Etudiante : Christine Le Mounier
Encadrante : Caroline Scotti

Responsable de Master : M. Aboubakry Sarr

Stage M2 2008

Phylogéographie du *Simarouba amara*, une espèce forestière tropicale, et proposition d'un scénario historique

Unité Mixte de Recherche
ECOlogie des FOrêts de Guyane

Institut National de la Recherche Agronomique

Master Professionel
Genetique et Gestion de la Biodiversite

Résumé

Pendant les périodes glaciaires, d'importants refroidissements puis réchauffements successifs ont du modeler les distributions d'espèces d'arbres de la forêt amazonienne. Il existe deux principales hypothèses concernant l'état de la forêt pendant le dernier pic glacière. L'une penche vers une couverture forestière totale du bassin amazonien. Les compositions d'espèces auraient changé en favorisant les espèces les mieux adaptées à de plus basses températures. Les espèces les plus résistantes à l'aridité auraient eu tendance à mieux persister en altitude. L'autre hypothèse est l'hypothèse des refuges : pendant les périodes froides, la forêt tropicale aurait en grande partie disparu et été remplacée par de la végétation du type savane, excepté en quelques zones plus humides. Dans ces zones, les espèces auraient divergées d'une zone refuge à une autre ce qui permettrait d'expliquer la richesse spécifique et intra spécifique retrouvée aujourd'hui dans le bassin amazonien.

Notre étude phylogéographique tente de repérer des éventuelles traces de ces zones riches en diversité à partir desquelles la recolonisation se serait effectuée. Pour cela on s'intéresse à la diversité génétique intra spécifique dans une espèce de la forêt amazonienne, le *Simarouba amara*.

Le polymorphisme observé dans nos séquences nous amène à distinguer deux lignées. La première se situe du côté Ouest des Andes et l'hypothèse des refuges pourrait expliquer son patron de distribution. La deuxième lignée se situe dans le bassin amazonien : en Guyane, l'idée d'une expansion récente (postérieur au réchauffement climatique de l'holocène) qui avait déjà été émise lors d'études précédentes est confirmée.

Mots clés : Phylogéographie, Amazonie, zones refuges

1. Introduction

1.1. Introduction générale

La diversité spécifique de la faune et de la flore des milieux tropicaux est étonnamment plus élevée que dans les plus hautes latitudes. On compte 73 espèces indigènes d'arbres en forêt tempérée française alors qu'on en compte jusqu'à deux cents sur seulement un hectare en Guyane. En Equateur, on peut dénombrer 1104 espèces sur 25 hectares de forêt humide (Valencia et al., 2004), environ 6 500 espèces au Guyana, 5 100 au Suriname et 5 400 en Guyane française, soit au total 9 500 espèces pour les trois Guyanes (Boggan et al., 1997). Dès 1960, Fischer explique qu'il existe deux processus principaux à l'origine des patrons biogéographiques. Le premier est l'évolution des organismes, l'autre, l'évolution de leurs habitats (Fischer, 1960). Il décrit différents modèles terrestres et marins en intégrant ces deux processus pour tenter d'expliquer la présence d'une telle diversité. Mais les explications de cette richesse décroissante sont encore très débattues. Deux modèles se distinguent :

- Un modèle attribue la richesse spécifique élevée a la stabilité de la forêt tropicale au cours de son l'histoire (Fischer, 1960). Cette stabilité aurait permis une accumulation des espèces au cours du temps associée à un faible taux d'extinction.
- Un modèle « dynamique » dans lequel le taux de spéciation est plus élevé qu'en milieu tempéré (Haffer, 1969). Pour expliquer ce taux de spéciations élevées, Bush et De Oliveira (2006) associent la forêt à un milieu perturbé qui maintiendrait une grande variabilité de micro-habitats et une diversité spécifique élevée en raison d'une forte compétition entre espèces pour l'occupation des niches. Wright (2006) observe que le taux de substitutions dans l'ADN des espèces d'arbres tropicales est plus de deux fois plus élevé que celui des espèces d'arbres de plus fortes latitudes. Il explique le taux d'évolution plus élevé dans les tropiques par l'influence du climat qui serait à l'origine d'un métabolisme cellulaire plus important lorsqu'on se rapproche de l'Equateur (absence des arrêts de croissance par exemple chez les plantes).

L'origine de la forte diversité des espèces dans les tropiques est une question centrale en écologie des communautés et des populations. A la différence de la démarche des écologues qui proposent des modèles basés sur la répartition et les interactions entre espèces au niveau de l'écosystème, la démarche du généticien est d'étudier les processus qui sont à l'origine de la structuration de la diversité à l'intérieur d'une espèce. L'accumulation de divergences génétiques entre les populations pouvant mener, dans le cas extrême, à l'isolement reproducteur, est le processus à l'origine de la formation de nouvelles espèces. L'étude de la structuration de la diversité génétique à l'intérieur des espèces voit alors son importance dans la compréhension des mécanismes à l'origine de la biodiversité dans les

tropiques. Un des principaux facteurs qui détermine la structuration actuelle de la diversité génétique dans les populations est l'Histoire. Le lien entre l'histoire d'une population ou d'une communauté et la composition actuelle de la biodiversité est connu depuis longtemps. En 1987, Ricklefs avait déjà introduit la notion d'histoire dans l'étude de la composition d'une communauté locale (Ricklefs, 1987). L'histoire regroupe ici l'ensemble des processus géologiques, climatiques, anthropiques qui auraient pu influencer la structuration de la diversité génétique à l'intérieur d'une espèce. L'étude de l'histoire des communautés regroupes des disciplines telles que la biogéographie, la palynologie, et depuis plus récemment, la phylogéographie (Avise, 1989).

Le but de la phylogéographie est d'obtenir des informations sur l'histoire des populations à partir d'observations génétiques au sein d'une espèce ou d'un genre et dans une zone géographique donnée. La phylogéographie est un domaine assez récent du fait de sa dépendance aux outils moléculaires et informatiques. Les résultats de l'étude phylogéographique à l'échelle de l'Amérique du Sud sur l'espèce *Simarouba amara* dont ce rapport fait l'objet peuvent apporter des indices historiques privilégiant l'hypothèse de la stabilité forestière, ou bien l'hypothèse des refuges. Tout d'abord, on se propose de prendre plus précisément connaissance de ces deux hypothèses.

1.2. Hypothèse de la persistance de la forêt au pléistocène

L'histoire de la forêt Amazonienne est ancienne. Elle commence à la fin du Jurassique et au début du Crétacé ce qui correspond à peu près à l'apparition de l'embranchement des angiospermes sur notre planète, il y a 130 à 90 millions d'années (Crane et al, 1995). Elle a subi quatre événements climatiques et géologiques majeurs qui sont l'isolation par rapport à l'Afrique et l'Amérique Centrale, le soulèvement des Andes, la fermeture de l'isthme du Panama et les variations climatiques du quaternaire (Burnham et Graham, 1999). Récemment, des études d'isotopes d'oxygènes présents dans les grains de pollen des anciens sédiments ont apporté des informations sur les compositions floristiques et le climat d'il y a plusieurs millions d'années (Colinvaux et al., 2000). Ces analyses de composition de graines présentes dans les sédiments du bassin amazonien montrent que la forêt recouvrait l'ensemble du territoire pendant les périodes de fluctuation climatiques du pléistocène et ne s'est jamais retirée. Seul des changements de composition auraient affecté la végétation. Et du fait de la diminution de température d'environ 7,5°C, les espèces des Andes auraient pu s'étendre vers l'Amazonie (Noonan et Gaucher, 2005). Ces auteurs font l'hypothèse que le réchauffement global du début Holocène (il y a 10 000) aurait conduit à l'expulsion des espèces qui supportaient le moins la chaleur, excepté dans les montagnes où cette végétation aurait pu

persister. Les espèces les plus résistantes à l'aridité auraient étendu leur population vers les hauteurs (Bush et De Oliveira, 2006). Ces dernières études paléo-écologiques et climatiques mettent en question l'hypothèse des refuges.

1.3. Hypothèse des refuges

Durant le pléistocène, la planète a connu des fluctuations climatiques. Des alternances de périodes froides et de périodes plus douces ont clairement influencé le milieu environnemental et ont pu modeler la distribution des populations de végétaux. Des variations d'humidité selon les périodes ont aussi participé à la modification de la distribution de la forêt tropicale (Hammond, 2005). Selon l'hypothèse des refuges, la forêt tropicale n'aurait pas subsisté sur l'ensemble du bassin amazonien. Une végétation herbacée aurait remplacé la forêt dans les zones où la pluviométrie était la plus faible. Seules les zones les plus humides auraient conservé une végétation de type tropical et auraient ainsi constitué des zones de refuges pour ces espèces (Dick et al., 2003). Selon Haffer (1969), ces zones refuges seraient localisées dans les régions où le climat actuel est le plus humide. Pour de nombreuses espèces, les individus présents dans une zone refuge auraient évolués indépendamment des individus d'une zone refuge voisine. Lors de la recolonisation de la forêt tropicale, chaque zone refuge aurait ainsi apporté une nouvelle diversité. C'est le principe de la divergence allopatrique (isolement par distance) qui donne lieu à une diversification écologique des individus, pouvant mener à l'isolement reproducteur, et donc à l'apparition de nouvelles espèces. En cas de contact secondaires des espèces menant à une répartition sympatrique (partagent la même zone géographique), une augmentation de la biodiversité locale est observée (Dick et al., 2003). Ce principe est proche de l'hypothèse de la vicariance proposée par Bush (1994). On parle de vicariance lorsqu'un même taxon occupe des habitats naturels similaires mais séparés géographiquement. Les zones refuges peuvent aussi être appelées zones d'endémismes (zones où la diversité présente est d'origine, sans avoir connu de phénomène de colonisation). La localisation des zones refuges est rendue complexe par le fait qu'il n'a pas existé un seul évènement de régression et d'expansion mais plusieurs : Durant le quaternaire, les aires forestières se sont peut être développées puis réduites plusieurs fois en suivant le rythme des périodes interglaciaires et glaciaires (Caron et al., 2000). Ces séquences d'expansion et de régression de la forêt ont certainement impacté les distributions actuelles géographiques de la diversité.

L'hypothèse des refuges a principalement été analysée par des études de répartition taxonomiques (réalisées à partir des phénotypes). Des études taxonomiques concernant par exemple les oiseaux (Haffer, 1969), les papillons (Hall et Harvey, 2002), les amphibiens, les

reptiles, les primates, les rongeurs, les marsupiaux, les chauves souris, et les serpents (Hall et Harvey, 2002 ; Bush et De Oliveira, 2006) ont pu mettre en relief des discontinuités de répartition d'espèces proches dans l'environnement forestier tropicale. D'après ces auteurs, les discontinuités sont liées à l'histoire car l'aspect uniforme de la forêt tropicale actuelle ne pourrait pas expliquer autant de phénomènes de spéciation. Ces études montrent parfois des répartitions congruentes entre les espèces pouvant s'interpréter comme des histoires communes (Hall et Harvey, 2002 ; Bush et De Oliveira, 2006). La carte ci-dessous résume les résultats obtenus pour les oiseaux, les papillons et quelques plantes, en prenant aussi en compte les natures de sols, les précipitations (Bush et De Oliveira, 2006).

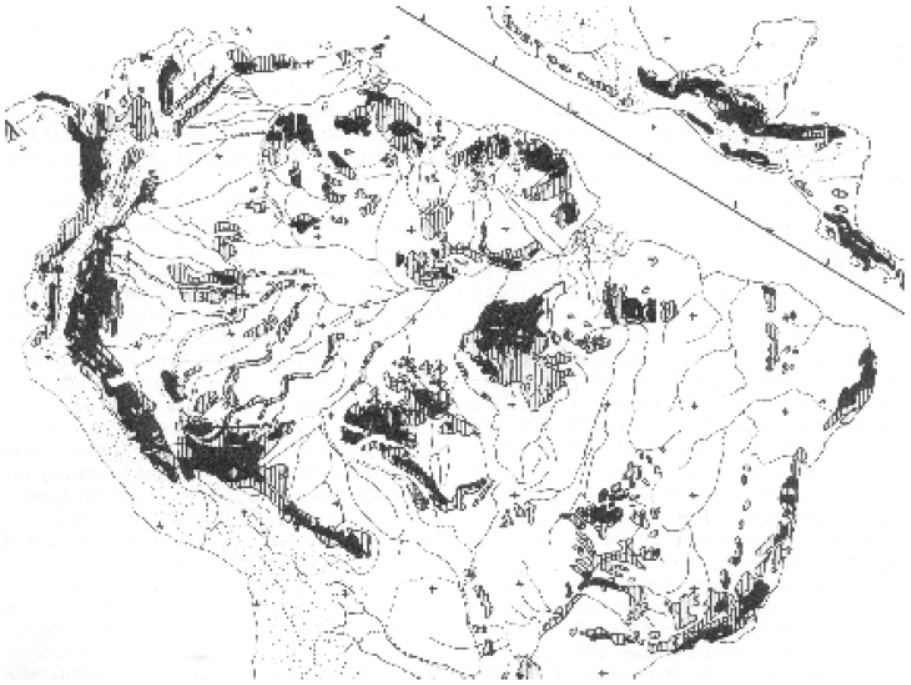


Figure 1 : Proposition de la distribution des zones refuges en Amérique du Sud et Centrale, basée sur les superpositions de zones postulées. L'identification de zones refuges est faite à partir d'études sur la distribution d'espèces d'oiseaux, de papillons et de plantes, tout en prenant en compte la nature des sols et les précipitations. Les marques représentent les probabilités de présence de zone refuge : en noir lorsque la probabilité est de 100 à 80%, en gris entre 60% et 80% (Bush et De Oliveira, 2006).

A l'issue de cette introduction sur l'impact du climat sur la forêt amazonienne nous notons une forte divergence entre les différentes méthodes paléo-écologiques réalisées par Colinvaux *et al*. qui montrent une persistance de la forêt durant les périodes glacières sur tout le bassin amazonien, et les différentes études taxonomiques (Haffer, 1969 ; Hall et Harvey, 2002 ; Bush et De Oliveira, 2006) qui présentent des répartitions en accord avec l'hypothèse des refuges. Depuis une dizaine d'années une nouvelle méthode, complémentaire des précédentes, a été utilisée afin d'affiner les connaissances sur l'évolution de la forêt tropicale durant ces périodes de changements climatiques. Cette méthode est la phylogéographie moléculaire.

1.4. Les outils moléculaires

1.4.1. Différentes applications apportent des informations historiques

D'après Hewitt (2000), les structures génétiques des populations actuelles auraient principalement été formées par les périodes glacières du quaternaire. La variation de la diversité génétique dans le temps et dans l'espace a d'abord été décrite en Europe avec notamment de nombreux travaux sur les génomes chloroplastique ou nucléaire chez le Chêne (Petit et al., 2002). Trois zones refuges ont été identifiées à partir desquelles la recolonisation s'est produite après le dernier maximum glaciaire. Les cycles récurrents de contraction-recolonisation des populations engendrés par des alternances de périodes glaciaires et interglaciaires au cours du pléistocène (il y a 13 000-18 000 ans) expliquent en grande partie la diversité et la distribution actuelle des espèces en Europe. En Amérique du sud, il est plus difficile de reconstituer l'histoire. La distribution très découpée des zones de refuges proposées par les études taxonomiques de Guyane laisse penser que le patron géographique du polymorphisme doit être plus complexe que celui du continent européen (Caron et al., 2000). Les études de phylogéographies réalisées en en Amérique de Sud n'ont toujours pas permis de localiser avec certitude des zones refuges, mais elles ont néanmoins fourni des informations sur les hypothèses de colonisation de la forêt amazonienne.

Phylogéographie et Isthme du Panama :

Des études phylogéographiques montrent que la diversité et les structurations observées sur le continent américain peuvent être liées à des évènements tectoniques plutôt que climatiques. Une étude basée sur des marqueurs de type RFLP (Restriction Fragment Length Polymorphisme) de Cavers et al. (2003) illustre très bien cette idée : du fait de son incontestable richesse spécifique, l'Amérique Centrale est souvent considérée comme une zone refuge. Les structurations de ces populations indiquent que la colonisation de cet espace s'est fait en deux temps : Une première colonisation aurait eu lieu avant la fermeture de l'Isthme du Panama du Nord (Mexique, Guatemala) vers le sud (Costa Rica-Panama) à travers la mer. Et après la formation du lien entre ces deux Amériques au niveau du Panama, il y a environ trois millions d'années, une grosse invasion se serait effectuée de l'Amérique du Sud vers l'Amérique Centrale (Cavers et al. 2003). Dans cet exemple, la structuration de la diversité génétique est plutôt expliquée par des phénomènes de colonisations multiples liées aux évènements tectoniques que par l'existence de zones refuges. Notons que la formation de l'Isthme du Panama peut aussi avoir une influence au niveau des séquences d'ADN du règne animal : une étude phylogéographique de coléoptères du genre Cephaloleia (McKenna et

Farrell, 2006) met en évidence une diversification des coléoptères avec la formation du pont entre les deux Amériques au niveau de l'Isthme du Panama.

Phylogéographie et formation des Andes :

Le soulèvement des Andes est un évènement géologique qui a aussi largement contribué à la structuration génétique des populations. Une étude basée sur le séquençage d'ADN mitochondrial d'abeilles (Apidae; Euglossini) (Dick et al., 2004) ainsi qu' une étude réalisée sur des séquences d'ADN nucléaire d'un arbre commun de la forêt néotropicale (*Symphonia globulifera*) (Dick et al., 2003) montrent une divergence entre les populations du bassin amazonien et les population à l'ouest des Andes. En effet, selon Dick et al. (2003), les montagnes sont des barrières biotiques particulièrement efficaces dans les milieux tropicaux : le gradient de températures constitue une plus grande contrainte que dans les pays tempérés puisque les espèces ne sont pas adaptées aux variations de températures saisonnières (Dick et al., 2003). Muellner et al. (2005) montrent aussi, dans une étude de génétique basée sur des marqueurs de type AFLP (Amplified Fragment Length Polymorphism) réalisée dans la partie sud de la cordillère des Andes et en Patagonie (Chili et Argentine), que les populations de plantes d'*Hypochaeris palustris* proches de la côte, le long des Andes, sont très différenciées. Cette zone est considérée par les auteurs comme une zone refuge, qui n'aurait pas participé à la recolonisation des autres régions des Andes en raison de sa forte différenciation avec les autres régions (Muellner et al., 2005).

Les phylogénies réalisées en Amérique du Sud sur différents modèles aussi bien animaux que végétaux permettent de mettre en avant l'impact de la formation des Andes et de la fermeture de l'Isthme du Panama sur la structuration de la diversité génétique. On peut alors penser, qu'en utilisant un outil moléculaire adapté à l'étude de l'histoire plus récente (Holocene : environ 10 000 ans), on pourrait aussi mettre en évidence l'impact de la fragmentation forestière durant le pléistocène si celle-ci a eu lieu.

L'histoire liée au réchauffement climatique du pléistocène :

Le séquençage des gènes de cytochrome b (ADN mitochondrial) réalisée sur 35 espèces dont des marsupiaux et des petits mammifères recouvrant l'Amazonie montre des patrons de distribution concordants pour certaines espèces (Da silva et Platton, 1998). Les divergences entre taxons sont estimées du début du pléistocène et leur répartition semble plus expliquée par l'histoire tectonique et par les positions des axes fluviaux majeurs, que par l'existence des éventuelles zones refuges (Da silva et Platton, 1998).

La faible structuration des populations d'abeilles en Amazonie (Apidae; Euglossini) (Dick et al., 2004) peut signifier que l'expansion de la forêt à l'ère quaternaire n'a pas été réalisée à

partir de plusieurs centres de diversité, ce qui est plutôt en faveur de l'hypothèse de transition floristique de Colinvaux et al. (2000) selon laquelle la forêt amazonienne n'aurait pas connu de grand changement durant les périodes glacières et interglacières du pléistocène.

La phylogéographie d'Amphibiens réalisée dans le bouclier guyanais sur le genre *Atelopus* (Noonan et gaucher, 2005) montre que la divergence entre les espèces est due à une hétérogénéité d'habitat significative. Aussi, l'émigration d'une lignée aurait eu lieu des Andes vers les montagnes du centre et du nord de la Guyane. Cette lignée originaire des Andes et adaptée au froid est aujourd'hui restreinte aux zones les plus montagneuses de Guyane et reste isolées des autres populations par les plaines environnantes. Au contraire de l'hypothèse des refuges, ces résultats prévoient une expansion d'espèces forestières pendant les périodes de grand froid.

En 2006, la phylogéographie des espèces de paresseux réalisée dans trois régions de la forêt Atlantique au Brésil montre des lignées mitochondriales très divergentes, avec un groupe phylogéographique au nord, et un autre au Sud (Moraes-Barros et al., 2006). On ne sait pas si cette divergence doit être expliquée par des changements de végétation répétitifs durant le quaternaire (hypothèse des refuges) ou par les faibles capacités de dispersion de ces espèces (Moraes-Barros et al., 2006).

L'hypothèse des refuges au niveau de l'Amérique du sud n'a pas pu clairement être mise en évidence par ces différentes études.

1.4.2. Les outils moléculaires et informatiques utilisés lors de l'étude

Il existe plusieurs méthodes pour apprécier la diversité génétique, on peut citer plusieurs méthodes qui ont déjà fait l'objet d'études en Amérique du Sud ou en Amérique Centrale : des méthodes sont basées sur l'analyse du polymorphisme de l'ADN nucléaire en adoptant une démarche de génétique des populations. Différents types de marqueurs génétiques peuvent être utilisés. On peut citer les marqueurs de types AFLP (Muellner et al., 2005) qui ont l'avantage d'être peu couteux et abondants mais qui sont dominants (de type présence absence) limitant les analyses de diversité, les marqueurs microsatellites (Duminil et al., 2006) qui ont l'avantage d'être très polymorphes et codominants, mais couteux. D'autres méthodes sont basées sur le polymorphisme de l'ADN chloroplastique (ou mitochondriale) qui est transmis par un seul parent. Ce génome qui ne recombine pas est transmis de façon clonal au cours des générations. Il permet donc de reconstruire l'histoire d'une espèce en faisant l'hypothèse que les mutations se sont accumulées au cours du temps. Jusqu'à ces deux dernières années, le polymorphisme de l'ADN chloroplastique était analysé en utilisant des marqueurs de type RFLP (Cavers et al, 2003, Caron et al. 2000 ; Dutech et al., 2000). Avec le

développement et l'automatisation des méthodes de séquençage le polymorphisme est de plus en plus étudié directement au niveau de la séquence de la région amplifiée.

Notre étude de phylogéographie utilise la méthode qui représente au mieux la diversité génétique, le séquençage. Pour une séquence donnée, un haplotype correspond à la combinaison des différents allèles retrouvés aux différents sites polymorphiques. La phylogéographie regroupe l'étude des liens évolutifs entre chaque haplotype observé, et l'étude de leur répartition géographique (Mardulyn, 2001). Cela doit nous permettre de comprendre les facteurs qui ont conduit aux distributions observées qui sont parfois très complexes et résultent d'une succession d'évènements anciens (Widmer et Lexer, 2001). Les statistiques traditionnelles de structure des populations comme les F statistiques ne nous permettent pas d'accéder à toute l'information nécessaire pour retracer ces évènements. La méthode testée sur nos échantillons, prend en compte des informations sur la généalogie des haplotypes et détecte des structurations de population significatives dans des cas où le F_{ST} statistique en serait incapable (Templeton, 1998).

Cette méthode est l'analyse des clades emboitées, appelée aussi l'analyse NCPA (Nesting Clade Phylogeny Analysis). Cette technique détecte plus sensiblement les signaux de flux de gènes liés à l'histoire des populations (restriction de flux de gène, fragmentations de population, expansion) que les flux de gènes récurrents (qui modèlent les structures des populations). On étudie tout d'abord la répartition des haplotypes obtenus à partir du polymorphisme de nos séquences et on évalue la diversité génétique (indice de Nei). On établit ensuite le lien entre haplotypes de la façon la plus parcimonieuse. L'analyse NCPA effectuée nous indique des inférences historiques qui pourraient expliquer les patrons de répartition observés. En raison des critiques associées à cette méthode (Petit, 2007) d'autres méthodes ont été proposées pour tester la présence d'un signal phylogéographique. Des tests de neutralité permettant de mettre en évidence des changements de taille des populations (Tajima et Fu) et la comparaison entre l'indice de différenciation génétique calculé avec (Nst) ou sans (Gst) l'information sur la similarité entre haplotypes (Pons et Petit, 1996) sont effectués.

1.5. Structure d'accueil

Mon stage est inscrit dans le cadre du projet SEEDSOURCE qui vise à apporter des outils pour une meilleure gestion de la forêt amazonienne. Ces améliorations à apporter concernent principalement la régénération naturelle et les provenances des graines à planter dans les forêts néotropicales.

2. Matériels et Méthodes

2.1. Matériel végétal : le *Simarouba amara*

2.1.1. Distribution

La famille des Simaroubaceae contient 6 sous familles avec 22 genres et une centaine d'espèces d'arbres ou de buissons (Clayton et al., 2007). Cette famille a une répartition pantropicale et les espèces qui la compose se retrouvent en Amérique, en Afrique, en Asie, en Malaisie, et au Nord-Est de l'Australie (Fernand, 2003). Le *Simarouba amara* est retrouvé dans les forêts humides à partir de l'Amérique centrale (dès l'Honduras et le Nicaragua) jusqu'au bassin amazonien (Hardersty et al., 2005).

2.1.2. Rôle socio économique

Le bois de cet arbre est utilisé localement pour le papier, les constructions de meubles, les constructions intérieures. Il est utilisé comme fébrifuge, tonifiant et vermifuge (Hardersty et al., 2005). La décoction d'écorce et de feuilles ou le simple fait de frotter les feuilles sur le corps semble avoir un effet répulsif contre les moustiques et les poux d'agoutis. Les jeunes feuilles de simaroubacées ont été utilisées par les Waikas de la haute Amazonie comme peinture corporelle noire (Latreille et al., 2004).

2.1.3. Caractéristiques physiologiques et Mode de dispersion

L'arbre *Simarouba amara* peut atteindre 35 mètres en hauteur, et le record enregistré de diamètre à hauteur de poitrine est de 70cm. Cet arbre est un arbre pionnier, qui a donc besoin d'une intense luminosité et est dioïque. Les données fournies depuis 15 ans d'observation sur une surface de 50 ha relatent que la floraison persiste globalement environ de 11 à 15 semaines, avec quelques décalages dans les dates de floraison entre adultes. Ce sont des insectes qui permettent sa pollinisation (insectes principalement généralistes comme de petites abeilles et papillons nocturnes). (Hardesty et al., 2005). Les graines, brillantes, n'entrent pas en dormance et attirent particulièrement les espèces de chachalacas, de gobemouches, de motmots, de grives (oiseaux tropicaux), de singes hurleurs, de singes atèles, et de tamarins. Le transport des graines est effectué sur de longues distances, la distance de recrutement est estimée à 39 m par méthode de piégeage des graines et est trouvée plus de 10 fois plus élevée par des méthodes d'analyses génétiques de paternité et de maternité par marqueurs microsatellites (Hardesty et al., 2006). Même quand les graines sont déposées en grande densité au pied de l'arbre mère, la germination peut avoir lieu préférentiellement sur d'autres sites plus éloignés où les ennemis naturels sont moins présents (Flores et al., 2000).

L'information sur la capacité de dispersion du *S. amara* est importante dans le cadre de notre étude puisque, la grande capacité de dispersion observée peut avoir des effets importants sur l'histoire de l'espèce et notamment sur son expansion démographique suite à la recolonisation post glaciaire (Hardesty et al., 2006).

2.2. Echantillonnage

Au cours de mon stage, j'ai analysé des individus échantillonnés par l'équipe ECOFOG et par des botanistes de l'UMR AMAP (JF Molino et D. Sabatier) (Individus de Guyane), des échantillons frais collectés et envoyés par des partenaires du projet européen SEEDSOURCE, et des échantillons secs provenant d'herbiers (herbiers de Cayenne, de Utrecht, MOBOT, INPA, EMBRAPA-INP). Au total, les séquences de 145 individus ont été obtenues, échantillonnées sur 24 populations d'Amérique du Sud. Le tableau ci-dessous indique la localisation des populations et le nombre de séquences obtenues pour chacune d'entre elles. Une seule séquence supplémentaire a pu être lisible à partir des échantillons d'herbier. En Guyane, la plupart des populations sont situées le long de la côte car plus facilement accessibles. Les populations du sud, accessibles seulement par hélicoptère, on été obtenues à l'issue d'inventaires botaniques réalisés par JF Molino et Daniel Sabatier.
Pour chaque individu, des feuilles ou un disque de cambium ont été prélevés et placés dans des tubes eppendorf contenant du silicagel et stockés en salle climatisée.

Tableau 1 : nombre de séquences obtenues par localisation

Pays	Population	Latitude	Longitude	Effectif
E	Pichincha	-14,7	-78,9	1
Pe	Pérou	-10,4	-75,6	1
B	Manaus	-3,1	-60	4
E	Orellana	-0,78	-76,4	25
FG	Mont Tumuc Humac	2,21	-54,5	1
FG	Mont Emerillon	3,26	-53,2	4
FG	Saul	3,58	-53,2	10
FG	Montagne Tortue	4,3	-52,4	1
FG	Belizon	4,5	-52,3	6
FG	Route de l'Est	4,5	-52,3	5
FG	La Mirande	4,86	-52,3	1

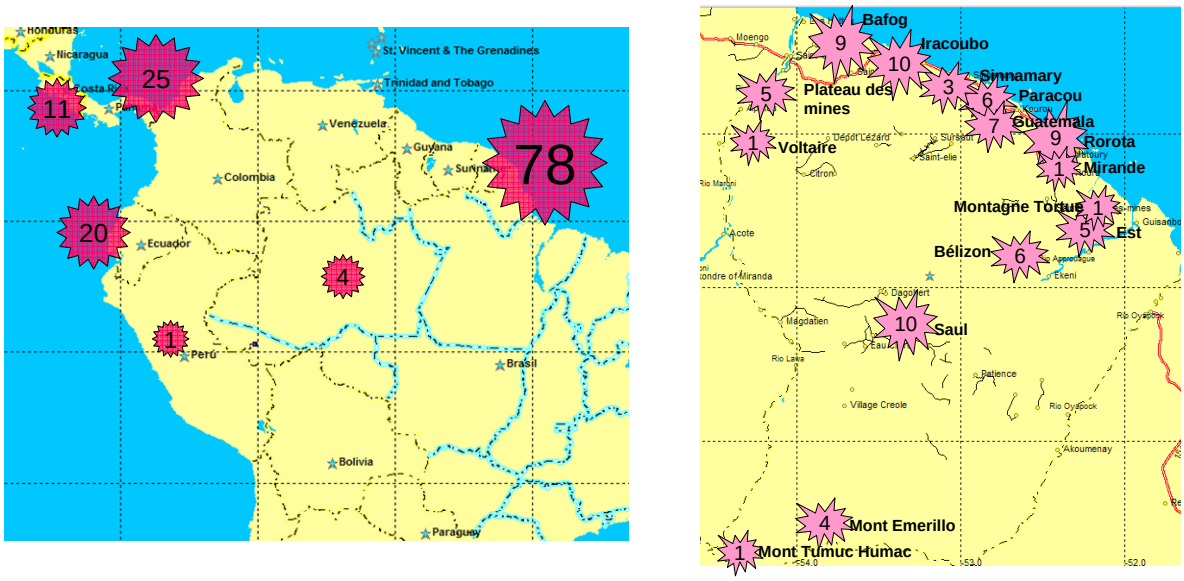
Pays	Population	Latitude	Longitude	Effectif
FG	Rorota	4,87	-52,4	9
FG	Chute Voltaire	5,03	-54,1	1
FG	Route de Guatemala	5,07	-52,7	7
FG	Paracou	5,26	-53	6
FG	Plateau des Mines	5,38	-54,1	5
FG	Sinnamary	5,42	-53,1	3
FG	Bafog	5,48	-54	9
FG	Iracoubo	5,54	-53,5	10
Pa	Pipeline	9,22	-80,1	9
Pa	Santa Rita	9,42	-80	9
Pa	Cerro jefe	9,42	-79,6	7
CR	autre	9,89	-83,7	8
CR	Sarapiqui	10,2	-83,9	3

Légende : E = Equateur ; Pe = Pérou ; B = Brésil ; FG = Guyane française ; Pa = Panama ; CR = Costa Rica

Figure 2 : zones échantillonnées.

a) Carte de l'Amérique du Sud et de l'Amérique Central

b) Carte de la Guyane française



En rose sont représentées les tailles et les localisations de nos échantillons par pays (à gauche) et par population en Guyane (à droite)

2.3. Extraction

2.3.1. Extraction et dosage de l'ADN total pour le matériel frais

Les ADN des individus du Panama et certains du Costa Rica sont arrivés au laboratoire déjà extraits, les autres échantillons frais ont été extraits selon le protocole de Doyle & Doyle (1990). Une méthode d'extraction particulière est réalisée pour les échantillons d'herbiers. L'ADN des échantillons a été dosé sur gel d'agarose 1,4%, grâce à des témoins de concentration connue en ADN. Afin de tester la qualité de l'extraction, l'ADN a été amplifié par PCR avec la paire d'amorces chloroplastique universelle TrnC/Ycf6, puis déposé sur un gel d'agarose 1,4%. Les migrations sont réalisées durant environ 45 min à 100 V/cm, et l'ADN est observé sous UV après révélation au Bromure d'éthidium (BET).

2.3.2. Extraction d'ADN appliquée pour les feuilles séchées d'herbiers

Pour les échantillons d'herbier, le broyage d'environ 100 mg de feuille est effectué dans l'azote liquide. L'ADN des feuilles est ensuite extrait en utilisant le kit invisorb, en suivant le protocole fourni dans le kit. On porte particulièrement attention aux précautions contre la contamination des échantillons et la dégradation d'ADN, celle-ci étant déjà entamée dans les feuilles séchées (port de gants, nettoyage de la paillasse et de la vaisselle à l'eau de javel et traitement aux UV des mortiers et pilons). Un échantillon frais d'une autre espèce

d'arbre (*Virola michelii)* est utilisé comme témoin positif de l'extraction. On ajoute par ailleurs, un témoin négatif qui consiste à effectuer l'extraction à partir d'un mortier qui ne contient pas de matériel végétal.

2.4. Choix du matériel génétique utilisé

2.4.1. L'ADN chloroplastique

Il s'agit d'une molécule d'ADN double brin circulaire dont la taille varie de 120 à 217 Kb chez les angiospermes (Serror, 1988). La structure de l'ADNcp s'organise en 4 régions : 2 séquences « inversées répétées » (SIR) qui délimitent 2 régions « simple copie », une grande de 80 Kb (LSC) et une petite de 12 kb (SSC). Les grandes variations entre les espèces sont essentiellement dues à la présence ou l'absence des SIR. La molécule d'ADN du chloroplaste est très conservée dans le règne végétal, tant dans sa séquence que dans sa structure (Hartl et Clark, 1989). Il est plus abondant que l'ADN nucléaire. L'ADNcp en stock haploïde (absence de recombinaison) est transmis de façon clonale chez la plupart des angiospermes par la mère uniquement comme l'ont montré des analyses de croisements contrôlés (Birky, 1995). Or, la dispersion des végétaux s'effectue exclusivement par les graines. La répartition géographique actuelle des différents variants de l'ADN chloroplastique doit donc témoigner des voies de migration empruntées par l'espèce. De plus, son taux de mutation de 1,1 à $2{,}9.10^{-9}$ mutations par site nucléotidique et par génération est beaucoup plus faible que pour l'ADN nucléaire (Wolfe *et al.*, 1987). Donc la variabilité observée avec le génome chloroplastique est souvent issue des évènements mutationnels très anciens. Tout cela fait de l'ADNcp, un excellent outil pour retracer l'histoire ancienne des populations (McCauley, 1995). Du fait de son évolution sur le modèle clonal, et de son faible taux de mutation, l'ADN chloroplastique est un très bon outils pour les études phylogéographiques et son polymorphisme peut révéler des évènements historiques majeures (routes de migration, goulot d'étranglement) (Caron et al., 2000 ; Dutech et al., 2000).

2.4.2. Le choix des amorces pour le matériel frais

L'ADN chloroplastique codant est très utilisé dans les études de systématique des plantes (Shaw et Al., 2005). Ces régions sont soumises à de fortes contraintes évolutives et permettent de distinguer seulement les taxons d'un niveau assez haut (famille, genre). Pour notre étude portant sur la phylogéographie d'une espèce, le *Simarouba amara*, il est plus approprié de séquencer des régions non codantes de l'ADN chloroplastiques, qui sont plus variables. Un grand nombre de séquences intergéniques ont aussi été étudiées pour utilisées des phylogéographies dans des espèces d'Angiospermes. Dans une étude comparative, Shaw

et al., (2005) classent ces séquences intergéniques en fonction de leur variabilité et du nombre de caractères informatifs pour la phylogénie et la phylogéographie. Sur la base des résultats de Shaw et al. (2005), le polymorphisme intra population et inter populations des douze espèces sélectionnées dans le projet de phylogéographie comparé (SEEDSOURCE) a été analysé au niveau d'une dizaine de séquences intergéniques. Cette étude préliminaire réalisée avant mon arrivée, a permis de sélectionner deux régions intergéniques qui sont utilisées sur l'ensemble des douze espèces. Ces régions sont les séquences intergéniques trnH-psbA et trnC-ycf6, qui présentent une grande variabilité et qui montrent un bon succès d'amplification dans toutes les espèces étudiées. Ces deux séquences font environ 465 pb et 690 pb respectivement.

Les amorces universelles des séquences intergéniques chloroplastiques étudiées lors de mon stage sont celles présentées dans le tableau ci-dessous.

$TrnH^{GUG}$	CGC GCA TGG TGG ATT CAC AAT CC	Amorce sens
psbA	GTT ATG CAT GAA CGT AAT GCT C	Amorce antisens
$trnC^{GCA}$	CCA GTT CRA ATC YGG GTG	Amorce sens
ycf6	GCC CAA GCR AGA CTT ACT ATA TCC AT	Amorce antisens

Les amorces universelles (séquences conservées) $TrnH^{GUG}$ et psbA sont respectivement localisées aux extrémités des gènes trnH et psba. De même, les amorces universelles $TrnC^{GCA}$ et ycf6 sont respectivement localisées aux extrémités des gènes TrnC et ycf6 (Shaw et Al., 2005).

2.4.3. Le choix des amorces pour les échantillons herbier

L'ADN des feuilles extrait à partir d'herbier est généralement très dégradé. La qualité dépend de plusieurs paramètres comme l'âge des herbiers, les conditions de séchage, la présence de traitements insecticides et les conditions de stockage. Les PCR réalisées sur cet ADN en utilisant les amorces universelles TrnH/psbA et TrnC/YCF6 ne fonctionnent généralement pas. Pour contourner ce problème, des amorces internes aux séquences d'origines sont conçues pour permettre l'amplification de séquences plus petites. Ces nouvelles amorces n'ont plus le caractère universelle des amorces d'origine, elles sont spécifiques de l'espèce étudiée.

2.5. Test de polymorphisme

Le test de polymorphisme est réalisé sur du matériel non dégradé en choisissant des individus qui maximisent la chance d'observer des sites variables (choisis de pays différents) Sur la base des sites polymorphes observés et de leur niveau d'information (une substitution

entre deux populations a plus de poids qu'un variant observé dans un seul individu) des amorces internes aux régions TrnHpsbA et TrnCYcf6 ont été conçues.

2.6. Les PCR

2.6.1. PCR appliquée pour les échantillons provenant de matériel frais

Les réactifs utilisés pour la PCR proviennent du kit de PCR Biolabs. La préparation des mélanges réactionnels est réalisée dans une pièce réservée uniquement aux produits non amplifiés pour éviter les contaminations. Les concentrations finales utilisées sont les suivantes :

Concentration finales pour la PCR dans un volume de 15µl	
Amorces sens F	0,25µM
Amorces anti sens R	0,25µM
Tampon	1X
DNTp	0,15mM
Taq	0,08 U/µl

Le principal facteur ayant limité l'amplification des séquences a été la qualité de l'ADN extrait. Pour maximiser le nombre d'individus étudiés, j'ai donc du appliquer plusieurs niveaux de dilution de l'ADN mère allant de 1/5 à 1/100.

2.6.2. Optimisation de la PCR appliquée pour les échantillons de matériel d'herbier

L'optimisation de la PCR sur ADN provenant d'herbier passe par plusieurs étapes :

_ J'ai dans un Premier temps testé les amorces nouvellement conçues sur des échantillons frais. Le programme PCR identique à celui utilisé pour les amorces TrnH/Psba et TrnC/Ycf6 m'a permis de vérifier la spécificité des amorces internes. J'ai ensuite établi la température d'hybridation optimale pour chacune des nouvelles amorces en réalisant un gradient de température d'hybridation sur quelques individus. Dans une dernière étape, j'ai vérifié la qualité de l'amplification par migration du produit PCR sur un gel d'agarose, les profils avec une bande unique et intense correspondant aux températures d'hybridation optimales.

_ Une deuxième étape de mise au point permet de trouver les conditions de PCR optimales concernant

- les concentrations de $MgCl_2$,
- le nombre de cycles et les temps de chaque étape du programme PCR,
- les dilutions d'ADN,
- et l'utilisation de réactifs supplémentaires pouvant amplifier le signal.

- La concentration en $MgCl_2$:

En solution, l'ion Mg^{2+} est un cofacteur essentiel pour l'action de la Taq Polymérase. Ce cation bivalent interagit également avec les charges négatives de la chaine d'ADN, limitant ainsi les forces de répulsion entre brins d'ADN et favorisant la stabilité de l'hybridation. Ainsi, plus sa concentration est importante, plus l'hybridation est facilitée, qu'elle soit spécifique ou non. Il faut donc être vigilant quant à la spécificité du signal obtenu lorsque la concentration en Mgcl2 est augmentée (Mebarki-Société Qiagen). Les différents tests de concentrations que j'ai réalisé on montré un meilleur résultat lorsque la concentration est augmentée de 2 mM à 3,25mM dans le volume final de réaction de PCR. Par ailleurs, afin d'assurer la stabilité de la réaction, la température d'hybridation est aussi augmentée de 51 à 53°C.

- Programme PCR :

En conditions normales, 35 cycles étaient effectués. Pour les ADN d'herbier, j'ai augmenté le nombre de cycle jusqu'à 50 cycles. La Taq ayant une efficacité diminuant avec le temps, j'ai par ailleurs réduit la durée de chaque cycle. La phase de dénaturation initiale reste à 2 minutes, la phase de dénaturation est réduite à 30 secondes (au lieu de 45), la phase d'hybridation est réduite à 30 secondes (au lieu de 45), et la phase finale de synthèse est réduite à 45 secondes (au lieu de 2 minutes). Environ 1000 nucléotides par minute sont synthétisés par la Taq (comm. pers., Scotti I), les séquences à amplifier faisant de 140 pb à 220 pb environ, des durée d'élongation de l'ordre de 20 secondes sont adéquates. La durée totale de la réaction est réduite à 1H30 environ. Enfin, j'ai augmenté la quantité de Taq polymérase de 0,027 U à 0,042 U dans le volume réactionnel afin que l'enzyme ne soit pas un facteur limitant l'amplification au delà de 35 cycles. Le passage de 35 à 50 cycles augmente le rendement de la PCR, mais l'amplification reste faible, à peine visible sur gel d'agarose. La dernière phase d'optimisation de l'amplification à été la réalisation de deux PCR successives. Les bandes observées sur gel agarose paraissent suffisamment nettes pour entreprendre leur séquençage.

- Les dilutions PCR :

Les concentrations d'ADN extrait de feuilles d'herbiers sont trop faibles pour être comparées avec des témoins de concentration sur gel d'agarose. j'ai donc testé l'amplification des échantillons d'herbiers en utilisant une deux niveaux de dilution de l'ADN : $1/10^{ème}$ et au $1/50^{ème}$. Je n'ai pas observé de différence entre les PCR réalisées à partir des deux concentrations.

-Les purifications

Nous avons aussi testé l'efficacité d'une purification sur colonne en utilisant le kit « Genelute » PCR clean-Up ». Cette purification a été appliquée entre les deux PCR appliquées à nos échantillons. Les échantillons purifiés ont donné le même signal que les non purifiés. Cela est peut être du à la dilution engendrée lors de la phase d'élution des ADN après purification. Une quantité moindre d'éluant serait peut être intéressante.

2.7. Séquençage

2.7.1. Purification

La première étape est la purification par l'exosap (Amersham) qui contient une exonucléase I et une phosphatase alkaline. Son action permet d'éliminer les fragments d'ADN simple brin (amorces) et les nucléotides qui n'ont pas été intégrés lors de la réaction PCR. Le mélange réactionnel de la purification de PCR comprend 2µl d'Exosap et 10µl de produit de PCR. Le programme utilisé dans le thermocycleur contient une première phase de quinze minutes à 37°C, correspondant à la température d'activation des enzymes, puis une deuxième phase de quinze minutes à 80°C pour l'inactivation des enzymes.

2.7.2. Réaction de séquence

La deuxième étape concerne la réaction de séquence. La réaction de séquence est réalisée en suivant le protocole associé au kit Big dye terminator 3.1 d'Applied Biosystem. Le kit de séquence contient un mélange de désoxyribonucléotides (dNTP) et de didésoxyribonucléotides (ddNTP) en quantité optimisée pour l'obtention de séquences de l'ordre de 600 paires de bases. Les ddNTP qui stoppent l'élongation de l'ADN lors de leur intégration dans la chaîne de synthèse sont marqués par quatre fluorochromes différents, chacun spécifique d'une des bases constitutives de l'ADN (A, C, G, T). A l'issue de la réaction de séquence, des fragments de tailles différentes (de 1 paire de base à N la taille du fragment amplifié) sont obtenus, chacun marqué par un fluorochrome spécifique de la base incorporée au niveau du ddNTP qui a stoppé la synthèse (méthode de Sanger). Les concentrations utilisées lors de la réaction de séquence sont les suivantes :

Concentration finales pour la PCR dans un volume de 10µl	
Amorces F ou R	0,5µM
Tampon	1X
DNTp	0,15mM
Big Dye	1X

4µl de produits PCR sont ajoutés. Le programme du thermocycleur utilisé commence par une phase d'1 minute à 96°C, puis enchaine 35 cycles de 10 secondes à 96°C suivies de 5 secondes à 50°C, et de 4 minutes à 60°C.

2.7.3. Purification

Une dernière purification est nécessaire avant la migration le séquenceur automatique. Les réactions de séquence sont filtrées sur une colonne de séphadex d'Amersham Biosciences. La résine séphadex ® G50 utilisée permet d'une part de dessaler les échantillons et d'autre part, d'éliminer les amorces PCR et les nucléotides non incorporés (limite d'exclusion de 20 bases). La résine se présente sous forme de poudre qu'on étale dans une plaque à colonnes à l'aide d'un racloir. On transvase les colonnes de résine dans une plaque séphadex dont les puits sont munis d'un filtre. On ajoute 300µl d'eau et on laisse imbiber au moins 3 heures. On élimine ensuite le surplus de liquide en centrifugeant 5 minutes à 910g. Les produits issus de la réaction de séquençage sont dilués de moitié (on ajoute 10 ul d'eau) et sont transférés sur les colonnes Sephadex. Le produit purifié est récupéré dans une plaque destinée au séquenceur après une centrifugation pendant 5 minutes à 910g. Le volume minimal détecté par le séquenceur étant 10µl il est parfois nécessaire de rajouter un peu d'eau au réactif purifié. Enfin, pour limiter l'évaporation pendant la phase de séquençage quelques goutes de Formamide sont rajoutées dans chaque puit. La plaque est recouverte d'une carpette percée permettant aux capillaires du séquenceur de plonger dans les puits.

2.7.4. Séquençage automatique

Les séquences sont analysées à l'aide d'un séquenceur 16 capillaires ABI 3130 XL. Les fragments d'ADN chargés négativement migrent en fonction de leur taille, selon le principe de l'électrophorèse, de la cathode vers l'anode en parcourant le capillaire du séquenceur. Le séquenceur détecte la fluorescence sortant des colonnes de chromatographie, repérant ainsi les fragments d'ADN et leur taille précise. Les quatre nucléotides (A, C, G, T) marqués par un fluorochrome différent sont détectés par une diode laser. Un algorithme permet alors de transcrire les données brutes d'absorbance en données de séquence. Le logiciel utilisé est Genetic Analyzer Data Collection Software v 3.0.

2.7.5. Correction des séquences

Les chromatogrammes sont ensuite vérifiés manuellement et analysées avec le logiciel Codoncode Aligner v 1.6.3. Ce logiciel permet d'aligner plusieurs séquences semblables. Les sites polymorphiques sont ainsi repérés.

2.8. Analyses

En raison de l'absence de recombinaison dans le génome chloroplastique, les séquences obtenues à partir des deux paires d'amorces ont été concaténées et sont donc considérées comme une séquence unique de 1033 paires de bases.

2.8.2. Construction du réseau d'haplotypes

Le logiciel Arlequin a été utilisé pour calculer une matrice de distance entre les différents haplotypes à partir du polymorphisme détecté au niveau des sites nucléotidiques. A partir de cette matrice, un réseau d'haplotypes est construit de la façon la plus parcimonieuse en utilisant la méthode « Minimum Spanning Tree ». Les relations entre les haplotypes sont déterminées de façon à minimiser le nombre de mutations qui les séparent. Afin de tenir compte d'une hétérogénéité du taux de substitution dans la séquence, notamment en raison de la présence d'un microsatellite, un poids différent est attribué aux différentes mutations observées lors de la réalisation de la matrice de distance. Les délétions sont pondérées par un facteur 1 alors que les transitions et les transversions sont respectivement pondérées par un facteur 2 et 2,5. Par ailleurs, une insertion/délétion composée de 4 nucléotides a été observée chez certains individus. L'absence de polymorphisme à l'intérieur de l'insertion nous a conduits à considérer ce type de polymorphisme comme un événement mutationnel unique.

2.8.3. Analyse par clades emboîtés : NCPA (Nested Clustering Phylogeny Analysis)

Ce test est réalisé avec le progamme Geodis inclus dans le logiciel ANeCA_v1.1.

2.8.3.1. Construction d'un réseau d'haplotypes (clade de niveau 0)-TCS

Le réseau d'haplotype est déterminé en utilisant le critère de parcimonie. Lorsque des boucles apparaissent dans le réseau, en raison d'un nombre équivalent de mutations séparant plusieurs haplotypes, le choix des branches à supprimer se base sur deux critères : 1- Les haplotypes rares sont plus susceptibles d'être trouvés en bout de branches et les haplotypes communs à l'intérieur. 2- Un singleton est plus probablement connecté aux haplotypes provenant de la même population que d'une population éloignée géographiquement (Panchal, 2007 ; Clement et al., 2000 ; Posada et al., 2000 ; Madrulyn et al., 2001)

2.8.3.2. Regroupement des clades de niveaux supérieurs

Une fois le réseau établi de manière optimale (élimination des boucles), les haplotypes (niveau 0) sont groupés dans des clades de niveau 1. Les clades de niveau 0 sont associés en prenant par ordre de priorité, les haplotypes situés à l'extrémité du réseau, puis en second lieu, les haplotypes qui ont une seule mutation de différence avec le groupe de l'extrémité. La même démarche est suivie pour la construction des clades de niveau 2 qui regroupent les clades de niveau 1 qui présentent une seule mutation de différence. Et ainsi de suite jusqu'au dernier niveau qui regroupe toute le réseau.

2.8.3.3. Test sur la structuration géographique de la phylogénie.

La structuration géographique de la phylogénie est estimée par un test d'indépendance basée sur le chi deux. Les informations sur la localisation des individus sont traitées comme des données catégorielles (on ne prend pas en compte les distances). Les valeurs de Khi² sont calculées à partir des tables de contingence opposant les clades (lignes) aux localisations géographiques (colonne) (Templeton et al., 1995 ; Mardulyn et al., 2001). La formule utilisée est la suivante :

$$\mathbf{Khi^2} = \sum_{i}^{L} = \mathbf{L} \sum_{K=L}^{i} \frac{(n_{ij} - n_i \cdot p_j)^2}{n_i \cdot p_j}$$

Où L est le nombre de localité dans le clade testé et K, le nombre d'haplotypes dans l'échantillon total, n_i est la taille de l'échantillon dans la localité i, n_{ij} est le nombre de fois où l'on observe l'haplotype j dans la localité i, et p_j est la fréquence de l'haplotype j dans l'échantillon total. C'est la méthode traditionnelle utilisée par Nei pour tester la différenciation génétique (Hudson et al., 1992).

La significativité du test est obtenue en calculant 10 000 fois la valeur de khi² après 10 000 redistributions aléatoires des individus dans le tableau de contingence. Si la valeur de khi² calculée après permutations des données est inférieure à la valeur calculée dans plus de 95% des cas, alors on rejette l'hypothèse nulle d'absence de structuration géographique des haplotypes ; par conséquence on conclue à la présence d'une structuration des haplotypes en fonction de la géographie.

Ce test de contingence peut détecter des associations significatives entre les regroupements géographiques et cladistiques, cependant, il ne prend pas en compte ni la taille des populations, ni la distance entre chaque populations (Templeton et al., 1994).

2.8.3.4. Test sur la distance de dispersion : Calcul des valeurs Dc et Dn

L'intérêt de ces valeurs réside dans le fait qu'elles tiennent compte de la distance entre les populations : les localisations sont ici des variables continues. Dc

correspond à l'étendue géographique d'un clade, tandis que Dn correspond à l'éloignement des individus d'un clade par rapport à un autre clade de même niveau avec lequel il est lié dans le niveau supérieur (Posada et al., 2006).

Soit une zone géographique A comprenant les individus d'un clade X d'un niveau n, qui est lui-même compris dans un clade Y de niveau n+1. Pour le calcul de Dc, la première étape consiste au calcul du centre géographique de tous les individus inclus dans le clade X. Une méthode d'estimation de ce centre géographique est de calculer la moyenne des longitudes et des latitudes sur tous les individus. Ensuite, pour chaque individu compris dans le clade X, on calcule la distance qui le sépare du centre géographique du clade X. La moyenne de ces distances sur tous les individus du clade X, donne la valeur Dc du clade X (Dc(X)).

Pour le calcul de Dn, le centre géographique du clade Y est déterminé. Puis pour chaque individus compris dans le clade X, on calcule la distance qui le sépare du centre géographique du clade Y. La moyenne de ces distances pour tous les individus du clade X, donne la valeur Dn(X) (Templeton et al., 1995). Pour savoir si les distances observées sont significativement différentes des distances que l'on observerait si la répartition des haplotypes était aléatoire, ces valeurs sont recalculées après 10000 permutations aléatoires dans le tableau de contingence opposant les clades aux localisations.

La comparaison des valeurs de Dc et Dn permet de faire des hypothèses sur la distance de dispersion : Si Dn>Dc, la distance de dispersion est élevée. En effet, si un haplotype est observé dans un petit nombre de populations géographiquement proches les unes des autres, et que cet haplotype est phylogénétiquement proche d'un haplotype très distant géographiquement, cela peut s'expliquer par une forte capacité de dispersion dans l'espèce étudié.

2.8.3.5. Tests des Hypothèses historiques : Calcul du I-T

A l'intérieur de chaque clade, on distingue maintenant les distances Dc du ou des sous clades présents à l'**intérieur** du réseau (qui ont le plus de connexions avec le reste du réseau), des distances Dc du ou des sous clades présents à l'**extrémité** du réseau (qui ont le moins de connexions avec le reste du réseau). De même, on distingue les distances Dn des clades intérieurs et extérieurs. Pour chaque clade Y, on calcule la moyenne des valeurs de Dc des sous clades intérieurs, puis on lui soustrait la valeur moyenne des Dc des sous clades extérieurs. Le même calcul est effectué pour Dn (Posada et al., 2006 ; Madrulyn, 2001). On obtient deux valeurs (I-T)c et (I-T)n. Les significativités de ces deux valeurs sont testées avec des permutations aléatoires du tableau de contingence.

En partant de la supposition que les clades intérieurs sont plus anciens que les clades des extrémités, différents scénarios historiques peuvent être proposés.

2.8.3.6. Principes généraux sur lesquels se basent l'analyse NCPA

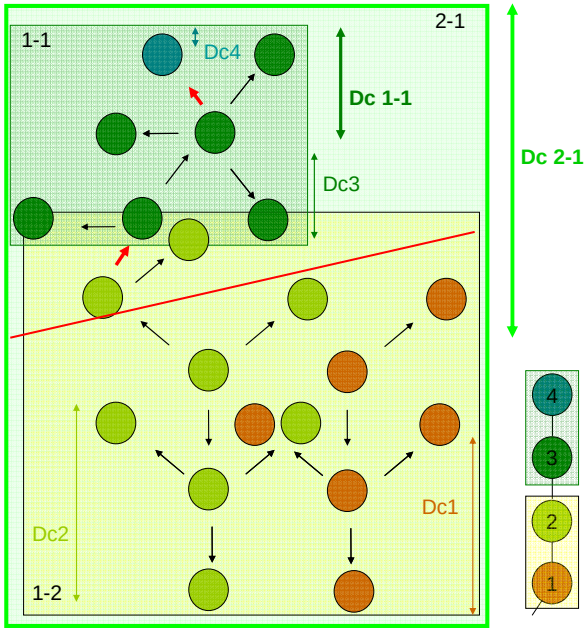
A l'issue de l'estimation des différents paramètres de distance décrits dans le paragraphe précédent associée à leur significativité, une clé d'inférence est utilisée pour proposer des scénarios historiques (restriction de flux de gènes, fragmentation passée, ou expansion de population) qui expliqueraient les patrons de distribution des distances. Cette clé publiée par Templeton et et al., (1995) repose sur les principes suivants :

Restriction aux flux de gènes

La signature d'une restriction aux flux de gènes dans un clade peut être observée de trois façons :

- Tout d'abord en regardant les valeurs de Dc des clades de niveaux successifs. En présence de restriction au flux de gènes on s'attend à ce que l'étendue géographique d'un clade soit corrélée avec son âge. Le clade de niveau supérieur à Dc(x) est donc plus étendu que Dc(x) lui-même. De ce fait, on arrive à repérer une barrière de flux de gène par une valeur de Dc(x), bien plus inférieure à une valeur de Dc(x+1).
- Une deuxième façon de repérer un clade qui a subit une restriction de flux de gènes consiste à comparer les valeurs de Dc des clades d'extrémités (les plus récents) des valeurs Dc des clades de l'intérieur (des haplotypes ancestraux). Suivant le même principe que précédemment, cela se traduit par des Dc plus faibles pour les clades de l'extrémité (plus jeune) par rapport aux clades de l'intérieur (plus ancien). La valeur I-T(c) prend une valeur significativement grande.
- La troisième méthode consiste à comparer les Dc avec les Dn d'un même clade : En présence d'une restriction aux flux de gènes associée à une faible distance de dispersion, les haplotypes dérivés et l' haplotype ancestral se trouvent à proximité géographique. Cela implique que les centres géographiques des clades récents et anciens sont proches les uns des autres. Cela se traduit par des valeurs de Dc et Dn proches (Templeton, 1998).

Figure 3 : Représentation d'un flux de gène restreint par l'isolation par la distance.



A droite est présenté le réseau d'haplotypes, avec l'haplotype 4 qui dérive de l'haplotype 3 qui dérive lui-même de l'haplotype 2 qui dérive lui-même de l'haplotype 1. La partie gauche représente la distribution des haplotypes dans l'espace géographique. Pour la clarté de la représentation, l'isolement par la distance de l'haplotype 2 est représenté par le trait rouge. Les haplotypes sont regroupés de telle façon que le clade 1-1 regroupe les haplotypes 3 et 4. Le clade 1-2 regroupe les haplotypes 1 et 2 et le clade 2-1 regroupe les clades 1-1 et 1-2. Le schéma illustre les différentes façons de mettre en avant la présence d'une restriction aux flux de gènes :
Dc1-1 << Dc 2-1
I-T (c) >>0
Dc1-1<<Dc1-2
Les flèches rouges représentent les évènements mutationnels récents, et les flèches noires représentent les phénomènes d'expansion. Le clade 1-1 est écarté pour des raisons de lisibilité du schéma, mais il pourrait être compris dans le clade 2-1.

Cas d'une fragmentation passée

Les haplotypes plus récents, apparus dans la zone fragmentée, sont d'une part génétiquement proches et d'autre part géographiquement proches entre eux. Cela engendre une valeur de Dc très faible pour le clade concerné, mais sa valeur de Dn est généralement très élevée (ceci est remarqué surtout dans les cas où des zones isolées sont éloignées géographiquement les unes des autres, ce qui n'est pas toujours le cas) (Templeton, 1995). Les clades concernés sont parfois connectés au reste du réseau par des branches plus longues qu'en moyenne.

Cas d'une expansion de population

Si la population est en expansion, alors, on s'attend à ce que les étendues des haplotypes les plus récents soient les plus élevées (cas inverse d'une restriction de flux de gènes). Les valeurs Dc et Dn des clades en expansion sont donc plus élevées que ceux des clades ancestraux. Par ailleurs, les valeurs I-T peuvent permettre de distinguer des expansions réalisées en continuité sur de courtes distances et des colonisations réalisées par sauts de puce sur de longues distances : I-T(c) est significativement petit en cas d'expansion continue, et

c'est I-T(n) qui sera significativement petit en cas de colonisation à longue distance. Les Dn des clades en expansion peuvent être élevés si l'expansion se fait sur une très longue distance.

2.8.4. Calcul des indices de diversité génétique

La richesse allélique *A* correspond au nombre d'allèles différents au sein d'une population. Ce paramètre présente le défaut de donner le même poids aux allèles rares et aux allèles les plus communs. Par ailleurs il est fortement dépendant de la taille de l'échantillon. Ce dernier problème est contournable en utilisant la méthode de raréfaction qui estime la richesse allélique A_r pour des effectifs standardisés (Comps *et al.*, 2001).

La diversité génétique de Nei (1973b) H_e, appliquée au génome haploide, estime la probabilité de tirer deux haplotypes différents dans la population

$$H_e = 1 - \sum_{i=1}^{n} \left(p_i\right)^2$$

Avec p_i la fréquence de chaque haplotype, n le nombre d'haplotypes observés.

. La diversité génétique de Nei est donc formellement identique à l'indice de diversité Simpson (1949) classiquement utilisé en écologie.

2.8.5. Analyse de la structuration génétique avec le logiciel Arlequin

2.8.5.1. Test AMOVA (Analysis of MOlecular VAriance)

On utilise le programme AMOVA inclut dans le logiciel Arlequin ver3.11. L'information sur la divergence entre les haplotypes est résumée sous la forme d'une matrice de distance qui est ensuite prise en compte dans l'analyse de variance Cette méthode permet d'estimer la diversité génétique et d'estimer les indices différenciation génétique à plusieurs niveaux hiérarchiques en utilisant d'une part les informations sur le contenu allélique des haplotypes et d'autre part, leurs fréquences (Excoffier et al., 1992).

Soit Va, la variance entre groupes, Vb, la variance entre populations à l'intérieur des groupes, Vc, la variance à l'intérieur des populations, et Vt, la variance totale (Va+Vb+Vc). L'indice de fixation qui nous permet d'étudier la structuration entre groupe est le paramètre F_{CT} =Va/Vt. L'indice qui nous rend comte de la structuration à l'intérieur des groupes est le paramètre F_{SC} = Vb/(Vb+Vc). L'indice F_{ST} = Va+Vb/Vt est un indice de différenciation global. La significativité des variances obtenues est testées par des permutations, séparément pour chaque niveau. Parmi ces trois indices, seuls les F_{CT} et F_{SC} sont présentés:

- F_{CT} nous informe sur le niveau de structuration entre groupes. Cela nous permet de confronter les résultats de l'AMOVA avec ceux issus du tableau de contingence testé dans l'analyse NCPA.

F_{SC} nous indique s'il y a structuration génétique entre les populations dans les groupes. Cette information sera prise en compte pour réaliser le test de Tajima présenté plus bas.

L'AMOVA a été réalisée en considérant cinq subdivisions géographiques.

- Est des Andes / Ouest des Andes.
- Amérique Centrale /Amérique du Sud.
- subdivision par pays.
- Subdivision des zones géographiques au sein de la Guyane.
- Ouest des Andes / groupe Equateur-Pérou.

2.8.5.2. Test TAJIMA

Le logiciel Arlequin v 3.01 a été utilisé pour effectuer le test de Tajima. Le test de Tajima s'applique aux données de séquences.

C'est un test de neutralité basé sur la diversité moléculaire des échantillons. le principe de ce test est de comparer l'estimation du paramètre de mutation $\theta = 4Nu$ obtenue à partir du nombre de sites polymorphes S (θ_S) à celle obtenue à partir du nombre moyen de différences entre deux séquences π (θ_π). A l'équilibre mutation-dérive θ_S est égal à θ_π.

La statistique du D de Tajima selon la méthode proposée par Tajima (1989) est :

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_S}{\sqrt{\mathrm{Var}(\hat{\theta}_\pi - \hat{\theta}_S)}}$$

En cas d'expansion de population, le nombre de sites polymorphes augmente relativement rapidement par rapport au nombre de différences entre séquences (en cas d'expansion, les nouveaux allèles ont une faible fréquence et donc une faible influence sur π). D sera alors négatif. En cas de contraction de population, le nombre de site polymorphe diminue, D sera positif.

2.8.5.3. Test Fs de Fu

Ce test est semblable à Tajima, et détecte parfois plus facilement les effets d'extension ou restriction de population. k est le nombre d'haplotype observé. $\hat{\theta}_\pi$ est l'estimation du paramètre de mutation à partir du nombre de site polymorphes entre les séquences. k_0 est le nombre d'haplotypes attendus sous le modèle neutre en choisissant $\Theta\hat{}\ \pi$

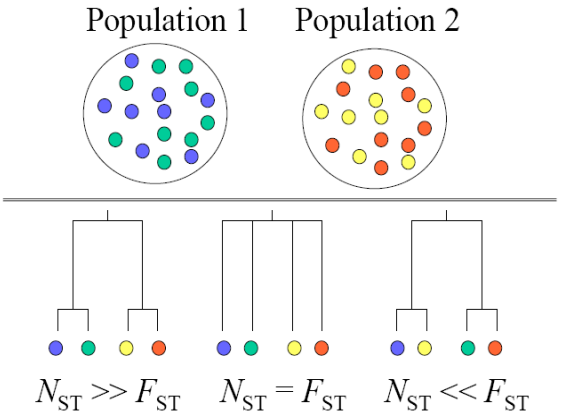
comme paramètre de diversité. Des simulations sont effectuées sous le modèle neutre et à chaque simulation une nouvelle valeur de k_0 est attribuée. S' est la probabilité d'avoir un nombre d'haplotypes attendus sous le modèle neutre (k_0) plus grand ou égal que notre nombre d'haplotype observé (Fu, 1997). $F_S=\ln(S'/(1-S'))$

En cas d'expansion, le nombre d'haplotypes k est élevé mais cette diversité haplotypique influe peu sur Θ^π (estimée à partir des différences entre séquences) puisque les nouveaux haplotypes sont en faibles fréquences. Les valeurs k_0 induites sous le modèle neutre vont donc être plus faibles que celles observées. S' va être proche de 0, le Fs est négatif (la fonction ln(S'/(1-S')) est une fonction croissante qui devient positive lorsque S'>0,5).

2.9. Analyse de la structuration géographique par comparaison des coefficients de différenciation.

La méthode utilisée est celle proposée par Pons et Petit (1996) qui consiste à comparer le coefficient de différenciation entre populations (F_{ST}) avec un coefficient de différenciation génétique qui prend en compte la similarité entre les haplotypes (N_{ST}). En présence d'un signal phylogéographique, la diversité génétique est structurée géographiquement, les haplotypes phylogénétiquement les plus proches se retrouvent plus souvent dans les mêmes populations. La valeur de N_{ST} devient alors supérieure à la valeur de F_{ST}. La significativité de la différence observée entre le N_{ST} et F_{ST} est obtenue en réalisant 1000 permutations de l'identité des haplotypes sur les populations. Les logiciels Spagedi (Hardy O et Vekemans X, 2002) et PERMUT (Petit R et al., 2002) permettent de réaliser ces tests. Dans cette étude j'ai seulement utilisé PERMUT.

Figure 4 : Illustration de la correspondance entre les phylogénies des haplotypes et leur distribution géographique



1) Présence d'un signal phylogéographique
2) Structuration génétique sans patron phylogéographique
3) Structuration génétique mais les individus les plus proches génétiquement se trouvent dans des populations différentes.
D'après Pons et Petit (1996).

3. Résultats

3.1. Polymorphisme dans les séquences et nombre d'haplotypes

Un total de 145 séquences qui se répartissent en 24 populations est obtenu. 13 sites polymorphiques sont repérés sur un total de 1033 paires de bases. 11 haplotypes résultent des différentes combinaisons observées aux sites polymorphiques.

Un individu provenant de nos extractions d'herbier a pu être amplifié à partir de nouvelles amorces dessinées. Pour cet individu du Venezuela, douze sites polymorphiques sur treize ont pu être lus. Aucune séquence entière n'est obtenue, alors aucun individu d'herbier n'est pris en compte dans les analyses de structure des populations.

3.2. Répartition des haplotypes

* Au niveau continental

Le nombre d'échantillons et d'haplotypes pour chaque population étudiée est représenté sur la figure 5. En Equateur, deux populations sont échantillonnées, une à l'Ouest de la barrière des Andes, et l'autre à l'Est. L'haplotype retrouvé à l'Ouest est aussi principalement retrouvé en Amérique Centrale.

* En Guyane

Le nombre d'haplotypes pour chaque population de Guyane est représenté sur la figure 6. Les populations de Bélizon et de route de l'Est sont celles où le plus d'haplotypes ont été retrouvés. Les haplotypes de 1 à 4 retrouvés en Guyane ont successivement une différence d'insertion délétion d'une Thymine au niveau d'un poly-T. L'haplotype 2, au sud, présente une Thymine de plus que l'haplotype 3 retrouvé à Saul au centre de la Guyane, qui lui présente une Thymine de plus que l'haplotype 4 retrouvé au Nord Est. Sans prendre en compte l'haplotype 1, on remarque que plus on avance vers le Nord-Est, moins la séquence est chargée en Thymine.

Tableau 2 : Les haplotypes de Guyane

Haplotype	Polymorphisme observé à l'extrémité du polyT	Localisation
H1	TTT	Nord Guyane + Belizon
H2	TT-	Extrême Sud
H3	T--	Sud ; Centre ; Bélizon
H4	---	Belizon

3.3. La diversité génétique

Les pays d'Amérique Centrale et le Brésil contiennent le plus de diversité. Le tableau 3 ci-dessous indique les nombres d'allèles en tenant comptes des différents effectifs et les indice de Nei.

Tableau 3 : Diversité génétique de séquences chloroplastiques chez le *S.amara.*

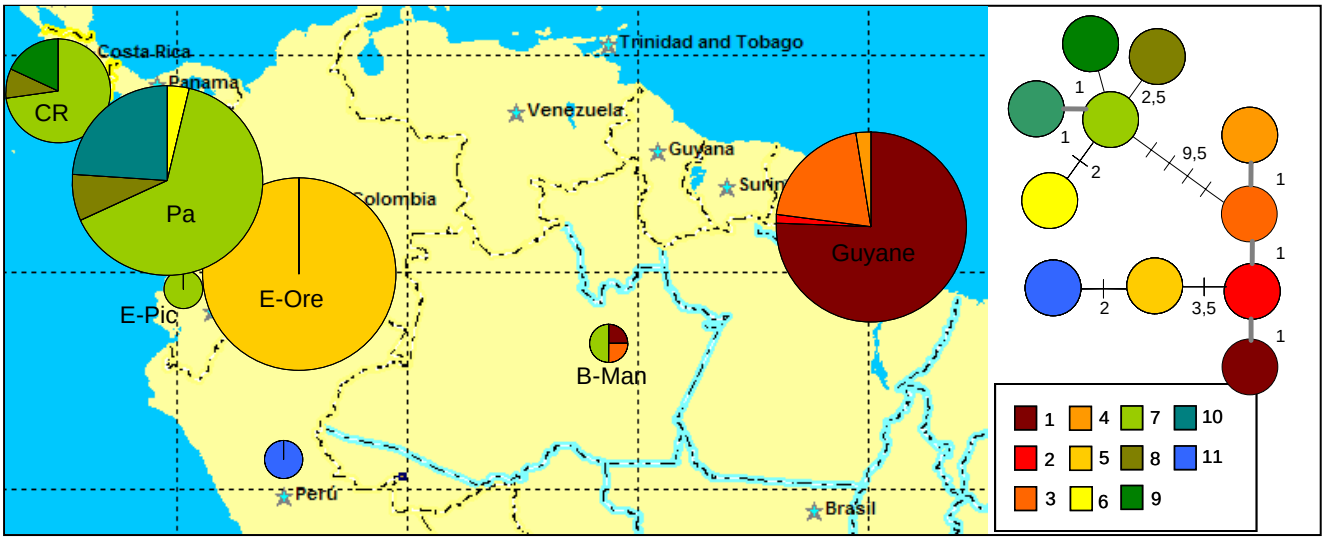
	Bresil	Panama	CostaRica	Guyane	Equateur	Perou
N_I	4	25	11	78	26	1
N_H	3	4	3	4	2	1
N_{Hr}	na	3,12	3	2,34	1,42	na
Indice de Nei	0,83+/- 0,22	0,54+/- 0,0928	0,47+/- 0,16	0,38 +/- 0,05	0,07 +/- 0,06	0,00 +/- 0,00

Avec N_I le nombre d'individus dans la population, N_H, le nombre d'haplotypes observés et N_{Hr}, le nombre d'haplotypes estimé à partir de la méthode raréfact, en considérant une taille standardisée des populations égale à 11 individus (taille de la population du Costa-Rica).

3.4. Relation phylogénétique entre les haplotypes

Le réseau d'haplotypes proposé par le programme Arlequin en utilisant la méthode « Minimum Spanning Tree » est représenté à gauche de la figure 5. Ce réseau d'haplotypes met en évidence deux lignées génétiques. La première comprend les haplotypes d'Amérique Centrale, de l'Ouest de l'Equateur et deux individus du Brésil (haplotype de 6 à 10) ; la deuxième, qui se différencie par cinq mutations, comprend les haplotypes de Guyane, du Brésil et de l'est de l'équateur.

Figure 5 : Répartition des haplotypes obtenus en Amérique Centrale et Amérique du Sud avec les séquences TrnH-PsbA et TrnC-Ycf6



A gauche, les camemberts représentent la fréquence de chaque haplotype observé par pays. Seul l'Equateur est représenté par ses deux populations car on privilégie l'effet de la barrière géographique des Andes plutôt que les frontières politiques. Légende : Costa Rica (CR) ; Panama (Pa) ; Equateur (E) ; Orellana (Ore) ; Pichincha (Pic) ; Brésil (B) ; Manaus (Man). Le diamètre des camemberts est proportionnel au nombre d'individus échantillonnés. Pour la lisibilité de la figure, une taille minimale et maximale est attribuée lorsque les populations contiennent respectivement moins de 4 ou plus de 20 individus.

A droite, est représenté le réseau d'haplotype obtenu de la façon la plus parcimonieuse en attribuant un coût de 2 pour les transitions, un coût de 2,5 est attribué pour les transversions, et un coût de 1 pour les délétions. Les distances estimées sont indiquées au niveau de chaque branche. Le nombre de segments représente en revanche le nombre minimum de mutations observées entre les différents haplotypes du réseau. Les haplotypes qui résultent de l'accumulation de mutations à l'intérieur d'un microsatellite (poly T) sont liées par un trait gris.

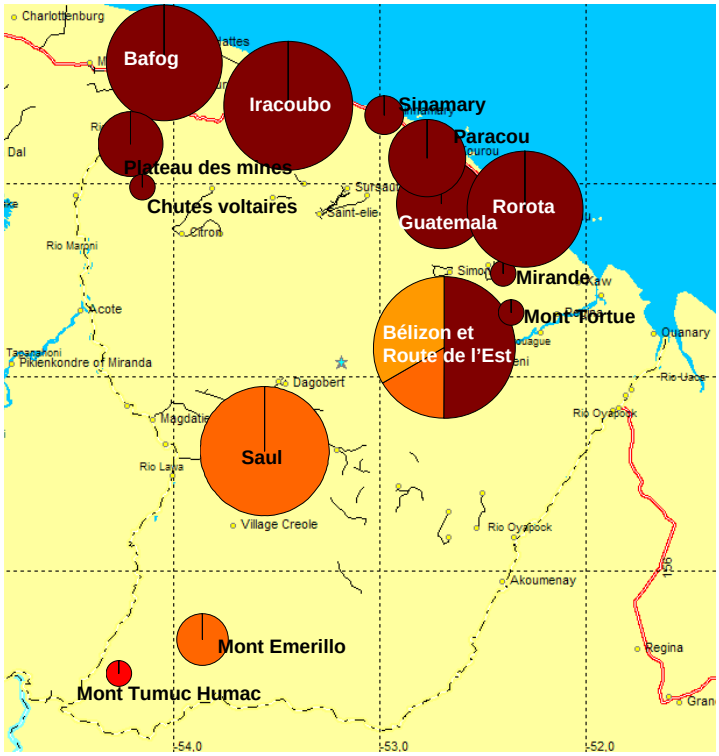


Figure 6 : Répartition des haplotypes obtenus en Guyane française à partir des séquences TrnH-PsbA et TrnC-Ycf6 :
Les camemberts représentent la fréquence de chaque haplotypes dans les populations guyanaises. Les noms des populations sont indiqués sur la carte. La taille des rayons est proportionnel au nombre de séquences lues.

3.5. L'analyse de clades emboîtés

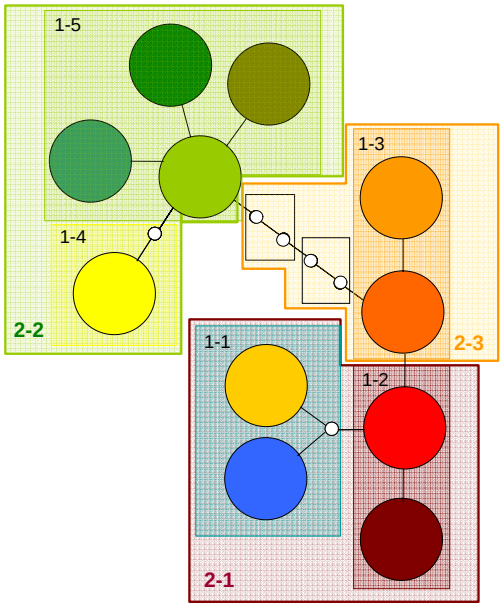
3.5.1. Le réseau de clade emboité

La figure 7 présente le réseau d'haplotype construit avec le logiciel TCS. Une boucle ambiguë a été enlevée suivant la règle de Templeton et al. (1998) (cf matériels et méthodes). Le réseau d'haplotype résultant est très légèrement différent de celui construit avec la méthode du Minimum Spanning Tree. La différence réside au niveau de l'haplotype 2 guyanais qui est maintenant lié à un haplotype ancestral manquant qui serait un ancêtre commun aux deux haplotypes 5 et 11 du Pérou et de l'Ouest de l'Equateur.

3.5.2 Résultats de l'analyse NCPA

Les tests de permutations sur les tableaux de contingence mettent en avant une structuration géographique significative des haplotypes dans quatre clades : clades 1-1, 1-5, 2-1 et le cladogramme total. Les résultats de l'analyse des clades emboîtés qui, à la différence du test de contingence, prend en compte les distances géographiques, sont donnés dans la figure 8. Par ailleurs, les scénarios proposés à l'issue de la clé d'inférence sont résumés dans le tableau 4. Des hypothèses de scénarios sont proposées pour les clades 1-2, 1-5, 2-1 et le cladogramme total.

Figure 7 : schéma du clade emboité proposé par le logiciel NCPA.



On retrouve avec le clade 2-2, la lignée d'Amérique Centrale qui se rattache à l'haplotype de l'Ouest de l'Equateur. L'haplotype le plus au Nord de la Guyane, et celui le plus au Sud (clade 1-2) se regroupent avec les haplotypes de l'Equateur de l'Ouest et du Pérou (clade 1-1) pour donner le clade 2-1. L'autre partie des haplotypes guyanais, (clade 1-3) se regroupent avec des haplotypes intermédiaires à la lignée de l'Amérique Centrale. Un échantillonnage couvrant le nord du Brésil, le Suriname, le Guyana, le Venezuela et la Colombie pourrait peut être nous permettre d'attribuer des localisations à ces hap lotypes intermédiaires.

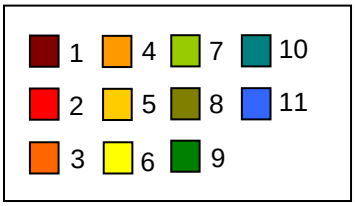


Tableau 4 : interprétations des valeurs calculées lors de l'analyse de clades emboités

clade	chaîne d'inférence	Hypothèse de scénario
1-1	1-19-20 NO	Région incorrectement échantillonnée
1-2	1-2-3-4 NO	Restriction de flux de gènes avec isolement par la distance
1-5	1-2-3-4 NO	Restriction de flux de gènes avec isolement par la distance
2-1	1-2-11-12 NO	Expansion continue
Total	1-2-11-12-13-14 NO	Colonisation sur longue distance et/ou fragmentation passée

En raison, d'un échantillonnage insuffisant dans le clade 1-1 (un seul individu dans chacune des deux populations) la clé d'inférence ne peut aboutir à aucun scénario.

Pour le clade 1-2, il est surprenant que la valeur trouvée pour le I-T(c) du clade 1-2 soit significativement grande alors que celle-ci est négative. Une erreur du logiciel est à envisager. Ce point sera discuté plus loin.

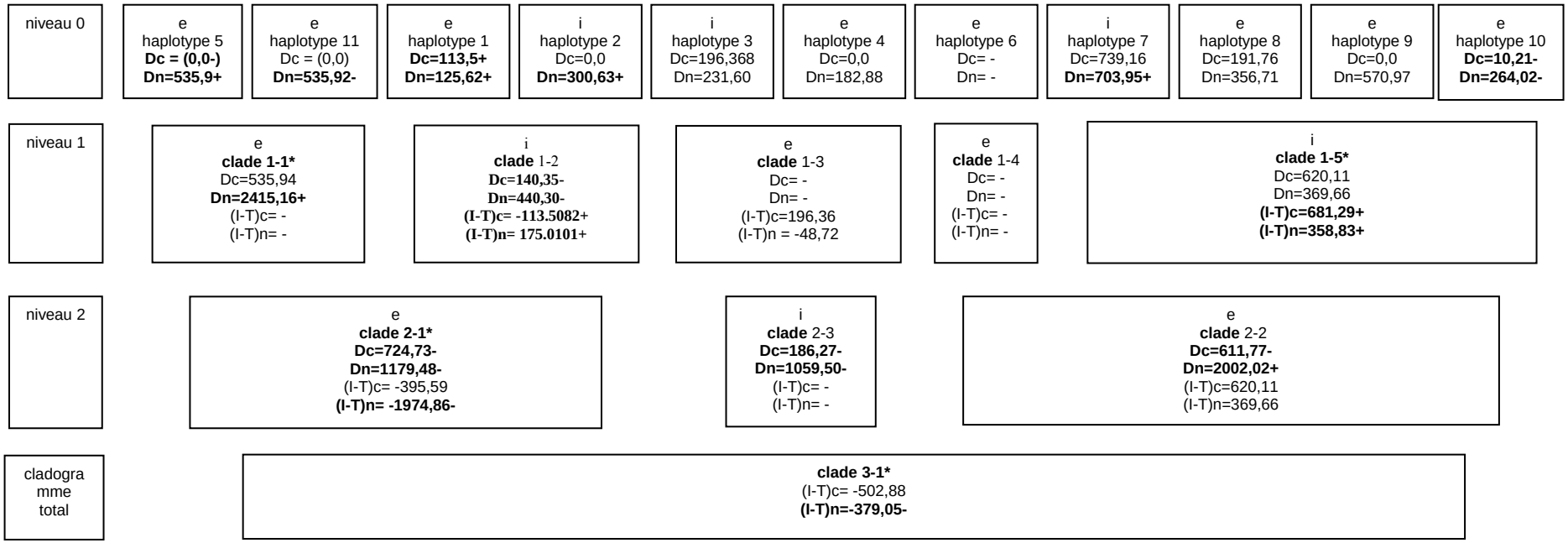
Dans le clade 1-5, la structuration géographique est expliquée par une restriction aux flux de gènes avec isolement par la distance.

Alors que dans le clade 2-1 la structuration géographique s'expliquerait par une variation de la taille des populations suite à une expansion continue.

Enfin deux scénarios sont proposés pour expliquer la structuration géographique observée au niveau du cladogramme global, le premier est une colonisation sur de longues distance le deuxième une fragmentation ancienne des populations.

.

Figure 8 : Récapitulatif des valeurs calculées par le programme NCPA.

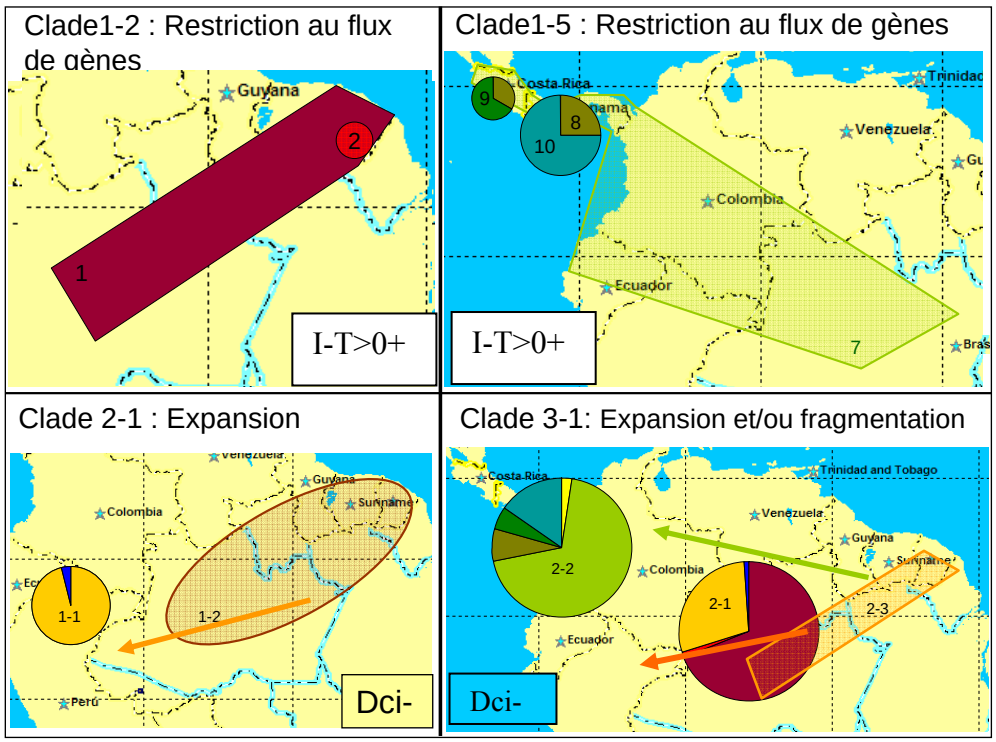


Les lettre e et i indiquent si l'haplotype (ou le clade) est en position interne (i) ou externe (e) dans le réseau.
Les valeurs significatives sont représentées en gras. Les valeurs significativement grandes sont indiquées par un signe +, les valeurs significativement petites sont signalées par un signe -. Les astérisques positionnés après les noms des clades indiquent que les valeurs de khi² du test de permutations dans le tableau de contingence sont significatives. Les valeurs de Dc entre parenthèses sont celles pour lesquelles on ne peut pas se rendre compte de l'étendue du clade du fait d'un trop faible échantillonnage. Les valeurs Dc, Dn, et I-T sont notées (-) lorsqu'elles ne peuvent pas être calculées. Pour le clade 1-4, les valeurs n'existent pas puisque que le clade est représenté par un seul haplotype. Le clade 1-1 ne permet pas de calculer les valeurs I-T(Dc) et I-T(Dn) puisque les haplotypes sont tous extérieurs. Les valeurs associées aux clades 2-3 manquent aussi car les haplotypes sont tous dans un clade extérieur.

La figure 9 illustre pour chaque clade, les distributions géographiques des sous clades intérieurs et des sous clades de l'extrémité du réseau. Ces figures mettent en relief les patrons phylogéographiques trouvés par l'analyse de clade emboité.

En raison des critiques multiples associées au test NCPA, notamment pour les faux positifs qu'il peut inférer (Petit, 2007) ; des méthodes alternatives et complémentaires ont été utilisées.

Figure 9 : Schématisation des inférences de l'analyse NCPA.



Les clades intérieurs sont représentés par les étendues de couleurs pales et les clades extérieurs par les graphiques en couleurs opaques. Pour les clades extérieurs, la fréquence des haplotypes qui les composent est représentée dans les diagrammes en camembert.
I-T(c) correspond à la moyenne des distances Dc calculées au niveau clades intérieurs à laquelle on soustrait la moyenne des distances Dc obtenues pour les clades extérieurs (cf. matériels et méthodes). Le signe + est placé après une valeur significativement positive, et le signe - après une valeur significativement négative.

3.6. Tests de structuration

3.6.1. Test AMOVA (Analyse Moléculaire de Variance)-Arlequin

Le test AMOVA a été en premier lieu réalisé en confrontant les populations situées à l'Ouest des Andes et les populations situées à l'Est, puis en confrontant les populations de l'Amérique Centrale et d'Amérique du Sud (Tableau 5).

Tableau 5 : Valeurs des F_{CT} et F_{SC} estimées par l'analyse AMOVA

Groupes Source de variation	Ouest-Andes/Est-Andes	Am.Centrale/ Am.Sud	FG/B/E/Pe/Pa/CR	Guyane NO/NE/S	Ouest des Andes/Pe EstE
Inter-groupes (F_{CT})	0.78074***	0.78768***	0.83610***	0.83350***	0.91936*
Intra-groupe et Inter-population (F_{SC})	0.73611***	0.73528***	0.51606***	0.13835	0.20717*

Une significativité à 1.10^{-3} est représentée par ***. Une significativité à 1.10^{-2} est représentée par **. Une significativité à $5.\ 10^{-2}$ est représentée par *. Avec FG=Guyane française, B=Brésil, E=Equateur, Pe=Pérou, Pa=Panama, CR=Costa Rica, NO=Nord Ouest, NE=Nord Est, S=Sud.

La variance génétique totale est expliquée à plus de 70 % par la variance génétique entre le groupe de l'Ouest des Andes et celui de l'Est ou encore, entre le groupe de l'Amérique Centrale et celui de l'Amérique du Sud. Il existe donc une structuration forte et significative des populations à l'échelle du continent. La valeur forte des F_{SC} (F_{SC}= 0,73) montre qu'il existe aussi une structuration à plus petite échelle entre les populations de ces groupes.

Dans un deuxième temps les tests AMOVA réalisés à l'échelle des pays mettent en avant une structuration génétique significative entre les pays. A l'intérieur des pays, les populations apparaissent encore structurées génétiquement (F_{SC}=0,52). Dans un troisième temps, les AMOVA ont été réalisée au niveau régional ; d'une part l'Ouest du continent puis d'autre part la région Guyane. Le test indique une forte structuration entre les régions Nord Ouest, Nord Est et Sud de Guyane (F_{CT}=0,83) par contre, la variance génétique entre populations de la même région n'est plus significative. En Amérique Centrale, la variance génétique entre le Costa-Rica et le Panama est négligeable (données non présentées), et le nombre de populations échantillonnées est trop faible pour mettre en évidence une structuration à l'intérieur même ce ces pays.

En prévision de la réalisation du test Tajima, il est intéressant de retenir que globalement, les populations à l'intérieur des régions Nord Ouest, Nord Est et Sud de Guyane ne sont pas

structurées génétiquement. De même, la dernière analyse effectuée en considérant, d'une part, le groupe de l'Ouest des Andes et, d'autre part, le groupe réunissant le Pérou et la population Est de l'Equateur indique une variance inter-populations à l'intérieur des groupes qui est significative, mais bien plus faible qu'en considérant les autres regroupements. C'est ce dernier schéma de regroupement qui est utilisé pour le test Tajima.

3.6.2. Test TAJIMA et Test du F_S de Fu

Tableau 6 : Valeurs du D de Tajima et du F_S de Fu

	Guyane			Ouest des Andes	Pe EstE	NO-NE
	NO	NE	S			
taille des échantillons	51	12	15	27	37	63
D de Tajima	0	7,77	0,17	-0,83**	0,90	0,38
F_S de Fu	0	3,24	0,55	3,18	0,94	1,50

Une significativité à 1.10^{-3} est représentée par ***. Une significativité à 1.10^{-2} est représentée par **. Une significativité à $5.\ 10^{-2}$ est représentée par *. NO=Nord Ouest, NE=Nord Est, S=Sud, B=Brésil, E=Equateur, Pa=Panama, Pip = Pipeline, San = Santa Rita, CR=Costa Rica FG=Guyane française.

Une valeur positive et significative de D est obtenue avec le test de Tajima dans le groupe contenant le Pérou et l'Equateur laissant supposer la présence d'une expansion de population. Néanmoins, le test de Fu ne confirme pas cette expansion.

Pour les autres groupes testés, ni la méthode de Tajima ni le Fs de Fu ne permettent de détecter des changements brusques de tailles de populations.

3.7. Test Analyse de la structuration géographique par comparaison des coefficients de différenciation.

Au niveau continental, N_{ST} est significativement supérieur à F_{ST} à 1%. Les haplotypes phylogénétiquement les plus proches se retrouvent plus souvent dans les mêmes populations. A l'échelle de la Guyane, la calcul n'est pas significatif.

4. Discussion

4.1. Intérêt du séquençage

Notre étude se caractérise par l'utilisation de la méthode du séquençage pour révéler le polymorphisme. C'est actuellement la technique la plus informative. Le séquençage de la zone d'intérêt permet de visualiser toutes les mutations y compris celles d'une base. L'ensemble de la séquence nucléotidique d'une région permet alors d'évaluer les distances nucléotidiques qui séparent les haplotypes et d'établir leurs relations phylogénétiques.

4.2. Diversité génétique et structuration

Notre étude indique qu'en Guyane, la diversité génétique intra-population au niveau du génome chloroplastique du *S.amara* est égale à 0,38. Cette diversité chloroplastique est d'un niveau plutôt élevé comparée à celle observée dans d'autres espèces tropicales. Chez le Cèdre rouge (*Cedrela odorata)*, et le Wacapou (*Vouacapoua americana)* les niveaux de diversité tous deux estimés à partir du polymorphisme de longueur de fragments dans les chloroplastes (méthode RFLP), sont bas, respectivement de 0,03 et 0,09 (Cavers *et al.*, 2003 ; Dutech *et al.*, 2001) alors que chez l'Angélique (*Dicorynia Guianensis*), la diversité génétique est proche de celle du *S. amara* (He= 0,41) (méthode RFLP) (Caron *et al.*, 2000).

Le *S. amara* montre la plus forte diversité au Brésil dans la population de Manaus (He=0,83) mais ce résultat nécessite d'être confirmé en raison du faible effectif d'individus échantillonnés (4). Il apparait par contre plus clairement que les populations d'Amérique Centrale (He=0,47 et He=0,54 pour le Costa Rica et le Panama respectivement) sont plus diversifiées que celles de la Guyane et de l'Equateur (He=0,38 et He=0,07 pour la Guyane et l'Equateur respectivement).

C'est aussi en Amérique Centrale que le niveau de diversité génétique dans le génome nucléaire est le plus élevé (He=0,730 au Costa Rica contre He=0,484 en Guyane) (Résultat du stage de master I de Stéphanie Barthe, 2008 ; basé sur le polymorphisme au niveau de quatre loci microsatellites).

La différenciation génétique estimée par les F statistiques en considérant différents niveaux hiérarchiques est significative, dans le *S. amara*, entre la partie Ouest et la partie Est des Andes, entre l'Amérique Centrale et l'Amérique du Sud, entre les pays, et entre les grandes régions de la Guyane. Au niveau nucléaire, la différenciation génétique estimée à partir de quatre loci microsatellites, est aussi retrouvée significative entre le groupe de l'Amérique centrale et celui de l'Amérique du Sud. On constate par contre, l'absence de différenciation génétique entre la plupart des populations guyanaises, à l'exception des paires Ouest-Centre et Ouest-Sud.

Dans le *S. amara* la structuration géographique par comparaison des coefficients de différenciation (F_{ST}) a montré une structuration significative uniquement à l'échelle continentale. La présence d'un signal phylogéographique au niveau continental a aussi été observé chez le *Jacaranda copaia* (Casalis, 2007) et dans le Carapa (Duminil *et al.*, 2006) .

Il est intéressant de voir maintenant si cette structuration peut révéler de changements de tailles significatifs au cours de l'histoire de ces populations.

4.3. Histoire démographique des populations de Guyane

Les tests de Tajima et de Fu effectués sur nos séquences de *S.amara* ne montrent aucun écart à l'équilibre mutation-dérive. Nous ne pouvons donc pas proposer de scénario historique basés sur des évènements démographiques (restriction ou expansion de population) à l'issue de ces tests. Il semble donc que les populations de *S. amara*, n'aient pas subit de fluctuations brusques de taille. La région du Nord Est de la Guyane présente pourtant une diversité beaucoup plus importante que dans le reste de la Guyane qui pourrait s'expliquer par de l'endémisme. Dans quelques études, cepandant, les populations des aires colonisées sont trouvée autant ou plus génétiquement variables que les refuges du fait de l'amalgame des immigrants de différentes zones refuges. (Muellner et al., 2005). L'hypothèse des refuges a été évoquée pour expliquer la structuration spatiale des populations de *Dicorynia guianensis* (Caron *et al.*, 2000) et de *Vouacapoua americana* (Dutech *et al.*, 2000) en Guyane. Selon cette hypothèse, comme nous l'avons expliqué dans l'introduction, les baisses de précipitations au cours des cycles glaciaires du quaternaire auraient entraîné un remplacement, dans de nombreux endroits, de la forêt par de la savane. La forêt se serait retrouvée restreinte à quelques zones de refuges. Durant cette période, la dérive génétique aurait fixé des allèles particuliers au sein des populations.

Le test de bottleneck réalisé pour tester l'équilibre mutation-dérive chez le *S.amara* avec quatre loci microsatellites montrent un écart à l'équilibre dans le sens d'un déficit en Hétérozygotes pour deux loci sur quatre. Cet écart à la neutralité pourrait être la signature d'une expansion récente des populations (Stage de S. Barthe, 2008), non détectable au niveau des séquences chloroplastiques. Chez le *J.copaia*, le test de Tajima réalisé au niveau des séquences chloroplastiques suggère que les populations de Guyane ont subi une phase d'expansion suite à un goulot d'étranglement (bottleneck) à un moment donné qui reste encore indéterminer (Stage de Maxime Casalis, 2007). Le fait que les populations de *Jacaranda copaia* en Guyane suivent une phase d'expansion suite à un goulot d'étranglement pourrait s'expliquer par l'hypothèse des refuges. Le schéma actuel observé pour le *S.amara* et

surtout pour le *J.copaia* correspondrait à une recolonisation, par les espèces, des espaces vides. Néanmoins les marques de goulots d'étranglement encore visibles aujourd'hui au niveau des microsatellites nucléaires laisse supposer que cet évènement est récent. Le taux de mutation plus élevé au niveau nucléaire associé à un fort brassage des gènes sont des éléments qui effacent rapidement la signature de fluctuations importantes de populations. Il est donc difficile de croire que le goulot d'étranglement subi par l'espèce date de l'époque de l'aire glaciaire, soit il y a 18 000 ans. Le schéma de réduction-expansion subit par l'espèce pourrait plutôt s'expliquer par des événements plus récents. L'activité humaine en foret amazonienne dans les siècles passés est encore peu connue. On pourrait imaginer que l'abandon de certaines zones cultivées aurait permis à des espèces pionnières comme le *S. amara* et le *Jacaranda copaia* de coloniser rapidement les espaces vides. L'hypothèse de brulis recolonisés est aussi évoquée (Servant et al., 1996).

4.4. Evènements historiques à l'échelle du continent

Le patron géographique de la diversité génétique est influencé par la structure des populations et leur histoire. D'après les inférences faites par la méthode NCPA (le tableau 4), l'histoire des populations a joué un rôle important dans l'établissement de la distribution actuelle des haplotypes obtenus avec les séquences neutres de chloroplastes chez le *Simarouba amara*. La méthode NCPA étant controversée, on choisi d'examiner si les inférences sont justifiées.

L'hypothèse nulle selon laquelle il n'y a pas de structuration des haplotypes en fonction de la géographie ne peut pas être rejetée

- si les séquences utilisées ne présentent pas assez de variations génétiques
- si l'échantillonnage n'est pas équilibré et continu sur la surface géographique car les calculs liés au tableau de contingences ne peuvent pas être effectués.
- et enfin, si les évènements d'expansion, de restrictions aux flux de gène et de fragmentations passées n'ont pas eu lieu car la répartition des populations approche une répartition aléatoire.

Dans notre étude, les régions de l'ADN chloroplastiques séquencées ont montré suffisamment de polymorphismes pour distinguer un total de 11 haplotypes dispersés parmi les six pays où au moins une population a été échantillonnée. La variabilité est donc suffisante. En revanche, l'irrégularité de notre échantillonnage parmi les différentes populations et sa couverture géographique discontinue nous demandent de prendre des précautions dans nos interprétations. En effet, plus le nombre d'échantillons obtenu dans une population est grand,

plus on a de chance d'y observer des allèles rares. Si les populations aux alentours sont moins échantillonnées mais contiennent aussi ces allèles rares, on ne les détecte pas. Les valeurs de Dc utilisées dans l'analyse NCPA pour ces haplotype rares seront sous estimées. De même, si certaines zones géographiques ne sont pas couvertes, les étendues des haplotypes (Dc) présents dans les zones non échantillonnées vont avoir tendance à être sous estimées. Un premier point semble montrer que les résultats obtenus ne sont pas trop affectés par cet échantillonnage irrégulier : nos populations les plus largement échantillonnées ne sont pas les populations qui présentent la plus grande richesse allélique. Par exemple, en Guyane, 4 haplotypes ont été observés pour 78 individus séquencés, alors qu'au Panama, autant d'haplotypes ont été trouvés pour seulement 25 individus séquencés. Le deuxième point qui nous permet d'affaiblir les doutes dus à l'échantillonnage concerne la méthode elle-même. Un grand avantage de la méthode NCPA réside dans les emboitements de clades. Ces regroupements d'haplotypes permettent de récupérer du pouvoir statistique pour les calculs concernant les clades de plus hauts niveaux (Templeton, 1995). Par exemple, notre échantillonnage réalisé au niveau du Pérou et de l'ouest de l'Equateur ne permet pas d'utiliser la clé d'inférence sur le clade 1-1. Néanmoins, au niveau supérieur, le clade 2-1 ne considère plus les étendues concernant les haplotypes, mais celles des clades 1-1 et 1-2 dans leurs ensembles et réitère les calculs en considérant les centres et étendues des nouveaux clades.
Ainsi, l'analyse des clades emboîtés a permis de révéler une forte structuration géographique dans quatre clades. Deux clades auraient été structurés à l'issue d'une restriction aux flux de gènes, un autre, à la suite d'une expansion démographique, et le dernier, correspondant au clade total, par une expansion et/ou une fragmentation passée.

4.4.1. Restriction aux flux de gènes en Amérique centrale (clade 1.5).

Le principe sur lequel se base cette prédiction est le suivant : en cas de restriction aux flux de gènes, les clades les plus anciens ont tendances à être géographiquement plus étendus que les plus récents, et les étendues des clades extérieurs sont généralement comprises dans les aires de répartitions des clades anciens. Dans le clade 1-5, l'haplotype 7 est celui qui est considéré comme étant l'haplotype le plus ancien. Il s'étend du Brésil jusqu'au Costa Rica. Les haplotypes présents en bout de chaine du réseau sont les haplotypes 8, 9 et 10 ; leurs étendus sont comprises dans l'aire de répartition de l'haplotype ancien (7) (Figure 9). La différence entre l'étendue de l'haplotype intérieur (7) et les moyennes des haplotypes de l'extérieur (8, 9, 10) est significativement positive. D'après le principe présenté ci-dessus, l'explication inférée à la distribution non aléatoire est une restriction aux flux de gènes.

Mais la présence de l'haplotype 7 dans la population de Manaus au Brésil nous interroge particulièrement et nécessitera une attention particulière dans la poursuite de cette étude. Un premier point important sera la validation de l'identité de la séquence en réamplifiant les deux individus suspects. Si la séquence est validée, il faudra alors intensifier l'échantillonnage pour savoir si cet haplotype est commun au Brésil ou s'il est seulement localisé à Manaus. Dans ce dernier cas une origine anthropique ne devra pas être exclue compte tenu de l'utilisation multiple du *S. amara* par l'homme. Même sans être présent au Brésil, l'haplotype 7 est étendu puisqu'on le retrouve en Equateur de l'Ouest, l'inférence n'est probablement pas erronée.

En tenant conte de la présence de l'haplotype 7 au Brésil, le cheminement exact suivi par le programme est noté sur le tableau 4 qui fait référence à la clé de Templeton présente en annexe 1. Le principe utilisé décrit plus haut est bien en accord avec les modélisations informatiques construites par ailleurs, mais sa validité sur des exemples réels reste discutée. Il semble néanmoins que dans la plupart des cas, le principe de l'analyse NCPA appliqué sur des espèces dont l'histoire est connue à travers des méthodes indépendantes, infère les bonnes conclusions (Templeton et al., 1995).

Cette observation de restriction aux flux de gènes nous semble donc vraisemblable et est particulièrement intéressante du fait qu'elle ne concorde pas avec la biologie de l'espèce : la répartition des graines du *S.amara* est effectuée sur de longues distances. Des facteurs historiques ou géographiques ont alors du influencer ce patron d'isolement en Amérique Centrale. Cette restriction au flux de gènes qui limite l'étendue des haplotypes extérieurs sur le réseau est localisée au Panama et au Costa Rica. Ceci pourrait être expliqué par

- L'isolement de l'Amérique Centrale par rapport à l'Amérique du Sud avant la formation de l'Isthme du Panama (avant 4,5 Ma)
- La formation des Andes il y a 14 millions d'années (14 Ma)
- La réduction des aires forestières pendant le dernier maximum glaciaire (14 000 ans)
- L'étroitesse même du passage au niveau du Panama

Une analyse exploratoire basée sur les probabilités bayésiennes et qui intègre un groupe externe (*Simaba cedron*, famille des Simaroubaceae) à permis de d'estimer le temps de divergence entre les lignées de l'Est et de l'Ouest des Andes aux alentours de 15 000 ans (données non présentées). Cette divergence, issue de la restriction aux flux de gènes, serait donc postérieure à la formation des Andes.

Un scénario possible pour expliquer la structuration géographique à l'Ouest des Andes serait que l'haplotype 7 localisé dans la partie Ouest des Andes se soit étendu de l'Equateur (peut être du Pérou), jusqu'au Costa Rica (et peut être plus haut, vers le Nicaragua).

Puis, suite à une barrière aux flux de gènes au sein même de la partie Ouest de la cordillère, les nouveaux haplotypes dérivés de cet haplotype 7 seraient restés confinés en Amérique Centrale. Les haplotypes dérivés de l'haplotype 7 ne divergent tous que d'une ou deux mutations ce qui révèle une diversification du *S.amara* plutôt récente en Amérique Centrale.
Une alternative à cette hypothèse serait d'imaginer la présence d'une ou plusieurs zones refuges en Amérique Centrale à partir desquelles les populations ont divergé ; Les zones refuges contenant l'haplotype 7 et les différents dérivés auraient ensuite participées à la recolonisation de l'Ouest des Andes. Un échantillonnage précis dans de plus nombreuses populations de l'Ouest des Andes, y compris l'Amérique Centrale aurait été nécessaire pour connaître les délimitations d'éventuelles zones monophylétiques.

Les explications précédentes considèrent la diversification récente à partir de l'haplotype 7. Une remarque est particulièrement frappante concernant la divergence entre la lignée issue de l'haplotype 7 et l'haplotype 7 lui-même, et le reste. La divergence d'environ 15 000 ans trouvée est postérieure à la présence de la mer au niveau du Panama et à la formation des Andes. Cela signifie que la présence d'une barrière au flux de gène est aussi très récente. Depuis sa formation il y a des millions d'années, la cordillère présentant incontestablement une barrière au flux de gène pour une espèce qui se trouverait de part et d'autre. Il semblerait que le *S.amara* soit resté du même côté de la cordillère jusqu'à il y à 15000 ans environ.

L'étude de Cavers *et al.* qui s'étend jusqu'au Mexique a révélé que la richesse retrouvée en Amérique Centrale est plutôt due à des évènements multiples de colonisation pré et post fermeture de l'isthme du Panama, qu'à la présence d'une zone refuge (Cavers et al. 2003). Si nos *S. amara* étaient du côté du bassin amazonien avant ces 15000 ans, la colonisation de l'Amérique Centrale a été bien plus récente que pour C odorata. On note que *C. odorata* est une espèce pionière, tout comme *le S.amara*, cependant, le *C. odorata* est une espèce dispersée par le vent (Cavers et al. 2003), tandis que le *S.amara* est dispersé par des oiseaux et des mammifères (Hardesty et al., 2005). Ceci pourrait expliquer en partie les histoires évolutives divergentes entre ces deux espèces : les animaux dispersant le *S.amara* seraient restés sur le même continent Sud américain avant la fermeture de l'Isthme (les espèces concernées sont des espèces forestières strictes et n'auraient pas traversé la mer). On se rend compte, au travers de cet exemple, de l'importance du mode de dispersion dans l'histoire de colonisation d'une espèce. On note qu'on ne sait pas de quel côté de la cordillère l'espèce était présente en premier. On pense qu'elle aurait pu traverser cette barrière à plusieurs époques bien antérieure à -15000 ans, et que les différentes lignées présentes de part et

d'autres auraient alors connu le phénomène de spéciation (la divergence entre le groupe extérieur, *Simaba cedron* et le *Simarouba amara* est évaluée à seulement 4 Ma (données non présentées) et ils sont pourtant de genre différents).

4.4.2. Restriction aux flux de gènes dans le clade 1-2 (Guyano-Brésilien)

Un scénario de restriction aux flux de gènes dans le clade 1-2 parait peu probable en raison de la grande étendue géographique de l'haplotype 1 dérivé (présent du brésil jusqu'au Nord de la Guyane) par rapport à son haplotype 2 ancestral restreint dans les Monts Tumuc Humac. Suivant le principe utilisé dans l'analyse NCPA, ce patron où l'haplotype dérivé est le plus étendu devrait signifier une expansion. Il apparait donc une incohérence entre la valeur I-T calculée (-113,5) et l'inférence proposée négative qui est prise en compte par le logiciel comme si elle était significativement grande. Ceci est probablement une erreur du programme.

De plus, il est très probable que l'aire de l'haplotype 1 s'étende jusqu'au Venezuela : les séquences obtenues au niveau de 12 sites polymorphiques sur 13 de l'individu du Venezuela (herbier) sont identiques à ceux de l'haplotype 1. Une valeur I-T encore plus petite correspondrait alors mieux à la situation réelle. Un échantillonnage plus large au sud de la Guyane et dans les états du Brésil adjacents à la Guyane (Amapa, Para et Roraima), ainsi que dans les pays voisins plus au Nord (Suriname, Guyana) est nécessaire pour confirmer la faible étendue géographique de l'haplotype 2.

4.4.3. Expansion démographique dans le clade 2-1

Le principe utilisé pour détecter les expansions de population se base, comme pour la détection de restriction aux flux de gènes, sur l'écart entre l'étendue géographique observée des clades internes et des clades externes du réseau. Une expansion de population entraine une étendue limitée chez les haplotypes internes (anciens) et une étendue bien plus grande pour les haplotypes externes (récents). Aussi les haplotypes récents qui dérivent de l'haplotype ancien peuvent être distants. C'est le cas dans le clade 2-1 pour lequel l'expansion aurait eu lieu de la partie orientale de l'Amazone (Guyane + Brésil) (clade 1-2) vers sa partie occidentale à l'est (Equateur ; Pérou) (clade 1-1). Cependant, la valeur faible trouvée pour l'étendue du clade 1-2 pourrait être liée à l'absence d'échantillonnage dans le reste du Brésil. Et peu d'études aussi bien théoriques (modélisations) qu'empiriques ont permis de vérifier la fiabilité de ce principe d'expansion (Templeton et al., 1995).

Aussi, la description des séquences contenues dans les clades 1-2 et 1-3 propres à la Guyane nous pousse à critiquer le résultat obtenu par cette analyse NCPA :

En Guyane française, les quatre haplotypes observés diffèrent tous par une, deux, ou trois mutations de type insertion/délétion, au niveau d'une même région répétée en Thymines. En progressant du Nord-Est jusqu'au Centre de la Guyane (Saül), le nombre de Thymines augmente dans la séquence. Le type correspondant à trois délétions « polyT + --- » (haplotype 4) est seulement observé au Nord-Est dans la zone de Bélizon. Le type à deux délétions « polyT + T-- » (haplotype 3) est aussi observé au Nord-Est dans la population de Bélizon mais s'étend vers le centre de la Guyane (Saul) et jusqu'au Sud puisqu'elle est retrouvée au Brésil. Le type qui à une seule délétion « polyT + TT- » (haplotype 2) est par contre très localisée dans l'extrême sud de la Guyane (Monts Tumuc Humac). En progressant du Nord-Est jusqu'au Sud-Est, la région répétée en Thymine gagne des Thymines. Une expansion en cette même direction est en effet possible, et du fait de la grande variabilité des régions répétées, on suppose que cet évènement est très récent. Il est intéressant de préciser que les haplotypes 5 de l'Ouest de l'Equateur et 11 du Pérou ont tous les deux, au niveau de ce poly-T, la forme de l'haplotype 2 localisé au Sud de la Guyane, et présentent aussi d'autres mutations dont une substitution. Le Pérou et l'Equateur feraient parti du mouvement de migration Amazonie orientale/Amazonie occidendale (Sud-Ouest/Nord-Est) à l'échelle de l'Amérique du Sud. La direction du mouvement qu'on décrit là est la même que celle qui est inférée par l'analyse NCPA, seulement, nos observations nous laissent penser que la migration se serait produite dans le sens contraire. On peut vraisemblablement suspecter que les mutations accumulées au niveau de la région du poly-T soient récentes en raison du taux de mutations plus élevé dans les régions microsatellites. L'expansion récente du *S. amara* en Guyane pourrait correspondre à l'expansion observée chez le *Jacaranda copaia.* (Maxime Casalis, 2007).

4.4.4. Expansion et/ou fragmentation du clade total

La dernière inférence proposée par l'analyse NCPA est formulée à partir de la valeur Dc du clade intérieur (clade 2-3) significativement faible. Cette valeur n'est pas représentative puisque la Colombie, le Venezuela, le Guyana et le Suriname n'ont pas été échantillonnés. Il serait très intéressant de savoir si les mutations intermédiaires représentées dans le réseau d'haplotype de la figure 7 peuvent se retrouver dans ces pays. Cela pourrait être un indice concernant d'une part les différents mouvements d'expansion et d'autre part, concernant la diversification des deux lignées.

5. Conclusion

Le polymorphisme retrouvé dans les séquences chloroplastiques intergéniques nous a permis de définir 11 haplotypes. Les haplotypes présents en Amérique Centrale forment une lignée bien éloignée du reste et est caractérisée par des richesses alléliques élevées.

L'analyse du réseau de clade emboité aboutit en deux principales inférences. L'une propose une restriction de flux de gène en Amérique Centrale, et l'autre une expansion à partir de la Guyane jusque vers l'ouest du bassin amazonien. La méthode NCPA, analyse à deux dimensions (évolution génétique, et répartition spatiale) est une méthode dont il faut se méfier si l'échantillonnage ne recouvre pas entièrement et équitablement toute la surface étudiée. D'autres analyses sont donc effectuées.

Les diverses analyses de nos séquences montrent qu'il y a bien une forte structuration entre l'Ouest et l'Est de la cordillère des Andes (F_{ST} et N_{ST}), et au sein même des populations de Guyane (F_{ST}), mais les inférences NCPA ne sont pas confirmées par les tests de fluctuations démographiques (Tajima, Fu). On retient quand même l'idée d'une expansion récente en Guyane du fait des caractéristiques observées au niveau de nos séquences et de leur répartition. Plusieurs autres éléments mettent en évidence l'influence de perturbations récentes dans l'histoire de la forêt amazonienne. Dans le *J. copaia,* tout d'abord où il a été mis en évidence une expansion récente de la population en Guyane, postérieur à la dernière glaciation (Stage M2 de Maxime Casalis). L'étude de microsatellite nucléaire sur le *Simarouba amara* (Stage M1 de Stéphanie Barthe) suggère aussi une expansion récente

La lignée d'Amérique Centrale, serait isolée par la cordillère des Andes, mais au sein même de la population à l'Ouest de la cordillère, une structuration expliquée par d'anciennes zones refuges n'est pas écartée.

Bien que notre étude représente une des rares études réalisées à l'heure actuelle sur une aussi grande échelle en Amérique Centrale et en Amérique du Sud, un meilleur échantillonnage pourrait préciser les passages empruntés lors des phénomènes d'expansion.

Des méthodes fiables et précises permettant de comparer des données de séquences et leurs répartitions géographiques sur plusieurs espèces ne sont pas encore disponibles. Ces méthodes de phylogéographies comparées pourraient intégrer des données écologiques concernant la biologie des espèces (temps de génération, distance de dissémination des graines, types d'habitats…), mais aussi des données sur les transformations géologiques du passé (Evolution des reliefs), et les modifications hydrographiques (Evolution des niveaux des fleuves et des mers).

Bibliographie

Avise JC, (1989). Gene trees and organismal histories : a phylogenetic approach to population biology. Evolution 43: 1192-1208

Boggan J, Funk V, Kelloff C, Hoff M, Cremers G, Feuillet C, (1997). Cheklist of the plants of Guianas (Guyana, Surinam, French Guiana). Smithsonian's Biological Diversity of the Guianas Program publication series : Washington

Birky CW (1995). Uniparental inheritance of mitochondrial and chloroplast genes : mechanisms and evolution. Proc Natl Acad Sci USA 92 : 11331-11338

Burnham RJ, Graham A, (1999) The history of neotropical vegetation: New developments and status Annals of the Missouri. Botanical Garden 86 : 546-589

Bush MB, (1994). Amazonian speciation: a necessarily complex model. Journal of Biogeography 21:5-17

Bush, MB and Oliveira PE, (2006). The rise and fall of the Refugial Hypothesis of Amazonian Speciation: a paleoecological perspective. Biota Neotrop. V6(1)

Caron H, Dumas S, Marque G, messier C, Bandou E, Petit RJ, Kremer A, (2000) Spatial and temperal distribution of chloroplast DNA polymorphism in a tropical tree species. Molecular Ecology 9 : 1089-1098

Cavers S, Navarro C, Lowe AJ, (2003) Chloroplast DNA phylogeography reveals colonization of a neotropical tree, Cedrela odorata L., in Mesoamerica. Molecular Ecology 12 : 1451-1460

Clayton JW, Fernando ES, Soltis PS, et Soltis DE, (2007). Molecular Phylogeny of the Tree-of-Heaven Family (Simaroubaceae) based on chloroplast and nuclear markers. Int. J. Plant Sci. 168(9):1325–1339

Clement M, Posada D, et Crandall KA (2000). TCS: A computer program to estimate gene genealogies. Molecular Ecology, 9(10):1657-1659

Colinvaux PA, De Oliveira PE, Bush MB, (2000) Amazonian and neotropical plant communities on glacial time-scale : The failure of the aridity and refuge hypothesis. Quaternary Science Reviews 19 : 141-169

Comps B, Gömöry D, Letouzey J, Thiébaud B, Petit RJ, (2001). Diverging Trends Between Heterozygosity and Allelic Richness During Postglacial Colonisation in the European Beech. Genetics 157 :389-397

Crane PR, Friis EM, Pedersen KR, (1995). The origin and early diversification of angiosperms. Nature 374. 27

Da silva MNF et Platton JL, (1998). Molecular phylogeography and the evolution and conservation of Amazonian mammals. Molecular ecology 7 : 475-486

Dick CW, Abdul-Salim K, Bermingham E, (2003). Molecular Systematic Analysis Reveals Cryptic Tertiary Diversification of a Widespread Tropical Rain Forest Tree. the american naturalist 162.6

Dick CW, Roubik DWn Gruber KF, Bermingham E, (2004). Long-distance gene flow and cross-Andean dispersal of lowland rainforest bees (Apidae: Euglossini) revealed by comparative mitochondrial DNA phylogeography. Molecular Ecology 13 : 3775–3785

Doyle JJ et Doyle JL, 1990. Isolation of plant DNA from fresh tissue. Focus 12 : 13-15

Duminil J, Caron H, Scotti I, Cazal SO, Petit RJ, (2006). Blind population genetics survey of tropical rainforest trees. Molecular Ecology : 1-9

Dutech C., Maggia L., Joly H.I. (2000) Chloroplast diversity in Vouacapoua americana (Caesalpinaceae), a neotropical forest tree. Molecular Ecology 9 : 1427–1432.

Excoffier L, Smouse PE et Quattro JM, (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes - application to human mitochondrial - DNA restriction data. Genetics 131:479-49

Fischer AG, (1960). Latitudinal Variations in Organic Diversity. Evolution 14:64-81

Flores O, Gourlet-Fleury S, Picard N (2006). Local disturbance, forest structure and dispersal effects on sapling distribution of light-demanding and shade-tolerant species in a French Guianian forest. Acta Oecologica 29:141-154

Fu YX (1997) Statistical tests of neutrality of mutations against population growth,hitchhiking and background selection. Genetics 147:915-925

Haffer J, (1969). Speciation in Amazonian Forest Birds. Science 165, 131-137

Hall JPW et HarveyJH, (2002). The phylogeography of amazonia revisited: new evidence from riodinid butterflies. Evolution 56(7):1489-1489

Hammond DS, (2005). Tropical forests of the guiana shield Ancient forest in a modern world. In: Hammond DS. Biophysical features of the Guiana Shield. Ed Hammond DS Pp15-194

Hardesty BD, Dick CW, Kremer A, Hubbell S et Bermingham E, (2005). Spatial genetic structure of Simarouba amara Aubl.(Simaroubaceae), a dioecious, nimal-dispersed Neotropical tree, on Barro Colorado Island, Panama. Heredity:1–8

Hardy OJ, Vekemans X. (2002).SPAGeDi : a versatile computer program to analyse spatial genetic structure at the individual or population levels. Molecular Ecology Notes 2 : 618-620

Hartl DL et Clark GA, (1989). Principle of population genetics. Second edition. Sinauer Associates, Inc., Sunderland, Mass., 01375 USA

Hewitt G, (2000). The genetic legacy of the Quaternary ice ages. Nature 405(20)

Hudson RR, Boos DD et Kaplw NL, (1992) A statistical test for detecting geographic subdivision. Mol. Biol. Evol. 9: 138-151

Latreille C - Office National des Forêts (2004). Guide de reconnaissance des arbres de Guyane - 120 essences décrites - ONF Guyane (FR) . 374 p. 74-75

Mardulyn P, (2001). Phylogeography of the Vosges mountains populations of Gonioctena pallida (Coleoptera: Chrysomelidae): a nested clade analysis of mitochondrial DNA haplotypes. Molecular Ecology 10, 1751-1763

McCauley DE (1995). The use of chloroplast DNA polymorphism in studies of gene flow in plants. Tree, vol. 10, n° 5 : 198-202

McKenna DD et Farell BD, (2006). The tropical forests are both evolutionary cradles and museums of leaf beetle diversity. PNAS 103(29):10947-10951

Moraes-Barros N, Silva JAB, Miyaki CY, Morgante JS, (2006). Comparative phylogeography of the Atlantic forest endemic sloth (Bradypus torquatus) and the widespread three-toed sloth (Bradypus variegatus) (Bradypodidae, Xenarthra). Genetica 126 :189-198

Muellner AN, Tremetsberger K, Stuessy T et Baeza CM, (2005) Pleistocene refugia and recolonization routes in the southernAndes: insights from Hypochaeris palustris (Asteraceae,Lactuceae). Molecular Ecology 14:203–212

Noonan BP et Gaucher P, (2005). Phylogeography and demography of guianan harlequin toads (Atelopus) : diversification within refuge. Molecular Ecology 14:3017-3031

Panchal M, (2007). The automation of Nested Clade Phylogeographic Analysis. Bioinformatics, 23:509-510

Petit J, (2007). The coup de grâce for the nested clade phylogeographic analysis? Molecular Ecology

Petit RJ, Brewer S, Bordács S, Burg K, Cheddadi R, Coart E, Cottrell J, Csaikl UM, Van DBC, Deans JD, Fineschi S, Finkeldey R, Glaz I, Goicoechea PG, Jensen JS, König AO, Lowe AJ, Madsen SF, Mátyás G, Munro RC, Popescu F, Slade D, Tabbener HdeVSMG, Ziegenhagen BBJL, Kremer, A, et Espinel S (2002) Identification of refugia and postglacial colonization routes of European white oaks based on chloroplast DNA and fossil pollen evidence. Forest Ecology Managenent 156 : 49-74

Pons O. et Petit R.J. (1996). Measuring and testing genetic differenciation with ordered versus unordered alleles. Genetics 144 : 1237-1245

Posada D, Crandall KA, Templeton AR, (2000). GeoDis: A program for the cladistic nested analysis of the geographical distribution of genetic haplotypes. Molecular Ecology, 9(4):487-488

Ricklefs RE, (1987). Community diversity: Relative Roles of Local and Regional Processes. Science 235, 167-171

Serror P. (1988). Etude de la variabilité de l'ADN chloroplastique dans le genre Helianthus : recherché de marqueurs de cytoplasmes. Université Blaise Pascal Clermont-Ferrand II, 149 p

Servant M, Servant-Vildary S (1996). Dynamique à long terme des écosystèmes forestiers intertropicaux. Editeurs scientifiques : Paris

Shaw J, Lickey EB, Beck JT, Farmer, SB, Wusheng L, Miller J, Kunsiri CS, Winder CT, Shilling EE, Small RL, (2005). The Tortoise and the Hare II : Relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. American Journal of Botany 92(1) : 142-166

Simpson EH, (1949). Measurement of diversity. Nature 163. 688.

Tajima F. (1989). Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics 123 : 585-595

Templeton AR, Routman E, et Phillips CA, (1995). Separating population structure from population history : a cladistic analysis of the geographical distribution of mitochondrial dna haplotypes in the tiger salamander,ambystoma tigrinum. Genetics, 140(2):767-782

Templeton, (1998). Nested clade analyses of phylogeographic data : testing hypotheses about gene flow and population history. Molecular ecology 7 : 381-397

Valencia R, Foster RB, Villa G, Condit R, Svenning JC, Hernàndez C, Romoleroux K, Losos E, Magard E, Balslev H, (2004). Tree species distributions and local habitat variation in the Amazon : large forest plot in weastern Ecuador. Journal of Ecology 92 : 214-229

Widmer A, Lexer C, (2001). Glacial refugia: sanctuaries for allelic richness, but not for gene diversity. Trends Ecol Evol 16 (6): 267-269

Wolfe K, Li WH, Sharp P (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear dans. Proc. Natl. Acad. Sci. USA 84 : 9054-9058

Wright S, Keeling J, Gillman L, (2006). The road from Santa Rosalia: A faster tempo of evolution in tropical climates

Remerciements

Je tiens tout d'abord à remercier sincèrement Caroline Scotti pour sa disponibilité, sa sympathie et pour m'avoir aidé quand cela était nécessaire, même le dimanche. Merci aussi pour ses connaissances et sa passion pour l'histoire des populations qu'elle a su me transmettre.

Merci à Ivan Scotti pour être toujours prêt à répondre aux questions.

Un merci à
Valérie Troispoux et Eliane Louisanna pour leur aide au laboratoire.
Saint Omer Cazal pour l'échantillonnage et les extractions.

Merci à tous les collègues de bureaux (Delphine, Sandra, Malia et Emilien) pour leur aide en génétique… ou en ornithologie.

Et enfin,
Merci aussi aux Parcouris (Mauricio, Pauline, Benoit et William) d'être ce qu'ils sont.

Annexe 1

Clé d'inférence de l'analyse de clades emboités (Templeton et al., 2005)

11 November 2005

1

Inference Key for the Nested Haplotype Tree Analysis of Geographical Distances

Start with haplotypes nested within a 1-step clade and work up to clades nested within the total tree. If the tree is not rooted through an outgroup or if none of the clades nested at the total tree level have the sum of the outgroup probabilities of their haplotypes greater than or equal to 0.95, regard all clades nested at the total tree level as tips. When rooting is deemed reliable, interiors should also refer to the older clades in a nesting category, and tips to their evolutionary descendants.

This key is applied only if there are some significant values for D_c, D_n, or I-T within the nesting clade. If there are no statistically significant distances within the clade, the null hypothesis of no geographical association of haplotypes cannot be rejected (either panmixia in sexual populations, extensive dispersal in non-sexual populations, small sample size, or inadequate geographical sampling). In that case, move on to another clade at the same or higher level.

1. Are all clades within the nesting clade found in separate areas with no overlap?
- NO – Go to step 2.
- YES - Go to step 19.

2. Is at least one of the following conditions satisfied?
   a. The D_c's for one or more tips are significantly small and the D_c's for one or more of the interiors are significantly large or non-significant.
   b. The D_c's for one or more tips are significantly small or non-significant and the D_c's for some but *not* all of the interiors are significantly small.
   c. The D_c's for one or more interiors are significantly large and the D_c's for the tips are either significantly small or non-significant
   d. The I-T D_c is significantly large.
- NO - Go to step 11.
- YES - Go to step 3.
- Tip/Interior Status Cannot be Determined - **Inconclusive Outcome.**

3. Is at least one of the following conditions satisfied?
   a. Some D_n and/or I-T D_n values are significantly reversed from the D_c values.
   b. One or more tip clades show significantly large D_n's.
   c. One or more interior clades show significantly small D_n's.
   d. I-T has a significantly small D_n with the corresponding D_c value non-significant.
- NO - Go to step 4.
- YES - Go to step 5.

4. Are both of the following conditions satisfied?
   a. The clades (or 2 or more subsets of them) with significantly small D_c or D_n values have ranges that are completely or mostly non-overlapping with the other clades in the nested group (particularly interiors).

   b. The pattern of completely or mostly non-overlapping ranges in the above condition represents a break or reversal from lower level trends within the nested clade series (applicable to higher-level clades only).
- NO - **Restricted Gene Flow with Isolation by Distance (Restricted Dispersal by Distance in Non-sexual species).** This inference is strengthened if the clades with restricted distributions are found in diverse locations, if the union of their ranges roughly corresponds to the range of one or more clades (usually interiors) within the same nested group (applicable only to nesting clades with many clade members or to the highest level clades regardless of number), and if the D_c values increase and become more geographically widespread with increasing clade level within a nested series (applicable to lower level clades only).
- YES - Go to step 9.

5. Are both of the following conditions satisfied?
   a. The clades (or 2 or more subsets of them) with significantly small D_c values have ranges that are completely or mostly non-overlapping with the other clades in the nested group (particularly interiors).
   b. The pattern of completely or mostly non-overlapping ranges in the above condition represents a break or reversal from lower level trends within the nested clade series (applicable to higher-level clades only).

- NO - Go to step 6.
- YES - Go to step 15.

6. Do clades (or haplotypes within them) with significant reversals or significant D_n values without significant D_c values define two or more geographically concordant subsets.

- No - Go to step 7.
- YES - Go to step 13.
- TOO FEW CLADES ($\leq$ 2) TO DETERMINE CONCORDANCE - **Insufficient Genetic Resolution to Discriminate between Range Expansion/Colonization and Restricted Dispersal/Gene Flow** - Proceed to step 7 to determine if the geographical sampling is sufficient to discriminate between short versus long distance movement.

7. Are the clades with significantly large D_n's (or tip clades in general when D_n for I-T is significantly small) separated from the other clades by intermediate geographical areas that were sampled?

- NO - Go to step 8.
- YES - **Restricted Gene Flow/Dispersal but with some Long Distance Dispersal.**

8. Is the species absent in the non-sampled areas?

- NO - **Sampling Design Inadequate to Discriminate between Isolation by Distance (Short Distance Movements) versus Long Distance Dispersal**
- YES - **Restricted Gene Flow/Dispersal but with some Long Distance Dispersal over Intermediate Areas not Occupied by the Species; or Past Gene Flow Followed by Extinction of Intermediate Populations**.

9. Are the different geographical clade ranges identified in step 4 separated by areas that have not been sampled?

- NO - **Allopatric Fragmentation.** (If inferred at a high clade level, additional confirmation occurs if the clades displaying restricted by at least partially non-overlapping distributions are mutationally connected to one another by a larger than average number of steps.)
- YES - Go to step 10.

10. Is the species absent in the non-sampled areas?

- NO - **Geographical Sampling Scheme Inadequate to Discriminate Between Fragmentation and Isolation By Distance.**
- YES - **Allopatric Fragmentation.** (If inferred at a high clade level, additional confirmation occurs if the clades displaying restricted by at least partially non-overlapping distributions are mutationally connected to one another by a larger than average number of steps.)

11. Is at least one of the following conditions satisfied?
   a. The D_c value(s) for some tip clade(s) is/are significantly large.
   b. The D_c value(s) for all interior(s) is/are significantly small.
   c. The I-T D_c is significantly small.

- NO - Go to step 17
- YES - **Range Expansion**, go to step 12.

12. Are the D_n and/or I-T D_n values significantly reversed from the D_c values?

- NO - **Contiguous Range Expansion.**
- YES - Go to step 13.

13. Are the clades with significantly large D_n's (or tip clades in general when D_n for I-T is significantly small) separated from the geographical center of the other clades by intermediate geographical areas that were sampled?

- NO - Go to step 14.
- YES - **Long Distance Colonization Possibly Coupled with Subsequent Fragmentation** (subsequent fragmentation is indicated if the clades displaying restricted but at least partially non-overlapping distributions are mutationally connected to one another by a larger than average number of steps) **or Past Fragmentation Followed by Range Expansion**. To see if secondary contact is involved, perform the supplementary tests given in Templeton, Molecular Ecology **10**: 779-791, 2001. To discriminate the type of movement leading to this pattern, go to step 21.

14. Is the species present in the intermediate geographical areas that were not sampled?

- YES - **Sampling Design Inadequate to Discriminate between Contiguous Range Expansion, Long Distance Colonization, and Past Fragmentation.**
- NO - **Long Distance Colonization and/or Past Fragmentation** (not necessarily mutually exclusive). If inferred at a high clade level, fragmentation rather than colonization is inferred if the clades displaying restricted but at least partially non-overlapping distributions are mutationally connected to one another by a larger than average number of steps. If the branch lengths are short, a colonization event is inferred, perhaps associated with recent fragmentation. To discriminate the type of movement leading to this pattern, go to step 21.

15. Are the different geographical clade ranges identified in step 5 separated by areas that have not been sampled?

- NO - **Past Fragmentation and/or Long Distance Colonization** (not necessarily mutually exclusive). If inferred at a high clade level, fragmentation rather than colonization is inferred if the clades displaying restricted but at least partially non-overlapping distributions are mutationally connected to one another by a larger than average number of steps. If the branch lengths are short, a colonization event is inferred, perhaps associated with recent fragmentation. To discriminate the type of movement leading to this pattern, go to step 21.
- YES - Go to step 16.

16. Is the species present in the intermediate geographical areas that were not sampled?

- YES - Go to step 18.
- NO - **Allopatric Fragmentation.** If inferred at a high clade level, additional confirmation occurs if the clades displaying restricted by at least partially non-overlapping distributions are mutationally connected to one another by a larger than average number of steps.

17. Are either of the following conditions satisfied?

a. The D_n values for tip or some (but not all) interior clades are significantly small.
b. The D_n for one or more interior clades is/are significantly large.
c. The I-T D_n value is significantly large.

- NO - **Inconclusive Outcome.**
- YES - Go to step 4.

18. Are the clades found in the different geographical locations separated by a branch length with a larger than average number of mutational steps.

- NO - **Geographical Sampling Scheme Inadequate to Discriminate Between Fragmentation, Range Expansion, and Isolation By Distance.**
- YES - **Geographical Sampling Scheme Inadequate to Discriminate Between Fragmentation and Isolation By Distance.**

19. Is the species present in the areas between the separated clades?

- NO – **Allopatric Fragmentation.** If inferred at a high clade level, additional confirmation occurs if the clades displaying restricted by at least partially non-overlapping distributions are mutationally connected to one another by a larger than average number of steps.
- YES - Go to step 20.

20. Was the species sampled in the areas between the separated clades?

- NO – **Inadequate Geographical Sampling.**
- YES – Go to step 2.

21. Are all of the following true?

a. Is it biologically realistic that the organism could have undergone long-distance movement?
b. Are the nested haplotypes that mark a potential long-distance colonization event within a clade that shows evidence of population growth by other methods (such as mismatch distributions)?
c. At the level of the entire cladogram, does the clade *not* inferred to have produced long-distance colonization *not* show evidence of past population growth with other methods?

- YES – **Long-distance movement.**
- NO – Insufficient evidence to discriminate between long-distance movements of the organism and the combined effects of gradual movement during a past range expansion and fragmentation. If the case against long-distance movement is compelling, then the inference is **past gradual range expansion followed by fragmentation.**