ce phénomène est répandu chez les oiseaux (Berg et al. 1995, Gandolfi et al. 2017, Moum & Bakke 2001, Mundy et al. 1996).

Un problème relativement mieux étudié chez l'ADN mitochondrial des Procellariiformes est la présence d'une région dupliquée. Abbott et al. (2005) a en effet montré que trois marqueurs, *nad6, cob* et la région de contrôle étaient présents en double exemplaire dans le génome mitochondrial d'une espèce d'albatros (*Thalassarche melanophrys*). La présence de deux copies divergentes de la région de contrôle a depuis été indirectement suggérée dans huit autres espèces de Procellariiformes (Lawrence et al. 2008, 2014; Rains et al. 2011, Smith et al. 2007), couvrant trois des quatre familles de l'ordre. La duplication d'une partie du génome mitochondrial semble donc relativement répandue au sein des Procellariiformes. Toutefois, depuis l'étude d'Abbott et al. (2005), seulement trois études ont séquencé la région dupliquée complète, toutes chez des albatros (Eda et al. 2010, Gibb et al. 2007, Lounsberry et al. 2015) Toutes ces études ont révélées une région dupliquée à la structure identique, et donc conservée au sein de la famille des albatros. Cependant, l'ordre des gènes mitochondriaux chez les autres familles de Procellariiformes n'a jamais été étudié.

Ces trois problèmes de multiple-copies –numts, hétéroplasmie et région dupliquée– posent des problèmes dans l'interprétation des résultats d'analyses de diversité et différenciation des populations, avec le séquençage Sanger, encore largement utilisé en général et chez les Procellariiformes en particulier. En effet si deux copies d'un même marqueur d'intérêt sont présentes dans l'individu séquencé, deux cas de figure se posent pendant l'amplification par PCR. Soit la PCR amplifie préférentiellement une des deux copies, sans que l'on sache laquelle et pas nécessairement la même pour tous les individus, soit la PCR amplifie pareillement les deux copies, ce qui mène au séquençage de deux copies différentes et donc la présence de double-pics dans les chromatogrammes de la séquence. Ces doubles pics peuvent alors être traités de différentes manières, soit en favorisant aléatoirement l'une des deux bases, soit en indiquant une base ambigüe à la position concernée, soit en supprimant la position voire la séquence concernée. La présence de multiples copies mène donc soit à l'obtention de séquençage difficilement comparables, soit au bruitage voire à la perte de données.

Tous ces problèmes sont pourtant rarement mentionnés dans les études de biologie évolutive des Procellariiformes. Sur les 40 études de ce type qui utilisent des marqueurs mitochondriaux publiées après Abbott et al. 2005, seulement 11 l'ont cité. Sur ces 11, trois auteurs ont décidé de ne pas utiliser la région contrôle dans leurs analyses, sept ont utilisées des amorces spécifiques à l'une des deux copies de la région contrôle et le dernier a seulement vérifié la présence d'ambiguïtés dans les séquences. De la même manière, seulement sept études ont cité la co-amplification de numts avec l'ADN mitochondrial et trois ont mentionné le phénomène d'hétéroplasmie comme source d'ambiguïté. Dans tous les cas, l'initiative prise par les auteurs était de vérifier la présence d'ambiguïtés ou de codons-stops pour vérifier l'origine mitochondriale des séquences. Cette vérification, cependant, est mentionnée par

plusieurs autres études et devrait être réalisée à chaque fois. Le traitement final de ces ambiguïtés est rarement précisé, mais certains auteurs choisissent de supprimer les sites présentant des ambiguïtés (e.g. Kerr & Dove 2013), d'éliminer les séquences complètes des individus incluant les ambiguïtés (e.g. Genovart et al. 2007) ou bien de coder les ambiguïtés comme des données manquantes dans leur jeu de données (e.g. Wallace et al. 2017). Ces différents traitements, ainsi que l'absence de traitement, peuvent entrainer une surestimation de la diversité et différenciation génétique des populations et donc avoir un impact majeur sur les mesures de conservation à prendre.

2. Modes de différenciation chez les Procellariiformes

Les oiseaux marins en général et les Procellariiformes en particulier constituent un excellent modèle d'étude pour étudier les mécanismes de différenciation de populations. Ils peuvent en effet voyager sur des centaines, voire des milliers de kilomètres (e.g. Croxall et al. 2005). Ils rencontrent donc très peu de barrières physiques/géographiques à la dispersion et aux flux de gènes, puisque des individus peuvent facilement visiter et se reproduire avec des individus provenant de colonies différentes de la leur (e.g. Inchausti & Weimerskirch 2002). Néanmoins des preuves directes et indirectes suggèrent que la différenciation entre populations peut être grande. Par exemple la variation géographique de la morphologie est grande (revue par Del Hoyo et al. 1992), des populations sympatriques peuvent même diverger génétiquement (Smith et al. 2007), mais aussi des populations éloignées géographiquement mais dans un même océan (Abbott & Double 2003, Silva et al. 2016). Plusieurs explications possibles pour cet apparent paradoxe ont été apportées (revues dans Friesen 2015, Friesen et al. 2007a).

La première explication est que, malgré leurs capacités de dispersion, les Procellariiformes restent des oiseaux pélagiques et côtiers. Donc des populations dans deux océans peuvent être isolées par de la glace arctique ou des masses continentales. Par exemple, des populations Atlantiques et Indo-Pacifiques sont séparées par l'Isthme de Panama et l'Afrique. Ces barrières physiques aux flux de gènes semblent infranchissables même pour des oiseaux marins. Cet argument parait valable pour certains groupes d'espèces, par exemple certaines populations d'océanites de Castro (*Oceanodroma castro)* sont séparées par l'Isthme de Panama qui avec seulement 35 km de large parait une barrière infranchissable pour ces oiseaux marins (Friesen et al. 2007b). Cependant cette explication seule ne suffit pas car des structurations de populations sont montrées au sein d'un même océan, sans barrière géographique visible (Abbott & Double 2003, Silva et al. 2016).

En théorie la philopatrie, comportement des individus qui reviennent se reproduire et nicher à la colonie où ils sont nés, pourrait restreindre suffisamment les flux de gènes pour mener à la différenciation des populations. Les Procellariiformes sont connus pour être fortement philopatriques (Schreiber & Burger 2001), et il s'agit probablement de la seconde barrière aux flux de gènes chez les oiseaux. Par exemple, aucune barrière physique ou écologique à la dispersion ne semble apparaitre entre les trois colonies de Tasmanie d'Albatros à cape blanche (*Thalassarche cauta*). Pourtant ces colonies diffèrent sur des marqueurs mitochondriaux et nucléaires (Abbott & Double 2003). A nouveau ce facteur seul ne peut expliquer la différenciation de tous les Procellariiformes puisque généralement la philopatrie est associée à des différences écologiques entre les populations et que certaines espèces très

philopatriques ne montrent pas de restriction de flux de gènes, comme l'Albatros à tête grise (*Thalassarche chrysostoma,* Burg & Croxall 2001). La philopatrie est un cas extrême d'isolement par la distance, où les individus échangent des gènes avec les colonies les plus proches géographiquement de leur colonie d'origine. Selon ce modèle, la distance génétique augmenterait avec la distance géographique entre les colonies. Certains exemples vont dans ce sens, ainsi la structuration de population est plus grande chez des espèces avec une aire de répartition plus large, donc une grande distance entre colonies, que chez des espèces sœurs à aire de répartition plus restreintes, par exemple l'Albatros hurleur (*Diomedea exulans*) et l'Albatros des Antipodes (*Diomedea antipodensis*) (Friesen et al. 2007a). Mais à nouveau certaines espèces montrent une structuration génétique siginficatives entre les différentes colonies d'un même archipel, voire d'une même île, par exemple le pétrel des Galápagos (*Pterodroma phaeopygia*, Friesen et al. 2006).

Des différences de pressions de sélection peuvent perturber les flux de gènes et potentiellement mener à l'isolement reproductif, comme on l'a vu précédemment. Ces différences peuvent résulter d'une myriade de facteurs environnementaux, tels que le climat, la disponibilité des proies, les compétiteurs… Chez les oiseaux marins, qui sont dépendants de la mer, toutes ces composantes sont liées au régime océanique dans lequel ils évoluent. Ces régimes peuvent varier en température de surface, en productivité, en saisonnalité, en upwelling, en proies, en compétiteurs, et en complexité de chaîne alimentaire. Ces différences pourraient inhiber les flux de gènes soit en réduisant la valeur sélective des individus migrants d'un régime à un autre, soit en empêchant la dispersion. Par exemple Gomez-Diaz et al. (2009) ont montré que la barrière aux flux de gènes entre les populations de puffins de Scopoli (*Calonectris diomedea*) en Méditerranée et de puffins cendrés (*Calonectris borealis*) d'Atlantique, n'était pas le détroit de Gibraltar mais le Front Océanographique Almeria-Oran, à l'intérieur de la Méditerranée. D'autres barrières de ce type comme la Convergence Subtropicale ou les courants autours du Cap-Vert pourraient en effet être à l'origine de différenciation de populations de Procellariiformes (Gangloff et al. 2013, Techow et al. 2009). Mais à nouveau des restrictions aux flux de gènes existent entre des populations vivant dans les mêmes conditions océaniques, cette explication seule ne peut donc pas suffire.

Si des individus restent près de leurs colonies ou ne se déplacent que dans des zones spécifiques à leurs colonies, ils ne rencontreront jamais d'individus d'autres colonies et n'échangeront pas de gènes avec eux, favorisant la diversification. En effet plusieurs espèces sœurs diffèrent dans leurs zones de trajets, par exemple les albatros à sourcils noirs (*Thalassarche melanophris*) et les albatros de Campbell (*Thalassarche impavida*, Burg & Croxall 2001). La forte philopatrie des Procellariiformes fait que même les populations de pétrel de Hawaii (*Pterodroma sandwichensis*) issues d'îles proches montrent des trajets différents et ont commencé à diverger (Welch et al. 2011, Wiley et al. 2012). Enfin, la différence de saison de reproduction, ou allochronie, peut expliquer à elle seule la différenciation de populations sympatriques de Procellariiformes. C'est par exemple le cas des océanites de Castro (*Oceanodroma castro,* Friesen et al. 2007b). Cependant la spéciation allochronique n'est probablement pas la plus commune chez les Procellariiformes, car peu d'autres cas ont été reportés et beaucoup d'espèces se reproduisent en synchronie.

Ces différents facteurs ont tous montré un rôle plus ou moins important dans la différenciation des Procellariiformes, cependant d'autres facteurs peuvent potentiellement interférer mais

n'ont jamais été testés. Notamment l'effet de la variabilité environnementale sur les populations a été peu exploré chez les Procellariiformes. Pourtant ces oiseaux marins, nichant sur des falaises et des îles, sont extrêmement sensibles aux variations du niveau de la mer, à la présence de proies, aux destructions de nids... Les distributions et les taille des populations ont certainement été modifiées en réponse aux changements climatiques, géologiques et écologiques qu'ils ont subis depuis leur apparition. Leur comportement philopatrique en a sans doute été affecté, de même que leur plasticité phénotypique. L'effet de ces changements mérite donc d'être exploré plus avant. La présence de compétiteurs, de parasites ou de prédateurs a également été peu explorée chez les Procellariiformes. Les interactions interspécifiques pourraient empêcher les flux de gènes, ou mener à des déplacements de caractères et a donc une importance probable dans la différenciation des Procellariiformes. Peu d'études se sont également intéressées à la sélection de partenaire et à ses effets. Des différences géographiques de coloration, vocalisations et caractères sexuels secondaires ont été montrées (e.g. Barbraud & Jouventin 1998, Bolton et al. 2008, Bretagnolle & Genevois 1997) mais sans être liées aux variations de flux de gènes.

3. Présentation du complexe d'étude

Parmi les Procellariiformes, nous nous sommes particulièrement intéressés au complexe d'espèces qui regroupe le puffin d'Audubon (*Puffinus lhermineri*) et le petit puffin (*Puffinus assimilis*). Il s'agit d'espèces de petite taille, qui nichent dans des régions tropicales, subtropicales, tempérées et subantarctiques. La systématique de ce complexe a subi de nombreuses modifications depuis une première taxonomie basée sur la morphologie en 1927 (Table 1.1). Plusieurs espèces ont subi plusieurs changements de noms et sont passées par des classifications différentes, considérées tour à tour comme des sous-espèces de *lherminieri*, d'*assimilis* ou comme des espèces à part entière. La seule taxonomie moléculaire est basé sur le seul gène cytochrome-b, réalisée par Austin et al. en 2004. La taxonomie la plus récente a été construite par Onley et Scofield en 2007, basée à la fois sur l'étude moléculaire d'Austin et des analyses morphologiques. Un arbre phylogénétique basé sur le cytochrome-b a été inféré par Kawakami et al. 2018 (Figure 1.1), cet arbre est cohérent avec la systématique instaurée par Onley et Scofield et n'apporte pas de nouvelles conclusions taxonomiques.

Trois lignées sont présentes dans l'Atlantique nord, *lherminieri* dans les Caraïbes et au large du Brésil, *boydi* au Cap Vert et *baroli* aux Açores, Canaries et Madère. Le complexe est également représenté dans l'Océan Indien, aux Seychelles avec *nicolae* et à la Réunion avec *bailloni*, où des populations distinctes vivent au nord et au sud de l'île avec des saisons de reproduction différentes (Bretagnolle et al. 2000). Dans le clade de ces deux espèces on retrouve une lignée à large aire de répartition dans le Pacifique central, *dichrous*, et deux lignées au nord de l'océan Indien, *persicus* et *temptator*. Deux autres commplexes de lignées sont également présents dans l'océan Pacifique, l'un au nord regroupant *auricularis*, *newelli*, *myrtae* et *bannermani*, et l'autre au sud regroupant *tunneyi*, *assimilis*, *kermadecensis*, *haurakiensis* et *elegans*. Deux lignées ne sont pour l'instant rattachées à aucun clade : *bryani* et *opisthomelas*.

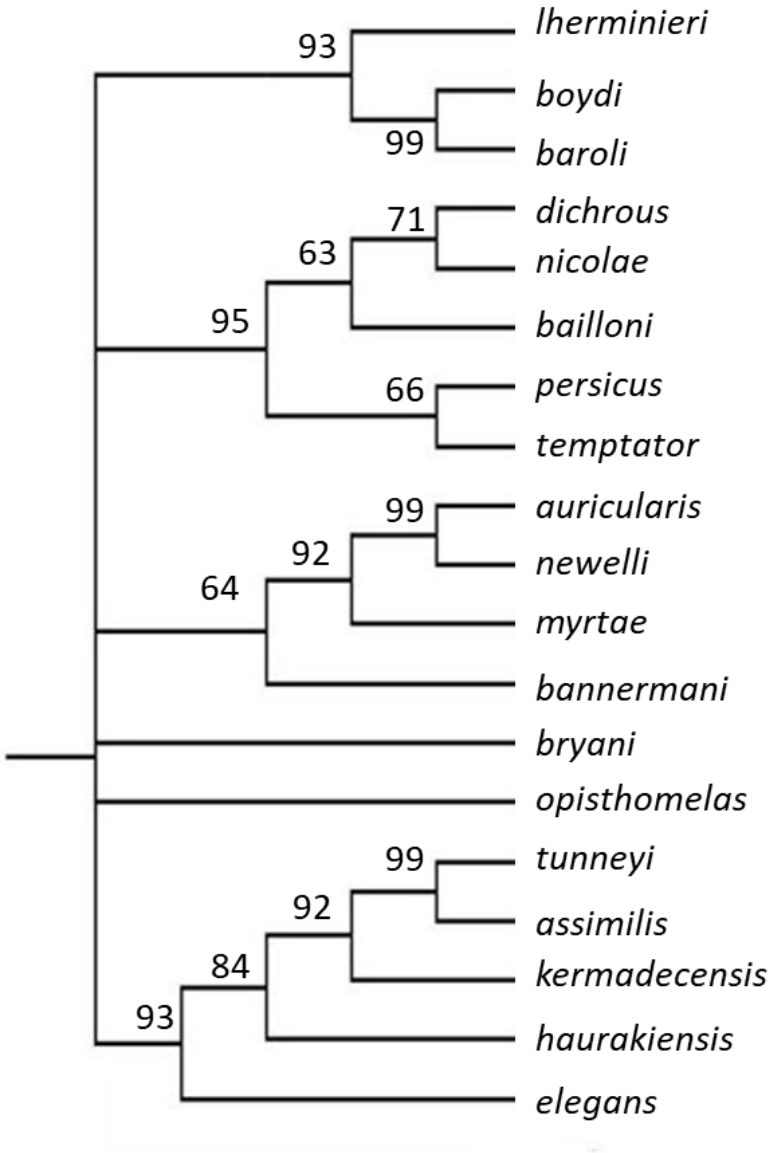
Table 1.1: Evolution de la systématique du complexe de lignées au cours du temps

Les lignées appartenant au complexe étudié sont présentées selon leur nom donné dans leur première description. Six systématiques du complexe sont présentées, indiquant pour chaque lignée le nom donné dans la systématique. Les espèces aujourd'hui considérées come synonymes, comme *nugax*, ne sont pas présentées. A partir de l'étude de Murphy (1927), toutes les lignées sont considérées comme appartenant au genre *Puffinus*. Les études suivies d'une astérisque ont menées à des actes nomenclaturaux.

Espèces et sous-espèces nominales	Description originales	Murphy 1927*	Bourne 1986	Sibley and Monroe 1990*	Shirihai et al. 1995	Austin et al. 2004	Onley et Scofield 2007
		9 sous –espèces en 2 espèces	17 sous-espèces en 1 espèce	5 espèces	16 sous-espèces en 6 espèces	13 sous-espèces en 4 espèces	14 sous-espèces en 7 espèces
Puffinus assimilis assimilis	Gould, 1838	*assimilis assimilis*	*assimilis assimilis*	*assimilis*	*assimilis assimilis*	*assimilis assimilis*	*assimilis assimilis*
Puffinus assimilis haurakiensis	Fleming et Serventy, 1943	*assimilis assimilis*	*assimilis haurakiensis*	*assimilis*	*assimilis aurakiensis*	*assimilis aurakiensis*	*assimilis aurakiensis*
Puffinus assimilis kermadecensis	Murphy, 1927	*assimilis kermadecensis*	*assimilis kermadecensis*	*assimilis*	*assimilis kermadecensis*	*assimilis kermadecensis*	*assimilis kermadecensis*
Puffinus assimilis tunneyi	Mathews, 1912	*assimilis assimilis*	*assimilis tunneyi*	*assimilis*	*assimilis tunneyi*	*assimilis tunneyi*	*assimilis tunneyi*
Puffinus bannermani	Mathews et Iredale, 1915	*lherminieri bannermani*	*assimilis bannermani*	*bannermani*	*bannermani*		
Puffinus dichrous	Finsch et Hartlaub, 1867	*lherminieri dichrous*	*assimilis dichrous*	*lherminieri*	*lherminieri dichrous*	*bailloni dichrous*	*bailloni dichrous*
Puffinus elegans	Giglioli et Salvadori, 1869	*assimilis assimilis*	*assimilis elegans*	*assimilis*	*assimilis elegans*	*assimilis elegans*	*assimilis elegans*
Puffinus heinrothi	Reichenow, 1919		*assimilis heinrothi*	*heinrothi*	*heonrothi*		*heinrothi*
Puffinus lherminieri boydi	Mathews, 1912	*lherminieri boydi*	*assimilis boydi*	*lherminieri*	*assimilis boydi*	*lherminieri boydi*	*lherminieri boydi*
Puffinus lherminieri lherminieri	Lesson, 1839	*lherminieri lherminieri*	*assimilis lherminieri*	*lherminieri*	*lherminieri lherminieri*	*lherminieri*	*lherminieri*
Puffinus lherminieri nicolae	Jouanin, 1971	*lherminieri dichrous*	*assimilis nicolae*	*lherminieri*	*nicolae nicolae*	*bailloni dichrous*	*bailloni dichrous*
Puffinus lherminieri polynesiae	Murphy, 1927	*lherminieri polynesiae*	*assimilis polynesiae*	*lherminieri*	*lherminieri polynesiae*	*bailloni dichrous*	*bailloni dichrous*
Puffinus lherminieri temptator	Louette et Herremans, 1985	*lherminieri lherminieri*	*assimilis temptator*	*persicus*	*persicus*	*bailloni temptator*	*persicus temptator*
Puffinus persicus	Hume, 1872	*lherminieri lherminieri*	*assimilis persicus*	*persicus*	*persicus*	*bailloni persicus*	*persicus*
Procellaria bailloni	Bonaparte, 1857	*lherminieri bailloni*	*assimilis bailloni*	*assimilis*	*nicolae bailloni*	*bailloni bailloni*	*bailloni bailloni*
Procellaria baroli	Bonaparte, 1857	*lherminieri lherminieri*	*assimilis baroli*	*lherminieri*	*assimilis baroli*	*lherminieri baroli*	*baroli*
Puffinus subalaris	Ridgway, 1897	*lherminieri subalaris*	*assimilis subalaris*	*lherminieri*	*lherminieri subalaris*	*subalaris*	*subalaris*

Figure 1.1: Phylogénie du complexe de puffins

Phylogénie consensuelle basée sur *cob*, obtenue par maximum de vraisemblance d'après Kawakami et al. 2018. Les nombres près des nœuds représentent les bootstraps.



VI. Objectifs de la thèse

Nous savons que chez les Procellariiformes, on peut trouver chez un même individu plusieurs copies différentes de portions du génome mitochondrial (Sorenson et Quinn 1998, Abbott et al. 2005). Ces multiples-copies résultent de processus d'évolution moléculaire qui ont certainement un impact sur l'évolution du génome mitochondrial et probablement sur la biologie, puisque la mitochondrie joue un rôle fondamental dans la respiration cellulaire, la production d'énergie mais aussi la communication, la différenciation, l'apoptose et la régulation du cycle cellulaire et la diversification des taxons eux-mêmes. Cette évolution du génome mitochondrial chez les oiseaux marins n'a pourtant été que peu étudiée. Dans la première partie de cette thèse nous allons donc nous demander :

- Quel est le patron d'évolution de ces multiples copies chez les oiseaux marins et quels sont les processus qui la régissent ? Nous avons reconstruit la région mitochondriale dupliquée complète ou incomplète chez 21 espèces de Procellariiformes afin d'étudier son évolution.

Lors d'une étude de la diversité génétique, différentes copies d'un même marqueur peuvent donc être séquencées par la méthode Sanger, ce qui peut apporter l'obtention de séquences chimériques entre deux paralogues et du bruit dans les inférences phylogéographiques. Ces séquences chimériques, qui sont ensuite utilisées pour des analyses de génétique de la conservation, peuvent en principe mener à une sous-estimation ou une surestimation de la diversité et de la différenciation génétique des populations. Pourtant l'impact exact de ces multiples-copies sur les analyses de génétique chez les oiseaux marins n'a jamais été étudié. Dans la deuxième partie de cette thèse nous allons donc nous demander :

- Quel est l'impact des multiples copies de marqueurs mitochondriaux sur les analyses de génétique de la conservation chez les oiseaux marins ? Nous avons réalisé des mesures de diversité et différenciation génétique des populations sur des jeux de données comprenant des séquences chimériques et des séquences orthologues et comparé les résultats pour déterminer l'impact des chimères.

Pour répondre à ces questions nous avons utilisé des marqueurs génétiques mitochondriaux et nucléaires. Nous avons vu dans la section précédente que ces marqueurs, porteurs de l'information génétique, sont eux-mêmes soumis à différents mécanismes qui sont à l'origine de la variation génétique et qui peuvent avoir un impact direct sur l'évolution des génomes et des organismes. Les molécules d'ADN peuvent être soumises à des phénomènes de recombinaison génétique, de duplication ou de transposition, qui peuvent modifier l'information portée. Ces mécanismes peuvent avoir des conséquences négatives sur les analyses phylogéographiques, en entrainant le séquençage de séquences paralogues et donc l'obtention de séquences chimériques. Dans cette thèse nous nous concentrons sur ce point mais la recombinaison, duplication et transposition peuvent aussi avois des conséquences fonctionnelles sur leur biologie (e.g. Hazkani-Covo et al. 2010, Kondrashov 2012, Stewart & Chinnery 2015).

Pour comprendre plus en profondeur les mécanismes qui régissent la différenciation des espèces, il serait intéressant d'étudier cette diversification chez des organismes à forte capacité de dispersion qui pourraient théoriquement s'affranchir de toute barrière géographique. Les oiseaux marins constituent un modèle intéressant dans ce sens. Les oiseaux marins sont capables de voler sur des milliers de kilomètres et ne rencontrent donc virtuellement aucune barrière physique aux flux de gènes. Par ailleurs les oiseaux marins sont dépendants à la fois du milieu marin pour se nourrir et du milieu terrestre pour se reproduire, et leur évolution est donc potentiellement soumise à différents paramètres écologiques et climatiques. De plus les oiseaux marins montrent généralement un comportement philopatrique qui pourrait théoriquement empêcher les flux de gènes. Enfin les oiseaux marins comportent plusieurs taxons à forts enjeux de conservation ; étudier et comprendre leur diversité génétique et les facteurs qui l'ont façonnée est donc essentiel. Pourtant peu d'études ont étudié l'importance relative des différents facteurs de différenciation chez les oiseaux marins. Dans la troisième partie de cette thèse nous allons donc nous demander :

- Comment est caractérisée et structurée la diversité génétique de populations d'oiseaux marins, en l'absence de barrières géographiques aux flux de gènes ? Pour y répondre, nous avons mené une étude de génétique des populations multi-locus sur un groupe de taxons tropical sur une très vaste aire de répartition en échantillonnant toutes les populations.
- Quelle est l'importance relative de différents facteurs de différenciation dans la mise en place de cette diversité génétique ? Pour y répondre, nous avons reconstruit le scénario phylogéographique du groupe de taxons étudié afin de tester l'importance relative du comportement supposé hautement philopatrique de ces taxons, de la distance des colonies chez des organismes hautement dispersifs, de la ségrégation des zones de foraging et des conditions climatiques sur des taxons dépendants de l'océan et des îles.

Chapitre II

Evidence for a duplicated mitochondrial region in Audubon's shearwater based on MinION sequencing

Mitochondrial DNA part A

Torres Lucas[1,2], Welch Andreanna J.[3], Zanchetta Catherine[4], Chesser R. Terry [5], Manno Maxime[4], Donnadieu Cécile[4], Bretagnolle Vincent[1], Pante Eric[2]

Affiliations
1 Centre d'Etudes Biologiques de Chizé, UMR 7372, CNRS & Université de La Rochelle, Villiers en Bois, France
2 Littoral, Environnement et Sociétés, UMR 7266 CNRS, Université de La Rochelle, La Rochelle, France
3 Department of Biosciences, Durham University, South Road, Durham, DH1 3LE, UK
4 US1426 Get-PlaGe, Centre INRA de Toulouse Midi-Pyrénées, Castanet-Tolosan, France
5 USGS Patuxent Wildlife Research Center, National Museum of Natural History, Smithsonian Institution, Washington, DC 20013

L'ADN mitochondrial est un marqueur très utilisé en études phylogéographiques, phylogénétiques, et de génétiques de population, car considéré comme ayant une vitesse d'évolution permettant de détecter une structuration inter et intra-population et de la distinguer de la variance intra-individuelle. L'ADN mitochondrial est également présent en de nombreuses copies chez un individu ce qui le rend plus facile à séquencer. L'ADN mitochondrial est enfin considéré comme simple-copie, c'est-à-dire que toutes les copies sont identiques au sein d'un même individu. Des phénomènes comme l'hétéroplasmie, la duplication ou les numts entrainent la formation de plusieurs copies potentiellement différentes de l'ADN mitochondrial, biaiser les inférences sur la diversité, la divergence, et la différentiation intra- et inter-populationnelle. Bien que ces phénomènes soient connus, leur impact sur les analyses de génétique des populations ont été peu étudiés (e.g. chez les insectes Cristiano et al. 2012, Haran et al. 2015, Song et al. 2008) et de nombreuses études ne mentionnent pas leur existence.

Chez les Procellariiformes, il a été montré l'existence d'une région mitochondriale dupliquée comprenant deux copies divergentes de la région de contrôle, nad6 et d'une partie du cytochrome-b (Abbott et al. 2005; Eda et al. 2010; Gibb et al. 2007; Lounsberry et al. 2015). Parmi les Procellariiformes, cette région dupliquée n'a été entièrement trouvée que chez les Albatros (Diomedeidae). Nous avons séquencé le génome mitochondrial complet d'une espèce d'une autre famille de Procellariiformes, *Puffinus lherminieri* (Procellariidae). Le but était de comprendre l'évolution de la région dupliquée le long de la phylogénie des Procellariiformes afin de mieux appréhender son impact sur les analyses de génétique des Procellariiformes.

Nous avons utilisé le séquençage par longs-fragments MinION (Oxford Nanopore Technologies, Royaume-Uni, Jain et al. 2016) afin d'obtenir des fragments suffisamment longs pour couvrir la totalité de la région dupliquée putative et fournir la preuve directe de son existence. Nous avons utilisé en complément le séquençage à courts-fragments Illumina afin d'obtenir la séquence précise de la région.

Nous avons ainsi obtenu deux mitogénomes complets, l'un de 18 kb ne présentant pas de région dupliquée et l'autre de 21 kb présentant une région dupliquée. La présence deux mitogénomes différents pourrait être due à un artéfact de PCR ou bien à un phénomène d'hétéroplasmie, ce qui a déjà été montré chez les oiseaux (Gandolfi et al. 2017) mais pas chez les Procellariiformes. La région similaire à celle trouvée chez les Albatros mais ne couvrant aucune partie du cytochrome-b. Le fait que la composition de la région dupliquée soit différente entre les deux familles suggère une évolution complexe de la duplication, qui mériterait d'être étudiée plus en détails afin de mieux appréhender son impact.

Abstract

Mitochondrial genetic markers have been extensively used to study the phylogenetics and phylogeography of many birds, including seabirds of the order Procellariiformes. Evidence suggests that part of the mitochondrial genome of Procellariiformes, especially albatrosses, is duplicated, but no DNA fragment covering the entire duplication has been sequenced. We sequenced the complete mitochondrial genome of a non-albatross species of Procellariiformes, *Puffinus lherminieri* (Audubon's shearwater) using the long-read MinION (ONT) technology. Two mito-genomes were assembled from the same individual, differing by 52 SNPs and in length. The shorter was 19kb-long while the longer was 21 kb, due to the presence of two identical copies of *nad6*, three tRNA, and two dissimilar copies of the control region. Contrary to albatrosses, *cob* was not duplicated. We further detected a complex repeated region of undetermined length between the control region and 12S. Long-read sequencing suggests heteroplasmy and a novel arrangement within the duplicated region, indicating a complex evolution of the mitogenome in Procellariiformes.

Keywords
Heteroplasmy, Control Region, Long-Range PCR, Tandem repeats, Cytochrome-b, *cob*

Many studies have been conducted on the evolutionary biology and phylogeography of Procellariiformes, a group of seabirds (albatrosses and petrels) that, like other seabirds, has high dispersal abilities and tends to be distributed over vast areas (Harrison 2000). Most such studies were based on mitochondrial markers, principally *co1*, *cob* and the Control Region (CR). MtDNA has been used for phylogenetic (e.g. Austin et al. 2004; Jesus et al. 2009; Welch et al. 2014), phylogeographic (e.g. Cagnon et al. 2004; Smith et al. 2007; Kerr & Dove 2013) and population genetics studies (e.g. Burg & Croxall 2001; Alderman et al. 2005; Ramírez et al. 2013) but use of mtDNA for such purposes relies on the assumption that the markers are single-copy (e.g. Brown 1985; Avise et al. 1987; Moritz et al. 1987; Boore 1999). However, it is known that heteroplasmy (e.g. Berg et al. 1995, Mundy et al. 1996, Moum & Bakke 2001, Gandolfi et al. 2017), mitochondrial pseudogenes or NUMTS (Sorenson & Quinn 1998), and recombination (e.g. Tsaousis et al. 2005, Sammler et al. 2011) occur in birds. For instance, Abbott et al. (2005) found evidence of a duplicated region in the mitochondrial genome of albatrosses, resulting in two copies of *nad6*, CR, and two fragments of *cob*. Two divergent copies of the mitochondrial control region have since been indirectly suggested in eight additional species of Procellariiformes (Smith et al. 2007; Lawrence et al. 2008, 2014; Burg et al. 2014; Rains et al. 2011), covering three of the four Procellariiformes families (Gangloff et al. 2013; Welch et al. 2014; Prum et al. 2015). The partial duplication of the mitochondrial genome is therefore apparently widespread within the Procellariiformes. However, apart from the study of Abbott et al. (2005), only three studies, all of albatrosses, have sequenced the complete duplicated region. Gibb et al. (2007) revised a previously-published genome of *Thalassarche melanophris* (AY158677, Slack et al. 2006), Eda et al. (2010) used primer-walking to sequence the duplicated region of three species from the genus *Phoebastria*, and Lounsberry *et al.* (2015) sequenced the complete mitochondrial genome of these same three *Phoebastria* species using Illumina and Sanger sequencing. Nearly-complete mitochondrial genomes of *Diomedea chrysostoma*, *Procellaria cinerea*, and *Pterodroma brevirostris*, which apparently lack the duplication, were also sequenced (Slack et al. 2006; Watanabe et al. 2006) (Table 2.1). Here we sequenced the complete mitochondrial genome of *Puffinus lherminieri*, providing the first complete mitogenome of a non-albatross species of Procellariiformes. We used the MinION sequencing platform (Oxford Nanopore Technologies, UK, Jain et al. 2016) to obtain reads long enough to encompass the entire duplicated region, providing direct evidence of its existence.

Table 2.1: Previously known mitochondrial duplications in Procellariiformes.

NA: information not available, CR: Control Region

Family	**Species**	***cob* duplicated**	***nad6* duplicated**	**CR duplicated**	**Object of the study**	**Genbank accession number**	**Sequencing method**	**Study**
Diomedeidae	*Diomedea amsterdamensis, D. exulans*	NA	NA	Yes	CR		PCR and Sanger-sequencing	Rains et al. 2011
Diomedeidae	*Diomedea chrysostoma*	NA	NA	NA	Nearly complete mitogenome	AP009193.1	PCR, primer-walking and shotgun sequencing	Watanabe et al. 2006
Diomedeidae	*Phoebastria albatrus*	NA	NA	Yes	CR		PCR and Sanger sequencing	Kuro-o et al. 2010
Diomedeidae	*Phoebastria albatrus, Ph. immutabilis, Ph. nigripes*	in part	Yes	Yes	Whole duplicated region	AB276044: AB276051	PCR, primer-walking and Sanger sequencing	Eda et al. 2010
Diomedeidae	*Phoebastria albatrus, Ph. immutabilis, Ph. nigripes*	in part	Yes	Yes	Complete mitogenome	KJ735512.1: KJ735514.1	Illumina sequencing, PCR and Sanger sequencing	Lounsberry et al. 2015
Diomedeidae	*Thalassarche cauta*	in part	Yes	Yes	Whole duplicated region		Restriction digest map, PCR, primer-walking, Sanger sequencing	Abbott et al. 2005
Diomedeidae	*Thalassarche melanophris*	in part	Yes	Yes	Complete mitogenome	AY158677.2	Re-check from Slack et al. (2006)	Gibb et al. 2007
Hydrobatidae	*Oceanodroma castro*	NA	NA	Yes	CR			Smith et al. 2007
Procellariidae	*Fulmarus glacialis*	NA	NA	Yes	CR		PCR and Sanger sequencing	Burg et al. 2014
Procellaridae	*Procellaria cinerea*	NA	NA	NA	Nearly complete mitogenome	AP009191.1	PCR, primer-walking and shotgun sequencing	Watanabe et al. 2006
Procellariidae	*Pterodroma brevirostris*	NA	NA	NA	Complete mitogenome	AY158678.1	PCR and Primer-walking, Sanger sequencing	Slack et al. 2006
Procellariidae	*Pterodroma macroptera gouldi*	NA	NA	Yes	CR		PCR and Illumina sequencing	Lawrence et al. 2014
Procellariidae	*Pterodroma magentae*	NA	NA	Yes	CR		PCR and Sanger sequencing	Lawrence et al. 2008

Material and Methods

We focused on a single individual of *Puffinus lherminieri* from Martinique, caught and bled within the framework of a long-term demographic program on South Martinique, 14°25'02.8"N 60°49'53.7"W (see Precheur et al. 2016 for details on the study site and study species). Genomic DNA was extracted from the blood sample using the NucleoSpin® Tissue XS Kit (Macherey & Nagel, Düren, Germany). The sample was incubated overnight in 4 mg of Proteinase K. Purified genomic DNA was eluted twice in 50 µL of TE buffer pre-heated to 70°C. To ensure optimal PCR amplification, DNA quality and quantity were measured using 1% Agarose gel electrophoresis and Nanodrop 1000 spectrophotometry.

1. *PCR amplification and Sanger sequencing of co1*

The mitochondrial genome was amplified using long-range PCR. First, we amplified 576 bp of *co1*. We chose *co1* because it was expected to be located opposite the genes previously shown to be duplicated in other Procellariformes (namely *cob*, *nad6*, CR and several tRNAs), and we wanted the duplication to occur in the middle of the long-range PCR product, so as to be sure to sequence it. Bird blood is poorly concentrated in mitochondria and is likely to contain NUMTS (Sorenson & Quinn 1998), i.e., nuclear copies of mitochondrial genes. Since the mitochondrial and nuclear genetic codes are not the same, these nuclear copies may be non-functional and may thus diverge from the mitogenome by genetic drift (Lopez et al. 1994). Co-sequencing of the nuclear and mitochondrial copies will thus lead to ambiguities in the sequences. To avoid such copies, we digested linear, nuclear genomic DNA prior to sequencing *co1* (Sorenson & Quinn 1998), using ExonucleaseV (NEB-M0345S), according to the following protocol, modified from the manufacturer's instructions and from the protocol described in Jayaprakash et al. (2015): One nanogram (ng) of DNA sample was heated to 70°C to inactivate putative Proteinase K residual of the extraction protocol. Digestion was then carried out, adding the following to the sample in a 15 µL volume: 1X NEB4 Buffer, 1 mM ATP, 0.3 U of ExoV, 0.24 mg/mL of BSA. The mix was then heated to 37°C for 48h then to 70°C for 30 min to inactivate the exonuclease. This protocol allowed us to remove SNPs in numerous individuals of the same species in an upcoming population genetics paper (Torres et al. in prep; Chapter V).

Shearwater-specific primers for *co1* were designed using Primer3 (Untergasser et al. 2012). PCRs were carried out in a total volume of 30 µL, using 1X Ex Taq Buffer (Mg2+ plus), 200 µM of dNTP, 0.8 µM of each primer, 0.025 U of TaKaRa Ex Taq® DNA Polymerase Hot-Start Version, and 60 ng of DNA extract. After an initial denaturation step of two min at 95°C, we ran 40 PCR cycles consisting of 1 min at 95°C, 1 min at 56°C and 1 min at 72°C. These cycles were followed by a 7-min final extension step at 72°C. PCR products were purified and sequenced on both DNA strands by Eurofins Genomics Munich, using the same PCR primers. Chromatograms were first visually inspected and edited using Sequencher v.5.4.1 (Gene Codes Corporation), and then sequences were assembled.

2. *Long-Range amplification of the mitogenome*

Based on *co1* sequences from this individual, as well as others sampled from the same population (Torres et al. in prep; Chapter V), we designed three long-range primers

conserved within this population, using Primer3 (Untergasser et al. 2012): Co1_test_L_221_1_LT (GCTCCTGCTTCTACTGTAGATGAGGCTAGTAGGAG) on the light strand, and Co1_test_H_344_1_EP (CGACCTAGCTATCTTCTCTCTTCACCTAGC) and Co1_test_H_296_1_EP (TTAGCCCATGCTGGAGCCTCAGTCGACCTAG) on the heavy strand.

Long-range PCR was carried out using TaKaRa LA Taq® DNA Polymerase Hot-Start Version, in a total volume of 25 µL, using 0.25 U of taq, 1x LA PCR Buffer II (Mg2+ plus), 0.21 mM of each dNTP, 0.125µM of each primer and 900 ng of purified (.e., free of nuclear DNA) mitochondrial DNA. After an initial denaturation step of one min at 94°C, we ran 35 PCR cycles consisting of 10s at 98°C, 30s at the primer-specific annealing temperature (with a temperature gradient from 60 to 72°C), and 17 to 17.5 min at 72°C. These cycles were followed by a 10-min final extension step at 72°C. PCR products were purified on agarose gel using the Monarch® DNA Gel Extraction Kit (New England Biolab) when multiple bands were visible. We followed the manufacturer's instructions, except that we added 1.5 times the volume of water and that we used 50 µL of elution buffer heated to 50°C.

3. Library preparation and sequencing

Given the small size difference expected between the two sets of amplified products (48 bp) all PCR products were pooled, and a final purification was performed using Beckman beads (Agencourt Bioscience). Quality control was performed using Qubit, Nanodrop and Fragment Analyzer. The Fragment Analyzer run revealed a peak of DNA concentration at around 12 kb, although we were expecting 18kb amplicons. The smaller size of the focus band of the Fragment Analyzer could be explained by folding of the PCR products.

Samples were prepared for sequencing following the 1D DNA protocol selecting for long reads (SQK-LSK108, ONT). DNA repair, end repair, and A-tailing (M6630 and E7546, NEB) were performed on PCR products (3.8µg input) and each step was followed by a purification using Beckman beads. Adapters were ligated using the Blunt/TA ligase master mix (M0367, NEB). A 0.6X Beckman beads purification followed the ligation step, and 620 ng of library was loaded into the flowcell. DNA was not sheared so as to maximize sequencing read length. MinION sequencing was performed as per manufacturer's guidelines using R9.5 flowcells (FLO-MIN107, ONT), by MinKnow v1.7.10 (ONT), and runs extended for up to 48 h. The MUX scan reported 982 active pores.

4. Bioinformatics

Base calling was performed with the ONT Albacore command line tool (v1.2.4) and reads were supplied in the fastq format. We trimmed adapter sequences using Porechop (v0.2.1, Wick R. Porechop, available at: https://github.com/rrwick/Porechop). We used Nanofilt (v 1.1.3, De Coster et al. 2018, available at https://github.com/wdecoster/nanofilt) to filter reads for which the average quality score was less than 11.

Read assembly was performed using Canu (Koren et al. 2016) with the ng6 platform (Mariette et al. 2012). Because we were expecting a genome 16 kb long (18 kb long if the duplication was present) we chose to exclude from the analysis all the reads longer than 25 kb. By default, Canu also removes reads shorter than 1 kb. We expected the duplicated

region to be complex and thus difficult to assemble, so assembly was run with a target coverage setting of 100X. Similarly, as the obtained reads seemed to be of relatively low quality, we increased the correction quality setting to "corMinCoverage=8." Resulting contigs and unitigs were annotated using MITOS (Bernt et al. 2013). We then compared them to the nearly-complete mitogenome of *Thalassarche melanophris* (Gibb et al. 2007) (AY158677.2) using Blast (MegaBLAST, nr database, E-value threshold: 10, identity threshold 85%) (Altschul et al. 1990).

Due to the presence of several stop-codons in the resulting contigs, we polished our Canu assembly using 100bp Illumina HiSeq 2500 reads previously obtained as a by-product of a separate, targeted enrichment project (Welch et al. in prep) on a separate individual from the same population (National Museum of Natural History, Smithsonian Institution, voucher: USNM 620721). The reads were aligned with the contigs using BWA (Li 2013), and then polished with Pilon (Walker et al. 2014). Illumina reads were mapped again on the resulting assembly sequence and drops in depth of coverage (hereon coverage) were observed at some positions. These were resolved manually by correcting the assembly with the corresponding Illumina reads with the greatest depth. The resulting sequence was again annotated using MITOS.

Results & Discussion

We obtained 148 644 raw MinION reads (SRA accession SRS3196799). After length and quality filtering, we ran Canu with 12 764 reads. To optimize running time the assembly is constructed so that 100x of the 18 expected kb is covered. This is why only 88 reads were retained for the final assembly; these had a total cumulative length of 1 748 405 bp. The median length of the retained reads was 21 203 bp. Of these 88 reads, 74 (84%) were used to assemble a contig 21 344 bp long (hereafter named Ct1; average coverage 57X, min 9X, max 74X) and 12 reads (14%) were used to assemble a contig of 18 884 bp (hereafter named Ct2; average coverage 10X, min 5X, max 12X). The two remaining orphan reads were removed from further analyses, as they corresponded to short mitochondrial sequences, slightly divergent from Ct1 and Ct2. We mapped the quality-filtered 12 764 MinION reads onto these two contigs: 88% mapped to Ct1 and 92% to Ct2. The average coverage was high, with more than 3 000 X for the two contigs (Table 2.2). We observed, however, that coverage of the first 8.4 kb of the two genomes was almost ten times lower than that of the rest of the genome (Figure 2.1a, Figure 2.2a, Table 2.2). This difference in coverage was not visible when mapping only the 88 reads used for the assembly, or when mapping the Illumina reads (see below). These two regions presented a similar GC content on the two assemblies (Table 2.2), suggesting that the first half was not more complex than the second one. We therefore suggest that this difference of coverage was due to a difficulty in sequencing encountered by the MinION device (e.g., due to secondary structures formed by mitochondrial DNA).

We mapped 277 693 Illumina reads on the two contigs (SRA acc. SRP141134), obtaining an average coverage of more than 1 300 X for both (Figure 2.1b, Figure 2.2b, Table 2.3). Using these data, we corrected 150 (on Ct1) and 177 (on Ct2) local SNPs on the assemblies. The two resulting genomes (GenBank acc. MH206162 and MH206163) were

21 144 bp long and 19 004 bp long, respectively, and consisted of 14 and 13 protein coding genes, respectively, 25 tRNA genes and 2 rRNA genes (Figure 2.1c, Figure 2.2c, Supplementary Material 2.1, Supplementary Material 2.2). The nucleotide composition of the complete genome was 32.0% A, 29.9% C, 11.8% G and 26.3% T for Ct1 and 31.1% A, 29.8% C, 12.8% G and 26.3% T for Ct2, as expected in AT-rich mitochondrial genomes (see Saccone et al. 1999). Most of the genes are encoded on the light strand, with only *nad6* and eight tRNA genes (Gln, Ala, Asn, Cys, Try, Glu, Pro and Ser) encoded on the heavy strand.

All protein-coding genes started with typical ATN codons, except for *cox2* and *nad5*, which began with the GTG codon. All protein coding genes finished with the TAA stop-codon, except for *nad4*, *nad5* and *nad1*, which ended with AGA, *nad2* and *nad6*, which ended with TAG, and *co1*, which ended with AGG (Supplementary Material 2.1 and Supplementary Material 2.2). We observed a supplementary base in the gene *nad3* that implies a translational frameshift (Mindell et al. 1998). This had already been observed in mitogenomes of several bird species, including Procellariiformes (e.g. Watanabe et al. 2006; Gibb et al. 2013).

Table 2.2 Coverage (X) of mapping of the 12 764 MinION reads used by Canu for the assembly, and percentage of GC, for the two contigs. The "first part" consists of the first 8 428 bp of Ct1 and the first 8 447 bp of Ct2, and the "second part" consists of the last 12 716 bp of Ct1 and the last 10 557 bp of Ct2.

Sequence	Average coverage	Minimum coverage	Maximum coverage	% of covered bases	% of GC
Ct1	3 189	433	5 927	100	41,76
Ct2	3 523	433	9 067	100	42,66
Ct1 first part	560	433	683	100	43,36
Ct2 first part	560	433	683	100	43,36
Ct1 second part	4 945	3 812	5 927	100	40,70
Ct2 second part	5 919	4 229	9 067	100	42,11

Table 2.3 Coverage (X) of mapping of the 277 693 Illumina reads and percentage of GC of the two genomes. "Ct without RR" is the genome sequence in which the RR sequence was manually deleted.

Sequence	Average coverage	Minimum coverage	Maximum coverage	% of covered bases	% of GC
Ct1	1 327	0	6 971	95	41.8
Ct2	1 386	0	4 554	93	42.7
RR in Ct1	575	0	6 971	41	32.3
RR in Ct2	40	0	745	12	42.7
Ct1 without RR	1 395	465	2 449	100	42.7
Ct2 without RR	1 547	574	4 554	100	42.8

Figure 2.1: Coverage of mapped MinION and Illumina reads and MITOS annotation for the longer mitogenome (Ct1). a. Coverage of mapping of the 12 764 MinION reads used by Canu for the assembly. Colored bars represent SNPs among reads, gray bars represent consensus bases among all the reads. b. Coverage of mapping of the 277 693 Illumina reads. c. proportion of GC. d. MITOS annotation. Red, green and blue arrows represent protein-coding genes, ribosomal genes, and tRNAs, respectively.

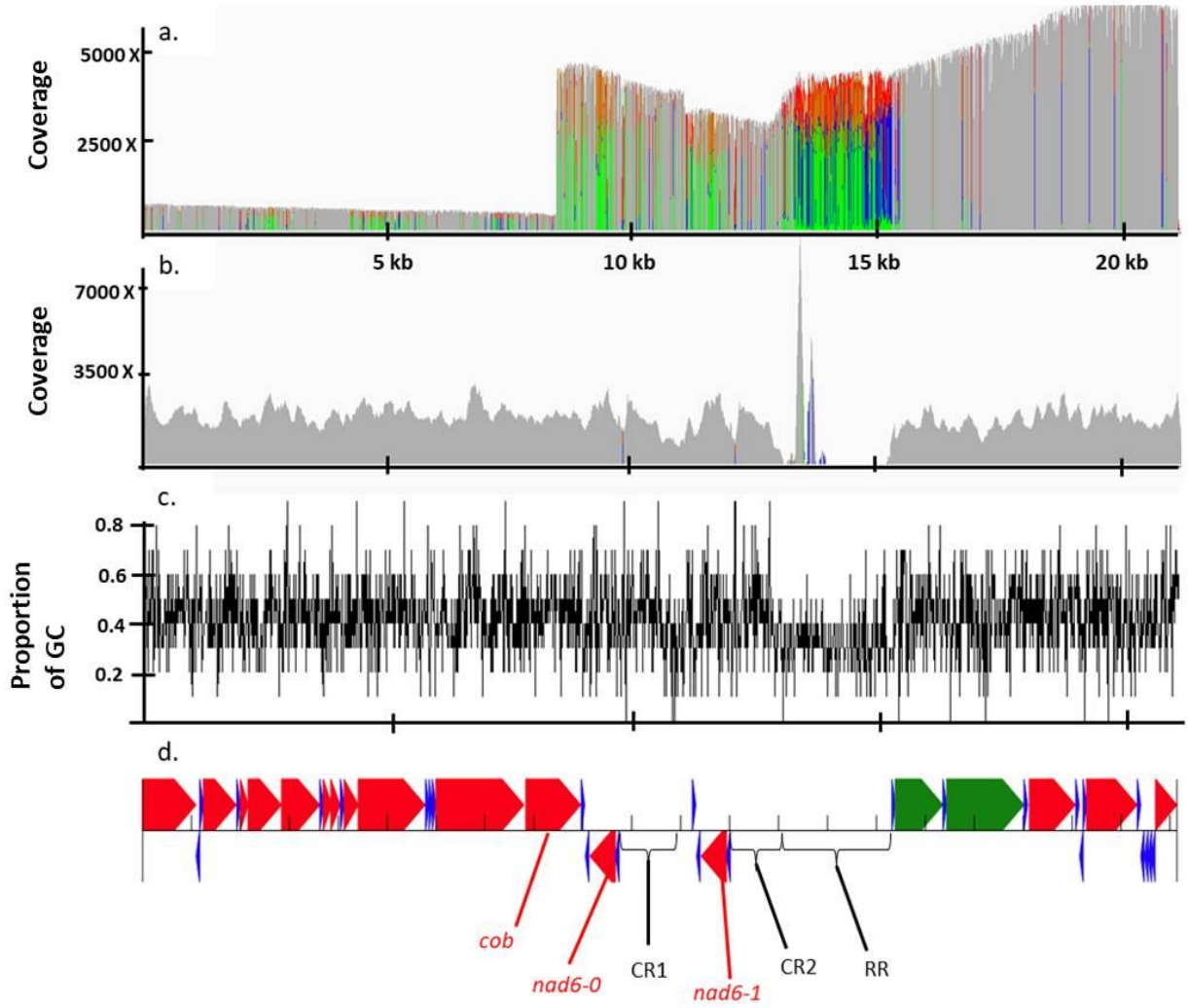
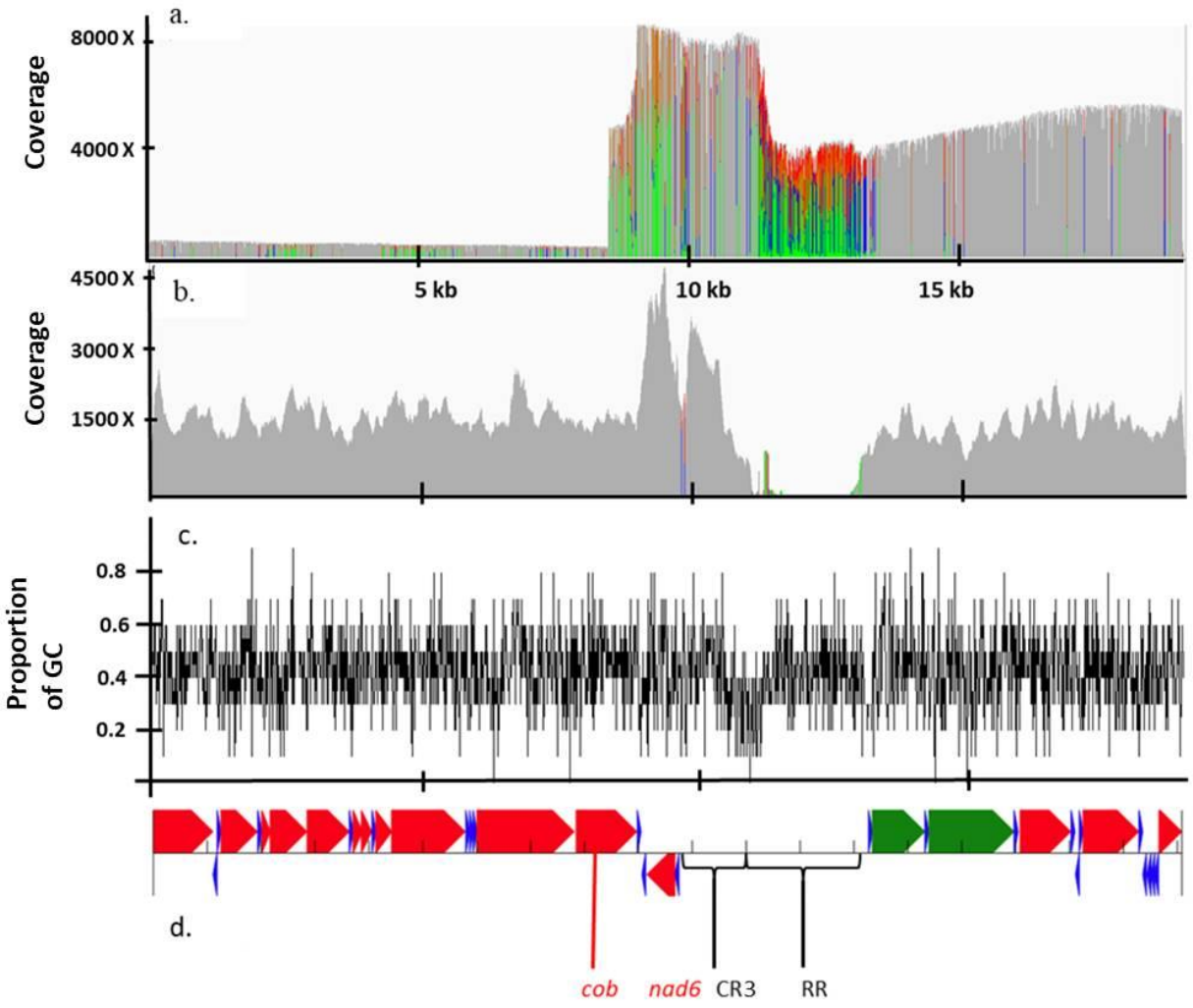


Figure 2.2: Coverage of mapped MinION and Illumina reads and MITOS annotation for the shorter mitogenome (Ct2). a. Coverage of mapping of the 12 764 MinION reads used by Canu for the assembly. Colored bars represent SNPs among reads, gray bars represent consensus bases among all the reads. b. Coverage of mapping of the 277 693 Illumina reads. c. proportion of GC. d. MITOS annotation. Red, green and blue arrows represent protein-coding genes, ribosomal genes, and tRNAs, respectively.



Ct1 and Ct2 were different in length and in composition. Although we cannot exclude that the DNA sample was contaminated by another one, we have strong evidence to contradict this hypothesis (see Supplementary Material 2.3). We observed that a duplicated region was present on Ct1, whereas Ct2 did not include a duplication. We used the contigs obtained with the MinION reads to study the divergences between the two mitogenomes. We corrected the contigs for indels but not for substitutions, with Illumina reads, using Pilon (Walker et al. 2014). We found 52 SNPs between the two contigs, only five of which were transversions. The distribution of the SNPs was not homogenous along the genome because *nad-4*, *nad-6*, the control region and *rnL* contained 35 (67%) of them. Of the 52 SNPs, 37 were located in coding regions, and 43% were on the third position of the codon and were synonymous mutations. We inferred the consensus sequence of the two contigs, using ambiguity codes where the sequences diverged. We mapped the 12 764 MinION reads on this consensus sequence and observed that 85% of the reads were consistent with Ct1, whereas 15% of the reads were consistent with Ct2. The distribution of the SNPs was not homogeneous along the genome and the divergence between the contigs was supported by all the reads. These characteristics made it unlikely for sequencing errors to be the sole explanation for the divergence between the two contigs. Based on these results, we suggest that the *Puffinus lherminieri* individual used for MinION sequencing showed mitochondrial heteroplasmy. Heteroplasmy has already been observed in birds (e.g. Mundy et al. 1996; Moum & Bakke 2001; Gandolfi et al. 2017). Because of likely PCR artefacts, such as preferential amplification of one mitogenome over the other, the proportion of MinION reads did not necessarily reflect the proportion of the two mitogenomes in the individual. When we mapped all the Illumina reads on these two contigs, we did not see any divergence among the reads, suggesting that heteroplasmy was not present in the individual sequenced by Illumina.

The duplication in Ct1 consisted of two identical copies of *nad6*, two identical copies of the tRNAs Phe, Trp and Cys, and two dissimilar copies of the CR (which we call CR1 and CR2). Contrary to previous results for albatrosses (Abbott et al. 2005, Gibb et al. 2007, Eda et al. 2010, Lounsberry et al. 2015), *cob* was not duplicated CR1 and CR2 included a 1 270 bp region in common; 24 mutations separated CR1 from CR2. This overlap region is followed in CR1 by 229 supplementary bases and in CR2 by 2 033 supplementary bases. These two supplementary sequences did not align to each other. The single CR found in Ct2, which we will call CR3, was more similar to CR1 than to CR2 in the overlap region but included a 2 250 bp supplementary region that aligned with the supplementary region of CR2.

CR2 (on Ct1) and CR3 (on Ct2) were followed by a 2 kb-long stretch of DNA (hereafter called RR, for repeat region) (Figure 2.1, 2.2) composed of 90 bp modules repeated 19 times. During polishing, Illumina reads mapped onto the beginning of the RR, but no read overlapped CR2/CR3 and the RR, nor did they overlap the RR and 12S; hence we have no proof that the RR was effectively present in the mitogenome of *P. lherminieri* individual USNM 620721 (Illumina). Moreover, the RR was poorly covered by the Illumina reads (Table 2.3), and in Ct1 this region had a lower GC content (32%) than the rest of the genome (average of 42%). This could explain why this region was more difficult to sequence with the Illumina technology. MinION reads were discordant in this particular region. When the RR was manually removed from the genome, no read linked CR2/CR3

and 12S. This means that no MinION read covered the entire mitochondrial genome if the RR was not present. Moreover, no significant BLAST match was found for this region (MegaBLAST & BLASTn, nr database, E-value threshold: 10, identity threshold 85%). Therefore, two hypotheses may explain the presence of this repeated sequence: (i) the RR is biologically present in the mitochondrial genome of this individual of *Puffinus lherminieri*, and potentially in other species of Procellariiformes, but was never sequenced until now. The large number of repetitions implied that MinION sequencing was difficult and led to differences among reads. Illumina sequencing was difficult in this region due to the high AT content and so, no Illumina read linked the RR to the rest of the mitogenome. Alternatively, (ii) the RR may be an artefact, due for example to chimeric amplification in the early stages of the PCR, containing parts of the mitochondrial genome and a nuclear sequence not catalogued on Genbank.

The duplicated region of *Puffinus lherminieri* was similar to all the duplicated regions observed in albatross species. The two Ct1 copies of *nad6* and the tRNA were identical, and the two control regions differed by several bases in the first 100 bp. The only major difference between shearwater and albatross mitogenome structures was that no duplication of *cob* was observed in the genome of *Puffinus lherminieri*. We also found evidence for two different mitogenomes co-existing in the same individual. These two mitogenomes differ by point mutations and gene duplication. Because albatross mitogenomes were sequenced using short fragments, heteroplasmy may have been harder to detect.

To conclude, we have provided direct evidence for a duplicated region in the mitogenome of *Puffinus lherminieri*. A similar duplication had already been observed in several albatross species (Abbott et al. 2005; Gibb et al. 2007; Eda et al. 2010; Lounsberry et al. 2015), phylogenetically distant from the shearwaters (Welch et al. in prep). This suggests that the duplication may be widespread within the Procellariiformes. The fact that the composition of the duplication is different between albatrosses and shearwaters suggests that this region has evolved during the history of the Procellariiformes. At least one event of deletion or addition of the copy of *cob* has occurred during diversification of Procellariiformes, and we know that other species show different versions of this duplication (Gibb et al. 2013). Investigating mitogenomes from more species is needed to better understand the evolutionary history of the mitochondrial genome of the Procellariiformes.

Discussion complémentaire

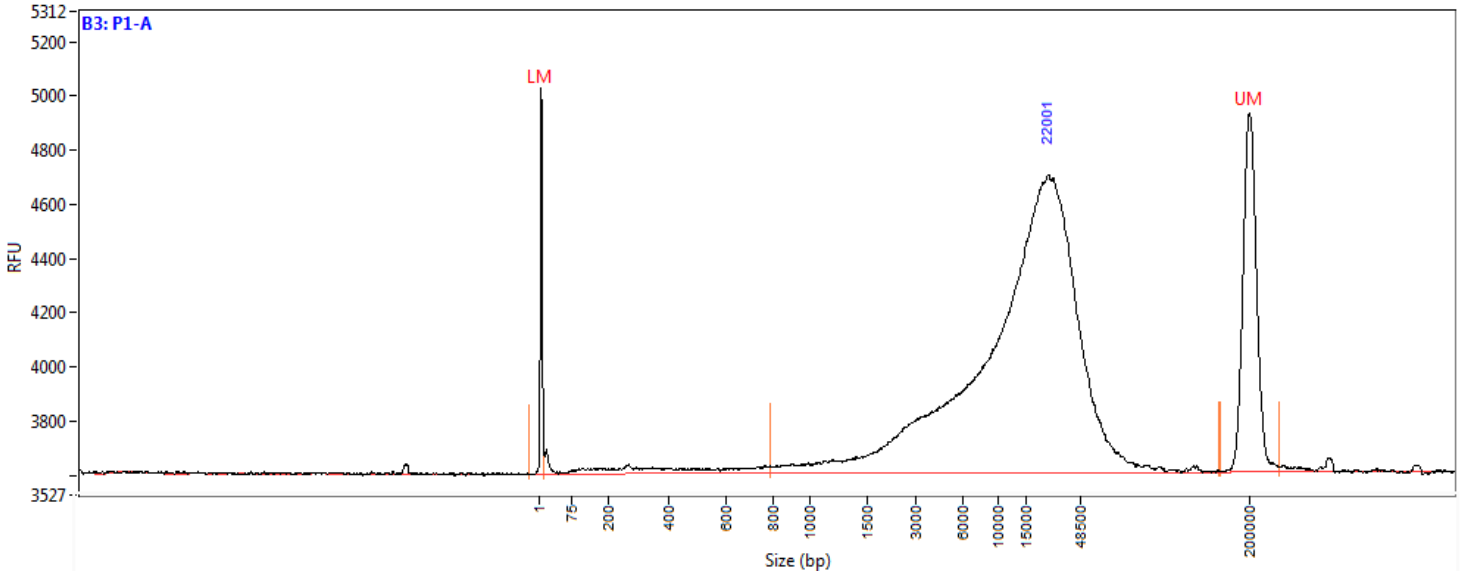
Afin de vérifier l'hypothèse avancée dans le chapitre précédent selon laquelle les deux mitogénomes assemblés seraient le résultat d'un phénomène d'hétéroplasmie, et non pas d'un artefact technique (e.g. produit PCR chimérique, artéfect d'assemblage), nous avons réalisé un contrôle de la taille de l'ADN génomique présent dans l'échantillon utilisé. Notre attendu théorique est la présence d'ADN de 19 et 21 kb.

Nous avons utilisé l'ADN extrait pour l'étude ainsi qu'une nouvelle extraction du même échantillon, réalisée selon le même protocole. Nous avons ensuite digéré l'ADN nucléaire présent dans des sous-échantillons de ces deux extractions. Pour cela nous avons utilisé un protocole de digestion par endonucléase V tiré de Jayaprakash et al. 2015 (voir Torres et al. in prep; Chapitre V). Les 4 échantillons ont ensuite été testés au Fragment Analyzer (Plateforme GenoToul, INRA Toulouse). Sur l'échantillon original, les résultats du contrôle-qualité montrent une courbe croissante à partir de 1,5 kb et culminant à 22 kb (Figure 2.3). Les tailles des deux mitogénomes attendus (respectivement 19 et 21 kb) ne peuvent être dissociés sur ce graphique. L'échantillon après digestion par exonucléaise, aucune trace d'ADN est détectable. Sur la seconde extraction les résultats du contrôle avant digestion montrent un pic à 37 kb et après digestion à 49 kb. Ces tailles, qui pourraient etre le résultat d'hybridation et formation d'hétérodimères entre molécules d'ADN mitochondriales pendant la phase de digestion (incubation à 37°C ptd 48h), ne permettent malheureusement pas de statuer sur l'origine hétéroplasmique des deux mitogénomes assemblés dans le cadre de notre étude.

Il est également possible que la digestion de l'ADN nucléaire ait été incomplete, et que l'un des deux ou les deux séquences mitogénomiques séquencés soient issus de numts, c'est-à-dire des copies de mitogénome transposées dans le génome nucléaire. Des copies de mitogénomes complets dans un génome nucléaire ont été trouvées par exemple chez le chien (Verscheure et al. 2015). Toutefois aucun des deux mitogénomes ne présente de codons-stops, ce qui rend la pseudogenisation peu probable.

L'hypothèse d'hétéroplasmie ne peut pas être totalement exclue puisque des phénomènes de transmission paternelle des mitochondries ont déjà été observés chez des oiseaux (Gandolfi et al. 2017). Si les processus d'hétéroplasmie mènent majoritairement à des différences courtes de tandem repeats, il existe également des événements d'hétéroplasmie menant à des mitogénomes différents par la présence d'une région dupliquée. Ce phénomène est connu chez les lézards et les humains (Poulton et al. 1989; Zevering et al. 1991). Pour confirmer le phénomène d'hétéroplasmie, le séquençage d'ADN génomique sans PCR préalable serait nécessaire, de préférence sur à partir de tissus non-sanguins plus riches en mitochondries.

Figure 2.3 : Analyse de la taille des mitogénomes par Fragment Analyzer



Annexe 1

Supplementary material for Evidence for a duplicated mitochondrial region in Audubon's shearwater based on MinION sequencing

Supplementary Material 2.1: Composition of Ct1

Name	Start	Stop	Strand	Length	Start-codon	Stop-codon
cox1_b	1	1092	+	1091		AGG
trnS2(tca)	1096	1169	-	73		
trnD(gac)	1172	1240	+	68		
cox2	1242	1916	+	674	GTG	TAA
trnK(aaa)	1927	1998	+	71		
atp8	2000	2161	+	161	ATG	TAA
atp6	2158	2838	+	680	ATG	TAA
cox3	2841	3623	+	782	ATG	TAA
trnG(gga)	3625	3693	+	68		
nad3_a	3694	3867	+	173	ATT	
nad3_b	3848	4042	+	194		TAA
trnR(cga)	4048	4116	+	68		
nad4l	4118	4411	+	293	ATG	TAA
nad4	4408	5775	+	1367	ATG	AGA
trnH(cac)	5786	5855	+	69		
trnS1(agc)	5856	5922	+	66		
trnL1(cta)	5922	5992	+	70		
nad5	5993	7795	+	1802	GTG	AGA
cob	7834	8958	+	1124	ATC	TAA
trnT(aca)	8968	9037	+	69		
trnP(cca)	9054	9123	-	69		
nad6-0	9147	9665	-	518	ATG	TAG
trnE(gaa)	9669	9741	-	72		
trnT(aca)	11240	11309	+	69		
trnP(cca)	11326	11395	-	69		
nad6-1	11419	11937	-	518	ATG	TAG
trnE(gaa)	11941	12013	-	72		
trnF(ttc)	15317	15386	+	69		
rrnS	15386	16361	+	975		
trnV(gta)	16361	16433	+	72		
rrnL	16434	18018	+	1584		
trnL2(tta)	18018	18091	+	73		
nad1	18129	19067	+	938	ATC	AGA
trnI(atc)	19075	19146	+	71		
trnQ(caa)	19156	19226	-	70		
trnM(atg)	19226	19294	+	68		
nad2	19295	20329	+	1034	ATG	TAG
trnW(tga)	20334	20403	+	69		
trnA(gca)	20405	20473	-	68		
trnN(aac)	20484	20556	-	72		
trnC(tgc)	20559	20625	-	66		
trnY(tac)	20626	20695	-	69		
cox1_a	20706	21143	+	437	ATC	

Supplementary Material 2.2: Composition of Ct2

Name	Start	Stop	Strand	Length	Start-codon	Stop-codon
cox1_b	2	1105	+	1103		AGG
trnS2(tca)	1109	1182	-	73		
trnD(gac)	1185	1253	+	68		
cox2	1255	1929	+	674	GTG	TAA
trnK(aaa)	1940	2011	+	71		
atp8	2013	2174	+	161	ATG	TAA
atp6	2171	2851	+	680	ATG	TAA
cox3	2854	3636	+	782	ATG	TAA
trnG(gga)	3638	3706	+	68		
nad3_a	3707	3880	+	173	ATT	
nad3_b	3861	4055	+	194		TAA
trnR(cga)	4061	4129	+	68		
nad4l	4131	4424	+	293	ATG	TAA
nad4	4421	5788	+	1367	ATG	AGA
trnH(cac)	5799	5868	+	69		
trnS1(agc)	5869	5935	+	66		
trnL1(cta)	5935	6005	+	70		
nad5	6006	7808	+	1802	GTG	AGA
cob	7847	8971	+	1124	ATC	TAA
trnT(aca)	8981	9050	+	69		
trnP(cca)	9067	9136	-	69		
nad6	9160	9678	-	518	ATG	TAG
trnE(gaa)	9682	9754	-	72		
trnF(ttc)	13272	13341	+	69		
rrnS	13341	14316	+	975		
trnV(gta)	14316	14388	+	72		
rrnL	14389	15973	+	1584		
trnL2(tta)	15973	16046	+	73		
nad1	16084	17022	+	938	ATC	AGA
trnI(atc)	17030	17101	+	71		
trnQ(caa)	17111	17181	-	70		
trnM(atg)	17181	17249	+	68		
nad2	17250	18284	+	1034	ATG	TAG
trnW(tga)	18289	18358	+	69		
trnA(gca)	18360	18428	-	68		
trnN(aac)	18439	18511	-	72		
trnC(tgc)	18514	18580	-	66		
trnY(tac)	18581	18650	-	69		
cox1_a	18661	19083	+	422	ATC	

Supplementary Material 2.3:

There are two explanations for the fact that the assembly of the MinION reads resulted in two contigs: either the DNA library was contaminated by another individual, so that the two contigs correspond to the mitochondrial genomes of two individuals, or two different mitogenomes were present in the one individual sequenced, i.e., heteroplasmy. We cannot rule out contamination, but we have strong evidence to contradict this hypothesis.

First, a Sanger sequence of *co1*of the same individual groups with the MinION sequences. This sequence was obtained as part of a population genetics study of 175 individuals (Torres et al. in prep; Chapter V), representing all populations of *Puffinus lherminieri* and three populations of the closely related species *P. bailloni* (Genbank acc. MH383332-MH383506). The *co1* sequences of the two MinION contigs were aligned with these 175 sequences using MAFFT (Katoh et al. 2002) and truncated to the 557 bp aligned for the other sequences. A phylogenetic tree was inferred using MrBayes v3.3.6 (Ronquist et al. 2012) and 500 million generations, with *P. yelkouan* (Genbank AY567884.1) as the outgroup. After deletion of 10% burn-in, convergence of the run was assessed using Tracer. The *co1* sequences of the two MinION contigs were part of the same clade as the sequences of the same lineage obtained by Sanger sequencing (Figure 2.S1), with a posterior probability of 0.80. The sequence from Ct1 was identical to the sequence from the individual of origin. The sequence of Ct2 was different to the original individual from 1 SNP.

Second, control region sequences from Ct1 and Ct2 are identical to those obtained using Sanger sequencing. Partial control region sequences (313 bp) of 224 sequences of the control region from the same sampling locations are extremely polymorphic, consisting of different haplotypes for most individuals. The two control regions of Ct1 and the one of Ct2 were aligned and truncated to correspond to these Sanger sequences. These three sequences are both identical among themselves and to the sequence of the individual obtained by Sanger.

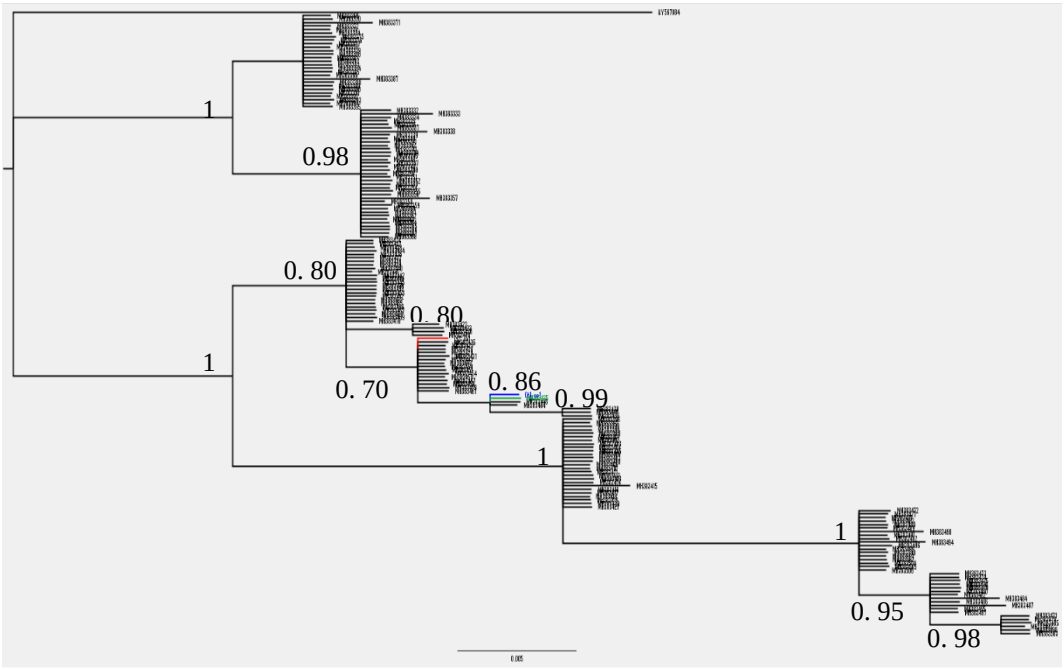
Thus, we compared two parts of the two mitochondrial genomes with sequences of other individuals, presenting numerous SNPs among individuals. These analyses show that both contigs are phylogenetically very close to the target individual. If the sample was contaminated it would be by a nearly identical sample phylogenetically, which is very unlikely.

Moreover, we know that heteroplasmy occurs in birds (Mundy et al. 1996, Moum & Bakke 2001, Gandolfi et al. 2017), and we have no reason to think that it is not present in Procellariiformes. Hence several different mitogenomes can be present in the same individual. A cross-contaminated sample would bear more likely four different mitogenomes instead of two. So the assembly of two different mitogenomes is not an evidence of cross-contamination.

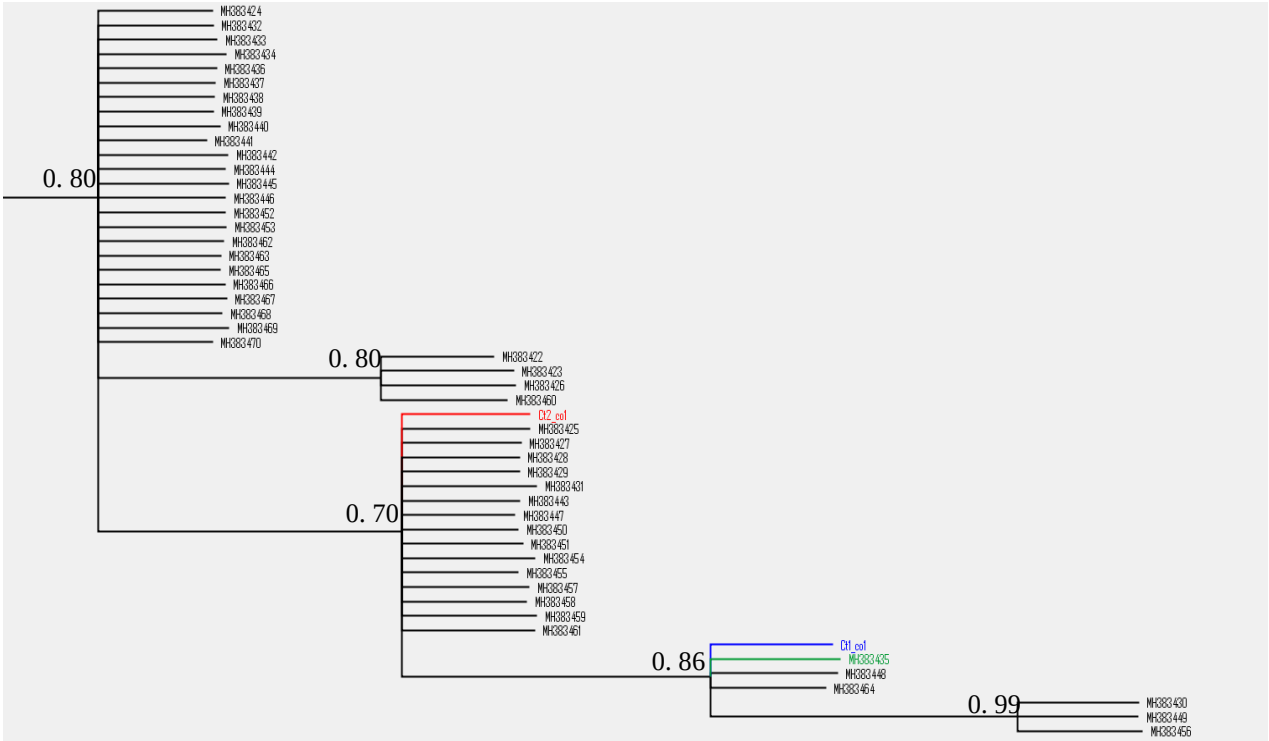
Therefore, we believe that the presence of two different contigs in our study is due to heteroplasmy. However, if the contamination is real, it would mean that one of the two individuals did not have the mitochondrial duplication (see Results & Discussion), and would suggest that the mitochondrial duplication is not present in every individual of this species. Therefore the evolutionary scenario of the duplicated region would be even more complicated than that suggested by the heteroplasmy hypothesis.

Supplementary Figure 2.1. a. Gene tree obtained from Sanger *co1* sequences and the two *co1* from the two MinION contigs. b. Zoom on the clade containing the two contigs sequences. The sequence from Ct1 is highlighted in blue, the one from Ct2 in red, and the one from the same individual sequenced by Sanger in green. Posterior probabilities superior to 0.70 are shown.

a.



b.



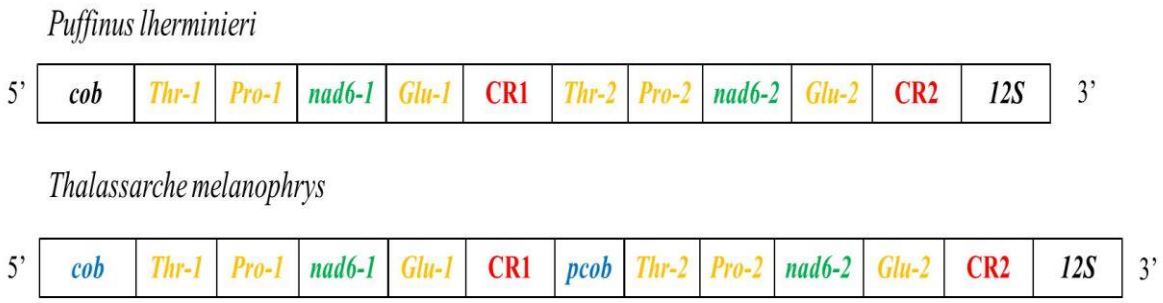
Chapitre III

Analyse préliminaire de l'évolution de la région dupliquée le long de la phylogénie des Procellariiformes

Dans le deuxième chapitre de cette thèse nous avons apporté la preuve directe de la présence d'une région dupliquée au sein du mitogénome d'un puffin d'Audubon *Puffinus lherminhieri*. Cette région est constituée des ARNt *Thr* et *Pro*, du gène *nad6* complet, de l'ARNt *Glu* et de la région de contrôle (CR) complète (Figure 3.1). Une région dupliquée similaire avait été trouvée dans le génome mitochondrial de quatre espèces d'albatros (Abbott et al. 2005; Lounsberry et al. 2015), considéré comme le groupe-frère de tous les Procellariiformes (Estandía et al. in prep). La région trouvée chez les albatros était toutefois différente puisque 159 pb en 3' du gène *cob* était également dupliquée en extrémité 5'. Plusieurs régions dupliquées similaires ont été trouvés chez d'autres oiseaux (voir Gibb et al. 2013). Une fois dupliqué, la copie d'un marqueur a quatre devenirs possibles (Zhang 2003) : la pseudogenisation ou perte de fonction puisque la copie n'est plus sous sélection (Lynch & Conery 2000), la conservation de sa fonction première si cela confère un avantage à l'individu (e.g. Piontkivska et al. 1997), la sous-fonctionnalisation, par exemple lorsque les deux copies sont exprimées différemment dans différents tissus (Force et al. 1999) ou l'acquisition d'une nouvelle fonction (e.g. Yokoyama & Yokoyama 1989). La région dupliquée des Procellariiformes semble montrer plusieurs de ses patrons. Une courte portion de *cob*, apparemment variable entre les espèces, ne semble pas fonctionnelle et a probablement subi un processus de pseudogenisation. Le gène *nad6* et les ARNt dupliqués présentent des copies identiques et pourraient donc être tous deux fonctionnels. Enfin les deux copies de CR semblent partiellement divergentes et pourraient présenter deux niveaux de fonctionnalisation différents. La présence d'une région mitochondriale dupliquée chez les oiseaux a été montrée comme corrélée à des données biologiques, comme la masse du corps (Urantowka et al. 2018). En effet les gènes mitochondriaux dupliqués sont impliqués dans la fonction respiratoire cellulaire et CR est liée à la réplication des gènes mitochondriaux. Deux copies de CR pourraient augmenter le nombre de copies de gènes mitochondriaux et augmenter la production d'énergie, comme supposé chez les Psittaciformes (Urantowka et al. 2018). Les données actuelles sur la région dupliquée ne sont présentes que pour quatre espèces de Procellariiformes. Une étude de l'évolution de cette région le long de la phylogénie des Procellariiformes pourrait apporter de nouveaux indices sur le niveau de fonctionnalisation des différents marqueurs impliqués et leur impact sur la biologie des espèces. Cette région dupliquée a également un impact méthodologique pour les études de phylogéographie, puisque le séquençage Sanger d'un gène dupliqué peut entrainer l'apparition d'ambigüités dans les séquences. Ces ambigüités apportent des biais dans les analyses de génétique de populations (voir Torres et al. in prep; chapitre IV). Afin de comprendre l'évolution de cette duplication, son importance biologique et comment limiter au mieux son impact sur la qualité des données il est donc important de connaitre la répartition de cette duplication au sein des Procellariiformes. Nous avons donc entrepris de reconstruire la région dupliquée pour 18 nouvelles espèces de Procellariiformes, couvrant toutes les grandes familles de l'ordre, afin de mieux comprendre son évolution. Cette section regroupe les résultats préliminaires obtenus uniquement à partir de fragments Illumina (précis mais courts). Cette étude sera complétée par l'obtention de fragments MinION (pouvant couvrir tout le génome) pour chacun des taxa étudiés ici.

Figure 3.1: Détail de la région mitochondriale dupliquée chez *Puffinus lherminieri* (Torres et al. 2018) et *Thalassarche melanophrys* (Abbott et al. 2005)

Les marqueurs en couleurs sont dupliqués, les ARNt sont indiqués en jaune. *pcob* indique la portion de *cob* dupliquée. Pour chaque espèce, les deux copies de *nad6* et de chaque ARNt sont identiques entre elles.



Matériel et méthodes

Nous disposons de fragments 100bp Illumina HiSeq 2500 appariés obtenus par Andreanna Welch (Durham University) pour 18 espèces de Procellariiformes et un groupe externe de l'ordre des Ciconiiformes (Matériel Supplémentaire 3.1), chez qui aucune région dupliquée n'a été trouvée (Gibb et al. 2013; Lee et al. 2017; Liu et al. 2016). Chacune des espèces est représentative d'un genre afin de couvrir toute la diversité taxinomique des Procellariiformes. Dans le riche genre *Pterodroma* (35 espèces Onley & Scofield 2007), quatre espèces ont été choisies pour représenter quatre clades génétiques distincts qui pourraient donc avoir une structure mitogénomique différente (Welch et al. 2014).

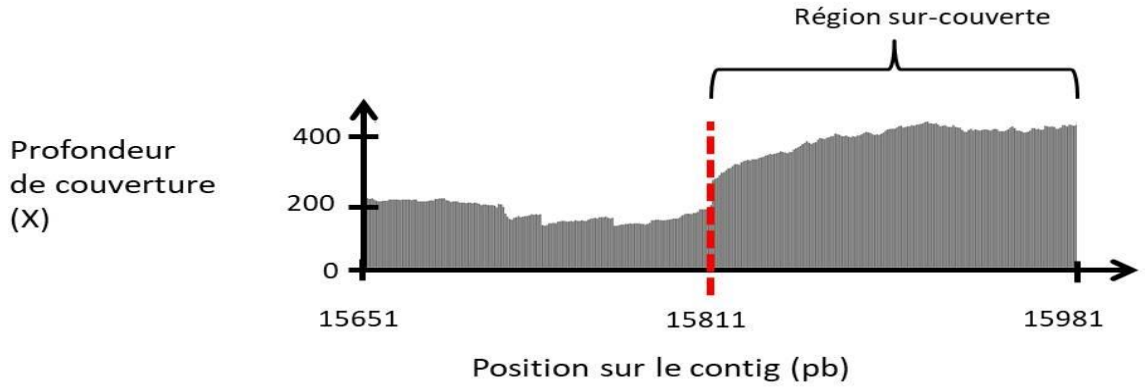
Les données de séquencage Illumina sont issus d'une expérience de capture de loci nucléaires (UCE Bejerano et al. 2004), à partir de tissus non-sanguins (Welch et al in prep). La capture de loci nucléaire mène généralement à la capture accidentelle de loci mitochondriaux (Raposo et al. 2015). Afin d'extraire les fragments correspondants à l'ADN mitochondrial nous avons utilisé le logiciel NOVOPlasty qui permet de comparer tous les fragments à un génome mitochondrial de référence puis d'assembler les fragments mitochondriaux en contigs. Nous avons utilisé comme référence le mitogénome de *Thalassarche melanophrys* (Genbank accession number AY158677.2) (Abbott et al. 2005) dont la région dupliquée contenait deux fragments de *cob*. Nous avons entré une taille de mitogénome visée entre 12 et 22 kb pour l'assemblage et une longueur de chevauchement entre les reads de 39 pb. Les fragments Illumina étant très courts (100 pb) par rapport à la duplication (environ 2000 pb, Abbot et al. 2005, Lounsberry et al. 2012, Torres et al. 2018) des génomes complets contenant la duplication n'ont pas toujours pu être obtenus. Nous avons utilisé utilisé une méthode de détection indirecte de duplication en alignant les fragments Illumina sur les contigs, pour chaque taxon, en utilisant BWA v 0.7.17.4 (Li & Durbin 2009), en utilisant les paramètres automatiques pour séquences Illumina appariées. Nous avons ensuite examiné sur IGV v2.3.90 (Robinson et al. 2012) la profondeur de couverture tout au long du contig obtenu. Si une région présentait une hausse de profondeur de couverture par rapport au reste du contig (voir Figure 3.2), nous avons calculé la

profondeur de couverture moyenne le long de cette région (en utilisant un script R, Matériel Supplémentaire 3.2) et l'avons comparée à la profondeur de couverture moyenne le long du contig. Les régions où la profondeur de couverture moyenne double par rapport à la moyenne du contig, indiquent que cette région est probablement en deux copies dans le mitogénome mais qu'elles n'ont pas pu être distinguées par les courts fragments Illumina. Nous considérons donc ces régions comme des régions dupliquées (voir Figure 3.2)

La nature des marqueurs compris dans ces régions putativement dupliquées a ensuite été inférée en les comparant aux mitogénomes de Procellariiformes déjà existant en utilisant BLAST (MegaBLAST, nr database, E-value threshold: 10, identity threshold 85%) (Altschul et al. 1990). La longueur et la composition de la région putativement dupliquée a ainsi pu être estimée pour chaque taxon. Nous avons ajouté à cela les longueurs et composition des régions dupliquées obtenues pour *Puffinus lherminieri* (MH206162.1, Torres et al. 2018), *Thalassarche melanophrys* (AY158677, Gibb et al. 2007) et *Phoebastria nigripes* (KJ735512.1 Lounsberry et al. 2015). La portion de *cob* dupliquée semble varier entre les espèces et la composition de la duplication pourrait avoir des impacts biologiques. Afin d'obtenir une estimation de l'évolution de la région dupliquée au sein des Procellariiformes nous avons réalisé une analyse de reconstruction d'état ancestral par maximum de vraisemblance grâce à la fonction fastAnc du package R *phytools*. L'état ancestral a été reconstruit pour la longueur totale de la région dupliquée et pour la longueur de la portion de *cob* dupliquée, qui est attendue pour être très variable au sein des Procellariiformes au vu des résultats existants sur les albatros (Abbott et al. 2005) et les puffins (Torres et al. 2018).

Figure 3.2 : Analyse de la profondeur de couverture pour détecter les régions dupliquées du mitogénome d'*Aphrodroma brevirostris*

Des fragments Illumina ont été alignés sur le contig assemblé. La profondeur de couverture correspond au nombre de fragments alignés. La barre rouge délimite une région dont la profondeur de couverture est double par rapport au reste du mitogénome, cette région est considérée comme dupliquée.



Résultats préliminaires

Nous avons obtenu un unique contig couvrant une copie de chacun des marqueurs du mitogénome pour toutes les espèces étudiées. Le nombre de fragments Illumina obtenus pour chaque espèce variait de 5 389 à 139 421 et la profondeur de couverture moyenne de chaque

mitogénome variait de 29X à 828X (Matériel Supplémantaire 3.1). Les contigs obtenus variaient de 15 731 pb à 20 167 pb. Tous les contigs présentaient une seule copie de chaque marqueur à l'exception de ceux assemblés pour *Calonectris leucomelas* et *Diomedea exulans*. Pour *Calonectris leucomelas* le contig obtenu présentait deux copies complètes de *nad6,* CR et des trois ARNt attendus dans la région dupliquée ainsi que 20 pb annotées comme identiques aux 20 pb en 3' de *cob* via BLAST. Pour *Diomedea exulans*, le contig présentait deux copies complètes de *nad6,* CR et des trois ARNt attendus dans la région dupliquée ainsi que 550 pb annotées comme identiques à *cob* via BLAST. Lorsque les fragments Illumina correspondants à ces deux contigs étaient alignés sur ces deux contigs respectifs la profondeur de couverture était homogène le long du contig. Pour chacun des contigs inférés pour les autres taxons, l'alignement des fragments Illumina sur le contig correspondant mettait en évidence une région du contig dont la profondeur de couverture moyenne était au moins deux fois plus élevée que la moyenne le long du contig complet. Nous considérons qu'il s'agit de régions dupliquées dans le mitogénome, dont l'analyse BLAST nous donne la composition (Figure 3.3.a).

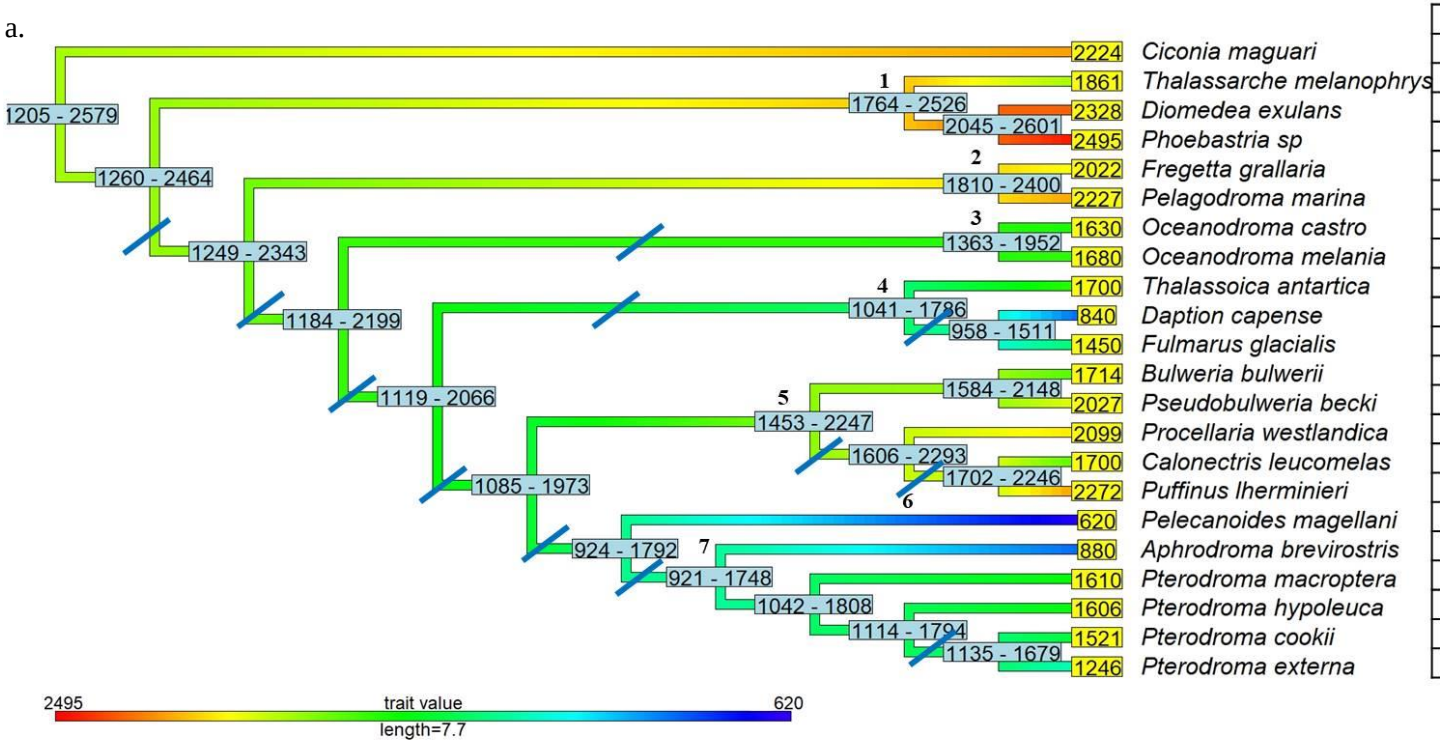
Nous divisons l'ordre en six clades, basés sur la taxonomie et le nombre d'espèces : les albatros (1), les océanites austraux (2), les océanites boréaux (3), les fulmarines (4), les puffins (5), les pétrels-plongeurs (6) et les pétrels (7) (Figure 3.3). Une région dupliquée est trouvée chez chacune de ces espèces. La taille de la région dupliquée varie de 620 pb à 2227 pb. Chez toutes les espèces, à l'exception de deux, le gène *nad6*, l'ARNt *Glu* et CR sont entièrement dupliqués. Pour chaque taxon, les deux copies du gène *nad6* trouvées sont identiques entre elles, de même que les deux copies trouvées de chacun des trois ARNt. Au niveau inter-individuel *nad6* est pourtant plus variable que *cob* (Sorenson 2003), le fait que les deux copies soient conservées au sein d'un même individu est un autre élément en faveur du fait qu'elles soient toutes deux fonctionnelles. Les deux copies de CR étaient toujours de taille équivalente mais présentaient souvent plusieurs mutations (jusqu'à 10). La séquence précise des deux copies n'a pas pu être reconstituée ce qui rend une analyse phylogénétique des deux copies pour l'instant impossible.

La reconstruction de l'état ancestral de la taille de la région dupliquée montre un patron de diminution de la taille de la duplication le long de la phylogénie des Procellariiformes. La taille de la région dupliquée entière ainsi que la taille de la portion de *cob* dupliquée semblent plus longues chez les taxons basaux à l'échelle de l'ordre entier des Procellariiformes (Figure 3.3).

On retrouve ce patron d'érosion de la région dupliquée et de la portion de *cob* dupliquée au sein de plusieurs clades de Procellariiformes. Ainsi au sein des fulmarines (4), la duplication complète, et la portion de *cob* en particulier, ont diminuées en taille au fur et à mesure des divergences. Les océanites boréaux (5) ont subi une forte diminution de la taille de la duplication relativement à l'état ancestral reconstruit. Cette réduction ancestrale se traduit chez les deux espèces actuelles par l'absence de duplication de *cob* et chez *Oceanodroma castro* de l'ARNt mitoyen à *cob*. Le patron est toutefois inversé chez les océanites austraux (6) et les albatros (7), chez qui la duplication semble s'être rallongée relativement aux états ancestraux, ce qui est corrélé à une portion de *cob* dupliquée plus longue qu'aux nœuds ancestraux chez ces deux clades.

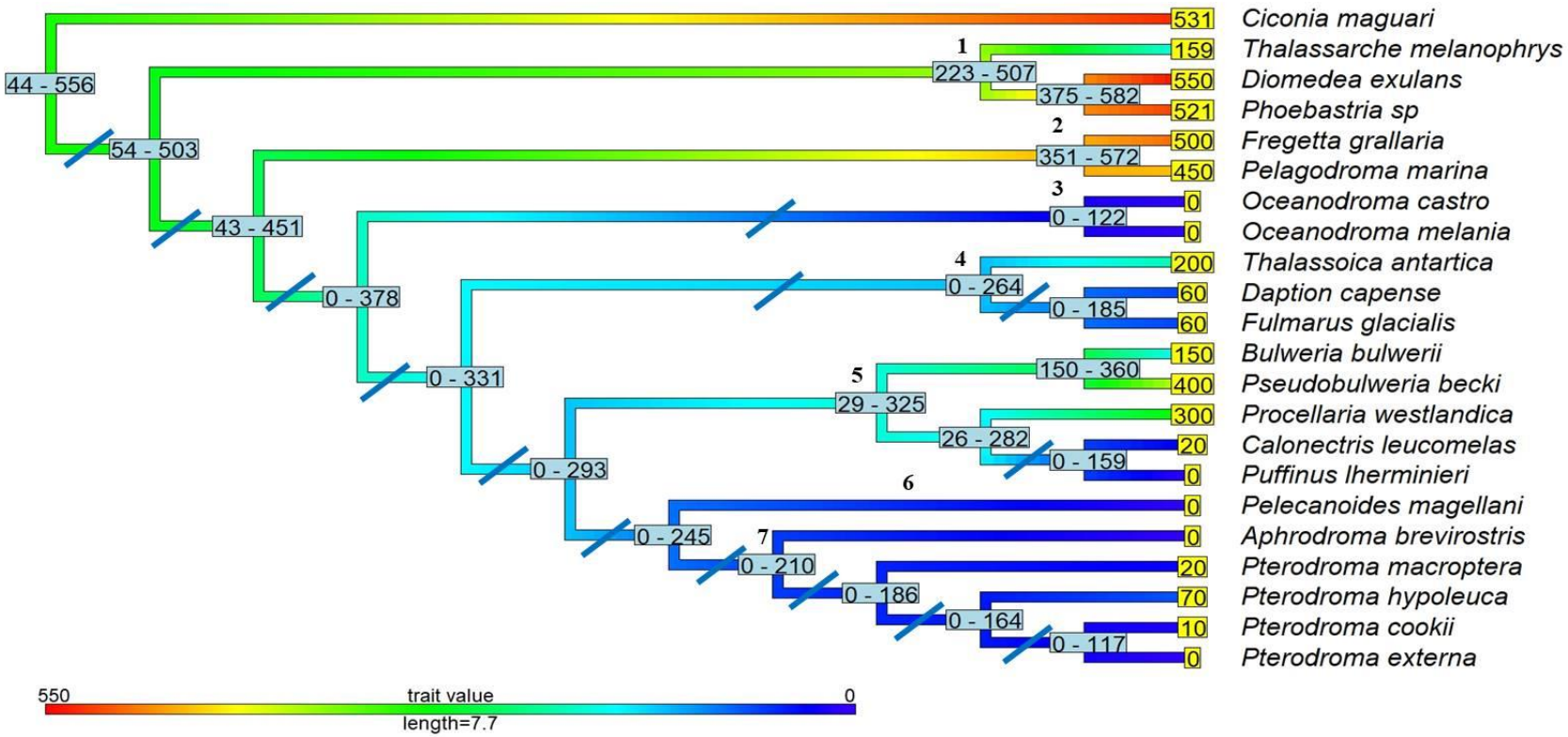
Figure 3.3 : Analyse des états ancestraux de la taille de la région dupliquée complète (a) et de la taille de le la portion de *cob* dupliquée (b)
L'arbre phylogénétique est basé sur Estandía et al. in prep. Les longueurs de branche ne sont pas proportionnelles au temps d'évolution
Les nombres indiqués sur les nœuds internes représentent les intervalles de confiance à 95% de l'état ancestral (en pb), sur les nœuds externes les données des espèces étudiées (en pb). Les barres bleues indiquent des événements probables d'érosion de la région sur les branches internes.
Le tableau indique la portion de *cob* dupliquée (en pb), si *nad6* et CR sont dupliqués (x, auquel cas la taille des deux copies est identique) ou non (-) et le nombre d'ARNt dupliqués.
Les Procellariiformes sont divisé en sept clades : les albatros (1), les océanites austraux (2), les océanites boréaux (3), les fulmarines (4), les puffins (5), les pétrels-plongeurs (6) et les pétrels (7)

a.



cob	*nad6*	CR	ARNt
531	x	x	3
159	x	x	3
550	x	x	3
521	x	x	3
500	x	x	3
450	x	x	3
0	x	x	2
0	x	x	3
200	x	x	3
60	x	x	3
60	x	x	3
150	x	x	3
400	x	x	3
300	x	x	3
20	x	x	3
0	x	x	3
0	-	partielle	0
0	-	x	0
20	x	x	3
70	x	x	3
10	x	x	3
0	x	x	1

b.


44 - 556
54 - 503
43 - 451
0 - 378
0 - 331
0 - 293
0 - 245
0 - 210
0 - 186
0 - 164
0 - 117
1
223 - 507
375 - 582
2
351 - 572
3
0 - 122
4
0 - 264
0 - 185
5
29 - 325
26 - 282
0 - 159
6
7
531
Ciconia maguari
159
Thalassarche melanophrys
550
Diomedea exulans
521
Phoebastria sp
500
Fregetta grallaria
450
Pelagodroma marina
0
Oceanodroma castro
0
Oceanodroma melania
200
Thalassoica antartica
60
Daption capense
60
Fulmarus glacialis
150
Bulweria bulwerii
150 - 360
400
Pseudobulweria becki
300
Procellaria westlandica
20
Calonectris leucomelas
0
Puffinus lherminieri
0
Pelecanoides magellani
0
Aphrodroma brevirostris
20
Pterodroma macroptera
70
Pterodroma hypoleuca
10
Pterodroma cookii
0
Pterodroma externa
550
trait value
0
length=7.7

Les puffins (3) montrent un patron intermédiaire puisque la duplication semble avoir subi une élongation le long de la branche ancestrale, puis la branche menant à *Bulweria bulwerii* et *Pseudobulweria becki* a également subi une élongation. Au contraire le clade regroupant *Procellaria westlandica, Calonectris leucomelas* et *Puffinus lherminieri* présente un patron d'érosion de la duplication au fur et à mesure des événements de divergence. L'évolution complexe de la longueur de la région dupliquée au sein des puffins est corrélée à l'évolution de la taille de la portion de *cob* dupliquée.
Les pétrels (1) présentent également un patron intermédiaire. La taille de la duplication complète semble avoir subi une élongation sur les branches ancestrales du genre *Pterodroma* puis parait avoir diminué à nouveau. Ce patron est le même pour la portion de *cob* dupliquée chez ce clade.

Le long de la branche menant à *Pelecanoides magellani* comme le long de la branche menant à *Aphrodroma brevirostris* on observe une diminution très forte de la région dupliquée. La taille de la duplication aux nœuds ancestraux est plus longue, ce qui nous indique que ces deux diminutions ont eu lieu indépendamment sur les deux branches. Le même patron est observé pour la diminution de la taille de la portion de *cob* dupliquée et la disparition de la duplication du gène *nad6* (Figure 3.3 et Matériel Supplémentaire 3.3). Chez *Pelecanoides magellani*, un contig couvrant tous les marqueurs mitochondriaux en un seul exemplaire a été obtenu. En alignant les fragments Illumina sur ce contig on observe sur une région particulière une profondeur de couverture six fois plus élevée que sur le reste du génome. La région sur-couverte mesure 620 pb dont 20 pb ont été assignées à CR par BLAST. On peut donc postuler qu'une partie de CR de cette espèce est répétée de multiple fois dans le génome. La répétition d'une courte portion de CR n'étant pas retrouvée chez d'autres espèces, il se pourrait également qu'aucune partie du génome ne soit dupliquée et que le patron observé provienne d'un artéfact de l'assemblage. Cette région est en effet faible en GC (30%) ce qui rend l'assemblage de fragments Illumina incertain (Minoche et al. 2011). Sur le contig inféré pour *Aphrodroma brevirostris*, on observe une région sur-couverte de 880 pb qui correspond à CR.

On remarque également que la région dupliquée est présente chez le groupe externe choisi *Ciconia maguari*. Cette duplication semble avoir augmentée en taille le long de la branche qui la sépare du nœud ancestral.

Discussion

Nos analyses préliminaires apportent le premier indice que la région mitochondriale dupliquée est présente chez tous les clades de Procellariiformes et que l'évolution de sa taille et de sa composition le long de la phylogénie de l'ordre est complexe. Nous avons en plus apporté le premier indice de son existence chez le groupe-externe des Ciconiiformes (Lee et al. 2017; Liu et al. 2016). La reconstruction de l'état ancestral semble indiquer que la région dupliquée a subi plusieurs réductions de taille entre les nœuds les plus ancestraux de la phylogénie. Toutefois le faible échantillonnage taxonomique du groupe externe peut entrainer une sous-évaluation de l'état ancestral aux nœuds les plus profonds (Cunningham 1999). Une

élongation de la région dupliquée semble mécaniquement complexe, d'autant qu'une longue région dupliquée a été trouvée chez plusieurs autres oiseaux (Gibb et al. 2013) et est considérée comme ancestrale chez les oiseaux (Urantowka et al. 2018). Il semble donc parcimonieux et cohérent que la duplication ancestrale était en réalité de taille comparable aux grandes tailles de région dupliquée des Ciconiiformes, des albatros et des océanites austraux.

Considérant la composition de la duplication chez tous ces taxons, le scénario le plus vraisemblable d'évolution de la région dupliquée le long de la phylogénie des Procellariiformes est le suivant :

1. Apparition d'une région dupliquée couvrant environ 550 pb de *cob*, *nad6,* CR et trois ARNt chez l'ancêtre des Procellariiformes et des Ciconiiformes.
2. Erosion progressive de la région dupliquée au fur et à mesure des divergences entre et au sein des différents clades de Procellariiformes.
3. Rétrécissement brusque de cette région chez les océanites boréaux
4. Elongation du fragment dupliqué de *cob* le long de la formation de *Bulweria bulwerii* et *Pseudobulweria becki*, érosion vers la diversification de *Procellaria*, *Calonectris* et *Puffinus*.
5. Quasi-disparition de toute la région chez au moins une espèce de *Pelecanoides*
6. Perte de la deuxième copie de *cob*, *nad6* et des ARNt chez *Aphrodroma brevirostris*
7. Elongation puis nouvelle érosion de la région dupliquée parallèlement à la diversification des espèces du genre *Pterodroma*.

Le fait que la longueur du fragment de *cob* dupliqué soit aussi variable au sein des Procellariiformes est cohérent avec le fait qu'il ne soit pas fonctionnel. Les faibles contraintes de sélection qu'il subit peuvent expliquer que les mutations soient nombreuses et persistantes sur cette copie ce qui conduit à son érosion. L'érosion entre espèces de la région dupliquée au niveau de la portion de *cob* a été montrée chez les fous (Morris-Pocock et al. 2011) et chez les Psittaciformes (Urantowka et al. 2018), chez qui la copie de *cob* ne parait pas non plus fonctionnelle, car la région 5' est supprimée par rapport à la copie fonctionnelle. Le fait que le gène *nad6* et les ARNt soient présents en deux copies identiques chez la majorité des espèces pourrait signifier que les deux copies sont pleinement fonctionnelles ce qui n'est pas le cas chez plusieurs espèces de Gruidae (Akiyama et al. 2016) et de Psittaciformes (Urantowka et al. 2018). Les deux copies de CR semblent de longueur et de composition proches chez la majorité des espèces mais les données préliminaires obtenues ici ne permettent pour l'instant pas d'affirmer qu'elles soient toutes deux fonctionnelles. Ce patron général de la composition de la région dupliquée chez les Procellariiformes correspond au patron ancestral chez les Psittaciformes (Urantowka et al. 2018). Dans leur étude Urantowka et al. (2018) avancent plusieurs indices qui montrent que ce patron est lié à un fort métabolisme, les deux copies des marqueurs étant fonctionnels et une grande longévité, de grandes capacités de dispersion et un taux de substitution élevé. Toutes ces caractéristiques sont particulièrement présentes chez les Procellariiformes (Brooke 2004; Onley & Scofield 2007). Des analyses plus précises sont nécessaires mais la région mitochondriale semble bien avoir des conséquences biologiques et pourrait être fortement liées à la diversification des Procellariiformes.

La disparition quasi-totale de la région dupliquée chez au moins une espèce de *Pelecanoides* est susceptible d'avoir un impact sur la biologie de cette famille. Notre jeu de données ne comportant qu'une espèce sur les quatre qui constituent le genre des pétrels-

plongeurs, il est impossible de dire si ce patron est caractéristique de tout le genre ou uniquement de cette espèce. Une étude plus poussée sur le genre serait nécessaire. Les espèces de pétrel-plongeurs ont en effet un métabolisme particulier au sein des Procellariiformes du fait de leur mode de vie basé sur la plongée. De même que des patrons de duplication particuliers sont liés à des physiologies particulières chez les Psittaciformes (Urantowka et al. 2018), le fait que ni *nad6* ni la totalité de CR ne soit dupliquée chez cette famille à la physiologie particulière au sein de l'ordre des Procellariiformes est un argument en faveur de l'importance fonctionnelle de cette deuxième copie dans la physiologie des oiseaux marins. La disparition d'une grande partie de la région dupliquée détectée chez *Aphrodroma brevirostris*, seule espèce du genre, dont la classification est encore incertaine (Nunn & Stanley 1998; Penhallurick & Wink 2004; Welch et al. 2014), appelle également une étude plus poussée de ces caractéristiques biologiques.

Pour être complète cette étude nécessitera un séquençage du génome mitochondrial en utilisant les longs fragments obtenus par MinION de tous les individus. Des longs fragments permettront d'abord de confirmer ou d'invalider la composition de chaque duplication et donc de consolider le scénario d'évolution. Les longs fragments permettront également d'obtenir la séquence précise des deux copies des marqueurs dupliqués tout au long de la phylogénie des Procellariiformes. Par exemple cela permettra de savoir si le cadre de lecture est conservé pour les deux copies de *cob* et donc d'en attester la fonctionnalité. Nous pourrons ainsi inférer l'évolution des deux copies des différents marqueurs indépendamment et avoir une idée plus précise des mécanismes qui régissent l'évolution moléculaire des oiseaux.

Annexe 2

Matériel Supplémentaire pour Analyse préliminaire de l'évolution de la région dupliquée le long de la phylogénie des Procellariiformes

Matériel Supplémentaire 3.1 : Données de couverture obtenues

*Mitogénome provenant de Genbank

** Duplication contenue dans un seul contig

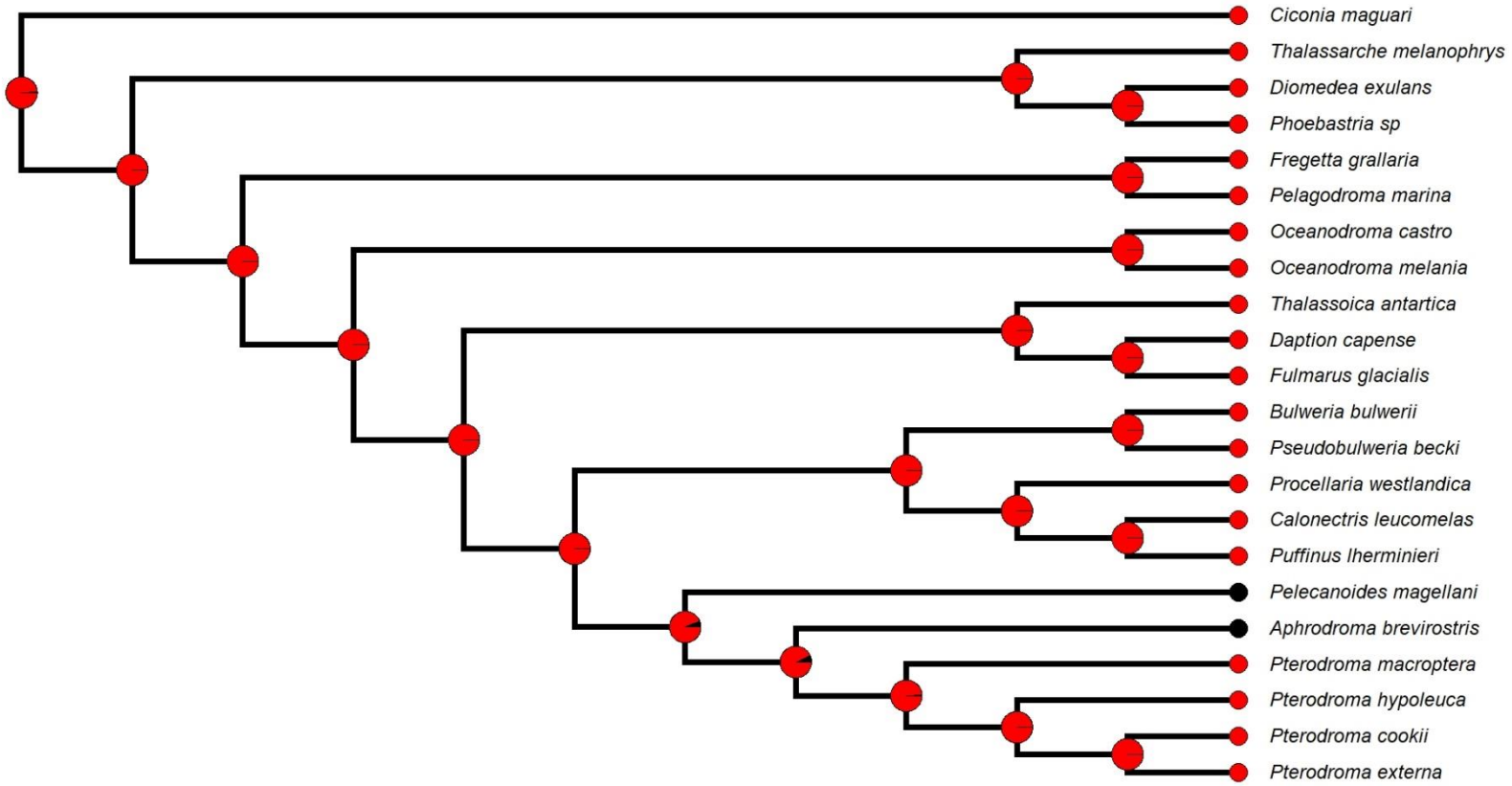
Espèce	Voucher	Taille du contig obtenu (pb)	Fragments correspondants au contig	Profondeur de couverture moyenne le long du contig (X)	Profondeur de couverture moyenne le long de la région surcouverte (X)
Ciconia maguari	USNM_614527	18112	14279	80	187
*Thalassarche melanophrys**	-	18967	-	-	-
Diomedea exulans	UWBM_81027	20167	37454	187	-
*Phoebastria nigripes**	USNM_630958	19072	-	-	-
Fregetta grallaria	AMNH_DOT3126	17573	6541	38	103
Pelagodroma marina	USNM_614205	16822	37031	222	414
Oceanodroma castro	USNM_602013	16697	39290	237	393
*Oceanodroma melania***	KU_9119	18412	5389	29	-
Thalassoica antartica	UWBM_81012	17096	47391	279	647
Daption capense	KU_21827	15731	13108	84	145
Fulmarus glacialis	AMNH_DOT3210	17636	15537	89	192
Bulweria bulwerii	USNM_631387	17026	39318	232	536
Pseudobulweria becki	Bretagnolle_becki S1	17022	18865	112	188
Procellaria westlandica	UWBM_82803	17466	7537	43	74
*Calonectris leucomelas***	LSU_B16967	18147	16065	89	-
*Puffinus lherminieri**	USNM_607632	21203	-	-	-
Pelecanoides magellani	AMNH_DOT3211	17320	39208	227	786
Aphrodroma brevirostris	UWBM_61673	16818	37737	226	549
Pterodroma macroptera	UWBM_80997	17719	24646	140	358
Pterodroma hypoleuca	UWBM_55680	16908	136658	812	2682
Pterodroma cookii	UWBM_70582	16909	139421	828	1659
Pterodroma externa	AMNH_DOT3106	16985	23877	142	428

Matériel Supplémentaire 3.2 : Script R pour le calcul de la profondeur de couverture moyenne

```
library(Rsamtools)

bamfile="file.bam"

# lecture du fichier
bam <- scanBam(bamfile)[[1]]
# filtre des fragments ne correspondant à aucune position
ind <- ! is.na(bam$pos)
bam <- lapply(bam, function(x) x[ind])

## inventaire de tous les fragments
ranges <- IRanges(start=bam$pos, width=bam$qwidth, names=make.names(bam$qname,
unique=TRUE))

## Profondeur de couverture moyenne le long du contig
mean(coverage(ranges))


## inventaire de tous les fragments correspondants à la région sur-couverte déterminée par
visualisation sur IGV. Par exemple ici la région sur-couverte couvre toute la région en 3’ de la
position 17187
ranges_dd<-ranges[start(ranges) >= 17187]

## Profondeur de couverture moyenne le long de la region sur-couverte
mean(coverage(ranges))
```

Matériel Supplémentaire 3.3 : Reconstruction de l'état ancestral pour la présence de la duplication de *nad6*

La proportion de rouge représente la vraisemblance que *nad6* soit entièrement dupliqué



Chapitre IV

Translocation of mitochondrial sequences into the nuclear genome may blur phylogeographic and conservation genetic studies in seabirds

In prep.

Torres Lucas[1,2], Bretagnolle Vincent[1], Pante Eric[2]

Affiliations
1 Centre d'Etudes Biologiques de Chizé, UMR 7372, CNRS & Université de La Rochelle, Villiers en Bois, France
2 Littoral, Environnement et Sociétés, UMR 7266 CNRS, Université de La Rochelle, La Rochelle, France

Les analyses de génétique de la population et phylogéographie sont essentielles pour comprendre le processus de spéciation. La majorité des études de ce type reposent au moins en partie sur des marqueurs mitochondriaux. Ceux-ci sont utilisés car l'ADN mitochondrial est connu pour avoir une vitesse d'évolution permettant de partitionner la variance du polymorphisme génétique intra- et inter-populationnel, permettant ainsi d'appréhender le processus de spéciation. L'ADN mitochondrial est également présent en de nombreuses copies chez un individu ce qui le rend plus facile à séquencer. Pourtant ces copies ne sont pas toutes identiques au sein d'un même individu. Des phénomènes comme l'hétéroplasmie, la duplication intra-mitochondriale, et la transposition de séquences mitochondriales vers le noyau (pseudogènes nucléaires d'origine mitochondriale, ou « numts ») entrainent la formation de paralogues dont les séquences peuvent diverger de celles du mitogénome. Ces phénomènes peuvent biaiser les inférences sur la diversité, la divergence, et la différentiation intra- et inter-populationnelle. Bien qu'ils soient connus, leur impact sur les analyses de génétique des populations ont été peu étudiés (e.g. chez les insectes ; Cristiano et al. 2012, Haran et al. 2015, Song et al. 2008) et de nombreuses études ne mentionnent pas leur existence.

L'hétéroplasmie, la duplication et les numts sont présents chez les Procellariiformes et dans notre complexe d'étude de puffins (Torres et al 2018, Chapitre II ; Torres et al in prep, Chapitre V). Dans ce chapitre nous comparons un jeu de données mitochondriales vierges de contaminant nucléaire, un jeu de données correspondant aux numts, et différents jeux de données dans lesquels ces deux sources de données sont amalgamées. La comparaison entre ces alignements nous a permis d'estimer l'impact des numts sur les analyses de génétique de la conservation et comment les gérer au mieux.

Pour cela, (1) nous avons obtenu des séquences du gène mitochondrial *cob* (encodant le cytochrome b) par amplification PCR et séquençage Sanger ; nous avons établi dans le chapitre V que ces séquences correspondent à un amalgame entre le marqueur mitochondrial *cob* et ses numts. (2) Nous avons ensuite obtenu les séquences du même marqueur des mêmes individus en digérant préalablement l'ADN nucléaire linéaire, éliminant ainsi tout contaminant nucléaire. Enfin (3), nous avons comparé in silico ces deux jeux de données afin d'en extraire les séquences correspondant aux numts. Nous comparons les relations phylogénétiques entre individus au sein et entre les cinq lignées du complexe *P. lherminieri*, en utilisant les séquences mitochondriales seules, les numts seuls, et les deux loci couplés afin de (1) comparer les niveaux de diversité génétique pour ces deux marqueurs, (2) tester l'hypothèse que les arbres de gènes inférés pour ces deux marqueurs sont congruent, en établissant les différences de patrons de divergence et différentiation inter-populationnelle, (3) tester l'hypothèse d'une transposition ancienne du *cob* vers le noyau. Cette hypothèse est motivée par l'observation que des séquences numts ont été détectées pour les cinq lignées du complexe *P. lherminieri*.

Lorsque qu'un locus mitochondrial et ses paralogues sont co-amplifiés par PCR, les séquences produites par méthode Sanger peuvent être caractérisées par de nombreuses ambiguïtés, traitées de différentes manières dans la littérature. Nous avons effectué des analyses de génétique des populations et de phylogéographie (estimations de diversité, divergence, et différentiation aux échelles intra- et inter-populationnelle) sur nos jeux de données (séquences mitochondriales homologues, et mélange d'homologues et de paralogues) en appliquant

différents traitements aux séquences caractérisées par des ambiguïtés. Puis, nous avons comparé les résultats afin de déterminer l'impact de ces différents traitements sur les analyses.
Les topologies des arbres de gènes inférés indépendamment pour *cob* et les séquences de numt sont similaires, la majorité des séquences nucléaires formant des clades correspondant aux lignées mitochondriales (à l'exception du couple *baroli*/*boydi* qui forme un groupe polyphylétique). En revanche, 61% des séquences de numts ont un placement incohérent avec l'attendu mitochondrial. Ce signal de polyphylie est visible lorsque les homologues et paralogues sont analysés conjointement. Ces patrons peuvent résulter de plusieurs événements de transposition de séquences mitochondriales vers le noyau, d'introgression, et/ou de tri incomplet des lignées. Ces deux derniers phénomènes étant également suspectés pour les six marqueurs nucléaires employés dans le chapitre V, cette hypothèse est probablement la plus parcimonieuse ; dissocier ces différents processus nécessitera le clonage d'amplicons du cob (la présence de plus de deux allèles [un mitohcondrial, un numt] étant indicatifs de plusieurs évènements de transposition) et l'utilisation d'un grand nombre de marqueurs nucléaires (distinction de l'introgression et du tri incomplet des lignées : voir Chapitre V pour détails).
Les différents traitements des numts ont un impact sur les analyses de génétique de conservation. En effet, le traitement qui consiste à retirer tous les sites présentant des ambigüités entraine la perte de toute l'information de différenciation des lignées. Retirer tous les individus présentant des ambigüités entraine une perte d'information sur la diversité et la différenciation des populations les moins échantillonnées mais n'a pas d'impact significatif sur les autres populations. Au contraire, garder les ambigüités ou les remplacer par des « N » dans le jeu de données mène à une différenciation globale moindre que le jeu de données vierge de contaminant nucléaire, du fait des séquences des numts supplémentaires. Les précédentes études sur les Procellariiformes qui n'ont pas relevé ces problèmes devraient donc être revues en tenant compte du biais qui a pu être occasionné par la présence de numts, mais également le bruit qui pourrait être occasionné par des évènements de duplication de gène intra-mitogénomique et d'hétéroplasmie.

Abstract

The translocation of mtDNA to the nuclear genome ("numts") can plague phylogeographic studies. To estimate the numt-related biases in estimates of population diversity, divergence and differentiation based on mtDNA, we targeted the mt cytochrome-b gene (*cob*) for 13 populations of the *Puffinus lherminieri* seabird species complex, including five lineages. *cob* homolog and paralog (numt) sequences were inferred by Sanger sequencing with and without exonuclease digestion of nuclear DNA. We inferred phylogenetic relationships among numt sequences. Numts formed monophyletic clades corresponding to three of the five mitochondrial lineages tested (the remaining two forming a paraphyletic group). 19% of numts alleles fell outside of their expected mitochondrial clade, a pattern consistent with multiple translocation events, incomplete lineage sorting, and / or introgression. We evaluated how numt-caused ambiguities in Sanger sequences impact statistics of genetic diversity, divergence and differentiation. Excluding individuals with ambiguities underestimates genetic diversity (4%) and differentiation (11%) among least-sampled populations. Removing all positions with ambiguities results in a 63% drop of the proportion of inter-lineage genetic variance. Co-analysing numts with mitochondrial sequences can therefore lead to severe bias in phylogeographic studies.

Keywords
Pseudogenes, phylogenetics, conservation, diversity, differentiation, numts, mito-numt discordance

Mitochondrial DNA (mtDNA) is the most popular marker for the study of molecular diversity in animals (Avise et al. 1987; Moritz et al. 1987; Zink & Barrowclough 2008). Reasons include the ease with which mtDNA is amplified, being 100-1000 times more abundant than genomic DNA (Bellis et al. 2003; Prado et al. 2007), levels of variation allowing to compare its signal within and across animal taxa (Brown 1985) , and the fact that these genes are single copy with no introns (but see (Ehara et al. 2000) and (Emblem et al. 2011) for counterexamples), with short intergenic regions (Gissi et al. 2008). MtDNA has played a major role in the study of evolution with thousands of studies per year published since 1998 (see (Desalle et al. 2017) for a review). However, the use of mtDNA as a marker of evolutionary history has also some drawbacks. The uniparental mode of inheritance of mitochondria, from the mother to the offspring, allows investigating only the genetic history and structure of female individuals (Avise et al. 1987). In addition, mtDNA can recombine, and is not always maternally inherited (e.g. (McCauley et al. 2005; Zouros et al. 1994)). Recombination has been detected in several taxa (see (White et al. 2008) for a review) as well as copies of mitochondrial markers in the nuclear genomes (Richly & Leister 2004b), the so-called nuclear-mitochondrial pseudogenes, or numts (Lopez et al. 1994). Heteroplasmy, i.e. the presence of several different mitogenomes within a single individual, may result from a biparental (or doubly uniparental, in the case of gonochoeric bivalves) transmission of mitochondria (e.g. (McCauley et al. 2005; Zouros et al. 1994)). Finally, some mtDNA markers appear to be duplicated within the mitogenome of several species (e.g. in birds (Gibb et al. 2013), salamanders (Mueller & Boore 2005), lizards (Moritz et al. 1987)). Whatever their causal origin, such multiple copies may affect sequence data interpretation. Indeed, copies may be equally amplified by PCR, generating double peaks and ambiguities in Sanger sequences. Alternatively, if only one of the copies is amplified, we may not know which one, and it may not necessarily be the same for all individuals. The presence of paralogs can therefore blur analyses of mtDNA genetic diversity, divergence and differentiation among populations. Numts, duplication and heteroplasmy have been found repeatedly in many taxa, among which birds (Gibb et al. 2013; Sorenson & Quinn 1998a). As bird blood is particularly poor in mitochondria (Sorenson & Quinn 1998b), numts can easily be co-amplified along with mtDNA (Sorenson & Quinn 1998b).

Petrels and albatrosses (order Procellariiformes) are a group of seabirds for which mitochondrial loci have been extensively used to study evolutionary history. Abbott et al (Abbott et al. 2005) first described a duplicated region within the mitogenome of a Procellariformes, the albatross *Thalassarche melanophris*. Other Procellariformes species were since found to have mitogenomes with duplications (Abbott et al. 2005; Eda et al. 2010; Gibb et al. 2007; Lounsberry et al. 2015; Torres et al. 2018). Although the composition of the duplication varies among taxa, the mitochondrial Control Region (CR) seems to always be present in at least two copies. Despite this widespread issue, only 11 of 40 studies that used mitochondrial markers on Procellariiformes and published after 2005 cited the study of Abbott (see Supplementary Material 1). These latter authors dealt with the issue of duplication of the CR either by simply removing mitochondrial CR sequences from analyses (n=3), designing PCR primers specific to one of the two copies of the CR to be sure to amplify and sequence only one copy (n=4), using the copy-specific primers designed in previous studies (n=3), the last only checking for the presence of ambiguities in the sequences as a sign of having

sequenced the two different copies. In addition, copy-specific primers were used in two other studies that did not cite Abbott et al (18). Apart from the issue of intra-chromosomal duplication events, seven other studies mentioned the likely co-amplification of numts along with mtDNA and three mentioned heteroplasmy as a source of ambiguities (see Supplementary Material 1). How to treat duplication, numts and heteroplasmy was actually rarely specified in these 40 studies: many authors (a) did not address the issue; when specified, some authors choose to (b) remove the sites presenting ambiguities (Kerr & Dove 2013), (c) to remove the individual sequences that presented ambiguities (Genovart et al. 2007). These four strategies may have major impacts on the estimation of genetic diversity, divergence and differentiation within and among populations. It has been shown in insects that *cox1* numts showed high degree of divergence with mitochondrial sequences (Haran et al. 2015) artificially rising the diversity and the number of estimated species (Song et al. 2008) and blurring the phylogenetic signal (Cristiano et al. 2012). However, the impact of different treatments of multiple copies on statistics used in phylogeography and conservation genetics was never quantified.

Here we use the petrel species *Puffinus lherminieri*, in which all three problems may occur simultaneously (Torres et al., 2018), as a worst-case scenario to investigate the effects of amalgamating mitochondrial loci and their paralogs when attempting to estimate genetic diversity, divergence and differentiation within and among populations. numts corresponding to several mitochondrial loci were recently detected in this group (Torres et al.). Moreover, this species is one of the many Procellariiformes that shows a duplicated region in the mitochondrial genome (Abbott et al. 2005; Eda et al. 2010; Gibb et al. 2007; Lounsberry et al. 2015; Torres et al. 2018). Finally, full mitogenome sequencing using Nanopore long-reads suggested the possibility of heteroplasmy in this species (Torres et al. 2018).

Here we first aim at evaluating the impact of these multiple-copies on population-level statistics used in genetic analyses, and second at evaluating to which extent numts affect the evolutionary scenarios obtained from mitochondrial cytochrome-b (*cob*) phylogenetic analyses. We generated three different datasets: we first Sanger-sequenced a 833-nt fragment of *cob*, retaining all ambiguities resulting from the putative presence of numts. We then treated all gDNA extracts with an exonuclease as to digest nDNA and eliminate numts, resulting in a second dataset free of ambiguities. We compared the first and second dataset to infer the numt sequences (third dataset), allowing direct comparisons of evolutionary history of the mitochondrial and numt loci. Numts being non-functional pseudogenes within the nuclear genome, they are expected to show different evolutionary dynamics in contrast to their functional counterparts in the mitochondrion (Bensasson et al. 2001). To confirm than the inferred sequences correspond to numts we looked for a reduced transversion bias compared with their corresponding mtDNA sequences, higher proportion of non-synonymous substitutions and more diffuse pattern of pairwise distances (Bensasson et al. 2001). Moreover we are expecting numts to form a monophyletic clade in the case of only one event of transposition, as they can then diverge from their mitochondrial counterpart, as already observed in birds (Arctander 1995; Zhang & Hewitt 1996b). Numts may further be considered as nuclear phylogenetic markers (Hazkani-Covo 2009; Zischler 2000). We therefore investigated the relative placement of numt sequences within the mitochondrial *cob* phylogeny, and reconstructed the phylogeographic history of the *P. lherminierii* complex

using numt sequences. To evaluate the effects of numt contamination on the analysis of mitochondrial sequences, we implemented four independent treatments to correct the ambiguities, reflecting common practice reported in the literature (see above): (a) removing the sites with ambiguities for all individuals, (b) removing all individuals with ambiguities, (c) keeping ambiguities. Then, we estimated population genetic parameters on all four datasets, treating them as uncontaminated mitochondrial sequences to test whether the artefactual merging of mitochondrial and numt loci can significantly bias common metrics used in conservation genetics.

Material and methods

1. DNA extraction, PCR amplification and sequencing of mitochondrial and numt loci

Our data set is an extension of a companion study (Torres et al.), in which blood samples from a total of 228 individuals were obtained from *Puffinus lheminieri* and *P. bailloni* shearwaters, totalling 13 different breeding populations. Extraction of total genomic DNA was carried out using NucleoSpin® Tissue XS Kit (Macherey & Nagel, Düren, Germany). Samples were incubated overnight in 4 mg of Proteinase K. Purified genomic DNA was eluted twice in 50 µL of TE buffer pre-heated at 70°C. DNA concentration was measured using Nanodrop (ND 1000 model) spectrophotometry. A portion of *cob* was amplified using shearwater-specific primers designed in (Torres et al.) (forward: Cytb-F1-Puf-CRI, GGCCTACTACTAGCYATACA; reverse: Cytb-R4-PUF-CRI, GTTARGATGAATAGGTTRGCG), with the proof-reading Ex-Taq polymerase (Takara Bio Europe). The PCR amplification protocol is detailed in (Torres et al.). We chose to target *cob* for three reasons. First, this marker was shown to be informative at the phylogeographic scale in this group of shearwaters (Torres et al.). Second, PCR amplification of *cob* can yield ambiguous DNA sequences, which have been attributed to numts based on the comparison of sequences obtained with and without digestion of linear gDNA. Third, this marker is apparently not duplicated in the mitochondrial genome of *P. lheminieri* (Torres et al 2018); intra-individual polymorphisms on mitochondrial sequences can therefore be attributed to the co-amplification of nuclear pseudogenes, rather than other types of mitochondrial paralogs.

To prevent numt co-amplification, we digested nuclear DNA with the ExonucleaseV (ExoV; NEB-M0345S), using the following protocol modified from Jayaprakash et al. (2015). One µg of gDNA was heated to 70°C to inactivate any residual Proteinase K from the extraction protocol. Digestion was then carried out, adding to the sample 1X NEB4 Buffer, 1 mM ATP, 0.3 U of ExoV, and 0.24 mg/mL of BSA. The mix was heated to 37°C during 48h, followed by 30 min at 70°C to inactivate the exonuclease. We compensated the lowered PCR yield by using BSA at a final concentration of 0.24 mg/mL. PCR products were sent to Eurofins (Anzinger Str. 7a / 85560 Ebersberg, Germany) for purification and Sanger sequencing in both directions. Chromatograms were checked visually and sequences were aligned to our sequences using MAFFT v 7.187 (Katoh et al. 2002). We then created two datasets of mitochondrial *cob* sequences. The first was called “CLEAN,” since all ambiguities due to numts were removed with the digestion protocol. The second, called “AMBIGUOUS,”

contained ambiguities due to the co-amplification of numts in our initial PCR reactions. Before conducting further analyses, we checked that the CLEAN and the AMBIGUOUS datasets did not contain any stop codon, indel or frame-shift mutations.

2. Mutation patterns within numts

Numt sequences were inferred by comparing numt-contaminated (AMBIGUOUS) and uncontamined (CLEAN) mitochondrial sequences, using a custom R script (R Core Team 2019: Supplementary Material 2), resulting in the NUMT dataset. Because numts are located in the diploid nuclear genome, this technique allows the determination, within an individual, of the consensus sequence of numt alleles with mutations that are unique to the nuclear genome (i.e. the "divergent" allele illustrated on Figure 1). We did not attempt phasing of the AMBIGUOUS dataset, as it is a mix of three alleles: two nuclear alleles and the mitochondrial haplotype.

The transition/transversion ratio within each lineage was calculated using the TiTvRatio function of the strataG R package (v 2.0.2, (Archer et al. 2017)). Nonsynonymous to synonymous mutation ratios (Ka/Ks) between DNA sequences from separate mitochondrial lineages were calculated (Nei & Gojoborit 1986) using DnaSP 5.10.1 (Librado & Rozas 2009). Non-synonymous to synonymous nucleotide diversity ratios (π_a/ π_s) were calculated within lineages. Both ratios were calculated for the CLEAN and NUMT datasets, using a 50 bp-wide sliding window, sliding every one bp. The same sliding window was used to calculate the Kimura 2-parameter pairwise distances (Kimura 1980) along the sequences using the R package ape (v 5.2 (Paradis et al. 2004)) and a custom R script. This was done for every lineage and for the pairwise comparison of each lineage, both for the CLEAN and NUMT datasets.

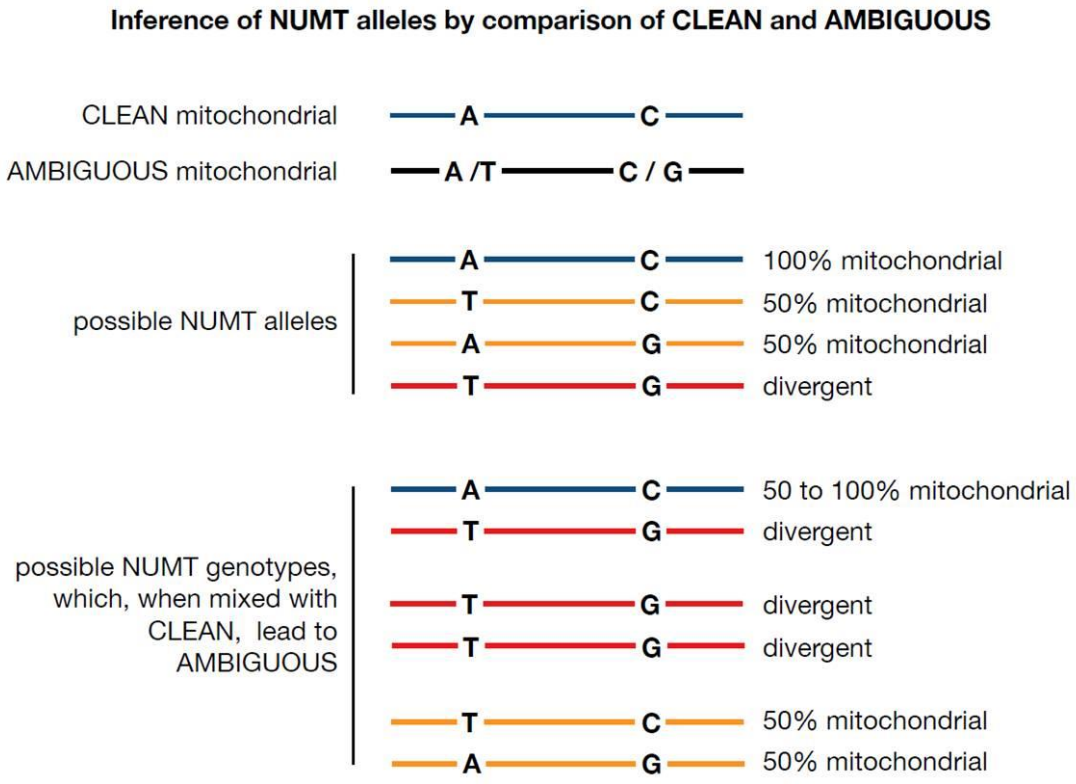
The number of nucleotide substitutions, at each codon position, was recorded in order to estimate codon position bias. Distributions obtained for the CLEAN and NUMT datasets were then compared, for each lineage, using a X^2 test (ref 42). As numt sequences are nonfunctional, they should accumulate mutations independently from codon positions, while mitochondrial sequences should preferentially accumulate mutations on the third position. Significant codon position bias among lineages is therefore indicative of multiple translocation events from the mitochondrial to the nuclear genome (Bensasson et al. 2001; Mundy et al. 1996). As we did not found any theoretical hypothesis about the neutral distribution of substitution among positions in birds, all distributions were also compared to an equal distribution of substitutions among positions (1-1-1). Distribution inferred for each lineage was finally compared to each other lineages for mitochondrial and numt sequences separately. Different codon position bias among lineages would imply that numts occurred from different functional ancestors and from independent transposition events. Codon usage bias (defined as differences in the frequency of synonymous codons) was calculated as the Relative Synonymous Codon Usage (RSCU) (Sharp et al. 1986) using the *uco* function of the seqinr R package (Charif & Lobry 2007). Frequency distributions inferred for each lineage were then compared to each other between mitochondrial and numt sequences using an ANOVA on a linear model. Distributions were compared among lineages with all mitochondrial sequences and all numt sequences. All distributions were finally compared to an equal distribution of codon usage (i.e. 1/64 for each codon). We also performed a Z-test of neutral evolution (Nei &

Gojoborit 1986) for the pair of mitochondrial/numt sequences for each individual using MEGA v.7.0.26 (Kumar et al. 2016). Significantly different selection pressures would indicate that numts sequences are under less-constrained pressure.

Phylogenetic relationships among CLEAN and the NUMT sequences were inferred using MrBayes v 3.2.6 (Ronquist et al. 2012b). The most likely model of evolution was inferred in JModeltest2 (Darriba et al. 2015) using BIC and LRT, and the parameters of this model were inputted in MrBayes. Phylogenetic inference was based on 500 000 generations sampled every 1000 generations with a 25% burn-in, with four chains. Stationarity of the chains was evaluated in Tracer v 1.7.0. The *cob* sequence for *Puffinus pacificus* (Genbank accession: AF076088.1) was used as an outgroup. The same phylogenetic analysis was conducted on the CLEAN and NUMT datasets separately, and trees were compared using the tanglegram function (Scornavacca et al. 2011) in

Figure 4.1: Inference of numt sequences from CLEAN and AMBIGUOUS

This figure gives an example of the pattern observed with an AMBIGUOUS sequence presenting ambiguities in two positions relatively to the CLEAN sequence of the same individual.
Four alleles are possible for the NUMT sequence corresponding to these two positions, with different levels of divergence relatively to the CLEAN sequence. The NUMT genotye leading to the ambiguities when mixed to the mitochondrial sequences have to be divergent of the mitochondrial sequence in the two positions. This leads to three different genotypes, which are summarized in the NUMT dataset in one sequence which is 100% divergent from the CLEAN dataset.



Dendroscope v3.5.10 (Huson & Scornavacca 2012). Kimura 2-parameter (K80) pairwise distances (Kimura 1980) were calculated between mitochondrial and numt sequences using the R package ape (Paradis et al. 2004).

Evaluating the impact of numts on statistics of diversity, divergence and differentiation

In the AMBIGUOUS dataset, three subsequent treatments were applied to correct the ambiguities, leading to three new independent datasets. First, we removed, for all individuals, the sites bearing ambiguities (hereon called the "SITE-LESS" dataset). Second, we removed all the individuals presenting ambiguities (hereon called the "INDIVIDUAL-LESS" dataset). We tested whether the distribution of ambiguous sequences among sampling populations followed a uniform distribution using the G-test (Sokal & Rohlf 1981) as implemented in the DescTools R package and tested the correlation between the average number of ambiguous sequences within populations and the number of ambiguities within sequences using Pearson's product-moment correlation. These tests were performed in R v. 3.6.0 (Team 2019) at alpha = 0.05.

We then evaluated whether these different treatments can introduce bias in population genetics studies by computing statistics of genetic diversity, sequence divergence, and population differentiation. Although numts are diploid sequences (unless mitochondrial DNA was translocated on the W chromosome), we calculated these statistics considering all data to be haploid (CLEAN, which contains haploid sequences and AMBIGUOUS, which contains a mix of haploid and diploid sequences), as to mimic a study in which numts went undetected. For each dataset, we calculated the number of parsimony-informative sites (S), haplotype diversity and nucleotide diversity (π), using the R packages *ips* v0.0.11 (Heibl 2008) and *pegas* v0.11 (Paradis 2010). For each population within each dataset, and for each statistic, 95% confidence intervals were computed using 1000 bootstrap replicates using the "sample" function in R. NeighborNet networks were inferred using SplitsTree v 4.14.2 (Huson & Bryant 2006) to further to visualize the effects of ambiguities (and their treatment) on genetic diversity and differentiation within and among populations. Global F_{ST} was calculated for each dataset using the Weir and Cockerham (Weir & Cockerham 1984) method implemented in the R *hierfstat* package (Goudet 2005). For each dataset, 1000 bootstrap replicates were produced with each population sampled. As the *hierfstat* method does not allow to input a model of nucleotide substitutions, we also estimated differentiation among populations by performing AMOVAs, and calculating pairwise Φ_{ST} using Arlequin v.3.1 (Excoffier et al. 2005). While this method does not associate confidence interval with *F* statistics, it does provide statistical significance of each variance component; this was computed based on 1000 bootstrap replicates. For AMOVAs, samples were stratified into five groups, corresponding to the five nominal lineages (*lherminieri*, *boydi* and *baroli* in the Atlantic, *nicolae* and *bailloni* in the Indian Ocean), and populations (i.e. sampling localities; see map in (Torres 2019)) within these groups. Pairwise Kimura 2-parameter distances, calculated for all pairs of haplotypes, were computed using the K2P model of substitution (Kimura 1980). The distributions of pairwise Φ_{ST} obtained by the different treatments were compared to the distribution obtained with the CLEAN dataset using the Kolmogorov-Smirnov test implemented in R.

Results

1. *Prevalence of ambiguities in mitochondrial cob sequences*

Two sequences of 833 bp, generated with and without the exonuclease treatment, were obtained for each of the 228 individuals. Four individuals were removed from the AMBIGUOUS dataset due to poor sequence quality (the corresponding CLEAN sequence was obtained for these four individuals, but the NUMT sequence could not be inferred). In the CLEAN dataset, 22 individuals, distributed among the five lineages, showed ambiguities (at 37 sites) after the digestion of nuclear DNA. Because these ambiguities could be due to incomplete digestion of linear DNA, PCR amplification and/or sequencing errors, or heteroplasmy (Torres et al. 2018), we removed them from the dataset.

The final CLEAN dataset contains therefore 206 sequences, which were submitted to GenBank (Acc. Numbers XXXXX); the NUMT dataset contain 202 sequences. In the AMBIGUOUS dataset, 75 chromatograms presented double-peaks (Table 1). The proportion of ambiguous sequences within a population varied from 0 (Funchal, taxon *baroli*) to 63% (*lherminieri* lineage from the Saint Barthélémy population, and *boydi* from Raso), with a

median of 29%. The average number of ambiguous sites per sequence varied from 0 (*baroli* from Funchal) to 61 sites (*nicolae* from the Seychelles) within populations, with a median value of 27 sites. Although the sequences with the highest number of ambiguities were from populations with the highest proportion of ambiguous sequences (e.g. *boydi* from Raso, *nicolae* from the Seychelles), these two variables were overall weakly correlated (Pearson's product-moment correlation: 0.55; t = 2.18, df = 11, p-value = 0.052). The observed geographical distribution of the ambiguous sequences and the ambiguous sites were not significantly different from a uniform distribution (G-test p-value = 0.99, G-statistic = 44). The CLEAN dataset presented no insertion, deletion, nonsense or stop-codon following translation. The NUMT dataset presented no frame-shift mutations. However, five numt sequences showed stop-codon following translation.

2. Comparisons of mutation patterns within the CLEAN and NUMT datasets

The transition/transversion ratio was much higher in CLEAN sequences than in NUMT sequences within each lineage, as 9 transversions were observed within the entire CLEAN dataset (Vs 57 transitions) and 30 transversions are present in the NUMT dataset (Vs 103 transitions) (Table 2). Codon position bias was significantly different between mt and numt sequences for every lineage (X^2 p-value $<2e^{-16}$, Supplementary Material 3 and Table 2) except for *boydi* (X^2 p-value: 0.99). In each lineage, except for *boydi*, the proportion of substitution in second position was nonzero in NUMT sequences, contrarily to CLEAN sequences (Table 2). In the latter dataset, codon position bias was significant for *lherminieri*, *boydi* and *baroli* (Supplementary Material 3ab). For NUMT sequences, codon position bias was significant for all lineages (Supplementary Material 3.a,b). The pairwise comparisons of codon position bias between each lineage, for CLEAN and NUMT sequences, respectively, led to non-significant differences (p-value of the X^2 test >0.77), with the exception of *lherminieri* / *boydi* for the NUMT dataset (p-value of the X^2 test $<2e^{-16}$). This pair represents the highest and the lowest $2^{nd}/3^{rd}$ position ratio, respectively.

All codon usage biases were significantly different from the equal theoretical distribution, but p-values were higher when testing numt sequences than mitochondrial sequences (Supplementary Material 3.c). Within each lineage mitochondrial and numt codon biases were not significantly different (all p-values >0.05, Supplementary Material 3.c). More pairwise comparisons of codon biases between lineages were significantly different in NUMT dataset (7 p-values >0.05) than in CLEAN dataset (5 p-values >0.05). On the basis of the distribution of mutations across codon positions, few translocation events may have taken place. The *boydi* population may contain recently-transposed numt sequences.

The number of nonsynonymous substitution is low in the entire CLEAN dataset, leading to Pi(a)/Pi(s) and Ka/Ks ratios constantly null all along the mt sequences (Table 2, Figure 2a). The only exceptions showed one (e.g. *boydi-bailloni* comparison) to five peaks (*lherminieri-bailloni* comparison) of proportion of nonsynonymous substitutions. More nonsynonymous substitutions are present in the NUMT dataset, leading to higher Pi(a)/Pi(s) and Ka/Ks ratios (Table 2, Figure 2a). The sliding window of K80 pairwise distances along the sequences revealed intra-lineage substitution hotspots within the mitochondrial sequences (Figure 2b). The distribution of the pairwise distance along the numt sequences is more homogeneous with wider and higher peaks in K80. This difference in the distribution of

pairwise sequence along the sequences is still visible when comparing populations having diverged for less than 300 ky, i.e. *boydi* Vs *baroli* and *bailloni* Vs *nicolae* (Torres et al.). The difference between the distributions of K80 distances fades when comparing all other populations (Figure

Table 4.1: Summary of the presence of ambiguous sequences within each tested lineage and population, among the 206 sequenced shearwater individuals

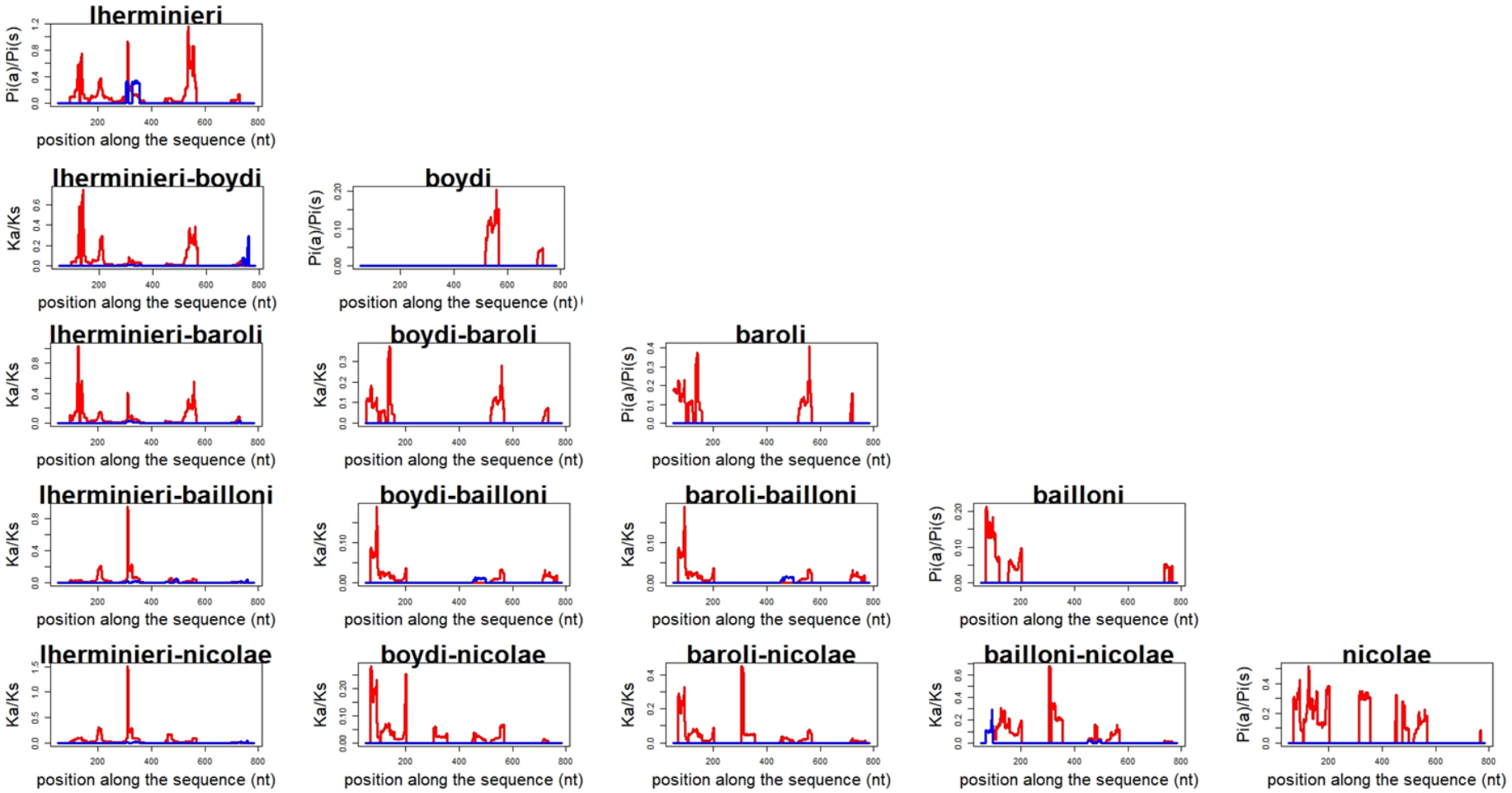
Lineage	Population	Number of sequenced individuals	Number of ambiguous sequences	Proportion of ambiguous sequences	Average number of ambiguous sites per sequence
lherminieri	Allencay	19	4	0.21	14
lherminieri	Longcay	19	3	0.16	22
lherminieri	Martinique	10	4	0.40	33
lherminieri	StBarthélémy	8	5	0.63	27
boydi	Raso	16	10	0.63	40
boydi	Cima	18	4	0.22	10
baroli	Mclara	14	7	0.50	19
baroli	Vila	16	5	0.31	36
baroli	Selvagem	5	1	0.20	10
baroli	Funchal	3	0	0	0
bailloni	North Reunion	25	4	0.16	29
bailloni	South Reunion	24	7	0.29	36
nicolae	Seychelles	29	12	0.41	61

Table 4.2: Difference of codon bias position and transition/transversion ratio between mitochondrial and numt sequences

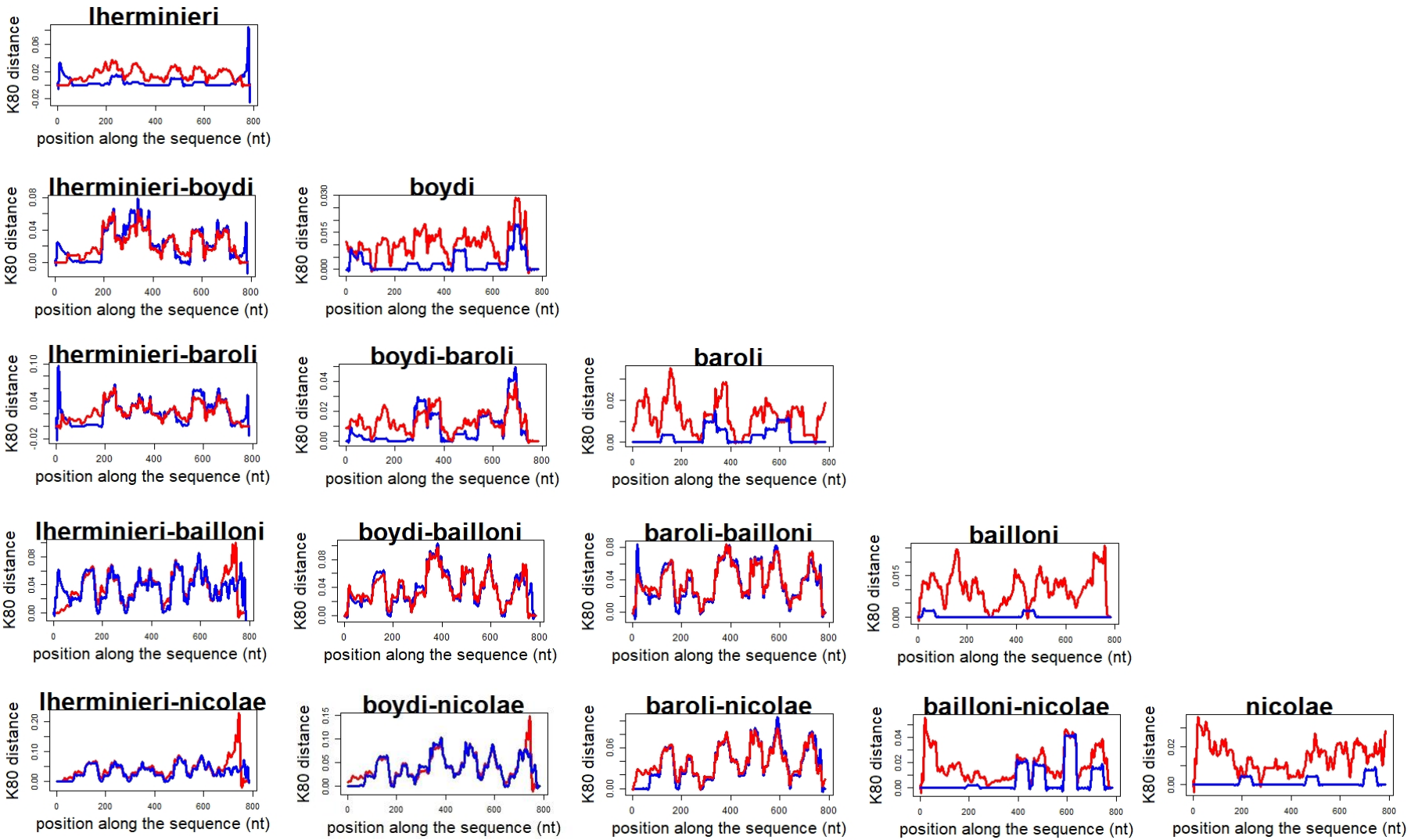
Lineage	Sequence number	CLEAN			NUMT			CLEAN			NUMT		
		1st position	2nd position	3rd position	1st position	2nd position	3rd position	Transitions	Transversions	Ti/Tv ratio	Transitions	Transversions	Ti/Tv ratio
lherminieri	18	5	0	10	20	10	39	13	2	6,5	54	15	3,6
boydi	17	2	0	8	8	0	36	9	1	9	41	5	8,2
baroli	13	1	0	6	9	1	27	7	0	Inf	33	4	8,3
bailloni	18	1	0	1	11	3	41	2	0	Inf	49	6	7,1
nicolae	9	1	0	2	9	3	30	3	0	Inf	38	6	6,3
All lineages	75	13	1	47	31	13	77	57	6	9,5	101	25	4

Figure 4.2: Sliding window of Pi(a)/Pi(s) and Ka/Ks ratios and Kimura 2-parameter pairwise distances along the CLEAN and NUMT sequences

a. Pi(a)/Pi(s) (intra-lineage) and Ka/Ks (inter-lineage) ratios are calculated along the sequence for CLEAN (blue) and NUMT (red) sequences



b. Average Kimura 2-parameter pairwise distances are calculated within and between each lineage along the sequence for CLEAN (blue) and NUMT (red) sequences



2b). The Z-test of neutral evolution performed on the pair of mitochondrial/numt sequences for each individual showed that 23 of the 75 numt sequences (31%) have significantly different selection pressures than their mitochondrial counterparts (Supplementary Material 3.d), hence being under less-constrained pressures.

3. Evolutionary history of numt compared to mitochondrial sequences

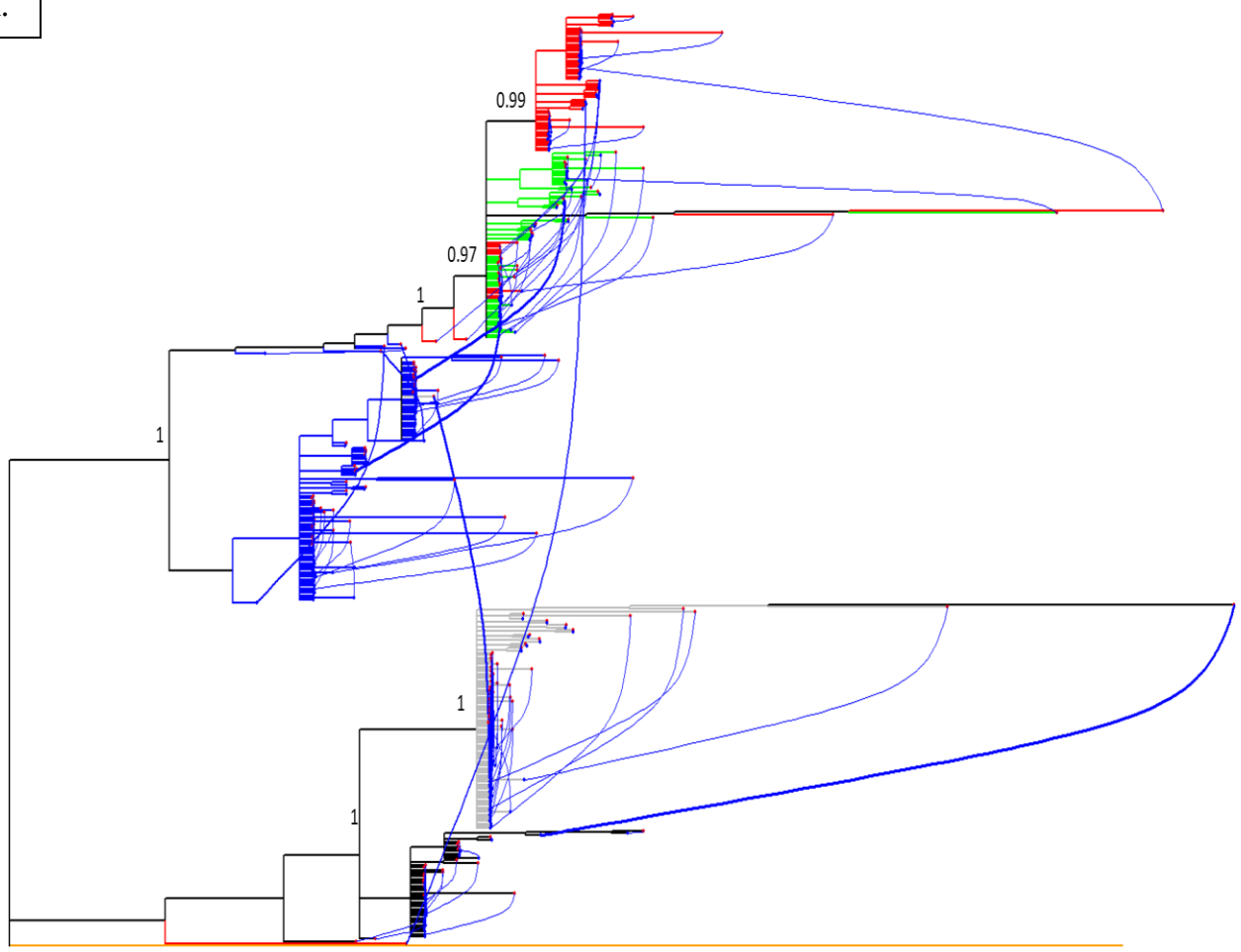
Numts were co-amplified for 75 out of the 202 individuals used in this study (37%). The numt sequences inferred by comparing the CLEAN and AMBIGUOUS datasets were submitted to GenBank (Acc. Numbers XXXXX). We analyzed the modification of topology of the phylogenetic tree by comparing the branch length of both a numt sequence and the mitochondrial sequence of the same individual. As the parameters of the models of substitution inferred by the Jmodeltest analyses were different for the mitochondrial and the Numt datasets (HKY with kappa = 26 for the mt and kappa = 13 for the numts) comparisons of branch lengths between these datasets should be done with caution. 29 of the numt sequences (39%) were placed close (<10% difference compared to the length of the branch calculated from the CLEAN alignment) to their associated mitochondrial sequence in the phylogenetic tree (Supplementary Material 4). The average pairwise distance between each of these numt sequences and their associated mitochondrial sequence was 0.76%. 10 numt sequences (13%) were in the same clade than their mt corresponding sequence but closer to the root of the tree with an average genetic distance of 0.33%. Conversely 24 numt sequences (32%) showed branches longer than their mitochondrial counterparts, with an average genetic distance of 0.88%. These values were superior to the intra-lineage average pairwise distance (0.13% to 0.34%) but inferior to the inter-lineage distance (0.95% to 3.83%) found with the mitochondrial sequences (Supplementary Material 5).

The phylogenetic position of numts and uncontaminated mitochondrial sequences were discordant in 14 individuals (19%; average genetic distance between the numt and mitochondrial sequence of 2%), numt sequences falling outside of the expected mitochondrial clade (Figure 3). Only two of these discordances corresponded to a shift in oceanic basin, a *bailloni* individual presenting a numt sequence placed in *lherminieri*, and a *baroli* individual presenting a numt sequence in the Indo-Pacific clade. Within the Atlantic lineages, five *lherminieri* individuals presented numt sequences in one of the two east Atlantic lineages: four *baroli* numt sequences were found in the *boydi* clade, while one *boydi* numt sequence was found in the *lherminieri* clade. Within the Indo-Pacific clade, two *nicolae* sequences were found in the *bailloni* clade. The tree inferred exclusively from numt sequences showed however the same topology as the cob tree, except for the individuals listed above.

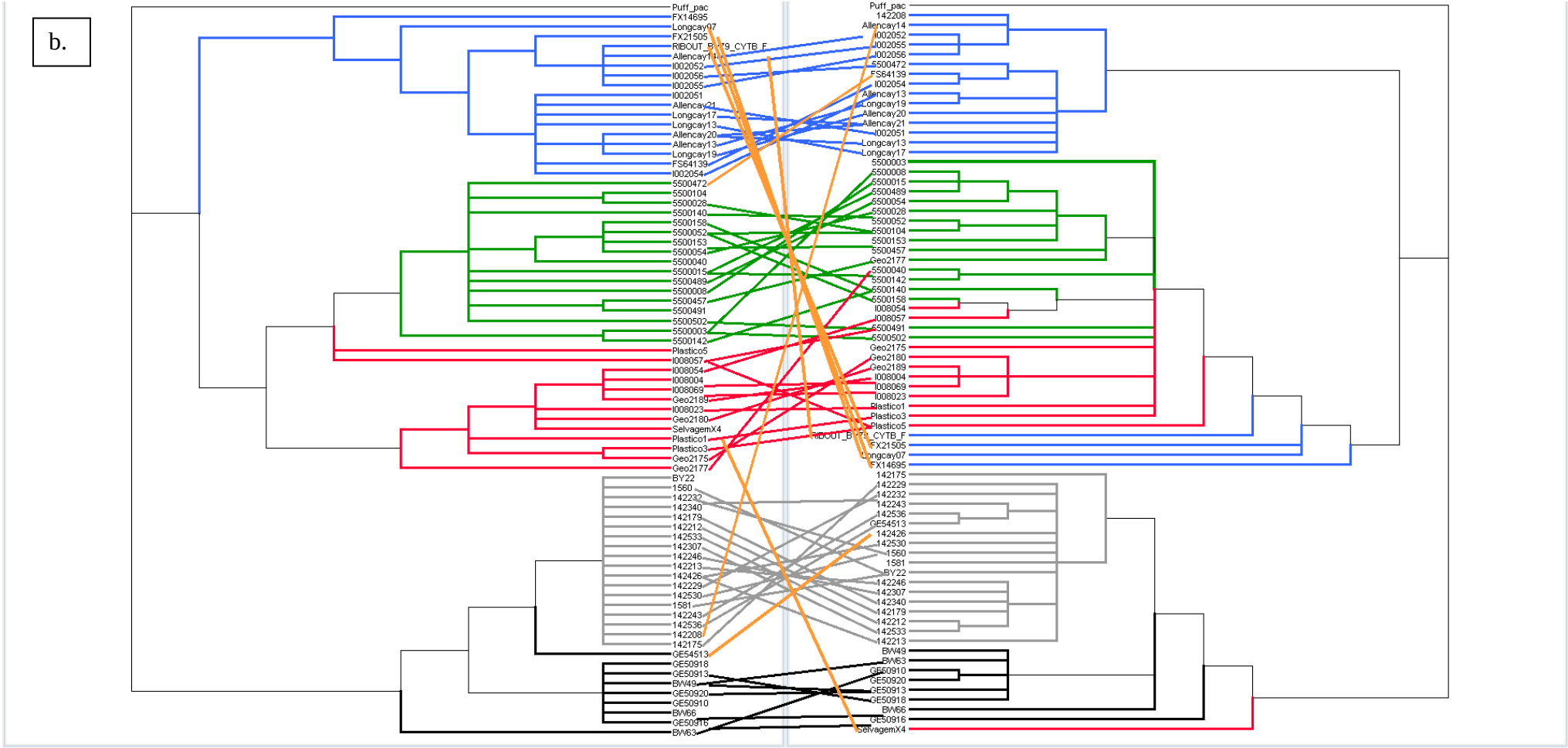
Figure 4.3: Phylogenetic relationships among mitochondrial and numt sequences

a. For each individual mitochondrial (blue) and numt (red) sequences are on the tree by a curve, which the width is proportional to the genetic distance between them. Discordant individuals which are also discordant in Torres et al. in prep (Chapter V) are indicated (see Supplementary material 4.6). Posterior probabilities >0.95 are indicated
b. Phylogenetic tree of mitochondrial sequences obtained by MrBayes.
c. Phylogenetic tree of numt sequences obtained by MrBayes.

a.



Only the individuals presenting different mitochondrial and numt sequences are used. Posterior probability superior to 0.95 are shown. Lineages are indicated based on mitochondrial data. *Puffinus pacificus* is indicated in orange.


b.
Puff_pac
FX14695
Longcay07
FX21505
Allencay14
I002052
I002056
I002055
I002051
Allencay21
Longcay17
Longcay13
Allencay20
Allencay13
Longcay19
FS64139
I002054
5500472
5500104
5500028
5500140
5500158
5500052
5500153
5500054
5500040
5500015
5500489
5500008
5500457
5500491
5500502
5500003
5500142
Plastico5
I008057
I008054
I008004
I008069
Geo2189
I008023
Geo2180
SelvagemX4
Plastico1
Plastico3
Geo2175
Geo2177
BY22
1560
142232
142340
142179
142212
142533
142307
142246
142213
142426
142229
142530
1581
142243
142536
142208
142175
GE54513
GE50918
GE50913
BW49
GE50920
GE50910
BW66
GE50916
BW63
Puff_pac
142208
Allencay14
I002052
I002055
I002056
5500472
FS64139
I002054
Allencay13
Longcay19
Allencay20
Allencay21
I002051
Longcay13
Longcay17
5500003
5500008
5500015
5500489
5500054
5500028
5500052
5500104
5500153
5500457
Geo2177
5500040
5500142
5500140
5500158
I008054
I008057
5500491
5500502
Geo2175
Geo2180
Geo2189
I008004
I008069
I008023
Plastico1
Plastico3
Plastico5
FX21505
Longcay07
FX14695
142175
142229
142232
142243
142536
GE54513
142426
142530
1560
1581
BY22
142246
142307
142340
142179
142212
142533
142213
BW49
BW63
GE50910
GE50920
GE50913
GE50918
BW66
GE50916
SelvagemX4

4. Impact of the different strategies to deal with numts

Only 113 sites (representing 14% of the sequence) were lost by removing all the sites presenting double-peaks. However, the number of parsimony informative sites dropped drastically between the CLEAN dataset and the SITE-LESS dataset, with 81% of the informative sites being lost (Table 3), 50-100% depending on the population (Supplementary Material 7). This indicates that most sites where numts diverged from original mitochondrial copies (i.e. intra-individual divergence) are the same as the sites of divergence among mitochondrial copies (i.e. inter-individual divergence), which is consistent with the sliding window analysis of average pairwise distance (Figure 2b). Indeed, on the 115 polymorphic sites between the numt and mitochondrial sequences, 60 (52%) were common to the mitochondrial and numt sequences, 43 (37%) specific to numts (with 26 singletons) and 12 (10%) specific to mitochondrial sequences (with 10 singletons). This loss of information impacted every statistic comparing the SITE-LESS dataset to the CLEAN dataset. The drop of haplotype and nucleotide diversity was significant both among (Table 3) and within populations (Supplementary Material 7). The same result is visible on haplotype networks of the SITE-LESS dataset, where almost all structuration was lost, compared to the CLEAN network (Figure 4). Most pairwise Φ_{ST} values were non-significant in the SITE-LESS dataset and consistently lower than in the CLEAN dataset (Table 4, Supplementary Material 7, Student's test: $p < 2.2e^{-16}$, $t = 11.8$). Moreover, most of the genetic variance was found due to within-lineage differentiation and variance due to inter-lineage differentiation was 70% lower than found with the CLEAN dataset (Table 4). No divergence times could be estimated since no significant structuration among lineages emerged.

Conversely, the other strategies (AMBIGUOUS and INDIVIDUAL-LESS) showed no significant change for any genetic diversity indexes (i.e. parsimony-informative sites, haplotype and nucleotide diversity; Table 3). However, in the INDIVIDUAL-LESS dataset, significant loss of diversity was observed for the populations showing the highest proportion of ambiguous sequences: Martinique and St Barthélémy from the *lherminhieri* lineage and Raso from the *boydi* lineage (Table 1 and Supplementary Material 7). The loss of diversity seemed not to cause a loss of diversification at the oceanic or lineage scale as shown by the network analyses where the closest results to CLEAN datasets were obtained with the INDIVIDUAL-LESS dataset (Figure 4). The proportion of variance explained by inter-lineage variation was higher than 0.05% compared to CLEAN (Table 4). However, lower diversity was associated with lower levels of population differentiation in the pairwise Φ_{ST} analysis, where most values involving Martinique, St Barthélémy or Raso populations decrease or are non-significant in the INDIVIDUAL-LESS dataset (Supplementary Material 8, Student's test: $p = 1$). The median and 95% confidence intervals of divergence times among lineages were remarkably similar to the one estimated with the CLEAN dataset (Table 5).

When comparing the AMBIGUOUS and the CLEAN datasets, all diversity (Table 4) and differentiation (Table 3, 4) statistics were not significantly different from each other, both among and within populations. The genetic variances explained at the inter-lineage scale were 1% to 0.5% lower, respectively. The only exception was the global F_{ST} analysis obtained by *hierfstat*, from which the AMBIGUOUS value was significantly lower than the CLEAN values (Table 4). This result was neither found in Φ_{ST} pairwise analyses nor in the

global Φ_{ST} analysis in the Arlequin software (Table 4 and Supplementary Material 8, Student's test: p = 0.34). This could be due to a difference of treating the missing data between the two methods.

Table 4.3: Diversity statistics

S represent the number of parsimony informative sites, h the haplotype diversity and π the nucleotide diversity. 95% confidence intervals (within brackets) were calculated based on 1000 sample bootstraps

Dataset	S	h	π	Sequences x nucleotides positions
CLEAN	59 [57-66]	0.95 [0.93-0.96]	0.0236 [0.0229-0.02383]	206 x 833
SITES-LESS	11 [9-24]	0.82 [0.37-0.85]	0.0018 [0.00157-0.00215]	206 x 720
INDIVIDUAL-LESS	54 [54-63]	0.95 [0.93-0.96]	0.0238 [0.02307-0.02411]	140x 833
AMBIGUOUS	59 [59-76]	0.94 [0.93-0.96]	0.0226 [0.02177-0.02312]	206 x 833
N	59 [58-74]	0.94 [0.93-0.96]	0.0226 [0.02186-0.02309]	206 x 833

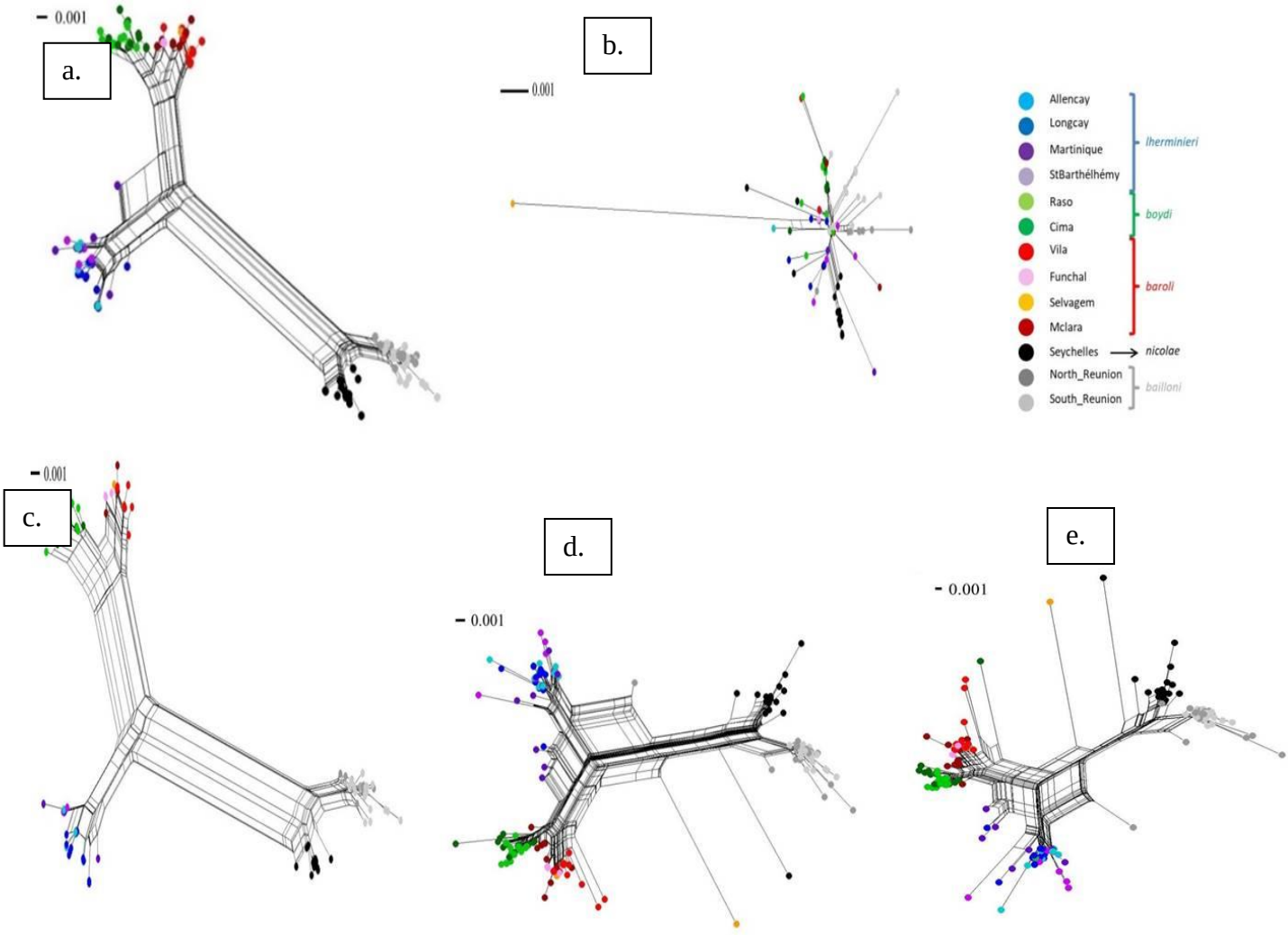
Table 4.4: Global differentiation statistics

Global F_{ST} was calculated both by the Weir and Cockerham method implemented in the R *hierfstat* package with 95% CI calculated from 1000 samples bootstrap and by an AMOVA in Arlequin. Va, Vb and Vc represent the percentage of variation among groups, among populations within groups and within populations of the AMOVA, respectively.

Dataset	Global F_{ST} hierfstat	Global Φ_{ST} Arlequin	Va	Vb	Vc
CLEAN	0.87 [0.88-0.92]	0.92	90.49***	1.75***	7.76***
SITES-LESS	0.69 [0.64-0.79]	0.37	27.98***	8.27***	63.75***
INDIVIDUAL-LESS	0.89 [0.89-0.9]	0.92	91.02***	1.59***	7.39***
AMBIGUOUS	0.77 [0.75-0.83]	0.92	90.19***	1.47***	8.34***
N	0.89 [0.88-0.92]	0.91	89.42***	1.72***	8.85***

Figure 4.4 Haplotype networks for the five datasets and all individuals

a. CLEAN dataset. b. SITES-LESS dataset. c. INDIVIDUAL-LESS dataset. d. AMBIGUOUS dataset. e. N dataset
The scale bars show how the length of a branch translates in sequence divergence. The unit is divergent nucleotides divided by the length of the sequence analysed.



Discussion

1. *Evolution of numt sequences*

The sequences inferred by comparison of the CLEAN and AMBIGUOUS datasets present several properties expected from numts. First, the fact that Sanger sequence ambiguities disappear when linear DNA is enzymatically digested indicates that these sequences are of nuclear origin. They show significantly lower transition/transversion ratio and codon position bias than their mitochondrial counterparts and significantly more nonsynonymous substitutions. The transition/transversion ratio, (values from *cob* in birds going from 1 to infinity (Belle et al. 2005)), have been shown to be lower in numts sequences than in their mitochondrial counterparts (e.g. in birds (Sato et al. 2001)). All of these clues are consistent with the fact that these sequences are under relaxed selection, contrarily to their mitochondrial counterpart (Bensasson et al. 2001). The results of the Z-tests of neutral evolution are partly consistent with this hypothesis. Weaker selection pressure makes more likely the appearance of non-synonymous substitutions, e.g. transversions in codon first or second positions, which leads to higher average pairwise distances within populations. However, following (Arctander 1995), the mutation rate of mitochondrial DNA is expected to be higher than nuclear DNA, including numts, in birds. The fact that we found more mutations among numts sequence than among mitochondrial sequences could be explained by the co-amplification and sequencing of multiple numt copies within individuals. Additional work to estimate the number of nuclear *cob* loci would help resolve this issue (see below).

These numts bear phylogenetic information, similar but not identical to the mitochondrial *cob* sequences. Three (*lherminieri*, *nicolae*, *bailloni*) of the five mitochondrial lineages were recovered as monophyletic in the numt tree. The east Atlantic *baroli* and *boydi* lineages appeared polyphyletic based on numts, as for other nuclear markers tested thus far (Torres et al.). Finally, multiple individuals displayed mitonuclear discordance.

In a previous study (Torres et al.), we have shown that nuclear markers (single copy introns: *βfib, csde, irf2, pax, rag1, tpm*) do not allow the discrimination of the *boydi* and *baroli* lineages, as found here with the numt sequences. We have also shown that several individuals presented mito-nuclear discordance (Toews & Brelsford 2012) for at least one out of the six nuclear markers tested. For example, two individuals of *boydi* (NE Atlantic) showed a *βfib* (nuclear) sequence placed in the Indo-Pacific clade (see Supplementary Material 6.a). Here, these two individuals presented numt sequences were basal to the East Atlantic population. Six other individuals showed a nuclear sequence placed in a discordant clade relatively to their geographic origin in (Torres et al.) and a numt sequence discordant from the mitochondrial sequence (Supplementary Material 6.a). Similarly, based on *βfib*, *pax*, *irf2* and *tpm*, an individual had been identified as a putative hybrid between the *nicolae* and *baroli* lineages. In the present study, a numt sequence is found in the Indo-Pacific clade, whereas its mitochondrial counterpart and its geographic origin indicate a *baroli* origin. The proportions of individuals being discordant both for numt sequences and for one nuclear marker (e.g. 25% of discordant individuals for *pax* are also discordant for numts) are similar to the proportions of individuals discordant for two non-numt nuclear markers (e.g. 25% of discordant individuals for *pax* are also discordant for *βfib*, see

Supplementary Material 6.b). Moreover, the patterns of discordance for numt sequences are similar to nuclear markers and could be explained by the same processes. We recently proposed that the aforementioned nuclear markers may reveal recent introgression and/or incomplete lineage sorting in this species complex. The same processes could affect a single nuclear copy of *cob*, resulting in mito-numt discordance. Numts, as part of the nuclear genome, are likely submitted to some of the same processes than nuclear loci used for phylogenetics and phylogeography. The polyphyletic placement of 37% of the numts sequences could therefore be the result of recent introgression and/or incomplete lineage sorting (ILS).

However, the observation that numts do not form a monophyletic clade (Bensasson et al. 2001) makes unlikely the hypothesis saying that all numt sequences occurred in only one translocation event. The more logical explanation is that multiple translocation events occurred from the mitochondrial to the nuclear genome. Under this scenario, the low divergence separating most numts from their mitochondrial counterpart suggests recent translocations followed by slow accumulation of mutations on nuclear *cob* copies (Zhang & Hewitt 1996b). Only two individuals bear numt signatures discordant with the mitochondrial clade at the ocean basin level (i.e. one Atlantic individual with an Indo-Pacific numt, and one Indo-Pacific individual with an Atlantic numt). These copies may have resulted from mitochondrial transposition events that occurred before the diversification of Atlantic and Indo-Pacific *Puffinus* lineages, at least 1 My ago (Torres et al.). This is possible, since 10-14 My translocation events have been reported (Lammers et al. 2017; Nacer & do Amaral 2017; Schiavo et al. 2017), but unlikely because it concerns only 2 out of 75 individuals for which numts were amplified.

Finding which process is responsible for the observed mito-numt discordance relies on determining the number of *cob* translocation event from the mitochondrial to the nuclear genome. All lineages present a significantly similar (but not identical – see exception for the *boydi* lineage) pattern of codon position bias and mitochondrial and numt sequences showed a similar pattern of codon usage bias. These results are a clue that all the numt sequences likely are the result of a unique, or few, event(s) of transposition from the mitochondrial genome to the nuclear genome (Bensasson et al. 2001). They may, however, suffer from insufficient statistical power. In addition, translocation events that happened in close succession may be hard to tease apart with these tests. Cloning of PCR products to estimate the number of numt alleles (e.g. (Hassanin et al. 2010)), digestion of PCR products with restriction enzymes (Zhang & Hewitt 1996a), looking for single-stranded conformation polymorphism (SSCP, (Sunnucks et al. 2000)), constant denaturant capillary electrophoresis (Li-Sucholeiki et al. 1999), or whole-genome NGS sequencing to map translocations (e.g. (Dayama et al. 2014)(Shokralla et al. 2014)) are techniques that would further help determine the number of mitonuclear translocation events. Of course, introgression and ILS could also have affected the molecular signature of numt sequences under the scenario of multiple translocations.

Numts could be helpful in phylogeographic study. Here, the phylogenetic relationships among mitochondrial lineages inferred from *cob* numts corroborates results from six nuclear markers suggestive of hybridization and introgression among these lineages (Torres et al.). Numts have previously been used to detect hybridization in mammals (Pérez et al. 2017), and have proved useful as phylogenetic markers (Hazkani-

Covo 2009; Ko et al. 2015). Focusing on Darwin's Finches, Sato et al proposed that mtDNA be used to resolve terminal nodes and numt deep nodes, as the first is more prone to saturation than the latter (Sato et al. 2001). Numts can bring a supplemental temporal information if their translocation to the nucleus occurred before evolutionary splits of interest (e.g. (Arctander & B 1995; Song et al. 2013)). However, numt characteristics, such as abundance, vary across taxa, even across populations (Bensasson et al. 2001). In mammals and birds, the nuclear genomes, including numts, evolve much slower than mitochondrial taxa (Arctander 1995), short numt sequences will show little inter-population divergence, and can therefore offer a snapshot in time of ancient mitochondrial DNA genetic structure. Inadvertent numt amplification can lead to incorrect phylogenies, for example by inferring a supplemental monophyletic group, which in fact corresponds to paralogs (Arctander 1995). Particular attention should therefore be ported to the amplification of numts with mitochondrial markers, which could false phylogenetic inferences but also population genetics analyses.

2. Impact of numts co-amplification on mitochondrial genetics analyses

Diversity and differentiation analyses on *cob* were inferred in a companion study, on the same individuals, using additional mitochondrial (*cox1* and CR) and nuclear (single copy introns: *βfib, csde, irf2, pax, rag1, tpm*) markers (Torres et al.). The evolutionary patterns inferred using *cob* are consistent with the analyses performed on *cox1* and CR (each of the five nominal groups form a monophyletic clade, no significant genetic structure within these groups). However, we have shown that nuclear markers were less resolutive within the *P. lherminieri* complex (separation of Indian and Atlantic Ocean lineages, East and West Atlantic lineages, no genetic structure evidenced within these groups). We proposed that the discordance between mitochondrial and nuclear data is due to incomplete lineage sorting and / or introgression events occurring within this complex. These processes may have an impact on the numt sequences as well.

Numt and mitochondrial datasets had similar levels of genetic diversity, but some numt sequences were more divergent from each other (45% of raw p distance >0.5%) than what was observed for mitochondrial sequences (e.g average *lherminieri* intra-lineage raw p distance =0.4%, Supplementary Material 4 and 5). Such sequences will bring high bias in analyses, artificially raising diversity within populations and accentuating differentiation (when using genetic distance-corrected statistics like Φ_{ST}) among populations. Only five of the numt sequences (7%) contain stop-codons; while verifying the absence of stop-codons in mitochondrial protein-coding sequences is essential, it does not guarantee weeding out all existing numts.

Mutations along numt sequences can coincide with the position of informative mutations on the mitochondrial locus, hence blurring the phylogeographic signal. The same pattern was found in ambiguities due to the duplication of the CR (60% of informative sites contain ambiguities in the CR, (Torres et al.) and data not shown). Hence, removing the sites comprising ambiguous data leads to a massive loss of information and must be avoided. In their study, Kerr & Dove (Kerr & Dove 2013) trimmed the alignment in the 3' extremity of the CR sequences, which seemed have little effect on their final results, since the analyses inferred with these sequences were similar to the analyses of the complete *cox1* sequences.

In our dataset, the effects of the INDIVIDUAL-LESS treatment were the most negative on the populations that were the most impacted by ambiguous sequences. In these populations, a significant proportion of the signal is loss in populations with >30% sequences bearing ambiguities. Genovart et al. (Genovart et al. 2007) used this method on well-sampled populations, and with a proportion of removed individuals < 30%, hence their results are likely little biased, but results from studies with lower sampling (<10 individuals per population) should be treated with caution. Using ambiguities led to reduced genetic information, although not significant in our dataset. Evidently, separating mitochondrial sequences from their paralogs is the best course of action. In this study, enzymatic digestion of linear DNA has proven an efficient and inexpensive strategy for doing so. If this is impractical however, we found that using either the INDIVIDUAL-LESS or AMBIGUOUS treatments of ambiguous datasets will minimize the bias in analyses of genetic diversity, population divergence and differentiation.

Annexe 3

Supplementary Material for Translocation of mitochondrial sequences into the nuclear genome may blur phylogeographic and conservation genetic studies in seabirds

Supplementary Material 4.1: Inventory of treatment of mitochondrial ambiguities in Procellariiformes

Studies that used mitochondrial markers on Procellariiformes genetics, published after the first paper describing a duplicated region within Procellariiformes mitogenomes (Abbott et al. 2005). The last column indicates the treatment that the authors applied to the ambiguities if mentioned.

Authors	Year	Mitochondrial marker	Cite Abbott et al. 2005	Cite Sorenson & Quinn 1998	Mention heteroplasmy	Treatment of ambiguities
Gómez-Díaz et al.	2006	CR				Ambiguities non mentionned
Peck	2006	CR	x			do not use Control Region
Friesen et al.	2007	CR				Ambiguities non mentionned
Genovart et al.	2007	*cob* and CR			x	remove individuals with ambiguities
Smith et al.	2007	CR	x			design specific primers
Lawrence et al.	2008	*cob* and CR	x			design specific primers
Zino et al.	2008	*cob*				Ambiguities non mentionned
Chambers et al.	2009	*cob*				Ambiguities non mentionned
Gómez-Díaz et al.	2009	CR	x		x	No ambiguities detected
Jesus et al.	2009	*cob*				No ambiguities detected
Techow et al.	2009	*cob*		x		No ambiguities detected
Ramirez et al.	2010	*cob*				Ambiguities non mentionned
Young	2010	CR	x			use specific primers
Brown et al.	2011	*cob*				Ambiguities non mentionned
Connan et al.	2011	*cob*				Ambiguities non mentionned
Pyle et al.	2011	*cob*		x		No ambiguities detected
Rains et al.	2011	CR	x			use specific primers
Robertson et al.	2011	*cob*				Ambiguities non mentionned
Welch et al.	2011	*cob*		x		Ambiguities non mentionned
Bicknell et al.	2012	CR				use specific primers
Gangloff et al.	2012	*co1* and *cob*				Ambiguities non mentionned
Kawakami et al.	2012	*cob*				Ambiguities non mentionned
Welch et al.	2012a	*cob*	x			do not use Control Region
Welch et al.	2012b	*cob*	x			do not use Control Region
Wiley et al.	2012	*cob*		x		No ambiguities detected

Deane	2011	CR				use specific primers
Gangloff et al.	2013	*co1* and *cob*		x		No ambiguities detected
Kerr & Dove	2013	*co1* and CR				trimm sites with ambiguities
Ramírez et al.	2013	CR				Ambiguities non mentionned
Tennyson et al.	2013	*cob*				Ambiguities non mentionned
Brace et al.	2014	*cob*				No ambiguities detected
Burg et al.	2014	CR	x			design specific primers
Lawrence et al.	2014	CR	x		x	design specific primers
Welch et al.	2014	*cob*		x		No ambiguities detected
Cibois et al.	2015	*cob*				Ambiguities non mentionned
Martínez-Gómez et al.	2015	*co1* and *cob*				Ambiguities non mentionned
Silva et al.	2015	CR				Ambiguities non mentionned
Wallace et al.	2017	*cob*		x		code "N"
Wold	2017	CR	x			use specific primers
Taylor et al.	2018	*cob*				Ambiguities non mentionned

Supplementary Material 4.2: R code

1. To extract numt sequences from ambiguous sequences

```
library(ape)
library(adephylo)
library(tidyverse)

# function to convert ambiguity into vector of corresponding nucleotides
iupac = function(amb) {
vec = NULL
if (amb == "M") vec <- c("A","C")
if (amb == "R") vec <- c("A","G")
if (amb == "W") vec <- c("A","T")
if (amb == "S") vec <- c("C","G")
if (amb == "Y") vec <- c("C","T")
if (amb == "K") vec <- c("G","T")
if (amb == "V") vec <- c("A","C","G")
if (amb == "H") vec <- c("A","C","T")
if (amb == "D") vec <- c("A","G","T")
if (amb == "B") vec <- c("C","G","T")
if (amb == "N") vec <- c("G","A","T","C")

if (amb == "m") vec <- c("a","c")
if (amb == "r") vec <- c("a","g")
if (amb == "w") vec <- c("a","t")
if (amb == "s") vec <- c("c","g")
if (amb == "y") vec <- c("c","t")
if (amb == "k") vec <- c("G","t")
if (amb == "v") vec <- c("a","c","g")
if (amb == "h") vec <- c("a","c","t")
if (amb == "d") vec <- c("a","g","t")
if (amb == "b") vec <- c("c","g","t")
if (amb == "n") vec <- c("g","a","t","c")

return (vec)
}

# for a given ambiguity (with 2 solutions), returns the nucleotide that is not in the clean
chromatogram
resolve = function(cl, mx){
        res = NULL
        iupac(mx) -> vec
        if(length(vec)>2) res <- NA
        if(length(vec)==2) res <- vec[vec != cl]
        return(res)
}

is.ambiguity = function(x){
        if(x == "A") return(FALSE) else(TRUE)
        if(x == "T") return(FALSE) else(TRUE)
        if(x == "C") return(FALSE) else(TRUE)
        if(x == "G") return(FALSE) else(TRUE)
        if(x == "-") return(FALSE) else(TRUE)
        if(x == "a") return(FALSE) else(TRUE)
        if(x == "t") return(FALSE) else(TRUE)
```

```
        if(x == "c") return(FALSE) else(TRUE)
        if(x == "g") return(FALSE) else(TRUE)
}
# build a data matrix of numt sequences.
matrix(data=NA, nrow=dim(Mito)[1], ncol=dim(Mito)[2]) -> Numt
for(i in 1:dim(Mito)[1]){                       # indiv par indiv
        for(j in 1:dim(Mito)[2]){               # pour un indiv, base par base
                        if (Mito[i,j] == mixed [i,j]) Numt[i,j] <- Mito[i,j]
                        if (Mito[i,j] != mixed [i,j]) {
                                if (is.ambiguity(mx)) Numt[i,j] <- resolve(Mito[i,j], mixed [i,j])
                                if (!is.ambiguity(mx)) Numt[i,j] <- mixed [i,j]
                                }
                        }
        }

## Create a file of both numt and mitochondrial sequence position on the tree
read.tree("infile.tree") -> tr
distRoot(tr)->d

lab <- as_tibble(tr$tip.label) %>%
  mutate(Position = 1:length(tr$tip.label)) %>%
  mutate(Dist = d)

lab2 <- lab %>%
  separate(col = value, into = c("Type","Code"), extra = "merge")

Mito <- lab2 %>%
  filter(Type == "Mito")

Numt <- lab2 %>%
  filter(Type == "Numt")

out <- Mito %>%
  left_join(Numt, by = "Code")

read.dna("infile.fas", format="fasta") -> ali
distances <- dist.dna(ali, as.matrix=TRUE)

# Get distances in "long" format
DNA_Dist <- distances %>%
  as_tibble() %>%
  mutate(First = colnames(.)) %>%
  gather(key = Second, value = Distance, -First) %>%    # Then extract code
  separate(col = First, into = c("First_Type","First_Code"), extra = "merge") %>%
  separate(col = Second, into = c("Second_Type","Second_Code"), extra = "merge") %>%
  filter(First_Code == Second_Code) %>%    # Retain only distances between a Mito sequence and
its "Numt" counterpart
  filter(First_Type == "Mito" & Second_Type == "Numt") %>%
  select(First_Code, Distance) %>%   # Cleaning and selecting only relevant variables
  rename(Code = First_Code, Numt_Mito_DNA_Dist = Distance)

DNA_Dist

# Join with first table
out <- out %>%
  left_join(DNA_Dist, by = "Code")
```

```
plot(tr,show.tip.label=F)
for (i in 1:nrow(out)) {
  points(out$Dist.y[i], out$Position.y[i], col="red", pch=19, cex=0.5)
  points(out$Dist.x[i], out$Position.x[i], col="blue", pch=19, cex=0.5)
  xspline(c(out$Dist.y[i],max(d)/2,out$Dist.x[i]),
        c(out$Position.y[i],median(c(out$Position.y[i],out$Position.x[i])),out$Position.x[i]),
        shape=1, lwd=0.1+100*out$Numt_Mito_DNA_Dist[i], border="blue")
}
```

2. To calculate bootstrapped values of genetic statistics

```
library(ips)
library(hierfstat)
library(pegas)

## Calculus of genetic statistic for bootstrapped populations ##

pop1<-read.dna("pop1.fas",format="fasta",as.matrix=T)

pisd<-matrix(ncol=1,nrow=1000) ## The aim is to obtain the number of parsimony informative
sites for one population from which the sequences are bootstrapped 1000 times

for (i in 1:1000)
{
  pisd[i,]<-pis(pop1[sample(nrow(pop1), replace = TRUE),], what="absolute") ## The sample
function allows to bootstrap the sequences of the populations and the pis function allows to
calculate the number of parsimony informative sites for one bootstrapped population of pop1
}
quantile(pisd,c(0.025,0.975))

## The same calculus is repeated for all populations and for other statistics, see below


hapdiv<-matrix(ncol=1,nrow=1000)
for (i in 1:1000)
{
  hapdiv[i,1]<-hap.div(pop1[sample(nrow(pop1), replace = TRUE),],pairwise.deletion=T) ##
Calculate the haplotype diversity for one bootstrapped population of pop1
}
quantile(hapdiv[,1],c(0.025,0.975))


nucdiv<-matrix(ncol=1,nrow=1000)
for (i in 1:1000)
{
  nucdiv[i,]<-nuc.div(pop1[sample(nrow(pop1), replace = TRUE),], pairwise.deletion = TRUE) ##
Calculate the nucleotide diversity for one bootstrapped population of pop1
}
quantile(nucdiv,c(0.025,0.975))


tajd<-matrix(ncol=1,nrow=1000)
for (i in 1:1000)
{
  tajd[i,1]<-tajima.test(pop1[sample(nrow(pop1), replace = TRUE),])$D ## Calculate the Tajima's
D value for one bootstrapped population of pop1
}
```

```
quantile(tajd[,1],c(0.025,0.975))


## Calculus of pairwise genetic distance among all sequences ##

allp<-read.dna(all_pops.fasta,format="fasta")

distdna<-dist.dna(allp,model="T92",pairwise.deletion = T) ## Calculate the pairwise genetic distances among sequences of all populations, the average pairwise distances within and among populations were then calculated manually


## Calculus of global F_ST for bootstrapped populations ##

pop1<-read.table(pop1.csv)
pop2<-read.table(pop2.csv)

##All the sequences of each population are displayed in a matrix in which the first column indicates the population of origin and following columns represent loci

bsp<-as.data.frame(matrix(NA,ncol=ncol(pop1),nrow = nrow(pop1)+nrow(pop2)))

bsfun<-function(bsp)
  {
    bsp<-rbind(pop1[sample(nrow(pop1), replace = TRUE),],pop2[sample(nrow(pop2), replace = TRUE),])
  }
## This function allows to bootstrap all the populations and to bind then in one dataset.

  wcd<-rep(NA,1000)
    for (i in 1:1000)
  {
    wcd[i]<-wc(bsfun(bsp),diploid = F)$FST ## Calculate the Weir and Cockerham estimates of global F_ST for one dataset bootstrapped for each population
  }
quantile(wcd,c(0.025,0.975))
```

3. To evaluate evolutionary tendencies on numt sequences

```
## Calculus of positions with substitutions in the alignment ##

ali<-read.dna("Puff_lherminieri_clean.fas",format="fasta")

posi=dista=distc=distg=distt=NULL
for(i in 1:833){
  sub.ali = ali[,i]
  base.freq((sub.ali)) -> d ## counts bases at position i
  ### d is a matrix ; choose comparison here
  d[1] -> da   ##indicates the number of "A" in the alignment
  d[2] -> dc   ##indicates the number of "C" in the alignment
  d[3] -> dg   ##indicates the number of "G" in the alignment
  d[4] -> dt   ##indicates the number of "T" in the alignment

  if (da!=1&&dc!=1&&dg!=1&&dt!=1)
  {
    posi = c(posi,i) da   ##concatenates all position with substitutions
```

```
    dista = c(dista, da)
    distc = c(distc, dc)
    distg = c(distg, dg)
    distt = c(distt, dt)
  }
}
count<-cbind(posi,dista,distc,distg,distt)

## Sliding window of average pairwise genetic distance along the alignments##

ali<- read.dna("Puff_lherminieri_clean.fas", format="fasta",as.character=T) # matrix of bases

le = length(ali[1,])
sw.size = 50  ### size of the sliding window
step = 1    ### distance between sliding windows
sw = seq(1,le, by=step)
vec = sw[1:(length(sw))-1]
sw.stop  = (vec + sw.size)[(vec + sw.size)<le]
sw.start = sw[1:(length(sw))-1][1:length(sw.stop)]

dist1=NULL
for(i in 1:length(sw.start)){
  sub.ali = ali[,sw.start[i]:sw.stop[i] ] ##reduces the alignment in the sliding window
  dist.dna(as.DNAbin(sub.ali), model="k80",as.matrix=T) -> d ## calculates the pairwise distance
  ### d is a matrix ; choose comparison here
  as.numeric(mean(d)) -> d1   ## averages the pairwise distance
  dist1 = c(dist1, d1)
}

smooth.spline(sw.start, dist1) -> spl1
```

Supplementary Material 4.3: Comparison of codon position bias between CLEAN and NUMT sequences

X^2 values and p-value results of X^2 test of comparison between CLEAN and NUMT sequences for each lineage are indicated in the diagonal. The comparisons of ratios between lineages for CLEAN dataset are indicated below the diagonal anf for the NUMT dataset above the diagonal.

The last two columns indicate the p-value of the X^2 test of comparison of each ratio with the 1-1-1 theroretical ratio.

	lherminieri	***boydi***	***baroli***	***bailloni***	***nicolae***	**1-1-1 CLEAN**	**1-1-1 NUMT**
lherminieri	2100400	2100400	0.56	0.23	0.13	10	19
boydi	0.11	0.01	0.05	0.06	0.09	10	49
baroli	0.29	0.03	73046	0.02	0.03	9	29
bailloni	0.11	0.36	0.51	297520	0.01	1	44
nicolae	1e-16	0.08	0.16	0.125	510200	2	29

	lherminieri	***boydi***	***baroli***	***bailloni***	***nicolae***	**1-1-1 CLEAN**	**1-1-1 NUMT**
lherminieri	<2e-16	<2e-16	0.77	0.89	0.93	0.007	8e-05
boydi	0.96	0.99	0.97	0.97	0.95	0.006	2e-11
baroli	0.86	0.98	<2e-16	0.98	0.98	0.01	5e-07
bailloni	0.94	0.83	0.77	<2e-16	0.99	0.61	3e-10
nicolae	1	0.96	0.92	0.93	<2e-16	0.37	5e-07

Supplementary Material 4.4 Genetic and Phyletic distance of numt sequences relative to mitochondrial sequence.

For each numt sequence pairwise distance between numt sequence and mitochondrial sequence is calculated with the dist.dna function in the R ape package; relative phyletic distance is calculated as $(D_{numt}-D_{mitochondrial})/D_{mitochondrial}$ D is the distance to the root calculated in the tree combining both numt and mt sequences of the Figure 4.3.a.

Individual	Genetic distance	Relative phyletic distance	Position of the numt sequence relatively to the mt sequence on the tree
BV79	0.0176	-0.45908	In another clade
FX21505	0.0184	-0.42987	In another clade
Longcay07	0.0134	-0.42173	In another clade
FX14695	0.0097	-0.33196	In another clade
5500491	0.0036	-0.21555	more basal in the clade
5500040	0.0063	-0.20969	more basal in the clade
Geo2175	0.0036	-0.18515	In another clade
Plastico3	0.0036	-0.13903	more basal in the clade
5500028	0.0024	-0.13508	more basal in the clade
Plastico1	0.0024	-0.12255	more basal in the clade
5500502	0.0036	-0.11015	more basal in the clade
5500457	0.0012	-0.10259	more basal in the clade
5500003	0.0012	-0.08349	more basal in the clade
Geo2177	0.0048	-0.07511	more basal in the clade
142229	0.0013	-0.03700	Close to the mtDNA
5500153	0.0053	-0.03275	Close to the mtDNA
5500054	0.0053	-0.02094	Close to the mtDNA
5500052	0.0053	-0.00419	Close to the mtDNA
Allencay13	0.0053	-0.00222	Close to the mtDNA
1581	0.0025	-0.00030	Close to the mtDNA
1560	0.0037	0.00353	Close to the mtDNA
142213	0.0013	0.02341	Close to the mtDNA
Longcay19	0.0012	0.02509	Close to the mtDNA
142232	0.0025	0.02537	Close to the mtDNA
Geo2189	0.0013	0.03477	Close to the mtDNA
Longcay17	0.0013	0.03594	Close to the mtDNA
Allencay21	0.0013	0.03602	Close to the mtDNA
142530	0.0013	0.03632	Close to the mtDNA
Allencay14	0.0014	0.03864	Close to the mtDNA
142426	0.0013	0.03943	Close to the mtDNA
I008023	0.0013	0.04053	Close to the mtDNA
GE50918	0.0013	0.04625	Close to the mtDNA
142208	0.0398	0.04822	In another clade
Plastico5	0.0012	0.05009	Close to the mtDNA
5500142	0.0024	0.05649	Close to the mtDNA
142340	0.0013	0.06156	Close to the mtDNA
142246	0.0013	0.06178	Close to the mtDNA
142307	0.0013	0.06316	Close to the mtDNA

142179	0.0013	0.06334	Close to the mtDNA
I008004	0.0027	0.07420	Close to the mtDNA
GE50913	0.0013	0.07426	Close to the mtDNA
BY22	0.0026	0.08403	Close to the mtDNA
GE50920	0.0014	0.08406	Close to the mtDNA
BW63	0.0013	0.09926	Close to the mtDNA
5500104	0.0024	0.11433	Longer branch
I002051	0.0026	0.13035	Longer branch
GE50910	0.0027	0.13252	Longer branch
BU80	0.0225	0.13716	In another clade
I002055	0.0052	0.16059	Longer branch
I008069	0.0053	0.16131	Longer branch
142212	0.0039	0.16679	Longer branch
BW66	0.0026	0.19241	In another clade
5500489	0.0036	0.20292	Longer branch
5500140	0.0079	0.20449	Longer branch
Geo2180	0.0066	0.21169	Longer branch
5500472	0.0267	0.22227	In another clade
BW49	0.004	0.22781	Longer branch
5500015	0.0062	0.22821	Longer branch
I002056	0.0066	0.23745	Longer branch
142175	0.0091	0.25182	Longer branch
I002052	0.0073	0.26223	Longer branch
142243	0.0103	0.30938	Longer branch
Longcay13	0.0121	0.32690	Longer branch
Allencay20	0.0133	0.32836	Longer branch
FS64139	0.0094	0.33320	Longer branch
5500008	0.0089	0.38320	Longer branch
142533	0.0119	0.39065	Longer branch
GE54513	0.038	0.52235	Longer branch & in another clade
142536	0.0186	0.52386	Longer branch
I008057	0.0148	0.57247	Longer branch & in another clade
GE50916	0.0118	0.57429	In another clade
I002054	0.0201	0.62825	Longer branch
SelvagemX4	0.0395	0.71217	In another clade
5500158	0.0201	0.83944	Longer branch
I008054	0.0199	0.94470	Longer branch & in another clade

Supplementary Material 4.5: Pairwise genetic distance between mitochondrial sequences

Average pairwise distances between mitochondrial sequences were calculated with the dist.dna function of the *ape* R package on the CLEAN dataset, standard errors are indicated in brackets.
Name of the mitochondrial lineages are displayed, n represents the number of sequences for each lineage

	lherminieri (n=56)	*lherminieri* (n=56)	*lherminieri* (n=56)	*lherminieri* (n=56)	*lherminieri* (n=56)
lherminieri (n=56)	0.4% (9^{e-5})				
boydi (n=34)	2.3% (8^{e-5})	0.3% (7^{e-5})			
baroli (n=38)	2.3% (6^{e-5})	1% (6^{e-5})	0.3% (7^{e-5})		
bailloni (n=49)	3.9% (4^{e-5})	4.1% (6^{e-5})	3.8% (5^{e-5})	0.2% (4^{e-5})	
nicolae (n=29)	3.5% (5^{e-5})	3.8% (6^{e-5})	3.7% (5^{e-5})	1% (5^{e-5})	0.1% (7^{e-5})

Supplementary Material 4.6: List of individuals displaying discordant phylogenetic patterns between mitochondrial and numt sequences (this study), and between mitochondrial and at least one nuclear markers (Torres et al in prep ; Chapter V).

The markers discordant in the other study as well as the geographic placement of the discordant sequence are indicated in the third and fourth column. The genetic and phyletic distances and time of divergence between the numt sequence and its mitochondrial counterpart found in this study are indicated in the last four columns.

		Nuclear marker Vs mitochondrial marker Torres et al. in prep; Chapter V		Numt sequence Vs mitochondrial sequence This study			
Individual	Lineage	marker discordant	Geographic placement of nuclear discordant sequence	Genetic distance	Relative phyletic distance	Time of divergence mt/numt	Position on the tree
142175	*bailloni*	*tpm*	Atlantic populations	0.0091	0.25182	[0.05;0.24]	Longer branch
142213	*bailloni*	*paxip1*	East Atlantic populations	0.0013	0.02341	[0;0]	Close to the mtDNA
BY22	*bailloni*	*tpm* and *irf2*	Atlantic populations	0.0026	0.08403	[0;0]	Close to the mtDNA
5500040	*boydi*	*βfib*	Indian populations	0.0063	-0.20969	[0;0]	more basal in the clade
5500491	*boydi*	*βfib*	Indian populations	0.0036	-0.21555	[0;0]	more basal in the clade
Longcay19	*lherminieri*	*βfib*	Indian populations	0.0012	0.02509	[0;0]	Close to the mtDNA
BW66	*nicolae*	*paxip1*	West Atlantic populations	0.0026	0.19241	[0.01;0.04]	more basal in the clade
GE50910	*nicolae*	*paxip1*	East Atlantic populations	0.0027	0.13252	[0.04;0.05]	Longer branch

Supplementary Material 4.7: Diversity statistics for each population, based on the CLEAN and SITES-LESS datasets.

a. Parsimony informative sites. b. Haplotype diversity. c. Nucleotide diversity (in %)

95% confidence intervals were calculated based on 1000 bootstrap replicates

a.	Allencay	Longcay	Martinique	St Bartélémy	Raso	Cima	Mclara	Vila	Selvagem	Funchal	North Reunion	South Reunion	Seychelles
CLEAN	0.71 [0.52-0.78]	0.82 [0.63-0.87]	0.87 [0.51-0.91]	0.64 [0.0-0.82]	0.86 [0.65-0.89]	0.87 [0.70-0.92]	0.86 [0.64-0.90]	0.72 [0.44-0.82]	0.8 [0.40-0.80]	0.67 [0.0-0.66]	0.84 [0.63-0.89]	0.43 [0.16-0.63]	0.52 [0.26-0.69]
SITES-LESS	0.11 [0.0-0.28]	0.38 [0.0-0.63]	0.73 [0.0-0.80]	0.46 [0.0-0.68]	0.13 [0.0-0.33]	0.4 [0.11-0.62]	0.14 [0.0-0.36]	0.24 [0.0-0.58]	0.4 [0.0-0.60]	0.67 [0.0-0.67]	0.73 [0.48-0.83]	0.08 [0.0-0.23]	0.46 [0.19-0.62]
INDIVIDUAL-LESS	0.66 [0.45-0.74]	0.83 [0.65-0.88]	0.6 [0.0-0.80]	0 [0-0]	0.6 [0.0-0.74]	0.9 [0.68-0.92]	0.81 [0.29-0.86]	0.76 [0.47-0.85]	1 [0.50-1]	0.67 [0.0-0.67]	0.85 [0.68-0.90]	0.4 [0.12-0.60]	0.57 [0.23-0.74]
AMBIGUOUS	0.67 [0.44-0.77]	0.82 [0.63-0.88]	0.84 [0.51-0.88]	0.64 [0.0-0.82]	0.83 [0.45-0.86]	0.84 [0.55-0.90]	0.81 [0.63-0.89]	0.72 [0.43-0.83]	0.9 [0.40-0.90]	0.67 [0.0-0.67]	0.86 [0.70-0.90]	0.16 [0.8-0.66]	0.61 [0.37-0.76]
N	0.67 [0.46-0.77]	0.82 [0.60-0.88]	0.84 [0.0-0.89]	0.64 [0.25-0.82]	0.83 [0.44-0.85]	0.84 [0.54-0.90]	0.81 [0.60-0.88]	0.72 [0.43-0.82]	0.9 [0.40-0.90]	0.67 [0.0-0.67]	0.86 [0.69-0.90]	0.08 [0.8-0.58]	0.61 [0.37-0.75]

b.	Allencay	Longcay	Martinique	StBartélémy	Raso	Cima	Mclara	Vila	Selvagem	Funchal	North Reunion	South Reunion	Seychelles
CLEAN	4 [3-7]	5 [3-9]	7 [1-20]	3 [0-5]	5 [2-7]	6 [3-9]	5 [2-7]	1 [1-6]	2 [0-2]	0 [0-0]	2 [2-7]	1 [0-3]	3 [1-6]
SITES-LESS	0 [0-2]	1 [0-4]	1 [0-4]	0 [0-1]	0 [0-1]	1 [0-3]	0 [0-1]	0 [0-2]	0 [0-6]	0 [0-0]	1 [1-6]	0 [0-1]	1 [1-6]
INDIVIDUAL-LESS	3 [3-8]	5 [3-7]	0 [0-8]	0 [0-0]	0 [0-4]	8 [3-23]	1 [0-4]	1 [1-4]	0 [0-2]	0 [0-0]	2 [2-7]	1 [0-2]	1 [1-6]
AMBIGUOUS	3 [3-8]	5 [3-9]	6 [2-12]	3 [0-7]	5 [3-7]	9 [4-24]	5 [2-8]	2 [1-7]	1 [0-15]	0 [0-0]	3 [2-16]	1 [0-4]	1 [1-11]
N	3 [3-8]	5 [3-9]	6 [2-11]	3 [0-5]	4 [2-6]	8 [3-23]	5 [2-6]	1 [1-6]	1 [0-15]	0 [0-0]	3 [2-15]	1 [0-3]	1 [1-9]

c.	Allencay	Longcay	Martinique	St Bartélémy	Raso	Cima	Mclara	Vila	Selvagem	Funchal	North Reunion	South Reunion	Seychelles
CLEAN	27 [16-32]	26 [15-33]	78 [23-120]	23 [03-33]	24 [16-29]	27 [17-34]	25 [15-31]	17 [07-26]	16 [04-16]	27 [0-27]	21 [13-28]	7 [2-11]	13 [6-19]
SITES-LESS	2 [0 -5]	5 [0-10]	14 [03-23]	04 [0-85]	2 [0 -5]	6 [2-11]	2 [0-5]	3 [0 -7]	48 [0-71]	1 [0 -1]	17 [11-23]	1 [0-3]	12 [5-19]
INDIVIDUAL-LESS	26 [15-33]	25 [15-29]	33 [0-52]	0 [0-0]	16 [0-26]	62 [18-113]	2 [6-278]	15 [6-212]	21 [06-21]	24 [0-244]	24 [15-31]	06 [02-11]	15 [04-25]
AMBIGUOUS	25 [16-31]	24 [14-29]	46 [20-61]	22 [04-32]	2 [11-26]	49 [15-97]	25 [16-31]	16 [06-25]	110 [04-160]	33 [0-33]	31 [16-48]	5 [1-09]	15 [7-23]
N	25 [15-32]	24 [15-29]	46 [21-61]	22 [03-32]	2 [11-26]	49 [15-95]	25 [16-30]	16 [065-25]	110 [05-158]	33 [0-33]	31 [16-49]	05 [01-08]	15 [07-23]

Supplementary Material 4.8: Pairwise Φ_{ST} for all populations, across all datasets

All values > 0 are significant with a p-value < 0.005

a. CLEAN dataset b. SITES-LESS dataset. c. INDIVIDUAL-LESS dataset. d. AMBIGUOUS dataset. e. N dataset.

Blue: *lherminieri*, green: *boydi*, red: *baroli*, grey: *bailloni*. Triple band indicates the separation between intra- and inter-ocean values.

a.	**Allencay**	**Longcay**	**Martinique**	**St Barthelemy**	**Raso**	**Cima**	**Mclara**	**Vila**	**Selvagem**	**Funchal**	**North Reunion**	**South Reunion**	**Seychelles**
Allencay	0.00												
Longcay	0.16	0.00											
Martinique	0.10	0.30	0.00										
StBarthelemy	0.00	0.48	0.00	0.00									
Raso	0.89	0.90	0.78	0.90	0.00								
Cima	0.89	0.90	0.79	0.90	0.00	0.00							
Mclara	0.89	0.90	0.77	0.91	0.70	0.69	0.00						
Vila	0.88	0.89	0.77	0.90	0.73	0.73	0.00	0.00					
Selvagem	0.87	0.89	0.69	0.90	0.71	0.71	0.00	0.00	0.00				
Funchal	0.86	0.88	0.65	0.88	0.53	0.53	0.27	0.42	0.39	0.00			
North Reunion	0.93	0.94	0.90	0.94	0.94	0.94	0.95	0.94	0.94	0.94	0.00		
North Reunion	0.95	0.96	0.92	0.97	0.96	0.96	0.97	0.97	0.98	0.98	0.09	0.00	
Seychelles	0.95	0.95	0.91	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.77	0.85	0.00

b.	Allencay	Longcay	Martinique	St Barthelemy	Raso	Cima	Mclara	Vila	Selvagem	Funchal	North Reunion	South Reunion	Seychelles
Allencay	0.00												
Longcay	0.00	0.00											
Martinique	0.00	0.00	0.00										
StBarthelemy	0.62	1.00	0.53	0.00									
Raso	0.00	0.00	0.00	0.59	0.00								
Cima	0.00	0.00	0.00	0.36	0.00	0.00							
Mclara	0.00	0.00	0.00	0.39	0.00	0.00	0.00						
Vila	0.00	0.00	0.00	0.59	0.00	0.00	0.00	0.00					
Selvagem	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00				
Funchal	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00			
North Reunion	0.11	0.11	0.00	0.21	0.10	0.12	0.00	0.10	0.00	0.00	0.00		
North Reunion	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.00	
Seychelles	0.58	0.60	0.48	0.60	0.56	0.53	0.55	0.56	0.41	0.46	0.46	0.63	0.00

c.	Allencay	Longcay	Martinique	St Barthelemy	Raso	Cima	Mclara	Vila	Selvagem	Funchal	North Reunion	South Reunion	Seychelles
Allencay	0.00												
Longcay	0.21	0.00											
Martinique	0.00	0.53	0.00										
StBarthelemey	0.00	0.65	0.00	0.00									
Raso	0.92	0.93	0.91	0.94	0.00								
Cima	0.83	0.84	0.79	0.78	0.00	0.00							
Mclara	0.92	0.93	0.91	0.94	0.81	0.60	0.00						
Vila	0.93	0.94	0.93	0.95	0.86	0.68	0.00	0.00					
Selvagem	0.92	0.93	0.92	0.00	0.85	0.60	0.00	0.00	0.00				
Funchal	0.91	0.93	0.90	0.00	0.73	0.46	0.00	0.52	0.43	0.00			
North Reunion	0.94	0.95	0.94	0.95	0.95	0.90	0.94	0.95	0.94	0.94	0.00		
North Reunion	0.96	0.96	0.96	0.98	0.97	0.91	0.97	0.97	0.97	0.97	0.07	0.00	
Seychelles	0.95	0.96	0.95	0.96	0.96	0.90	0.96	0.96	0.97	0.96	0.82	0.88	0.00

d.	Allencay	Longcay	Martinique	St Barthelemy	Raso	Cima	Mclara	Vila	Selvagem	Funchal	North Reunion	South Reunion	Seychelles
Allencay	0.00												
Longcay	0.14	0.00											
Martinique	0.17	0.35	0.00										
StBarthelemy	0.13	0.43	0.00	0.00									
Raso	0.91	0.91	0.85	0.91	0.00								
Cima	0.84	0.84	0.75	0.82	0.00	0.00							
Mclara	0.90	0.90	0.83	0.90	0.70	0.57	0.00						
Vila	0.91	0.91	0.86	0.92	0.80	0.68	0.13	0.00					
Selvagem	0.83	0.83	0.68	0.77	0.67	0.56	0.17	0.00	0.00				
Funchal	0.91	0.91	0.81	0.91	0.69	0.49	0.00	0.42	0.00	0.00			
North Reunion	0.93	0.93	0.91	0.93	0.93	0.91	0.93	0.93	0.87	0.92	0.00		
North Reunion	0.97	0.97	0.96	0.98	0.97	0.94	0.97	0.97	0.93	0.98	0.05	0.00	
Seychelles	0.96	0.96	0.94	0.96	0.96	0.93	0.96	0.96	0.92	0.97	0.78	0.88	0.00

e.	Allencay	Longcay	Martinique	St Barthelemy	Raso	Cima	Mclara	Vila	Selvagem	Funchal	North Reunion	South Reunion	Seychelles
Allencay	0.00												
Longcay	0.15	0.00											
Martinique	0.16	0.36	0.00										
StBarthelemy	0.00	0.42	0.00	0.00									
Raso	0.90	0.90	0.84	0.91	0.00								
Cima	0.84	0.84	0.75	0.83	0.00	0.00							
Mclara	0.89	0.89	0.80	0.89	0.69	0.56	0.00						
Vila	0.91	0.91	0.85	0.92	0.81	0.70	0.17	0.00					
Selvagem	0.84	0.84	0.70	0.81	0.71	0.59	0.00	0.00	0.00				
Funchal	0.90	0.90	0.78	0.90	0.72	0.54	0.00	0.38	0.00	0.00			
North Reunion	0.92	0.92	0.89	0.92	0.93	0.90	0.92	0.93	0.88	0.92	0.00		
North Reunion	0.96	0.96	0.95	0.97	0.97	0.94	0.97	0.97	0.95	0.98	0.07	0.00	
Seychelles	0.94	0.94	0.92	0.95	0.95	0.93	0.95	0.96	0.93	0.96	0.77	0.90	0.00

Chapitre V

Sea surface temperature, rather than land mass or geographical distance, may drive genetic differentiation in a species complex of highly-dispersive seabirds

Submitted in Molecular Phylogenetics and Evolution

Lucas Torres[1,2], Eric Pante[2], Jacob González-Solís[3], Amélia Viricel[2], Cécile Ribout[1], Francis Zino[4], Will MacKin[5], Carine Precheur[6], Julie Tourmetz[7], Licia Calabrese[8, 9], Teresa Militão[3], Laura Zango[3], Hadoram Shirihai[10] & Vincent Bretagnolle[1]

[1] Centre d'Etudes Biologiques de Chizé, UMR 7372, Université de La Rochelle – CNRS, Beauvoir sur Niort, 79360, France
[2] Laboratoire LIENSs, UMR 7266 University of La Rochelle – CNRS, 2 rue Olympe de Gouges, La Rochelle, 17000, France
[3] Institut de Recerca de la Biodiversitat (IRBio) and Department de Biologia Evolutiva, Ecologia i Ciències Ambientals (BEECA), Universitat de Barcelona, Av Diagonal 643, Barcelona 08028, Spain
[4] Rua Dr Pita 7, 9000 Funchal, Madeira
[5] 3913 Sterling Ridge Ln, Durham, NC 27707
[6] Karibiod, 2 impasse hibiscus, Place d'Armes, 97232 Lamentin, Martinique
[7] Société d'Etudes Ornithologiques de La Réunion, 13, ruelle des Orchidées, Cambuston, 97440 Saint André, Reunion Island, France
[8] Island Conservation Society, BP 775, Pointe Larue, Mahé, Seychelles
[9] Island Biodiversity & Conservation center, Faculty of Business & Sustainable Development, University of Seychelles
[10] Erga & Dudi Rivis c/o Hadoram Shirihai, Hela 24, Kfar Truman 731500, Israel

Etudier quels sont les facteurs de la différenciation des organismes est essentiel afin de mieux comprendre comment la diversité des organismes s'est formée. Comme nous l'avons vu, les oiseaux marins montrent de très fortes capacités de dispersion et il est difficile d'imaginer des barrières suffisantes pour empêcher la dispersion de tels organismes (e.g. Croxall 2005). Pourtant plusieurs populations d'oiseaux marins montrent des patrons de structuration génétique, à la fois entre espèces et au sein d'une même espèce (e.g. Genovart et al. 2007, 2012; Gómez-Díaz et al. 2009). Plusieurs processus ont été mis en avant pour expliquer ce phénomène, notamment la présence de barrières continentales, l'isolement par la distance, la ségrégation entre zones de reproduction et de non-reproduction ou encore le comportement philopatrique de ces oiseaux (voir Friesen 2015, Friesen et al. 2007a). Toutefois aucune étude n'a tenté de comparer ces différents facteurs simultanément afin de dissocier le rôle de chacun d'eux.

Nous avons mené une étude phylogéographique multi-locus sur un complexe d'espèces de puffin (Puffinus *lherminieri/bailloni*) nichant sur des îles tropicales et subtropicales dans les océans Indien et Atlantique. Les différentes lignées qui la composent sont écologiquement et morphologiquement très similaires et leur systématique est encore méconnue, les précédentes études moléculaires ne s'étant basées que sur un seul marqueur (Austin et al. 2004; Heidrich et al. 1997). Notre étude comporte plus de 20 individus pour les 14 populations existantes des 5 lignées représentatives du complexe. Nous avons séquencé trois marqueurs mitochondriaux et six marqueurs nucléaires afin d'étudier l'importance relative de différents facteurs de différenciation possibles chez ces oiseaux marins.

Nous avons tout d'abord constaté une nette différenciation entre les populations d'Atlantique et les populations de l'océan Indien, à la fois avec les marqueurs mitochondriaux et nucléaires. Ce qui pourrait suggérer un fort impact du continent Africain comme barrière aux flux de gènes. Toutefois le faible temps de divergence trouvé (il y a 1.28 à 1.43 Million d'années) implique que les populations des deux océans échangeaient des flux de gènes jusqu'à très récemment. Nous attribuons la séparation à des changements de température de surface de l'océan, facteur qui n'avait pas été mis en avant jusqu'à présent. L'importance de ce facteur est confirmée par la reconstruction du scénario de divergence des lignées nominales au sein des océans, qui correspond à des changements de température déterminés. La différenciation que l'on observe entre les lignées et qui a été formée par des changements de température océanique a pu être maintenue par des ségrégations de zones de reproduction et de non-reproduction entre les lignées. De plus aucune différenciation n'est observée entre les populations d'un même archipel, ce qui remet en question le haut niveau de philopatrie observé chez ces oiseaux marins. Nous avons donc mis en avant dans cette étude multi-locus l'importance de la ségrégation des zones de reproduction et de non-reproduction et surtout l'importance des changements de températures de surface des océans comme facteurs de différenciation chez les oiseaux marins.

Parallèlement à la reconstruction du scénario phylogéographique nous avons montré que les marqueurs mitochondriaux montraient une différenciation des lignées plus forte que les marqueurs nucléaires. Nous expliquons cette discordance d'abord par un tri de lignée incomplet des marqueurs nucléaires (McKay & Zink 2010; Toews & Brelsford 2012). Nous montrons également que cette discordance pourrait être expliquée par une possible hybridation entre plusieurs lignées et une dispersion femelle-biaisée dans les populations les plus larges (Petit & Excoffier 2009). Nous montrons ainsi l'importance de différents phénomènes moléculaires ou démographiques sur l'information génétique utilisée dans les analyses phylogéographiques.

Abstract

Seabirds, particularly Procellariformes, are highly mobile organisms with a great capacity for long dispersal, though simultaneously showing high philopatry, two conflicting characteristics that may lead to contrasted patterns of genetic population structure. Landmasses were suggested to explain differentiation patterns observed in seabirds, but philopatry, isolation-by-distance, segregation between breeding and non-breeding zones, and oceanographic conditions (sea surface temperatures) may also contribute to differentiation patterns. No study has simultaneously contrasted the multiple factors contributing to the diversification of seabird species, especially in the grey zone of speciation. We conducted a multi-locus phylogeographic study on a widespread shearwater species complex (*Puffinus lherminieri*/*bailloni*), showing highly homogeneous morphology. We sequenced three mitochondrial and six nuclear markers on all extant populations (five nominal lineages, 14 populations). We found sharp differentiation among populations separated by the African continent with both mitochondrial and nuclear markers, while only mitochondrial markers allowed characterizing the five nominal lineages. No differentiation could be detected within these five lineages, questioning the strong level of philopatry showed by these shearwaters. Finally, we propose that Atlantic populations likely originated from the Indian Ocean. Within the Atlantic, a stepping-stone process accounts for the current distribution. Based on our divergence times estimates, we suggest that the observed pattern of differentiation mostly resulted from variation in sea surface temperatures.

Keywords
Puffinus, divergence, phylogeography, multi-locus, philopatry, mito-nuclear discordance

Introduction

Species are cornerstone in evolution and are used as study units in both evolutionary and conservation biology (Butlin et al. 2012, Wake 2006). Though divergence with gene-flow has been theorized and observed (review in Pinho & Hey 2010), the model of allopatric speciation predominates by far in the literature (Stroud & Losos 2016). In this model, a physical barrier to gene flow catalyzes genetic differentiation between populations, through selection and/or genetic drift, eventually followed by other pre- or post-zygotic barriers (Coyne & Orr 2004). In practice however, the mechanisms that impede gene flow and promote differentiation are multifactorial and still poorly understood (Butlin et al. 2012, Coyne & Orr 2004, Price 2008). In particular, geographic barriers alone may not explain the differentiation of populations in highly dispersive species, e.g. marine birds (e.g. Genovart et al. 2007), birds of prey (e.g. Doyle et al. 2016), mammals (Hassanin et al. 2018) or plants (Sanz et al. 2014). Seabirds are a case in point: their wide geographic distribution and dispersal ability should theoretically maintain high levels of gene flow, but many seabirds show surprisingly strong geographic population structure, a pattern attributed to their high degree of philopatry (Abbott & Double 2003, Deane 2013, Smith et al. 2007a).

Present or historic landmasses were identified as the most important barriers to gene flow in seabirds (Friesen 2015, Friesen et al. 2007a) though they cannot explain every differentiation patterns, and other factors such as philopatry, isolation-by-distance, or segregation between breeding and non-breeding zones may play a role. So far, however, no study has contrasted all these processes simultaneously. Moreover, most previous studies were based on maternally-inherited mitochondrial markers (Friesen 2015), while gene flow is expected to vary between sexes given stronger male philopatry in seabirds and other birds in general (Clarke et al. 1997). Sex-specific patterns of divergence are expected, and indeed mito-nuclear discordance has been detected in seabirds, with more genetic structure in mt than in nuDNA (Burg & Croxall 2001, Deane 2013, Gangloff et al. 2013, Silva et al. 2015, Welch et al. 2011, Friesen et al. 2006, Welch Fleicher 2012, Genovart 2012, Sonsthagen et al. 2016; but see Pons et al. 2014). Such discordance was suggested to result from incomplete lineage sorting in nuDNA due to a higher effective sample size than mtDNA (McKay & Zink 2010), but other mechanisms were proposed such as adaptive introgression of mtDNA, demographic disparities and sex-biased asymmetries (review in Toews & Brelsford 2012). Disentangling these latter processes can be difficult (McKay & Zink 2010), thus a combined use of mitochondrial and biparentally-inherited nuclear markers, as well as coalescent theory-based methods using the lineage species concept coupled with population genetics within a phylogeographic context, are advocated (Hudson & Coyne 2002).

Shearwaters (order Procellariiformes) are long-lived birds showing slow demographic rates (Warham 1996). They breed in large colonies on remote oceanic islands, are pelagic (González-Solís et al. 2007) and highly philopatric (Brooke 2004). Geographic structuring of populations can be strong (Genovart et al. 2012, Gómez-Díaz et al. 2009), and the systematics of some taxa is still highly controversial, especially for the little and Audubon's Shearwaters Puffinus assimilis-lherminieri complex (Austin et al. 2004). This

widespread small-sized, black-and-white shearwater breeds from equatorial to sub-arctic seas (see map in Fig. 1). Previous studies, based only on the mitochondrial gene *cytb*, indicated that population genetic clusters matched with geographic distribution (Austin et al. 2004, Kawakami et al. 2018). Using this single marker, three distinct lineages were recognized in the North Atlantic: lherminieri in the Caribbean and off Brazil, baroli in the Azores, Canaries and Madeira, and boydi in Cape Verde (Fig. 1). These taxa are characterised by non-overlapping breeding and non-breeding distributions at sea (Ramos et al. in review), and are thus geographically and genetically (for at least one marker) separated. Still, they are morphologically and ecologically highly similar (Precheur et al. 2016, Calabrese et al. in review), a pattern typical of the first stages of the speciation process, the so-called grey-zone (De Queiroz 2007). Unsurprisingly, their taxonomic ranking has been hotly debated, e.g. baroli being considered as belonging to assimilis (Shirihai et al. 1995), lherminieri (Austin et al. 2004) or a species of its own (Sangster et al. 2005). Lineages belonging to this complex are also found in the Indian and Pacific Ocean, with populations breeding in the Seychelles (nicolae), Réunion (bailloni) and many islands in the Pacific Ocean (dichrous). Breeding populations of bailloni are characterised by different breeding phenologies on the northern and southern parts of Réunion (Bretagnolle & Attie 1996), potentially impacting genetic structuration (Friesen et al. 2007b). Indian Ocean birds were alternatively considered as a P. lherminieri subspecies (Warham 1990) or subspecies of a bailloni pan-tropical taxon (Austin et al. 2004). The exact taxonomic status of these six lineages is thus largely unresolved.

Here we consider these six lineages, covering Atlantic and the Indo-Pacific branches of the Puffinus assimilis-lhermineri complex. We sampled birds from all but one known breeding localities in the North Atlantic (four Caribbean breeding sites for lherminieri, two main breeding sites in Cape Verde for boydi, one breeding site in Azores, Madeira, Selvagens Canary Islands for baroli) and in the northern Indian Ocean (northern and southern populations of bailloni in Réunion and Seychelles for nicolae; Fig. 1). Only a single breeding locality was unsampled, the Fernando de Noronha archipelago, off Brazil (Fig. 1). We analysed three mitochondrial and six nuclear markers to delineate genetic units and investigated patterns of genetic differentiation and divergence among populations. As conflicting geographic patterns between mitochondrial and nuclear markers (hereon referred to as mito-nuclear discordance; Toews & Brelsford 2012) was systematically found in petrels so far (references above), we inferred population structure among females and males independently to test for sex-biased dispersal on nuDNA. We then used multispecies coalescent inference and an ABC framework to investigate evolutionary scenarios of breeding site colonisation over the last million years, trying to disentangle contrasted processes shaping evolutionary history and the contemporary population structure of this complex, such as landmass presence, isolation-by-distance, oceanographic conditions (sea surface temperature) and climatic oscillations.

Material and Methods

2.1 Sampling, extraction and PCR amplification of gDNA

A total of 276 birds (all adults, i.e. having already dispersed) spanning the entire known breeding distribution of *Puffinus lheminieri* and four populations of interest of *P. bailloni* were selected (Supplementary Material 1). Four individuals from the Pacific Ocean (taxon *dichrous*, by far the most widespread lineage in the Pacific (Onley & Scofield 2007); Fig. 1) were also used. Individuals were sexed using PCR amplification (with the 2250F and 2781R primers (Fridolfsson & Ellegren 1999)). The overall sex ratio was unbiased (120 females, 107 males; Pearson's Chi² with Yates' continuity correction, p=0.42; 49 individuals could not be sexed successfully, see Supplementary Material 1).

Total genomic DNA was extracted from blood samples (except for the population of the Bahamas, for which samples were derived from toepads collected on dead birds) using NucleoSpin® Tissue XS Kit (Macherey & Nagel, Düren, Germany). Samples were incubated overnight in 4 mg of Proteinase K. Purified genomic DNA was eluted twice in 50 µL of TE buffer pre-heated at 70°C. DNA concentration was measured using Nanodrop spectrophotometry. Three mitochondrial markers (*cox1*, *cytb*, and the mitochondrial Control Region, CR) and six nuclear markers (Beta-fibrinogen exon6 through 8, *βfib*; Cold Shock Domain-containing E1 intron5, *csde*; Interferon Regulatory Factor 2 intron2, *irf2*; PAX interacting protein 1 intron20, *pax*; Recombination activating protein 1, *rag1*; and Tropomyosin 1 alpha exon7, *tpm*) were targeted (primer sequences, PCR profile and conditions: Supplementary Material 2). These markers were previously shown to be polymorphic within and between petrel species (Gangloff et al. 2013, Silva et al. 2011).

2.2 Quality control of genetic data

While checking chromatograms we found several cases of double peaks on the sequences of *cox1*, *cytb* and CR (Supplementary Material 3). Bird blood contains relatively few mitochondria, and it is therefore likely to amplify nuclear copies of mitochondrial markers, or "numts" (Sorenson & Quinn 1998a). Such nuclear copies may diverge from the original mitochondrial genes since they are non-coding, which result in double-peaks on the chromatograms. To check this scenario and avoid such copies, we digested nuclear DNA with the ExonucleaseV (ExoV; NEB-M0345S) and sequenced again the mitochondrial markers for all individuals showing double-peaks for *cox1*, plus 5 individuals showing no double-peaks randomly chosen, using a protocol modified from (Jayaprakash et al. 2015, see Supplementary Material 3). Before running analyses, we checked that coding sequences contained no stop codons or indels. Some analyses required phased data (e.g. *BEAST analysis), so the gametic phase of nuclear markers was determined probabilistically using PHASE 2.1 (Stephens et al. 2001) implemented in DNAsp v.5.10.01 (Librado & Rozas 2009). Additional Genbank sequences (Supplementary Material 1) were aligned to our sequences using MAFFT v 7.187 (Katoh et al. 2002).

2.3 Population diversity, differentiation, and divergence

Each population was described by calculating haplotype frequencies, inter-haplotype distances, haplotype diversity and nucleotide diversity (π), as well as expected and observed heterozygosity for nuclear markers using DNAsp v.5.10.01 (Librado & Rozas

2009) and Genetix v4.05.2 (Belkhir et al. 2004). We evaluated signals for departures from neutrality or demographic changes by estimating Tajima's D (Tajima 1989) and Fu's Fs (Fu 1997) for each locus, with Arlequin v.3.1 (Excoffier et al. 2005). Differentiation among populations was estimated by performing AMOVAs, and calculating pairwise F_{ST} and Φ_{ST} and the population average pairwise differences D_{XY}, using Arlequin. For AMOVAs, samples were stratified into five groups, corresponding to the five nominal lineages (*lherminieri*, *boydi* and *baroli* in the Atlantic, *nicolae* and *bailloni* in the Indian Ocean), and populations (i.e. sampling localities; Fig. 1) within these groups. The matrix of genetic distances among all pairs of haplotypes was computed using the K2P model of substitution for concatenated mitochondrial markers, and TN93 for concatenated nuclear markers, as determined using jModelTest2. We used a Mantel test to measure the level of correlation among genetic distances and geographic distances (Smouse et al. 1986). Geographic distance was calculated as the shortest distance between two populations without crossing land. Statistical significance (AMOVAs, Pairwise F_{ST} and Mantel tests) was estimated using 1000 permutations. To visualize relationships among lineages, we inferred NeighborNet networks using SplitsTree v 4.14.2 (Huson & Bryant 2006), with different dataset combinations: all markers independently, concatenated mitochondrial markers, concatenated nuclear markers.

2.4 Estimation of sex-biased dispersal using nuclear markers

To detect sex-biased dispersal, we separated sequences from females and males into two separate datasets, excluding three populations (Saint-Barthélémy, Funchal and Selvagem) represented by fewer than five individuals from each sex. Sex-biased dispersal was tested at both intra- and inter-lineage scales, expecting females to be less structured than males since males are presumably more philopatric. We calculated average pairwise relatedness for each sex, within each population, using the triadic likelihood estimator (Wang 2007) implemented in Coancestry (Wang 2011). To test if the difference of mean relatedness between males and females of each population was significant, we used the test of difference between sex by bootstrapping samples 1000 times and recalculating difference in means between sex for each bootstrap. Observed and simulated differences was then compared, and if the observed difference fell outside of the 95% confidence interval, we considered it to be significant.

If females disperse more than males, females sampled from a single population will be a mixture of residents and immigrants. The female sample will therefore deviate from the Hardy-Weinberg equilibrium and show a deficit of heterozygotes (Wahlund effect). F_{IS} calculated for the female sample is thus excepted to be larger than the male F_{IS} (Goudet et al. 2002). We estimated F_{IS} separately for females and males for all tested populations, and evaluated its significance using 1000 permutations with Genetix. Conversely, we expect F_{ST} (Goudet et al. 2002) to be higher in philopatric males than in females. We calculated F_{ST} for each pair of populations within the two datasets, with Arlequin.

2.5 Phylogeographic scenarios

Reciprocal monophyly was inferred on all markers as partition by gene trees, which were used to evaluate the degree of divergence among lineages using MrBayes v. 3.2.6 (Ronquist et al. 2012; Supplementary Material 4). To reconstruct the scenario of divergence among the different populations, we used species trees, inferred using two different methods. We first ran an analysis with *BEAST v 2.2.0 (Bouckaert et al. 2014) using the three mitochondrial markers. We choose to link time-trees for the three mitochondrial markers, since they are on the same plasmid, where no recombination is expected at the time scale considered in this study. However, the three markers have different composition and different evolutionary rates, so we did not link the molecular clock and evolution models. We tested the hypothesis of molecular clock with the Clock Test using ML implemented in MEGA v7.0.20 (Kumar et al. 2016). This hypothesis was rejected (p-value < 0.0005). We therefore used, for each marker, an uncorrelated lognormal relaxed clock model. The clock rate was fixed to 0.0772 ± 0.0076 substitution per site per million year for *cox1* (rates inferred for Procellariiformes by Pereira & Baker 2006) and 0.0945 ± 0.0175 for *cytb* (rates inferred for Procellariiformes by Weir & Schluter 2008). The rate for the control region was estimated by the model as no rate is published for Procellariiformes. Existing rates for other birds could not be used either because they belong to groups that are too distant (e.g. Moas (Baker et al. 2005) or Peafowls (Kimball et al. 1997)), and the control region is highly variable among groups (Ruokonen & Kvist 2002). Published rates for the CR were however set in *BEAST as priors and did not affect the results (data not shown). We used for each marker a model consistent with the result of jModelTest2, and a Yule process species tree prior with a continuous population size model. As for the MrBayes analysis, we ran each MCMC chain with $50*10^6$ generations, sampled every 1,000 generations, the first 25% of generations were discarded as burn-in and we inspected the stationarity of the chains using Tracer. To test whether the colonization of the Atlantic could result from birds of the Pacific passing through the Panama Isthmus, individuals from the central Pacific (Marquises archipelago, taxon *dichrous*) were added to the *BEAST inference. This taxon can be considered as the best representative taxon for the central Pacific, as it is the most widespread and numerous, since *polynesiae* is considered synonym to *dichrous* (Austin et al. 2004), while *bannermani* is best recognised as a species on its own (Kawakami et al. 2018). The taxon *gunax* (from Vanuatu) has never been sequenced, and actually the location of its breeding colonies is unknown. Three *dichrous cytb* sequences retrieved from Genbank (AY219949-AY219951) and three individuals from our own collection were used. We ran a second *BEAST analysis by adding all the nuclear markers independently to the three mitochondrial markers, using the same MCMC parameters. Clock rates priors were set to 0.019 substitutions per site per million year for *βfib* and 0.013 substitution per site per million year for *rag1*, since these rates were estimated for birds (Groth & Barrowclough 1999; Prychitko & Moore 1997). The clock rate for nuclear markers was estimated by the model as no rate has been produced for petrels. In this latter case, we kept only the individuals for which we had sequences for all markers. To investigate the demographic history of lineages, we estimated population size through time by estimating Extended Bayesian

Skyline plot as implemented in *BEAST, for each genetic unit previously delimited, considering a generation time of 15 years (Precheur et al. 2016).

We also used a coalescent-based ABC approach to explore the best demographic scenario describing the dataset of the combined mitochondrial and nuclear markers using the program DIYABC v. 2.1.0 (Cornuet et al. 2014). ABC methods consist in the simulation of datasets similar to the real dataset in terms of population and marker sizes. First, in the Indian Ocean, we tested if the three populations emerged simultaneously in a radiation event or in two disjoint events by comparing the posterior probability of these three scenarios, and in the case of the latter, which population was basal to the two others and which one of the two remaining populations was basal to the other. This hierarchical strategy was applied to each lineage independently, then to the three Atlantic lineages, and finally considering the five lineages together (see Supplementary Material 5 for a description of all tested scenarii). For each possible scenario, 10^6 pseudo-observed datasets were simulated, with the same ploidy and number of loci per population as observed in the real dataset. We fixed uniform priors for population sizes, times of size variation and divergence and mutation and admixture rates priors (see Supplementary Material 4 for details), from which we simulated the datasets. Summary statistics were calculated from the simulated datasets and compared to the same statistics obtained from the real dataset. The Euclidean distance was calculated between the statistics obtained for each normalized simulated dataset and those for the observed dataset (Beaumont et al. 2002). Posterior probability of each scenario was then calculated using a logistic regression on summary statistics produced by the 1% of the simulated datasets closest to the real dataset. To reduce the dimensionality of the data, a linear discriminant analysis was preliminarily applied to the summary statistics (Estoup et al. 2012). The scenario with the highest posterior probability value with a non-overlapping 95% confidence interval (95% CI) was selected.

Results

3.1 Patterns of genetic diversity, numts and the presence of a duplicated region

We obtained an average of 192 sequences per marker (length range 307-1323 bp; Table 1). Mitochondrial data produced 148 polymorphic sites yielding 150 haplotypes, while nuclear data exhibited a total of 111 variable sites and 150 alleles (see Table 1 for summary of polymorphic sites, haplotypes and diversities per marker). Mitochondrial markers were twice as variable as nuclear markers, though nuclear haplotype and nucleotide diversity in *βfib* reached a level similar to mitochondrial haplotype diversity (Table 1).

None of the coding markers (*cox1*, *cytb* and *pax*) presented any insertion, deletion, nonsense or stop-codon following translation (see Supplementary Material 1). Double peaks on Sanger chromatograms were however detected for each of the three mtDNA markers. While all double peaks at *cox1* were removed by the exonuclease treatment, 60 CR sequences (33%) still showed double-peaks at 73 positions, as well as 37 positions for 22 individuals (10%) for *cytb*. Double-peaks were not specific to any population or sex, and were not linked to the position of the individuals in the sequencing plate (see Supplementary Material 3). Only 12 (5%) individuals showed double-peaks both at CR and *cytb*, so the presence of double peaks seemed unlinked between the two markers.

Replicating DNA extractions, PCR and sequencing confirmed these results, making laboratory contamination unlikely. Contamination in the field was also unlikely since new sampling supplies were used for every sample. Given that only 10% of the *cytb* sequences presented such ambiguities (which may be due to heteroplasmy, Torres et al 2018), we removed such sequences for further analyses expecting little impact on the analyses. However, for the CR, since a third of the total sequences were involved, we kept all CR data in further analyses, considering two haplotypic phases, which correspond to the two copies of the duplicated CR, for MrBayes and *BEAST analyses (although this is a violation of the assumption that mixed sequences to be phased are under Hardy-Weinberg equilibrium; a robustess analysis showed that there is no significant impact to use these phased data comparatively to remove the ambiguous CR sequences see Supplementary Material 6).

3.2 Population structure and sex-biased dispersal

Mitochondrial and nuclear results from Fu's Fs indicated no deviation from neutrality (Supplementary Material 7). However, for Tajima's D tests, three localities displayed significant negative Tajima's D for all mitochondrial and nuclear markers: South Reunion (Indian Ocean), Saint-Barthélemy (W Atlantic) and Raso (E Atlantic). In addition, three localities presented significant negative Tajima's D at mitochondrial loci only (Selvagem and Funchal, E Atlantic; South Reunion) and two at nuclear loci only (Vila, E Atlantic; North Reunion). Patterns of population structure at seven out of 13 localities might therefore be influenced by selection and/or recent demographic changes, in addition to neutral processes.

Gene trees and phylogenetic inference with *BEAST and MrBayes revealed a hierarchical structure composed of two well-supported (posterior probabilities PP $\geq$ 0.95) reciprocally monophyletic clades corresponding to the two oceans, within which individuals from the five lineages further clustered into monophyletic sub-clades (Fig. 2b,c and S7). All except one of these sub-clades (E Atlantic *boydi*) were supported in *BEAST (PP $\geq$ 0.95) using all concatenated markers. For both mtDNA markers, and all markers concatenated, the central-Pacific *dichrous* lineage was nested within the *bailloni*/*nicolae* clade, although node supports for *dichrous* position within Indian Ocean clade were weak (between 0.27 and 0.73). Assignment to an ocean basin based on nuclear haplotype networks was however discordant from the mitochondrial data for 33 individuals, for at least one nuclear locus (Fig. 3b and S8): 15 Atlantic individuals fell closely to the Indian phylogroup, and 18 Indian Ocean individuals clustered within the Atlantic group. All of these 33 individuals showed the mitochondrial signature expected based on their geographical sampling location. Interestingly, a *baroli* individual showed one haplotypic phase clustering with the *baroli* phylogroup (mother) while the other haplotypic phase (father) clustered with the *nicolae* phylogroup for four nuclear markers (the two remaining could not be assigned to any particular lineages). This individual might be the result of hybridization, although further analyses based on additional markers would be necessary to detect more robustly hybridization among lineages. The ambiguous assignment of the other individuals might be due to introgression or incomplete lineage sorting (see below).

In parallel, we used an AMOVA framework with the five nominal lineages now defined a priori, to examine how genetic variants partitioned among and within these taxonomic units. Most of the genetic variance was due to inter-lineage differentiation (88.5% and 58.4% for mitochondrial and nuclear markers, respectively). The variance among sampling localities within lineages accounted for 0.5 to 4.1%, while variance within sampling localities represented 11.0 to 37.5%. Pairwise F_{ST} showed consistently higher values among, than within lineages for both marker types, with mostly non-significant values within each lineage (Table 2a). Indeed, 24 nuclear F_{ST} values were found non-significant versus 10 mitochondrial Φ_{ST} values (Table 2a). Population average pairwise differences led to similar results, with high structuration for the five nominal lineages (Supplementary Material 7). Genetic distance increased clearly with geographic distance (Fig. 4a), but Mantel tests were performed only between pairs within the same Ocean (given that each Ocean taxon is likely different species). Tests confirmed that genetic and geographic distances were strongly correlated to each other, both for mtDNA and nuclear markers when analyzing pairs of populations within an Ocean basin (r=0.88 and 0.70, n=45, p <0.005 for mtDNA and nuDNA, respectively, Fig. 4a,b). Between breeding sites and within lineages, isolation by distance could not be reliably tested as the number of populations was too low, but visually it seemed that there was no relation between geographic and genetic distances (Fig. 4).

Indian Ocean populations (*nicolae* and *bailloni*) showed stronger female dispersion as indicated by significantly stronger deficit of heterozygotes and a significantly lower average relatedness in females (Table 3). Conversely, in *baroli*, F_{IS} was significantly higher for males and they were less related to each other than females, suggestive of male-biased dispersal. Finally, an ambiguous pattern was found for both *lherminieri* (males had stronger deficit of heterozygotes and were significantly more related to each other than females) and *boydi* (female F_{IS} was significantly higher than male F_{IS}, though females were more related to each other than for males at one sampling locality). In addition, population structure within lineages, as measured with F_{ST}, was similar between sexes, but between oceans a larger range of F_{ST} values was observed for males with higher maximum values, suggesting that males were more structured at least for some pairs of populations (e.g. *lherminieri* vs. Indian Ocean lineages; Table 2b). Overall males seemed more structured than females between oceans, suggesting that females disperse farther, but genetic signal for sex-biased dispersal varied geographically: female-biased in the Indian Ocean, male-biased or inconclusive in the Atlantic Ocean.

3.3 Reconstructing scenario of breeding site colonization

A split between basal Atlantic and Indian populations (Fig. 2a) occurring 1.06 My ago (95% CI range 0.5 -1.70) was inferred based on mitochondrial and nuclear species trees. We choose to keep the estimation based on all markers since it seemed less influenced by the ambiguous sequences described above (see robustness analysis in Table S3). West and east Atlantic basal populations split around 0.43 My ago (0.20-0.73), *baroli* and *boydi* split at 0.32 My ago (0.12-0.53), and *nicolae* and *bailloni* at 0.22 My ago (0.08-0.38). The inferred species tree based on all nine markers showed the same topology

though with generally lower divergence times and higher confidence intervals (Supplementary Material 8). ABC analyses also supported a similar scenario of basal population divergence: best retained topologies using mtDNA markers and all nine markers suggested, starting from oldest to newest splits, *nicolae* being basal, leading to the appearance of *boydi* (Fig 2d, Supplementary Material 5). Then *lherminieri* diverged from *boydi*, and *baroli* diverged from *boydi*. Finally, *bailloni* diverged from *nicolae* (Fig. 2d Supplementary Material 4). Our phylogenetic trees placed the Central Pacific taxon *dichrous* within the Indian clade, thus supporting the putative scenario of Atlantic lineages diversifying from Indian Ocean rather than from Pacific ancestors (Fig. 2b,c). Within *lherminieri*, ABC analyses suggested a northerly stepping stone colonization process, from Martinique to the Bahamas (Supplementary Material 5). Similarly, the most likely scenario of population divergence within *baroli* was colonization from the Canaries to the more northerly Azores.

Finally, Bayesian Skyline analyses inferred current effective population sizes to be around 10^4 individuals (see Supplementary Material 10). Mean Ne seemed to increase slowly over time, but high confidence intervals precluded detection of any brutal change in population sizes. Confidence intervals on current population sizes were also very large, in particular for the individuals breeding on Réunion (see Supplementary Material 10). The demographic parameter estimations were consistent with a constant population size over time for each lineage, as suggested from Fu's Fs results (which are more suitable than Tajima's D to estimate population expansion; Ramirez-Soriano et al. 2008; Supplementary Material 7).

Discussion

4.1 Mito-nuclear discordance and sex-biased dispersal

At the inter-lineage scale, we observed more genetic structure at mitochondrial than at nuclear loci. This dissimilarity has been observed for numerous Procellariiformes species (e.g. Gangloff et al. 2013, Silva et al. 2015, Welch et al. 2011) as well as other organisms (see Toews & Brelsford 2012 for a review). We suspect that incomplete lineage sorting and retention of ancestral polymorphisms at nuclear loci also contribute. Indeed, effective population size of mitochondrial DNA is four times smaller than that of nuclear DNA due to uniparental inheritance. Lineage sorting will therefore be faster in mtDNA than in nuDNA, being inversely proportional to the effective population size (Funk & Omland 2003). Incomplete lineage sorting is actually thought to be the main cause of mito-nuclear discordance when associated to a pattern of loss of geographic differentiation on nuclear markers (McKay & Zink 2010; Toews & Brelsford 2012). We also found patterns suggestive of introgression in nuclear markers. Hybridisation with introgression has been documented in shearwaters (Genovart et al. 2007, Gómez-Díaz et al. 2009), other Procellariiformes (Brown et al. 2010,) and other seabirds (Gay et al. 2009; Morris-Pocock et al. 2011; Pons et al. 2014). The likelihood of Indian Ocean petrels visiting breeding Atlantic Petrels may be supported by recent tracking of *Pterodroma arminjoniana* breeding on Round Island (Mauritius), which showed that some individuals foraged around South Trinidad Is, off Brazil, and even in the northern Atlantic (Booth Jones et al. 2017),

although flight capacities of *Pterodroma* are far higher than *Puffinus*. Introgression can also blur phylogeographic signals by mixing alleles from distinct populations, and is considered as the second main cause of mito-nuclear discordance (McKay & Zink 2010). Incomplete lineage sorting and introgression are however difficult to distinguish, and additional unlinked markers would be required to disentangle these phenomena. Finally, as the mitochondrial markers represent only the female evolutionary history, sex-biased dispersal favoring females may alternatively explain why the population structure inferred based on nuclear markers conflicts with female-inherited mtDNA markers (see Petit & Excoffier (2009). These authors suggested that the markers associated with the most dispersing sex should better delimitate species, as they will show stronger intra-specific gene flow from colonizing lineages, reducing the effects of genetic drift and lowering the probability of fixating introgressed alleles. Dispersal was indeed stronger in females in some populations, particularly in the larger and the putatively basal lineage, *nicolae*. Sex-biased dispersal was however more uncertain for *lherminieri* and *boydi*, while for *baroli* dispersal was inferred to be male-biased. The sample size for *baroli* was theoretically large enough to robustly detect a bias in F_{ST}, F_{Is}, and Relatedness (Goudet et al. 2002). Sex-biased dispersal may therefore have further contributed to the observed mito-nuclear discordance, at least in some lineages.

4.2 Sequencing artifacts due to mtDNA duplication and uncertainties about molecular clock rates

The mitochondrial genome of several Procellariiformes presents tandem repeats (Abbott et al. 2005; Lounsberry et al. 2015), including *P. lherminieri* (Torres et al. 2018). In this taxon, a duplicated region comprising two copies of the CR was found, yielding the presence of double peaks on chromatograms. By treating gDNA with an exonuclease (which effectively removed all linear DNA), we did not observe triple- or quadruple-peaks on chromatograms, and thus hypothesized that only two copies of the CR sequences were amplified by PCR. Therefore we considered two haplotypic phases for all CR sequences, assuming that removing all CR sequences would have led to a loss of information and of statistical power in the analyses. To check the robustness of this approach however, we replicated all our analyses using two other data subsets: one in which all ambiguous CR sequences were removed, and another one from which all CR sequences were removed. Removing all CR sequences led to a strong loss of information, an increase of the estimated differentiation and a decrease of the estimated divergence times (Supplementary Material 6). Removing only the ambiguous sequences led to estimations close to the estimations of the complete dataset. This suggests that the noise caused by the multiple copies of CR was swamped by the signal contained in that marker and so our analyses using all the individuals are not significantly biased by these sequences.

Another major issue concerns the choice of a molecular clock rate for dating the splitting events. We found a clear hierarchical structuration of populations in the Atlantic and Indo-Pacific Oceans, which diverged around 1.28 to 1.43 My ago either using all markers or only mtDNA markers, respectively. In contrast, using the same taxa (but fewer specimens and only *cytb*), Austin et al. (2004) dated this split at 3.2-3.8 My, and rather

suggested that the closure of the Isthmus of Panama erected a barrier to gene flow between Indo-Pacific and Atlantic populations. This difference in divergence times might be due to taxon sampling, gene sampling, or the calibration of the molecular clock (1.89%/My here, vs. 0.9%/My in Austin et al). This latter parameter is probably a major contributor to the difference observed between our studies, since using the same value as Austin et al, and only sequences from *cytb*, we found a divergence time of 2.2-3.8 My. Many substitution rates for both Mt DNA markers were proposed for petrels, ranging from 0.19%/My (Pacheco et al. 2011) to 1.544%/My (Pereira & Baker 2006) for *cox1*, and from 0.18%/My (Pacheco et al. 2011) to 0.88-0.92%/My (Nunn & Stanley 1998; the latter rate was used by Austin et al), 1.022%/My (Pereira & Baker 2006) and 1.89%/My (Weir & Schluter 2008) for *cytb*. Using lowest rates (Pacheco et al. 2011) resulted in a divergence time between Atlantic and Indo-Pacific lineages around 12.8 My ago (Supplementary Material 10), an unlikely value given the presence of an ocean between the Americas before the Isthmus of Panama erected, despite the fact that *Puffinus* is dating back to the Oligocene (Henderson & Gill 2010). Here we used the substitution rates that were estimated with the highest number of calibration points, i.e. 90 for *cox1* (Pereira & Baker 2006) and 3 for *cytb* (Weir & Schluter 2008). There are two further arguments against Austin et al. (2004) scenario: first, the closure of the Isthmus of Panama has been actually dated at 2.8 My ago (O'Dea et al. 2016), thus later than Austin et al's scenario. Second, we found that the pacific taxon, *dichrous*, was not basal to the Atlantic taxon, but was embedded within Indian ocean taxa. The use of precise fossil calibration would bring a more robust estimation of divergence times and so a supplemental evidence of the origin of the colonization.

4.3 Inferring key drivers of diversification in the small Puffinus

Based on our inferred date of the Indian –Atlantic lineage split, gene exchange between Indian and Atlantic birds have apparently occurred after the closure of the Isthmus of Panama: we suggest that this happened through individuals passing off South Africa, since African continent is an insurmountable barrier for Procellariformes (Silva et al. 2015). Indeed, these shearwaters are tropical or subtropical species (at least currently). Off South Africa, until 1 My ago, sea surface temperatures (SST) were approximately 2°C higher than today (Bell et al. 2015), suggesting that migration between Atlantic and Indian oceans may have remained possible for such birds. From 2.0, a strong decrease of SST occurred in both oceans (Bell et al. 2015), and gene flow between Indian and Atlantic may thus have ceased, in agreement with our estimated time of divergence (1.06 My ago). Once the Atlantic birds were isolated from Indo-Pacific populations, differentiation started to occur among Atlantic lineages 0.43 My ago (respectively for the 9 markers or only the mtDNA markers). This period also corresponds to a further decrease of the SST in the Atlantic (Bell et al. 2015), a southward shift of the subtropical front and warmer waters in the Southern Ocean (Maiorano et al. 2009), and sea ice development in the North extending southwards from the Arctic to the current Great lakes in the USA (Webb & Bartlein 1992). Cold temperatures, preventing the colonization of potential northern breeding sites such as the Azores, may have forced shearwaters to spread over the two sides of the Atlantic (Fig. 2d). The third step in the colonization of Atlantic breeding grounds concerns the

divergence of *baroli* from *boydi*, which would have occurred around 0.32 My ago, a period that corresponds to a stabilization of the SST at the current level in the North Atlantic. The ice melt may then have allowed northward colonization on both sides of the Atlantic, from Cape Verde to the Canaries, Madeira and the Azores, and from the lesser Antilles to the Bahamas and Bermuda. Similar timing of divergence has been suggested between *Calonectris edwardsii* (Cape Verde) and *C. diomedea* (North Atlantic and Mediterranean) 0.7-0.9 My ago (Gómez-Díaz et al. 2006) and *Puffinus olsoni* (Canaries) and *Puffinus puffinus* (North Atlantic) 0.2-1.0 My (Ramirez et al. 2010). In the Indian Ocean, a reversed pattern of southward colonization occurred, shearwaters colonizing from Seychelles, partly continental in origin, to Réunion 0.22 My ago, precisely at the time strong volcanic activity ended on Réunion (Gillot & Nativel 1989). Mauritius was probably colonized long before (Mauritius age: 8My; McDougall & Chamalaun 1969), then Rodrigues and Réunion (both about the same age, 2 My; McDougall 1971), but shearwaters are now extinct on Mauritius and Rodrigues, following human colonization since the 17th century, so no sample is available. The southward movement was however not necessarily related to changes in SST, but rather to availability of volcanic islands that eventually emerged in a southward direction.

Mantel tests showed a strong correlation between geographic and genetic distance between lineages, i.e. at large scale. Indeed Little shearwaters are poor flyers compared to other Procellariiformes. However, isolation-by-distance within each lineage was not detected, therefore distance alone could not be a factor of population divergence at this smaller scale; we argue that sea temperature could rather be the most important factor of divergence in our case at this scale. Seabirds indeed depend on both sea and the islands where they breed, thus sea temperature has strong impacts on seabird phenology, breeding, survival and abundance (see Sydeman et al. 2012 for a review). We suggest that foraging ecology, which strongly depends on SST, is an important process shaping divergence among lineages. The segregation of foraging areas among populations is an important factor of differentiation among seabirds (Friesen 2015; Friesen et al. 2007a), as shown for other shearwaters (Genovart et al. 2007; Gómez-Díaz et al. 2006), petrels (Gangloff et al. 2013; Welch et al. 2012a), storm-petrels (Deane 2013; Smith et al. 2007) and albatrosses (Alderman et al. 2005; Burg & Croxall 2001). Assessment of non-breeding and breeding distributions at sea of the little shearwaters complex revealed that all three Atlantic taxa show rather separated foraging and wintering areas (Ramos et al. in review), and further suggest that *boydi* rather than *lherminieri* was basal in the North Atlantic. Indeed, *boydi* is more flexible in its foraging ecological niche, suggesting basal behavior, with a far larger potential distribution at sea covering all central North Atlantic (Zajková et al. 2017, see map in Ramos et al. in review). In addition, over the last My, SST oscillations gradually increased in amplitude with up to five degree difference in range (Bell et al. 2015; Herbert et al. 2011). These oscillations were showed to correlate with changes in marine productivity (Martı et al. 2009), prey species diversity (Yasuhara & Cronin 2008), atmospheric circulation (Chang et al. 2000) and sea level (e.g. in Atlantic Nascimento et al. 2011, Zazo et al. 2010). It is likely that these oscillations also contributed to divergence in the North Atlantic, and we suggest that significant Tajima's D tests found in almost half of

the populations studied represent traces of past bottlenecks and population expansions as it has already been shown in other taxa (Ramakrishnan et al. 2005; Weber et al. 2004; Zhu et al. 2006).

A last interesting side effect of our study is that it suggests that the small black and white shearwaters have shown a very recent radiation speciation event, with not less than 13 species radiating in just 1.46 million years since *P. puffinus*, *P. assimilis* and *P. newelli* clades are either embedded in *lherminieri* or *bailloni* clades. All these species are rather coastal shearwaters (compared to more pelagic species such as the larger shearwaters), and such high rate or speciation may be the result of the high climatic oscillation that occurred over the last 2 million years which may have favored high rates of colonization and extinction on coastal islands.

4.4 Sea surface temperature as a major diversification driver in marine organisms?

Conversely to pelagic fishes that show no or low inter-ocean structure (Díaz-Jaimes et al. 2010, Ely et al. 2005) and no intra-ocean structure (Nomura et al. 2014, Taguchi et al. 2015), sea mammals' gene flow is shaped by sea temperature which drives structuration between ocean basins and among breeding areas (Alexander et al. 2016, Jackson et al. 2014, Richard et al. 2018, Fontaine et al. 2014, Viricel & Rosel 2014). SST plays also a role in the diversification of sea turtles, since only cold-adapted species are able to exchange genes among oceans (Dutton et al. 1999). Finally, organisms with a pelagic larval phase show globally low structuration (Kelly & Palumbi 2010) but when detected, structuration is often linked to sea temperature (Benestan et al. 2015; Teske et al. 2005, 2018). Therefore, SST appears as a generic driver of diversification in marine organisms, though their patterns of structuration are generally considerably weaker (see Bowen et al. 2016 for a review) than those found here. We suggest that the reason for this discrepancy lies in the fact that these seabirds are central place foragers, i.e. they still depend on terrestrial habitats for breeding, the latter being impacted for instance by glaciations. They are therefore highly sensitive to any latitudinal change of SST in comparison to island distribution, which acts as a constraint since an optimal area in regard to SST may lack any island for breeding. Indeed, in marine organisms with low dispersal abilities, patterns of structuration and divergence are more similar to the patterns found here on shearwaters. For instance, timing of population divergence between Atlantic and Indian Ocean lineages and within the Atlantic in the seahorse *Hippocampus* 'kuda complex' (Floeter et al. 2007) fits to our estimates. This species has no planktonic larval duration (Lourie et al. 2005) and long dispersal events are considered rare and implying a few individuals (Teske et al. 2005). Moreover, Cape Agulhas is known to be a phylogeographic break among several costal species, due to the difference of currents and sea temperatures between the two oceans (review in Teske et al. 2011). On the other extreme, seabirds do not compare either to terrestrial organisms living on islands, despite being highly philopatric, simply because they can disperse easily, if an island is available to being colonised.

In this small shearwater complex, geographical barriers and/or isolation by distance may have been a major driver of differentiation at large scale (typically, between Oceans)

while SST has been a more important driver at smaller scale (within Oceans), with shearwaters shifting their breeding latitudes with a changing SST. However, since these seabirds depend on the geographical distribution of their breeding islands and because these seabirds are not the best seabird fliers, this distribution becomes a major constraint resulting in the present geographical structure, promoting local adaptation to small scale ecological constraints and reducing gene flow. Therefore, petrels and shearwaters present an interesting case study where diversification processes rely more (or at least equally) on ecological factors, in particular sea surface temperature, rather than distance or continental barriers, in contrast to either "true" marine organisms or terrestrial organisms. Strict marine organisms can disperse far more, or alternatively are unconstrained by island distribution, and thus show much less geographical structure within taxa. The terrestrial organisms tend to disperse far less, and isolation-by-distance tends to be a main driver of population differentiation (Meirmans 2012; Vekemans & Hardy 2004). Indeed terrestrial organisms, such as lizards or birds in Macaronesia (AlmalkI et al. 2017; Brehm et al. 2003), geckos in Cape Verde (Arnold et al. 2008) or birds in America (Patel et al. 2011), have revealed strong splits between islands with no shared haplotypes for the same mitochondrial markers.

Table 1: Summary statistics of polymorphisms for the nine markers used in this study.
N is the number of sequences obtained (with numbers in parentheses referring to the sequences obtained from this study, obtained from previous studies and downloaded from Genbank); L is the length of the sequences in bp; S is the number of polymorphic segregating sites; h / a: h is the number of corresponding haplotypes for mitochondrial markers, and a is the number of alleles for nuclear markers; hd is the haplotype diversity and π is the nucleotide diversity. Due to the presence of ambiguities in the sequence for CR, two haplotype phases were considered here.

Marker	**N**	**L**	**S**	**h / a**	**hd**	**π**	**He**	**Ho**
cox1	225 (212, 10, 3)	577	18	12	0.843	0.01301	-	-
cytb	230 (205, 3, 22)	877	50	48	0.951	0.01394	-	-
CR	181	307	80	98	0.993	0.07497	-	-
tpm	184	427	2	3	-	0.00011	0.0041	0.0008
irf2	182	542	6	6	-	0.00048	0.0133	0.000
csde	223	542	15	13	-	0.0006	0.0007	0.0003
pax	255 (227, 28)	515	9	10	-	0.00159	0.0055	0.0008
rag1	162	1323	29	65	-	0.00337	0.0109	0.0026
βfib	92	1067	50	53	-	0.0083	0.0074	0.0014

Table 2: Population differentiation, according to the types of genetic markers and sex.
a. Pairwise Φ_{ST} values for mitochondrial markers (below diagonal) and F_{ST} for nuclear markers (above diagonal).
Border indicates the separation between intra and inter lineage comparisons. Triple band indicates the separation between intra- and inter-ocean comparisons. *: p<0.05; ***: p<0.001

a.		*lherminieri*	*lherminieri*	*lherminieri*	*lherminieri*	*boydi*	*boydi*	*baroli*	*baroli*	*baroli*	*baroli*	*bailloni*	*bailloni*	*nicolae*
		Allencay	**Longcay**	**Martinique**	**St Barthélémy**	**Raso**	**Cima**	**Funchal**	**Mclara**	**Selvagem**	**Vila**	**North Reunion**	**South Reunion**	**Seychelles**
lherminieri	**Allencay**	-	0	0.18*	-0.09	0.47***	0.25***	0.45*	0.5***	0.46***	0.47***	0.69***	0.68***	0.41***
lherminieri	**Longcay**	0.08*	-	0.11	-0.12	0.42***	0.19***	0.45***	0.44***	0.39***	0.41***	0.69***	0.65***	0.37***
lherminieri	**Martinique**	0	0.22***	-	-0.24	0.16*	-0.11	0.38*	0.39***	0.42***	0.43***	0.78***	0.65***	0.61***
lherminieri	**St Barthélémy**	0.19*	0.35***	-0.13	-	-0.28	-0.21	-0.11	-0.07	-0.43	-0.48	0.52***	0.71***	0.39*
boydi	**Raso**	0.81***	0.83***	0.82***	0.77***	-	-0.11	-0.02	0.29***	0.37*	0.35***	0.79***	0.77***	0.52***
boydi	**Cima**	0.76***	0.78***	0.74***	0.71***	0.01	-	-0.22	0.15*	0.15*	0.11*	0.66***	0.7***	0.42***
baroli	**Funchal**	0.81***	0.84***	0.85***	0.72***	0.7***	0.56***	-	-0.04	-0.48	-0.3	0.78***	0.77***	0.51*
baroli	**Mclara**	0.78***	0.81***	0.82***	0.73***	0.67***	0.61***	0.08	-	0.33***	0.06	0.76***	0.75***	0.48***
baroli	**Selvagem**	0.8***	0.83***	0.83***	0.73***	0.64***	0.53***	0.36	0.10	-	0.14*	0.68***	0.69***	0.39*
baroli	**Vila**	0.85***	0.86***	0.84***	0.81***	0.69***	0.65***	0.06	-0.04	0.03	-	0.75***	0.73***	0.47***
bailloni	**North Reunion**	0.93***	0.93***	0.89***	0.91***	0.87***	0.84***	0.89***	0.85***	0.88***	0.89***	-	-0.05	0.05
bailloni	**South Reunion**	0.93***	0.94***	0.9***	0.92***	0.89***	0.87***	0.91***	0.88***	0.88***	0.91***	-0.16	-	-0.13
nicolae	**Seychelles**	0.93***	0.93***	0.9***	0.91***	0.89***	0.87***	0.91***	0.89***	0.9***	0.92***	0.76***	0.79***	-

b. Pairwise F_{ST} for nuclear markers for females (below diagonal) and males (above diagonal).

b.		*lherminieri*	*lherminieri*	*lherminieri*	*boydi*	*boydi*	*baroli*	*baroli*	*bailloni*	*bailloni*	*nicolae*
		Allencay	**Longcay**	**Martinique**	**Raso**	**Cima**	**Mclara**	**Vila**	**North Reunion**	**South Reunion**	**Seychelles**
lherminieri	**Allencay**	-	0	0.33*	0.55***	0.49***	0.5***	0.47***	0.81***	0.83***	0.8**
lherminieri	**Longcay**	0	-	0	0.47***	0.4**	0.42***	0.43***	0.81***	0.83***	0.78***
lherminieri	**Martinique**	0.45**	0.38*	-	0.2*	0.35**	0.42**	0.41***	0.71***	0.77***	0.7**
boydi	**Raso**	0.6***	0.48***	0.24*	-	0	0	0.29*	0.81***	0.84***	0.78**
boydi	**Cima**	0.49***	0.34***	0.19*	0	-	0	0	0.77*	0.82***	0.75*
baroli	**Mclara**	0.67***	0.56***	0.52**	0.34**	0	-	0	0.76***	0.81***	0.74**
baroli	**Vila**	0.71***	0.61***	0.69**	0.44**	0	0	-	0.77***	0.79***	0.7**
bailloni	**North Reunion**	0.8***	0.77***	0.86***	0.9***	0.81***	0.92***	0.94***	-	0	0.6*
bailloni	**South Reunion**	0.69***	0.64***	0.65***	0.75***	0.69***	0.73***	0.77***	0.3*	-	0.53***
nicolae	**Seychelles**	0.78***	0.75***	0.8***	0.85***	0.73***	0.86***	0.89***	0.67***	0.31***	-

Table 3: Sex-specific F_{IS} and Relatedness indices
Mean and confidence interval of F_{IS} and relatedness are indicated for each population with a sufficient sample size and for each lineage The observed difference of mean is considered as significant when not in the confidence interval (indicated in bold and by an asterisk)

Lineage	Population	Female Sample Size	Male Sample Size	F_{IS} Female	F_{IS} Male	Observed Relatedness Female	Observed Relatedness Male	Observed difference of relatedness	95% CI of inferred difference
lherminieri	Allencay	8	10	**0.79 [0.77-0.90]**	**0.35 [0.27-0.43]**	0.45	0.55	-0.10	-0.14 : 0.13
lherminieri	Longcay	8	12	**0.19 [0.13-0.20]**	**0.60 [0.55-0.69]**	0.62	0.71	-0.09	-0.08 : 0.09
lherminieri	Bahamas	16	22	**0.63 [0.59-0.67]**	**0.76 [0.75-0.79]**	0.55	0.62	-0.07	-0.07 : 0.07
lherminieri	Martinique	19	22	0.71 [0.64-0.71]	0.62 [0.59-0.66]	**0.70***	**0.63***	**0.07**	**-0.04 : 0.04**
boydi	Raso	11	7	0.53 [0.49-0.58	0.51 [0.49-0.55]	**0.49***	**0.38***	**0.11**	**-0.11 : 0.10**
boydi	Cima	10	9	**0.77 [0.76-0.78]**	**0.66 [0.64-0.73]**	0.56	0.57	-0.01	-0.08 : 0.09
baroli	Mclara	6	9	**0.36 [0.29-0.39]**	**0.58 [0.55-0.60]**	0.59	0.55	0.04	-0.08 : 0.08
baroli	Vila	10	8	**0.48 [0.44-0.54]**	**0.64 [0.63-67]**	0.47	0.45	0.02	-0.09 : 0.09
bailloni	North Reunion	14	8	**0.77 [0.75-0.82]**	**0.33 [0.23-0.43]**	**0.68***	**0.78***	**-0.10**	**-0.06 : 0.07**
bailloni	South Reunion	17	11	**0.58 [0.58-0.66]**	**0.42 [0.34-0.49]**	**0.6***	**0.72***	**-0.12**	**-0.07 : 0.08**
nicolae	All populations	14	10	**0.61 [0.59-0.68]**	**0.42 [0.39-0.46]**	**0.6***	**0.72***	**-0.12**	**-0.09 : 0.09**

Figure 1: World map distribution of *Puffinus lherminieri* complex (see Austin et al. 2004 for *assimilis*).

Numbers represent breeding localities that were sampled for this study (colour codes identical across figures). The size of filled circles corresponds to sample size (number of individuals). "X" represents the only breeding colony from North Atlantic that was not sampled here. Other letters (A to E) for other taxa within the complex.

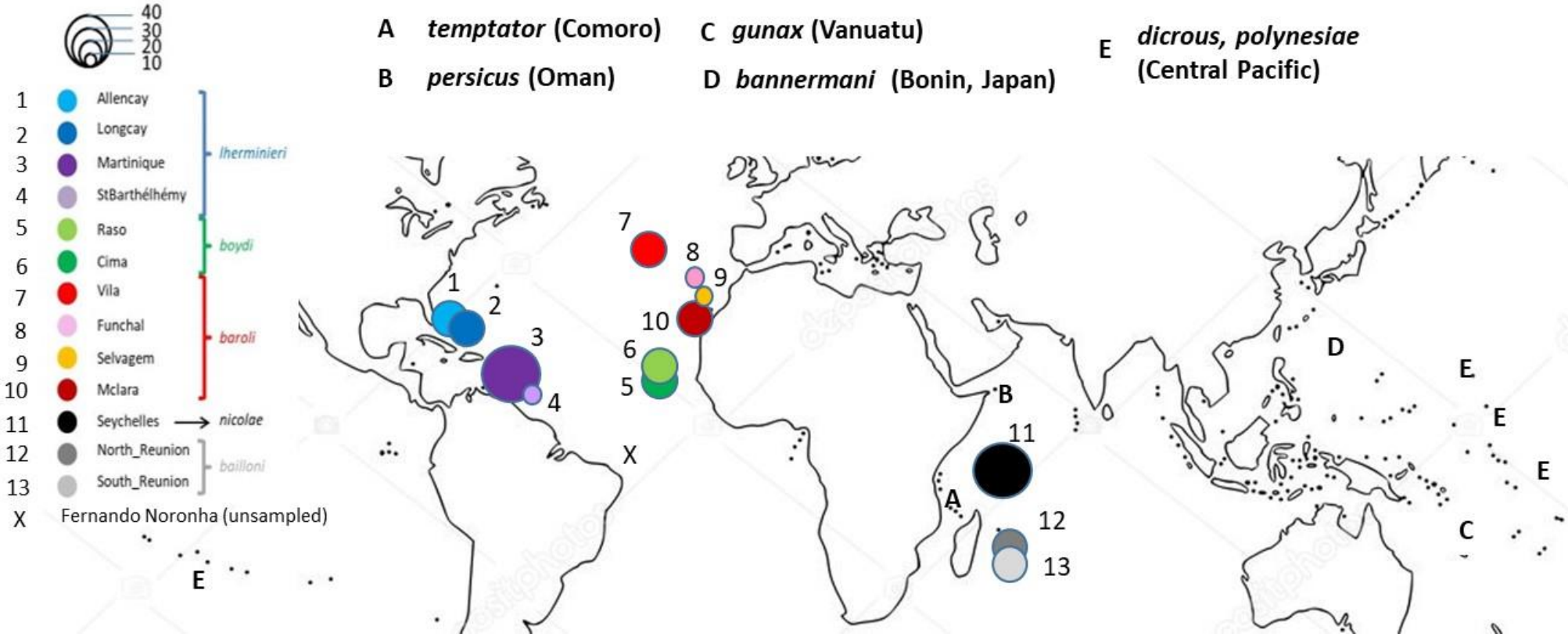


Figure 2: Gene trees and scenario of breeding site colonization

a. Species trees obtained by Bayesian inference for all mitochondrial and nuclear markers, node bars correspond to the 95% confidence interval of the estimated divergence times. The scale corresponds to time before present in Million years (My). b. Gene trees obtained using *BEAST for all mitochondrial markers with dichrous haplotypes in yellow. c. Gene trees obtained using *BEAST for all mitochondrial and nuclear markers with dichrous sequences in yellow. In b. and c. only individuals sequenced for all mitochondrial markers and all markers respectively are showed. Only the posterior values >0.90 are shown. d. Scenario of colonization inferred based on DIYABC. Branch colours correspond to basal populations, dates to mean divergence times of trees in My inferred by *BEAST analyses. Each bifurcation corresponds to a divergence-colonization event. * Indicates the split between Atlantic and Indian basal populations. The analysis excludes the population of Funchal, due to low sample size (4 individuals).

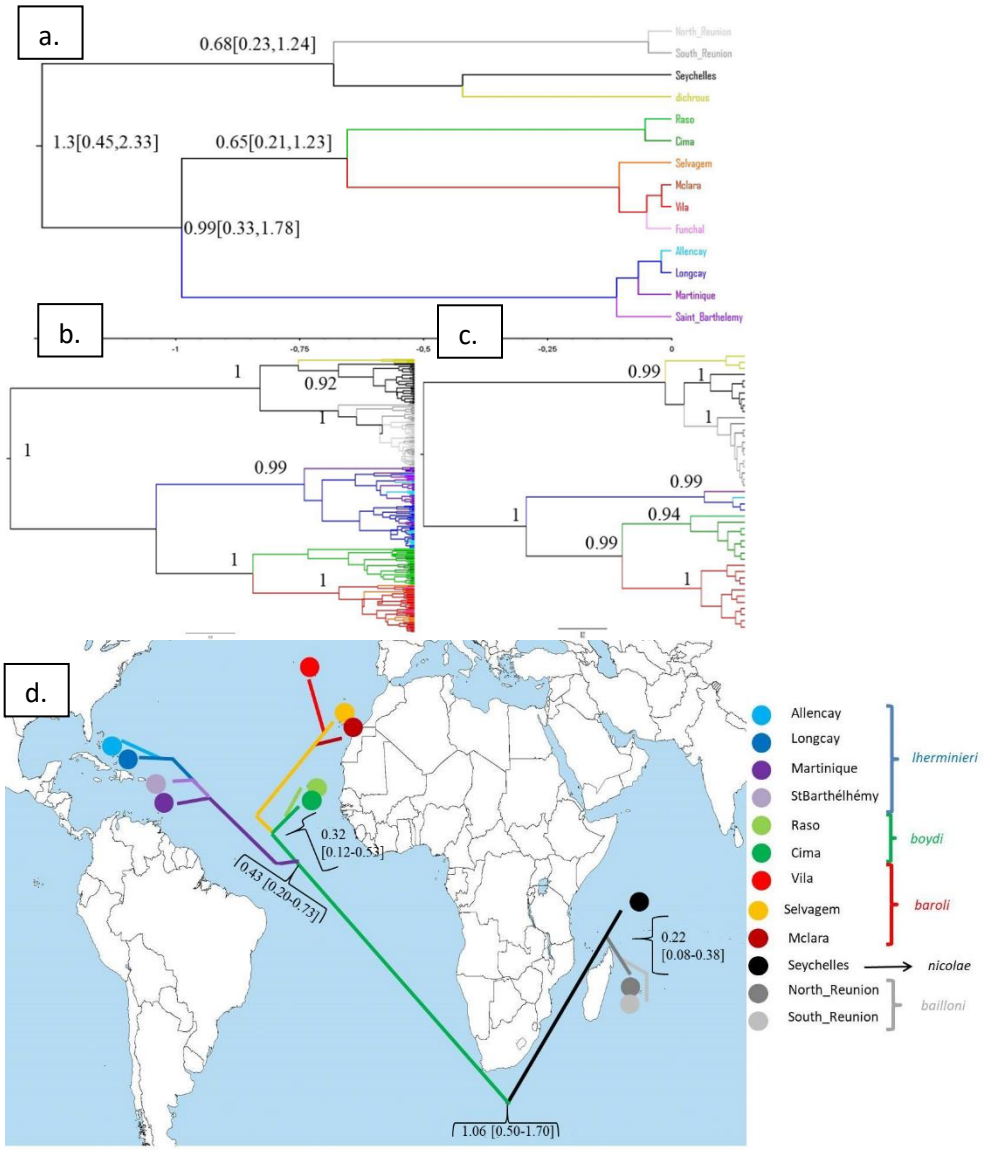


Figure 3: NeigborNet networks obtained using mitochondrial markers (a.) and nuclear markers (b.).
The scale bars indicate the sequence divergence (number of substitutions per site) represented by the length of a branch.

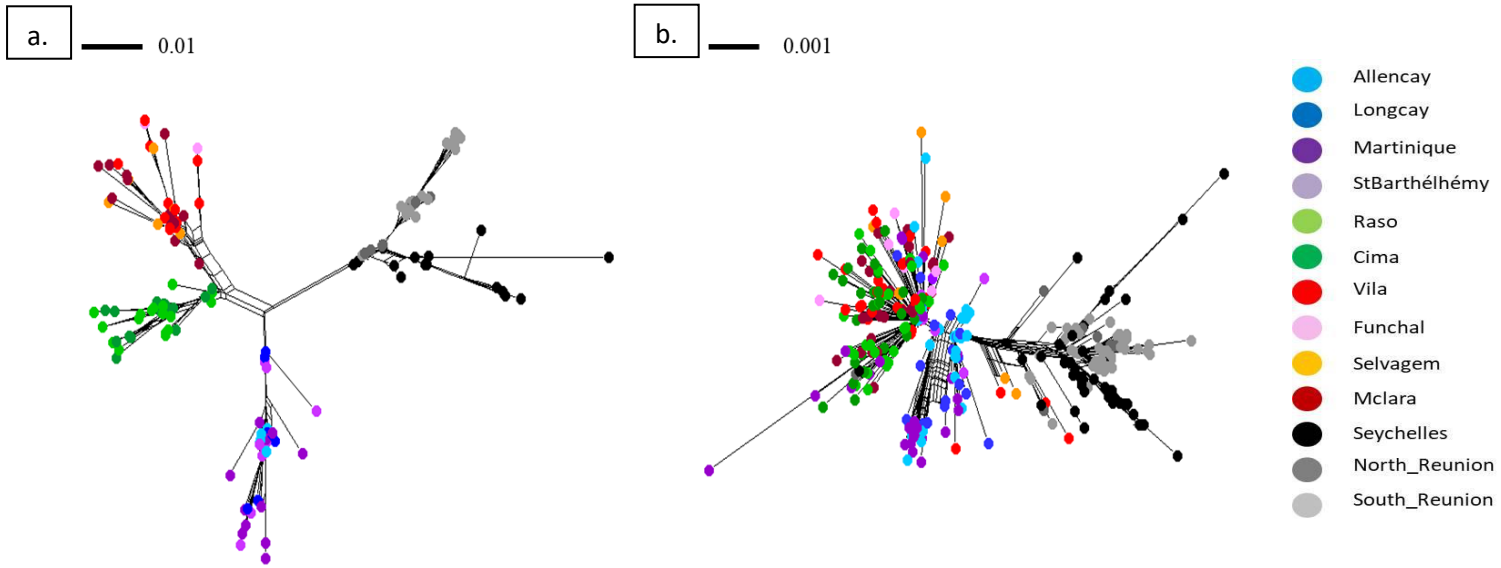
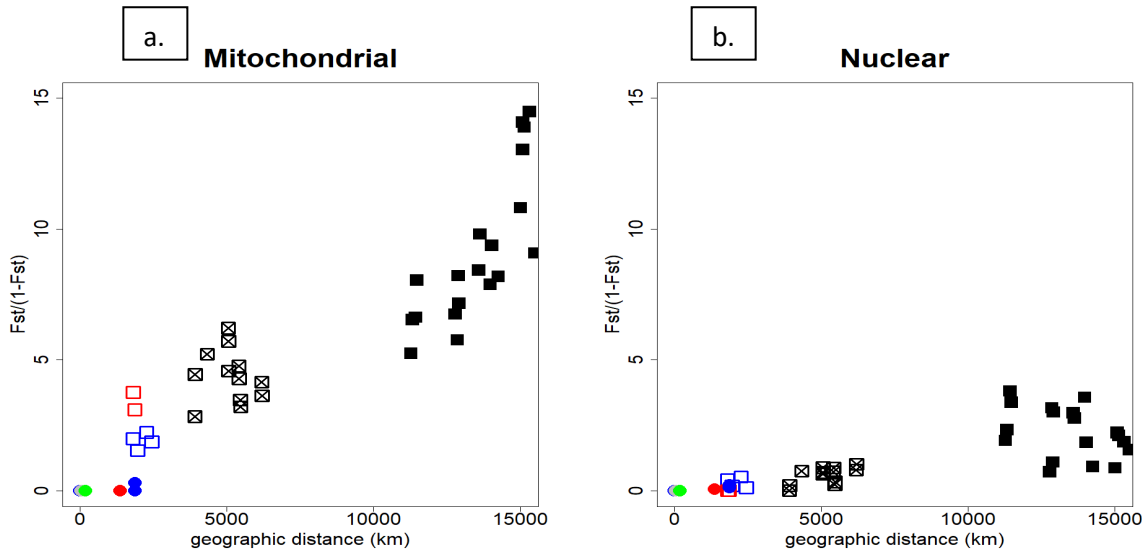


Figure 4: Correlation between gentic and geographic distance
Relationships between genetic (F_{ST}/(1-F_{ST})) and geographic distances. Genetic distances were calculated for all mitochondrial markers (a.) and all nuclear markers (b.). Geographic distances were calculated as the shortest distances between pairs of populations without crossing land. Black squares: pairs of populations from Atlantic and Indian oceans. Light squares: pairs of population between *lherminieri* and *baroli-boydi (crossed)*, *boydi* and *baroli (in blue)* and *bailloni* and *nicolae (in red)*. Coloured circles include only pairs of populations belonging to the same lineage.



Annexe 4

Supplementary Material for Sea temperature, rather than land mass or geographical distance, drives genetic differentiation in a highly-dispersive seabird species complex

Supplementary Material 5.1: List of sequences used in this study.

a. Summary table: sample size per lineage, colony, and sex. b. Additional sequences obtained from Genbank

Lineage	Colony	Sample size	Number of females	Number of males	Genbank accession numbers
lherminhieri	Allencay	19	8	10	XXXX
lherminhieri	Longcay	20	8	12	XXXX
lherminhieri	Martinique	44	19	22	XXXX
lherminhieri	StBathélémy	8	NA	NA	XXXX
boydi	Cima	18	10	9	XXXX
boydi	Raso	18	11	7	XXXX
baroli	Funchal	4	NA	NA	XXXX
baroli	Mclara	15	6	9	XXXX
baroli	Selvagem	10	3	1	XXXX
baroli	Vila	19	10	8	XXXX
bailloni	North Réunion	28	14	8	XXXX
bailloni	South Réunion	32	17	11	XXXX
nicolae	Seychelles	41	14	10	XXXX
dichrous	Marquises	3	NA	NA	XXXX
pacificus	Réunion	1	NA	NA	XXXX

Genus	species	Marker	Genbank accession	Reference	Voucher
Puffinus	*lherminieri*	*co1*	DQ434015	Kerr et al. 2007	USNM 620720
Puffinus	*lherminieri*	*co1*	JQ176049	Schindel et al. 2011	USNM:Birds:607634
Puffinus	*lherminieri*	*co1*	JQ176050	Schindel et al. 2011	USNM:Birds:607633
Puffinus	*lherminieri*	*cob*	AF076085	Nunn and Stanley 1998	
Puffinus	*baroli*	*cob*	AY219935	Austin et al. 2004	tissue sample Pabr93
Puffinus	*baroli*	*cob*	AY219936	Austin et al. 2004	tissue sample Pabr91
Puffinus	*boydi*	*cob*	AY219937	Austin et al. 2004	museum skin BMNH 1936.2.21.87
Puffinus	*lherminieri*	*cob*	AY219940	Austin et al. 2004	museum skin BMNH 1913.12.26.75
Puffinus	*lherminieri*	*cob*	AY219941	Austin et al. 2004	museum skin BMNH 1932.4.13.1
Puffinus	*lherminieri*	*cob*	AY219942	Austin et al. 2004	museum specimen MZUSP 75186
Puffinus	*lherminieri*	*cob*	AY219943	Austin et al. 2004	museum tissue sample LSM B20918
Puffinus	*lherminieri*	*cob*	AY219944	Austin et al. 2004	tissue sample Pllh_EP7
Puffinus	*lherminieri*	*cob*	AY219945	Austin et al. 2004	tissue sample Pllh_EP8
Puffinus	*loyemilleri*	*cob*	AY219946	Austin et al. 2004	

Puffinus	*lherminieri*	*cob*	AY219947	Austin et al. 2004	
Puffinus	*lherminieri*	*cob*	AY219948	Austin et al. 2004	
Puffinus	*nicolae*	*cob*	AY219956	Austin et al. 2004	
Puffinus	*nicolae*	*cob*	AY219957	Austin et al. 2004	
Puffinus	*nicolae*	*cob*	AY219960	Austin et al. 2004	
Puffinus	*bailloni*	*cob*	AY219963	Austin et al. 2004	
Puffinus	*bailloni*	*cob*	AY219964	Austin et al. 2004	tissue sample Plba_VB
Puffinus	*boydi*	*cob*	L43024	Austin 1996	
Puffinus	*boydi*	*cob*	L43025	Austin 1996	
Puffinus	*lherminieri*	*cob*	L43047	Austin 1996	
Puffinus	*lherminieri*	*cob*	U57815	Austin 1996	
Puffinus	*pacificus*	*co1*	JF498895.1	Kerr et al. 2001	USNM:643456
Puffinus	*pacificus*	*cob*	AF076088.1	Nunn & Stanley 1998	

Supplementary Material 5.2: Primer sequences, PCR profile and conditions

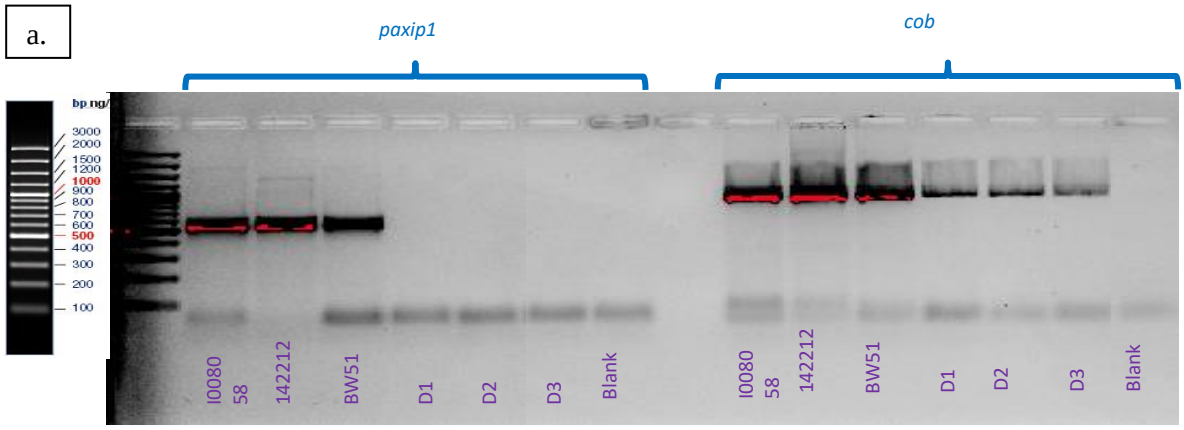
Shearwater-specific primers were designed using Primer3 (Untergasser et al. 2012) to amplify and sequence the three mitochondrial markers and *irf2*, *rag1* and *tpm*. We used primers designed from avian genomes to amplify *paxip1* and *csde1* (Kimball et al. 2009). Due to their short size and low level of polymorphism, *paxip1*, *tpm* and *irf2* were sequenced only in one direction (forward strand). Because of issues of length polymorphism resulting in unreadable chromatograms, we designed internal primers for the *βfib* marker, in addition to the two designed by Gangloff et al. (2013). Polymerase Chain Reactions (PCR) were carried out using TaKaRa ExTaq® Polymerase Hot-Start Version, in a total volume of 30 µL, using 1X Ex Taq Buffer (Mg2+ plus) with a final concentration of 200 µM of dNTP, 0.8 µM of each primer, 0.015 U of Taq and 60 ng of DNA extract. For all markers, after an initial denaturation step of two min at 95°C, we ran 40 PCR cycles consisting in 1 min at 95°C, 1 min at the primer-specific annealing temperature (varying between 52°C and 64°C) and 1 min at 72°C. These cycles were followed by a 7 min final extension step at 72°C. PCR products were purified and sequenced by Eurofins Genomics Munich. Chromatograms were checked and assembled into contigs using Sequencher v.5.4.1 (Gene Codes Corporation).

Primer name	Sequence	Origin	Strand	Marker	Melting temperature
Co1-F1-PUF-CRI	CTCAGCCTACTCATCCGTG	this study	Forward	*co1*	58.8°C
Co1-R3-PUF-CRI	TGTTGRTATAGGACTGGGTC	this study	Reverse	*co1*	56.3°C
Cytb-F1-Puf-CRI	GGCCTACTACTAGCYATACA	this study	Forward	*cob*	56.3°C
Cytb-R4-PUF-CRI	GTTARGATGAATAGGTTRGCG	this study	Reverse	*cob*	55.9°C
RCM-PUF-CRI-F	GGGTTGCTGATTTCTCGTGA	this study	Forward	Control region	57.3°C
RCM-PUF-CRI-R	GGCAAACACATTCAATGCATG	this study	Reverse	Control region	55.9°C
PAX-20F	CCCTCAGACACTGGATTAYGAATCAT	Kimball et al. 2009	Forward	*paxip1*	62.4°C
PAX-21R	CCAAGGATTCCGAAGCAGTAAG	Kimball et al. 2009	Reverse	*paxip1*	60.3°C
CSDE-5F	CTGGTGCTGTAAGTGCTCGTAAC	Kimball et al. 2009	Forward	*csde1*	64.6°C
CSDE-6R	CCAGGCTGTAAGGTTTCTAGGTCAC	Kimball et al. 2009	Reverse	*csde1*	62.4°C
TPM-F1-CRI	TGCAACCCAAGTCTTTCAGC	this study	Forward	*tpm*	57.3°C
TPM-R2-CRI	TTCGGAAGGAAGGCAGGAAA	this study	Reverse	*tpm*	57.3°C
IRF2-F-PUF-CRI	TGAAATTGAAAACCTAAGGCGAA	this study	Forward	*irf2*	55.3°C
IRF2-R-PUF-CRI	TGGAACTCTCTTTCTGCAGGA	this study	Reverse	*irf2*	55.9°C
BFIB-BI7U	GGAGAAAACAGGACAATGACAATTCAC	Gangloff et al. 2013	Forward	*βfib* (first half)	61.9°C
BFIB-R2-CRI	ACAATTGAGCTCCTGTCTTCTG	this study	Reverse	*βfib* (first half)	58.4°C
BFIB-BI7L	TCCCCAGTAGTATCTGCCATTAGGGTT	Gangloff et al. 2013	Forward	*βfib* (second half)	64.6°C
BFIB-F3-CRI	CAGAAGACAGGAGCTCAATTGT	this study	Reverse	*βfib* (second half)	58.4°C
RAG1-PUF-CRI-F	TCGCTCCAGATTTTCAGCATG	this study	Forward	*rag1*	57.3°C
RAG1-PUF-CRI-R	TCTTGCCAGATCTGTGAGCA	this study	Reverse	*rag1*	57.9°C

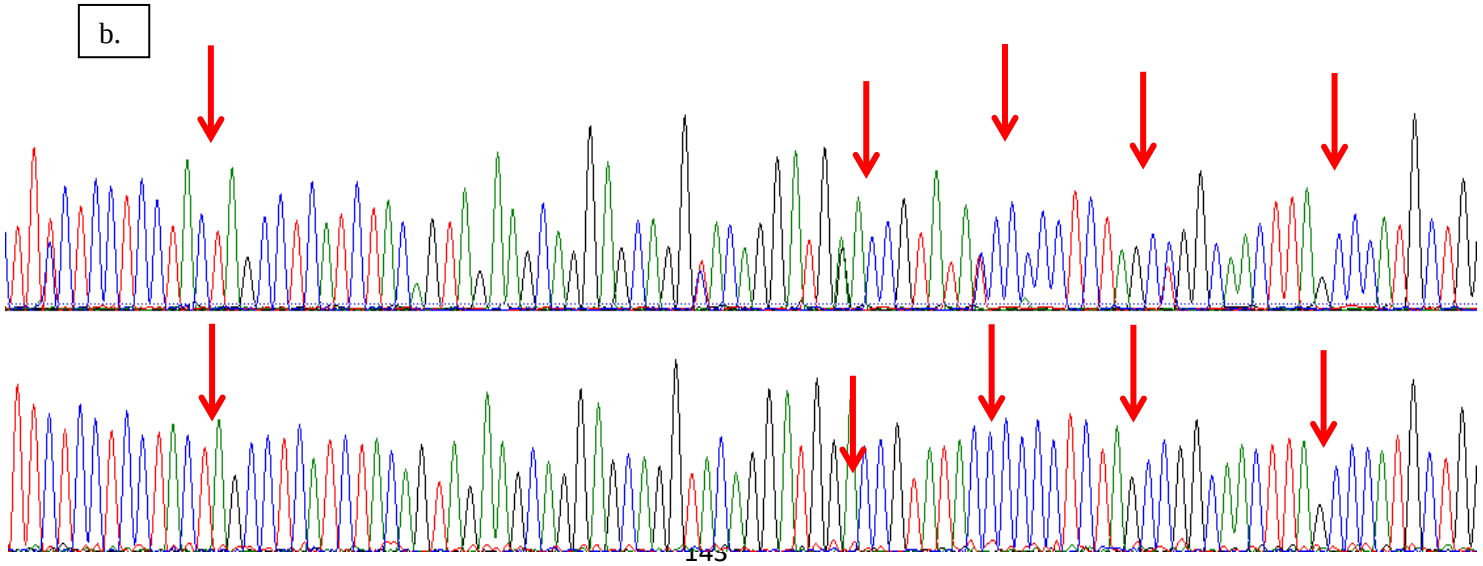
Supplementary material 5.3: Detection and removal of nuclear pseudogenes of mitochondrial origin (numts)

To avoid numts, we digested nuclear DNA with the ExonucleaseV (ExoV; NEB-M0345S) prior to PCR amplification and Sanger sequencing of mitochondrial markers, using the following protocol inspired by (Jayaprakash et al. 2015). One ng of DNA sample was heated to 70°C to inactivate any residual Proteinase K from the extraction protocol. Digestion was then carried out, adding to the sample 1X NEB4 Buffer, 1 mM ATP, 0.3 U of ExoV, and 0.24 mg/mL of BSA. The mix was heated to 37°C during 48h, followed by 30 min at 70°C to inactivate the exonuclease.
To test whether the digestion protocol effectively removed all traces of nuclear DNA, we performed PCR amplifications on the same three individuals before and after ExoV digestion, targeting a nuclear marker (*PAXIP1*; 515 bp) and a mitochondrial marker (*cob;* 877 bp). Our expectation was that ExoV treatment would destroy all template for *PAXIP1* amplification, while sparing mtDNA. We compensated the lowered PCR yield by using BSA at a final concentration of 0.24 mg/mL. PCR products were sent to Eurofins for sequencing. The results on an agarose gel are showed here (a.). The three individuals are I008058 (*baroli*), 142212 (*bailloni*) and BW51 (*nicolae*). D1, D2 and D3 correspond to the same individual after applying the digestion protocol. The two markers indicated here are *paxip1* and *cob*. As expected, the nuclear marker is not amplified after digestion, whereas the mitochondrial marker is still amplified for the three individuals. Below are showed the chromatograms before (b.) and after the digestion step (c.) of *co1* marker base 198 to 291 from 142208 (*bailloni*) individual. The double peaks present in the chromatograms before digestion are not present after the digestion step (position indicated by arrows).

a.



b.



Supplementary Material 5.4: Bayesian inference of gene trees and species tree

Among- and within lineage divergence, as well as reciprocal monophyly were evaluated using gene trees inferred using MrBayes v 3.2.6 (Ronquist et al. 2012). We ran two independents MCMC chains of 50x10^6 generations each, sampled every 100 generations, using three heated and one cold chain. The first 25% were discarded as burn-in. We used the model resulting from jModelTest2 (Darriba et al. 2015), with *Puffinus pacificus* as an outgroup for mitochondrial markers (Genbank sequences: JF498895.1 for *co1* (Kerr et al. 2007) and AF076088.1 for *cob* (Nunn & Stanley 1998) and sequences that we obtained of *Puffinus pacificus* for nuclear markers. We investigated the stationarity, visualizing the log-likelihood across generations in Tracer v 1.6 (Rambaut & Drummond 2007), checking the Effective Sample Size (ESS) and the convergence of multiple independent runs. Trees were inferred for all mitochondrial markers, and for both mitochondrial and nuclear markers, each marker considered independently. For each marker, the optimal model of substitution was estimated by jModeltest. These models were also inputted in both *BEAST and DiyABC. For each gamma model, four categories were inputted. For *BEAST, substitution rates were fixed to 1 for each marker. For DiyABC all population size priors followed a uniform distribution between 10^3 and 10^5 individuals, all times of divergence priors followed a uniform distribution between 10^4 and 2.10^6 years ago.

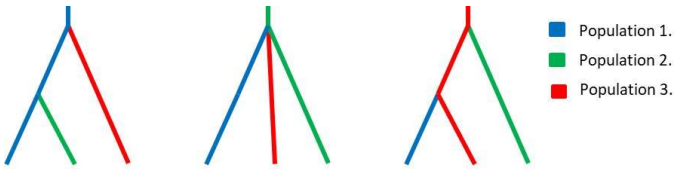
Marker	Gamma shape	Proportion of invariant sites	Substitution model	Model param	Model param	Model param	Model param	Model param	Model param
co1	0.02	0.8	HKY	11.5					
cob	1	0.8	TN93	3.2	5.8				
CR	0.75	0.45	GTR	0.1	14.3	0.1	1	7.4	1
paxip1	1	0.5	HKY	1.88					
csde1	0.5	0.9	TN93	4.6	4.7				
tpm	0.5	0.5	HKY	8.7					
irf2	0.5	0.5	HKY	4.2					
βfib	0.5	0.8	TN93	3.2	5.8				
rag1	0.5	0.9	GTR	1	2	0.2	0.2	2	1

Supplementary Material 5.5: Divergence scenario inferred by DiyABC

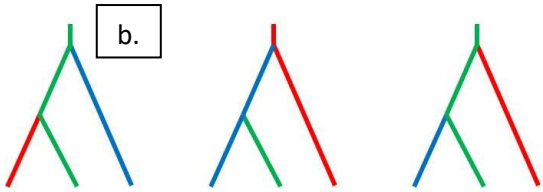
We used a coalescent-based ABC approach to explore the demographic scenario best describing the dataset of the combined mitochondrial and nuclear markers using the program DIYABC v. 2.1.0 (Cornuet et al. 2014). ABC methods consist in the simulation of datasets similar to the real dataset in terms of population and marker sizes. For each possible scenario, 10^6 pseudo-observed datasets were simulated, with the same ploidy and number of loci per population as observed in the real dataset. We fixed uniform priors for population sizes and divergence times priors. All population size priors followed a uniform distribution between 10^3 and 10^5 individuals, all times of divergence priors followed a uniform distribution between 10^4 and 2.10^6 years ago. For each marker, the optimal model of substitution was estimated by jModeltest. These models were inputted in both *BEAST and DiyABC. For each gamma model, four categories were inputted. We simulated the datasets with these parameters. Summary statistics were calculated from the simulated datasets and compared to the same statistics obtained from the real dataset. The Euclidean distance was calculated between the statistics obtained for each normalized simulated dataset and those for the observed dataset (Beaumont et al. 2002). Posterior probability of each scenario was then calculated using a logistic regression on summary statistics produced by the 1% of the simulated datasets the closest to the real dataset based on this Euclidian distance. To reduce the dimensionality of the data, a linear discriminant analysis was preliminarily applied to the summary statistics (Estoup et al. 2012). The scenario with the highest posterior probability value with a 95% confidence interval (95% CI) non-overlapping with other scenarios was selected.

To test the evolutionary scenario including only two groups, only two scenarios were possible and tested. When the scenario included 3 or 4 populations, we ran a hierarchical analysis by comparing first the posterior probability given by different topologies including all the populations. The different topologies tested for 3 populations is presented as an example below (a.).

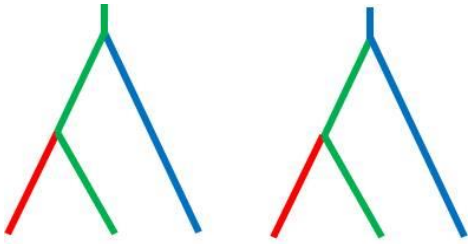
a.



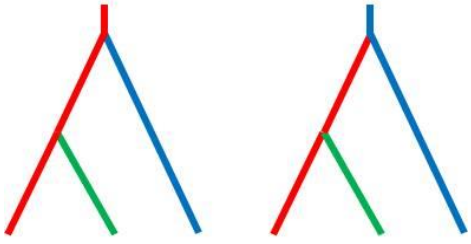
b.



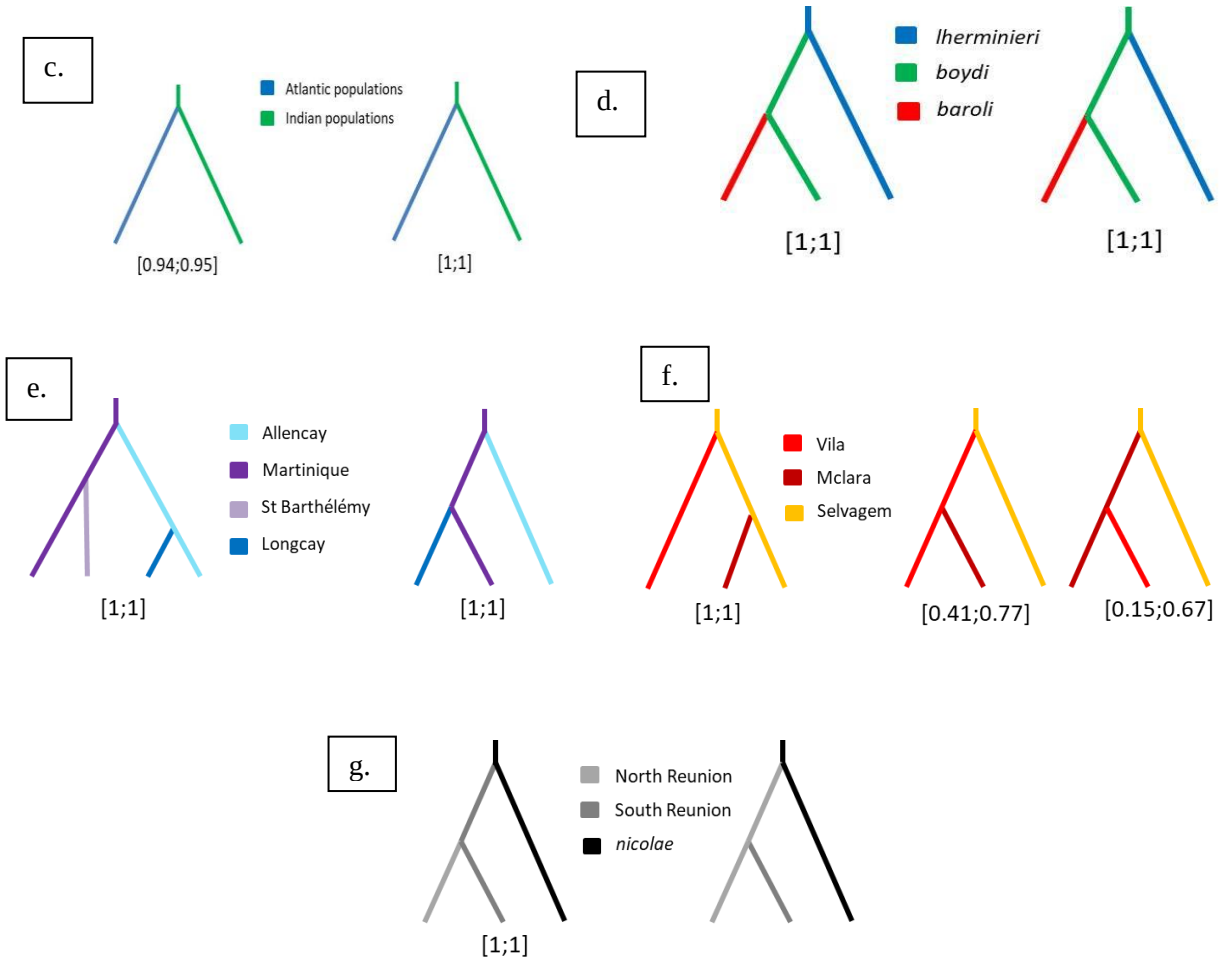
Once the best-fit topology was found, we compared the posterior probability given by different basal population for each node. The tested scenarios (with mitochondrial markers only) are represented for North Atlantic lineages, *lherminieri* populations, *baroli* populations and Indian Ocean populations in panel b.



We present hereafter the best-fit scenario for all populations. In each case, the left scenario was inferred for all mitochondrial markers and the right scenario was inferred for all markers. Posterior probabilities of each scenario were calculated using a logistic regression on summary statistics

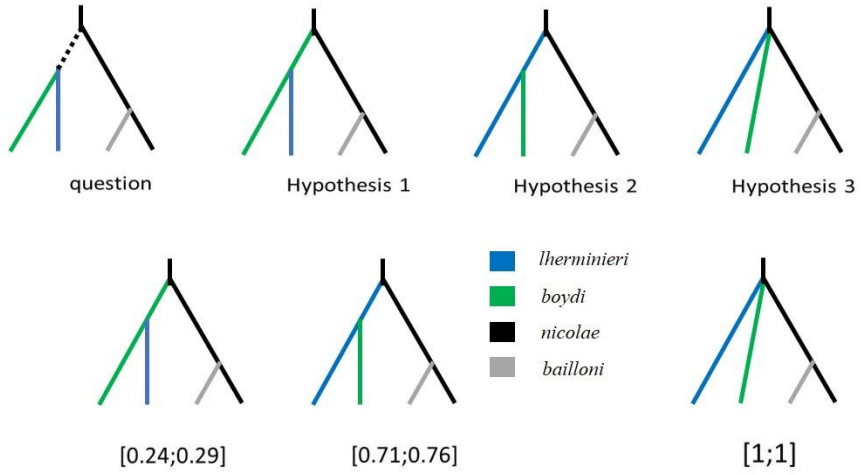


produced by the 1% of the simulated datasets the closest to the real dataset. The 95% confidence intervals of probability of the selected scenario after 10000 samples of the regression are indicated. We selected scenarios which 95% confidence interval did not include 0. (c.) All populations; (d.) Atlantic lineages; (e.) *lherminhieri* populations; (f.) *baroli* populations; (g.) Indian lineages. The selected scenarios show that Indian populations are basal over Atlantic populations. Mitochondrial markers show that *boydi* diverged from *lherminieri* whereas the use of both mitochondrial and nuclear markers show that *lherminieri* diverged from *boydi*. This incongruence is resolved later. Within *lherminieri* all markers show that Martinique (the southern) is the most basal population and Longcay (the northern) is the most recent. However, we cannot determine if Longcay diverged from Allencay in a stepping stone process or from Martinique. Similarly, within *baroli*, Selvagem population was the most basal but uncertainty remains in the order of appearance of the other populations. Within the Indian ocean, all markers show that *nicolae* is the basal lineage and the use of all marker show that the first population that have appeared is the North Reunion.



Finally, we tried to define the population that originally colonized the Atlantic Ocean from the Indian Ocean (dashed branch in h.). We have previously defined that *nicolae* was the basal lineage in the Indian Ocean and that either *boydi* either *lherminieri* was the basal lineage in the Atlantic Ocean. We tested if the basal lineage in the Atlantic Ocean was either *boydi*, *lherminieri*, or *nicolae*. The hypothesis of *nicolae* being the basal lineage in the Atlantic Ocean i.e. the dashed line being black could not be tested as such in the Diyabc software. We tested the hypothesis of the two Atlantic lineages as appearing simultaneously from *nicolae* as an equivalent. Results are presented in the lower panel of h. The selected scenario is considering *nicolae* as the basal lineage in the Atlantic Ocean.

h.



Supplementary material 5.6: Comparative analysis of the impact of the supplemental CR sequences

Due to the presence of numts and a duplicated region, we observed the presence of double peaks in the mitochondrial CR sequences (see Supplementary Material 3 for the removal of numts). As these sequences represented a third of the CR dataset, we decided to evaluate the effect of noise in phylogeographic analyses. We made all the Φ_{ST}, Tajima's D, Fu's Fs, MrBayes and *BEAST analyses using three datasets. The first dataset included all mitochondrial markers and all individuals. In this dataset, CR sequences either contained ambiguities (the Φ_{ST}, Tajima's D, Fu'Fs analyses) or were phased (MrBayes and *BEAST analyses). The second dataset contained only ambiguity-free *co1* and *cob* sequences. The third dataset was built by removing from the first dataset all 60 individuals that displayed double-peaks at the CR locus (*BEAST analyses), or for CR sequences only, encoding missing data for that locus (Arlequin and MrBayes analyses). *BEAST analyses were run for on mitochondrial markers only and on mitochondrial and nuclear markers. Dropping the CR marker altogether resulted in a strong loss of information. Most Φ_{ST} values are strongly higher in the analysis of the three mitochondrial markers than in the analysis on only two (Table S1). Similarly, if no Fu's Fs value was significant, regardless of the dataset, Cima, Mclara and North Reunion Tajima's D values were significant using only two mitochondrial markers rather than three (Table S2).

Table S1. Pairwise Φ_{ST} values for a. all mitochondrial markers and all individuals, b. *co1* and *cob* only for all individuals and c. all mitochondrial markers but individuals presenting ambiguities for the control region were removed. Border indicates the separation between intra and inter lineage. Triple band indicates the separation between intra- and inter-ocean comparisons. *: p<0.05; ***: p<0.001

a.		*lherminieri*	*lherminieri*	*lherminieri*	*lherminieri*	*boydi*	*boydi*	*baroli*	*baroli*	*baroli*	*baroli*	*bailloni*	*bailloni*	*nicolae*
		Allencay	**Longcay**	**Martinique**	**St Barthélémy**	**Raso**	**Cima**	**Funchal**	**Mclara**	**Selvagem**	**Vila**	**North Reunion**	**South Reunion**	**Seychelles**
lherminieri	**Allencay**	-												
lherminieri	**Longcay**	0.08*	-											
lherminieri	**Martinique**	0	0.22***	-										
lherminieri	**St Barthélémy**	0.19*	0.35***	0	-									
boydi	**Raso**	0.81***	0.83***	0.82***	0.77***	-								
boydi	**Cima**	0.76***	0.78***	0.74***	0.71***	0	-							
baroli	**Funchal**	0.81***	0.84***	0.85***	0.72***	0.7***	0.56***	-						
baroli	**Mclara**	0.78***	0.81***	0.82***	0.73***	0.67***	0.61***	0	-					
baroli	**Selvagem**	0.8***	0.83***	0.83***	0.73***	0.64***	0.53***	0	0	-				
baroli	**Vila**	0.85***	0.86***	0.84***	0.81***	0.69***	0.65***	0	0	0	-			
bailloni	**North Reunion**	0.93***	0.93***	0.89***	0.91***	0.87***	0.84***	0.89***	0.85***	0.88***	0.89***	-		
bailloni	**South Reunion**	0.93***	0.94***	0.9***	0.92***	0.89***	0.87***	0.91***	0.88***	0.88***	0.91***	0	-	
nicolae	**Seychelles**	0.93***	0.93***	0.9***	0.91***	0.89***	0.87***	0.91***	0.89***	0.9***	0.92***	0.76***	0.79***	-

b.		*lherminieri*	*lherminieri*	*lherminieri*	*lherminieri*	*boydi*	*boydi*	*baroli*	*baroli*	*baroli*	*baroli*	*bailloni*	*bailloni*	*nicolae*
	.	**Allencay**	**Longcay**	**Martinique**	**St Barthélémy**	**Raso**	**Cima**	**Funchal**	**Mclara**	**Selvagem**	**Vila**	**North Reunion**	**South Reunion**	**Seychelles**
lherminieri	**Allencay**	-												
lherminieri	**Longcay**	0	-											
lherminieri	**Martinique**	0	0.26***	-										
lherminieri	**St Barthélémy**	0.21*	0.45***	0	-									
boydi	**Raso**	0.89***	0.91***	0.92***	0.92***	-								
boydi	**Cima**	0.89***	0.91***	0.91***	0.92***	0	-							
baroli	**Funchal**	0.88***	0.91***	0.94***	0.91***	0.86***	0.85***	-						
baroli	**Mclara**	0.9***	0.91***	0.92***	0.92***	0.85***	0.85***	0	-					
baroli	**Selvagem**	0.88***	0.9***	0.94***	0.92***	0.87***	0.86***	0	0	-				
baroli	**Vila**	0.86***	0.9**	0.94**	0.91**	0.85***	0.84***	0	0	0	-			
bailloni	**North Reunion**	0.94***	0.95***	0.93***	0.96***	0.96***	0.95***	0.95***	0.95***	0.95***	0.95***	-		
bailloni	**South Reunion**	0.95***	0.96***	0.94***	0.97***	0.97***	0.96***	0.96***	0.96***	0.96***	0.96***	0.05*	-	
nicolae	**Seychelles**	0.79***	0.8***	0.72***	0.78***	0.8***	0.81***	0.8***	0.83***	0.78***	0.75**	0.54***	0.6***	-

c.		*lherminieri*	*lherminieri*	*lherminieri*	*lherminieri*	*boydi*	*boydi*	*baroli*	*baroli*	*baroli*	*baroli*	*bailloni*	*bailloni*	*nicolae*
		Allencay	**Longcay**	**Martinique**	**St Barthélémy**	**Raso**	**Cima**	**Funchal**	**Mclara**	**Selvagem**	**Vila**	**North Reunion**	**South Reunion**	**Seychelles**
lherminieri	**Allencay**	-												
lherminieri	**Longcay**	0	-											
lherminieri	**Martinique**	0	0.3***	-										
lherminieri	**St Barthélémy**	0	0.48***	0	-									
boydi	**Raso**	0.85***	0.86***	0.86***	0.85***	-								
boydi	**Cima**	0.83***	0.84***	0.81***	0.82***	0	-							
baroli	**Funchal**	0.83***	0.85***	0.86***	0.82***	0.71***	0.69***	-						
baroli	**Mclara**	0.88***	0.89***	0.89***	0.89***	0.74***	0.73***	0	-					
baroli	**Selvagem**	0.87***	0.89***	0.92***	0.91***	0.74***	0.67***	0	0	-				
baroli	**Vila**	0.82***	0.83***	0.85***	0.78***	0.73***	0.64***	0	0	0	-			
bailloni	**North Reunion**	0.94***	0.95***	0.94***	0.96***	0.92***	0.9***	0.9***	0.94***	0.96***	0.89***	-		
bailloni	**South Reunion**	0.94***	0.94***	0.92***	0.95***	0.92***	0.91***	0.91***	0.94***	0.94***	0.92***	0	-	
nicolae	**Seychelles**	0.92***	0.93***	0.91***	0.93***	0.91***	0.9***	0.91***	0.93***	0.93***	0.9***	0.78***	0.8***	-

Table S2. Results of Fu's Fs and Tajima's D for a. all mitochondrial markers and all individuals, b. *co1* and *cob* only for all individuals and c. all mitochondrial markers but individuals presenting ambiguities for the control region were removed.

a.	Allencay	Longcay	Martinique	St Barthélémy	Raso	Cima	Mclara	Vila	Selvagem	Funchal	North Reunion	South Reunion	Seychelles	Mean	s.d.
Tajima's D test															
Sample size	19	20	44	8	18	18	15	19	10	4	28	32	41	21	12
S	17	14	1	48	61	64	49	49	37	24	14	47	14	34	21
Pi	2053	1930	2302	1901	2279	1349	1914	1963	2845	2051	2495	2413	2543	2157	379
Tajima's D	-0.47	-0.27	7.39	-2.59	-2.23	-0.61	-0.95	-1.35	-2.40	-1.12	-1.18	-2.40	0.34	-0.60	2.57
p-value	0.352	0.419	1	0	0	0.285	0.193	0.071	0	0	0.104	0.002	0.718	0.24	0.32
Fu's FS test															
No. of alleles	19	20	44	8	18	18	15	19	10	4	28	32	41	21	12
Theta_pi	2053	1930	2302	1901	2279	1349	1914	1963	2845	2051	2495	2413	2543	2157	379
Exp. no. of alleles	19	20	44	8	18	18	15	19	10	4	28	32	41	21	12
FS	2.45	2.27	0.68	4.21	2.67	2.12	2.88	2.40	4.14	5.83	1.81	1.48	0.97	2.61	1.42
p-value	0.624	0.602	0.537	0.731	0.722	0.533	0.599	0.624	0.834	0.608	0.752	0.667	0.658	0.65	0.087

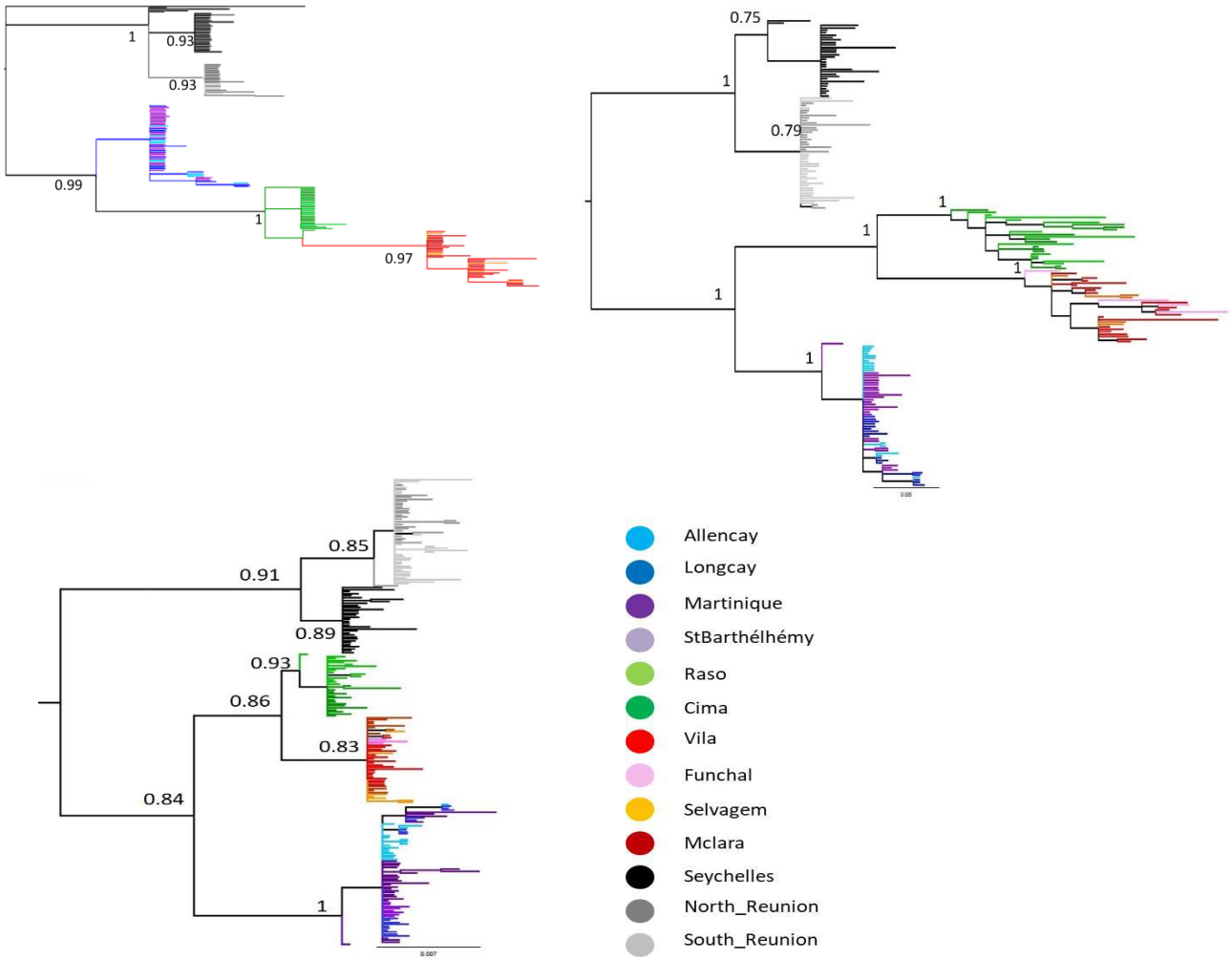
Annexe 4: Supplementary Material for Sea surface may drive genetic differentiation in seabirds

b.	Allencay	Longcay	Martinique	St Barthélémy	Raso	Cima	Mclara	Vila	Selvagem	Funchal	North Reunion	South Reunion	Seychelles	Mean	s.d.
Tajima's D test															
Sample size	17	19	25	8	16	18	14	19	8	3	25	30	36	18.31	9.19
S	13	10	3	6	9	12	4	5	6	5	9	12	7	7.77	3.3
Pi	292.24	348.43	551.72	269.89	513.18	409.56	437.45	384.96	687.46	931.67	443.28	321.19	481.76	467.14	180.33
Tajima's D	-1.04	-0.84	0.1	-0.85	-2.09	-2	0.79	1.24	-2.14	0	-1.77	-2.23	11.18	0.03	3.55
p-value	0.16	0.21	0.63	0.23	0.01	0.01	0.79	0.89	0	1	0.02	0	1	0.38	0.41
Fu's FS test															
No. of alleles	17	19	25	8	16	18	14	19	8	3	25	30	36	18.31	9.19
Theta_pi	292.24	348.43	551.72	269.89	513.18	409.56	437.45	384.96	687.46	931.67	443.28	321.19	481.76	467.14	180.33
Exp. no. of alleles	16.55	18.53	24.47	7.9	15.77	17.64	13.8	18.57	7.96	3	24.35	28.72	34.75	17.85	8.81
FS	9.93	4.5	0	2.22	9.76	7.93	6.99	2.63	7.32	5.74	10.41	4.87	20.8	0.8	0.4
p-value	1	0.97	1	0.52	1	1	1	0.83	1	0.84	1	0.98	1	0.93	0.14

c.	Allencay	Longcay	Martinique	St Barthélémy	Raso	Cima	Mclara	Vila	Selvagem	Funchal	North Reunion	South Reunion	Seychelles	Mean	s.d.
Tajima's D test															
Sample size	19	20	26	8	18	18	15	19	8	4	18	33	36	18.62	9.22
S	13	10	3	6	35	44	28	5	6	22	15	15	7	16.08	12.72
Pi	618	579	681	416	769	546	639	534	857	926	388	606	620	629	154
Tajima's D	-0.9	-0.51	1.17	-0.42	-2.64	-2.01	-1.54	5.43	-0.85	-1.28	-2.91	-1.87	1.14	-0.55	2.18
p-value	0.2	0.34	0.87	0.37	0	0.01	0.04	1	0.24	0	0	0.01	0.87	0.3	0.37
Fu's FS test															
No. of alleles	19	20	26	8	18	18	15	19	8	4	18	33	36	18.62	9.22
Theta_pi	618	579	681	416	769	546	639	534	857	926	388	606	620	629	154
Exp. no. of alleles	18.73	19.68	25.53	7.93	17.8	17.73	14.84	18.69	7.97	3.99	17.62	32.16	35.02	18.28	8.94
FS	15.76	3.18	25.69	2.67	10.3	1.14	1.73	0.99	7.76	5.04	7.73	1.36	4.15	6.73	7.14
p-value	1	0.9	1	0.57	1	0.48	0.63	0.5	1	0.73	1	0.72	0.98	0.81	0.21

If the topology of the Bayesian tree is the same in the three analyses, posterior probabilities values at lineage nodes are lower using only two markers (Dataset 2; Fig. S1). The analyses based on datasets 3 led to results closer to the dataset 1. Φ_{ST} values are quite similar between the two datasets except for some intra-lineage values that are significant in one case and not in the other (Table S1). Again there were differences in the Tajima's D values but not in the Fu's Fs values (Table S2). The differences observed between the analyses of the datasets 1 and 3 mainly include the populations with fewer individuals (St-Barthélémy, Selvagem and Funchal). This could be due to a loss of statistical power when too many individuals are removed. Posterior probabilities values are highly similar between the two bayesian trees.

Figure S1. MrBayes trees for a. all mitochondrial markers and all individuals, b. *co1* and *cob* only for all individuals and c. all mitochondrial markers but individuals presenting ambiguities for the control region were removed. Posterior probabilities superior to 0.75 are showed. The scale bars show how the length of a branch translates in sequence divergence. The unit is divergent nucleotides divided by the length of the sequence analysed.



However the divergence times between the lineages in the *BEAST analysis seems overestimated using dataset 1 (Table S3). This could be due to the supplemental information brought by the multiple copies of the control region in the dataset. This difference among the divergence time is lowered using also all the nuclear markers and the information they bring.

The artificial diversity brought by the duplicated sequences seems not bring major bias in the analyses, since the difference in the results are low when these sequences are removed and could be due to a loss of statistical power. These differences are lowered when adding supplemental markers to the analyses. We consider that the analyses performed using dataset 1 are valid.

Table S3. *BEAST estimations of divergence times, for the three datasets defined in Suppl. M6. Median divergence time and their 95% confidence interval are showed.

	Atlantic/Indian split	East Atlantic/West Atlantic split	*boydi*/*baroli* split	*nicolae*/*bailloni* split
Three mt markers, all individuals	3.92 [1.52-7.06]	2.4 [0.97-4.21]	1.43 [0.48-2.71]	1.32 [0.54-2.32]
Three mt markers, all individuals except 60	1.97 [0.80-3.12]	1.12 [0.37-1.92]	0.55 [0.20-0.94]	0.40 [0.15-0.68]
co1 and *cob* only	10.49 [2.14-23.72]	5.48 [0.94-12.29]	2.51 [0.45-5.64]	2.85 [0.54-6.84]
all markers, all individuals	1.06 [0.50-1.70]	0.43 [0.20-0.73]	0.32 [0.12-0.53]	0.22 [0.08-0.38]
all markers, all individuals except 60	1.15 [0.67-1.71]	0.73 [0.44-1.08]	0.35 [0.20-0.53]	0.22 [0.12-0.36]
all markers except CR	1.10 [0.47-1.82]	0.53 [0.25-0.88]	0.28 [0.11-0.52]	0.23 [0.07-0.42]

Supplementary Material 5.7: Pairwise distance results and Fu'Fs and Tajima's D results

Population average pairwise differences with above diagonal: average number of pairwise differences between populations (P_{XY}), diagonal cells: average number of pairwise differences within population (P_X), below diagonal: corrected average pairwise difference ($P_{XY}-(P_X+P_Y)/2$)
a. All mitochondrial markers and all individuals. b. All nuclear markers and all individuals.
c. Results of Fu's Fs and Tajima's D for all mitochondrial markers and all individuals.
d. Results of Fu's Fs and Tajima's D for all nuclear markers and all individuals
Border indicates the separation between intra and inter lineage comparisons. Triple bands indicate the separation between intra- and inter-ocean comparisons.
*: p<0.05; ***:p<0.001

a.		*lherminieri*	*lherminieri*	*lherminieri*	*lherminieri*	*boydi*	*boydi*	*baroli*	*baroli*	*baroli*	*baroli*	*bailloni*	*bailloni*	*nicolae*
		Allencay	Longcay	Martinique	St Barthélémy	Raso	Cima	Vila	Mclara	Selvagem	Funchal	North Reunion	South Reunion	Seychelles
lherminieri	Allencay	4	4	4	8***	26***	27***	41***	35***	29***	31***	39***	48***	38***
lherminieri	Longcay	0	3	1***	2***	26***	27***	34***	28***	23***	22***	37***	45***	37***
lherminieri	Martinique	0	4***	3	6	23***	20***	33***	36***	29***	32***	22***	30***	26***
lherminieri	St Barthélémy	1*	9***	-1	9	28***	28***	45***	29***	24***	26***	37***	49***	38***
boydi	Raso	32***	32***	29***	37***	8	10	26***	29***	23***	31***	30***	40***	36***
boydi	Cima	35***	34***	28***	39***	0	13	29***	30***	23***	29***	33***	43***	39***
baroli	Vila	35***	39***	27***	36***	18***	19***	8	0	0	0*	38***	48***	47***
baroli	Mclara	27***	35***	28***	40***	19***	18***	9	11	11	13	31***	41***	41***
baroli	Selvagem	23***	29***	23***	33***	15***	13***	8	1	9	16	27***	32***	32***
baroli	Funchal	23***	30***	24***	37***	20***	16***	10	1	5	13	25***	39***	35***
bailloni	North Reunion	42***	40***	25***	43***	35***	40***	43***	38***	32***	33***	2	2	10***
bailloni	South Reunion	51***	48***	34***	55***	45***	51***	54***	48***	38***	47***	0	3	13***
nicolae	Seychelles	41***	40***	29***	43***	41***	47***	52***	48***	38***	42***	8***	11***	3

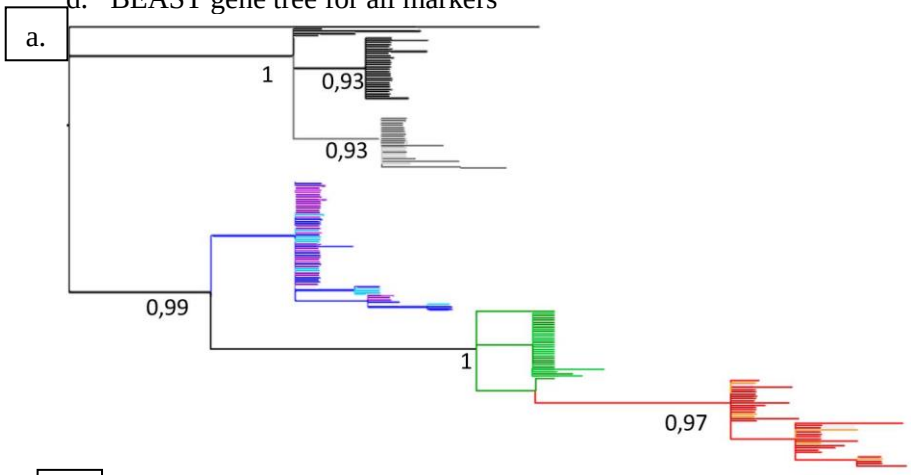
b.		*lherminieri*	*lherminieri*	*lherminieri*	*boydi*	*boydi*	*baroli*	*baroli*	*bailloni*	*bailloni*	*nicolae*
		Allencay	Longcay	Martinique	Raso	Cima	Vila	Mclara	North Reunion	South Reunion	Seychelles
lherminieri	Allencay	0	0	2.72***	0	4.65***	3.93***	3.66***	5.34***	5.17***	4.32***
lherminieri	Longcay	0.05***	1.43***	1.95***	2.77***	3.53***	2.95***	3.4***	3.88***	4.58***	3.11***
lherminieri	Martinique	0	0	1.03***	0	2.47***	2.07***	3.65***	3.35***	3.93***	2.7***
boydi	Raso	0.39***	0.63***	0	0	0	0	0	0	0	0
boydi	Cima	1.28***	0.47***	-0.39***	0	4.69***	0	0	0	0	0
baroli	Vila	1.78***	1.12***	0.44***	0	0	0	0	3.61***	4.76***	3.18***
baroli	Mclara	0***	1.06***	1.51***	0	0	0.23***	0	0	0	0
bailloni	North Reunion	2.8***	1.67***	1.33***	0	0.7***	1***	0	3***	5.19***	0
bailloni	South Reunion	2.25***	1.98***	1.52***	0	1.12***	1.76***	0	0	3.78***	3.79***
nicolae	Seychelles	2.27***	1.38***	1.16***	0	0***	1.04***	-0.19***	0.13***	0.88***	2.04***

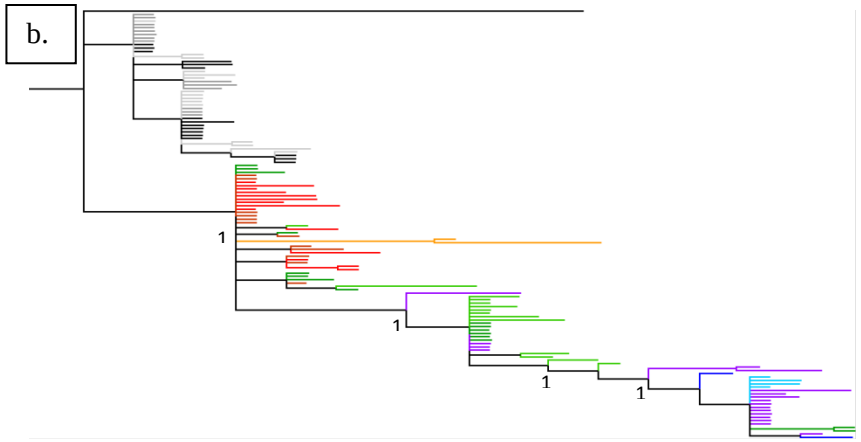
c.	**Allencay**	**Longcay**	**Martinique**	**St Barthélémy**	**Raso**	**Cima**	**Mclara**	**Vila**	**Selvagem**	**Funchal**	**North Reunion**	**South Reunion**	**Seychelles**	**Mean**	**s.d.**
Tajima's D test															
Sample size	19	20	44	8	18	18	15	19	10	4	28	32	41	21	12
S	17	14	1	48	61	64	49	49	37	24	14	47	14	34	21
Pi	2053	1930	2302	1901	2279	1349	1914	1963	2845	2051	2495	2413	2543	2157	379
Tajima's D	-0.47	-0.27	7.39	-2.59	-2.23	-0.61	-0.95	-1.35	-2.40	-1.12	-1.18	-2.40	0.34	-0.60	2.57
p-value	0.352	0.419	1	0	0	0.285	0.193	0.071	0	0	0.104	0.002	0.718	0.24185	0.31477
Fu's FS test															
No. of alleles	19	20	44	8	18	18	15	19	10	4	28	32	41	21	12
Theta_pi	2053	1930	2302	1901	2279	1349	1914	1963	2845	2051	2495	2413	2543	2157	379
Exp. no. of alleles	19	20	44	8	18	18	15	19	10	4	28	32	41	21	12
FS	2.45	2.27	0.68	4.21	2.67	2.12	2.88	2.40	4.14	5.83	1.81	1.48	0.97	2.61	1.42
p-value	0.624	0.602	0.537	0.731	0.722	0.533	0.599	0.624	0.834	0.608	0.752	0.667	0.658	0.65315	0.08694

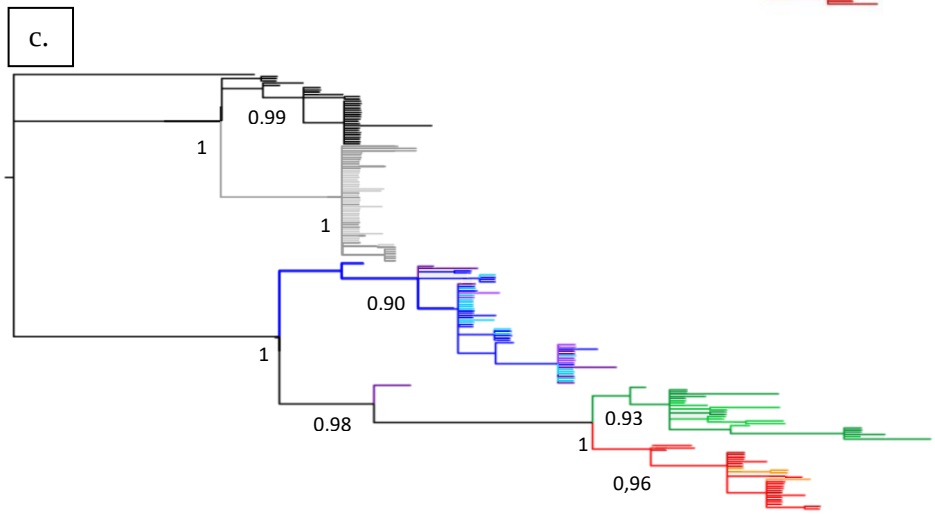
d.	Allencay	Longcay	Martinique	St Barthélémy	Raso	Cima	Mclara	Vila	Selvagem	Funchal	North Reunion	South Reunion	Seychelles	Mean	s.d.
Tajima's D test															
Sample size	19	20	41	5	18	18	14	18	10	4	18	33	41	19.92	11.85
S	4	4	1	6	20	11	13	24	4	2	11	1	14	8.85	7.39
Pi	1434	1348	1629	916	1515	891	1042	1348	1940	1124	1753	1723	1803	1420	347
Tajima's D	0.14	-0.45	1.58	-2.82	-3.04	0.78	-1.57	-2.84	0.02	0.59	-2.84	9.49	-1.35	-0.18	3.3
p-value	0.62	0.35	0.95	0	0	0.81	0.06	0	0.55	0.84	0	1	0.09	0.4	0.4
Fu's FS test															
No. of alleles	19	20	41	5	18	18	14	18	10	4	18	33	41	19.92	11.85
Theta_pi	1434	1348	1629	916	1515	891	1042	1348	1940	1124	1753	1723	1803	1420	347
Exp. no. of alleles	18.88	19.86	40.5	4.99	17.9	17.83	13.91	17.89	9.98	3.99	17.91	32.7	40.55	19.76	11.7
FS	20.67	19	4	4.51	5.42	4.41	2.4	12.47	3.75	5.23	4	11.4	21	6	4
p-value	1	1	1	0.59	0.97	0.92	0.56	1	0.8	0.56	1	1	1	0.88	0.18

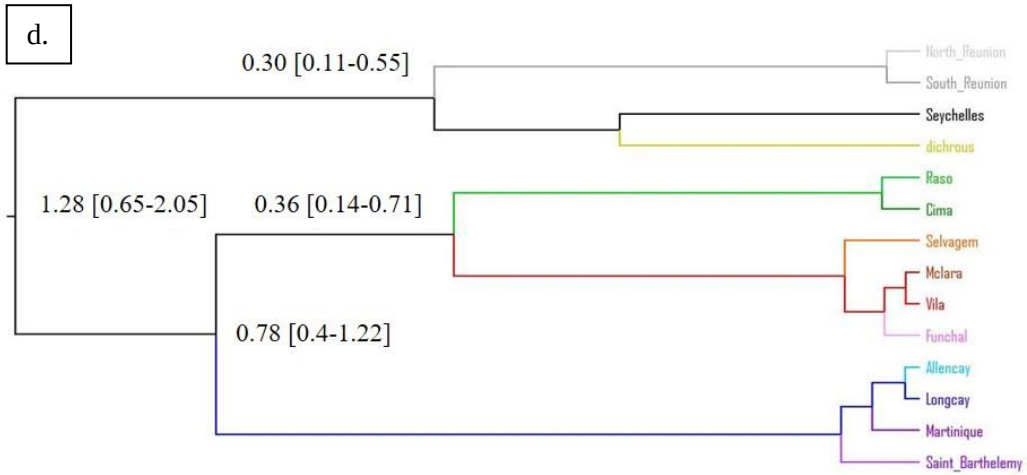
Supplementary Material 5.8: Gene trees obtained by MrBayes and *BEAST

Gene trees obtained by MrBayes for a. All mitochondrial markers, b. All nuclear markers, c. All markers.
Blue: *lherminieri*, green: *boydi*, red: *baroli*, grey: *bailloni*, black: *nicolae*, yellow: *dichrous*
d. *BEAST gene tree for all markers

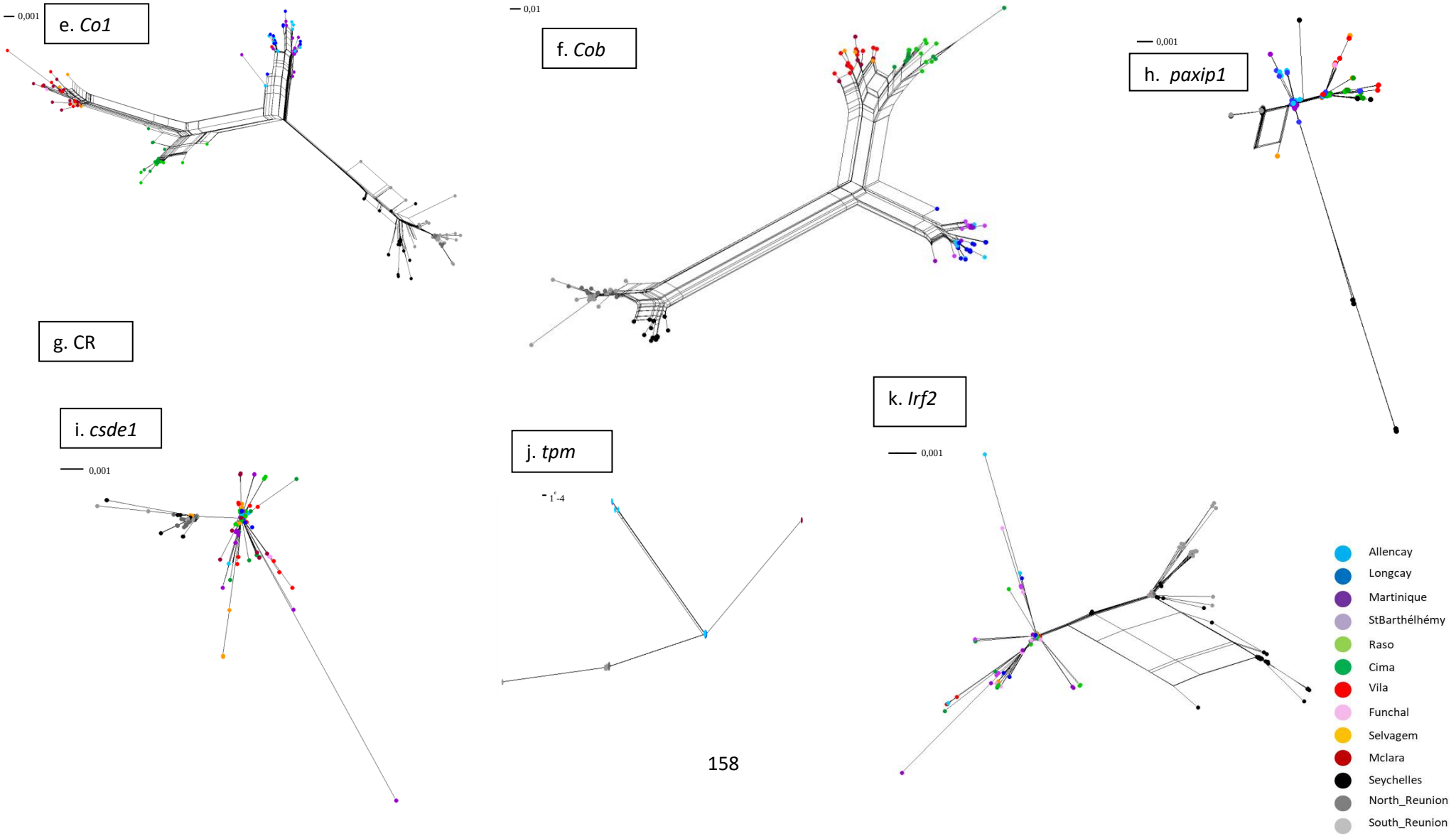


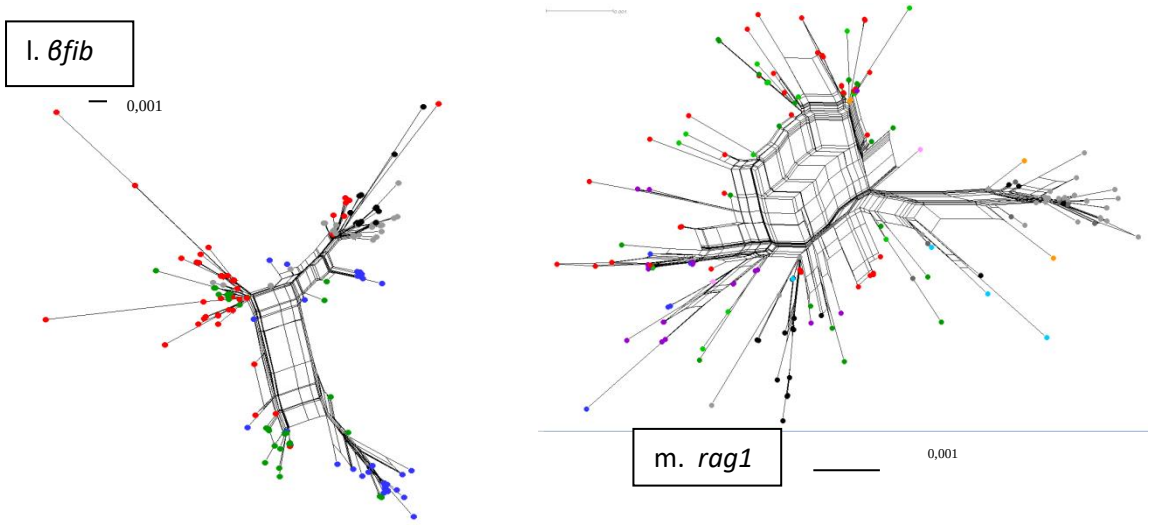






Neighbor networks on e. *co1*, f. *cob*, g. the Control Region, h. *paxip1*, i. *csde1*, j. *tpm*, k. *irf2*, l. *βfib*, m. *rag1*. The scale bars show how the length of a branch translates in sequence divergence. The unit is divergent nucleotides divided by the length of the sequence analysed.




l. βfib
0,001
m. rag1
0,001

Supplementary Material 5.9: Discordant mito-nuclear data

Assignment to an ocean basin based on nuclear and mitochondrial data were discordant for 33 individuals, for at least one nuclear locus. All of these individuals showed the mitochondrial signature expected based on their geographical sampling location. The haplotype networks of the two haplotypic phases of each nuclear marker showed us than one or two phases of these individuals could be discordant for one to three markers. The fact that two haplotypic phases of one individual could be discordant does not allow us to exclude contamination during lab work as a cause of this pattern. However recent hybridization as well as retention of ancestral polymorphism can also cause these patterns. For one individual, one haplotypic phase is found discordant for the more discriminant nuclear markers but not for mitochondrial markers, which could be indicative of first-generation hybridization. For other individuals, the discordant haplotypic phase at one or several markers could be a remaining of ancestral form of the loci. Further analyses are required to investigate this pattern more in detail, including the sequencing of more loci and correspondance and assignment analyses.

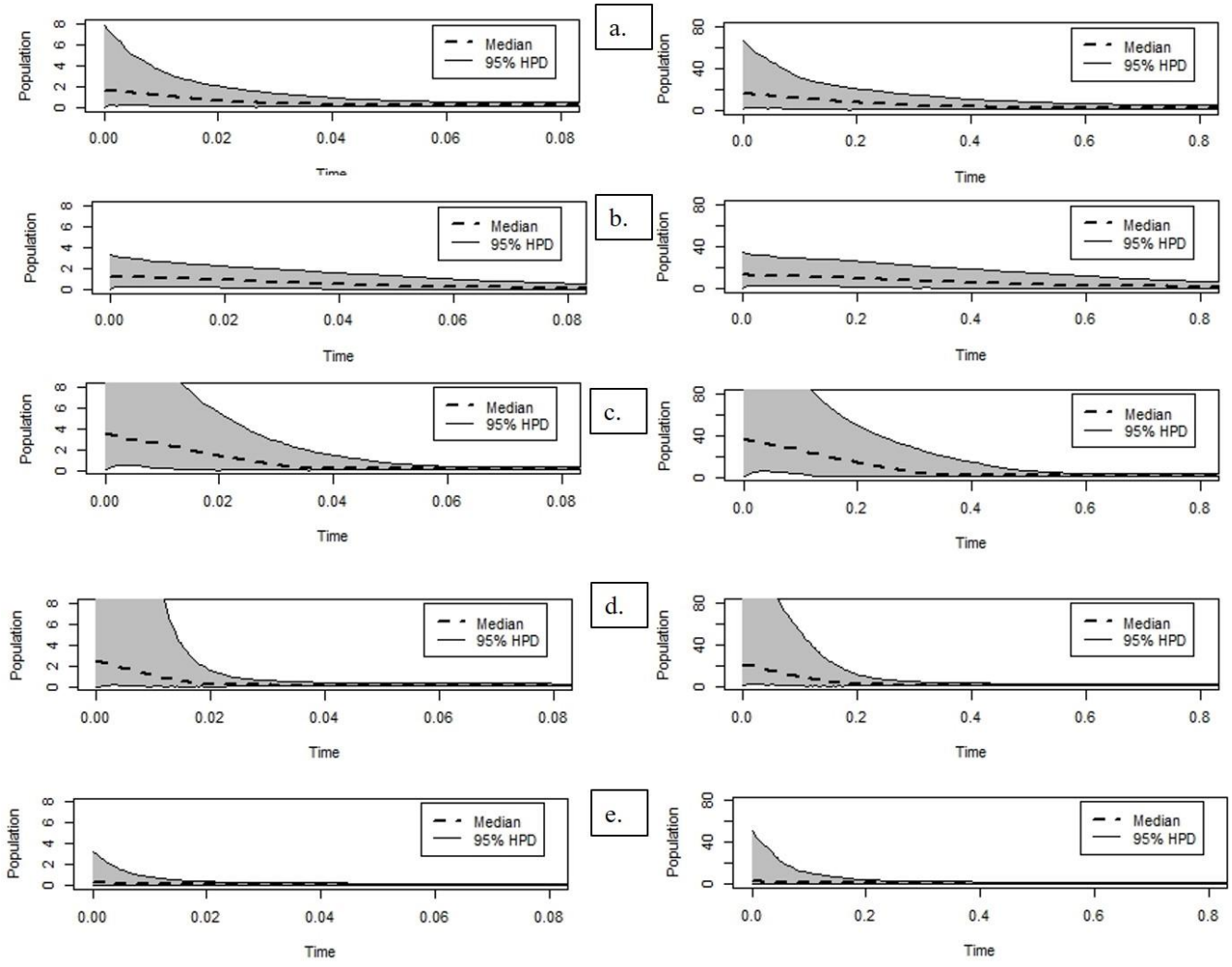
Inventory of discordant individuals in nuclear markers. For each population individuals presenting one or two haplotypic phases that was grouped with the individuals of the other ocean for at least one nuclear marker are indicated here. For each nuclear marker, an "x" indicates that the two haplotypic phases of the individual are discordant with its sampling location, a "v" indicates that only one phase is discordant. The sex is indicated if known.

Lineage	Population	Individual	*paxip1*	*csde1*	*tpm*	*irf2*	*βfib*	Sex	Geographic placement of dicordant phase
lherminieri	Allencay	Allencay 18					x	M	Indian populations
lherminieri	Allencay	Allencay 19					x	M	Indian populations
lherminieri	Longcay	Longcay19					v	M	Indian populations
lherminieri	Longcay	Longcay2					v	F	Indian populations
lherminieri	Martinique	BU83					x	M	Indian populations
boydi	Raso	5500040					v	M	Indian populations
boydi	Cima	5500491					v	M	Indian populations
baroli	Vila	I008058	v			v	v	M	Indian populations
baroli	Vila	I008072			v	v	x	M	Indian populations
baroli	Vila	I008098				v		F	Indian populations
baroli	Vila	I008099					x	M	Indian populations
baroli	Selvagem	Selvagem 1				v	x	F	Indian populations
baroli	Selvagem	Selvagem 2					x	F	Indian populations
baroli	Selvagem	SelvagemX1		v				NA	Indian populations
baroli	Selvagem	SelvagemX2	v					NA	Indian populations
bailloni	North Reunion	142169			v	v		M	Atlantic populations
bailloni	North Reunion	142213	v					F	East Atlantic populations
bailloni	North Reunion	BY13				v	x	F	East Atlantic populations
bailloni	North Reunion	BY15			x			F	Atlantic populations
bailloni	North Reunion	BY17	v				v	M	East Atlantic populations
bailloni	North Reunion	BY22			x	v		M	Atlantic populations
bailloni	North Reunion	BY23	v			v	x	F	East Atlantic populations
bailloni	North Reunion	BY25			x			F	Atlantic populations
bailloni	North Reunion	BY27			x			M	Atlantic populations
bailloni	South Reunion	1587	x					F	East Atlantic populations
bailloni	South Reunion	142175			v			F	Atlantic populations
bailloni	South Reunion	142529			x			F	Atlantic populations
nicolae	Seychelles	BW44	v					NA	West Atlantic populations
nicolae	Seychelles	BW66	v					F	West Atlantic populations
nicolae	Seychelles	BW68	v					M	West Atlantic populations
nicolae	Seychelles	GE50909	x					NA	West Atlantic populations
nicolae	Seychelles	GE50910	x					NA	East Atlantic populations
nicolae	Seychelles	GE50921	v					NA	West Atlantic populations

Supplementary material 5.10 : Bayesian skyline analyses

The four rows correspond to the four sets of clock rates used for *co1* and *cob*. The first one is the set used in our main analyses: 0.01588 ± 0.00115 per site per million year for *co1* (Pereira and Baker 2006) and 0.0189 ± 0.0035 per site per million year for *cob* (Weir and Schluter 2008). The second uses both rates of Pacheco et al. (2011), 0.00184 for *co1* and 0.00178 for *cob*. Population size is displayed as θ. Ne is obtained by multiplying it by the mean substitution rate.

a. *lherminieri* b. *boydi* c. *baroli* d.*bailloni* e. *nicolae*



Chapitre VI

Discussion générale

Nous avons investigué au cours de cette thèse une région mitochondriale dupliquée, en reconstruisant totalement ou partiellement la duplication le long de la phylogénie des Procellariiformes et montré une évolution complexe de cette région dupliquée et nos résultats pourraient aller dans le sens d'une relation entre cette région et la physiologie des Procellariiformes (Chapitre II et III). Nous avons mené une étude comparative en appliquant des mesures de diversité et différenciation génétique sur des jeux de données biaisées ou non par des problèmes de copies multiples. Nous avons ainsi mis en avant le traitement optimal de ces problèmes de copies multiples (Chapitre IV). Enfin notre étude phylogéographique nous a permis de reconstruire le scénario de différenciation des puffins et de mieux comprendre les facteurs de diversification chez ces organismes à forte capacité de dispersion (Chapitre V). Nous avons mis en avant pour la première fois l'importance apparemment primordiale des changements de température océanique passés dans le processus de diversification des oiseaux marins en les corrélant à des événements de divergence. Nous suggérons que les zones d'alimentation en mer ségrégées semblent contribuer à maintenir cette différenciation entre les lignées.

I. Patrons d'évolution moléculaire et processus sous-jacents

1. Discordance mito-nucléaire

En comparant la différenciation génétique de loci mitochondriaux et nucléaires, que ce soit les marqueurs traditionels ou les numts, chez les puffins, l'ADN mitochondrial a montré une plus forte structuration des lignées que l'ADN nucléaire. Cette discordance entre l'information mitochondriale et nucléaire a déjà été observée chez plusieurs espèces de Procellariiformes (Burg & Croxall 2001; Deane 2013; Gangloff et al. 2013; Silva et al. 2015), mais aussi chez d'autres oiseaux (e.g. Crochet et al. 2003, Martínez-Cruz et al. 2007, Vallender et al. 2007) et d'autres organismes terrestres (Bull et al. 2006; Morgan-Richards & Wallis 2003; Won & Hey 2005).

Une explication possible serait le tri incomplet des lignées. L'ADN mitochondrial évoluant plus vite, la variation ancestrale est plus vite perdue que pour l'ADN nucléaire (Funk et Omland 2003). De plus, du fait de son mode de transmission uniparentale et de son haploïdie, l'ADN mitochondrial a une taille efficace de population quatre fois plus petite que l'ADN nucléaire et est ainsi plus soumis à la dérive génétique (Funk et Omland 2003). L'ADN nucléaire est donc plus susceptible de porter des allèles ancestraux, peu différenciés, ce qui abaissera artificiellement le niveau de différenciation. Le tri de lignée incomplet est ainsi proposé comme principale cause de discordance mito-nucléaire (McKay & Zink 2010; Toews & Brelsford 2012).

La discordance des signaux mitochondriaux et nucléaires peut également résulter de processus biologiques tels que l'hybridation récente. La présence, dans certaines populations, d'allèles nucléaires parentaux provenant d'autres populations est en effet susceptible de diminuer le signal de différenciation porté par les marqueurs nucléaires. Nous suggérons la présence d'hybridation dans le complexe étudié, comme cela a déjà été mis en évidence chez des puffins (Genovart et al. 2007), des Procellariiformes (Brown et al. 2010; Gómez-Díaz et al. 2009) et plus largement chez des oiseaux marins (Gay et al. 2009; Morris-Pocock et al. 2011; Pons et al. 2014). D'autres mécanismes peuvent jouer sur le niveau de différenciation des marqueurs mitochondriaux et nucléaires. Dans notre étude nous mettons par exemple en avant

le fait que l'ADN mitochondrial porte l'histoire évolutive des femelles et celles-ci sont moins philopatriques que les mâles. Suivant l'hypothèse de Petit & Excoffier (2009), l'introgression interlignée mitochondriale sera de fait moins forte. Toutefois des patrons de discordance inverse, structuration plus forte chez les marqueurs nucléaires, ont été observés chez des oiseaux marins (e.g. Pons et al. 2014), mais aussi chez d'autres organismes dont les mâles sont moins philopatriques que les femelles. Si l'hypothèse de Petit et Excoffier ne s'applique pas à tous les cas, de nombreux autres mécanismes démographiques ou sélectifs ont été proposés pour expliquer les patrons de discordance mito-nucléaire (revus dans Hedrick 2010). Par exemple les contraintes de taille et de taux d'évolution des marqueurs nucléaires microsatellites font que ceux-ci sont sujets à de l'homoplasie, la ressemblance d'origine indépendante chez plusieurs individus (Balloux et al. 2000). Des différences de pression de sélection entre les différents marqueurs pourraient favoriser l'introgression de certains loci (voir Toews & Brelsford 2012). L'ADN mitochondrial retraçant uniquement l'histoire des femelles, dans une situation où uniquement les femelles d'une espèce se reproduisent avec une autre espèce, l'information portée par l'ADN mitochondrial sera bien différente de celle portée par l'ADN nucléaire, comme cela a été montré chez des éléphants (Roca et al. 2007).

2. Numts, hétéroplasmie et région dupliquée

Nous avons mis en évidence trois phénomènes de multiple-copie de marqueurs mitochondriaux dans notre complexe de puffins. Le premier est la présence de numts, copies nucléaires de marqueurs mitochondriaux. Le sang d'oiseau étant particulièrement pauvre en mitochondrie (Sorenson & Quinn 1998), les numts y sont facilement amplifiés et séquencés, cependant les numts sont présents chez une grande diversité d'organismes (Bensasson et al. 2001). Le deuxième phénomène est la présence de deux mitogénomes différents au sein d'un seul individu, ou hétéroplasmie. L'hétéroplasmie a été mise en évidence chez plusieurs oiseaux (e.g. Gandolfi et al. 2017, Guerrini et al. 2007), et est bien connue chez d'autres organismes (voir Barr et al. 2005 pour une synthèse). Enfin nous avons apporté la preuve que plusieurs marqueurs sont dupliqués au sein du mitogénome de *Puffinus lherminieri*. Le séquençage par longs fragments, par la technique MinION (ONT), nous a permis de montrer que la région de contrôle mitochondriale et *nad6*, ainsi que plusieurs ARNt, sont présents en double exemplaire dans le génome de ce puffin. Cette région dupliquée est différente de celle trouvée chez les albatros (Abbott et al. 2005; Lounsberry et al. 2015) suggérant une évolution complexe de la duplication au sein des Procellariiformes. Les phénomènes de duplication sont présents chez plusieurs Procellariiformes, mais ont été montré également chez plusieurs autres oiseaux (voir par exemple Gibb et al. 2013) et de nombreux autres organismes (Lynch & Conery 2000). Comprendre comment les gènes évoluent est essentiel pour pouvoir les utiliser au mieux.

Ces trois phénomènes de copies multiples résultent de mécanismes de modification de l'ADN complexes qu'il est important de prendre en compte afin d'étudier au mieux l'évolution des organismes. Les numts résultent de la transposition et de l'assimilation de marqueurs mitochondriaux dans le noyau (Blanchard & Lynch 2000). La translocation de sequences mitochondriales dans le génome nucléaire peut être très ancienne. Nous avons mis en évidence la présence de numts depuis au moins 1 million d'années chez ces puffins mais des translocations plus anciennes ont été observées chez d'autres organismes, jusqu'à 14 Ma (Lammers et al. 2017; Nacer & do Amaral 2017; Schiavo et al. 2017). L'héteroplasmie peut

résulter de transmission biparentale ou uni-parentale double de génomes mitochondriaux (voir Zouros et al. 1994) et est souvent associée à de l'hybridation (voir Barr et al. 2005, Brannock et al. 2013). Les numts peuvent être de différentes tailles (de quelques pb à 2000 kb, Hazkani-Covo et al. 2010) et certains englobent le génome mitochondrial en entier (Verscheure et al. 2015). Les duplications peuvent impliquer un génome complet (polyploidisation, e.g. Wolfe & Schields 1997), mais les duplications de gènes sont les plus fréquentes (Lynch et Conery 2000). Les numts comme les duplications géniques sont dues à des mécanismes de transposition et de recombinaison ectopique (Montgomery et al. 1987).

La présence de multiples copies de certains marqueurs peut avoir une influence directe sur les organismes eux-mêmes. Il a par exemple été montré que les numts sont responsables de maladies génétiques chez l'homme (voir Hazkani-Covo et al. 2010), de même que l'hétéroplasmie (voir Stewart & Chinnery 2015). Au contraire les multiples copies pourraient conférer un avantage sélectif. Par exemple, les gènes mitochondriaux sont responsables de la respiration cellulaire. Les Procellariiformes sont des oiseaux capables de voler sur de très grandes distances. Il est envisageable que la présence de plusieurs copies des gènes mitochondriaux chez ces oiseaux ait un impact sur leur physiologie si les deux copies sont fonctionnelles et exprimées. Le fait que la duplication soit apparemment absente chez le genre *Pelecanoides*, qui a une physiologie différente de celle des autres Procellariiformes pourrait tendre à confirmer cette hypothèse, mais reste à démontrer (Chapitre III). De tels phénomènes de surfonctionnalisation existent chez d'autres organismes, entraînant par exemple une plus grande protection contre les changements de température, ou un meilleur transport des nutriments, mais à notre connaissance rien n'a été montré pour une augmentation du métabolisme chez les oiseaux marins (voir Kondrashov 2012 pour une synthèse).

II. Impact de l'évolution moléculaire sur le signal porté par les données génétiques

1. Utilisation des numts comme marqueurs phylogénétiques

La majorité des numts détectés dans notre étude semblent résulter de translocations récentes, permettant d'inférer une phylogénie proche de celle des données mitochondriales, à l'exception d'événements remarquables par leur placement dans l'arbre ou par les longueurs de branches impliquées. Ces résultats ont été observés chez d'autres espèces (Hazkani-Covo 2009; Ko et al. 2015), parfois les séquences de numts permettent de découvrir des divergences entre clades plus précisément que les marqueurs mitochondriaux, en datant les événements de translocation (Song et al. 2013). Du fait de leur évolution complexe, notamment de l'hétérogénéité des taux de substitution et de leur transmission uniparentale, les génomes mitochondriaux peuvent porter un signal phylogénétique qui ne reflètent pas l'évolution des espèces, contrairement aux numts (Schmitz et al. 2005). Les relations phylogénétiques inférées entre les séquences de numts détectés dans notre étude peuvent être dues à de multiples événements de transpositions, à du tri incomplet des lignées, à de l'introgression récente ou à une combinaison de plusieurs de ces méncanismes. Les numts peuvent être utiles pour détecter des événements démographiques ou de flux de gènes passés (Miraldo et al. 2012) ou permettre de calculer un indice d'hybridation (Pérez et al. 2017). Les numts peuvent varier en taille (de quelques pb à 2000 kb, Hazkani-Covo et al. 2010), en abondance (de 0 à 0.1% du génome nucléaire, Richy et Leister 2004) et en similarité avec leurs paralogues

(jusqu'à être totalement éliminés, e.g. (Sheppard & Timmis 2009). L'utilisation de numts comme marqueurs phylogénétiques ne peut donc pas être faite de la même manière chez tous les taxons. De plus, de telles études devraient être réalisées en sachant que les séquences sont des numts, mais la plupart des études qui incluent des numts n'en n'ont pas conscience ce qui peut mener à plusieurs biais dans les analyses (Haran et al. 2015; Thalmann et al. 2004).

2. Impact des copies multiples sur les analyses phylogéographiques

Les problèmes de copies multiples peuvent avoir un fort impact sur les analyses de diversité et de diversification en entrainant une perte ou un bruitage de l'information portée par l'ADN. En effet, lors de l'amplification de l'ADN par PCR en présence de multiples copies d'un même marqueur, deux cas peuvent se présenter. Dans le premier cas la PCR amplifie préférentiellement une des deux copies, sans que soit déterminé laquelle des deux copies. Il est alors possible que deux copies différentes soient séquencées chez deux individus différents. Comparer ces deux séquences revient alors à comparer des paralogues ce qui biaise les résultats. Dans le second cas la PCR amplifie équitablement les deux copies. Lors du séquençage, si ces copies diffèrent, des doubles pics apparaitront dans le chromatogramme de la séquence, correspondant aux deux bases portées par les deux copies.

Ces doubles pics peuvent être traités de différentes manières. Nous avons montré que retirer pour tous les individus, les sites où les doubles-pics sont présents entrainait une perte majeure du signal de diversité et de différenciation présent dans l'alignement. Ce résultat est dû au fait que les ambigüités issues des multiples copies se situent sur les mêmes sites que les sites divergents entre les copies orthologues. Nous avons estimé que le meilleur traitement des doubles-pics revenait à estimer, pour chaque population, la proportion de séquences présentant des ambigüités et de retirer les séquences des populations dont la proportion était la plus faible. Notre étude montre qu'en dessous de 40% d'individus contaminés par des numts, les analyses ne sont pas significativement biaisées. Pour les séquences restantes, le fait de laisser les ambigüités ou de les remplacer par des « N » dans les séquences semble avoir peu d'effet sur les mesures de diversité et de différenciation, mais les ambiguïtés semblent avoir un impact sur les estimations de temps de divergence. Nous conseillons globalement de laisser les ambigüités qui apportent moins de bruit que les « N » (deux états possibles au lieu de quatre). La présence de numts ne semble pas avoir eu d'impact majeur sur les analyses de génétique chez le Procellariiformes mais les études ultérieures devront prendre des précautions sur ces problèmes, notamment car les séquences déjà soumises à Genbank peuvent comporter des numts non-repérés (2% des 7603 séquences mitochondriales de Procellariiformes présentent ainsi des ambiguités.

III. Processus à l'origine de la différenciation

1. *Considérations taxonomiques des lignées du complexe de petit puffin*

Les critères morphologiques ont désigné tour à tour les cinq lignées étudiées comme des sous-espèces de *lherminieri* (Murphy 1927), d'*assimilis* (Bourne et al. 1988) ou des synonymes de *lherminieri* ou *assimilis* (Sibley & Monroe 1990). Une étude du *cob* définit les trois lignées Atlantiques comme trois sous-espèces de *lherminieri* et les deux lignées Indo-Pacifiques comme deux sous-espèces de *bailloni*, *nicolae* étant synonymisée avec *dichrous* (Austin et al. 2004). Cette dernière taxonomie sert de référence mais est encore débattue aujourd'hui et aucun autre critère n'a été utilisé pour réaliser une étude taxonomique plus poussée depuis.

L'étude de plusieurs marqueurs moléculaires nous permet de préciser le statut taxonomique de ces lignées. Nous considérons ici le concept général d'espèce (De Queiroz 2007) et le concept d'espèce internodal (Samadi et Barberousse 2009). Les marqueurs mitochondriaux et nucléaires montrent que les lignées de petit puffin Atlantique d'une part et Indo-Pacifique d'une autre part forment des branches évolutives distinctes et peuvent être considérées comme des espèces séparées. Au sein de l'océan Atlantique, *lherminieri* apparait comme divergente de *boydi* et *baroli* tant sur les marqueurs mitochondriaux que nucléaires et peut donc être considérée comme une espèce à part entière *Puffinus lherminieri*. Les lignées *boydi* et *baroli* sont quant à elles bien distinguées par les marqueurs mitochondriaux mais pas par les marqueurs nucléaires. Nous les considérons donc comme des espèces en cours de formation, dans la zone grise de spéciation (de Queiroz et al. 2007). Les lignées de l'océan Indien *bailloni* et *nicolae* peuvent également être considérées comme des espèces à part entière en se basant sur les marqueurs moléculaires. Ces considérations sont appuyées par le fait que des estimations de distance génétique, de différenciation et de temps de divergence ont été trouvées à des niveaux comparables, voire inférieurs à notre étude entre deux espèces distinctes de Procellariiformes (e.g. Genovart et al. 2006, 2007, Gomez-Diaz et al. 2006, Gangloff et al. 2013, Welch et al. 2011).

Les données moléculaires semblent donc plus résolutives que les données morphologiques employées jusqu'ici et leur utilisation permettraient une diagnose certaine. Les données utilisées, notamment les marqueurs nucléaires peuvent ne pas être suffisamment résolutifs pour détecter une divergence entre *boydi* et *baroli* par exemple. De même nous mettons en avant la possibilité d'introgression récente entre les différentes lignées et surtout entre des lignées Atlantiques et Indiennes. La détection claire d'hybridation récente entre les différentes lignées nécessiterait de toutes les considérer comme appartenant à une seule espèce. L'ajout de nombreux marqueurs, par exemple au moyen de NGS, permettrait de lever ces ambiguïtés. L'occurrence d'hybridation devra être particulièrement scrutée afin de confirmer que l'évolution des lignées est totalement divergente. L'ajout de données morphologiques et comportementales inédites, permettrait également de mener une analyse de taxonomie intégrative plus profonde afin de préciser davantage ces statuts taxonomiques et éventuellement de proposer une nouvelle description des taxons.

2. *Facteurs de différenciation chez des organismes à haute capacité de dispersion*

Les puffins sont capables de voler sur des centaines, voire des milliers de kilomètres (e.g. 1500 km pour *Puffinus*, Ramos et al. in prep, 4000 km pour *Calonectris* Paiva et al. 2013), pourtant nous avons mis en évidence une différenciation nette entre des lignées séparées par à peine 2000 km (*boydi* et *baroli* en Atlantique-est, *bailloni* et *nicolae* dans l'océan Indien) et en l'absence de toute barrière physique apparente. Ce patron a déjà été observé chez d'autres

Procellariiformes (Gangloff et al. 2013; Gómez-Díaz et al. 2009; Rayner et al. 2011). Ce résultat, en accord avec des résultats similaires chez d'autres espèces, montre qu'une distinction entre les modes de spéciation basée uniquement sur la géographie (allopatrique Vs sympatrique) n'est pas suffisante. De nombreux autres organismes à fortes capacités de dispersion montrent peu ou pas de structure génétique entre popoulations là où les puffins en montrent. Les flux de gènes entre les localités Atlantiques sont donc possibles pour les organismes à haute capacité de dispersion, mais sont atténués chez les puffins, mais aussi d'autres Procellariiformes (Gangloff et al. 2012, Gómez-díaz et al. 2009). Ces flux de gènes restreints ne sont pas dus à des barrières géographiques et il faut en chercher la cause ailleurs.

Les puffins sont considérés comme fortement philopatriques (Rabouam et al. 1998; Wooler et al. 1990) pourtant nous n'avons pas trouvé de ségrégation génétique au sein de chaque lignée. Une forte philopatrie conduirait à une structuration claire au sein de chaque lignée entre les différentes colonies, similaire à celle que l'on trouve chez les organismes terrestres dans les mêmes archipels (AlmalkI et al. 2017; Arnold et al. 2008; Brehm et al. 2003; Patel et al. 2011). Un tel patron de structuration est trouvé chez des Procellariiformes qui ont divergé plus récemment que les petits puffins (par exemple *Pterodroma sandwichensis* dans l'archipel d'Hawaï avec des marqueurs mitochondriaux Welch, Fleischer, et al. 2012). Une forte philopatrie aurait donc un impact visible de nos jours sur la structuration des puffins. On peut donc penser que les petits puffins ne sont pas aussi fortement philopatriques que ce que l'on pensait. Par ailleurs plusieurs espèces d'oiseaux marins philopatriques ne montrent pas de structuration génétique du fait de comportements de nourrissage particuliers (e.g. Burg & Croxall 2001).

Les flux de gènes de nombreux animaux marins sont façonnés par la température d'eau de mer (SST), c'est par exemple le cas chez les Mysticetes où les routes migratoires et la fidélité au site sont liées à la température de la mer et dirigent la structuration entre bassins océaniques (Jackson et al. 2014, Alexander et al. 2016, Richard et al. 2018). Pour les Odontocètes la structuration est reliée aux conditions écologiques telles que la SST ou la profondeur (Fontaine et al. 2014, Viricel & Rosel 2014). La SST dirige également la diversification des tortues de mer, puisque seules les espèces adaptées au froid sont capables d'échanger des gènes entre océans (Dutton et al. 1999). Enfin, la structuration des organismes à dispersion larvaire pélagique est souvent liée à la SST (Teske et al. 2005, Benestan et al. 2015, Teske et al. 2018). Nous avons montré une correlation entre variations des niveaux de température à une échelle globale, lors des 3 derniers Ma, et la structuration génétique des puffins, bien que celle-ci soit considérablement plus forte que chez les autres organismes marins (voir Bowen et al. 2016 pour une synthèse). L'impact fort des conditions climatiques et océaniques sur la différenciation des puffins s'explique par la forte dépendance de ceux-ci à la fois au milieu terrestre et océanique. Les oiseaux marins sont très sensibles à tout changement de conditions sur les îles qui leur servent de lieu de reproduction. Ainsi une forte hausse du niveau de la mer pourra les forcer à trouver de nouveaux sites de reproduction et une brusque baisse des températures pourra leur interdire le passage entre deux points auparavant reliés (e.g. Silva et al. 2015 chez les Procellariiformes, Friesen et al. 1996, Liebers et al. 2004, Ritchie et al. 2004 chez d'autres oiseaux marins).

La ségrégation des zones de nourrissage entre populations est également un facteur de différenciation des oiseaux marins (Friesen 2015, Friesen et al. 2007a) et peut être expliquée par le besoin de limiter la compétition pour une ressource donnée (Cairns 1989; Wakefield et

al. 2013). La compétition intra-spécifique peut engendrer un coût qui implique de limiter les déplacements, ce qui favorise la ségrégation de l'utilisation de l'habitat. Cette ségrégation peut être spatiale (González-Solís et al. 2007), temporelle (Friesen et al. 2007a) ou alimentaire (Wilson 2010). De plus chez les petits puffins, les individus d'une même population ont tendance à exploiter une zone de nourrissage commune. Cette homogénéisation intra-population, couplée à la ségrégation inter-lignées, pourrait avoir favorisé l'adaptation de chaque lignée à des conditions océaniques précises, ce qui contribuerait au processus de différenciation. Ainsi les espèces proches d'oiseaux marins ont souvent des zones de nourrissage différentes (e.g. Patterson et al. 2011). Ce phénomène est d'ailleurs retrouvé chez d'autres oiseaux qui montrent des routes de migration différente et des différences morphologiques (e.g. Rolshausen et al. 2009). L'impact de la SST et de la ségrégation des zones de nourrissage sont tous les deux liés au fait que les puffins, comme tous les oiseaux marins, sont des « central-place foragers », c'est-à-dire qu'ils sont dépendants d'habitats terrestres pour la reproduction. Nous avons d'ailleurs montré que ces différents facteurs avaient mené les petits puffins à un scénario de divergence semblable à celui retrouvé chez des organismes à faible capacité de dispersion, comme des hippocampes (Floeter et al. 2007). Le Cape Agulhas au sud de l'Afrique est d'ailleurs connu pour être un point de rupture biogéographique entre plusieurs espèces côtières, du fait de la différence de courants et de température entre les deux océans (voir Teske et al. 2011 pour une synthèse).

IV. Conclusion

Nous avons mis en évidence plusieurs phénomènes qui influencent directement la composition des séquences d'ADN. La discordance mito-nucléaire, les numts, l'hétéroplasmie et les duplications résultent d'une évolution complexe des molécules d'ADN à différents niveaux. Nous avons montré que la discordance mito-nucléaire observée était le signe de tri incomplet des lignées, d'hybridation et/ou d'une dispersion femelle-biaisée. Les phénomènes de copies multiples sont le résultat de processus de transposition de certaines parties de l'ADN ou de transmission de deux copies de mitogénomes. Tous ces processus causaux ont eux-mêmes une influence sur la diversité et la différenciation génétique des populations qui mérite d'être étudiée. Ils entrainent aussi un bruitage des données génétiques. Nous avons montré quel était le traitement optimal à effectuer afin de gérer au mieux un tel bruitage si l'éviter n'était pas possible.

Nous avons ensuite montré comment utiliser au mieux ces données génétiques afin de reconstituer le scénario de diversification des populations. La reconstitution d'un tel scénario nous apprend d'abord comment sont structurées les populations actuelles, quel est leur niveau de diversité et de diversification. L'histoire évolutive des populations nous apprend également quels sont les principaux facteurs de diversification chez les espèces étudiées. Nous avons ainsi mis en avant le faible impact de l'absence de barrière physique, de la distance géographique et de la supposée philopatrie sur la différenciation des organismes à haute capacité de dispersion. Nous avons au contraire mis en exergue l'importance de l'écologie, de la démographie et du comportement chez toutes ces espèces.

V. Perspectives

Pour étudier plus en profondeur l'impact de la région mitochondriale dupliquée sur la biologie des espèces et les analyses de génétique, la première étape serait de compléter l'étude menée dans le chapitre III afin de reconstruire la duplication complète chez plusieurs taxons le long de la phylogénie des Procellariiformes en utilisant une méthode de séquençage par longs fragments. Ce type de séquençage NGS permettrait de reconstruire l'évolution de la région mitochondriale dupliquée le long de la phylogénie des Procellariiformes puis d'analyser pour chaque copie de chaque marqueur dupliqué pour chaque taxon, son niveau d'expression et de sélection. Cette étude apporterait des informations sur (1) les processus qui sont à l'origine de l'apparition et la propagation de cette région dupliquée, (2) les mécanismes de conservation ou de divergence de cette région au sein de l'ordre des Procellariiformes et (3) la corrélation entre la présence et la composition de cette région et les performances physiologiques. Enfin l'évolution de cette région dupliquée pourrait également être associée à des événements de spéciation, comme cela a été vu chez la drosophile par exemple (Ting et al. 2004). Cet exemple précis pourrait être appliqué aux mêmes phénomènes de multiples copies chez d'autres organismes.

A propos de la divergence des populations de puffins, l'utilisation des NGS qui fournissent un plus grand nombre de marqueurs et donc d'information, permettrait de confirmer ou d'infirmer le scénario de différenciation trouvé. Un plus grand nombre de marqueurs permettrait une meilleure puissance statistique sur les analyses de reconstruction de scénario, d'évolution de taille de populations et les analyses de migration. Un grand nombres de marqueurs nucléaires permettraient également de différencier l'introgression récente de la variation ancestrale retenue comme cause de discordance mito-nucléaire. L'ajout de nombreux marqueurs ainsi obtenus permettrait par exemple de déterminer avec plus de certitude le statut de populations hybrides (Twyford & Ennos 2011). Une combinaison de séquençage NGS à courts et longs fragments, permettrait également de supprimer les ambigüités dues aux multiples-copies. Par exemple la combinaison de courts et longs fragments permetrait de déterminer à la fois la position et la composition précise de deux copies d'un même marqueur, comme nous l'avons fait pour la région dupliquée de *Puffinus* (Chapitre II). Ces séquences pourraient permettre ensuite l'obtention d'amorces spécifiques à chaque copie et ainsi d'éviter les problèmes de multiple copie en séquençage Sanger.

L'oscillation de la température de la mer est probablement liée à la divergence de nombreuses espèces d'oiseaux marins. Des études similaires à la nôtre pourraient être faites sur des taxons présentant différents niveaux de différenciation morphologique, comportementale ou génétique afin de déterminer l'importance de la température par rapport à d'autres facteurs biologiques. Le puffin fouquet (*Puffinus pacificus*) par exemple présente une aire de répartition semblable aux espèces du complexe du petit puffin (Marchant & Higgins 1990), est apparue pendant la même période (Welch et al. 2014) et a probablement subi les mêmes oscillations climatiques mais est monomorphe (aucune sous espèce décrite). Une étude génétique permettrait de comprendre si des flux de gènes entre les populations de puffin ont été maintenus entre les différentes populations de puffin fouquet et si c'est le cas pourquoi est-ce différent des espèces du complexe de petit puffin. De la même manière différentes espèces de Procellariiformes sont présents dans l'est de l'Atlantique, montrant des niveaux de différenciation plus forts que chez les puffins. Par exemple les différentes populations de ptérodromes de Macaronésie sont clairement différenciées (Gangloff et al. 2013) et des populations allochroniques d'océanites sont également observées dans cette région (Friesen et

al. 2007b). Il serait intéressant de comprendre quels facteurs sont à l'origine d'une différenciation de ces populations plus fortes que chez les puffins. Par ailleurs, ces changements de températures en mer ont été associés à des changements de température sur terre. Il a été montré que les changements de conditions climatiques ont théoriquement un impact sur la diversification des organismes terrestres (Hua & Wiens 2014) et des preuves vont dans ce sens (e.g. Demenocal 2004) mais le lien entre changement climatique et diversification terrestre doit encore être approfondi, par exemple par une étude phylogéographique similaire à la nôtre.

Enfin il a été montre chez des poissons d'Afrique du Sud que des gènes nucléaires associés aux températures montraient une structuration génétique absente chez les marqueurs mitochondriaux (Teske et al. 2018). L'utilisation de marqueurs similaires pourrait confirmer une structuration liée aux températures chez les puffins. Si les différentes lignées de puffins et les environnements dans lesquels ils évoluent nous semblent similaires, des différences morphométriques, de coloration et de vocalisations peuvent être observées entre les lignées. Il serait intéressant d'étudier plus en profondeur ces différences, de les comparer à la différenciation génétique et à des données écologiques afin de déterminer un impact éventuel de la sélection naturelle ou de la sélection sexuelle. L'ajout de données de suivi permettrait également de constater avec plus de certitude la possibilité d'hybridation actuelle entre les différentes lignées (e.g. Brown et al. 2015), et/ou une philopatrie plus forte chez un des deux sexes que chez l'autre. Des données phénotypiques pourraient éventuellement permettre de distinguer des hybrides, morphologiquement intermédiaires (Seehausen 2004).

Références bibliographiques

Abbott CL, Double MC, Trueman JWH, Robinson A, Cockburn A. 2005. An unusual source of apparent mitochondrial heteroplasmy: Duplicate mitochondrial control regions in Thalassarche albatrosses. *Mol. Ecol.* 14(11):3605–13

Akiyama T, Nishida C, Momose K, Onuma M, Takami K, Masuda R. 2016. Gene duplication and concerted evolution of mitochondrial DNA in crane species. *Mol. Phylogenet. Evol.* 106:158–63

Alderman R, Double MC, Valencia J, Gales RP. 2005. Genetic affinities of newly sampled populations of Wandering and Black-browed Albatross. *Emu*. 105:169–79

Alexander A, Steel D, Hoekzema K, Mesnick SL, Engelhaupt D, et al. 2016. What influences the worldwide genetic structure of sperm whales (Physeter macrocephalus)? *Mol. Ecol.* 25(12):2754–72

AlmalkI M, Kupán K, Carmona-Isunza MC, López P, Veiga A, et al. 2017. Morphological and Genetic Differentiation Among Kentish Plover *Charadrius alexandrinus* Populations in Macaronesia. *Ardeola*. 64(1):3–16

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215(3):403–10

Archer E, Adams P, Schneiders B, Fernandez S, Asfazadour W. 2017. Package ' strataG '

Arctander P. 1995. Comparison of a mitochondrial gene and a corresponding nuclear pseudogene. *Proc. R. Soc. B Biol. Sci.* 262:13–19

Arctander P, B PRSL. 1995. Comparison of a Mitochondrial Gene and a corresponding Nuclear Pseudogene. . 13–19

Arnold EN, Vasconcelos R, Harris DJ, Mateo JA, Carranza S. 2008. Systematics, biogeography and evolution of the endemic Hemidactylus geckos (Reptilia, Squamata, Gekkonidae) of the Cape Verde Islands: Based on morphology and mitochondrial and nuclear DNA sequences. *Zool. Scr.* 37(6):619–36

Austin JJ, Bretagnolle V, Pasquet E. 2004. A global molecular phylogeny of the small Puffinus shearwaters and implications for systematics of the Little-Audubon's Shearwater complex. *Auk.* 121(3):647–864

Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T, et al. 1987. Intraspecific Phylogeography; the Mitochondrial DNA Bridge Between Population Genetics and Systematics. *Annu. Rev. Ecol. Syst.* 18:489–522

Baer CF, Miyamoto MM, Denver DR. 2007. Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nat. Rev. Genet.* 8(8):619

Baker AJ, Huynen LJ, Haddrath O, Millar CD, Lambert DM. 2005. Reconstructing the tempo and mode of evolution in an extinct clade of birds with ancient DNA : The giant moas of New Zealand. *PNAS*. 102(23):8257–62

Balloux F, Brünner H, Lugon-Moulin N, Hausser J, Goudet J. 2000. Microsatellites can be misleading: An emprical and simulation study. *Evolution (N. Y).* 54(4):1414–22

Barr CM, Neiman M, Taylor DR. 2005. Inheritance and recombination of mitochondrial genomes in plants, fungi and animals. *New Phytol.* 168(1):39–50

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. *Genetics*. 162(4):2025–35

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, et al. 2004. Ultraconserved Elements in the Human Genome. *Science (80-.).* 304(May):1321–25

Belkhir K, Porsa P, Chikhi L, Raufaste N, Bonhomme F. 1996. *GENETIX 4.05, Logiciel Sous Windows TM Pour La Génétique Des Populations.*

Bell DB, Jung SJA, Kroon D. 2015. The Plio-Pleistocene development of Atlantic deep-water circulation and its influence on climate trends. *Quat. Sci. Rev.* 123:265–82

Belle EMS, Piganeau G, Gardner M, Eyre-walker A. 2005. An investigation of the variation in the transition bias among various animal mitochondrial DNA. *Gene*. 355:58–66

Bellis C, Ashton KJ, Freney L, Blair B, Griffiths LR. 2003. A molecular genetic approach for forensic animal species identification. *Forensinc Sci. Int.* 134:99–108

Benestan L, Gosselin T, Perrier C, Sainte-Marie B, Rochette R, Bernatchez L. 2015. RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (Homarus americanus). *Mol. Ecol.* 24(13):3299–3315

Bensasson D, Zhang DX, Hartl DL, Hewitt GM. 2001. Mitochondrial pseudogenes: Evolution's misplaced witnesses. *Trends Ecol. Evol.* 16(6):314–21

Berg T, Moum T, Johansen S. 1995. Variable numbers of simple tandem repeats make birds of the order Ciconiiformes heteroplasmic in their mitochondrial genomes. *Curr. Genet.* 27(3):257–62

Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, et al. 2013. MITOS: Improved de novo metazoan mitochondrial genome annotation. *Mol. Phylogenet. Evol.* 69(2):313–19

Bertheau C, Schuler H, Krumböck S, Arthofer W, Stauffer C. 2011. Hit or miss in phylogeographic analyses: The case of the cryptic NUMTs. *Mol. Ecol. Resour.* 11(6):1056–59

Bicknell AWJ, Knight ME, Bilton D, Reid JB, Burke T, Votier SC. 2012. Population genetic structure and long-distance dispersal among seabird populations: Implications for colony persistence. *Mol. Ecol.* 21(12):2863–76

Blanchard JL, Lynch M. 2000. Organellar genes why do they edd up in the nucleus? *Tig*. 16(7):315–20
Boore JL. 1999. Animal mitochondrial genomes. *Nucleic Acids Res.* 27(8):1767–80
Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, et al. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput. Biol.* 10(4):e1003537
Bourne WRP, Mackrill EJ, Paterson a M, Yesou P. 1988. The Yelkouan Shearwater Puffinus (puffinus?) yelkouan. *Br. Birds*. 817260(July):306–19
Bowen BW, Gaither MR, DiBattista JD, Iacchei M, Andrews KR, et al. 2016. Comparative phylogeography of the ocean planet. *Proc. Natl. Acad. Sci.* 113(29):7962–69
Brace S, Barnes I, Kitchener AC, Serjeantson D, Turvey ST. 2014. Late Holocene range collapse in a former British seabird species. *J. Biogeogr.* 41(8):1583–89
Brehm A, Jesus J, Spínola H, Alves C, Vicente L, Harris DJ. 2003. Phylogeography of the Madeiran endemic lizard Lacerta dugesii inferred from mtDNA sequences. *Mol. Phylogenet. Evol.* 26(2):222–30
Bretagnolle V, Attie C, Mougeot F. 2000. Audubon's Shearwaters Puffinus lherminieri on Réunion Island, Indian Ocean: Behaviour, census, distribution, biometrics and breeding biology. *Ibis (Lond. 1859).* 142:399–412
Brooke M. 2004. *Albatrosses and Petrels across the World*. Oxford University Press
Brown RM, Jordan WC, Faulkes CG, Jones CG, Bugoni L, et al. 2011. Phylogenetic relationships in Pterodroma petrels are obscured by recent secondary contact and hybridization. *PLoS One*. 6(5):e20350
Brown RM, Nichols RA, Faulkes CG, Jones CG, Bugoni L, et al. 2010. Range expansion and hybridization in Round Island petrels (Pterodroma spp.): evidence from microsatellite genotypes. *Mol. Ecol.* 19(15):3157–70
Brown RM, Techow NMSM, Wood AG, Phillips R a. 2015. Hybridization and Back-Crossing in Giant Petrels (Macronectes giganteus and M. halli) at Bird Island, South Georgia, and a Summary of Hybridization in Seabirds. *PLoS One*. 10(3):e0121688
Brown WM. 1985. The mitochondrial genome of animals. In *Molecular Evolutionary Genetics*
Bull V, Beltrán M, Jiggins CD, McMillan WO, Bermingham E, Mallet J. 2006. Polyphyly and gene flow between non-sibling Heliconius species. *BMC Biol.* 4:1–17
Burg TM, Bird H, Lait L, de M. 2014. Colonization pathways of the northeast Atlantic by northern fulmars: A test of James Fisher's "out of Iceland" hypothesis using museum collections. *J. Avian Biol.* 45(3):209–18
Burg TM, Croxall JP. 2001. Global relationships amongst black-browed and grey-headed albatrosses: analysis of population structure using mitochondrial DNA and microsatellites. *Mol. Ecol.* 10:2647–60
Cagnon C, Lauga B, Hémery G, Mouchès C. 2004. Phylogeographic differentiation of storm petrels (Hydrobates pelagicus) based on cytochrome b mitochondrial DNA variation. *Mar. Biol.* 145(6):1257–64
Cairns DK. 1989. The Regulation of Seabird Colony Size: A Hinterland Model. *Am. Nat.* 134(1):141–46
Calvignac S, Konecny L, Malard F, Douady CJ. 2011. Preventing the pollution of mitochondrial datasets with nuclear mitochondrial paralogs (numts). *Mitochondrion*. 11(2):246–54
Case TJ, Taper ML. 1986. On the Coexistence and Coevolution of Asexual and Sexual Competitors. *Evolution (N. Y).*
Cavalli-Sforza LL, Bodmer W. 1971. Human population genetics. *San Fr. CA Free.*
Chambers GK, Moeke C, Steel R, Trueman JWH. 2009. Phylogenetic analysis of the 24 named albatross taxa based on full mitochondrial cytochrome b DNA sequences. *Notornis*. 56:82–94
Chang P, Saravan R, Ji L, Heger GC. 2000. The Effect of Local Sea Surface Temperatures on Atmospheric Circulation over the Tropical Atlantic Sector. *J. Clim.* (1985):2195–2216
Charif D, Lobry JR. 2007. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In *Structural Approaches to Sequence Evolution*, pp. 207–32. Springer
Cibois A, Thibault J, Lecroy M, Bretagnolle V. 2015. Molecular analysis of a storm petrel specimen from the Marquesas Islands, with comments on specimens of Fregetta lineata and F. guttata. *Bull. Br. Ornithol. Club*. 135(3):240–46
Connan M, Kelly CMR, Mcquaid CD, Bonnevie BT, Barker NP. 2011. Morphological versus molecular identification of Sooty (Phoebetria fusca) and Light-mantled (P. palpebrata) albatross chicks. *Polar Biol.* 34:791–98
Cornuet JM, Pudlo P, Veyssier J, Dehne-Garcia A, Gautier M, et al. 2014. DIYABC v2.0: A software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics*. 30(8):1187–89
Crandall KA, Bininda-emonds ORP, Mace GM, Wayne RK. 2000. Considering evolutionary processes in conservation biology. *Trends Ecol. Evol.* 15(7):290–95
Cristiano M, Fernandes-salomão T, Yotoko K. 2012. Nuclear mitochondrial DNA : an Achilles'heel of molecular systematics, phylogenetics, and phylogeographic studies of stingless bees. *Apidologie*. 43:527–38
Crochet P-A, Chen JZ, Pons JM, Lebreton J-D, Hebert PDN, Bonhomme F. 2003. Genetic Differentiation At Nuclear and Mitochondrial Loci Among Large White-Headed Gulls : Sex-Biased Interspecific Gene

Flow ? *Evolution (N. Y).* 57(12):2865–78
Croxall JP. 2005. Global Circumnavigations: Tracking Year-Round Ranges of Nonbreeding Albatrosses. *Science (80-.).* 307(5707):249–50
Cunningham CW. 1999. Some Limitations of Ancestral Character-State Reconstruction When. *Syst. Biol.* 48(3):665–74
Darriba D, Taboada GL, Doallo R, Posada D. 2015. jModelTest 2 : more models , new heuristics and high-performance computing. *Nat. Methods*. 9(8):6–9
Darwin C. 1859. L ' Origine des espèces
Darwin C. 1871. *The Descent of Man*
Dayama G, Emery SB, Kidd JM, Mills RE. 2014. The genomic landscape of polymorphic human nuclear mitochondrial insertions. *Nucleic Acids Res.* 42(20):12640–49
Deane P. 2013. *What traits predispose the Band-rumped Storm-petrel, Oceanodroma castro, to ecological speciation inthe absence of physical barriers to gene flow?*
Demenocal PB. 2004. African climate change and faunal evolution during the Pliocene-Pleistocene African climate change and faunal evolution during the Pliocene ^ Pleistocene. *Earth Planet. Sci. Lett.* (March 2004):
Desalle R, Schierwater B, Hadrys H. 2017. MtDNA : The small workhorse of evolutionary studies. *Front. Biosci.* 22:873–87
Díaz-Jaimes P, Uribe-Alcocer M, Rocha-Olivares A, García-de-León FJ, Nortmoon P, Durand JD. 2010. Global phylogeography of the dolphinfish (Coryphaena hippurus): The influence of large effective population size and recent dispersal on the divergence of a marine pelagic cosmopolitan species. *Mol. Phylogenet. Evol.* 57(3):1209–18
Dobzhansky T. 1937. Genetic nature of species differences. *Am. Nat.* 71(735):404–20
Dudoit 'Ale'alani, Iacchei M, Coleman RR, Gaither MR, Browne WE, et al. 2018. The little shrimp that could: phylogeography of the circumtropical *Stenopus hispidus* (Crustacea: Decapoda), reveals divergent Atlantic and Pacific lineages. *PeerJ*. 6:e4409
Dutton PH, Bowen BW, Owens DW, Barragan A, Davis SK. 1999. Global phylogeography of the leatherback turtle (Dermochelys coriacea). *J. Zool.* 248(3):397–409
Eda M, Kuro-o M, Higuchi H, Hasegawa H, Koike H. 2010. Mosaic gene conversion after a tandem duplication of mtDNA sequence in Diomedeidae (albatrosses). *Genes Genet. Syst.* 85(2):129–39
Ehara M, Watanabe KI, Ohama T. 2000. Distribution of cognates of group II introns detected in mitochondrial cox1 genes of a diatom and a haptophyte. *Gene*. 256(1–2):157–67
Ely B, Viñas J, Alvarado Bremer JR, Black D, Lucas L, et al. 2005. Consequences of the historical demography on the global population structure of two highly migratory cosmopolitan marine fishes: The yellowfin tuna (Thunnus albacares) and the skipjack tuna (Katsuwonus pelamis). *BMC Evol. Biol.* 5:1–9
Emblem A, Karlsen BO, Evertsen J, Johansen SD. 2011. Mitogenome rearrangement in the cold-water scleractinian coral Lophelia pertusa (Cnidaria, Anthozoa) involves a long-term evolving group I intron. *Mol. Phylogenet. Evol.* 61(2):495–503
Estoup A, Lombaert E, Marin J-M, Guillemaud T, Puldo P, et al. 2012. Estimation of demo-genetic model probabilities with Approximate Bayesian Computation using linear discriminant analysis on summary statistics. *Mol. Ecol. Resour.* 12(5):846–55
Excoffier L, Laval G, Schneider S. 2005. Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evol. Bioinforma.* 1:47–50
Floeter SR, Rocha L., Roberston DR, Joyeux JC, Smith-Vaniz WF, et al. 2007. Atlantic reef fish biogeography and evolution. *J. Biogeogr.* 35(1):
Fontaine MC, Roland K, Calves I, Austerlitz F, Palstra FP, et al. 2014. Postglacial climate changes and rise of three ecotypes of harbour porpoises, Phocoena phocoena, in western Palearctic waters. *Mol. Ecol.* 23(13):3306–21
Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*
Frankham R, Ballou JD, Briscoe DA. 2004. *A Primer of Conservation Genetics*. Cambridge University Press
Frankham R, Bradshaw CJA, Brook BW. 2014. Genetics in conservation management: revised recommendations for the 50/500 rules, Red List criteria and population viability analyses. *Biol. Conserv.* 170:56–63
Franklin IR, Frankham R. 1998. How large must populations be to retain evolutionary potential? *Anim. Conserv.* 1:69–73
Fridolfsson A-K, Ellegren H. 1999. A simple and universal method for molecular sexing of non-ratite birds. *J. avian Biol.* 116–21
Friesen VL. 2015. Speciation in seabirds: why are there so many species…and why aren't there more? *J. Ornithol.*
Friesen VL, Baker AJ, Piatt JF. 1996. Phylogenetic relationships within the Alcidae (charadriiformes: Aves) inferred from total molecular evidence. *Mol. Biol. Evol.* 13(2):359–67

Friesen VL, Burg TM, McCOY KD. 2007a. Mechanisms of population differentiation in seabirds. *Mol. Ecol.* 16(9):1765–85

Friesen VL, Smith AL, Gómez-Díaz E, Bolton M, Furness RW, et al. 2007b. Sympatric speciation by allochrony in a seabird. *Proc. Natl. Acad. Sci. U. S. A.* 104(47):18589–94

Fu YX. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*. 147(2):915–25

Funk DJ, Omland KE. 2003. Species-Level Paraphyly and Polyphyly: Frequency, Causes, and Consequences, with Insights from Animal Mitochondrial DNA. *Annu. Rev. Ecol. Evol. Syst.* 34(1):397–423

Gandolfi A, Crestanello B, Fagotti A, Simoncelli F, Chiesa S, et al. 2017. New Evidences of Mitochondrial DNA Heteroplasmy by Putative Paternal Leakage between the Rock Partridge (Alectoris graeca) and the Chukar Partridge (Alectoris chukar). *PLoS One*. 12(1):4–11

Gangloff B, Shirihai H, Watling D, Cruaud C, Couloux A, et al. 2012. The complete phylogeny of Pseudobulweria, the most endangered seabird genus: systematics, species status and conservation implications. *Conserv. Genet.* 13:39–52

Gangloff B, Zino F, Shirihai H, González-Solís J, Couloux A, et al. 2013. The evolution of north-east Atlantic gadfly petrels using statistical phylogeography. *Mol. Ecol.* 22(January):495–507

Gay L, Neubauer G, Zagalska-Neubauer M, Pons JM, Bell DA, Crochet PA. 2009. Speciation with gene flow in the large white-headed gulls: Does selection counterbalance introgression? *Heredity (Edinb).* 102(2):133–46

Genovart M, Juste J, Contreras-Diaz H, Oro D. 2012. Genetic and Phenotypic Differentiation between the Critically Endangered Balearic Shearwater and Neighboring Colonies of Its Sibling Species. *J. Hered.* 103(3):330–41

Genovart M, Oro D, Juste J, Bertorelle G. 2007. What genetics tell us about the conservation of the critically endangered Balearic Shearwater? *Biol. Conserv.* 137(2):283–93

Gibb GC, Kardailsky O, Kimball RT, Braun EL, Penny D. 2007. Mitochondrial genomes and avian phylogeny: Complex characters and resolvability without explosive radiations. *Mol. Biol. Evol.* 24(1):269–80

Gibb GC, Kennedy M, Penny D. 2013. Beyond phylogeny: Pelecaniform and ciconiiform birds, and long-term niche stability. *Mol. Phylogenet. Evol.* 68(2):229–38

Gillot P-Y, Nativel P. 1989. Eruptive history of the Piton de la Fournaise volcano, Reunion Island, Indian Ocean. *J. Volcanol. Geotherm. Res.* 36(1–3):53–65

Gissi C, Iannelli F, Pesole G. 2008. Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. *Heredity (Edinb).* 101:301–20

Gómez-Díaz E, González-Solís J, Peinado MA. 2009. Population structure in a highly pelagic seabird, the Cory's shearwater Calonectris diomedea: an examination of genetics, morphology and ecology. *Mar. Ecol. Prog. Ser.* 382(Palumbi 1994):197–209

Gómez-Díaz E, González-Solís J, Peinado MA, Page RDM. 2006. Phylogeography of the Calonectris shearwaters using molecular and morphometric data. *Mol. Phylogenet. Evol.* 41(2):322–32

González-Solís J, Croxall JP, Oro D, Ruiz X. 2007. Trans-equatorial migration and mixing in the wintering areas of a pelagic seabird. *Front. Ecol. Environ.* 5(6):297–301

Goudet J. 2005. HIERFSTAT , a package for R to compute and test hierarchical F -statistics. *Mol. Ecol. Notes*. 5:184–86

Goudet J, Perrin N, Waser P. 2002. Tests for sex-biased dispersal using bi-parentally inherited genetic markers. *Mol. Ecol.* 11(6):1103–14

Groth JG, Barrowclough GF. 1999. Basal Divergences in Birds and the Phylogenetic Utility of the Nuclear RAG-1 Gene. *Mol. Phylogenet. Evol.* 12(2):115–23

Guerrini M, Panayides P, Hadjigerou P, Taglioli L, Dini F, Barbanera F. 2007. Lack of genetic structure of cypriot alectoris chukar (Aves, Galliformes) populations as infered from mtDNA sequencing data. *Anim. Biodivers. Conserv.* 30(1):105–14

Haran J, Koutroumpa F, Magnoux E, Roques A, Roux G. 2015. Ghost mtDNA haplotypes generated by fortuitous NUMTs can deeply disturb infra-specific genetic diversity and phylogeographic pattern. *J. Zool. Syst. Evol. Res.* 53(2):109–15

Harrisson P. 2000. *Seabirds: An Identification Guide (No. QL 673. H38 2000).* Houghton Mifflin CO., New York

Hassanin A, Bonillo Lc, Nguyen BX, Cruaud C. 2010. Comparisons between mitochondrial genomes of domestic goat (Capra hircus) reveal the presence of numts and multiple sequencing errors. *Mitochondrial DNA*. 21(August):68–76

Hazkani-Covo E. 2009. Mitochondrial insertions into primate nuclear genomes suggest the use of numts as a tool for phylogeny. *Mol. Biol. Evol.* 26(10):2175–79

Hazkani-Covo E, Zeller RM, Martin W. 2010. Molecular poltergeists: Mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet.* 6(2):

Heather JM, Chain B. 2016. The sequence of sequencers : The history of sequencing DNA. *Genomics*. 107(1):1–8

Hedrick PW. 2010. Cattle ancestry in bison: Eexplanations for higher mtDNA than autosomal ancestry. *Mol. Ecol.* 19(16):3328–35
Heibl C. 2008. PHYLOCH: R language tree plotting tools and interfaces to diverse phylogenetic software packages. http://www.christophheibl.de/Rpackages.html
Heidrich P, Amengual J, Wink M. 1997. Phylogenetic relationships in mediterranean and North Atlantic shearwaters (Aves: Procellariidae) based on nucleotide sequences of mtDNA. *Biochem. Syst. Ecol.* 26(2):145–70
Henderson N, Gill BJ. 2010. A mid-Pliocene shearwater skull (Aves : Procellariidae : Puffinus) from the Taihape Mudstone , central North Island , New Zealand. *New Zeal. J. Geol. Geophys.* 53(4):
Herbert TD, Years PM, Herbert TD, Peterson LC, Lawrence KT, Liu Z. 2011. Tropical Ocean Temperatures Over the Past 3.5 Million Years. . 1530(2010):
Hill GE. 2017. The mitonuclear compatibility species concept. *Auk*. 134(Mayr 1940):393–409
Hua X, Wiens JJ. 2014. How Does Climate Influence Speciation ? *Am. Nat.* 182(1):1–12
Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23(2):254–67
Huson DH, Scornavacca C. 2012. Dendroscope 3 : An Interactive Tool for Rooted Phylogenetic Trees and Networks. *Syst. Biol.* 61(6):1061–67
Jackson JA, Steel DJ, Beerli P, Congdon BC, Olavarria C, et al. 2014. Global diversity and oceanic divergence of humpback whales (Megaptera novaeangliae). *Proc. R. Soc. B Biol. Sci.* 281(1786):20133222–20133222
Jain M, Olsen HE, Paten B, Akeson M. 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* 17(1):256
Jamieson IG, Allendorf FW. 2012. How does the 50 / 500 rule apply to MVPs ? *Trends Ecol. Evol.* 27(10):
Jayaprakash AD, Benson EK, Gone S, Liang R, Shim J, et al. 2015. Stable heteroplasmy at the single-cell level is facilitated by intercellular exchange of mtDNA. *Nucleic Acids Res.* 43(4):2177–87
Jesus J, Menezes D, Gomes S, Oliveira P, Nogales M, Brehm A. 2009. Phylogenetic relationships of gadfly petrels Pterodroma spp. from the Northeastern Atlantic Ocean: molecular evidence for specific status of Bugio and Cape Verde petrels and implications for conservation. *Bird Conserv. Int.* 19(03):199
Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30(14):3059–66
Kawakami K, Eda M, Horikoshi K, Suzuki H, Chiba H, Hiraoka T. 2012. Bryan's shearwaters have survived on th Bonin islands northwestern Pacific. *Condor*. 114(3):507–12
Kawakami K, Eda M, Izumi H, Horikoshi K, Suzuki H. 2018. Phylogenetic Position of Endangered *Puffinus lherminieri bannermani*. *Ornithol. Sci.* 17(1):11–18
Kelly RP, Palumbi SR. 2010. Genetic structure among 50 species of the northeastern pacific rocky intertidal community. *PLoS One*. 5(1):
Kerr KCR, Dove CJ. 2013. Delimiting shades of gray: phylogeography of the Northern Fulmar, Fulmarus glacialis. *Ecol. Evol.* 3(7):1915–30
Kerr KCR, Stoeckle MY, Dove CJ, Weigt LA, Francis CM, Hebert PDN. 2007. Comprehensive DNA barcode coverage of North American birds. *Mol. Ecol. Notes.* 7(4):535–43
Kimball RT, Braun EL, Barker FK, Bowie RCK, Braun MJ, et al. 2009. A well-tested set of primers to amplify regions spread across the avian genome. *Mol. Phylogenet. Evol.* 50(3):654–60
Kimball RT, Braun EL, Ligon JD. 1997. Resolution of the phylogenetic position of the Congo peafowl , Afropavo congensis : a biogeographic and evolutionary enigma. *Proc. R. Soc. London B Biol. Sci.* 264(1387):1517–23
Kimura M. 1980. A Simple Method for Estimating Evolutionary Rates of Base Substitutions Through Comparative Studies of Nucleotide Sequences. *J. Mol. Evol.* 16:11–120
Ko Y-J, Yang EC, Lee J-H, Lee KW, Jeong J-Y, et al. 2015. Characterization of cetacean Numt and its application into cetacean phylogeny. *Genes and Genomics*. 37(12):1061–71
Kondrashov FA. 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc. R. Soc. B Biol. Sci.* 279(1749):5048–57
Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2016. Canu : scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27(5):722–36
Kryazhimskiy S, Plotkin JB. 2008. The Population Genetics of dN/dS. *PLoS Genet.* 4(12):
Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* 33(7):1870–74
Kuro-o M, Yonekawa H, Saito S, Eda M, Higuchi H, et al. 2010. Unexpectedly high genetic diversity of mtDNA control region through severe bottleneck in vulnerable albatross Phoebastria albatrus. *Conserv. Genet.* 11(1):127–37
Lammers F, Janke A, Rücklé C, Zizka V, Nilsson MA. 2017. Screening for the ancient polar bear mitochondrial genome reveals low integration of mitochondrial pseudogenes (numts) in bears. *Mitochondrial DNA Part B Resour.* 2(1):251–54
Lawrence HA, Lyver POB, Gleeson DM. 2014. Genetic panmixia in New Zealand's Grey-faced Petrel:

Implications for conservation and restoration. *Emu*. 114(3):249–58
Lawrence HA, Taylor GA, Millar CD, Lambert DM. 2008a. High mitochondrial and nuclear genetic diversity in one of the world's most endangered seabirds, the Chatham Island Taiko (Pterodroma magentae). *Conserv. Genet.* 9(5):1293–1301
Lawrence HA, Taylor GA, Millar CD, Lambert DM. 2008b. High mitochondrial and nuclear genetic diversity in one of the world's most endangered seabirds, the Chatham Island Taiko (Pterodroma magentae). *Conserv. Genet.* 9(5):1293–1301
Lee M-Y, Jeon HS, Kim Y-J, An J. 2017. Complete mitochondrial genome of *Ciconia nigra* (Ciconiiformes: Ciconiidae). *Mitochondrial DNA Part B*. 2(1):230–31
Li-Sucholeiki X-C, Khrapko K, André PC, Marcelino LA, Karger BL, Thilly WG. 1999. Applications of constant denaturant capillary electrophoresis / high-fidelity polymerase chain reaction to human genetic analysis Nucleic acids. *Electrophoresis*. 20:1224–32
Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv Prepr. arXiv*. 00(00):1–3
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 25(14):1754–60
Librado P, Rozas J. 2009. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*. 25(11):1451–52
Liebers D, De Knijff P, Helbig AJ. 2004. The herring gull complex is not a ring species. *Proc. R. Soc. B Biol. Sci.* 271(1542):893–901
Liu M, Kang C, Yan C, Huang T, Song X, et al. 2016. Phylogenetic analysis of the Black Stork *Ciconia nigra* (Ciconiiformes: Ciconiidae) based on complete mitochondrial genome. *Mitochondrial DNA*. 27(1):261–62
Lohr JN, Haag CR. 2015. Genetic load , inbreeding depression , and hybrid vigor covary with population size : An empirical evaluation of theoretical predictions. *Evolution (N. Y).* 69(12):3109–22
Lopez J V., Yuhki N, Masuda R, Modi W, O'Brien SJ. 1994. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *J. Mol. Evol.* 39(2):174–90
Lounsberry ZT, Brown SK, Collins PW, Henry RW, Newsome SD, Sacks BN. 2015. Next-generation sequencing workflow for assembly of nonmodel mitogenomes exemplified with North Pacific albatrosses (Phoebastria spp.). *Mol. Ecol. Resour.* 15(4):893–902
Lourie SA, Green DM, Vincent ACJ. 2005. Dispersal, habitat differences, and comparative phylogeography of Southeast Asian seahorses (Syngnathidae: Hippocampus). *Mol. Ecol.* 14(4):1073–94
Lynch M, Conery JS. 2000. The Evolutionary Fate and Consequences of Duplicate Genes. *Science (80-.).* 1151:
Maiorano P, Marino M, Flores JA. 2009. The warm interglacial Marine Isotope Stage 31: Evidences from the calcareous nannofossil assemblages at Site 1090 (Southern Ocean). *Mar. Micropaleontol.* 71(3–4):166–75
Marchant S, Higgins PJ. 1990. *Handbook of Australian, New Zealand & Antarctic Birds. Vol. 1, Ratites to Ducks, P. AB*. Oxford University Press
Mariette J, Escudié F, Allias N, Salin G, Noirot C, et al. 2012. NG6: Integrated next generation sequencing storage and processing environment. *BMC Genomics*. 13(1):462
Martínez-Cruz B, Godoy JA, Negro JJ. 2007. Population fragmentation leads to spatial and temporal genetic structure in the endangered Spanish imperial eagle
Martínez-Gómez JE, Matías-Ferrer N, Sehgal RNM, Escalante P. 2015. Phylogenetic placement of the critically endangered Townsend's Shearwater (Puffinus auricularis auricularis): Evidence for its conspecific status with Newell's Shearwater (Puffinus a. newelli) and a mismatch between genetic and phenotypic differentiation. *J. Ornithol.* 156(4):1025–34
Martı nez-GA, Rosell-mele A, Geibert W, Gersonde R. 2009. Links between iron supply , marine productivity , sea surface temperature , and CO 2 over the last 1 . 1 Ma. *Paleoceanography*. 24:1–14
McCauley DE, Bailey MF, Sherman NA, Darnell MZ. 2005. Evidence for paternal transmission and heteroplasmy in the mitochondrial genome of Silene vulgaris, a gynodioecious plant. *Heredity (Edinb).* 95:50–58
McDougall I. 1971. The geochronology and evolution of the young volcanic island of Réunion, Indian Ocean. *Geochim. Cosmochim. Acta*. 35(3):261–88
McDougall IAN, Chamalaun FH. 1969. Isotopic dating and geomagnetic polarity studies on volcanic rocks from Mauritius, Indian Ocean. *Geol. Soc. Am. Bull.* 80(8):1419–42
McKay BD, Zink RM. 2010. The causes of mitochondrial DNA gene tree paraphyly in birds. *Mol. Phylogenet. Evol.* 54(2):647–50
Meirmans PG. 2012. The trouble with isolation by distance. *Mol. Ecol.* 21:2839–46
Mindell DP, Sorenson MD, Dimcheff DE. 1998. LAn Extra Nucleotide Is Not Translated in Mitochondrial ND3 of Some Birds and Turtles. *Mol. Biol. Evol.* 15(11):1568–71
Minoche AE, Dohm JC, Himmelbauer H. 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biol.* 12:
Miraldo A, Hewitt GM, Dear PH, Paulo OS, Emerson BC. 2012. Numts help to reconstruct the demographic history of the ocellated lizard (Lacerta lepida) in a secondary contact zone. *Mol. Ecol.* 21(4):1005–18

Montgomery E, Charlesworth B, Langley C. 1987. Transposable, A test for the role of natural selection in the stabilization of element copy number in a population of Drosophila melanogaster. *Genet. Res.* 49(1):

Morgan-Richards M, Wallis GP. 2003. A comparison of five hybrid zones of the weta Hemideina thoracica (Orthoptera: Anostostomatidae): Degree of cytogenetic differentiation fails to predict zone width. *Evolution (N. Y).* 57(4):849–61

Moritz C, Dowling TE, Brown WM. 1987. Evolution of Animal Mitochondrial DNA: Relevance for Population Biology and Systematics. *Annu. Rev. Ecol. Syst.* 18:269–92

Morris-Pocock J a., Anderson DJ, Friesen VL. 2011. Mechanisms of global diversification in the brown booby (Sula leucogaster) revealed by uniting statistical phylogeographic and multilocus phylogenetic methods. *Mol. Ecol.* 20(13):2835–50

Moum T, Bakke I. 2001. Mitochondrial control region structure and single site heteroplasmy in the razorbill (Alca torda ; Aves). *Curr. Genet.* 39:198–203

Mueller RL, Boore JL. 2005. Molecular Mechanisms of Extensive Mitochondrial Gene Rearrangement in Plethodontid Salamanders. *Mol. Biol. Evol.* 22(10):2104–12

Mundy NI, Winchell CS, Woodruff DS. 1996. Tandem Repeats and Heteroplasmy in the Mitochondrial DNA Control Region of the Loggerhead Shrike (Lanius ludovicianus). *J. Hered.* 87(1):1–6

Murphy RC. 1927. *On Certain Forms of Puffinus Assimilis and Its Allies*. American Museum of Natural History

Nacer DF, do Amaral F. 2017. Striking pseudogenization in avian phylogenetics: Numts are large and common in falcons. *Mol. Phylogenet. Evol.* 115:1–6

Nascimento L De, Delgado JD, Garcı E, Whittaker RJ. 2011. A reconstruction of Palaeo-Macaronesia , with particular reference to the long-term biogeography of the Atlantic island laurel forests. . 226–46

Nei M, Gojoborit T. 1986. Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions ’. *Mol. Biol. Evol.* 3(5):418–26

Nomura S, Kobayashi T, Agawa Y, Margulies D, Scholey V, et al. 2014. Genetic population structure of the Pacific bluefin tuna Thunnus orientalis and the yellowfin tuna Thunnus albacares in the North Pacific Ocean. *Fish. Sci.* 80(6):1193–1204

Nunn GB, Stanley SE. 1998. Body size effects and rates of cytochrome b evolution in tube-nose(d seabirds. *Mol. Biol. Evol.* 15(10):1360–71

O’Dea A, Lessios HA, Coates AG, Eytan RI, Restrepo-Moreno SA, et al. 2016. Formation of the Isthmus of Panama. *Sci. Adv.* 2(8):1–12

Onley D, Scofield RP. 2007. *Albatrosses, Petrels and Shearwaters of the World.* Princeton field guides.

Pacheco MA, Battistuzzi FU, Lentino M, Aguilar RF, Kumar S, Escalante AA. 2011. Evolution of modern birds revealed by mitogenomics: Timing the radiation and origin of major orders. *Mol. Biol. Evol.* 28(6):1927–42

Paiva VH, Geraldes P, Ramirez I, Werner AC, Garthe S, Ramos JA. 2013. Overcoming difficult times : the behavioural resilience of a marine predator when facing environmental stochasticity. . 486:277–88

Paradis E. 2010. pegas: an {R} package for population genetics with an integrated--modular approach. *Bioinformatics*. 26:419–20

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*. 20(2):289–90

Patel S, Weckstein JD, Patané JSL, Bates JM, Aleixo A. 2011. Temporal and spatial diversification of Pteroglossus araçaris (AVES: Ramphastidae) in the neotropics: Constant rate of diversification does not support an increase in radiation during the Pleistocene. *Mol. Phylogenet. Evol.* 58(1):105–15

Patterson SA, Morris-Pocock JA, Friesen VL. 2011. A multilocus phylogeny of the Sulidae (Aves: Pelecaniformes). *Mol. Phylogenet. Evol.* 58(2):181–91

Peck DR. 2006. *Local adaptation in the wedge-tailed shearwater (Puffinus pacificus). PhD thesis, James Cook University.*

Penhallurick J, Wink M. 2004. Analysis of the taxonomy and nomenclature of the Procellariiformes based on complete nucleotide sequences of the mitochondrial cytochrome b gene. *Emu.* 104(2):125–47

Pereira SL, Baker AJ. 2006. A mitogenomic timescale for birds detects variable phylogenetic rates of molecular evolution and refutes the standard molecular clock. *Mol. Biol. Evol.* 23(9):1731–40

Pérez T, Rodriguez F, Fernandez M, Albornoz J, Dominguez A. 2017. Ancient mitochondrial pseudogenes reveal hybridization between distant lineages in the evolution of the Rupicapra genus. *Gene*. 628:63–71

Petit RJ, Excoffier L. 2009. Gene flow and species delimitation. *Trends Ecol. Evol.* 24(7):386–93

Piontkivska H, Rooney AP, Nei M. 1997. Purifying Selection and Birth-and-death Evolution in the Histone H4 Gene Family. *Mol. Biol. Evol.* 19(5):689–97

Pons JM, Sonsthagen S, Dove C, Crochet PA. 2014. Extensive mitochondrial introgression in North American Great Black-backed Gulls (Larus marinus) from the American Herring Gull (Larus smithsonianus) with little nuclear DNA impact. *Heredity (Edinb).* 112(3):226–39

Poulton J, Deadman ME, Gardiner RM. 1989. Duplications of mitochondrial DNA in mitochondrial myopathy. *Lancet*. 236–40

Prado M, Calo-Mata P, Villa TG, Cepeda A, Barros-Velazquez J. 2007. Co-amplification and sequencing of a

cytochrome b fragment affecting the identification of cattle in PCR-RFLP food authentication studies. *Food Chem.* 105:436–42

Precheur C, Barbraud C, Martail F, Mian M, Nicolas J, et al. 2016. Some like it hot : effect of environment on population dynamics of a small tropical seabird in the Caribbean region. *Ecosphere*. 7(10):1–18

Price TD. 2008. *Speciation in Birds*

Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, et al. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature*. 526(7574):569–73

Prychitko TM, Moore WS. 1997. Comparative Evolution of the Mitochondrial Cytochrome b Gene and Nuclear -Fibrinogen Intron 7 in Woodpeckers. *Mol. Biol. Evol.* 17(7):1101–11

Pyle P, Welch AJ, Fleischer RC. 2011. A New Species of Shearwater (Puffinus) Recorded from Midway Atoll, Northwestern Hawaiian Islands. *Condor*. 113(3):518–27

Rabouam C, Thibault J, Bretagnolle V. 1998. Natal Philopatry and Close Inbreeding in Cory ' s Shearwater (Calonectris diomedea) Author (s): Corinne Rabouam , Jean-Claude Thibault and Vincent Bretagnolle Reviewed work (s): Published by : University of California Press on behalf of the American. *Auk*. 115(2):483–86

Rains D, Weimerskirch H, Burg TM. 2011. Piecing together the global population puzzle of wandering albatrosses: Genetic analysis of the Amsterdam albatross Diomedea amsterdamensis. *J. Avian Biol.* 42(1):69–79

Ramakrishnan U, Hadly EA, L MJ. 2005. Detecting past population bottlenecks using temporal genetic data. *Mol. Ecol.* 14:2915–22

Rambaut A, Drummond DA. 2007. *Tracer v 1.4*

Ramirez O, Illera JC, Rando JC, González-Solís J, Alcover JA, Lalueza-Fox C. 2010. Ancient DNA of the Extinct Lava Shearwater (Puffinus olsoni) from the Canary Islands Reveals Incipient Differentiation within the P. puffinus Complex. *PLoS One*. 5(12):e16072

Ramírez O, Gómez-Díaz E, Olalde I, Illera JC, Rando JC, et al. 2013. Population connectivity buffers genetic diversity loss in a seabird. *Front. Zool.* 10(1):28

Raposo F, Neves LG, Resende MFR, Mobili F. 2015. Ultraconserved Elements Sequencing as a Low-Cost Source of Complete Mitochondrial Genomes and Microsatellite Markers in Non- Model Amniotes. *PLoS One*. 10:1–9

Raquin A, Depaulis F, Lambert A, Galic N, Brabant P, Goldringer I. 2008. Experimental Estimation of Mutation Rates in a Wheat Population With a Gene Genealogy Approach. *Genetics*. 2211(August):2195–2211

Rayner MJ, Hauber ME, Steeves TE, Lawrence HA, Thompson DR, et al. 2011. Contemporary and historical separation of transequatorial migration between genetically distinct seabird populations. *Nat. Commun.* 2(May):332

Richard G, Pâques F. 2000. Mini- and microsatellite expansions : the recombination connection. *EMBO Rep.* 1(2):122–26

Richard G, V Titova O, D. Fedutin I, Steel D, Meschersky I, et al. 2018. Cultural Transmission of Fine-Scale Fidelity to Feeding Sites May Shape Humpback Whale Genetic Diversity in Russian Pacific Waters. *J. Hered.* 1–11

Richly E, Leister D. 2004a. NUMTs in Sequenced Eukaryotic Genomes. . 21(6):

Richly E, Leister D. 2004b. NUMTs in sequenced eukaryotic genomes. *Mol. Biol. Evol.* 21(6):1081–84

Ritchie PA, Millar CD, Gibb GC, Baroni C, Lambert DM. 2004. Ancient DNA Enables Timing of the Pleistocene Origin and Holocene Expansion of Two Adélie Penguin Lineages in Antarctica. *Mol. Biol. Evol.* 21(2):240–48

Robertson BC, Stephenson BM, Goldstien SJ. 2011. When rediscovery is not enough: Taxonomic uncertainty hinders conservation of a critically endangered bird. *Mol. Phylogenet. Evol.* 61(3):949–52

Robinson JT, Thorcaldsdottir H, Winckler W, Guttman M, Lander ES, et al. 2012. Integrative Genomics Viewer. *Nat Biotechnol*. 29(1):24–26

Roca AL, Georgiadis N, O'Brien SJ. 2007. Cyto-nuclear genomic dissociation and the African elephant species question. *Quat. Int.* 169–170(SPEC. ISS.):4–16

Rolshausen G, Segelbacher G, Hobson KA, Schaefer HM. 2009. Contemporary Evolution of Reproductive Isolation and Phenotypic Divergence in Sympatry along a Migratory Divide. *Curr. Biol.* 19(24):2097–2101

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, et al. 2012a. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Syst. Biol.* 61(3):539–42

Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, et al. 2012b. Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61(3):539–42

Ruokonen M, Kvist L. 2002. Structure and evolution of the avian mitochondrial control region. *Mol. Phylogenet. Evol.* 23:422–32

Saccone C, De Giorgi C, Gissi C, Pesole G, Reyes A. 1999. Evolutionary genomics in Metazoa: The mitochondrial DNA as a model system. *Gene*. 238(1):195–209

Samadi S, Barberousse A. 2009. Species : towards new , well-grounded practices. *Biol. J. Linn. Soc.* 97:217–22

Sammler S, Bleidorn C, Tiedemann R. 2011. Full mitochondrial genome sequences of two endemic Philippine hornbill species(Aves: Bucerotidae) provide evidence for pervasive mitochondrial DNA recombination. *BMC Genomics*. 12(1):35

Sato A, Tichy H, Colm O, Grant PR, Grant BR, Klein J. 2001. On the Origin of Darwin ' s Finches. *Mol. Biol. Evol.* 18(3):299–311

Scherer RP, Bohaty SM, Dunbar RB, Esper O, Flores JA, et al. 2008. Antarctic records of precession-paced insolation-driven warming during early Pleistocene Marine Isotope Stage 31. *Geophys. Res. Lett.* 35(3):1–5

Schiavo G, Hoffmann OI, Ribani A, Utzeri VJ, Ghionda MC, et al. 2017. A genomic landscape of mitochondrial DNA insertions in the pig nuclear genome provides evolutionary signatures of interspecies admixture. *DNA Res.* 24(5):487–98

Schmitz J, Piskurek O, Zischler H. 2005. Forty million years of independent evolution: A mitochondrial gene and its corresponding nuclear pseudogene in primates. *J. Mol. Evol.* 61(1):1–11

Scornavacca C, Zickmann F, Huson DH. 2011. Tanglegrams for rooted phylogenetic trees and networks. *Bioinformatics*. 27:248–56

Seehausen O. 2004. Hybridization and adaptive radiation. *Trends Ecol. Evol.* 19(4):

Sharp PM, Tuohy TMF, Mosurski KR. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.* 14(13):5125–43

Sheppard AE, Timmis JN. 2009. Instability of Plastid DNA in the Nuclear Genome. *PLoS Genet.* 5(1):1–8

Shokralla S, Gibson JF, Nikbakht H, Janzen DH. 2014. Next-generation DNA barcoding : using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. *Mol. Ecol. Resour.* 14:892–901

Sibley CG, Monroe BL. 1990. *Distribution and Taxonomy of Birds of the World*. Yale University Press

Silva MC, Duarte M a, Coelho MM. 2011. Anonymous nuclear loci in the white-faced storm-petrel Pelagodroma marina and their applicability to other Procellariiform seabirds. *J. Hered.* 102(3):362–65

Silva MC, Matias R, Wanless RM, Ryan PG, Stephenson BM, et al. 2015. Understanding the mechanisms of antitropical divergence in the seabird White-faced Storm-petrel (Procellariiformes: Pelagodroma marina) using a multilocus approach. *Mol. Ecol.* 24(12):3122–37

Slack KE, Jones CM, Ando T, Harrison GL, Fordyce RE, et al. 2006. Early penguin fossils, plus mitochondrial genomes, calibrate avian evolution. *Mol. Biol. Evol.* 23(6):1144–55

Smith AL, Monteiro L, Hasegawa O, Friesen VL. 2007. Global phylogeography of the band-rumped storm-petrel (Oceanodroma castro; Procellariiformes: Hydrobatidae). *Mol. Phylogenet. Evol.* 43(3):755–73

Smouse PE, Long JC, Sokal RR. 1986. Multiple Regression and Correlation Extensions of the Mantel Test of Matrix Correspondence. *Syst. Zool.* 35(4):627

Sokal RR, Rohlf FJ. 1981. *Biometry*. CA. W. H. Free ed.

Song H, Buhay JE, Whiting MF, Crandall KA. 2008. Many species in one : DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *PNAS*. 105(36):13486–91

Song H, Moulton MJ, Hiatt KD, Whiting MF. 2013. Uncovering historical signature of mitochondrial DNA hidden in the nuclear genome: The biogeography of Schistocerca revisited. *Cladistics*. 29(6):643–62

Sorenson MD. 2003. Avian mtDNA primers

Sorenson MD, Quinn TW. 1998a. Numts: a challenge for avian systematics and population biology. *Auk.* 115:214–21

Sorenson MD, Quinn TW. 1998b. Numts: A Challenge for Avian Systematics and Population Biology. *Auk.* 115(1):214–21

Stephens M, Smith NJ, Donnelly P. 2001. A New Statistical Method for Haplotype Reconstruction from Population Data. *Am. J. Hum. Genet.* 68(4):978–89

Stewart JB, Chinnery PF. 2015. The dynamics of mitochondrial DNA heteroplasmy: Implications for human health and disease. *Nat. Rev. Genet.* 16(9):530–42

Sunnucks P, Wilson ACC, Beheregaray LB, Zenger K, French J, Taylor AC. 2000. SSCP is not so difficult : the application and utility of single-stranded conformation polymorphism in evolutionary biology and molecular ecology. *Mol. Ecol.* 9:1699–1710

Taguchi M, King JR, Wetklo M, Withler RE, Yokawa K. 2015. Population genetic structure and demographic history of Pacific blue sharks (Prionace glauca) inferred from mitochondrial DNA analysis. *Mar. Freshw. Res.* 66(3):267–75

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. 123(3):585–95

Taylor RS, Bailie A, Gulativa R, Birt T, Aarvak T, et al. 2018. Sympatric population divergence within a highly pelagic seabird species complex (Hydrobates spp.). *J. Avian Biol.* 49(1):

Team RC. 2019. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Techow NMS, Ryan PG, O'Ryan C. 2009. Phylogeography and taxonomy of White-chinned and Spectacled Petrels. *Mol. Phylogenet. Evol.* 52(1):25–33

Tennyson AJD, Zealand N, Papa T, Box PO, Zealand N, et al. 2013. A hybrid gadfly petrel suggests that soft-plumaged petrels (Pterodroma mollis) had colonised the Antipodes Islands by the 1920s. *Notornis*. 60:290–95

Teske PR, Hamilton H, Palsbøll PJ, Choo CK, Gabr H, et al. 2005. Teske et al 2005 Long distance dispersal. . 286:249–60

Teske PR, Sandoval-Castillo J, Golla TR, Emami-Khoyi A, Tine M, et al. 2018. Thermal selection drives biodiversity origination across the Atlantic/Indian Ocean boundary. *bioRxiv*. 1–10

Teske PR, Von der Heyden S, McQuaid CD, Barker NP. 2011. A review of marine phylogeography in southern Africa. *S. Afr. J. Sci.* 107(5/6):1–11

Thalmann O, Hebler J, Poinar HN, Pääbo S, Vigilant L. 2004. Unreliable mtDNA data due to nuclear insertions: A cautionary tale from analysis of humans and other great apes. *Mol. Ecol.* 13(2):321–35

Ting C-T, Tsaur S-C, Sun S, Browne WE, Chen Y-C, et al. 2004. Gene duplication and speciation in Drosophila: Evidence from the Odysseus locus. *Proc. Natl. Acad. Sci.* 101(33):12232–35

Toews DPL, Brelsford A. 2012. The biogeography of mitochondrial and nuclear discordance in animals. *Mol. Ecol.* 21(16):3907–30

Torres L. 2019. *Phylogéographie et évolution moléculaire chez les Procellariiformes : Apport à la diversification des oiseaux marins*

Torres L, Pante E, González-Solís J, Viricel A, Ribout C, et al. Sea surface temperature, rather than land mass or geographical distance, may drive genetic differentiation in a species complex of highly-dispersive seabirds

Torres L, Welch AJ, Zanchetta C, Chesser RT, Manno M, et al. 2018. Evidence for a duplicated mitochondrial region in Audubon's shearwater based on MinION sequencing. *Mitochondrial DNA Part A*. 1–8

Tsaousis AD, Martin DP, Ladoukakis ED, Posada D, Zouros E. 2005. Widespread recombination in published animal mtDNA sequences. *Mol. Biol. Evol.* 22(4):925–33

Twyford AD, Ennos RA. 2011. Next-generation hybridization and introgression. *Heredity (Edinb).* 108(3):179–89

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, et al. 2012. Primer3-new capabilities and interfaces. *Nucleic Acids Res.* 40(15):1–12

Urantowka AD, Kroczak A, Silva T, Zamora R, Fern N, et al. 2018. New Insight into Parrots ' Mitogenomes Indicates That Their Ancestor Contained a Duplicated Region. *Mol. Biol. Evol.* 35(12):2989–3009

Vallender R, Robertson RJ, Friesen VL, Lovette IJ. 2007. Complex hybridization dynamics between golden-winged and blue-winged warblers (Vermivora chrysoptera and Vermivora pinus) revealed by AFLP, microsatellite, intron and mtDNA markers. *Mol. Ecol.* 16(10):2017–29

Van Loenen AL. 2013. *Reconstructing the Genetic Legacy of Cook's Petrels (Pterodroma cookii)*

Vekemans X, Hardy O. 2004. New insights from fine-scale spatial genetic structure analyses in plant populations. *Mol. Ecol.* 13:921–35

Verscheure S, Backeljau T, Desmyter S. 2015. In silico discovery of a nearly complete mitochondrial genome Numt in the dog (Canis lupus familiaris) nuclear genome. *Genetica*. 143(4):453–58

Viricel A, Rosel PE. 2014. Hierarchical population structure and habitat differences in a highly mobile marine species: the Atlantic spotted dolphin. *Mol. Ecol.* 23(20):

Wakefield ED, Bodey TW, Bearhop S, Blackburn J, Colhoun K, et al. 2013. Space partitioning without territoriality in gannets. *Science (80-.).* 341(6141):

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, et al. 2014. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*. 9(11):

Wallace SJ, Morris-Pocock JA, González-Solís J, Quillfeldt P, Friesen VL. 2017. A phylogenetic test of sympatric speciation in the Hydrobatinae (Aves: Procellariiformes). *Mol. Phylogenet. Evol.* 107:39–47

Wang J. 2007. Triadic IBD coefficients and applications to estimating pairwise relatedness. *Genet. Res.* 89(3):135–53

Wang J. 2011. Coancestry: A program for simulating, estimating and analysing relatedness and inbreeding coefficients. *Mol. Ecol. Resour.* 11(1):141–45

Watanabe M, Nikaido M, Tsuda TT, Kobayashi T, Mindell D, et al. 2006. New candidate species most closely related to penguins. *Gene*. 378(1–2):65–73

Webb T, Bartlein PJ. 1992. Global Changes During the Last 3 Million Years: Climatic Controls and Biotic Responses. *Annu. Rev. Ecol. Syst.* 23(1):141–73

Weber DS, S SB, Ehman NL. 2004. Genetic Consequences of a Severe Population Bottleneck in the Guadalupe Fur Seal (Arctocephalus townsendi). *J. Hered.* 95(2):144–53

Weir BS, Cockerham CC. 1984. Atistics for the analysis of population structure. *Evolution (N. Y).* 38(6):1358–70

Weir JT, Schluter D. 2008. Calibrating the avian molecular clock. *Mol. Ecol.* 17(10):2321–28

Welch AJ, Fleischer RC, James HF, Wiley AE, Ostrom PH, et al. 2012a. Population divergence and gene flow in an endangered and highly mobile seabird. *Heredity (Edinb).* 109(1):19–28

Welch AJ, Olson SL, Fleischer RC. 2014. Phylogenetic relationships of the extinct St Helena petrel, Pterodroma rupinarum Olson, 1975 (Procellariiformes: Procellariidae), based on ancient DNA. *Zool. J. Linn. Soc.*

170(3):494–505
Welch AJ, Wiley AE, James HF, Ostrom PH, Stafford TW, Fleischer RC. 2012b. Ancient DNA reveals genetic stability despite demographic decline: 3,000 years of population history in the endemic hawaiian petrel. *Mol. Biol. Evol.* 29(12):3729–40
Welch AJ, Yoshida AA, Fleischer RC. 2011. Mitochondrial and nuclear DNA sequences reveal recent divergence in morphologically indistinguishable petrels. *Mol. Ecol.* 20(7):1364–77
West-Eberhard MJ. 1983. Sexual Selection, Social Competition, and Speciation
White DJ, Wolff JN, Pierson M, Gemmell NJ. 2008. Revealing the hidden complexities of mtDNA inheritance. *Mol. Ecol.* 17:4925–42
Wiley AE, Welch AJ, Ostrom PH, James HF, Stricker CA, et al. 2012. Foraging segregation and genetic divergence between geographically proximate colonies of a highly mobile seabird. *Oecologia*. 168(1):119–30
Wilson RP. 2010. Resource partitioning and niche hyper-volume overlap in free-living Pygoscelid penguins. *Funct. Ecol.* 24(3):646–57
Wold JR. 2017. *Phylogenetic relationships, population connectivity, and the development of genetic assignment testing in Buller's Albatross (Thalassarche bulleri)*
Wolfe K, Schields D. 1997. D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*. 387(6634):708–13
Won YJ, Hey J. 2005. Divergence population genetics of chimpanzees. *Mol. Biol. Evol.* 22(2):297–307
Wooler RD, Bradley J, Skira I, Serventy D. 1990. Reproductive success of short-tailed shearwaters puffinus tenuirostris in relation to their age and breeding experience. *Br. Ecol. Soc.* 59(1):161–70
Yasuhara M, Cronin TM. 2008. CLIMATIC INFLUENCES ON DEEP-SEA OSTRACODE (CRUSTACEA) DIVERSITY FOR THE LAST THREE MILLION YEARS. *Ecology*. 89(11):53–65
Yokoyama S, Yokoyama R. 1989. Molecular Evolution of Human Visual Pigment Genes '. *Mol. Biol. Evol.* 6(2):186–97
Young LC. 2010. Inferring colonization history and dispersal patterns of a long-lived seabird by combining genetic and empirical data. *J. Zool.* 281:232–40
Zazo C, Goy JL, Hillaire-marcel C, Dabrio CJ, González-delgado JA, et al. 2010. Sea level changes during the last and present interglacials in Sal Island (Cape Verde archipelago). *Glob. Planet. Change*. 72(4):302–17
Zevering CE, Moritz C, Heideman A, Sturm RA. 1991. Parallel Origins of Duplications and the Formation of Pseudogenes in Mitochondrial DNA from Parthenogenetic Lizards (Heteronotia binoei ; Gekkonidae). *J. Mol. Evol.* 33(5):431–41
Zhang D-X, Hewitt GM. 1996a. Highly conserved nuclear copies of the mitochondrial control region in the desert locust Schistocerca gregaria: some implications for population studies. *Mol. Ecol.* 5(2):295–300
Zhang DX, Hewitt GM. 1996b. Nuclear integrations: challenges for itochondrial DNA markers. *Trends Ecol. Evol.* 11:247–51
Zhang J. 2003. Evolution by gene duplication: an update. *Trends Ecol. Evol.* 18(6):292–98
Zhu Q, Zheng X, Luo J, Gaut BS, Ge S. 2006. Multilocus Analysis of Nucleotide Variation of Oryza sativa and Its Wild Relatives : Severe Bottleneck during Domestication of Rice. *Mol. Biol. Evol.* 24(3):857–88
Zink RM, Barrowclough GF. 2008. Mitochondrial DNA under siege in avian phylogeography. *Mol. Ecol.* 17(9):2107–21
Zino F, Brown RM, Biscoito M. 2008. The separation of Pterodroma madeira (Zino's Petrel) from Pterodroma feae (Fea's Petrel) (Aves: Procellariidae). *Ibis (Lond. 1859).* 150(2):326–34
Zischler H. 2000. Nuclear integrations of mitochondrial DNA in primates: inference of associated mutational events. *Electrophor. An Int. J.* 21(3):531–36
Zouros E, Ball AO, Saavedra C, Freeman KR. 1994. Mitochondrial DNA inheritance. *Nature*. 368:818

Annexe 5

Article non directement relatif aux travaux présentés dans ce manuscrit

Should we pursue RAD sequencing for phylogenetics?

In prep.

Amélia Viricel, Lucas Torres and Eric Pante

Systematics and phylogenetics have entered a new era characterized by accelerated innovation in DNA sequencing technologies paralleled by significant methodological development (e.g. Pyron 2015). Sequencing of restriction-site associated DNA (RAD) has become a popular method in evolutionary biology, with applications ranging from building genetic linkage maps to inferring phylogenies. This method provides a reduced representation of variation in the genome, by sequencing DNA flanking the recognition sites of enzymes used to digest template DNA and detecting single nucleotide polymorphisms (SNPs) in these regions (Baird et al. 2008). High-throughput sequencing platforms such as the Illumina Hiseq allow obtaining hundreds to thousands of loci.

Variants of the original protocol were developed to improve scalability, reduce costs, or fine-tune phylogenetic resolution (e.g. ddRAD, 2bRAD, ezRAD; Peterson et al. 2012, Wang et al. 2012, Toonen et al. 2013). Nevertheless, a major limit of the methods resides in a phenomenon called "locus drop-out:" as divergence increases, conservation of restriction sites erodes and the number of orthologous loci decreases (Davey et al. 2013). Two articles were published to test the phylogenetic resolution of RAD-tags by comparing genomes of *Drosophila* (Rubin et al. 2012, Cariou et al. 2013), both suggesting that the method would reach its limits when divergence exceeds 60 Mya. Several studies subsequently offered empirical data to test this prediction in oaks (Hipp et al. 2014), beetles (Cruaud et al. 2014), corals (Pante et al. 2015), barnacles (Herrera et al. 2015), and salmonids (Gonen et al. 2015), offering contrasting results on the phylogenetic resolution and wide applicability of these methods in phylogenetics. In addition to these empirical tests, significant efforts were invested to develop bioinformatic tools to analyze RAD-tag data at phylogenetic scales (e.g. PyRAD, Eaton et al. 2014; RADami, Hipp et al. 2014).

In parallel to the development of RAD-tag sequencing and its testing in phylogenetics, several other genome sampling strategies were proposed to take advantage of next-generation sequencing technologies to boost resolutive power in phylogenetic inference. Targeted enrichment methods such as sequence capture of conserved (Lemmon et al. 2012) and ultraconserved elements (Faircloth et al. 2012) were proposed. While RAD-tag sequencing provides loci picked at random, these methods use genomic data to design molecular probes that will capture specific loci of interest ("targeted enrichment" strategy, hereon TE). This is, *a priori*, more repeatable than RAD-tag sequencing. The potential of TE to resolve phylogenetic relationships at deep evolutionary time scales (>200My divergence, e.g. Faircloth et al. 2015) raises the question of whether it remains pertinent to invest significant efforts using RAD-sequencing, a method initially described to resolve problems at shallow evolutionary scales (Baird et al. 2008), for reconstructing phylogenies.

New studies, using empirical and *in silico* data, enlighten this issue by comparing RAD and TE strategies at large evolutionary scales (Leaché et al. 2015, Collins et al. 2015). We conducted a literature review of studies that used at least one of these methods for phylogenomics above the species level. The aim of the proposed article is to present the development of the RAD and TE strategies as outlined above, and describe their differences in phylogenetic sensitivity. Notably, the following points will be discussed: (1) the total number of loci recruited by both methods and how this number erodes with increasing time of divergence, (2) the level of topological congruence between phylogenies inferred with RAD and TE, (3) the congruence between estimates of clade age, and (4) the applicability of these method for degraded DNA samples. We will provide a synthesis of the pros and cons of both methods.

References relatives à cette annexe :

Baird NA et al. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* 3:e3376.
Cariou M et al. 2013. Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. *Ecology and Evolution* 3:846–852.
Collins RA and Hrbek T 2015. An in silico comparison of reduced-representation and sequence-capture protocols for phylogenomics. *bioRxiv*.
Cruaud A et al. 2014. Empirical assessment of RAD sequencing for interspecific phylogeny. *Molecular Biology and Evolution* 31:1272–1274.
Davey JW et al. 2013. Special features of RAD Sequencing data: implications for genotyping. *Molecular Ecology* 22:3151–3164.
Eaton D 2014. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics* 30:1844–1849.
Faircloth BC et al. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology* 61:717–726.
Faircloth BC et al. 2015. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. *Molecular Ecology Resources* 15:489–501.
Gonen S et al. 2015. Exploring the utility of cross-laboratory RAD-sequencing datasets for phylogenetic analysis. *BMC Research Notes* 8:299.
Herrera S et al. 2015. Evolutionary and biogeographical patterns of barnacles from deep-sea hydrothermal vents. *Molecular Ecology* 24:673–689.
Hipp AL et al. 2014. A framework phylogeny of the American oak clade based on sequenced RAD data. *PLoS ONE* 9:e93975.
Leaché AD et al. 2015. Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. *Genome Biology and Evolution* 7:706–719.
Lemmon AR et al. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology* 61:727–744.
Pante E et al. 2015. Use of RAD sequencing for delimiting species. *Heredity* 114, 450–459.
Pyron RA 2015. Post-molecular systematics and the future of phylogenetics. *Trends in ecology & evolution* 30:384–389.
Peterson BK 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
Rubin BER et al. 2012. Inferring phylogenies from RAD sequence data. *PLoS ONE* 7:e33394.
Toonen RJ et al. 2013. ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ* 1:e203.
Wang S et al. 2012. 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nature methods* 9:808–810.

Annexe 6

Inventaire des échantillons utilisés, provenance et numéros d’accession Genbank

Annexe 6: Inventaire des échantillons utilisés, provenance et numéros d'accession Genbank

			Numéros d'accession GenBank (XXXX: sequncé mais non enregistré, NA : non séquencé)										Collection		
Specimen ID	Lignée	Colonie	*co1*	*Cob (CLEAN /SITESLESS / AMBIGUOUS / NUMTS*	CR	*paxip1*	*csde1*	*tpm*	*irf2*	*βfib*	*rag1*	Sexe	date	Latitude	Longitude
Allencay1	*lherminieri*	Allencay, Bahamas	MH383430	XXXX	NA	XXXX	XXXX	XXXX	NA	NA	NA	F	NA	23°43'07.1"N	76°09'59.7"W
Allencay10	*lherminieri*	Allencay, Bahamas	NA	XXXX	NA	XXXX	XXXX	XXXX	NA	NA	NA	M	NA	23°43'07.1"N	76°09'59.7"W
Allencay11	*lherminieri*	Allencay, Bahamas	NA	NA	NA	XXXX	XXXX	XXXX	NA	NA	NA	M	NA	23°43'07.1"N	76°09'59.7"W
Allencay12	*lherminieri*	Allencay, Bahamas	NA	NA	NA	XXXX	XXXX	XXXX	NA	NA	NA	F	NA	23°43'07.1"N	76°09'59.7"W
Allencay13	*lherminieri*	Allencay, Bahamas	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	F	NA	23°43'07.1"N	76°09'59.7"W
Allencay14	*lherminieri*	Allencay, Bahamas	MH383424	XXXX	XXXX	NA	XXXX	NA	XXXX	NA	NA	F	NA	23°43'07.1"N	76°09'59.7"W
Allencay15	*lherminieri*	Allencay, Bahamas	MH383425	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	M	NA	23°43'07.1"N	76°09'59.7"W
Allencay16	*lherminieri*	Allencay, Bahamas	MH383426	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	F	NA	23°43'07.1"N	76°09'59.7"W
Allencay17	*lherminieri*	Allencay, Bahamas	MH383427	XXXX	XXXX	XXXX	XXXX	NA	XXXX	NA	NA	F	NA	23°43'07.1"N	76°09'59.7"W
Allencay18	*lherminieri*	Allencay, Bahamas	MH383428	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	NA	23°43'07.1"N	76°09'59.7"W
Allencay19	*lherminieri*	Allencay, Bahamas	MH383429	XXXX	XXXX	XXXX	XXXX	XXXX	NA	XXXX	NA	M	NA	23°43'07.1"N	76°09'59.7"W
Allencay2	*lherminieri*	Allencay, Bahamas	MH383422	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	M	NA	23°43'07.1"N	76°09'59.7"W
Allencay20	*lherminieri*	Allencay, Bahamas	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	M	NA	23°43'07.1"N	76°09'59.7"W
Allencay21	*lherminieri*	Allencay, Bahamas	MH383431	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	M	NA	23°43'07.1"N	76°09'59.7"W
Allencay3	*lherminieri*	Allencay, Bahamas	NA	XXXX	NA	XXXX	XXXX	XXXX	NA	NA	NA	F	NA	23°43'07.1"N	76°09'59.7"W
Allencay5	*lherminieri*	Allencay, Bahamas	NA	XXXX	NA	XXXX	XXXX	XXXX	NA	NA	NA	M	NA	23°43'07.1"N	76°09'59.7"W
Allencay6	*lherminieri*	Allencay, Bahamas	NA	XXXX	NA	XXXX	XXXX	XXXX	NA	NA	NA	M	NA	23°43'07.1"N	76°09'59.7"W
Allencay7	*lherminieri*	Allencay, Bahamas	MH383423	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	XXXX	F	NA	23°43'07.1"N	76°09'59.7"W
Allencay9	*lherminieri*	Allencay, Bahamas	NA	XXXX	NA	XXXX	XXXX	XXXX	NA	NA	NA	NA	NA	23°43'07.1"N	76°09'59.7"W
5500423	*boydi*	Cima, Cape Verde	MH383406	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	F	26/07/10	14°58'27.8"N	24°38'05.2"W
5500444	*boydi*	Cima, Cape Verde	MH383407	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	16/11/13	14°58'27.8"N	24°38'05.2"W
5500456	*boydi*	Cima, Cape Verde	MH383408	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	29/10/12	14°58'27.8"N	24°38'05.2"W
5500457	*boydi*	Cima, Cape Verde	MH383409	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	10/08/11	14°58'27.8"N	24°38'05.2"W
5500472	*boydi*	Cima, Cape Verde	MH383410	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	F	28/10/11	14°58'27.8"N	24°38'05.2"W
5500473	*boydi*	Cima, Cape Verde	MH383411	XXXX	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	30/10/15	14°58'27.8"N	24°38'05.2"W
5500474	*boydi*	Cima, Cape Verde	MH383412	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	30/10/15	14°58'27.8"N	24°38'05.2"W
5500475	*boydi*	Cima, Cape Verde	MH383413	XXXX	NA	XXXX	XXXX	XXXX	XXXX	NA	NA	F	09/08/12	14°58'27.8"N	24°38'05.2"W
5500489	*boydi*	Cima, Cape Verde	MH383414	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	13/08/12	14°58'27.8"N	24°38'05.2"W
5500490	*boydi*	Cima, Cape Verde	MH383415	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	14/08/12	14°58'27.8"N	24°38'05.2"W
5500491	*boydi*	Cima, Cape Verde	MH383416	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	14/08/12	14°58'27.8"N	24°38'05.2"W
5500502	*boydi*	Cima, Cape Verde	XXXX	NA	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	14/08/09	14°58'27.8"N	24°38'05.2"W
5500503	*boydi*	Cima, Cape Verde	NA	NA	NA	NA	NA	NA	NA	NA	XXXX	M	14/08/09	14°58'27.8"N	24°38'05.2"W
5500518	*boydi*	Cima, Cape Verde	MH383417	XXXX	NA	XXXX	XXXX	XXXX	XXXX	NA	NA	F	02/08/10	14°58'27.8"N	24°38'05.2"W
5500520	*boydi*	Cima, Cape Verde	MH383418	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	27/07/10	14°58'27.8"N	24°38'05.2"W
5500534	*boydi*	Cima, Cape Verde	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	F	13/11/13	14°58'27.8"N	24°38'05.2"W
5500537	*boydi*	Cima, Cape Verde	XXXX	NA	XXXX	XXXX	XXXX	NA	XXXX	NA	NA	F	14/11/13	14°58'27.8"N	24°38'05.2"W
5500538	*boydi*	Cima, Cape Verde	MH383421	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	XXXX	M	15/11/13	14°58'27.8"N	24°38'05.2"W

Annexe 6: Inventaire des échantillons utilisés, provenance et numéros d'accession Genbank

5500540	*boydi*	Cima, Cape Verde	MH383419	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	F	15/11/13	14°58'27.8"N	24°38'05.2"W
I001664	*baroli*	Funchal, Madeira	MH383488	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	XXXX	M	10/06/03	32°43'53.3"N	16°55'14.3"W
I001951	*baroli*	Funchal, Madeira	NA	XXXX	XXXX	XXXX	XXXX	NA	XXXX	NA	NA	NA	NA	32°43'53.3"N	16°55'14.3"W
I001954	*baroli*	Funchal, Madeira	NA	NA	XXXX	XXXX	XXXX	NA	XXXX	NA	NA	NA	NA	32°43'53.3"N	16°55'14.3"W
I002182	*baroli*	Funchal, Madeira	MH383471	NA	XXXX	XXXX	XXXX	NA	XXXX	NA	NA	NA	NA	32°43'53.3"N	16°55'14.3"W
longcay1	*lherminieri*	Longcay, Bahamas	MH383456	XXXX	NA	XXXX	XXXX	XXXX	NA	NA	NA	M	NA	23°41'40.5"N	76°06'58.2"W
Longcay10	*lherminieri*	Longcay, Bahamas	NA	NA	NA	XXXX	XXXX	XXXX	NA	NA	NA	M	NA	23°41'40.5"N	76°06'58.2"W
longcay11	*lherminieri*	Longcay, Bahamas	NA	XXXX	NA	XXXX	XXXX	XXXX	NA	NA	NA	M	NA	23°41'40.5"N	76°06'58.2"W
longcay12	*lherminieri*	Longcay, Bahamas	MH383449	XXXX	NA	XXXX	XXXX	XXXX	NA	NA	NA	F	NA	23°41'40.5"N	76°06'58.2"W
longcay13	*lherminieri*	Longcay, Bahamas	MH383450	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	NA	F	NA	23°41'40.5"N	76°06'58.2"W
longcay14	*lherminieri*	Longcay, Bahamas	MH383451	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	M	NA	23°41'40.5"N	76°06'58.2"W
longcay15	*lherminieri*	Longcay, Bahamas	MH383452	NA	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	F	NA	23°41'40.5"N	76°06'58.2"W
longcay16	*lherminieri*	Longcay, Bahamas	MH383453	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	NA	M	NA	23°41'40.5"N	76°06'58.2"W
longcay17	*lherminieri*	Longcay, Bahamas	MH383454	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	F	NA	23°41'40.5"N	76°06'58.2"W
longcay18	*lherminieri*	Longcay, Bahamas	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	M	NA	23°41'40.5"N	76°06'58.2"W
longcay19	*lherminieri*	Longcay, Bahamas	MH383455	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	NA	23°41'40.5"N	76°06'58.2"W
longcay2	*lherminieri*	Longcay, Bahamas	MH383447	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	F	NA	23°41'40.5"N	76°06'58.2"W
longcay20	*lherminieri*	Longcay, Bahamas	MH383457	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	NA	23°41'40.5"N	76°06'58.2"W
longcay3	*lherminieri*	Longcay, Bahamas	NA	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	XXXX	M	NA	23°41'40.5"N	76°06'58.2"W
longcay4	*lherminieri*	Longcay, Bahamas	MH383458	NA	NA	XXXX	XXXX	XXXX	NA	NA	NA	M	NA	23°41'40.5"N	76°06'58.2"W
longcay5	*lherminieri*	Longcay, Bahamas	MH383459	XXXX	NA	NA	XXXX	XXXX	NA	NA	NA	M	NA	23°41'40.5"N	76°06'58.2"W
longcay6	*lherminieri*	Longcay, Bahamas	MH383460	XXXX	NA	XXXX	XXXX	XXXX	NA	NA	NA	F	NA	23°41'40.5"N	76°06'58.2"W
longcay7	*lherminieri*	Longcay, Bahamas	MH383448	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	M	NA	23°41'40.5"N	76°06'58.2"W
longcay8	*lherminieri*	Longcay, Bahamas	MH383461	XXXX	NA	XXXX	XXXX	XXXX	NA	NA	NA	F	NA	23°41'40.5"N	76°06'58.2"W
longcay9	*lherminieri*	Longcay, Bahamas	NA	XXXX	NA	XXXX	XXXX	XXXX	NA	NA	NA	M	NA	23°41'40.5"N	76°06'58.2"W
BU72	*lherminieri*	Martinique	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	M	06/06/12	14°25'02.6"N	60°49'54.0"W
BU76	*lherminieri*	Martinique	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	F	17/04/12	14°25'02.6"N	60°49'54.0"W
BU78	*lherminieri*	Martinique	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	M	08/06/12	14°25'02.6"N	60°49'54.0"W
BU79	*lherminieri*	Martinique	MH383462	NA	NA	NA	NA	NA	NA	NA	NA	M	09/06/12	14°25'02.6"N	60°49'54.0"W
BU80	*lherminieri*	Martinique	MH383432	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	17/04/12	14°25'02.6"N	60°49'54.0"W
BU81	*lherminieri*	Martinique	MH383463	NA	NA	XXXX	NA	NA	NA	NA	NA	F	02/06/12	14°25'02.6"N	60°49'54.0"W
BU83	*lherminieri*	Martinique	MH383433	NA	XXXX	XXXX	NA	XXXX	NA	XXXX	XXXX	M	NA	14°25'02.6"N	60°49'54.0"W
BU90	*lherminieri*	Martinique	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	M	02/06/12	14°25'02.6"N	60°49'54.0"W
BU96	*lherminieri*	Martinique	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	M	17/04/12	14°25'02.6"N	60°49'54.0"W
BU97	*lherminieri*	Martinique	MH383464	NA	NA	XXXX	NA	NA	NA	NA	NA	M	NA	14°25'02.6"N	60°49'54.0"W
BV01	*lherminieri*	Martinique	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	F	02/06/12	14°25'02.6"N	60°49'54.0"W
BV02	*lherminieri*	Martinique	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	F	02/06/12	14°25'02.6"N	60°49'54.0"W
BV07	*lherminieri*	Martinique	MH383466	NA	NA	XXXX	NA	NA	NA	NA	NA	M	02/06/12	14°25'02.6"N	60°49'54.0"W
BV08	*lherminieri*	Martinique	MH383467	NA	NA	XXXX	NA	NA	NA	NA	NA	F	17/04/12	14°25'02.6"N	60°49'54.0"W
BV09	*lherminieri*	Martinique	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	F	17/04/12	14°25'02.6"N	60°49'54.0"W
BV11	*lherminieri*	Martinique	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	M	18/04/12	14°25'02.6"N	60°49'54.0"W
BV14	*lherminieri*	Martinique	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	F	02/06/12	14°25'02.6"N	60°49'54.0"W
BV20	*lherminieri*	Martinique	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	F	02/06/12	14°25'02.6"N	60°49'54.0"W
BV22	*lherminieri*	Martinique	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	M	02/06/12	14°25'02.6"N	60°49'54.0"W
BV23	*lherminieri*	Martinique	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	F	02/06/12	14°25'02.6"N	60°49'54.0"W

Annexe 6: Inventaire des échantillons utilisés, provenance et numéros d'accession Genbank

BV26	*lherminieri*	Martinique	MH383468	NA	NA	XXXX	NA	NA	NA	NA	NA	M	02/06/12	14°25'02.6"N	60°49'54.0"W
BV29	*lherminieri*	Martinique	MH383469	NA	NA	NA	NA	NA	NA	NA	NA	F	02/06/12	14°25'02.6"N	60°49'54.0"W
BV32	*lherminieri*	Martinique	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	M	19/04/12	14°25'02.6"N	60°49'54.0"W
BV35	*lherminieri*	Martinique	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	F	17/04/12	14°25'02.6"N	60°49'54.0"W
BV58	*lherminieri*	Martinique	MH383470	NA	NA	NA	NA	NA	NA	NA	NA	M	18/04/12	14°25'02.6"N	60°49'54.0"W
BV61	*lherminieri*	Martinique	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	M	02/06/12	14°25'02.6"N	60°49'54.0"W
BV62	*lherminieri*	Martinique	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	F	17/04/12	14°25'02.6"N	60°49'54.0"W
BV63	*lherminieri*	Martinique	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	M	02/06/12	14°25'02.6"N	60°49'54.0"W
BV66	*lherminieri*	Martinique	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	M	02/06/12	14°25'02.6"N	60°49'54.0"W
BV77	*lherminieri*	Martinique	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	M	18/04/12	14°25'02.6"N	60°49'54.0"W
BV79	*lherminieri*	Martinique	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	F	19/04/12	14°25'02.6"N	60°49'54.0"W
eto3	*lherminieri*	Martinique	MH383434	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	M	NA	14°25'02.6"N	60°49'54.0"W
FS52824	*lherminieri*	Martinique	MH383435	XXXX	XXXX	XXXX	NA	NA	XXXX	NA	NA	F	NA	14°25'02.6"N	60°49'54.0"W
FS52882	*lherminieri*	Martinique	MH383436	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	NA	14°25'02.6"N	60°49'54.0"W
FS64139	*lherminieri*	Martinique	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	F	NA	14°25'02.6"N	60°49'54.0"W
FX14695	*lherminieri*	Martinique	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	NA	14°25'02.6"N	60°49'54.0"W
FX14706	*lherminieri*	Martinique	MH383437	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	NA	14°25'02.6"N	60°49'54.0"W
FX14746	*lherminieri*	Martinique	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	M	NA	14°25'02.6"N	60°49'54.0"W
FX21485	*lherminieri*	Martinique	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	12/05/13	14°25'02.6"N	60°49'54.0"W
FX21505	*lherminieri*	Martinique	MH383438	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	11/04/13	14°25'02.6"N	60°49'54.0"W
GE72031	*lherminieri*	Martinique	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	NA	14°25'02.6"N	60°49'54.0"W
GE72106	*lherminieri*	Martinique	MH383439	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	NA	14°25'02.6"N	60°49'54.0"W
GE72201	*lherminieri*	Martinique	MH383440	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	M	NA	14°25'02.6"N	60°49'54.0"W
GE72238	*lherminieri*	Martinique	NA	NA	XXXX	NA	XXXX	XXXX	NA	NA	XXXX	F	10/01/13	14°25'02.6"N	60°49'54.0"W
Mayotte		Mayotte	NA	XXXX	XXXX	NA	XXXX	NA	XXXX	NA	NA	NA	02/12/18	12°48'35.4"S	45°09'43.1"E
Geo2175	*baroli*	Montana Clara, Canaries	MH383490	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	24/02/07	29°18'06.9"N	13°32'09.2"W
Geo2177	*baroli*	Montana Clara, Canaries	MH383491	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	24/02/07	29°18'06.9"N	13°32'09.2"W
Geo2180	*baroli*	Montana Clara, Canaries	MH383492	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	F	23/02/07	29°18'06.9"N	13°32'09.2"W
Geo2182	*baroli*	Montana Clara, Canaries	MH383493	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	23/02/07	29°18'06.9"N	13°32'09.2"W
Geo2184	*baroli*	Montana Clara, Canaries	MH383494	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	24/02/07	29°18'06.9"N	13°32'09.2"W
Geo2185	*baroli*	Montana Clara, Canaries	MH383495	XXXX	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	21/02/07	29°18'06.9"N	13°32'09.2"W
Geo2186	*baroli*	Montana Clara, Canaries	MH383486	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	XXXX	F	23/02/07	29°18'06.9"N	13°32'09.2"W
Geo2187	*baroli*	Montana Clara, Canaries	MH383496	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	23/02/07	29°18'06.9"N	13°32'09.2"W
Geo2188	*baroli*	Montana Clara, Canaries	NA	NA	XXXX	NA	NA	NA	NA	NA	XXXX	M	23/02/07	29°18'06.9"N	13°32'09.2"W
Geo2189	*baroli*	Montana Clara, Canaries	MH383497	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	M	23/02/07	29°18'06.9"N	13°32'09.2"W
Plastico1	*baroli*	Montana Clara, Canaries	MH383499	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	21/02/07	29°18'06.9"N	13°32'09.2"W
Plastico2	*baroli*	Montana Clara, Canaries	MH383500	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	F	21/02/07	29°18'06.9"N	13°32'09.2"W
Plastico3	*baroli*	Montana Clara, Canaries	MH383501	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	M	21/02/07	29°18'06.9"N	13°32'09.2"W
Plastico4	*baroli*	Montana Clara, Canaries	MH383502	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	24/02/07	29°18'06.9"N	13°32'09.2"W
Plastico5	*baroli*	Montana Clara, Canaries	MH383503	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	25/02/07	29°18'06.9"N	13°32'09.2"W
5500003	*boydi*	Raso, Cape Verde	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	11/11/09	16°37'05.2"N	24°35'11.4"W
5500008	*boydi*	Raso, Cape Verde	MH383396	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	05/11/09	16°37'05.2"N	24°35'11.4"W
5500014	*boydi*	Raso, Cape Verde	MH383420	XXXX	NA	XXXX	XXXX	XXXX	XXXX	NA	XXXX	M	06/11/07	16°37'05.2"N	24°35'11.4"W
5500015	*boydi*	Raso, Cape Verde	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	06/11/07	16°37'05.2"N	24°35'11.4"W
5500028	*boydi*	Raso, Cape Verde	MH383397	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	07/11/07	16°37'05.2"N	24°35'11.4"W

Annexe 6: Inventaire des échantillons utilisés, provenance et numéros d'accession Genbank

5500040	*boydi*	Raso, Cape Verde	XXXX	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	15/11/08	16°37'05.2"N	24°35'11.4"W
5500052	*boydi*	Raso, Cape Verde	MH383398	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	NA	F	08/11/09	16°37'05.2"N	24°35'11.4"W
5500054	*boydi*	Raso, Cape Verde	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	F	09/11/07	16°37'05.2"N	24°35'11.4"W
5500104	*boydi*	Raso, Cape Verde	MH383399	NA	XXXX	XXXX	XXXX	XXXX	XXXX	NA	XXXX	F	05/11/09	16°37'05.2"N	24°35'11.4"W
5500114	*boydi*	Raso, Cape Verde	MH383400	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	23/01/09	16°37'05.2"N	24°35'11.4"W
5500115	*boydi*	Raso, Cape Verde	NA	NA	XXXX	NA	XXXX	XXXX	NA	NA	XXXX	M	23/01/09	16°37'05.2"N	24°35'11.4"W
5500140	*boydi*	Raso, Cape Verde	MH383401	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	F	11/11/09	16°37'05.2"N	24°35'11.4"W
5500142	*boydi*	Raso, Cape Verde	NA	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	16/11/08	16°37'05.2"N	24°35'11.4"W
5500143	*boydi*	Raso, Cape Verde	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	F	15/11/08	16°37'05.2"N	24°35'11.4"W
5500147	*boydi*	Raso, Cape Verde	MH383402	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	15/11/08	16°37'05.2"N	24°35'11.4"W
5500150	*boydi*	Raso, Cape Verde	MH383403	XXXX	XXXX	XXXX	XXXX	NA	XXXX	NA	NA	F	12/11/08	16°37'05.2"N	24°35'11.4"W
5500153	*boydi*	Raso, Cape Verde	MH383404	XXXX	NA	XXXX	XXXX	XXXX	NA	NA	NA	F	11/11/08	16°37'05.2"N	24°35'11.4"W
5500158	*boydi*	Raso, Cape Verde	MH383405	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	F	21/11/08	16°37'05.2"N	24°35'11.4"W
142133	*bailloni*	North Reunion	NA	NA	XXXX	NA	XXXX	XXXX	NA	NA	XXXX	M	NA	20°52'49.0"S	55°27'08.6"E
142151	*bailloni*	North Reunion	MH383332	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	XXXX	F	NA	20°52'49.0"S	55°27'08.6"E
142169	*bailloni*	North Reunion	MH383333	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	NA	20°52'49.0"S	55°27'08.6"E
142208	*bailloni*	North Reunion	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	NA	20°52'49.0"S	55°27'08.6"E
142223	*bailloni*	North Reunion	MH383336	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	NA	20°52'49.0"S	55°27'08.6"E
BY13	*bailloni*	North Reunion	MH383350	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	07/02/16	20°52'49.0"S	55°27'08.6"E
BY14	*bailloni*	North Reunion	MH383352	NA	NA	XXXX	XXXX	XXXX	NA	NA	NA	F	14/02/16	20°52'49.0"S	55°27'08.6"E
BY15	*bailloni*	North Reunion	MH383353	XXXX	NA	XXXX	XXXX	XXXX	NA	NA	NA	F	27/02/16	20°52'49.0"S	55°27'08.6"E
BY16	*bailloni*	North Reunion	MH383354	XXXX	NA	XXXX	XXXX	XXXX	NA	NA	NA	F	14/02/16	20°52'49.0"S	55°27'08.6"E
BY17	*bailloni*	North Reunion	MH383355	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	24/02/16	20°52'49.0"S	55°27'08.6"E
BY18	*bailloni*	North Reunion	MH383357	NA	NA	XXXX	XXXX	XXXX	NA	NA	NA	F	24/02/16	20°52'49.0"S	55°27'08.6"E
BY19	*bailloni*	North Reunion	MH383358	XXXX	NA	XXXX	XXXX	XXXX	NA	NA	NA	M	24/02/16	20°52'49.0"S	55°27'08.6"E
BY20	*bailloni*	North Reunion	MH383359	NA	NA	XXXX	XXXX	XXXX	NA	NA	NA	M	24/02/16	20°52'49.0"S	55°27'08.6"E
BY21	*bailloni*	North Reunion	MH383360	XXXX	NA	XXXX	XXXX	XXXX	NA	NA	NA	F	27/02/16	20°52'49.0"S	55°27'08.6"E
BY22	*bailloni*	North Reunion	MH383361	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	26/02/16	20°52'49.0"S	55°27'08.6"E
BY23	*bailloni*	North Reunion	MH383363	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	05/03/16	20°52'49.0"S	55°27'08.6"E
BY24	*bailloni*	North Reunion	MH383365	XXXX	NA	XXXX	XXXX	NA	NA	NA	NA	F	05/03/16	20°52'49.0"S	55°27'08.6"E
BY25	*bailloni*	North Reunion	MH383366	XXXX	NA	XXXX	XXXX	XXXX	NA	NA	NA	F	05/03/16	20°52'49.0"S	55°27'08.6"E
BY26	*bailloni*	North Reunion	MH383367	NA	NA	XXXX	XXXX	XXXX	NA	NA	NA	F	03/04/16	20°52'49.0"S	55°27'08.6"E
BY27	*bailloni*	North Reunion	MH383368	NA	NA	XXXX	XXXX	XXXX	NA	NA	NA	M	03/04/16	20°52'49.0"S	55°27'08.6"E
BY28	*bailloni*	North Reunion	NA	XXXX	NA	NA	NA	NA	NA	NA	NA	M	03/04/16	20°52'49.0"S	55°27'08.6"E
BY30	*bailloni*	North Reunion	NA	XXXX	NA	NA	NA	NA	NA	NA	NA	M	NA	20°52'49.0"S	55°27'08.6"E
BY31	*bailloni*	North Reunion	NA	XXXX	NA	NA	NA	NA	NA	NA	NA	M	22/05/16	20°52'49.0"S	55°27'08.6"E
BY32	*bailloni*	North Reunion	NA	XXXX	NA	NA	NA	NA	NA	NA	NA	F	22/05/16	20°52'49.0"S	55°27'08.6"E
BY33	*bailloni*	North Reunion	NA	XXXX	NA	NA	NA	NA	NA	NA	NA	F	22/05/16	20°52'49.0"S	55°27'08.6"E
BY37	*bailloni*	North Reunion	NA	XXXX	NA	NA	NA	NA	NA	NA	NA	F	24/05/16	20°52'49.0"S	55°27'08.6"E
1560	*bailloni*	South Reunion	MH383348	XXXX	XXXX	XXXX	XXXX	NA	XXXX	NA	NA	M	26/05/16	21°22'44.7"S	55°37'47.7"E
1581	*bailloni*	South Reunion	NA	XXXX	XXXX	XXXX	XXXX	NA	XXXX	NA	NA	M	28/05/16	21°22'44.7"S	55°37'47.7"E
1587	*bailloni*	South Reunion	XXXX	XXXX	XXXX	XXXX	NA	XXXX	XXXX	NA	NA	F	28/05/16	21°22'44.7"S	55°37'47.7"E
15171	*bailloni*	South Reunion	MH383349	XXXX	NA	XXXX	NA	XXXX	NA	NA	NA	F	NA	21°22'44.7"S	55°37'47.7"E
15175	*bailloni*	South Reunion	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	F	NA	21°22'44.7"S	55°37'47.7"E
15184	*bailloni*	South Reunion	NA	NA	XXXX	XXXX	XXXX	NA	XXXX	NA	NA	M	NA	21°22'44.7"S	55°37'47.7"E

Annexe 6: Inventaire des échantillons utilisés, provenance et numéros d'accession Genbank

142175	*bailloni*	South Reunion	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	NA	21°22'44.7"S	55°37'47.7"E
142178	*bailloni*	South Reunion	XXXX	XXXX	XXXX	XXXX	XXXX	NA	XXXX	NA	NA	M	NA	21°22'44.7"S	55°37'47.7"E
142179	*bailloni*	South Reunion	MH383334	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	NA	21°22'44.7"S	55°37'47.7"E
142212	*bailloni*	South Reunion	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	NA	21°22'44.7"S	55°37'47.7"E
142213	*bailloni*	South Reunion	MH383335	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	NA	21°22'44.7"S	55°37'47.7"E
142229	*bailloni*	South Reunion	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	NA	21°22'44.7"S	55°37'47.7"E
142232	*bailloni*	South Reunion	MH383337	XXXX	NA	XXXX	XXXX	NA	XXXX	NA	NA	F	NA	21°22'44.7"S	55°37'47.7"E
142243	*bailloni*	South Reunion	MH383338	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	F	NA	21°22'44.7"S	55°37'47.7"E
142246	*bailloni*	South Reunion	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	M	NA	21°22'44.7"S	55°37'47.7"E
142247	*bailloni*	South Reunion	MH383339	XXXX	NA	XXXX	XXXX	NA	XXXX	NA	NA	F	NA	21°22'44.7"S	55°37'47.7"E
142307	*bailloni*	South Reunion	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	NA	21°22'44.7"S	55°37'47.7"E
142327	*bailloni*	South Reunion	MH383340	XXXX	NA	XXXX	XXXX	NA	XXXX	NA	NA	F	NA	21°22'44.7"S	55°37'47.7"E
142328	*bailloni*	South Reunion	NA	XXXX	XXXX	XXXX	XXXX	NA	XXXX	NA	NA	F	NA	21°22'44.7"S	55°37'47.7"E
142340	*bailloni*	South Reunion	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	NA	21°22'44.7"S	55°37'47.7"E
142382	*bailloni*	South Reunion	MH383341	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	NA	21°22'44.7"S	55°37'47.7"E
142390	*bailloni*	South Reunion	NA	NA	XXXX	NA	XXXX	XXXX	NA	NA	XXXX	M	NA	21°22'44.7"S	55°37'47.7"E
142425	*bailloni*	South Reunion	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	NA	21°22'44.7"S	55°37'47.7"E
142426	*bailloni*	South Reunion	MH383342	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	NA	21°22'44.7"S	55°37'47.7"E
142465	*bailloni*	South Reunion	MH383343	XXXX	XXXX	NA	NA	XXXX	NA	NA	NA	F	NA	21°22'44.7"S	55°37'47.7"E
142529	*bailloni*	South Reunion	MH383344	XXXX	NA	XXXX	NA	XXXX	XXXX	NA	NA	F	NA	21°22'44.7"S	55°37'47.7"E
142530	*bailloni*	South Reunion	MH383345	XXXX	NA	XXXX	XXXX	NA	XXXX	NA	NA	F	NA	21°22'44.7"S	55°37'47.7"E
142532	*bailloni*	South Reunion	NA	XXXX	XXXX	XXXX	XXXX	NA	XXXX	NA	NA	M	NA	21°22'44.7"S	55°37'47.7"E
142533	*bailloni*	South Reunion	MH383346	XXXX	XXXX	XXXX	XXXX	NA	XXXX	NA	NA	F	NA	21°22'44.7"S	55°37'47.7"E
142536	*bailloni*	South Reunion	MH383347	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	F	NA	21°22'44.7"S	55°37'47.7"E
I001654	*baroli*	Selvagem, Madeira	MH383487	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	XXXX	M	25/02/08	30°08'31.5"N	15°51'53.0"W
I007151	*baroli*	Selvagem, Madeira	MH383474	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	F	25/02/08	30°08'31.5"N	15°51'53.0"W
I012060	*baroli*	Selvagem, Madeira	NA	NA	NA	XXXX	XXXX	NA	XXXX	NA	NA	NA	NA	30°08'31.5"N	15°51'53.0"W
I012063	*baroli*	Selvagem, Madeira	MH383498	XXXX	XXXX	XXXX	XXXX	NA	XXXX	NA	NA	NA	NA	30°08'31.5"N	15°51'53.0"W
selvagem-1	*baroli*	Selvagem, Madeira	MH383504	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	F	NA	30°08'31.5"N	15°51'53.0"W
selvagem-2	*baroli*	Selvagem, Madeira	MH383505	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	NA	30°08'31.5"N	15°51'53.0"W
SelvagemX1	*baroli*	Selvagem, Madeira	NA	NA	NA	XXXX	XXXX	NA	NA	NA	NA	NA	NA	30°08'31.5"N	15°51'53.0"W
selvagemX2	*baroli*	Selvagem, Madeira	MH383506	NA	NA	XXXX	XXXX	NA	NA	NA	NA	NA	NA	30°08'31.5"N	15°51'53.0"W
selvagemX3	*baroli*	Selvagem, Madeira	NA	XXXX	XXXX	NA	XXXX	NA	XXXX	NA	NA	NA	NA	30°08'31.5"N	15°51'53.0"W
selvagemX4	*baroli*	Selvagem, Madeira	NA	XXXX	NA	NA	XXXX	NA	NA	NA	NA	NA	NA	30°08'31.5"N	15°51'53.0"W
BW44	*nicolae*	Aride Island, Seychelles	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	NA	NA	4°12'45.3"S	55°39'53.1"E
BW45	*nicolae*	Aride Island, Seychelles	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	M	NA	4°12'45.3"S	55°39'53.1"E
BW47	*nicolae*	Aride Island, Seychelles	MH383369	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	XXXX	F	02/07/12	4°12'45.3"S	55°39'53.1"E
BW48	*nicolae*	Aride Island, Seychelles	NA	XXXX	NA	XXXX	NA	NA	NA	NA	NA	M	24/07/12	4°12'45.3"S	55°39'53.1"E
BW49	*nicolae*	Aride Island, Seychelles	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	26/07/12	4°12'45.3"S	55°39'53.1"E
BW50	*nicolae*	Aride Island, Seychelles	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	M	26/07/12	4°12'45.3"S	55°39'53.1"E
BW51	*nicolae*	Aride Island, Seychelles	MH383376	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	02/08/12	4°12'45.3"S	55°39'53.1"E
BW52	*nicolae*	Aride Island, Seychelles	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	07/08/12	4°12'45.3"S	55°39'53.1"E
BW53	*nicolae*	Aride Island, Seychelles	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	M	21/08/12	4°12'45.3"S	55°39'53.1"E
BW54	*nicolae*	Aride Island, Seychelles	MH383377	NA	NA	XXXX	NA	NA	NA	NA	NA	F	22/08/12	4°12'45.3"S	55°39'53.1"E
BW55	*nicolae*	Aride Island, Seychelles	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	NA	4°12'45.3"S	55°39'53.1"E

Annexe 6: Inventaire des échantillons utilisés, provenance et numéros d'accession Genbank

BW56	*nicolae*	Aride Island, Seychelles	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	F	22/08/12	4°12'45.3"S	55°39'53.1"E
BW57	*nicolae*	Aride Island, Seychelles	NA	NA	XXXX	XXXX	XXXX	XXXX	NA	NA	XXXX	F	NA	4°12'45.3"S	55°39'53.1"E
BW58	*nicolae*	Aride Island, Seychelles	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	M	12/03/13	4°12'45.3"S	55°39'53.1"E
BW59	*nicolae*	Aride Island, Seychelles	MH383382	NA	NA	XXXX	NA	NA	NA	NA	NA	M	18/04/13	4°12'45.3"S	55°39'53.1"E
BW61	*nicolae*	Aride Island, Seychelles	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	F	01/05/13	4°12'45.3"S	55°39'53.1"E
BW62	*nicolae*	Aride Island, Seychelles	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	28/05/13	4°12'45.3"S	55°39'53.1"E
BW63	*nicolae*	Aride Island, Seychelles	MH383383	XXXX	NA	XXXX	XXXX	XXXX	XXXX	XXXX	NA	F	28/05/13	4°12'45.3"S	55°39'53.1"E
BW64	*nicolae*	Aride Island, Seychelles	MH383370	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	28/05/13	4°12'45.3"S	55°39'53.1"E
BW65	*nicolae*	Aride Island, Seychelles	MH383384	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	28/05/13	4°12'45.3"S	55°39'53.1"E
BW66	*nicolae*	Aride Island, Seychelles	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	29/05/13	4°12'45.3"S	55°39'53.1"E
BW67	*nicolae*	Aride Island, Seychelles	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	F	29/05/13	4°12'45.3"S	55°39'53.1"E
BW68	*nicolae*	Aride Island, Seychelles	NA	NA	NA	XXXX	NA	NA	NA	NA	NA	M	21/08/12	4°12'45.3"S	55°39'53.1"E
GE50905	*nicolae*	Aride Island, Seychelles	MH383386	XXXX	XXXX	XXXX	XXXX	NA	NA	NA	NA	NA	01/04/14	4°12'45.3"S	55°39'53.1"E
GE50906	*nicolae*	Aride Island, Seychelles	MH383387	XXXX	XXXX	NA	XXXX	NA	XXXX	NA	NA	NA	27/03/14	4°12'45.3"S	55°39'53.1"E
GE50907	*nicolae*	Aride Island, Seychelles	MH383388	NA	NA	NA	XXXX	NA	XXXX	NA	NA	NA	27/03/14	4°12'45.3"S	55°39'53.1"E
GE50908	*nicolae*	Aride Island, Seychelles	MH383389	NA	NA	XXXX	XXXX	NA	XXXX	NA	NA	NA	27/03/14	4°12'45.3"S	55°39'53.1"E
GE50909	*nicolae*	Aride Island, Seychelles	MH383390	XXXX	XXXX	XXXX	XXXX	NA	XXXX	NA	NA	NA	31/03/14	4°12'45.3"S	55°39'53.1"E
GE50910	*nicolae*	Aride Island, Seychelles	NA	XXXX	XXXX	XXXX	XXXX	NA	XXXX	NA	NA	NA	01/04/14	4°12'45.3"S	55°39'53.1"E
GE50911	*nicolae*	Aride Island, Seychelles	NA	XXXX	XXXX	XXXX	XXXX	NA	XXXX	NA	NA	NA	01/04/14	4°12'45.3"S	55°39'53.1"E
GE50912	*nicolae*	Aride Island, Seychelles	MH383391	XXXX	NA	XXXX	NA	NA	XXXX	NA	NA	NA	01/04/14	4°12'45.3"S	55°39'53.1"E
GE50913	*nicolae*	Aride Island, Seychelles	NA	XXXX	XXXX	XXXX	XXXX	NA	XXXX	NA	NA	NA	01/04/14	4°12'45.3"S	55°39'53.1"E
GE50915	*nicolae*	Aride Island, Seychelles	MH383392	XXXX	NA	NA	XXXX	NA	XXXX	NA	NA	NA	03/04/14	4°12'45.3"S	55°39'53.1"E
GE50916	*nicolae*	Aride Island, Seychelles	MH383393	XXXX	NA	XXXX	XXXX	NA	XXXX	NA	NA	NA	03/04/14	4°12'45.3"S	55°39'53.1"E
GE50918	*nicolae*	Aride Island, Seychelles	MH383394	XXXX	XXXX	XXXX	XXXX	NA	XXXX	NA	NA	NA	07/04/14	4°12'45.3"S	55°39'53.1"E
GE50919	*nicolae*	Aride Island, Seychelles	NA	XXXX	XXXX	XXXX	XXXX	NA	XXXX	NA	NA	NA	07/04/14	4°12'45.3"S	55°39'53.1"E
GE50920	*nicolae*	Aride Island, Seychelles	MH383395	XXXX	XXXX	XXXX	XXXX	NA	XXXX	NA	NA	NA	07/04/14	4°12'45.3"S	55°39'53.1"E
GE50921	*nicolae*	Aride Island, Seychelles	NA	XXXX	XXXX	XXXX	XXXX	NA	XXXX	NA	NA	NA	09/04/14	4°12'45.3"S	55°39'53.1"E
GE54501	*nicolae*	Aride Island, Seychelles	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	F	NA	4°12'45.3"S	55°39'53.1"E
GE54513	*nicolae*	Aride Island, Seychelles	XXXX	XXXX	NA	XXXX	XXXX	XXXX	XXXX	NA	NA	F	NA	4°12'45.3"S	55°39'53.1"E
I002051	*lherminieri*	Saint Barthélémy	MH383441	XXXX	NA	XXXX	XXXX	NA	NA	NA	NA	NA	NA	17°53'52.2"N	62°49'15.2"W
I002052	*lherminieri*	Saint Barthélémy	MH383442	XXXX	XXXX	XXXX	XXXX	NA	XXXX	NA	NA	NA	NA	17°53'52.2"N	62°49'15.2"W
I002053	*lherminieri*	Saint Barthélémy	MH383443	XXXX	XXXX	XXXX	XXXX	NA	XXXX	NA	NA	NA	NA	17°53'52.2"N	62°49'15.2"W
I002054	*lherminieri*	Saint Barthélémy	MH383444	XXXX	NA	XXXX	XXXX	NA	XXXX	NA	NA	NA	NA	17°53'52.2"N	62°49'15.2"W
I002055	*lherminieri*	Saint Barthélémy	NA	XXXX	XXXX	XXXX	NA	NA	XXXX	NA	NA	NA	NA	17°53'52.2"N	62°49'15.2"W
I002056	*lherminieri*	Saint Barthélémy	MH383445	XXXX	XXXX	NA	NA	NA	XXXX	NA	NA	NA	NA	17°53'52.2"N	62°49'15.2"W
I002057	*lherminieri*	Saint Barthélémy	MH383446	XXXX	NA	NA	NA	NA	NA	NA	NA	NA	NA	17°53'52.2"N	62°49'15.2"W
I002058	*lherminieri*	Saint Barthélémy	NA	XXXX	XXXX	NA	NA	NA	XXXX	NA	NA	NA	NA	17°53'52.2"N	62°49'15.2"W
I003314	*baroli*	Vila, Azores	MH383472	NA	XXXX	XXXX	XXXX	XXXX	NA	XXXX	XXXX	M	17/02/07	36°57'22.8"N	25°09'17.8"W
I006055	*baroli*	Vila, Azores	MH383473	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	05/03/02	36°57'22.8"N	25°09'17.8"W
I006067	*baroli*	Vila, Azores	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	M	23/08/02	36°57'22.8"N	25°09'17.8"W
I008004	*baroli*	Vila, Azores	MH383475	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	F	15/02/07	36°57'22.8"N	25°09'17.8"W
I008023	*baroli*	Vila, Azores	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	M	18/02/07	36°57'22.8"N	25°09'17.8"W
I008044	*baroli*	Vila, Azores	MH383476	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	25/02/08	36°57'22.8"N	25°09'17.8"W
I008054	*baroli*	Vila, Azores	MH383477	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	F	14/02/07	36°57'22.8"N	25°09'17.8"W
I008055	*baroli*	Vila, Azores	NA	XXXX	NA	NA	NA	NA	NA	NA	NA	F	20/02/04	36°57'22.8"N	25°09'17.8"W

Annexe 6: Inventaire des échantillons utilisés, provenance et numéros d'accession Genbank

I008056	*baroli*	Vila, Azores	MH383489	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	XXXX	M	20/02/04	36°57'22.8"N	25°09'17.8"W
I008057	*baroli*	Vila, Azores	MH383478	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	F	20/02/04	36°57'22.8"N	25°09'17.8"W
I008058	*baroli*	Vila, Azores	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	20/02/04	36°57'22.8"N	25°09'17.8"W
I008063	*baroli*	Vila, Azores	XXXX	NA	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	F	21/02/04	36°57'22.8"N	25°09'17.8"W
I008067	*baroli*	Vila, Azores	MH383479	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	21/02/04	36°57'22.8"N	25°09'17.8"W
I008069	*baroli*	Vila, Azores	MH383480	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	NA	NA	F	21/02/04	36°57'22.8"N	25°09'17.8"W
I008072	*baroli*	Vila, Azores	MH383481	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	22/02/04	36°57'22.8"N	25°09'17.8"W
I008078	*baroli*	Vila, Azores	MH383482	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	23/02/04	36°57'22.8"N	25°09'17.8"W
I008096	*baroli*	Vila, Azores	MH383483	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	14/02/07	36°57'22.8"N	25°09'17.8"W
I008098	*baroli*	Vila, Azores	MH383484	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	F	14/02/07	36°57'22.8"N	25°09'17.8"W
I008099	*baroli*	Vila, Azores	MH383485	NA	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	M	15/02/07	36°57'22.8"N	25°09'17.8"W
Kerr et al. 2007	*lherminieri*	North Carolina, USA	DQ434015										15/09/95	36°N	74.45 W
Schindel et al. 2011	*lherminieri*	Bocas Del Toro,Panama	JQ176049										24/04/990	NA	NA
Schindel et al. 2011	*lherminieri*	Bocas Del Toro,Panama	JQ176050										24/04/90	NA	NA
Nunn and Stanley 1998	*lherminieri*	NA		AF076085									NA	NA	NA
Austin et al. 2004	*baroli*	Tenerife, Canary Islands		AY219935									16/06/09	NA	NA
Austin et al. 2004	*baroli*	Tenerife, Canary Islands		AY219936									16/06/09	NA	NA
Austin et al. 2004	*boydi*	Rombos Island		AY219937									12/03/09	NA	NA
Austin et al. 2004	*lherminieri*	Myra-Provost Island, Bahamas		AY219940									06/03/09	NA	NA
Austin et al. 2004	*lherminieri*	Tobago, Carribean		AY219941									16/04/09	NA	NA
Austin et al. 2004	*lherminieri*	Fernando de Noronha, Brazil		AY219942									23/06/09	NA	NA
Austin et al. 2004	*lherminieri*	near Oregon Inlet		AY219943									NA	NA	NA
Austin et al. 2004	*lherminieri*	Martinique		AY219944									19/06/09	NA	NA
Austin et al. 2004	*lherminieri*	Martinique		AY219945									19/06/09	NA	NA
Austin et al. 2004	*lherminieri*	Bocas del Toro, Panama		AY219946									12/05/09	NA	NA
Austin et al. 2004	*lherminieri*	Fernando de Noronha, Brazil		AY219947									23/06/09	NA	NA
Austin et al. 2004	*lherminieri*	Fernando de Noronha, Brazil		AY219948									23/06/09	NA	NA
Austin et al. 2004	*nicolae*	Seychelles		AY219956									24/04/09	NA	NA
Austin et al. 2004	*nicolae*	Seychelles		AY219957									19/06/09	NA	NA
Austin et al. 2004	*nicolae*	Maldives		AY219960									11/05/09	NA	NA
Austin et al. 2004	*bailloni*	Reunion		AY219963									18/06/09	NA	NA
Austin et al. 2004	*bailloni*	Reunion		AY219964									18/06/09	NA	NA
Austin 1996	*boydi*	Raso, Cape Verde		L43024									NA	NA	NA
Austin 1996	*lherminieri*	near Oregon Inlet		L43025									NA	35°30'N	75°W
Austin 1996	*lherminieri*	near Oregon Inlet		L43047									NA	35°30'N	75°W
Austin 1996	*lherminieri*	NA		U57815									NA	NA	NA

Phylogéographie et évolution moléculaire chez les Procellariiformes : Apport à la diversification des oiseaux marins

Résumé :

La génétique de la conservation a pour but de protéger à la fois la diversité génétique et les processus qui l'ont forgée, et nécessite de comprendre ces derniers. Les Procellariiformes (puffins, pétrels et albatros) représentent de nombreuses espèces avec de hautes capacités de dispersion mais un fort comportement philopatrique et avec un fort enjeu de conservation. Nous avons mené une étude multi-locus sur le complexe du puffin d'Audubon, *Puffinus lherminieri*, du nord de l'Atlantique et ses deux lignées-sœurs en océan Indien. Nous avons d'abord montré que les marqueurs génétiques génétiques appliqués ici, et typiquement utilisés pour étudier la phylogéographie des oiseaux marins, étaient probablement soumis à l'introgression, l'hybridation, l'hétéroplasmie, ainsi que la duplication et la pseudogénisation de loci mitochondriaux. Tous ces phénomènes ont un impact sur la qualité et la quantité d'information produite par ces marqueurs. Nous avons réalisé une étude comparative et montré comment gérer au mieux certains de ces problèmes de données et leur traitement. Nous avons également approfondi la composition de la région mitochondriale dupliquée et montré qu'elle avait une évolution complexe au sein des Procellariiformes et pourrait avoir une influence sur leur biologie et leur conservation. Enfin, nous avons montré l'influence des barrières continentales mais surtout de la température de surface de la mer sur la différenciation des oiseaux marins. Nous avons également mis au jour une structuration chez le complexe de puffins qui nécessite de définir des nouvelles priorités de conservation.
Mots clés : Procellariiformes, génétique de la conservation, différenciation, duplication génétique, ADN mitochondrial

Phylogeography and molecular evolution in Procellariiformes: Input to the diversification of seabirds

Summary:

The purpose of conservation genetics is to protect both genetic diversity and the processes that shaped it, and to understand them. Procellariiformes (shearwaters, petrels and albatrosses) represent many species with high dispersal capacities but a strong philopatric behavior and with strong conservation needs. We conducted a multi-locus study on the Audubon shearwater complex, *Puffinus lherminieri*, from the North Atlantic and its two sister lineages in the Indian Ocean. We first showed that the genetic markers used here, and typically used to study the phylogeographu of seabirds, were subject to introgression, hybridization, heteroplasmy, and the duplication and pseudogenisation of some mitochondrial genes.All of these phenomena have an impact on the quality and quantity of information produced by these markers. We led a comparative study and showed how best to manage some of these data problems and their treatment. We have also investigated the composition of the mitochondrial duplicated region and show that it have a complex evolution within the Procellariiformes and may have an influence on their biology and conservation. Finally, we have shown the influence of continental barriers but especially of sea surface temperatureon the differentiation of seabirds. We also uncovered a structuration of the shearwater complex that needed to define new conservation priorities.
Keywords: Procellariiformes, conservation genetics, differentiation, genetic duplication, mitochondrial DNA

CEBC (Centre d'Etudes Biologiques de Chizé)

405 Route de La Canauderie

79360 Villiers-en-Bois

LIENSs (LIttoral ENvironnement et Sociétés)

2 Rue Olympe de Gouges

17000 La Rochelle