Université de Montréal

Identification des bases évolutives de la diversité génétique des populations de naseux des rapides (*Rhinichthys cataractae*) dans l'est du Canada

par

Philippe Girard

Département des Sciences Biologiques
Faculté des Arts et des Sciences

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de Ph.D.
en Sciences Biologiques

Août 2007

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée:

Identification des bases évolutives de la diversité génétique des populations de naseux des rapides (*Rhinichthys cataractae*) dans l'est du Canada

présentée par:
Philippe Girard

a été évaluée par un jury composé des personnes suivantes:

Daniel Boisclair, président-rapporteur
Bernard Angers, directeur de recherche
Pierre Legendre, membre du jury
Louis Bernatchez, examinateur externe
Sabin Lessard, représentant du doyen de la FES

Résumé

L'identification des mécanismes évolutifs qui établissent et maintiennent la diversité des gènes fonctionnels à l'intérieur et entre les populations représente une question importante en biologie de la conservation. Cependant, décrire les processus qui agissent sur un seul gène est un défi de taille. La diversité génétique résulte d'une multitude de processus neutres ou non agissant à différentes échelles temporelles. Les effets à long et court terme de la sélection naturelle sont en effet entremêlés avec ceux de la dérive, de la migration, des mutations et de la fragmentation allopatrique. La compréhension du portrait évolutif complet d'un gène donné nécessite donc la décortication de tous ces effets. L'objectif général de cette thèse était de quantifier les impacts de ces processus à l'intérieur et entre les populations d'une espèce de Cyprinidae, le naseux des rapides (*Rhinichthys cataractae*) de manière à mettre en lumière les pressions de sélection locales agissant sur des gènes spécifiques: le complexe majeur d'histocompatibilité (CMH), l'hormone de croissance (GH) et la trypsine. Les populations de naseux ont été échantillonnées sur un large territoire de la péninsule du Québec. L'interprétation de la diversité génétique de ces populations a nécessité i) l'identification de leurs relations historiques et ii) la détermination de leurs processus démographiques inhérents. Des marqueurs mitochondriaux et des microsatellites ont été développés dans ce but. La diversité des gènes fonctionnels a ensuite été analysée conjointement à celle observée sur ces marqueurs neutres, permettant d'identifier les effets de la sélection naturelle et d'inférer les mécanismes biologiques qui en sont responsables. Cette étude a démontré l'importance des mécanismes neutres dans la détermination de la diversité des gènes chez les populations naturelles.

Mots-clés: CMH, colonisation postglaciaire, dérive, hormone de croissance, microsatellite, mitochondrie, POST, sélection naturelle, structure génétique, trypsine.

Abstract

The identification of the evolutionary mechanisms that established and maintained the diversity of functional genes within and among populations is an important issue in conservation biology. However, describing the processes that act on a specific gene is not an easy task. The genetic polymorphism is the result of processes (neutral and/or non-neutral) acting on different evolutionary time scales. The short and long term effects of natural selection are entangled with effects of drift, gene flow, mutation and allopatric fragmentation. The comprehension of the complete evolutionary portrait of a given gene requires the disentanglement of these numerous processes. The general objective of this project was to quantify the impacts of these processes within and among populations of a common Cyprinidae species, the Longnose dace (*Rhinichthys cataractae*), to highlight the local influence of selection acting on specific genes: the major histocompatibility complex (MHC IIβ), the growth hormone (GH) and the trypsin. Populations were sampled over a large territory of Quebec peninsula. The interpretation of the genetic diversity of these populations has required i) the identification of the historical relationships between them and ii) the determination of the demographic history of each population. Mitochondrial and microsatellite markers were developed in this order. The diversity of functional genes was then jointly analysed with the neutral diversity observed previously, allowing the quantification of the effects of natural selection on populations differentiation. The conclusions of this study have demonstrated the importance of neutral mechanisms into the determination of the genetic diversity of genes within and among populations.

Keywords: drift, GH, MHC, microsatellites, mitochondrial markers, POST, postglacial colonisation, selection, structure, tripsin.

Table des matières

Liste des Tableaux

Liste des Figures

Liste des Sigles et Abréviations

ACC (CCA): Analyse canonique des correspondances

ADN (DNA): Acide désoxyribonucléique

AMOVA: Analyse de variance moléculaire

BP: *Before present*

bp: paires de bases

CCorA: Analyse des corrélations canoniques

CI: Interval de confiance

CMH (MHC): Complèxe majeur d'hitocompatibilité

D: Statistique de Tajima

Dc: Diamètre de la distribution géographique d'un haplotype (NCA)

Dn: Diamètre de la distribution géographique d'un clade supérieur (NCA)

D_{CE}: Distance de Chord

F: Statistique de Fu et Li

F_{CT}: Indice de diversité entre groupes de populations

F_{NULL}: Fréquence d'allèles nuls (paramètre)

$^F_{NULL}$: Fréquence d'allèles nuls (estimation)

F_{SC}: Indice de diversité à l'intérieur d'un groupe de populations

F_{ST}: Indice de diversité inter-population

ϕ_{ST}: Indice de diversité moléculaire inter-population

GH: Hormone de croissance

H_E: Diversité génétique ou hétérozygosité

H_{E-TOT}: Hétérozygosité calculé sur les allèles nuls et visibles (paramètre)

$\hat{H}_{E-TOT}$: Hétérozygosité calculé sur les allèles nuls et visibles (estimation)

H_{E-VIS}: Hétérozygosité calculé sur les allèles visibles (paramètre)

$\hat{H}_{E-VIS}$: Hétérozygosité calculé sur les allèles visibles (estimation)

HIM: Modèle hiérarchique de migration en îles.

HKA: Test Hudson-Kreitmen-Aguade

H_O: Fréquence d'hétérozygotes (paramètre)

$\hat{H}_O$: Fréquence d'hétérozygotes (estimation)

HWE ou HW: Équilibre de Hardy-Weinberg

IAM: modèle de mutations à infinité d'allèles

ISM: modèle de mutations à infinité de sites

k: nombre d'allèles

μ: Taux de mutations

maxF: Statistique de la méthode POST

MCMC: Simulations de Monte-Carlo de chaînes de markov

MCT: Test Micro-Checker

MK: Test McDonald et Kreitman

MD: Fonction de probabilité de max F pour des populations mélangées

mtDNA: ADN mitochondrial

N_{VIS}: nombre d'invidus ayant au moins un allèle visible (paramètre)

$\hat{}N_{VIS}$: nombre d'invidus ayant au moins un allèle visible (estimation)

NCA: Analyse cladistique emboitée

NI: Indice de neutralité

N_E: Taille efficace

π: Diversité moléculaire

PCR : Réaction en chaîne de polymérase

POST: Méthode POpulation STructure

PD: Fonction de probabilité de max F pour des populations pures

R: Richesse allélique

r_{BROOK1}: Estimateur d'allèles nuls 1 de Brookfield

r_{BROOK2}: Estimateur d'allèles nuls 2 de Brookfield

r_{CHAK} : Estimateur d'allèles nuls de Chakraborty

RDA: Analyse de redondance

R_{ST}: Indice de diversité inter-population en tenant compte de la taille des microsatellites

SAMOVA: Analyse de variance moléculaire spatiale

SMM: modèle de mutations pas à pas

SSCP: Polymorphisme de conformation par simple brin

STR: courts tandems répétés

θ_K: Estimation de la diversité en fonction du nombre d'allèles

θ_π: Estimation de la diversité en fonction des différences entre paires d'allèles

θ_S: Estimation de la diversité en fonction du nombre de sites polymorphes

TPM: modèle de mutation en deux phases

TRY: trypsine

UT: Test de U

W: Avantage sélectif absolue

w: Avantage sélectif relatif

Il n'est pas facile de distinguer le futile de l'essentiel,
l'anecdotique de l'exemplaire,
les sentiers borgnes des vrais chemins.
Mais j'avancerai les yeux ouverts.

-Amin Maalouf

Remerciements

Mes premiers remerciements vont évidemment à mon directeur, Bernard Angers, qui m'a dirigé de façon impeccable tout au long de ce projet. Ce sont, en grande partie, ses conseils judicieux ainsi que la confiance qu'il m'a démontrée, qui m'auront permis d'amener ce projet à terme. Sa rigueur exemplaire, surtout à la fin, m'aura donné la force d'éviter les raccourcis.

Je voudrais aussi remercier les étudiants du laboratoire pour leur support constant et leur humour grinçant. Énumérer l'ensemble des visages qui sont passés devant moi durant ces cinq dernières années serait trop long. Mais, je vous promets qu'en ce moment, je me souviens de chacun de vous, des plus anciens aux plus récents, de Marie-Claire à Émilie en passant par Rolland et Fred. L'apport qu'aura eu votre présence dans ma vie tant personnelle que professionnelle est indéniable et surtout, irremplaçable.

Nombreux sont ceux qui m'auront aidé soit sur le terrain ou au laboratoire. Le fait que vous m'ayez survécu (!) démontre votre force de caractère. Je remercie donc Kim St-Pierre, Geneviève Roy, Rolland Vergilino, Isabelle Bouthillier, Isabelle Gaudet, Bénédicte Poncet et Camille Madec. J'offre également un remerciement tout spécial à Pascale Gibeau, pour la passion qu'elle m'aura démontrée pendant le bout de chemin que nous aurons partagé.

Je ne peux passer sous silence mon amitié profonde envers Olivier Gauthier, Naïty Jacel et Judith Bouchard, ainsi que l'amour que je porte à ma mère. Sans eux, je n'aurais pas passé au travers le chemin cahoteux que la vie m'aura obligé à suivre en parallèle de mon doctorat. À ce sujet, passer sous silence l'aide que m'aura apportée Pascale Desrosiers serait tout à fait ingrat. Un simple merci est insuffisant. Mais je souhaite tout de même en immortaliser la teneur dans ces quelques lignes.

Finalement, l'aboutissement de cette thèse n'aurait pu être possible sans la présence dans ma vie d'une jeune femme formidable. Aussi, je souhaite dédier ces dernières lignes à

Marie-France Dalcourt. Merci pour ta simplicité, ta complicité, ta confiance et ta compréhension. Tu es la plus belle et je t'aime sincèrement, tout simplement.

Ce projet de recherche a été financé par des subventions de recherche provenant du Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), et du Fond Québécois de la Recherche sur la Nature et les Technologies (FQRNT). J'aurai également été personnellement financé par des bourses d'études du FQRNT, du Fond Étienne-Magnin, du Département des Sciences Biologiques et de la Faculté des Études Supérieures.

Introduction

Les mécanismes évolutifs à la source du niveau de diversité génétique dans une population sont nombreux. Or, l'identification de ces mécanismes permettrait de mettre en lumière les processus par lesquels les populations peuvent s'adapter à leur environnement (Mayr 1963, Krimbas 1984, Taylor 1991). L'acquisition de telles connaissances, au point de vue de la préservation des populations naturelles, serait évidemment primordiale puisqu'elle permettrait d'entrevoir une meilleure gestion des populations naturelles. De telles stratégies de conservation, qui tiendraient compte de l'histoire évolutive des populations ainsi que de l'influence des pressions locales de sélection, auraient ainsi l'avantage d'être efficaces et réalistes.

L'histoire évolutive d'une population comporte plusieurs compartiments, chacun associés à des échelles temporelles et/ou spatiales différentes. C'est ainsi que la composition génétique d'une population résultera de l'interaction des processus agissant à l'intérieur ou entre les populations, à court ou long terme (Figure 1). Certaines régions précises se verront en plus modeler par les effets historiques et récents de la sélection naturelle. Le bagage génétique résultant passera alors à nouveau par le filtre de la sélection naturelle et dictera la capacité d'adaptation d'une population aux pressions de sélection locales (Figure 1).

L'analyse des données moléculaires combinée aux théories de génétique des populations permet de croire que les effets de ces différents processus pourraient être dissociés par l'analyse conjointe du polymorphisme des régions sous sélection avec celui de régions neutres (Lynch et al. 1999). Pour ce, il est cependant primordial de décrire i) les sources du polymorphisme observé sur le génome des organismes, ii) les effets des processus agissant à l'intérieur des populations, iii) les impacts des facteurs agissant entre les populations et iv) les formes de sélection. Ces sujets couvrent tous de larges évantails d'aspects. Aussi, dans les sections suivantes, de manière à assurer la compréhension des chapitres qui suivront, seuls ceux qui sont abordés dans la thèse y sont traités.

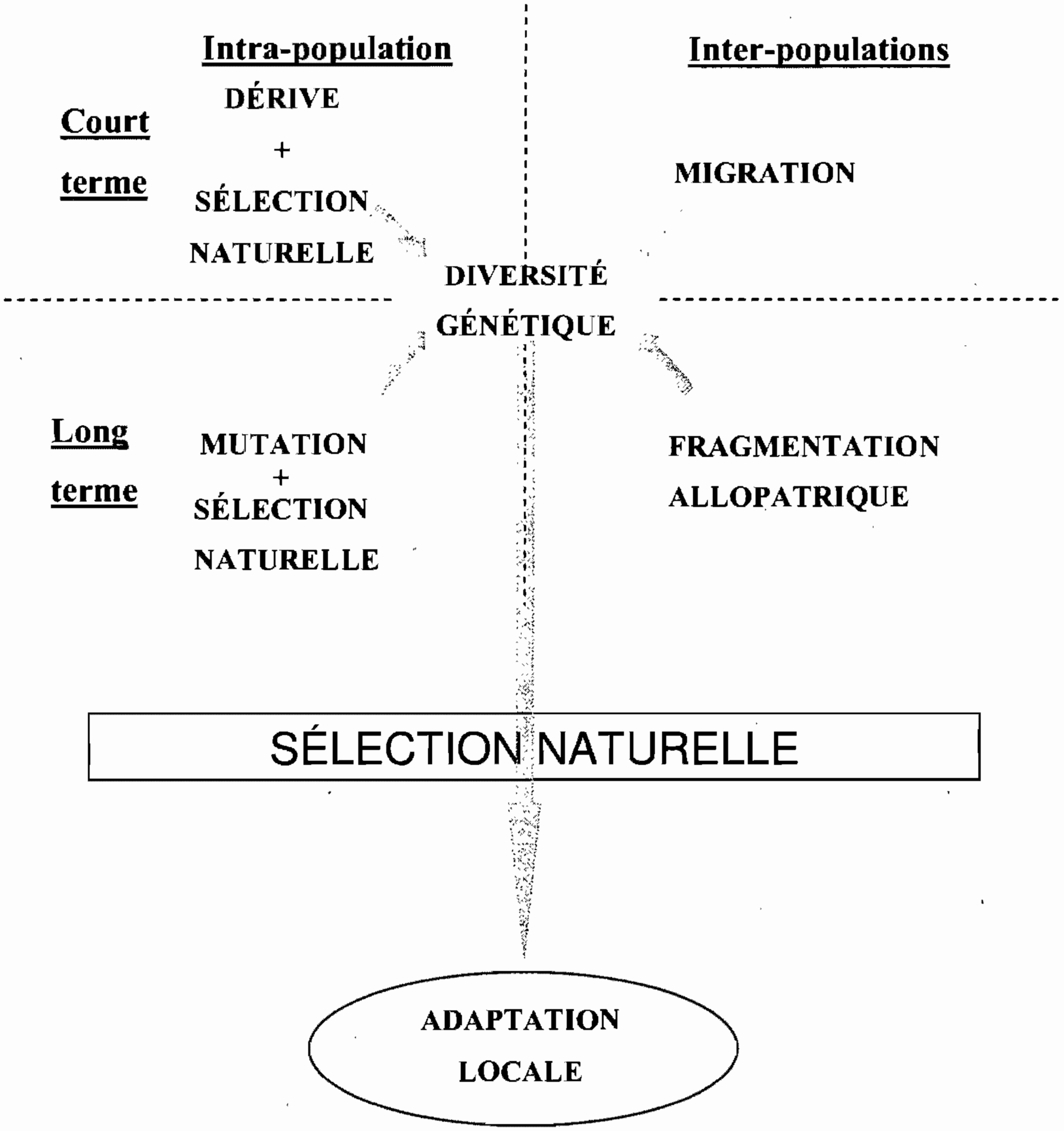


Figure 1: Les différents mécanismes évolutifs intra- et inter- populations ayant façonné à court et long terme la diversité génétique d'une population. Le résultat de leurs interactions passe alors à travers le filtre de la sélection naturelle, menant ainsi la population à l'adaptation aux conditions environnementales locales.

Les sources du polymorphisme

Sur une région donnée du génome (locus), différentes séquences nucléotidiques peuvent être observées. Ces différentes séquences sont des allèles (Stephens 2001). Au niveau d'un locus codant, les allèles assumeront tous la même fonction mais chacun avec leurs propres modalités (Russell 2002). Un gène comportera en général toujours plus d'un allèle permettant ainsi l'émergence d'une certaine diversité au niveau de son expression.

L'analyse de la structure de ce polymorphisme génétique passe par la compréhension de ses différentes sources. L'apparition de nouveaux allèles ainsi que les changements au niveau de leurs combinaisons proviennent des modifications héréditaires du matériel génétique. Dans un cadre de génétique des populations, ces modifications prennent leur source de deux processus fondamentaux, les mutations géniques et les recombinaisons génétiques.

Les mutations géniques

Les mutations géniques correspondent à des changements dans une séquence d'ADN par substitutions, délétions ou insertions de nucléotides. Sur les régions codantes, les séquences sont organisées par triplets de nucléotides appelés codons. La combinaison des quatre nucléotides permet la formation de 64 codons, dont 61 codent pour l'un ou l'autre des 20 acides aminés composant les protéines, les trois derniers servant d'indication pour l'arrêt de la traduction (Horton et al. 1993). Les deux premières positions suffisent souvent à spécifier un acide aminé donné, la troisième position étant souvent redondante. Cette redondance du code génétique nous permet de classer les substitutions sur une région codante du génome en deux catégories: celles qui ne provoquent pas de changement d'acide aminé (synonymes) et celles qui en provoquent (non synonymes). L'expression d'un allèle provenant d'une mutation synonyme sera la même que celle de l'allèle duquel il provient. Au contraire, une mutation non synonyme amènera l'émergence d'un allèle dont l'expression sera différente.

Les délétions et insertions sont en général surtout observées sur les régions non-codantes. Ces phénomènes ont souvent pour effet de modifier le cadre de lecture des régions codantes lors de la traduction, résultant en la création de protéines la plupart du temps non fonctionnelles ce qui mène à leur élimination (Halliburton 2004). Ainsi, à l'intérieur d'une même espèce, un locus codant différera surtout par des substitutions (Kreitman 1983, McDonald et Kreitman 1991)

Le taux de mutation

Le taux de mutations (μ) est défini comme étant la probabilité d'apparition d'un type donné de mutation par unité de temps, lui-même défini en générations de cellules, de réplications cellulaires ou d'organismes (Hartl et Clark 1997). Ce taux de mutation varie en fonction du type de mutation. Par exemple, les mutations synonymes observées sont plus fréquentes (10^{-9} mutation/ nucléotide/génération) que les mutations non-synonymes (10^{-10} mutation/nucléotide/ génération; Li 1997). Le taux varie également en fonction du type de séquences dans lesquelles les mutations se produisent. Le taux de mutations du génome mitochondrial est de l'ordre de 10^{-8} mutation/nucléotide/génération (Russell 2002).

Bien que ce taux semble bas, il a été démontré que cette source de variations a un effet non négligeable au niveau de la création de nouveaux allèles dans les populations avec un grand effectif. Par exemple, pour un taux aussi bas que 10^{-9} mutation/ nucléotide/génération et pour un organisme comme l'humain dont les gamètes contiennent environs $3x10^{9}$ nucléotides, on s'attend à l'apparition de 3 nouvelles mutations en moyenne par gamète à chaque génération. Les humains étant des organismes diploïdes, un embryon humain contiendra donc, en moyenne, 6 mutations absentes chez ses parents. Ainsi, dans le cas de la population humaine totale (6 milliards d'individus), on s'attend à la présence de 36 milliards de mutations qui n'étaient pas présentes à la génération précédente. Cet exemple simplifié, tiré de Hartl et Clark (1997), montre bien que les mutations géniques, bien qu'au taux relativement faible, puissent être une source de polymorphisme allélique importante.

Modèles de mutation

Il existe plusieurs modèles permettant de décrire la dynamique des mutations et de quantifier la variation allélique attendue. Les prédictions des différents modèles varieront tant au niveau du nombre d'allèles attendu que de leur distribution de fréquences. Or, l'estimation de plusieurs des paramètres reliés aux populations se fait à l'aide de la diversité allélique et dépendra ainsi du choix du modèle de mutation. Le choix de l'un ou l'autre de ces modèles se fera en fonction du type de locus étudié.

Modèle à infinité d'allèles

Le modèle à infinité d'allèles (IAM) proposé par Kimura et Crow (1964) reflète le nombre astronomique potentiel d'allèles associés à une région donnée du génome. Ainsi, toute mutation, dont la fréquence dépendra de μ, provoquera la création d'un nouvel allèle préalablement absent de la population. Les allèles sont considérés génétiquement équidistants. Ce modèle est principalement utile lors de l'analyse de distributions de fréquences d'allèles pour lesquels aucune information moléculaire n'est disponible (Neuhauser 2001 et les références à l'intérieur).

Modèle à infinité de sites

Le modèle à infinité de sites (ISM), proposé par Watterson (1975), est basé sur l'idée que les sites nucléotidiques du génome ne sont pas indépendants. Selon ce modèle, une mutation ne peut se produire deux fois sur un même site. Un site ne peut donc être représenté que par deux états, l'état original et l'état mutant. Les K sites divergents sont répartis entre de 2 à 2^K allèles. Ce modèle est utilisé dans le cas d'analyse du polymorphisme entre des allèles nucléaires ou mitochondriaux dont la séquence est connue (Tajima 1989, Neuhauser 2001).

Modèles de mutation pas à pas

Le modèle précédent est cependant inapproprié si le locus sous études n'évolue pas par modifications ponctuelles de nucléotides. Or, c'est le cas des loci appelés microsatellites. Ces loci sont formés de séquences répétées d'unités de 1 à 5 nucléotides. Le polymorphisme de ces marqueurs est représenté par des variations du nombre d'unités (Wyman et White 1980). À cause de leur structure par répétitions, l'analyse du polymorphisme des microsatellites doit se faire en utilisant des modèles de mutations pas à pas. Le modèle de mutation pas à pas (SMM; Ohta et Kimura 1973, Kimura et Ohta 1978) exprime les allèles par des entités discrètes (... A_{-1}, A_0, A_1...). Chacun de ces allèles ne peut muter que d'un pas vers la gauche ou d'un pas vers la droite. Ainsi, A_0 ne peut qu'être modifié en A_{-1} ou A_1, A_1 ne peut qu'être modifié en A_0 ou A_2, etc. Le taux de mutation est également réparti ($\mu/2$) entre les deux directions possibles. Un allèle peut donc apparaître plusieurs fois de manières indépendantes (homoplasie) et les mutations inverses sont possibles. Cependant, certains microsatellites semblent montrer une grande instabilité au niveau de leur taille. Dans ce contexte, Di Rienzo et al. (1994) ont proposé une modification au modèle SMM en y introduisant un paramètre représentant la probabilité (entre 5 et 10%) d'une mutation de plusieurs pas. Selon ces auteurs, la diversité de la plupart des microsatellites s'ajuste mieux à ce modèle en deux phases (TPM) qu'à un modèle SMM strict, tel que celui décrit précédemment.

Les recombinaisons génétiques

Un organisme est composé de plusieurs loci, liés à différents niveaux soit par une proximité physique sur un même chromosome (Russel 2002) ou bien par des interactions fonctionnelles telles que l'épistasie (Beaudry 1985). Ainsi, un portrait adéquat de la dynamique évolutive des organismes vivants ne peut découler de la seule étude du nombre et de la distribution de fréquences des allèles d'un unique locus. L'analyse de plusieurs loci simultanément invoque alors le concept de recombinaison génétique. Ce concept est défini comme tout processus permettant un réarrangement de la combinaison des allèles chez la descendance (Halliburton 2004). Or, l'importance des processus permettant la recombinaison génétique variera en fonction de la distance séparant les loci.

L'assortiment aléatoire

L'assortiment aléatoire est un processus qui permet la recombinaison allélique de loci situés sur des chromosomes non homologues chez les organismes diploïdes et sexués (Halliburton 2004). Lors de la méiose, les paires de chromosomes sont brisées et ces derniers sont distribués aléatoirement dans les gamètes (Lodish et al. 1995). Cet assortiment aléatoire des chromosomes non homologues permet ainsi la formation de différentes combinaisons d'allèles. Le nombre de possibilités d'arrangements chromosomiques est proportionnel au nombre de paires de chromosomes. Ainsi, pour un organisme qui comporte 23 paires de chromosomes comme l'humain, il y a 2^{23} assortiments possibles. Chacun de ces assortiments est en théorie équiprobable (Hart et Clark 1997).

Crossing-over et conversion génique

La recombinaison génétique de locus se situant sur le même chromosome ne peut se faire que via un processus d'échange entre chromatides homologues. Le *crossing-over* correspond à un échange réciproque (Russel 2002) alors que l'échange est unidirectionnel au cours d'une conversion génique (Schibler et al. 2000). Il en résulte une paire de chromosomes comportant un mélange de segments provenant des chromosomes maternel et paternel originaux et donc une combinaison d'allèles différente de celle retrouvée chez les parents.

Polymorphisme intra- populationnel.

Les facteurs démographiques modulant le polymorphisme génétique sont une composante majeure de l'évolution des populations. Au niveau démographique, le polymorphisme génétique d'une population sera déterminé par l'équilibre entre les mutations et la dérive génétique (Hartl et Clark 1997, Neuhauser 2001), de sa taille efficace ainsi que de sa stationnarité dans le temps (Avise et al. 1984, Harpending et al. 1998, Fay et Wu 1999, Hay et Harris 1999).

La dérive génétique

Le nombre d'individus composant une population étant fini, le nombre de gamètes disponibles pour la reproduction est également fini. Or, les processus reliés d'un côté à l'assortiment aléatoire des chromosomes lors de la formation des gamètes ainsi que l'échantillon aléatoire de ces dernières lors de la formation de la génération suivante font que les fréquences alléliques peuvent varier d'une génération à l'autre par la seule force du hasard. Cette dérive génétique agit sur l'ensemble des loci de manière indépendante.

Une population peut prendre plusieurs directions évolutives différentes par dérive. La stochasticité de ce processus empêche la prédiction exacte de l'évolution des fréquences alléliques d'une génération à l'autre. Cependant, il est possible de simuler les effets de la dérive à l'aide de modèles stochastiques permettant d'établir la probabilité des différentes directions évolutives possibles en fonction des fréquences alléliques actuelles. Le modèle le plus connu est celui développé par Wright et Fisher (Fisher 1930, Wright 1931) qui représente, de manière simplifiée, la transmission des allèles d'une génération à l'autre. Selon ce modèle, les fréquences alléliques d'une génération de N individus diploïdes sont déterminées par l'échantillonnage aléatoire avec remise de leurs 2N allèles dans le *pool* allélique de la génération précédente. Les simulations basées sur ce modèle de transmission allélique ont permis de déterminer les effets de la dérive sur le polymorphisme génétique ainsi que l'importance relative du phénomène en fonction de la taille des populations.

Dans le cadre de ce modèle, plus le nombre de copies d'un allèle est grand, plus la probabilité qu'il soit échantillonné lors de la formation des gamètes est élevée. La dérive mène donc inévitablement, lorsqu'il s'agit de la seule force évolutive en jeu dans une population de taille finie, à une diminution graduelle du nombre d'allèles (Buri 1956). Lorsqu'il n'en reste plus qu'un, ce dernier allèle est alors fixé. La probabilité qu'un allèle présent dans une population se fixe après un temps t est égale à sa fréquence relative.

L'intensité de la dérive est inversement proportionnelle à la taille efficace de la population (N_E; Kimura et Ohta 1969, Neuhauser 2001). Les individus d'une population ne participent pas tous à la reproduction. Ainsi, les effets de la dérive ne sont pas fonction de la taille totale de la population mais bien seulement du nombre d'individus qui participent à la reproduction. En termes théoriques, N_E désigne la taille d'une population idéale ayant un niveau de dérive génétique similaire à celui observé dans la population sous étude (Hart et Clark 1997). N_E sera fonction de différentes variables soit le ratio mâles/femelles, la variance du succès reproducteur, la taille des générations précédentes et la dispersion des individus.

L'équilibre mutation-dérive

Dans les populations naturelles, l'effet de la dérive génétique s'opposera à celui des mutations. L'importance de ces deux phénomènes dictera le nombre d'allèles pouvant être maintenus dans une population. En l'absence d'autres forces évolutives, les deux processus mèneront vers un point d'équilibre (l'équilibre mutation-dérive). Dans ce cas, le nombre d'allèles ne change pas car ceux perdus par dérive sont remplacés par de nouveaux, créés par les mutations. Ainsi, sous cet équilibre, la fréquence d'un allèle peut changer mais la distribution de fréquences de l'ensemble des allèles d'un locus ne changera pas (Ewens 1972). Cependant, un déséquilibre peut aller d'un côté (absence des allèles moins fréquents menant à une diversité allélique plus basse) ou de l'autre (distribution uniforme caractéristique d'une diversité plus élevée; Watterson 1978). Un tel déséquilibre peut être détecté à l'aide d'une comparaison entre le polymorphisme observé et celui prédit sous un équilibre mutation-dérive. Le polymorphisme génétique d'une population est souvent décrit par l'hétérozygotie (H_E). H_E correspond à la probabilité d'échantillonner deux allèles différents dans le *pool* d'allèles d'une population (Nei 1987). Sous un équilibre mutation-dérive, H_E ne dépend que de N_E et μ. Or, ces deux paramètres ne sont pas faciles à estimer. Le problème est contourné à l'aide du paramètre θ, dont la valeur théorique correspond à $4N_e\mu$ pour des organismes diploïdes et sexués (Hart et Clark 1997), mais qui peut être estimé de plusieurs autres façons. Ce paramètre θ est à la base de nombreux tests statistiques dont certains seront abordés plus loin.

Histoire démographique des populations

Un déséquilibre mutation-dérive peut résulter de différentes sources. L'identification de ces sources passe par la présence ponctuelle ou généralisée de ce déséquilibre. En effet, si un seul locus montre un déséquilibre mutation dérive, un phénomène de sélection peut possiblement agir sur ce locus. Par contre, si ce déséquilibre est détecté sur l'ensemble du génome, on doit mettre en cause un processus démographique historique.

Les goulots d'étranglement et les expansions démographiques

Un goulot d'étranglement correspond à une diminution drastique de la taille d'une population en un nombre restreint de générations (Hart et Clark 1997). La sévérité d'un tel événement sera fonction de la magnitude de la diminution ainsi que le nombre de générations sur lequel celle-ci s'est effectuée. Au niveau génétique, un goulot d'étranglement provoquera une diminution de la diversité génétique, tant au niveau du nombre d'allèles que de H_E (Nei et al. 1975). Cependant, le nombre d'allèles sera affecté de façon plus importante que H_E en raison de la disparition des allèles rares, qui sont plus propices à disparaître compte tenu de leur fréquence plus faible. En effet, la disparition de ces allèles affectera directement le nombre d'allèles, mais très peu H_E qui tient compte implicitement des fréquences alléliques. Il y aura ainsi un excès de H_E par rapport à ce que prédit l'équilibre mutation-dérive pour le nombre d'allèles observé.

L'expansion d'une population correspond à une augmentation de la taille d'une population en fonction du temps (Fay et Wu 1999). Au niveau génétique, ce phénomène démographique provoque également un déséquilibre entre les mutations et la dérive menant cette fois-ci à une augmentation de la diversité tant au niveau du nombre d'allèles que de H_E. Cependant, dans ce cas-ci, le nombre d'allèles augmentera plus rapidement que H_E. Ainsi, l'expansion provoquera un déficit de H_E par rapport à ce que prédit l'équilibre mutation dérive.

La détection des événements démographiques

La détection des événements démographiques est basée sur l'évaluation du H_E de la population et de sa probabilité sous un équilibre mutation dérive. Cette probabilité est déterminée à l'aide de la distribution de fréquences de H_E sous cet équilibre selon le nombre d'individus N (2N copies d'allèles), le nombre d'allèles et le modèle de mutation assumé pour le locus sous étude. La distribution de fréquences de cette hétérozygotie est construite à partir de simulations basées sur la théorie de la coalescence.

La théorie de la coalescence

Selon le modèle de Wright-Fisher, un individu peut donner plusieurs copies d'un même allèle à la génération suivante. Ces copies d'allèles sont alors identiques par descendance au niveau de la génération de cet individu. Les générations se succédant, les lignées de cette copie ancestrale peuvent soit se perpétuer dans le temps, subissant ou non des mutations, ou bien disparaître par dérive génétique (Hartl et Clark 1997). Par extension, une population peut être exprimée comme un ensemble d'allèles liés à divers degrés à un ou des allèles ancestraux (Nordborg 2001).

La description des relations entre les allèles d'une population en remontant les générations jusqu'à l'atteinte de l'ancêtre commun à tous ceux-ci est la base de la théorie de la coalescence. Un événement de coalescence est une fusion de deux lignées d'allèles dans le temps (Nordborg 2001). Sous un équilibre mutation-dérive, ce processus de coalescence est stochastique et indépendant pour chaque génération (Neuhauser 2001). Des calculs probabilistes sous un modèle de Wright-Fisher ont permis de tirer deux grandes règles sur le nombre de générations menant à un événement de coalescence dans une population à l'équilibre mutation-dérive. 1) La probabilité d'un évènement de coalescence à la génération précédente est inversement proportionnelle à la taille de la population pour un nombre de lignées d'allèles donné. Les événements de coalescence sont donc plus rapprochés dans le temps dans les populations de petite taille (Figure 2; Nordborg 2001). 2) La probabilité de coalescence à la génération précédente est inversement proportionnelle au nombre de lignées d'allèles. Ainsi, plus on remonte une généalogie, plus le nombre de

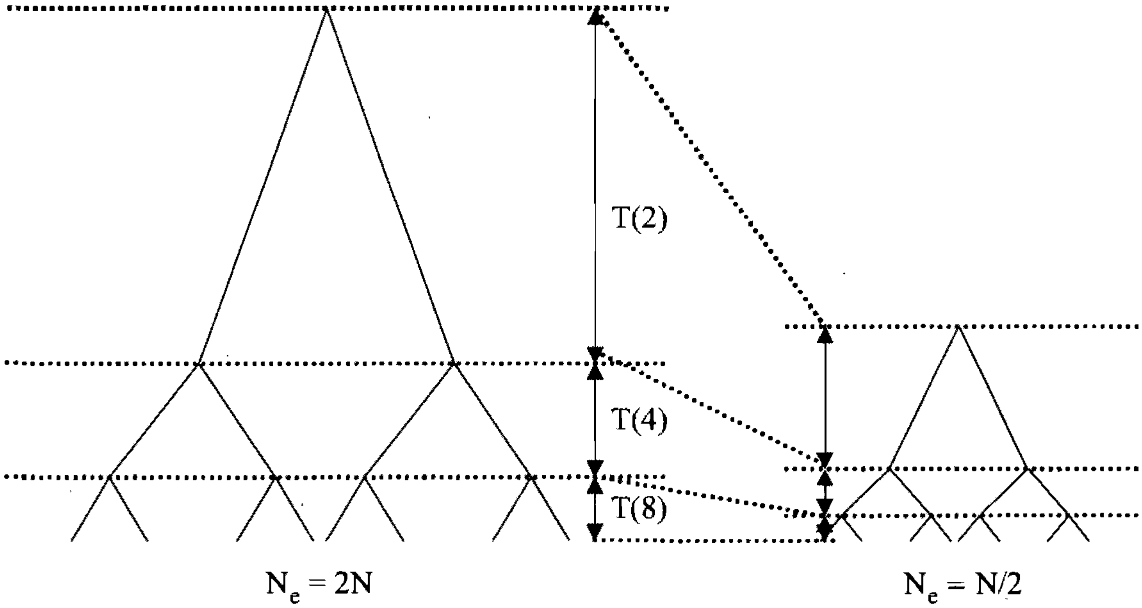


Figure 2: Généalogie de 8 lignées d'allèles jusqu'à leur ancêtre commun dans deux populations à l'équilibre mutation dérive mais de taille efficace différente.

lignées diminue, et plus le nombre de générations entre chaque événement de coalescence est important (Nordborg 2001). Ainsi, le temps de coalescence moyen T entre deux lignées d'allèles augmente de manière exponentielle lorsque l'on remonte la généalogie (Figure 2). Cependant, la stochasticité du processus de coalescence fait que la variance associée à T est près du carré de sa valeur rendant le nombre d'arbres de coalescence d'une population à l'équilibre mutation-dérive astronomique (Neuhauser 2001).

Sous un équilibre mutation-dérive les processus de coalescence et les mutations sont indépendants (Nordborg 2001). Il est donc possible de combiner l'information généalogique à celles des mutations. Celles-ci se produisent de façon aléatoire de long des branches d'une généalogie. Or, en assumant que le taux de mutations est constant, la probabilité d'observer des mutations sera proportionnelle à la longueur des branches séparant les événements de coalescence (Cornuet et Luikart 1996). Ainsi, les nombreuses possibilités d'arbres de coalescence pour une population stationnaire servent de base à la génération des distributions de fréquences de H_E pour un locus sous équilibre mutation-dérive.

Les microsatellites

Par leur très grand polymorphisme (Weber et Wong 1993), leur abondance dans le génome des eucaryotes (Jarne et Lagoda 1996) et la quasi absence d'emprise directe de la sélection (voir cependant Wiener et al. 2003 et Thuillet et al. 2007), les microsatellites apparaissent comme des marqueurs génétiques de choix pour l'analyse des processus démographiques (Avise 2004). Or, il existe plusieurs méthodes utilisant un modèle de type SMM pour détecter les événements démographiques. Celle de Luikart et Cornuet (1998) compare la valeur observée de H_E sur un locus donné à la moyenne de H_E calculée à partir de simulations de coalescence sous un équilibre mutation-dérive. Cette analyse est faite de manière indépendante sur plusieurs loci. Sous un équilibre mutation dérive, la probabilité d'avoir une valeur observée plus faible ou plus élevée que la moyenne obtenue par simulation est équivalente. Cependant, la présence d'un évènement démographique provoquera un écart entre la proportion de valeurs plus élevées et plus faibles que la moyenne attendue. Luikart et Cornuet (1998) ont donc développé une statistique permettant de déterminer la probabilité d'avoir *l* loci sur L avec un excès (ou un déficit) de H_E dans une population à l'équilibre. Un rejet de l'hypothèse nulle permet de conclure à un excès (ou un déficit) de H_E généralisé à plus de loci que ce que prévoit l'équilibre mutation-dérive.

Polymorphisme inter- populations

Les individus d'une espèce se reproduisent rarement de manière aléatoire. Au contraire, ils auront tendance à se reproduire avec une plus grande fréquence avec ceux qui sont à proximité, qui se reproduiront en même temps ou qui occuperont des milieux semblables. Une espèce sera donc subdivisée en unités génétiquement différentiées arrangées spatialement, temporellement ou écologiquement de manière hiérarchique ou non (Excoffier 2001). Cette organisation non-aléatoire de la diversité génétique entre les unités est nommé structure génétique. Si elle n'est pas adéquatement reconnue au préalable, la structure génétique pourrait biaiser l'interprétation du polymorphisme génétique (ou biologique) observé entre les unités. Il est donc primordial de détecter les structures, de décrire leur organisation et de comprendre les processus historiques qui sont responsables de leur mise en place.

Détection des structures génétiques

Effets des structures génétiques

L'équilibre de Hardy-Weinberg (HW) prédit la relation théorique entre les fréquences alléliques et génotypiques d'une population d'organismes diploïdes et sexués. Si les individus d'une population de taille infinie s'accouplent de façon aléatoire, cet équilibre est atteint et en l'absence de sélection, de migration et de mutation, les fréquences alléliques ne varient plus d'une génération à l'autre (Hart et Clark 1997). Bien que les prémisses sous-jacentes à ce principe ne soient généralement pas respectées dans une population naturelle, l'équilibre de HW peut tout de même servir d'hypothèse nulle afin de détecter les structures génétiques dans des populations de grandes tailles ($N_E > 500$ individus selon Hartl et Clark 1997). En effet, la reproduction non aléatoire due à la présence d'une structure génétique aura comme effet d'augmenter le nombre d'homozygotes au détriment des hétérozygotes. La relation génotypes/allèles prédite par l'équilibre de HW ne tient donc plus. Ce déficit en hétérozygotes est appelé l'effet Wahlund.

En plus de l'effet Wahlund l'écart par rapport à l'équilibre de HW en présence d'une structure génétique sera dépendant des processus évolutifs inhérents à chacune des populations. En effet, deux populations en isolement reproducteur verront les processus démographiques et potentiellement la sélection naturelle moduler leur polymorphisme génétique de manière indépendante. Cette modulation mènera à l'acquisition de fréquences alléliques et/ou génotypiques différentes dans chacune des populations. Il s'agit de la différentiation génétique (Raymond et Rousset 1995a). Le niveau de différentiation génétique entre populations sera fonction du flux migratoire qui les unit, et donc de la solidité de la structure. Dans ce contexte, deux populations qui s'échangent un nombre suffisant de migrants efficaces (qui se reproduisent dans leur nouvelle population) subiront une homogénéisation de leur composition génétique. Ces échanges court-circuiteront les processus de différentiation et diminueront la structure.

Quantification et test de la structure

Plusieurs méthodes existent afin de quantifier et tester le niveau de structure séparant des groupes d'individus. Bien qu'utilisant des statistiques différentes, les raisonnements sous-jacents à ces tests convergent tous vers l'une ou l'autre des deux signatures génétiques associées à l'isolement reproducteur: l'effet Wahlund et la différentiation génétique. Il est cependant possible de diviser les tests statistiques permettant d'identifier une structure en deux catégories: 1) les tests de qualité d'ajustement (*goodness of fit*) et 2) les tests basés sur les indices de fixation.

Les tests de qualité d'ajustement

Les tests de qualité d'ajustement permettent généralement de tester l'écart à l'équilibre de HW. Il existe plusieurs statistiques permettant de quantifier cet écart (Levene 1949, Haldane 1954, Hernandez et Weir 1989). Cependant, Rousset et Raymond (1995) ont démontré que, dans le cas d'une hypothèse contraire unilatérale (ici un déficit en hétérozygotes), leur statistique U est la plus puissante. Cette statistique détermine l'écart à l'équilibre de HW à partir des fréquences des homozygotes. La différentiation entre deux populations peut quant à elle être analysée simplement à l'aide d'une table de contingence et d'une statistique χ^2 (Raymond et Rousset 1995a, Goudet et al. 1996). Le calcul de la probabilité associée à cette statistique est généralement fait en utilisant les méthodes de Monte Carlo ou des chaînes de Markov proposées par Guo et Thompson (1992). Une conclusion englobant l'ensemble de ces analyses indépendantes est faite en utilisant la méthode de combinaison de probabilités proposée par Fisher (1954).

Les indices de fixation

La différentiation génétique peut également être décrite via le déséquilibre de liaison entre allèles décrit par l'effet Wallund. Imaginons un système panmictique à 4 allèles (*a*, *b*, *c* et *d*). Dans un tel système, l'équilibre de HW prévoit la présence d'hétérozygotes pour toutes les combinaisons possibles d'allèles, soit *ab*, *ac*, *ad*, *bc*, *bd* et *cd*. Or, imaginons maintenant un même système à 4 allèles, mais composé de deux unités différentiées, soit l'unité 1 où l'on retrouve les allèles *a* et *b* et l'unité 2 composée des allèles *c* et *d*. Comme il y a peu ou pas d'échanges migratoires entre les deux unités (d'où leur différentiation génétique), les hétérozygotes composés d'allèles provenant des deux unités (*ac*, *ad*, *bc*, et *bd*) seront absents contrairement aux prédictions sous un équilibre de HW dans une large population panmictique. À l'inverse, les hétérozygotes de formes *ab* et *cd* seront beaucoup plus fréquents que ce à quoi nous nous attendrions sous HW. Ce phénomène a mené Wright (1951) et Cockerham (1969) à parler de corrélation (ou de covariance) entre les allèles d'un locus. Ainsi, *a* et *b* seraient corrélés positivement dans le système à deux unités, alors que, dans le système panmictique, ils ne le seraient pas, puisqu'on les retrouverait couplés avec les deux autres allèles selon les prédictions de HW. Les indices de fixation se basent sur ce principe pour quantifier le niveau de structure entre deux ou plusieurs unités. Ils quantifient donc le niveau de corrélation des allèles d'une unité génétique donnée (par exemple *a* et *b* de l'unité 1 dans l'exemple précédent) par rapport aux allèles du même locus observés dans un regroupement d'unités génétiques (les allèles *a*, *b*, *c* et *d*).

L'avantage des indices de fixation, c'est qu'ils permettent d'explorer plusieurs niveaux de structure en fonction des définitions de l'unité génétique (un individu, une population, un groupe de populations) et du regroupement d'unité auquel sa composition allélique est comparée (une population, un groupe de populations, l'ensemble des allèles). Les combinaisons d'unités et de groupements permettent d'identifier différents indices de fixations, décrivant des composantes indépendantes et additives de la covariance totale des fréquences alléliques. L'analyse de la corrélation allélique d'une population par rapport à l'ensemble des allèles est désignée par F_{ST} et est souvent utilisée comme indice de différentiation des populations. Si les populations sont organisées en groupes de populations, le F_{ST} peut être divisé en deux composantes: le F_{SC} qui correspond à la covariance des populations à l'intérieur des groupes et le F_{CT} qui correspond à la corrélation allélique à l'intérieur des groupes de populations par rapport à l'ensemble des allèles. Il est également possible d'étendre ce principe à un niveau inférieur à la population, soit aux individus. Dans ce cas, la corrélation allélique chez les individus par rapport à celle de la population est désignée par $F_{IS.}$ Cet indice correspond à l'indice de consanguinité de Wright (la présence de structure étant ici associée à de la consanguinité). Finalement, la corrélation des allèles des individus par rapport à l'ensemble des allèles est désignée par F_{IT}. Cette valeur correspond à la covariance génétique couvrant tous les niveaux mentionnés précédemment ($F_{IT} = F_{IS} + F_{ST}$). Cependant, l'analyse de l'organisation de la structure au niveau des populations se concentrera généralement sur l'analyse du F_{ST} et de ces deux composantes F_{SC} et F_{CT}.

Organisation de la structure des populations

L'analyse de variance moléculaire

Ce type de structure hiérarchique peut être détecté à l'aide d'une analyse proposée par Cockerham (1969) et Cockerham (1973). Cette analyse consiste à extraire la corrélation allélique associée à différents niveaux de subdivisions (populations, groupes de populations) à l'aide des indices de fixation. La signification de ces indices de fixation peut être évaluée à l'aide de permutations selon un patron approprié au niveau hiérarchique auquel l'indice fait référence (Excoffier 2001). Par exemple, la signification du F_{ST} se fera en permutant les individus entre l'ensemble des populations, alors que la signification du F_{CT} se fera en permutant des populations complètes entre les groupes.

Ces analyses n'utilisent que l'information provenant des fréquences alléliques. Excoffier et al. (1992) ont développé une analyse de variance moléculaire (AMOVA) qui permet de détecter des structures hiérarchiques non seulement à partir des fréquences alléliques mais également à partir de leur séquence en incorporant le polymorphisme moléculaire qui les distingue. Ainsi, s'il y a une structure à un niveau hiérarchique donné, deux allèles échantillonnés dans une unité devraient avoir une similarité moléculaire plus grande, que deux allèles échantillonnés au hasard dans l'ensemble des unités. Les sommes des carrés des écarts sont donc une fonction de la distance génétique séparant chacun des allèles. La variance génétique à chacun des niveaux est évaluée de la même façon que la précédente méthode. Bien que leurs indices de fixation soient équivalents aux F quant à leur signification, Excoffier et al. (1992) propose une notation différente afin d'éviter toute confusion. Ainsi ϕ remplace F lorsque la divergence moléculaire est prise en compte lors de l'analyse de la structure. Bien qu'offrant beaucoup de libertés quant aux types de marqueurs pouvant être analysés, ces analyses nécessitent une connaissance préalable de l'organisation de la structure étudiée pour la construction des hypothèses. Si de telles connaissances sont inexistantes, l'utilisation de ces méthodes est donc limitée.

Analyses canoniques

Les analyses canoniques s'avèrent des solutions puissantes et malléables dans les situations où les connaissances sur l'organisation d'une structure sont fragmentaires. Bien que surtout utilisées dans des contextes d'analyse des communautés, Angers et al. (1999) ont démontré leur utilité dans le but de mettre en relation une matrice **Y** décrivant les fréquences alléliques des populations pour un locus donné et une matrice **X** contenant des variables environnementales et géographiques. Ce type d'analyse permet ainsi de faire ressortir les variables externes qui expliquent le mieux la variation génétique entre les populations.

Facteurs historiques

L'ADN mitochondrial

La structure génétique actuelle des populations est le résultat de facteurs évolutifs tels les mutations, la dérive et la sélection. Cependant, la mise en place du bagage génétique modelé par ces processus est le fruit des facteurs historiques qui ont mené à la colonisation des différents milieux habités par les populations. L'analyse du bagage génétique actuel des populations peut permettre la reconstruction des relations historiques qui les relient. L'information génétique, combinée à la distribution géographique des populations sous étude, est à la base des inférences phylogéographiques.

Avise et al. (1987) ont montré que les marqueurs situés sur l'ADN mitochondrial sont propices aux analyses phylogéographiques des populations. En effet, l'ADN mitochondrial est haploïde et, pour la majorité des organismes, a une origine strictement maternelle. Ainsi, la taille efficace de ces marqueurs est théoriquement quatre fois plus petite que celle d'un marqueur nucléaire. Les marqueurs mitochondriaux sont donc particulièrement sensibles à la dérive génétique, menant ainsi à une fixation rapide des allèles. De plus, le taux de mutation relativement bas (10^{-8} mutations/sites/génération) par rapport aux loci microsatellites (de 10^{-3} à 10^{-6} mutation/allèles/génération; Weber et Wong 1993) combiné à la faible taille du génome mitochondrial (16 800 paires de bases) permet d'estimer en moyenne une mutation toutes les 500 paires de bases à tous les 100 000 ans (Angers et Bernatchez 1998). Ces marqueurs sont donc très peu variables. Ainsi, des divergences observées entre les populations contemporaines peuvent être interprétées comme le résultat d'un isolement reproducteur ancien. L'analyse du génome mitochondrial est donc utile pour reconstruire, par exemple, l'origine ainsi que les étapes de la colonisation d'un territoire.

L'analyse cladistique emboîtée (*nested clade analysis*; NCA)

La NCA est l'une des analyses les plus utilisées pour inférer les facteurs historiques (ou récents) ayant façonné l'organisation génétique des populations à partir de marqueurs mitochondriaux (Templeton et al. 1995, Templeton 1998). Ces inférences historiques se font à partir d'un regroupement hiérarchique des allèles mitochondriaux (haplotypes) échantillonnés. La procédure de regroupement des haplotypes suit les règles d'emboîtement proposées par Templeton et Sing (1993) et Templeton et al. (1987). Selon ces règles, les haplotypes divergents d'une mutation sont regroupés en clade. La procédure est répétée avec les clades résultant jusqu'à ce que l'ensemble du réseau soit emboîté (voir Figure 11). Combinant la composition en haplotypes (et en clades) de chacune des populations à leur position géographique, trois distances sont alors estimées puis testées par permutations. 1) La distance moyenne qui sépare chaque individu du centre géographique de l'haplotype ou du clade auquel il appartient. 2) La distance moyenne entre le centre géographique de chaque haplotype et le centre géographiques du clade de niveau supérieur. 3) La différence entre ces deux distances pour les haplotypes (ou clades) situés à l'intérieur et au bout du réseau. Selon Templeton (1998), chaque processus historique (i.e. fragmentation allopatrique, expansion par colonisation, isolation par distance, etc.) aura une signature différente au niveau de ces distances. Dans cette perspective, Templeton a créé une clé dichotomique permettant l'identification du processus historique responsable de la structure étudié et ce à chacun des niveaux hiérarchiques délimités dans le réseau d'haplotypes (Posada et Templeton 2004). Bien que faisant l'objet d'un certain débat dans la littérature au niveau de son exactitude (Petit et Grivet 2002, Templeton 2002), la NCA demeure la seule technique disponible permettant l'incorporation directe d'une dimension géographique à l'analyse cladistique de séquences mitochondriales.

Sélection naturelle

La sélection naturelle repose sur la relation entre les organismes et leur milieu. Selon cette théorie, dans des conditions données, les organismes montreront des différences au niveau de leur survie et/ou de leur capacité à se reproduire. Ainsi, les caractéristiques des organismes qui auront une survie et une reproduction optimale dans ces conditions verront leur fréquence augmentée dans la population. Au contraire, les caractéristiques qui nuiront à des degrés divers à la survie et/ou à la reproduction verront leur fréquence diminuée menant parfois à leur élimination de la population. Ce processus mène à l'adaptation locale des populations.

Pour un trait qui subit les effets de la sélection, ses états (chacun défini par un génotype ou un phénotype donné) ne sont donc pas tous propagés également à la génération suivante. Ces variations entre états sont quantifiées par la valeur adaptative associée à chacun. Cette valeur peut varier selon diverses composantes. Tout d'abord, il y a la viabilité qui représente la probabilité de survie jusqu'à la maturité sexuelle (Salvane et Balino 1998, Merila et Hemborg 2000). Ensuite, il y a la fécondité d'un individu qui représente la quantité absolue de gamètes produites qui formeront un zygote (Pitnick et Markow 1994). Il est également possible de définir une composante reliée à la capacité d'un individu à s'accoupler, certains individus pouvant être victimes de discrimination provenant du sexe opposé ou de différences comportementales (Heckel et von Helversen 2002, McMahon et Bradshaw 2004).

L'étude de la sélection naturelle peut se faire selon différents axes tout dépendant de l'emphase qui est mise sur l'analyse des différences phénotypiques ou génotypiques. Or, l'identification des effets de la sélection naturelle sur les traits phénotypiques n'est pas simple. La plasticité de ces traits due simplement aux conditions environnementales est reconnue pour être très grande (Allendorf et al. 1987; Lynch et al.1999; Bronikowski 2000), ce qui rend difficile l'évaluation de la portion héréditaire de la variation et la quantification des effets de la sélection. Différentes stratégies ont été élaborées pour les mesurer en nature, mais chacune fait face à de nombreux problèmes logistiques. Les transplantations réciproques d'organismes (Bertness et Gaines 1993), les suivis d'organismes marqués sur l'ensemble de leur cycle vital (Plourde et al. 2001) et les mesures en milieux contrôlés en sont quelques exemples (Bronikowski 2000).

Sous prétexte que les effets de la sélection naturelle se reflètent directement sur le polymorphisme génétique des loci sous son influence (Otto 2000), l'analyse directe de la diversité génétique de certains gènes fonctionnels (Johnson et Black 1996; Stanton et al. 1997; Landry et Bernatchez 2001) offre potentiellement une solution aux problèmes associés aux études de génétiques quantitatives décrites plus haut. C'est cette voie qui sera privilégiée dans cette thèse.

Formes de sélection

En fonction de leur impact sur la diversité génétique, plusieurs formes de sélection naturelle peuvent être distinguées. Certaines réduiront le polymorphisme, alors que d'autres le maintiendront. Il est possible de diviser ces formes en deux grands groupes selon qu'elles agissent sur les génotypes ou sur les allèles. Les sélections agissant sur les génotypes peuvent se résumer en la sélection directionnelle, la sélection balancée et l'infériorité des hétérozygotes (revue dans Travis 1989). Deux autres types de sélection seront mis de l'avant en fonction de leurs effets sur les allèles, soit les sélections purificatrice et diversificatrice. Évidemment, la force de la sélection (ainsi que ses formes) varieront au gré des pressions environnementales au cours du temps. Pour fins de simplicité, les descriptions suivantes auront comme prémisses que ces conditions resteront constantes au cours du temps.

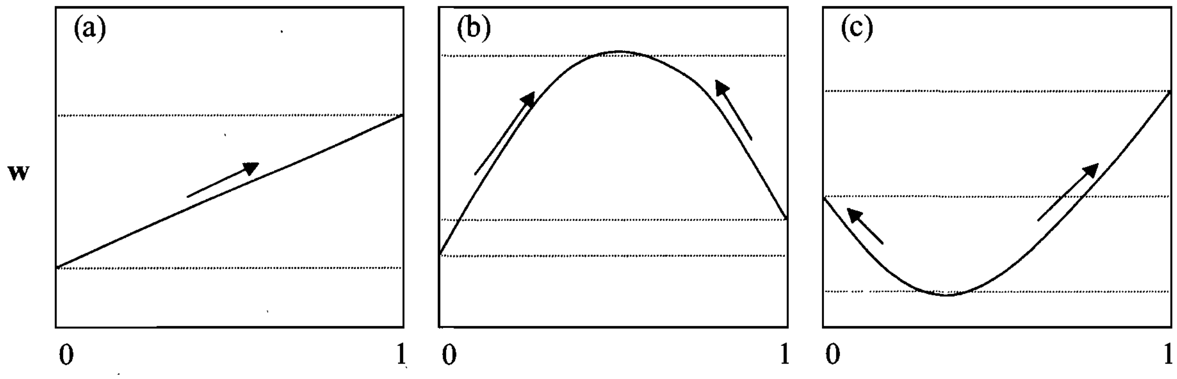
Sélection sur les génotypes

La sélection sur les génotypes fera varier les fréquences alléliques d'une génération à l'autre jusqu'à l'atteinte d'un point d'équilibre, à partir duquel celles-ci ne changeront plus. Dans un système à deux allèles, il existe trois points d'équilibre qui permettent mathématiquement l'obtention d'une différence nulle entre les fréquences alléliques d'une génération à l'autre. Les deux premiers se situent au niveau de la fixation de l'un ou l'autre des deux allèles. Le troisième, dépend des valeurs adaptatives de l'ensemble des génotypes et permet de maintenir un certain degré de polymorphisme dans la population en permettant la présence des deux allèles. Il s'agit d'un équilibre qui n'est présent que dans certaines formes de sélection. Ces trois équilibres sont stables ou instables. Un équilibre stable est un point d'attraction vers lequel les fréquences alléliques s'approcheront sous sélection et qui se traduira par une valeur adaptative moyenne plus élevée (Wright 1931, Coyne et al. 1997). Au contraire, un équilibre instable représente un point de répulsion duquel les fréquences alléliques s'éloigneront par sélection car il se traduit par une diminution de la valeur adaptative moyenne de la population (Wright 1931, Coyne et al. 1997). Ces principes peuvent également être étendus à des systèmes composés de plus de deux allèles.

Les différentes formes de sélection sur les génotypes varieront en fonction de la valeur adaptative des différents génotypes présents dans une population. Ces formes se distingueront par la stabilité ou l'instabilité de leurs trois points d'équilibre ce qui permettra de prédire l'évolution des fréquences alléliques sous chacune d'elles (Figure 3).

Sélection directionnelle

La sélection directionnelle favorisera les génotypes qui comportent un seul des allèles présents dans une population. Le point d'équilibre stable de ce type de sélection se retrouve au niveau de la fixation de l'allèle favorisé. La fixation de l'allèle défavorisé est un équilibre instable (Figure 3a). Cette forme de sélection réduira la variabilité observée sur un



Fréquence allèle favorisé

Figure 3: Représentation des relations entre la valeur adaptative moyenne de la population (w) et la fréquence de l'allèle favorisé dans un système à deux allèles. Les lignes pointillées correspondent aux points d'équilibre des fréquences alléliques pour les sélections directionnelle (a), balancée (b) et pour l'infériorité des hétérozygotes (c). Les flèches pointent dans la direction des équilibres stables.

locus par rapport à ce qui serait observé sous un équilibre mutation dérive (Kimura 1983). Suite à cette réduction (l'atteinte de l'équilibre stable), l'apparition de nouvelles mutations est possible, provoquant la formation d'une généalogie d'allèles en forme d'étoile, à l'image d'une population en expansion (Figure 4b; Tajima 1989).

Sélection balancée

De manière générale, cette forme de sélection favorisera le maintien de plusieurs allèles dans une population. Or une des façons permettant un tel maintien passe par un avantage de l'état hétérozygote (Nordborg 2001). Le point d'équilibre stable de la sélection balancée permet le maintient de plusieurs allèles. D'un autre côté, les points d'équilibre reliés à la fixation de l'un ou l'autre des allèles représentent des équilibres instables (Figure 3b). L'effet de la sélection se traduira par une distribution de fréquences alléliques uniforme. Ce maintien des allèles sur plusieurs générations fera en sorte que leur polymorphisme sera similaire à celui détecté sur les loci d'une population ayant subi un goulot d'étranglement (Figure 4a; Tajima 1989).

Infériorité des hétérozygotes

Comme son nom l'indique, cette forme de sélection défavorisera les hétérozygotes. Dans ce cas-ci, il existe plusieurs points d'équilibre stables. Dans un système à deux allèles, les équilibres stables seront la fixation de l'un ou l'autre des deux allèles. La présence de plusieurs allèles est possible mais représente un équilibre instable (Figure 3c). L'infériorité des hétérozygotes correspond à une forme de sélection qui dépend positivement des fréquences alléliques (Cherry 2003). Ce mécanisme n'est pas relié à la valeur adaptative des allèles. Ainsi, par l'action seule de cette forme de sélection, un allèle avantageux ne proliférera pas dans une population si sa fréquence allélique est basse. Dans ce contexte, l'infériorité des hétérozygotes est parfois considérée comme une barrière à la fixation d'allèles qui amélioraient la valeur adaptative moyenne de la population (Cherry

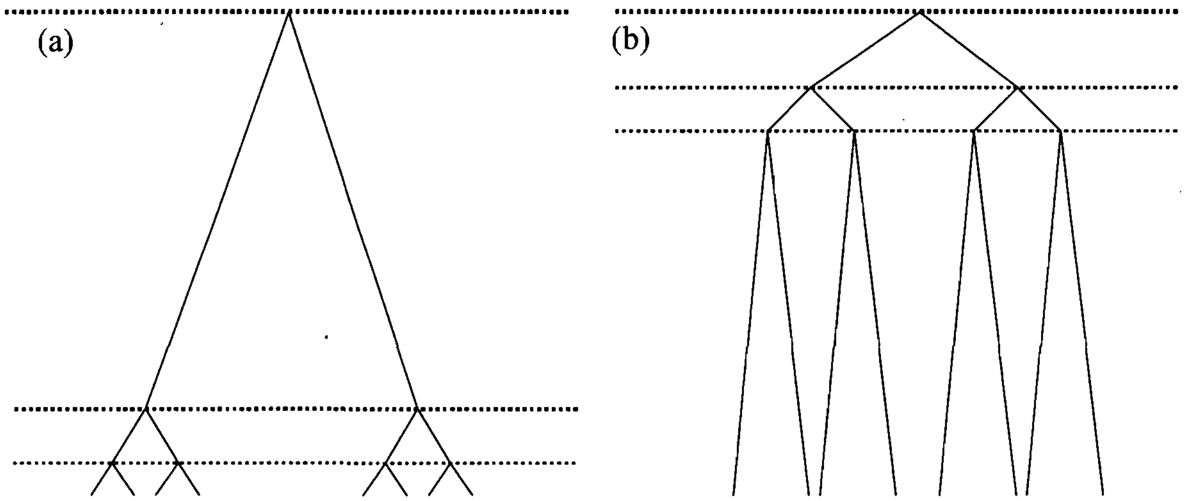


Figure 4: Effets de la sélection balancée (a) et de la sélection directionnelle (b) sur la généalogie des allèles d'un locus.

2003). Si l'équilibre stable est atteint (fixation de l'un ou l'autre des allèles), ce type de sélection ne peut être différentié de la sélection directionnelle.

Sélection des allèles

Sélection purificatrice

Les mutations créent continuellement de nouveaux allèles. Or, sur un locus codant donné, les mutations ne sont pas toutes avantageuses. Au contraire, certaines mutations peuvent faire diminuer la valeur adaptative des individus qui les portent, faisant du coup baisser la valeur adaptative moyenne de la population (fardeau génétique ou *genetic load;* (Whitlock 2002, Paland et Schmid 2003). Ce sont des allèles délétères. Ces allèles ne prolifèrent que rarement dans une population grâce à la sélection purificatrice. Ainsi, en l'absence d'autres forces, ces allèles seront progressivement éliminés de la population à un rythme qui dépendra de la force de la sélection, ou en d'autres termes, de l'effet délétère d'un allèle (Otto 2000). Ainsi, plus un allèle est délétère, plus la sélection purificatrice sera forte et plus il sera éliminé rapidement de la population. La sélection purificatrice diminuera la variabilité du locus sous sélection au niveau des sites non-synonymes (Kreitman et Hudson 1991). Cette forme de sélection n'est pas mutuellement exclusive avec celles décrites précédemment, son action étant parfois combinée avec celle de ces dernières, surtout avec la sélection directionnelle (Kim et Stephan 2000).

Sélection diversificatrice

De façon générale, cette forme de sélection favorise les génotypes différents (qui sont composés d'allèles différents). Elle diffère légèrement de la sélection balancée par le fait qu'elle ne favorise pas seulement les hétérozygotes mais bien les allèles différents. Ainsi, la sélection diversificatrice pourra favoriser la descendance provenant d'accouplements entre homozygotes si ceux-ci sont composés d'allèles différents (AA et BB). De plus, si la sélection balancée favoriserait l'accouplement entre des hétérozygotes AB, ça ne ne serait pas nécessairement le cas pour la sélection diversificatrice qui favoriserait par exemple des accouplements entre AB et CC avant ceux entre AB simplement parce que cela permettrait l'apparition de mélange d'allèles inédit. Il y aura donc une pression de sélection qui se traduira par la présence d'un polymorphisme élevé au niveau du nombre de mutations non-synonymes (Hughes et Nei 1988, Takahata et Satta 1998). Des exemples qui reviennent fréquemment dans la littérature sont les systèmes CMH chez les vertébrés (Hughes et Hughes 1995, Hughes et Nei 1988) et les systèmes d'auto-incompatibilité chez les végétaux (Ioerger et al. 1991). La sélection diversificatrice agit souvent de concert avec la sélection balancée dans le maintien du polymorphisme dans les populations naturelles (Cereb et al. 1997).

Sélection fréquence-dépendante

Il existe essentiellement deux formes de sélection fréquence-dépendante (SFD): la SFD positive et la SFD négative. Les SFD se distinguent des autres formes par le fait qu'elles agissent en fonction de la composition génétique de la population et non en fonction de conditions environnementales. La SFD positive favorise l'allèle qui a la plus forte fréquence dans une population, menant ainsi à sa fixation. De son côté, la SFD négative favorisera l'allèle qui aura la plus faible fréquence, jusqu'à ce que cet allèle perde ce statut. Ce type de sélection très dynamique puisque sa cible change continuellement, mène à un maintien de la diversité. Si la SFD positive reste théorique (on ne peut distinguer ses effets de ceux des sélections directionnel et purificatrice), certaines études ont suggéré la présence de SFD négative dans les cas d'auto-incompatibilité (Schueler et al. 2006, Billiard et al. 2007) et de relations hôte-pathogènes (Borghans et al. 2004, Meyer-Lucht et Sommer 2005).

Détection des formes de sélection

La détection des effets de la sélection sur un locus est une tâche complexe compte tenu des différents facteurs qui modulent le polymorphisme du génome (Otto 2000). Afin de dissocier les effets de la démographie et de la structure génétique des populations de ceux de la sélection, il importe d'analyser conjointement la variabilité du locus sous étude avec celui de régions neutres. Ainsi, si des patrons de polymorphisme différents sont détectés pour ces deux types de régions, une forme quelconque de sélection naturelle sera déclarée responsable de cette différence.

En fonction de l'échelle de temps sur laquelle elle agit, la sélection naturelle laissera des signatures différentes sur le génome des organismes qui lui sont soumis. De manière générale, les effets de la sélection sur les séquences nucléotidiques ne sont observés qu'après plusieurs centaines de milliers d'années d'évolution. À l'opposé, les effets observés sur les fréquences alléliques et génotypiques sont associés à des échelles de temps qui vont d'une dizaine de milliers d'années à quelques générations. Il est donc possible de partitionner les effets à court et long terme de la sélection en analysant les signatures moléculaires une à la fois.

La théorie neutre

Tout test statistique fonctionne à l'aide d'une hypothèse nulle. Ainsi, pour tester l'effet de la sélection naturelle, l'hypothèse qu'il n'y a aucun effet de la sélection doit être rejetée. Cette hypothèse nulle est représentée par la théorie neutre de Kimura (1983). Cette théorie prédit que les changements évolutifs ne sont causés que par la dérive d'allèles mutants sélectivement équivalents. Le polymorphisme d'une population ne dépend donc que de l'équilibre (ou le déséquilibre) mutation-dérive. Ce sont les prédictions de cette théorie qui permettent le développement des hypothèses nulles des différents tests statistiques. Le non-rejet de l'hypothèse de neutralité ne permet pas de dissocier le polymorphisme observé sur un locus de celui qui aurait été observé sur un locus neutre modelé par les processus démographiques ainsi que par la structure des populations. De l'autre côté, le rejet de cette hypothèse démontre que le polymorphisme observé est soit plus faible ou plus élevé que ce que prédit la théorie neutre et que cet écart ne peut pas être expliqué seulement par ces processus. Une certaine forme de sélection naturelle peut alors être proposée comme explication. Il s'agit alors de déterminer laquelle est en jeu.

Détection sur les fréquences génotypiques

La sélection balancée et l'infériorité des hétérozygotes influenceront à la hausse ou à la baisse la fréquence des hétérozygotes. Ces effets sur les fréquences génotypiques pourraient être détectés par un test permettant de tester les écarts par rapport à l'équilibre de Hardy-Weinberg. Les tests proposés par Rousset et Raymond (1995) qui permettent de tester le déficit ou l'excès d'hétérozygotes pourraient permettre de différentier laquelle des deux formes de sélection agit sur le locus. Il est cependant primordial de vérifier au préalable que la population est panmictique afin de s'assurer que les écarts observés sont spécifiques au locus étudié et non pas dus à une certaine forme de structure des populations (Rousset et Raymond 1995).

Détection sur les fréquences alléliques

La détection de la sélection à partir des fréquences alléliques peut se faire en déterminant la probabilité de H_E observée sous un équilibre mutation dérive. Le test de Ewens-Watterson (Watterson 1978) est sans doute le plus simple. Sa statistique correspond à la somme des carrés des fréquences alléliques. La probabilité de cette statistique est évaluée à l'aide de sa distribution de fréquences produite par simulations sous équilibre mutation-dérive. Ces simulations sont générées pour un échantillon de même taille et avec le même nombre d'allèles en utilisant un algorithme de coalescence suivant un modèle de mutations IAM. Un excès (sélection balancée et/ou diversificatrice) ou un déficit de diversité allélique (sélection directionnelle et/ou purificatrice) peuvent alors être détectés.

Des données sous formes de fréquences alléliques peuvent également permettre de tester certaines hypothèses de sélection à l'aide d'analyses de variance génétique telle que celle décrite dans la section sur la structure des populations. Dans ce cas, si la sélection est présente, une structure différente de celle observée sur des loci neutres pourrait être observée sur les loci sous sélection. L'intérêt de telles analyses, comme mentionné précédemment, est que différentes hypothèses de structure reliées entre autres à des facteurs écologiques peuvent être testées (Excoffier 2001).

Détection sur les séquences nucléotidiques

Le test de McDonald-Kreitman

Ce test est fondé sur une analyse conjointe d'une région codante chez deux espèces soeurs (McDonald et Kreitman 1991). Selon l'hypothèse de neutralité, le ratio de mutation synonymes/non-synonymes devrait être le même entre deux espèces et à l'intérieur d'une espèce car proportionnel au taux de mutation. En d'autres termes, un locus très polymorphique chez une espèce devrait également montrer une grande divergence entre espèces. Ce test prend donc la forme d'une table de contingence 2 x 2 testée à l'aide d'un rapport de vraisemblance (test de G) suivant une loi du χ^2 à 1 degré de liberté où les mutations sont classées en fonction de leur état synonyme/non-synonyme et en fonction de leur polymorphisme à l'intérieur et entre les espèces. Un rejet de l'hypothèse nulle signifie que le polymorphisme d'un locus à l'intérieur d'une espèce n'est pas représentatif de celui observé entre deux espèces, pour un des types de mutations. Des tests post-hoc peuvent alors être utilisés afin de déterminer quel type de mutations est responsable de l'écart à la neutralité. Ainsi, un excès de polymorphisme synonyme à l'intérieur d'une espèce serait associé à une sélection purificatrice, alors qu'un excès de polymorphisme non-synonyme serait associé à une sélection diversificatrice.

Le D de Tajima

La statistique D de Tajima (1989) permet de comparer deux estimations du même paramètre θ en faisant la différence entre le nombre moyen de différences entre paires d'allèles (θ_π) et le nombre de sites polymorphes (θ_S). Sous un équilibre mutation dérive, ces deux estimations mènent à la même estimation de θ, et D égale 0. Cependant, dans le cas d'une sélection diversificatrice, le grand nombre de différences entre chaque allèle mènera à une augmentation de θ_π par rapport à θ_S, ce qui rendra la valeur de D positive. À l'opposé, dans le cas d'une sélection purificatrice, la faible différence entre les allèles diminuera θ_π menant à une valeur négative de D. L'évaluation de l'intervalle de confiance de cette statistique à partir de simulations répondant à un équilibre mutation-dérive permet ensuite de déterminer sa signification.

Le F de Fu et Li

Le test de Fu et Li (1993) compare aussi deux estimations de θ. Par contre, plutôt que de prendre le nombre de sites polymorphes comme une estimation de ce paramètre, il utilise le nombre d'allèles (θ_k) et le compare à θ_π. En d'autres termes, ce test évalue la probabilité, sous un équilibre mutation-dérive, d'obtenir le nombre d'allèles observés dans un échantillon en fonction du nombre de mutations qui les différencient. De manière similaire au D, la relation entre les deux estimations permettra de distinguer les effets des sélections purificatrice et diversificatrice. Il s'agit ainsi d'une variante sur un même thème, mais dont l'utilisation combinée (D et F) permet d'explorer la diversité moléculaire sous différents angles.

Le test HKA

Ce test développé par Hudson et al. (1987) est fondé sur le même principe que le test de McDonald et Kreitman mais en analysant plusieurs locus simultanément. Ainsi, selon la théorie neutre, deux locus devraient montrer des niveaux de polymorphismes intra et inter-spécifiques équivalents. Contrairement au test précédent, le caractère synonyme ou non-synonyme des mutations n'est pas défini. Ainsi, il est possible, à l'aide d'une statistique du χ^2 de comparer le polymorphisme à l'intérieur d'une espèce et entre deux espèces d'un locus neutre avec celui d'un locus possiblement sous sélection. Tout écart provenant d'un polymorphisme réduit ou élevé à l'intérieur d'une espèce ou entre deux espèces sur un des deux locus mènera à un rejet de l'hypothèse nulle. Un tel rejet peut se traduire par une sélection directionnelle si un excès de polymorphisme entre espèce est observé ou par une sélection balancée s'il y a un excès de polymorphisme à l'intérieur d'une espèce.

Détection sur la structure entre les populations

Par leurs effets sur l'hétérozygotie à l'intérieur des populations, les formes de sélection influenceront également le niveau de différentiation entre les populations. La sélection directionnelle et/ou purificatrice augmentera le niveau de différentiation, alors que les différentes formes de sélection balancée le minimiseront. Incidemment, les loci montrant des niveaux de différentiation entre les populations très faibles ou très fortes sont souvent associés à des effets sélectifs. Il est donc possible de détecter ces loci sous sélection puisque ceux-ci apparaîtront comme des valeurs aberrantes par rapport à la relation sous neutralité entre l'hétérozygotie moyenne et le F_{ST} global. Beaumont et Nichols (1996) ont développé une méthode de simulation permettant de générer cette relation sous neutralité ainsi que son intervalle de confiance et ce pour différent modèle de mutations. La superposition des valeurs observées à cette relation permet alors de déterminer les loci qui sont plus (sélection directionnelle/purificatrice) ou moins (sélection balancée) différentiés que ce que prédit l'hypothèse de neutralité en fonction de leur hétérozygotie.

Objectifs

Objectif principal

Identifier et quantifier l'impact des mécanismes associés à l'histoire évolutive des populations de naseux des rapides (*Rhinichthys cataractae*) afin de mettre en lumière l'influence des pressions de sélection locales par une analyse conjointe de marqueurs nucléaires neutres, de gènes fonctionnels sous sélection, et de marqueurs mitochondriaux. D'un point de vue fondamental, ce type de connaissances mènerait à une meilleure compréhension des processus par lesquels les populations deviennent des entités évolutives uniques. Quant au point de vue pratique, cela permettrait d'entrevoir une gestion génétique des populations naturelles qui tiendrait compte autant de leur histoire évolutive que des pressions locales de sélection.

Objectifs secondaires et structure de la thèse

Chap. 1: Caractérisation de microsatellites spécifiques au naseux des rapides.

Les analyses proposées dans cette thèse pour identifier les mécanismes agissant sur les gènes fonctionnels nécessitent l'utilisation de marqueurs neutres comme les microsatellites. Or, aucun marqueur de ce type n'est disponible dans la littérature. L'objectif de ce chapitre est donc de développer ces marqueurs.

Chap. 2: Quantifier la puissance d'outils permettant la détection d'allèles nuls.

Les allèles nuls représentent un des artéfacts les plus courants des analyses génétiques utilisant les microsatellites. Il existe plusieurs outils statistiques permettant de détecter et de corriger les biais encourus par la présence de ce type d'allèle. Or, la précision et la puissance de ces outils n'ont jamais été établies. Nous avons profité de la présence d'un tel allèle sur un des loci développés au chapitre 1 pour évaluer ces outils expérimentalement et par simulations.

Chap.3: La colonisation post-glaciaire du naseux des rapides.

En étudiant deux marqueurs mitochondriaux à l'aide d'une NCA et d'une analyse canonique des correspondances, nous décrivons dans ce chapitre, l'origine, les étapes ainsi que les différents paramètres géographiques et historiques ayant influencé la colonisation postglaciaire de la péninsule québécoise par le naseux des rapides. Le résultat de cette étude permet ainsi d'avoir une idée précise de la structure des populations de naseux des rapides.

Chap.4: La méthode POST (POpulation STructure).

Les AMOVA ont deux principaux défauts qui limitent leur utilisation. 1) Elles nécessitent des connaissances *a priori* de l'organisation de la structure des populations. 2) Elles ne permettent pas de détecter des variations continues des fréquences alléliques telles que celles produites par les zones de contact. Nous proposons dans ce chapitre une méthode permettant de palier à ces deux défauts.

Chap.5: La phylogéographie du naseux des rapides à l'aide de marqueurs nucléaires

Dans ce chapitre, les marqueurs microsatellites développés au chapitre 1 sont utilisés afin de raffiner les connaissances des processus historiques ayant façonné la diversité génétique du naseux des rapides sur la Péninsule Québécoise.

Chap.6: Identifier les sources du polymorphisme du complex majeur d'histocompatibilité (CMH).

Utilisant les marqueurs et les résultats obtenus dans les chapitres précédents, nous tentons de décortiquer les effets à court et long terme de la sélection de naturelle sur le CMH de ceux associés aux processus neutres chez des populations de naseux des rapides.

Espèce modèle

Le naseux des rapides fait partie de la famille des cyprinidés. Sa distribution géographique couvre la presque totalité des habitats tempérés de l'Amérique du Nord, parcourant d'est en ouest le nord des États-Unis et le sud du Canada, du 28è au 69è parallèle (Scott et Crossman 1973; Figure 10a). Le naseux des rapides habite la plupart des cours d'eau du Québec, à l'exception de ceux situés sur la péninsule gaspésienne ainsi qu'à l'est du Saguenay (Bernatchez and Giroux 2000).

D'une longueur maximale de 17,8 cm selon (Hubbs et Lagler 1949), il a en moyenne une longueur à l'âge adulte de 7,5 cm (Bernatchez et Giroux 1996). Le naseux des rapides possède une forte capacité natatoire et une faible tolérance à l'eau chaude et mal oxygénée (Scott et Crossman 1973). On le retrouve dans les eaux turbulentes, bien qu'on l'ait aussi observé parfois dans la zone littorale de certains lacs (Scott et Crossman 1973, Page et Burr 1991). Il s'agit d'un poisson très généraliste dont la diète variera en fonction de la disponibilité des proies (Thompson et al. 2001) bien qu'elle se compose normalement d'éphémères, de mouches noires et de moucherons à l'état larvaire ou adulte (Scott et Crossman 1973). La ponte se déroule de mai à août mais est maximale en juin et juillet (McPhail et Lindsey 1970). Les femelles pondent de 200 à 1200 œufs préférentiellement sur des lits de gravier ou rocailleux (Scott et Crossman 1973). Selon McPhail et Lindsey (1970) un territoire gardé par un des deux parents est établi autour du nid où les œufs sont incubés de 7 à 10 jours à une température optimale de 16 °C. Les larves benthiques (Balon 1990) produisent des juvéniles qui seront retrouvés en banc dans les habitats calmes situés près des berges pour une période d'environ 4 mois pour ensuite passer aux habitats typiques des eaux turbulentes au centre des ruisseaux et rivières (Scott et Crossman 1973).

Sa très grande distribution géographique, ainsi que la multitude d'habitats différents dans lesquels on le retrouve, rend le naseux des rapides propice à l'étude de l'adaptation locale des populations. De plus, sa distribution peu affectée par les activités humaines permet de croire que le patron de polymorphisme génétique contemporain reflète les effets naturels des processus évolutifs depuis la colonisation postglaciaire de l'espèce.

CHAPITRE I

Characterization of microsatellite loci in Longnose dace (*Rhinichthys cataractae*) and interspecific amplification in five other Leuciscinae species

Publié dans Molecular Ecology Notes (2006) **6**:69-71

Résumé

Le naseux des rapides (*Rhinichthys cataractae*) représente un modèle pertinent dans le cadre de problématiques environnementales, écologiques et évolutives. Dans le cadre de cette étude, onze microsatellites spécifiques au naseux ont été caractérisés. Huit de ces marqueurs étaient hautement polymorphiques. Entre quatre et dix de ces marqueurs ont également été amplifiés avec succès chez cinq espèces apparentées du naseux des rapides. Ces résultats laissent croire que ces marqueurs représentent des outils utiles pour les analyses de génétique des populations du naseux des rapides ainsi que pour d'autres espèces appartenant à la même sous-famille (Leuciscinae).

Abstract

The Longnose dace (*Rhinichthys cataractae*) appears as a relevant model to address environmental and ecological issues in an evolutionary perspective. Eleven microsatellite markers were characterized for this species. Eight of these loci were highly polymorphic for populations of this species. Between four to ten loci were also successfully amplified in five closely related species. These markers are believed to be valuable tools for genetic analysis of populations of Longnose dace and other Leuciscinae species.

The Longnose dace (*Rhinichthys cataractae*) belongs to the Cyprinidae, the most diversified families of freshwater fishes (Simons et al. 2003). This species is widely distributed over North America and can be found in habitats of different characteristics and history (Scott and Crossman 1973). Those characteristics make this species a relevant model to address environmental and ecological issues in an evolutionary perspective. Microsatellites or short tandem repeats (STR) become a considerable source of highly polymorphic markers for population genetic-based analysis. However, no such markers are available for the Longnose dace. The objective of this study is to characterize microsatellite loci for the *R. cataractae* and evaluate their variability in natural populations. The usefulness of those markers was also evaluated in five other species of the Leuciscinae sub-family.

The isolation of microsatellite loci was performed on a single *R. cataractae* individual. A partial genomic library was obtained by digestion with the *MspI* endonuclease. According to Refseth et al. (1997), the fragments were joined to a cohesive adaptor, amplified by PCR, and inserted into T-vectors (Promega). Plasmids were transformed into E. *coli* (JM109 strain) and plated on ampiciline-selective media. Approximately 600 clones were screened by PCR combining plasmid primers and $(GATA)_5$, $(CATA)_5$, $(CT)_{10}$, or $(GT)_{10}$ probes. Positive clones were sequenced on both strands using a CEQ 2000XL DNA Analysis System (Beckman Coulter). Specific primers were designed for eleven positive clones that contained a microsatellite with more than eight uninterrupted repeats and for which flanking regions were present. PCR conditions were optimized for each locus. A 12.5 µl reaction volume contained 1.5 mM of $MgCl_2$ (4.0 mM for Rhca31), 2.5 nM of each dNTP, 0.2 unit of *Taq* polymerase, 1.25 µl of 10x *Taq* polymerase buffer (Invitrogen), 10 pmol of each primer and approximately 10 ng of DNA. Rhca16, Rhca20, Rhca23, Rhca34, Rhca46 and Rhca52 also necessitate the addition of 12 mM of BSA. The PCR program (Touchgene gradient thermal cycler; Techne) consisted of an initial denaturating step of 30s at 92°, followed by 45 cycles of the following profile: 10s of denaturation at 92°, 15s at annealing temperature (Table 1) and 5s at 68° (except for Rhca15b, Rhca23, Rhca24 and Rhca52 for which 15s of extension is recommended). The PCR reactions ended with a final extension of 2m at 68.

Table 1: Caracteristics of the 11 Longnose dace microsatellite loci. For each locus, forward and reverse primers, sequence of the microsatellite region, the reference size and the annealing temperature (T_A) are presented. Genetic diversity over the five Longnose dace populations and averaged by population (in parenthesis) are given in terms of number of alleles (A), size range of alleles in base pairs (bp) and Nei's gene diversity (H_E).

Locus	Primer sequences (5'- 3')	Repeats in original clone	Predicted	T_A	A	Size	H_E
(Genebank #)			size (bp)			(bp)	
Rhca 7	GTCCACCTCATACAAACTTCC	$(CA)_{10}$	113	50	1	113	-
(DQ106911)	ATGAGGCAACCACTGGAGC						
Rhca 9	TAGAACAATGGACGGATGG	$(GATA)_{11}$	188	52	1	188	-
(DQ106912)	CGGTTGCACTACAAATTATCC						
Rhca 15b	CTCACAGACTACCTGCCC		260	50	16	125-341	0.87
(DQ106913)	CAGAGGTCAAACAGTAGTAGG	$(CTAT)_{12}(CTATCATAT)_8(CTAT)_3$			(4.4)	(14.2)	(0.55)
Rhca 16	GAGAACGAGTGGACATCC	$(GA)_{12}$ GCGT$(GT)_{12}$	123	48	10	110-128	0.78
(DQ106914)	AGTGAGTGGTTGAGTAGG				(3.4)	(4.6)	(0.38)
Rhca 20	CTACATCTGCAAGAAAGGC	$(GA)_{17}$	113	50	12	93-121	0.85
(DQ106915)	CAGTGAGGTATAAAGCAAGG				(4)	(10.2)	(0.59)
Rhca 23	TTCGTCCATATCTAGAGG	$(CA)_7(CT)_4$ CACT	244	52	8	220-262	0.65
(DQ106916)	TCATGAATGCAGTACTGG	$(CA)_8(CT)_3$			(2.8)	(7.2)	(0.31)
Rhca 24	GTGGTGTTAGCAGAAACCCG	$(GA)_{27}$	300	54	28	296-424	0.94
(DQ106917)	CTGCTGTTTAATATGTCAC				(7)	(41.6)	(0.76)
Rhca 31	GTTACACCCACTTATTCG	$(GA)_{15}$	172	52	7	151-185	0.71
(DQ106918)	GTTACCGGCCAGAATGTC				(2.8)	(6.4)	(0.43)
Rhca 34	TCCTGGACGTTATCGTCC	$(GT)_{19}$	142	48	9	122-148	0.83
(DQ106919)	GTGATGAGGACCCAGAGGC				(3.8)	(6.4)	(0.59)
Rhca 46	GTGCCTGACTTAATTAGG	$(CA)_9$	113	48	1	113	-
(DQ106920)	CACGTGAAATTTAAGCCC						
Rhca 52	TTAATGCTGAATCCTTTGGG	$(CT)_9$	280	56	7	278-306	0.75
(DQ106921)	CAATGAGACAGATTCGATTC				(1.8)	(4.4)	(0.10)

A screening was performed on a total of 55 *R. cataractae* individuals from five populations located within the St. Lawrence River (Canada, QC) and the Atlantic (USA, CT) watersheds. Eight of the 11 markers were polymorphic (Table 1). The results revealed that the number of alleles varied from 7 to 28 and gene diversity (Nei 1987) from 0.65 to 0.94 (Table 1). No significant linkage disequilibrium was observed between pair of loci. No deviation from Hardy-Weinberg equilibrium was significant following Bonferroni correction, except for Rhca23 for which an excess of homozygotes was observed in a single population ($p<0.01$). Amplification problem was detected (MICRO-CHECKER v.2.2.1; van Oosterhout et al. 2004), but appears to be restricted to this population. Alleles of each locus were measured in steps of two nucleotides except for Rhca15b for which a combination of steps of four and nine nucleotides was observed (Table 1). The more variable loci, Rhca15b and Rhca24, showed a size variance of 216 and 128 bp respectively (Table 1). The sequence of the smallest and the largest alleles of both loci revealed that size differences are exclusively due to variation in the number of repeats (data not shown). Those results indicate that the polymorphic markers characterized in this work can be useful for various applications, from moderately polymorphic loci suitable for *R. cataractae* population studies to highly polymorphic loci for analysis on individuals.

The 11 microsatellites were then tested on four individuals from each of those following species: Blacknose dace (*R. atratulus*), Cutlips minnow (*Exoglossum mixillingua*), Pearl dace (*Margariscus margarita*), Northern redbelly dace (*Phoxinus eos*) and Fathead minnow (*Pimephales promelas*). Two populations were used for *R. atratulus* and *P. eos* while a single population was tested for the other species. The homology of cross-species amplifications was confirmed when one of these three conditions was met. 1- Both stutter shape and size of a cross-species amplification product were similar to those observed on source species. The stutter patterns are expected to be relatively constant over the range of the allele sizes, which is usually true for microsatellites with approximately the same number of repeats (Miller and Yuan 1997). 2- Locus extracted from the gel is successfully amplified with one primer of the flanking region and the other corresponding to the repeated motif of the loci. 3- Sequencing of the locus revealed flanking regions homologous to the one of source species.

Results revealed that 35 of the 55 cross-species tests provided amplification of homologous locus (Table 2). Rhca20, Rhca34 and Rhca46 were amplified in all species while all other loci were amplified in at least one species. Ten of the 11 loci were amplified in *R. atratulus*. Interestingly, two of the three loci monomorphic in the source species were variable in *R. atratulus*. The number of successful amplifications in other species ranged between four (*P. promelas*) to nine (*M. margarita*). The amplification of these markers across species from distinct genera makes them promising for population genetic studies in several Leuciscinae species.

Table 2: Results of the cross-species amplifications of *Rhinichthys cataractae* loci in five Leuciscinae species. For each species, the size range in bp and the number of alleles (A) are given except where no amplification of the homologous locus was detected

	Rhinichthys atratulus		*Margariscus margarita*		*Exoglossum maxillingua*		*Phoxinus eos*		*Pimephales promelas*	
	size	A	size	A	Size	A	size	A	size	A
Rhca 7	115-125	2	121	1	113-119	2	115	1	-	-
Rhca 9	180-200	4	192	1	-	-	180-184	2	-	-
Rhca 15b	204-226	2	246-254	3	-	-	-	-	-	-
Rhca 16	114	1	-	-	-	-	-	-	-	-
Rhca 20	108-122	3	142	1	122	1	104-132	2	122-136	5
Rhca 23	237-249	2	219-235	2	-	-	-	-	-	-
Rhca 24	-	-	442	1	384-420	4	-	-	-	-
Rhca 31	161	1	-	-	-	-	172	1	157	1
Rhca 34	132-148	3	126-148	4	132	1	175	1	126-134	3
Rhca 46	113	1	109-113	2	113	1	113	1	113	1
Rhca 52	304	1	302	1	-	-	316	1	-	-

CHAPITRE II

Assessment of power and accuracy of methods developed for detection and frequency-estimation of null alleles

Publié en ligne dans Genetica le 3 janvier 2008

Résumé

Les allèles nuls représentent un artefact commun des analyses génétiques utilisant des marqueurs microsatellites. Cependant, des méthodes permettant de les détecter rapidement et d'en estimer la fréquence à l'intérieur d'un échantillon ont été développées. L'objectif de ce chapitre est de quantifier la puissance ainsi que la précision de ces outils statistiques en utilisant des jeux de données simulées et réelles. Nos résultats ont démontré qu'aucun des tests analysés ne donne des résultats complètement satisfaisants. Cependant, il est possible d'augmenter la confiance des conclusions en combinant l'utilisation de plus d'un test. Une comparaison entre différents estimateurs des fréquences d'allèles nuls a clairement démontré que l'indice Brookfield2, qui tient compte des fréquences d'individus non-amplifiés, donne des estimations beaucoup plus précises. L'ensemble des analyses effectuées confirme par contre l'avantage de ce genre d'outils sur le développement d'amorces alternatives puisque les allèles nuls demeurent parfois indétectables malgré l'utilisation de telles procédures. Suivant ces résultats, nous proposons un certain nombre de recommandations permettant de détecter et de corriger avec confiance les jeux de données contenant des allèles nuls.

Abstract

Null alleles represent a common artefact of microsatellite-based analyses. Rapid methods for their detection and frequency estimation have been proposed to replace the existing time-consuming laboratory methods. The objective of this paper is to assess the power and accuracy of these statistical tools using both simulated and real datasets. Our results revealed that none of the tests developed to detect null alleles are perfect. However, combining tests allows the detection of null alleles with high confidence. Comparison of the estimators of null allele frequency indicated that those that account for unamplified individuals, such as the Brookfield2 estimator, are more accurate than those that do not. Altogether, the use of statistical tools appeared more appropriate than testing with alternative primers as null alleles often remain undetected following this laborious work. Based on these results, we propose recommendations to detect and correct datasets with null alleles.

Introduction

Microsatellite markers are short tandem repeats of 1–6 nucleotides (Wyman and White 1980). Abundant in the nuclear genome of most eukaryotes (Jarne and Lagoda 1996) and highly polymorphic (Weber and Wong 1993), these neutral markers are now extensively used in population genetics (Avise 2004). Despite these advantageous characteristics, some problems arise with their use. For example, null alleles represent a recurrent artefact of microsatellite based analyses. A null allele can be defined as any allele that consistently fails to amplify with the polymerase chain reaction (PCR) using a given pair of primers because of local mutation(s) on the DNA template (Dakin and Avise 2004). As a result, the microsatellite in question will exhibit inexact allele frequencies as well as non-Mendellian inheritance characterized by an excess of homozygote. The presence of such alleles may thus introduce serious complications in numerous genetic studies such as parentage assignment (Callen et al. 1993) and population diversity analyses (Pemberton et al. 1995; Bowling et al. 1997; Avise 2004; Chapuis and Estoup 2007).

It has been reported that null alleles are common at microsatellite loci and across taxa (Callen et al. 1993; Ardren et al. 1999), showing the importance of considering this amplification problem. To manage biases due to null alleles, several alternative statistical procedures have been developed to indirectly detect their presence by analyzing the visible allele scores of a locus within a population. While some methods require repeated genotyping of each individual (Valière 2002; Miller et al. 2002), procedures such as the test implemented in MICRO-CHECKER v.2.2.3 (MCT; Shipley 2003; van Oosterhout et al. 2004) and the U test found in GENEPOP v.3.4 (UT; Raymond and Rousset 1995), are appealing because they do not require excess additional laboratory work. MCT and UT are both based on the assumption that a null allele generates homozygote excess in a population dataset. However, they differ in the way that Hardy–Weinberg equilibrium (HWE) is tested. The UT assesses the presence of a null allele when it detects that the total number of homozygotes exceeds the number expected under random mating. On the other hand, the MCT compares the observed homozygote frequencies to the ones expected under random mating in each allele size class independently and the presence of a null allele is confirmed if significant homozygote excesses are evenly observed across the allelic size distribution.

Correcting datasets that include null alleles often requires extensive lab work. For instance, the use of alternative primers (Holm et al. 2001) has been suggested, but it does not guarantee identification of all null alleles (Walter and Epperson 2004). Considering the time cost and the dubious gain of information associated with this strategy, it is generally recommended that efforts be concentrated on the use of a large number of highly polymorphic loci (Estoup et al. 2002) and that loci showing null alleles be discarded (De Sousa et al. 2005). However, the number of known microsatellite loci remains limited for most of the organisms not currently used as genetic models. In this context, estimators have been proposed to quantify null allele frequency by analyzing the apparent deficit of heterozygotes caused by those alleles (Chakraborty et al. 1992; Brookfield 1996).

Considering the potential errors introduced by null alleles, an evaluation of the power and accuracy of methods for the detection of null alleles and for correction of the biases that they introduce is relevant. While the differential capacity of MCT and UT to detect null was never established, the accuracy of several null allele frequency estimators was recently performed by Chapuis and Estoup (2007). However, while the effects of sample size and genetic diversity were kept fixed in their study, their effect on the reliability of these procedures and estimators remain largely unknown. In this study, we quantified the effect of these parameters on the power of MCT and UT, and on the accuracy of three largely used estimators using simulated datasets. These tools were finally used to evaluate the capacity of an alternative primer procedure to recover null alleles within a real dataset of 29 populations.

Materials and methods

Sampling Simulated dataset

Frequency distributions that differed in their total number of alleles (k; 5, 10 or 15), their expected heterozygosity under HWE estimated with both null and visible alleles (H_{E-TOT}; 0.2, 0.5, 0.7 or 0.8) and with visible alleles only (H_{E-VIS}; 0.1–0.8) were simulated following a simple procedure. For given levels of k and H_{E-TOT}, an allelic frequency distribution in accordance with mutation-drift equilibrium under the stepwise mutation model expectations was constructed (Fig. 5a). A given frequency of null alleles was then distributed across all allelic size classes, keeping H_{E-TOT} constant but changing the H_{E-VIS} (Fig. 5b,c). A minimum frequency of 0.001 (one visible allele for 1,000 individuals) was set for each size class to keep constant the number of visible alleles. A total of 34 allelic distributions representing the widest range H_{E-VIS}, H_{E-TOT} and k throughout the complete range of F_{NULL} were selected for the analyses (Table 3).

Populations of 1,000 individuals were then constructed for each distribution following the HWE predictions. To mimic electrophoresis data, null homozygotes were then considered as unamplified samples while null heterozygotes were considered as homozygotes for the associated visible allele. One hundred random samplings of 20, 50 and 100 individuals were performed within each population. For each sample, $\hat{H}_{E-VIS}$, $\hat{H}_O$ (the frequency of individuals having two visible alleles) and $\hat{}N_{VIS}$ (the number of individuals having at least one visible allele) were computed. The hat symbol was added to distinguish estimations performed on samples from the parameters of the allelic distributions from which they were taken. Samples where $\hat{}N_{VIS} < 4$ individuals or $\hat{H}_{E-VIS} = 0$ were removed from the analyses because the UT and the MCT cannot be performed in those conditions.

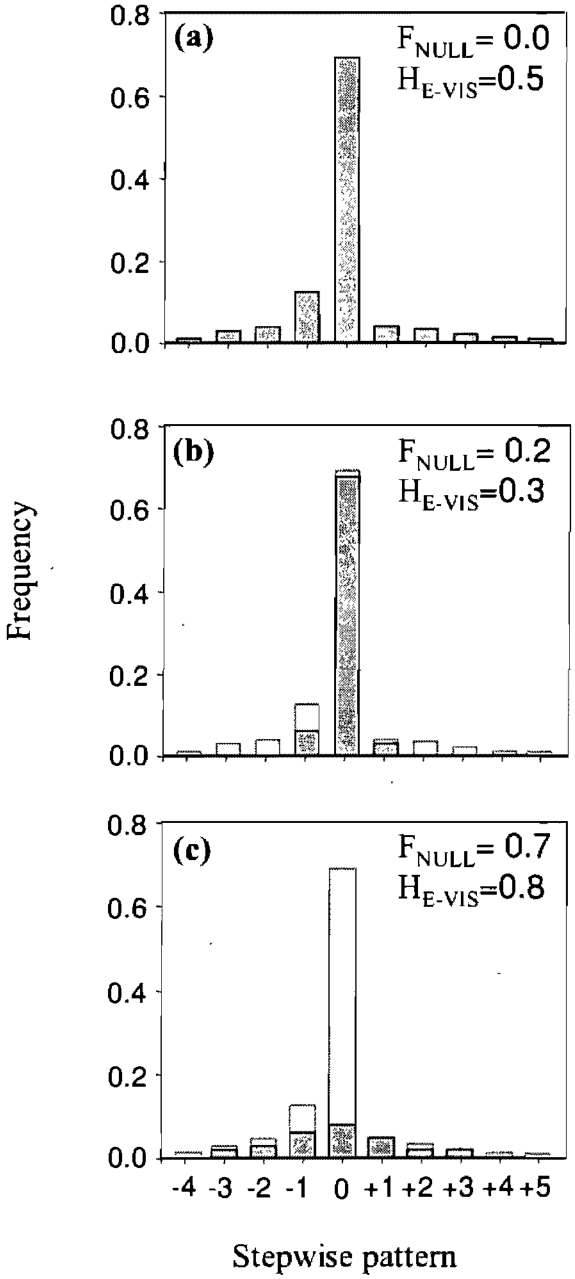


Figure 5: Examples of simulated frequency distributions. All distributions have the same $H_{E\text{-}TOT}$ (0.5) and k (10) but varied in function of their F_{NULL} and $H_{E\text{-}VIS}$. Grey and white bars represent visible and null alleles respectively.

Table 3: Null allele frequency (F_{NULL}), k (number of alleles), heterozygosity computed with both null and visible alleles (H_{E-TOT}) and heterozygosity computed only with visible alleles (H_{E-VIS}) of each simulated population. The power of each analysis (computed as the proportion of 100 samples for which an analysis detected a null allele) for three sample sizes is also presented.

F_{NULL}	k	$H_{E\text{-}TOT}$	$H_{E\text{-}VIS}$	Power								
				UT			MCT			TBR2		
				N=20	N=50	N=100	N=20	N=50	N=100	N=20	N=50	N=100
0.0	5	0.2	0.2	0.02	0.02	0.01	0.01	0.03	0.00	0.00	0.00	0.00
0.0	5	0.7	0.7	0.06	0.04	0.04	0.02	0.04	0.01	0.01	0.00	0.02
0.0	10	0.5	0.5	0.03	0.06	0.03	0.00	0.01	0.00	0.00	0.00	0.00
0.0	15	0.8	0.8	0.06	0.02	0.01	0.02	0.00	0.00	0.01	0.00	0.01
0.1	5	0.3	0.2	0.15	0.41	0.63	0.07	0.26	0.44	0.03	0.15	0.44
0.1	5	0.5	0.5	0.17	0.59	0.70	0.09	0.41	0.69	0.05	0.27	0.76
0.1	10	0.5	0.4	0.31	0.52	0.79	0.14	0.28	0.56	0.30	0.33	0.60
0.1	15	0.8	0.8	0.53	0.90	0.99	0.27	0.70	0.95	0.15	0.57	0.90
0.2	5	0.5	0.2	0.27	0.63	0.87	0.15	0.59	0.79	0.07	0.60	0.95
0.2	10	0.5	0.3	0.34	0.80	0.98	0.17	0.69	0.97	0.65	0.70	1.00
0.2	5	0.8	0.7	0.64	1.00	1.00	0.49	0.93	1.00	0.62	0.97	1.00
0.2	15	0.8	0.8	0.85	1.00	1.00	0.74	0.99	1.00	0.70	1.00	1.00
0.3	5	0.5	0.2	0.37	0.73	0.99	0.26	0.75	0.97	0.31	0.95	1.00
0.3	10	0.5	0.2	0.56	0.68	0.95	0.26	0.57	0.90	0.71	0.85	1.00
0.3	5	0.6	0.5	0.62	0.96	1.00	0.52	0.98	1.00	0.71	1.00	1.00
0.3	15	0.8	0.8	0.99	1.00	1.00	0.94	1.00	1.00	1.00	1.00	1.00
0.4	10	0.5	0.1	0.81	0.79	0.90	0.09	0.54	0.81	0.88	0.98	1.00
0.4	5	0.7	0.7	0.92	1.00	1.00	0.89	1.00	1.00	0.99	1.00	1.00
0.4	15	0.8	0.9	0.98	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00
0.5	5	0.6	0.2	0.71	0.90	0.97	0.35	0.81	0.97	0.90	1.00	1.00
0.5	5	0.6	0.5	0.85	0.99	1.00	0.85	0.99	1.00	0.99	1.00	1.00
0.5	5	0.7	0.7	0.96	1.00	1.00	0.96	1.00	1.00	1.00	1.00	1.00
0.6	5	0.5	0.2	0.69	0.94	0.97	0.34	0.85	0.99	1.00	1.00	1.00
0.6	5	0.6	0.5	0.93	1.00	1.00	0.81	0.99	1.00	1.00	1.00	1.00
0.6	5	0.6	0.7	0.98	1.00	1.00	0.94	1.00	1.00	1.00	1.00	1.00
0.7	5	0.4	0.2	0.50	0.92	0.94	0.17	0.71	0.94	1.00	1.00	1.00
0.7	5	0.5	0.7	0.97	1.00	1.00	0.89	1.00	1.00	1.00	1.00	1.00
0.7	10	0.5	0.8	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00
0.8	5	0.3	0.2	0.32	0.90	0.79	0.14	0.58	0.79	1.00	1.00	1.00
0.8	5	0.3	0.5	0.89	0.98	0.99	0.54	0.95	1.00	1.00	1.00	1.00
0.8	5	0.3	0.7	0.91	1.00	1.00	0.88	1.00	1.00	1.00	1.00	1.00
0.9	5	0.2	0.2	0.05	0.60	0.55	0.02	0.23	0.55	1.00	1.00	1.00
0.9	5	0.2	0.5	0.75	0.79	0.94	0.16	0.66	0.95	1.00	1.00	1.00
0.9	5	0.2	0.7	0.38	0.94	1.00	0.56	0.96	1.00	1.00	1.00	1.00

Model species and molecular analyses

A real dataset was obtained from 578 Longnose dace (*Rhinichthys cataractae*) sampled in 29 rivers of the Quebec peninsula in Canada. Sample sizes averaged 20 individuals per population and ranged between 9 and 24 individuals (Table 4). A piece of each individual's caudal fin was removed and stored in 95% ethanol for molecular analyses. Genetic analyses were performed on the nuclear genome of each individual using seven unlinked polymorphic microsatellites designed for Longnose dace (Girard and Angers 2006a). Six loci displayed Mendelian inheritance while the Rhca23 locus exhibited amplification problems consistent with the segregation of null alleles. The PCR were performed using a Touchgene gradient thermal cycler (Techne) according to the conditions described in Girard and Angers (2006a). Polymorphism of the microsatellite markers was evaluated with 6% denaturing urea-polyacrylamide gel electrophoreses.

We attempted to eliminate the effects of technical problems (e.g., poor quality DNA, PCR artefacts) or population processes (e.g., non-random mating) from the dataset that could mimic the effects of null alleles. First, populations that showed a departure from HWE (over all loci except Rhca23) were discarded because the tests and the estimators assume such equilibrium. Second, individuals that failed to amplify at four or more loci were also discarded as DNA extraction was likely of poor quality. Third, a second round of PCR with Rhca23 primers was performed to confirm that the amplification failure of individuals during the first PCR was not the result of a technical problem. After this, the Rhca23 locus was re-amplified using alternative primers (5'-CTCAGGAACATTTCATGC-3' and 5'-GGCCTACATTCACAGTTG-3') for all individuals possibly carrying a null allele: (1) those that failed to amplify in the third step of the previous test or (2) those that were scored as homozygotes using the initial primers. The following reaction conditions were used: a 12.5 µl reaction containing 1.5 mmol l^{-1} of MgCl2, 2.5 nmol l^{-1} of each dNTP, 0.2 units of Taq polymerase, 1.25 µl of 10X Taq polymerase buffer (Invitrogen Corp., Burlington, Ontario) and approximately 20 ng of DNA. Amplification conditions included an initial denaturation of 30 s at 92°C followed by 45 cycles of 10 s at 92°C, 15 s at 50°C and 30 s at 68°C, and a final extension of 120 s at 68°C.

Statistical analyses

MCT and UT

The MCT was constructed to discriminate between the presence of null alleles and the presence of a large allele dropout or genotyping errors due to stuttering. Note that these artefacts were not analyzed in the present study. To fulfill this task, MCT randomly constructs genotypes by sampling the visible alleles within a population. In this study, this procedure was repeated 999 times per sample, allowing the evaluation of the average and the 95% confidence interval (CI) for the expected homozygote frequencies for each size class. The presence of a null allele was assessed using the decision rules proposed by van Oosterhout et al. (2004). Accordingly, if at least (1) 50% of the observed homozygote frequencies are above the upper limit of the corresponding CI or (2) 90% of the observed homozygote frequencies are above the expected average, the presence of a null allele is confirmed. UT is a more general statistical analysis used to determine homozygote excess. Tested with Markov chains, this statistic uses homozygote frequencies to test HWE. Even though UT is more powerful than the bilateral probability test (Rousset and Raymond 1995), it is useless in identifying the causes of homozygote excess. Type I (α) and type II (β) errors were evaluated for both tests with the samples performed previously on simulated frequency distribution with MICRO-CHECKER and GENEPOP.

The capacity of detecting a null allele when present (power: $1 - \beta$) of the MCT and the UT was modeled afterward using a stepwise multivariate logistic regression in which the above parameters were regressed to the conclusions of each method (0 or 1 for failure or success of null allele detection respectively) for all samples containing at least one copy of the null allele.

F_{NULL} estimators

The frequency of null alleles (r) was evaluated with three estimators: r_{CHAK} (Chakraborty 1992), r_{BROOK1} and r_{BROOK2} (Brookfield 1996). These estimators are based on the assumption that the apparent heterozygote deficit is due to null alleles, but they differ in the way they deal with samples whose amplification failed. The r_{CHAK} and r_{BROOK1} estimators discard these samples while the r_{BROOK2} estimator considers that a certain proportion of unamplified samples corresponds to null homozygotes and thus takes them into account. F_{NULL} was estimated for all samples of a given simulated population and the mean and 95% confidence interval were computed. The width of the confidence interval and whether or not it included the real value were then used as indications of the accuracy of each estimator.

Results

MCT and UT

Both tests revealed a low type I error. They rejected the null hypothesis in a proportion below (or close to) the 0.05 threshold in all populations without a null allele (Table 3). Consequently, the rejection of the null hypothesis by either UT or MCT can be confidently interpreted to be due to the presence of a null allele (or due to another problem, such as allele dropout or a scoring error that could give a similar signal in a real dataset; Shaw et al. 1999). On the other hand, the acceptance of the null hypothesis should be interpreted cautiously as the power of both tests to detect a null allele was quite low when F_{NULL} was below 0.4 or above 0.6 (Fig. 6, Table 3). Furthermore, for a given F_{NULL}, the power appears to be positively correlated with $H_{E\text{-}TOT}$ (partial $r_{UT} = 0.43$ and $r_{MCT} = 0.50$; both $p < 0.0001$), $H_{E\text{-}VIS}$ (partial $r_{UT} = 0.36$ and $r_{MCT} = 0.36$; both $p < 0.005$) and N (partial

r_{UT} = 0.46 and r_{MCT} = 0.53; both $p < 0.0001$). However, no significant partial correlation was observed between the tests' power and k (r_{UT} = 0.14; p = 0.19 and r_{MCT} = 0.09; p = 0.41).

The positive relation between N_{VIS} and $H_{E\text{-}VIS}$ and the tests' power was confirmed by the stepwise logistic regression while a significant negative correlation was observed between H_O and power of both tests. These regression models are as follows:

$$\text{Power} = 1/(1 + e^{-\lambda})\ (4)$$

where

$$\lambda_{UT} = -2.62 + 0.05 \cdot N_{VIS} + 16.17 \cdot H_{E\text{-}VIS} - 17.71 \cdot H_O$$

$$\lambda_{MCT} = -3.65 + 0.06 \cdot N_{VIS} + 16.48 \cdot H_{E\text{-}VIS} - 18.61 \cdot H_O$$

All the parameters included in these models are significant at the 0.001 threshold. The predictions of these models clearly demonstrate the increased power when $\hat{H}_{E\text{-}VIS}$ and ^N_{VIS} are high as well as when the difference between $\hat{H}_{E\text{-}VIS}$ and $\hat{H}_O$ increases (Fig. 7). It is noteworthy that UT revealed similar or better performances than MCT.

F_{NULL} estimators

The accuracy of the estimators was variable (Fig. 8). The r_{CHAK} and r_{BROOK1} estimators displayed important weaknesses regardless of sample size. r_{BROOK1} consistently underestimated null allele frequencies (Fig. 8a–c). In contrast, estimations with r_{CHAK} were accurate, but their confidence intervals were very wide (Fig. 8d–f). Higher sample size appears to increase the accuracy of this estimator, but its high variance makes it useless in most conditions. The results obtained with r_{BROOK2} are far better since null allele frequen-

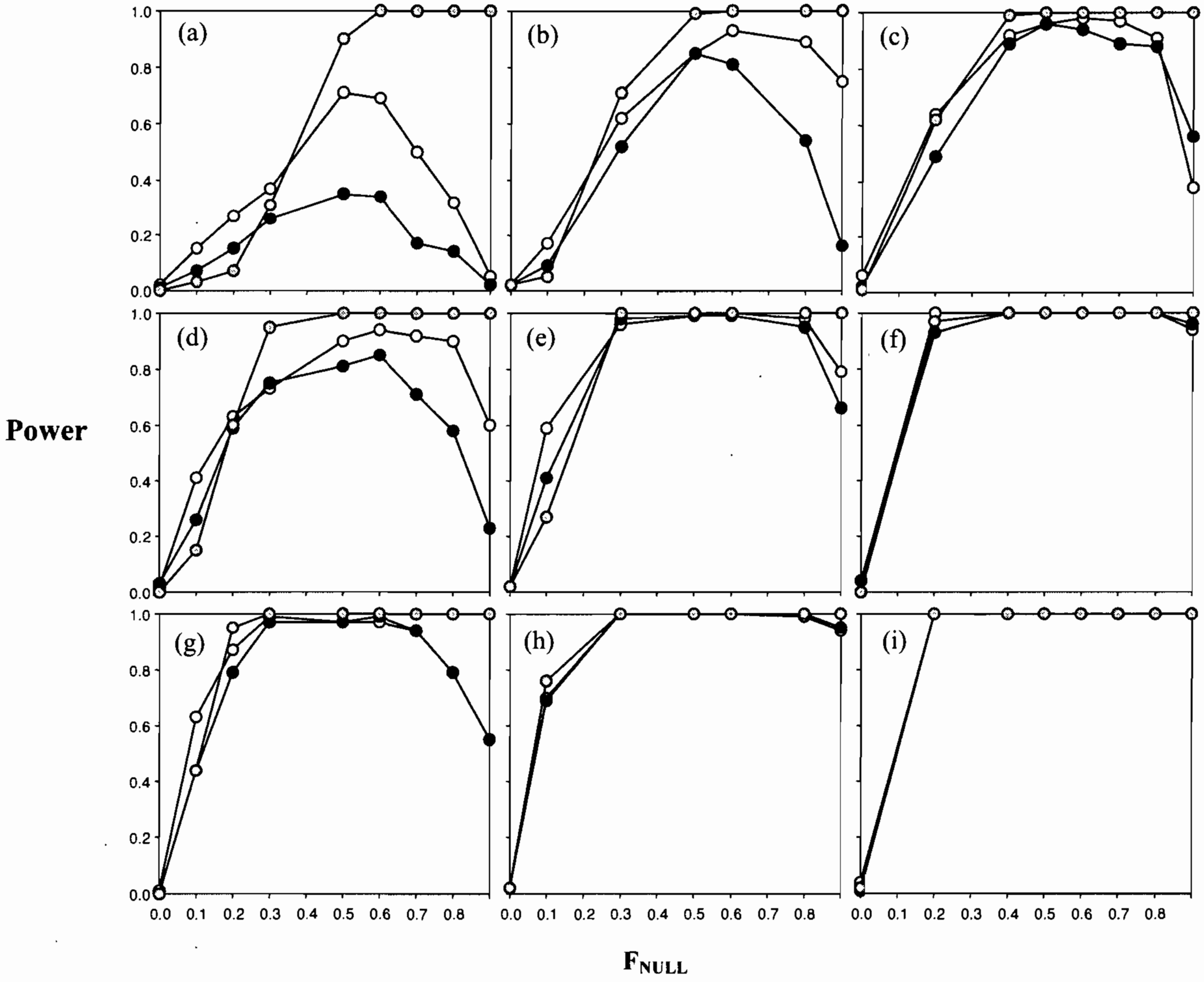


Figure 6: Power of the UT (white), the MCT (black) and the TBR2 (grey) as a function of F_{NULL} when $H_{E\text{-}VIS}$ is 0.2 (a, d, g), 0.5 (b, e, h) or 0.7 (c, f, i), and for N of 20 (a, b, c), 50 (c, d, e) or 100 (g, h, i) individuals. Results are those obtained with populations with k = 5 alleles.

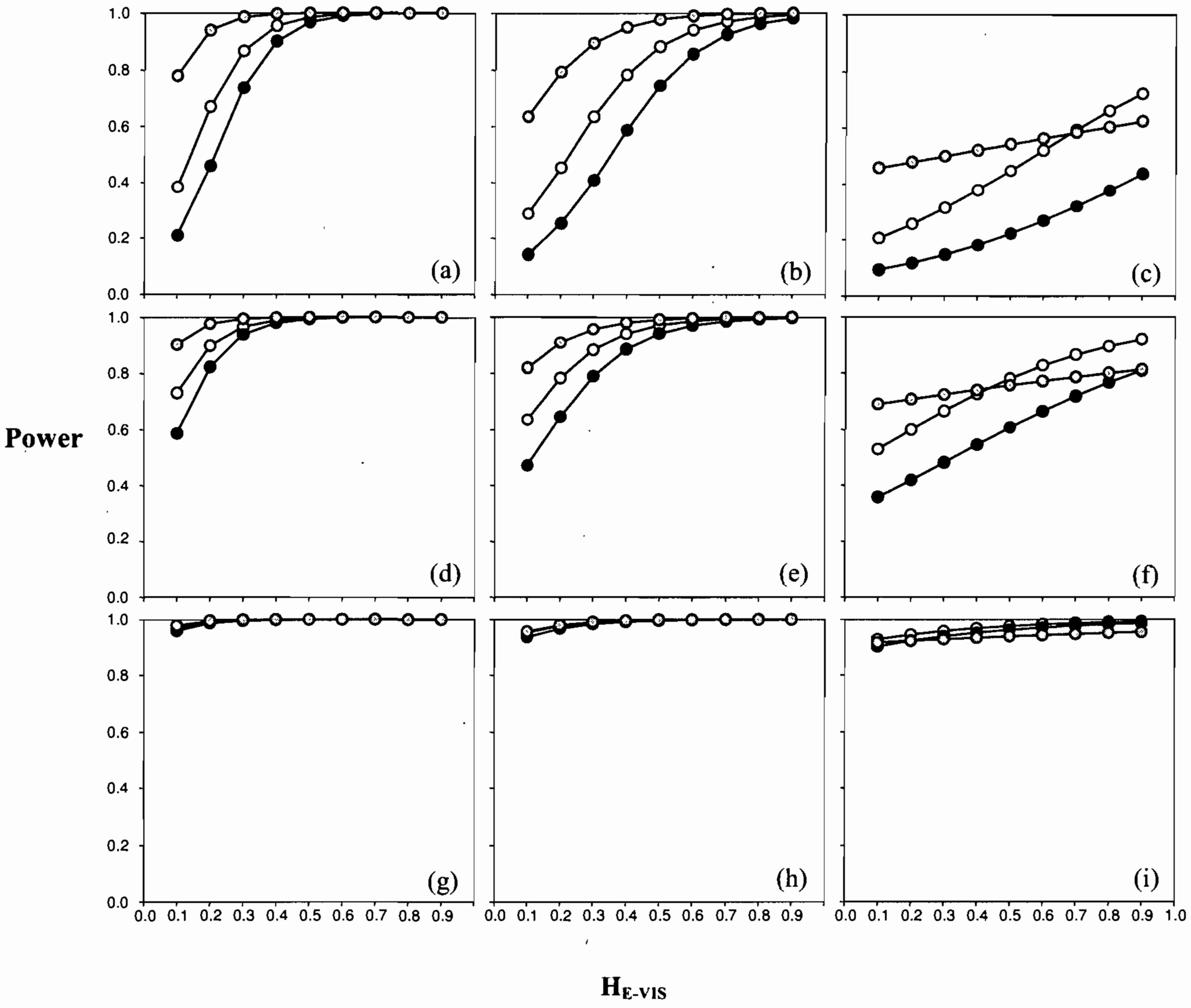


Figure 7: Power of the UT (white), the MCT (black) and the TBR2 (grey) as a function of $H_{E\text{-}VIS}$ when H_O is $0.25 \cdot H_{E\text{-}VIS}$ (a, d, g), $0.50 \cdot H_{E\text{-}VIS}$ (b, e, h) or $0.75 \cdot H_{E\text{-}VIS}$ (c, f, i), and for N of 20 (a, b, c), 50 (c, d, e) or 100 (g, h, i) individuals.

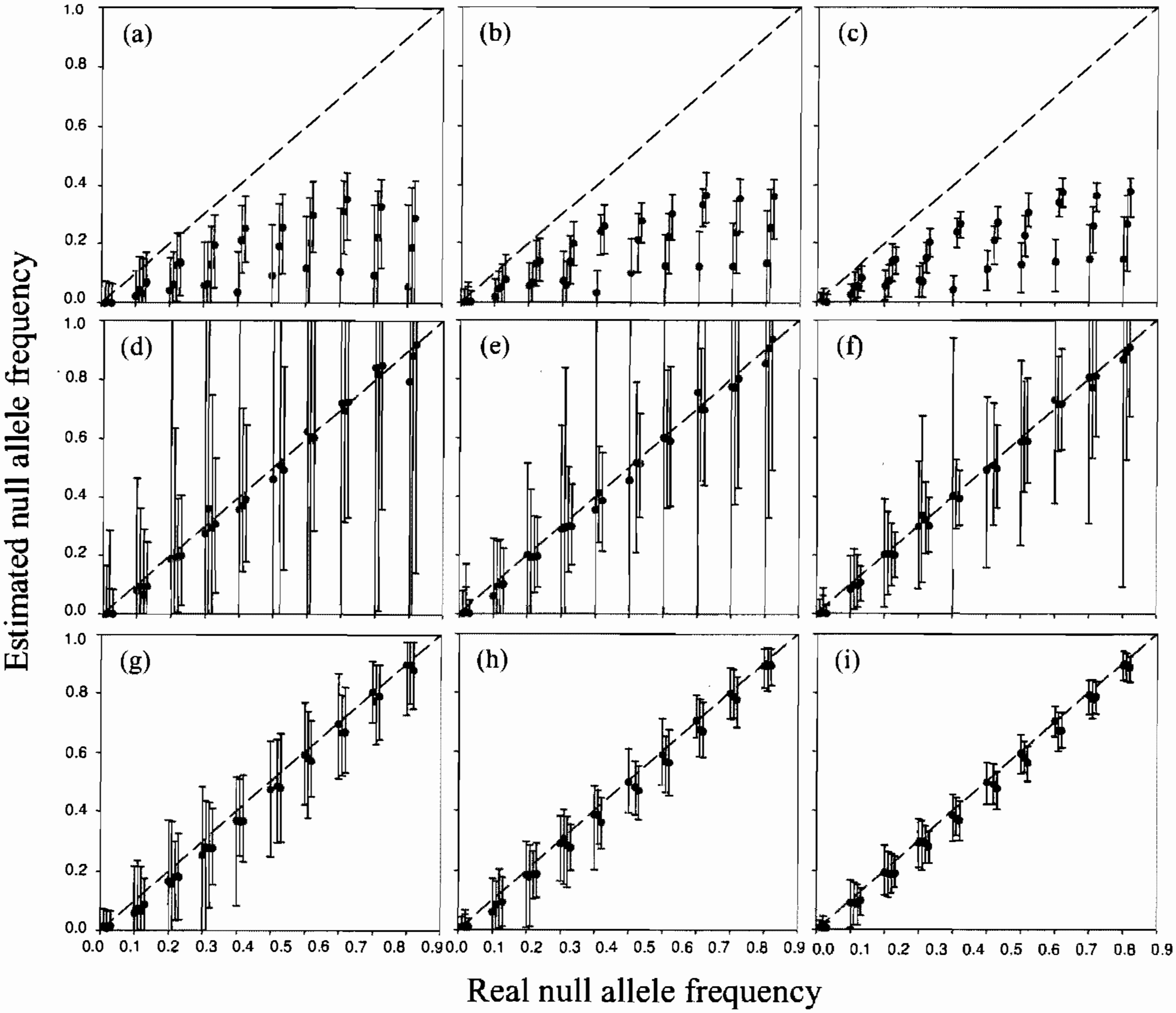


Figure 8: Simulation results showing the accuracy and precision of r_{BROOK1} (a, b, c), r_{CHAK} (d, e, f) and r_{BROOK2} (g, h, i), for N of 20 (a, b, c), 50 (c, d, e) and 100 (g, h, i) individuals performed in the 34 simulated populations described in Table 3. Simulated populations are placed in the same order in these figures as in Table 3.

cies are accurately estimated regardless of the parameters values (Fig. 8g–i).

Considering the high accuracy of r_{BROOK2}, we decided to assess its capacity to detect a null allele within a sample (this new test will be referred as TBR2). For a given sample, the confidence interval of r_{BROOK2} was computed using 1,000 bootstraps on individuals. If this interval excluded 0, the presence of a null allele was suspected in the sample. Types I and II errors for this analysis were evaluated following the same procedure described above for UT and MCT. The results are shown in Table 3 and Fig. 6 and the predictions of the following logistic model are presented in Fig. 7.

$$\lambda_{TBR2} = -0.88 + 0.03 \cdot N_{VIS} + 22.11 \cdot H_{E\text{-}VIS} - 28.39 \cdot H_O$$

The type I error of the TBR2 was lower than the 0.05 threshold as observed for the UT and the MCT (Table 3). However, the relationship between the power of TBR2 and F_{NULL} was different than that observed with previous tests. For a given $H_{E\text{-}VIS}$ condition, the power of TBR2 is lower than that of UT when F_{NULL} is low, but the reverse situation occurred when $F_{NULL} > 0.2$ for a sample size of 100 individuals (or $F_{NULL} > 0.4$ for 20 individuals). In addition, the power of TBR2 did not decrease with high F_{NULL}, thus contrasting with other tests (Table 3).

Alternative primers

According to the Fisher probability over all Mendelian loci, six of the 29 Longnose dace populations significantly deviated from the HWE (<0.001) and were no longer considered in this study. Screening of homozygotes and unamplified individuals with alternative primers allowed the detection of null alleles in 12 of the 23 populations. Four different null alleles were observed. After correcting for size differences resulting from use of new primers, the null allele size was found to vary from 242 to 258 bp, thus completely overlapping the range of the visible alleles (Fig. 9). For the purpose of the

Table 4: Results of genetic analyses with Rhca23 locus in each of the Longnose dace populations under HWE according to the Fisher method computed on the six Mendellian microsatellites (see text). Results of the UT, the MCT and the TBR2 are shown as well as the genetic diversity in terms of H_O and H_E. Estimations of r_{BROOK2} are also presented for each population. The populations in which null alleles were detected with the alternative primers are also presented. N represents the total number of individuals per population. N_{VIS} represents the frequency of individuals that were successfully amplified.

Sites	N	Original primers			Detection and estimation value of null alleles				Alternative primers		
		N_{VIS}	H_O	H_E	MCT	UT	TBR2	r_{BROOK2}	F_{NULL}	N_{VIS}	UT
7	19	0.95	0.06	0.25	**yes**	**yes**	**yes**	0.28	0.10	0.95	**<0.01**
8	22	0.41	0.11	0.57	**yes**	**yes**	**yes**	0.73	0.77	1.00	0.08
16	18	0.89	0.38	0.66	**yes**	**yes**	**yes**	0.29	0.14	0.94	**0.04**
19	22	1.00	0.46	0.76	**yes**	**yes**	**yes**	0.17	0.14	1.00	0.22
27	22	1.00	0.18	0.58	**yes**	**yes**	**yes**	0.24	0.04	1.00	**<0.01**
29	21	0.62	0.00	0.79	**yes**	**yes**	**yes**	0.67	0.62	0.95	0.3
10	19	0.26	0.20	0.69	no	**yes**	**yes**	0.77	0.84	1.00	0.53
14	20	0.70	0.57	0.70	no	**yes**	**yes**	0.35	0.30	0.90	**0.02**
15	23	0.87	0.55	0.74	no	**yes**	**yes**	0.24	0.14	1.00	**<0.01**
26	24	0.96	0.52	0.70	no	no	**yes**	0.16	0.08	1.00	**0.04**
30	16	0.94	0.07	0.19	no	no	**yes**	0.27	0.22	1.00	0.06
28	21	1.00	0.38	0.54	no	no	no	0.09	0.09	1.00	0.47
24	15	1.00	0.47	0.78	**yes**	**yes**	**yes**	0.16	no	1.00	--
25	24	0.96	0.26	0.35	no	**yes**	**yes**	0.18	no	1.00	--
21	18	0.83	0.60	0.74	no	no	**yes**	0.25	no	1.00	--
1	24	1.00	0.46	0.47	no	no	no	0.00	no	1.00	--
2	24	1.00	0.75	0.74	no	no	no	0.00	no	1.00	--
3	15	1.00	0.73	0.80	no	no	no	0.02	no	1.00	--
6	20	1.00	0.00	0.00	--	--	no	0.00	no	1.00	--
9	22	1.00	0.64	0.67	no	no	no	0.01	no	1.00	--
20	9	1.00	0.22	0.21	no	no	no	0.00	no	1.00	--
22	16	1.00	0.06	0.18	no	no	no	0.09	no	1.00	--
23	13	1.00	0.92	0.88	no	no	no	0.00	no	1.00	--

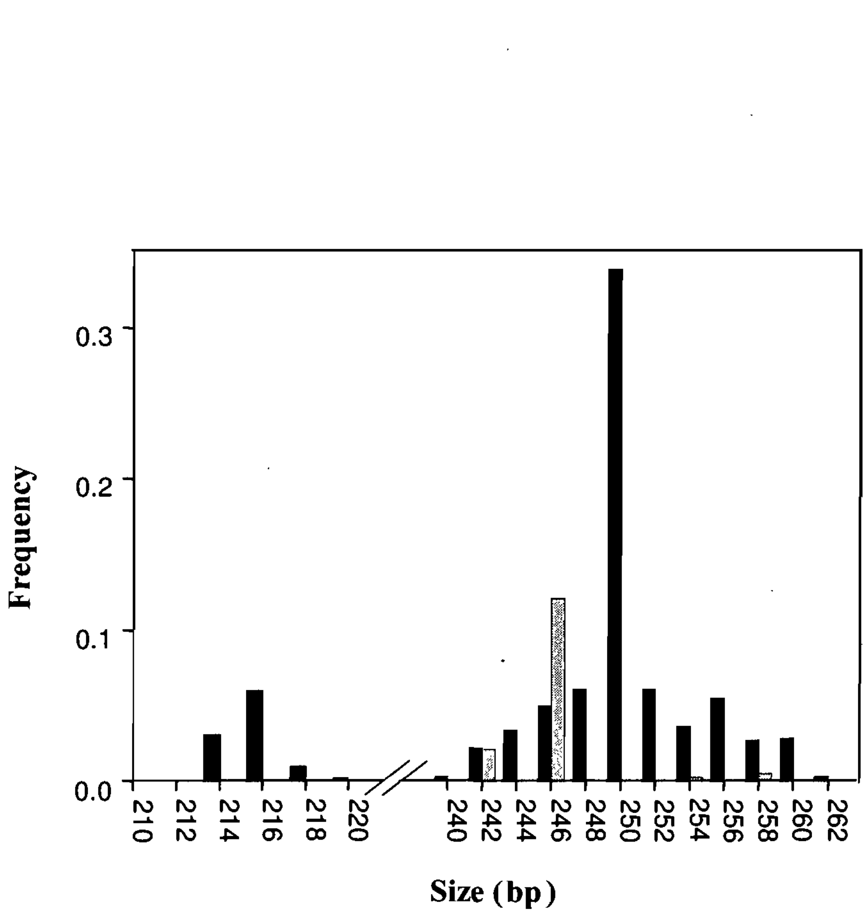


Figure 9: Size and frequency of visible (solid bars) and null (grey bars) alleles detected with alternative Rhca23 primers over all populations of the real dataset.

analyses, the combined frequency of these alleles was considered as representative of a single null allele. The variation of ^F_{NULL} among populations was large (Table 4) and varied from 0.04 (pop. 27) to 0.84 (pop. 10).

Tests correctly predicted the presence of a null allele in eleven (TBR2), nine (UT) and six (MCT) of the populations (Table 4). All tests missed the null alleles in population 28, but their frequency was very low (0.09). UT missed two additional populations (26 and 30) and MCT three additional ones (10, 14 and 15). Note that both MCT and UT were not performed on population 6 because it was fixed.

Interestingly, a null allele was detected by at least one of the tests in three populations (21, 24 and 25) in which alternative primers had failed to amplify any additional alleles. The F_{NULL} estimated in these populations were quite high and varied from 0.16 to 0.25. The low type I error of these tests suggests that the null alleles remained undetected by the alternative primers. This last hypothesis is supported by three points. First, the difference between $\hat{H}_O$ and $\hat{H}_{E\text{-}VIS}$ remained high. Second, UT performed afterward on populations in which alternative alleles were found remained significant in six populations where additional alleles were detected, indicating that an excess of homozygotes remained in several populations (Table 4). Third, the amplification failures (between 5 and 10%) remained in four populations (7, 14, 16 and 29), suggesting the presence of still undetected null homozygotes. This hypothesis may also explain the overestimation of F_{NULL} by r_{BROOK2} obtained in some of these populations.

Discussion

The results of this study revealed that the different tests developed to detect null alleles have limitations, especially when sample size is low. Unfortunately, this is often the case with population genetics studies that rely on these analytical tools (e.g. Roberts et al. 2005; Yang et al. 2005; Clay Green et al. 2006). One of the most important limitations of the UT and the MCT is observed when proportion of missing data is high. Because unamplified individuals are discarded in these tests, the analyses are performed on a reduced fraction of the sample, resulting in a marked decrease in power. Even though the easy and fast solution to this situation is the discard of the problematic microsatellite locus, researchers who study organisms with limited number of available microsatellites might be forced to keep them in their genetic analyses. In this context, TBR2 is a suitable alternative.

Interestingly, both simulated and real datasets revealed that combining UT and TBR2 gave a high probability of detecting null alleles. Simulations clearly demonstrated that TBR2 performed better in almost all situations except in those where both $H_{E\text{-}VIS}$ and H_O were high or where F_{NULL} was low. In these cases, the UT was the most powerful analysis. MCT must be used with caution when $N_{VIS} < 50$ individuals or when $H_{E\text{-}VIS} < 0.5$ (this situation is unlikely with highly variable microsatellite loci) since the probability of missing a null allele is higher. As a result, this test never detected a null allele that had not already been detected by either UT or TBR2.

Once null alleles have been detected, the use of the r_{BROOK2} estimator appears more suitable than time-consuming procedures with alternative primers. This estimator provides an accurate estimation of F_{NULL} in both simulated and real datasets. In contrast, amplification with alternative primers does not guarantee successful detection of all the null allelic states within a sample, as observed in this study and previous ones (e.g. Walter and Epperson 2004). Because they are inaccurate or have a large variance, other estimators are in most cases inappropriate. Furthermore, they are never more accurate than r_{BROOK2}. This advantage of r_{BROOK2} over r_{CHAK} is consistent with the findings of Chapuis and Estoup (2007). However, while our results clearly stated that r_{BROOK2} is accurate throughout the F_{NULL} range, Chapuis and Estoup (2007) simulations revealed that this estimator slightly overestimates the null allele frequency and they strongly suggested the use of another estimator (Dempster et al. 1977), which appeared as a superior alternative according to their analyses.

To verify the reliability of this last estimator on our simulated datasets, a supplementary analysis using the Dempster et al. (1977) statistic was performed using the samples of 20 individuals, a condition for which the variance of r_{BROOK2} was the highest. At the opposite of the Chapuis and Estoup (2007) results, both estimators were equally accurate and Dempster et al. (1977) show a slightly higher variance than r_{BROOK2} (data not shown). In our opinion, both estimators can be used with equal confidence. However, the longer computation time related to a maximum likelihood algorithm such as the Dempster et al. (1977) estimator appears as an argument for the use of the heuristic r_{BROOK2} estimator.

Clearly, the reliability of a method for detection and frequency-estimation of null alleles is correlated to its capacity to integrate the missing data. However, including such data implies a lot of preliminary cautions. The use of several markers is thus necessary to assure that all individuals who did not amplify because of improper DNA extraction be identified and withdrawn. An additional round of PCR must then be performed on the remaining individuals that failed to amplify so that null homozygotes can be discriminated from data resulting from technical errors (the information from these individuals is then incorporated to the dataset). This corrected dataset should be examined with MCT to discriminate the effects of null alleles from those due to large allele dropout or genotyping errors. Even though its power has not been assessed when used for this purpose, MCT still remains the only test that gives another possible interpretation for the presence of null alleles.

In conclusion, it worth noticing that all estimators evaluated in this study, as well as all the tests currently available to detect null alleles, necessitate the accordance with HWE as a prerequisite. Now, retaining only populations strictly in accordance with this equilibrium (as performed in this study) cannot be considered as a realistic solution when working in natural environment. van Oosterhout et al. (2006) have recently proposed a novel estimator, based on fixation indices, that aims at estimating null allele frequency in non-equilibrium populations. Even though the accuracy of this estimator still needs to be evaluated, it represents a good start. For instance, using this estimator with a bootstraps procedure as the one used for TBR2 might be part of a solution that will jointly estimate null allele frequency along with other factors that could cause deviation of HWE.

CHAPITRE III

The impact of post-glacial marine invasions on the genetic diversity of an obligate freshwater fish, the Longnose dace (*Rhinichthys cataractae*), on the Quebec peninsula.

Publié dans: Canadian Journal of Fisheries and Aquatic Sciences (2006) **63**,1429-1438

Résumé

Les mers postglaciaires ont possiblement eu les effets significatifs sur la structure génétique des populations de poissons d'eau douce. Afin de vérifier cet effet, la variabilité mitochondriale a été évaluée dans 32 populations de naseux des rapides (*Rhinichthys cataractae*), situées à l'intérieur et à l'extérieur des frontières des mers de Champlain et de Laflamme, qui ont inondé la péninsule québécoise au cours de la dernière déglaciation. Trois clades d'haplotypes divergeant d'une à deux mutations ont été trouvés. Les populations situées à l'extérieur des zones inondées par les invasions marines ont montré une forte structure spatiale. Cependant, une plus grande diversité génétique due au mélange des trois clades et d'une lignée évolutive supplémentaire a été observée à l'intérieur des limites des mers postglaciaires. La faible divergence entre les trois principaux clades suggère une origine commune malgré la présence de cette espècedans plusieurs des refuges glaciaires connus. Les invasions marines agissant comme barrières à la colonisation, surtout depuis le refuge de l'Atlantique, sont proposées comme explication possible à ce résultat. Cette étude représente ainsi un argument pertinent pour l'intégration des invasions marines dans les modèles décrivant la colonisation postglaciaire des espèces d'eau douce dans le nord-est de l'Amérique du Nord.

Abstract

Postglacial seas are expected to have had significant effects on the genetic structure of populations of obligate freshwater fishes. To assess this influence, mitochondrial DNA variability was evaluated in 32 populations of longnose dace (*Rhinichthys cataractae*) of the Quebec peninsula located within and outside of the maximum extent of marine invasions of the Champlain and Laflamme seas. Three clades of haplotypes diverging from one to two mutations were defined. Despite this low divergence, a clear and significant spatial genetic structure was observed outside of the extent of marine invasions. However, a higher genetic diversity was observed in populations located within the extent of marine invasions because of the admixture of these clades with an additional lineage restricted almost exclusively to those areas. The low genetic divergence between the main haplotypes suggests a single origin, despite the known presence of this species in various refuges. Marine invasions preventing entry to the peninsula, especially from Atlantic refuge, are proposed as a possible explanation to this particular result. This study is a relevant argument for integrating postglacial marine invasions into postglacial colonization models of freshwater species in the north eastern part of North America.

Introduction

The spatial genetic structure of populations results from various ecological and demographic causes (Walker and Avise 1998). For species distributed over territories covered by the ice sheet of Pleistocene glaciations, the postglacial dispersal remains one of the major processes that shaped the genetic structure and diversity (Avise et al. 1998; Bernatchez and Wilson 1998; Taberlet et al. 1998). Among potential factors influencing colonization opportunities, the important modifications of the hydrological network that occurred during deglaciation are of prime importance (Rempel and Smith 1998). For instance, the presence of postglacial seas in the northeastern part of North America is expected to have had profound effects on colonization. Both high salinity and size of these seas are believed to act as barriers for a lot of organisms.

A large portion of the Quebec peninsula was covered by postglacial seas about 8000 – 11 000 years BP (Elson 1969). Those seas were formed over territories that were isostatically depressed below sea level after the melting of the ice sheet, allowing the inland invasion of saltwater from the Atlantic Ocean. Those marine invasions are defined by two main seas. Laflamme Sea inundated a large territory including the actual Lake Saint-Jean and Saguenay River (Elson 1969; Fig. 10). The southwest part of the peninsula was covered by the Champlain Sea, which formed when ocean water invaded the Saint Lawrence lowlands (Occhietti 1989; Fig. 10). Impacts of those seas on the colonization opportunities are expected to be important as both the Saint Lawrence and Saguenay hydrological networks are believed to be major tributaries that allowed the colonization of the eastern and northeastern parts of the peninsula (Legendre and Legendre 1984; Bernatchez 1997). Although these marine invasions at their maximum extent provided dispersal routes inland for marine, estuarine, and salt marsh species (McAllister et al. 1988), they probably provided an insurmountable barrier for obligate freshwater species, preventing or delaying their colonization of newly deglaciated territory.

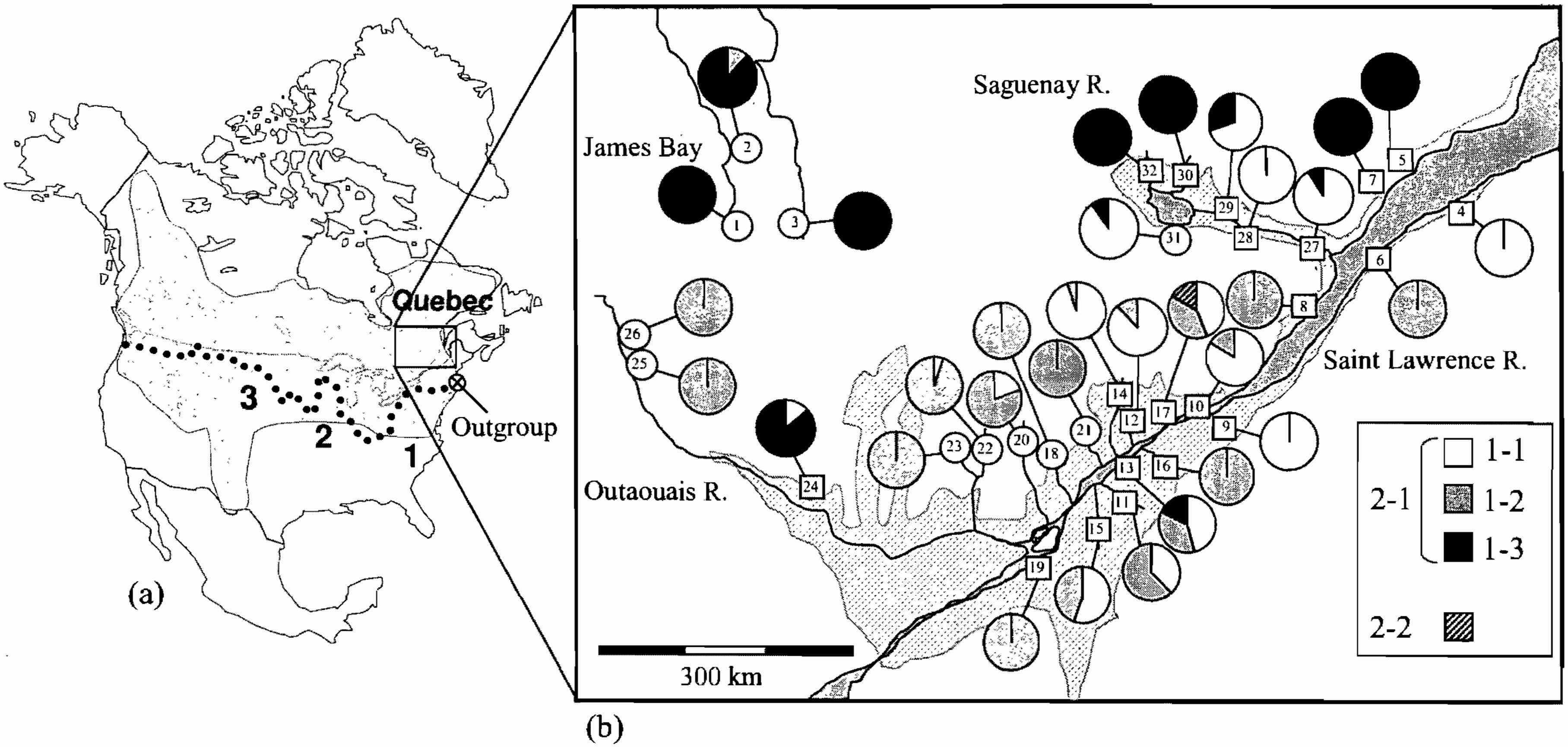


Figure 10: (a) Maximum extent (dotted line) of the last Pleistocene glaciation over the North American territory, location of the main glacial refuges that putatively contributed to the colonization of fish in the Quebec Peninsula (1, Atlantic; 2, Mississippian; 3, Missourian) and actual geographic distribution of longnose dace (Rhynicthys cataractae) (shaded region). (b) Geographic position of populations combined with their clade composition. Hatched region corresponds to the maximum extensions of the Champlain and Laflamme seas (after Occhietti 1989). Sites covered or not by the marine extent are respectively identified by squares or circles. Denomination of clades and lineages is based on the nested design presented in Fig. 11. Refer to Table 5 for population and drainage identifications.

Fishes were extensively used as models to understand postglacial colonization (Billington and Hebert 1988; Bernatchez and Wilson 1998 and references therein). As a result, the genetic structure of numerous species was studied over the northeastern part of North America. However, most of the species surveyed over the Quebec peninsula were either anadromous or salt-tolerant Salmonidae (Bernatchez and Dodson 1991; Wilson and Hebert 1996; Turgeon and Bernatchez 2001), and the few surveys of exclusive freshwater species had a limited sampling (Lafontaine and Dodson 1997; Senanan and Kapuscinski 2000). Hence, the hypothesis that Champlain and Laflamme postglacial seas played a significant role on the genetic structure of populations of stenohaline fish species has never been addressed.

The Cyprinidae exhibit many valuable characteristics that allow the evaluation of the potential impacts of marine invasions on the postglacial colonization of fish. As one of the most diversified families, the Cyprinidae colonized almost all of the freshwater habitats available in the Northern Hemisphere (Scott and Crossman 1973) and covered most of the Quebec peninsula, including the areas once inundated by postglacial seas (Legendre and Legendre 1984). Furthermore, both their intolerance to saltwater and their low dispersal capacity across large water bodies (Scott and Crossman 1973) make them especially sensitive to the presence of marine invasions. Finally, the combination of a short reproductive cycle and large population size is expected to result in high genetic diversity, which may increase the accuracy of the dispersal model (Avise et al. 1987).

The objective of this study was to assess the influences of the postglacial Champlain and Laflamme seas on a common Cyprinidae species in North America, the longnose dace (*Rhinichthys cataractae*). In this perspective, the geographic organization of the mitochondrial DNA (mtDNA) diversity was determined from 32 populations located within and outside the extent of those marine invasions. To complement the description of fish dispersal in the northeastern part of North America, arguments are given for the pertinence of integrating postglacial seas into phylogeographic studies of freshwater species in North America in the future.

Materials and methods

Model species and sampling

Longnose dace is typical of turbulent rivers located between the Atlantic and Pacific coasts, from Mexico to the Yukon (Scott and Crossman 1973; Fig. 10). Its present-day distribution overlaps a large fraction of the territory covered by Pleistocene glaciations and results from colonization from various glacial refuges (McPhail and Lindsey 1970; Fig. 10). Adult longnose dace occupies riffles (Scott and Crossman 1973) and exhibits high site fidelity (Hill and Grossman 1987). This fish has little interaction with humans but is known to be used occasionally as bait by sportfishers. Populations are expected to be almost undisturbed by stocking activities or translocations in Canada (Scott and Crossman 1973).

A total of 640 individuals were sampled by electrofishing (LR-24, Smith-Root Inc., Vancouver, Washington) from 32 rivers of the Quebec peninsula in Canada (Table 5). These rivers are located within four drainage basins (James Bay, Saint Lawrence River, Outaouais River, and Saguenay River) believed to be important colonization routes used by fishes after the deglaciation (Legendre and Legendre 1984). Efforts were made to sample populations both within and outside the maximum extent of marine invasions (Fig. 10). Both Champlain and Laflamme seas rose to an altitude of 200 m (Lasalle and Tremblay 1978; Occhietti 1989). As a result, 21 sampling sites below an altitude of 200 m and located within the inferred maximum extent of these marine invasions (Occhietti 1989) were assumed to be once covered by a postglacial sea. The other populations were considered outside the maximum extent of the marine invasions (Table 5; Fig. 10). To increase the probability of sampling a single panmictic population, each river was sampled only once within a single riffle. Sample sizes ranged from 9 to 26 individuals for each population (Table 5). For each individual, a piece of the caudal fin was removed and stored in 95% ethanol for genetic analysis. A population from the Eight-Mile River (Connecticut, USA), located in the Atlantic watershed (Fig. 10), was analyzed as the outgroup.

Table 5: The geographic position and genetic diversity of Longnose dace populations analysed in this study. The longitude (west of the prime meridian) and latitude (north of the equator) of the sample sites, the number of individuals (N), the number of haplotypes (h), the number of clades (k; associated to the nested design presented at Figure 11) and both gene (H_E) and nucleotide (π) diversities are presented. Numbers next to each population name will be used to identify populations in text and figures.

Population	**Longitude**	**Latitude**	**Alt.**	**n**	**h**	**H_E**	**k**	**π**
			(m)			**($x10^{-1}$)**		**($x10^{-4}$)**
James Bay drainage								
1-Martel	77° 54' 33"	48° 31' 05"	320	25	2	1.53	1	6.55
2-Coigny	77° 57' 56"	49° 05' 43"	200	25	3	2.27	2	4.99
3-Mégiscane	77° 07' 08"	48° 22' 43"	320	18	2	4.25	1	18.20
Saint Lawrence River drainage								
4-Mitis	68° 04' 56"	48° 31' 15"	100	9	2	2.22	2	4.75
5-Betsiamite	68° 57' 30"	49° 00' 40"	15	24	2	4.31	1	18.40
6-Trois-Pistoles	69° 12' 49"	48° 05' 39"	15	19	1	0.00	1	0.00
7-Sault-aux-Cochons	69° 15' 41"	48° 47' 33"	120	18	2	3.66	1	15.60
8-Malbaie	70° 11' 49"	47° 41' 38"	30	24	3	2.36	1	10.40
9-Chaudière	71° 12' 58"	46° 35' 16"	120	20	1	0.00	1	0.00
10-Jacques-Cartier	71° 44' 56"	46° 41' 35"	40	24	3	3.01	2	19.50
11-St-Francois	72° 29' 24"	45° 53' 16"	90	24	2	4.89	2	20.90
12-Des Envies	72° 30' 13"	46° 40' 23"	120	26	2	2.12	2	4.54
13-Nicolet	72° 32' 12"	46° 09' 16"	20	22	3	6.36	3	27.20
14-Mékinak du Nord	72° 38' 51"	46° 44' 52"	170	19	2	1.05	2	6.75
15-Yamaska	72° 56' 55"	45° 37' 16"	30	20	3	6.26	2	18.30
16-Bécancour	72° 09' 27"	46° 12' 52"	60	25	3	6.60	1	26.10
17-Ste-Anne	72° 09' 30"	46° 38' 17"	30	18	4	6.73 (2.12)*	2	145.00 (4.54)*
18-Noire	73° 39' 14"	46° 20' 16"	240	19	1	0.00	1	0.00
19-Châteauguay	73° 48' 00"	45° 15' 18"	40	22	2	4.55	1	19.40
20-Assomption	73° 58' 22"	46° 25' 12"	300	10	2	3.27	2	14.00
21-Du Loup	73° 09' 01"	46° 33' 44"	180	17	1	0.00	1	0.00
Outaouais River drainage								
22-Diable	74° 30' 45"	46° 22' 17"	330	17	2	1.11	2	2.37

23-Rouge	74° 47' 36"	46° 21' 31"	230	13	1	0.00	1	0.00
24-Kazabazua	76° 03' 04"	45° 56' 42"	150	15	2	2.48	1	10.60
25-Lavallée	79° 21' 30"	47° 10' 17"	200	24	2	0.83	1	3.56
26-Loutre	79° 21' 04"	47° 24' 42"	230	23	2	0.83	1	1.78
Saguenay River drainage								
27-Petit-Saguenay	70° 03' 50"	48° 12' 10"	30	23	2	1.66	2	7.09
28-À Mars	70° 56' 06"	48° 19' 08"	50	22	1	0.00	1	0.00
29-Shipshaw	71° 15' 11"	48° 33' 10"	110	21	3	5.14	2	36.20
30-Péribonka	71° 45' 44"	48° 50' 32"	150	15	3	3.62	1	13.40
31-Métabetchouane	71° 59' 55"	48° 19' 18"	260	20	2	1.90	2	16.20
32-Aux Rats	72° 14' 36"	48° 53' 35"	120	19	2	3.51	1	15.00

Molecular analysis

This study focused on the control region of and the cytochrome b of the mitochondrial DNA (mtDNA), which are expected to display high polymorphism within species (McMillan and Palumbi 1997). Primers were designed on preserved sequences of different fish mtDNA genomes available on GenBank. The primers 5′-GTGACTTGAAAAACCACCGTTG- 3′and 5′-AANNGTTGGTNGKYTCTTACT-3′successfully amplified, by polymerase chain reaction (PCR), a segment of approximately 1500 bp, covering the cytochrome b, tRNA-Thr, tRNA-Pro, and the first 300 bp of the control region. The following reaction conditions were used: a 12.5 μL reaction aliquot containing 1.5 mmol·L^{-1} of $MgCl_2$ and 2.5 nmol·L^{-1} of each dNTP, 0.2 unit of Taq polymerase, 1.25 μL of 10× Taq polymerase buffer (Invitrogen Corp., Burlington, Ontario), and approximately 20 ng of DNA. Reaction conditions included an initial denaturation of 30 s at 92 °C, followed by 45 cycles combining 10 s at 92 °C, 15 s at 50 °C, and 2 m at 68 °C, and a final extension of 10 m at 68 °C. Both strands were sequenced with a CEQ 2000XL DNA Analysis System (Beckman Coulter Inc., Fullerton, Calif.) on 20 individuals from 10 distant populations. According to these sequences, primers specific to longnose dace were designed to amplify shorter segments on cytochrome b (232 bp; 5′-CATCTGTCGAGACGTTAAC-3′and 5′-TAATAACGGTAGCGCCTC-3′) and control region (236 bp; 5′-ACCCCTGGCTCCCAAAGC-3′and 5′-GGTCTATGTACGTCTTAG-3′). PCR reactions that optimize the amplification of those loci contained the same products as above. However, PCR conditions include an initial denaturation of 30 s at 92 °C, followed by 45 cycles combining 10 s at 92 °C, 15 s at 48 °C, and 5 s at 68 °C, and a final extension of 2 m at 68 °C.

The polymorphism of these loci was then screened using the single strand conformation polymorphism (SSCP, Orita et al. 1989). Both loci were conjointly electrophoresed on a 6% nondenaturing gel for 10 h at 20 W in 0.5× TBE (Angers and Bernatchez 1998). The sequence of both loci was determined for individuals showing variation at either cytochrome b or control region segments. To confirm the reliability of the SSCP protocol, sequencing of individuals coming from different populations but displaying the same migration pattern was performed. This procedure was performed on six individuals for each of the most frequent haplotypes (I, III, and IX; see Results).

Data analysis

Genetic diversity within population was calculated using the number of haplotypes, haplotype frequencies (Nei 1987), and sequence information (π). Partition of the genetic diversity within and among populations was evaluated with both haplotypes frequencies (F_{ST}) and sequences (ϕ_{ST}) using Arlequin v.2.2 (Schneider et al. 2000). ϕ_{ST} distances were computed using pairwise differences. Significance of both statistics was determined by comparison with distributions obtained after 999 random permutations of the data.

A minimum spanning network was inferred from the observed haplotypes. Two repeat arrays were observed on the control region locus. Each dinucleotide deletion in these repeats arrays was considered as a single fifth state. Haplotypes were then grouped following the rules used in Templeton et al. (1987) and Templeton and Sing (1993). Accordingly, haplotypes or groups of haplotypes that are one mutation apart on the network are united, producing a hierarchical series of nested clades. This nesting procedure is specifically designed for haplotypes separated by few mutations. However, haplotypes found in this study can be grouped into two highly divergent major clades (hereafter referred as lineages). To counteract this problem, the nesting procedure was done independently in each of these lineages. As an indication of the support of this haplotype grouping, bootstraps (1000) were performed following by a parsimony analysis using the SEQBOOT, DNAPARS, and CONSENSE modules of PHYLIP 3.62 (Felsenstein 2004).

Processes defining the population structure

A nested clade analysis (NCA) was conducted to investigate processes that created the spatial organization of the observed genetic variation. The method detects geographic associations of haplotypes and explains them in terms of contributions from either historical or present-day processes that have played a role in defining the patterns of population structure (Templeton et al. 1995; Templeton 1998). The geographical range of each clade (Dc), the distance of each clade from the geographical centre of the next higher level clade (Dn), and contrasts of these distances between tips and interior clades (I–T Dc and I–T Dn) were computed using GeoDis 2.2 (Posada et al. 2000). The significance of these different estimators was evaluated by 1000 permutations.

Geographic organisation

To assess the relative importance of geographical factors in the spatial organisation of genetic diversity among populations, a canonical correspondence analysis (CCA) was performed using Canoco for Windows version 4.02 (ter Braak and Smilauer 1999). This method, designed for relating species composition to different predictive variables (ter Braak 1986), was successfully used to describe relationships between environmental variables and genetic composition (Angers et al. 1999; Costello et al. 2003; Volis et al. 2004). CCA was computed using a matrix Y representing frequency of clades in each population and a matrix X in which standardized geographical variables associated to the sampling point were compiled. Latitude, longitude, altitude, and three binary variables coding for the four main drainages were used as geographical variables. A cubic polynomial function of the geographic coordinates was also constructed to allow a full representation of the longitude and the latitude effects on the genetic variation. Thus, in addition to their first degree forms, the latitude^2, the longitude^2, the latitude^2 and the latitude^3 were added to the previous matrix X, along with the combinations longitude*latitude, longitude^2*latitude and longitude*latitude^2. The matrix X was thus constituted of 13 geographic variables. Test of significance of this CCA was performed using 999 random permutations (detailed procedure described in Angers et al. 1999).

The importance of these geographic factors in shaping the diversity within population was also evaluated. Multiple regressions were performed using either gene or nucleotide diversity computed for each population as the dependant variable versus the geographic matrix X described above. Tests of significance of the total regression model and of the partial regression coefficients were performed with 999 permutations of residuals under full model using a program for multiple linear regressions developed by Legendre (2002).

Structuring effects of postglacial seas

The structuring effects of the postglacial seas were evaluated on genetic diversity within and among populations. Populations were separated in relation to whether or not their present locations were covered by a postglacial sea during the deglaciation processes. The structuring effects on the variation among populations were evaluated by comparing the average clade frequencies of the two groups with a t statistic tested with 999 permutations. Each clade was analyzed separately. The structuring effects of those seas on the variation within population were evaluated by comparing the average genetic diversity (haplotype and nucleotide) of those two groups with the same statistical procedure.

Results

Diversity and haplotypes network

A total of 14 haplotypes, characterized by 13 and 17 mutations on cytochrome b and control region, respectively (Table 6), was detected in this survey. Sequencing and SSCP results were fully reliable. The haplotype network shows that haplotypes can be assigned to two distinct lineages (2-1 and 2-2; Fig. 11) separated by 21 mutations (4.5% divergence; Table 6). Twelve haplotypes belong to lineage 2-1 and two to lineage 2-2 (Fig. 11). Repeat arrays in the control region show two types of mutation. Single dinucleotide deletion was observed in both arrays but only in lineage 2-1 (haplotypes VII, VIII, and X; Table 6). The two haplotypes of lineage 2-2 show a base substitution in the last repeat of the AC array (position 175; Table 6). Interestingly, the haplotype observed in the population from the Atlantic watershed (outgroup) differed by only two or three mutations from haplotypes of lineage 2-2 (Table 6). Diversity within those lineages is 0.3% and 0.5% for 2-1 and 2-2, respectively. Lineage 2-1 can be divided into three clades characterized by an abundant interior haplotype (I, III, and IX for 1-1, 1-2, and 1-3, respectively) and one to five scarce tip haplotypes (Fig. 11). Bootstrap results showed high support of the differentiation between lineages 2-1 and 2-2 (Fig. 11), suggesting a high separation time between them. However, lack of support was observed within lineage 2-1, in which each clade appeared in less than 25% of the bootstrapped trees (Fig. 11). This result may suggest a recent differentiation of the three clades, all originating from a single refuge.

Haplotypes from lineage 2-1 were detected in 637 of the 640 individuals (99.5%) from all populations sampled in Quebec, whereas those of lineage 2-2 were restricted to three individuals from population 17 (Fig. 10). The number of clades (i.e., 1-1 to 1-3) observed within populations ranged from 1 to 3, but 18 of the 32 populations exhibited a single clade. Gene diversity and nucleotide diversity ranged from 0 to 0.67 (average 0.27) and from 0 to 0.014 (average 1.5 × 10–3; Table 5), respectively. Because they strongly inflate the nucleotide diversity of population 17, individuals that shared haplotypes belonging to lineage 2-2 were removed from further analyses. Analysis of molecular variance revealed a significant structure at the population level for both haplotype frequencies ($F_{ST} = 0.67$, $p < 0.001$) and sequences ($\phi_{ST} = 0.61$, $p < 0.001$).

Table 6: Variable sites for each haplotype in both mtDNA regions sequenced in this work. Numbers are associated to site positions of the cytochrome B (Cytb) and control region (CR) sequences of the zebrafish (*Danio rerio*; genbank accession number AC024175).

		Cytb													CR														Populations
		2	2	2	2	2	2	2	3	3	3	3	3	4	1	1	1	1	1	1	1	1	1	1	1	11	22	2	
		1	2	3	3	4	5	8	0	5	6	7	8	0	0	2	2	2	3	4	4	5	5	6	6	77	00	4	
Lineage	Haplotype	4	6	2	5	1	5	0	7	5	7	3	2	3	0	5	6	9	1	2	8	4	6	3	6	45	67	3	(Access # DQ400450-DQ400479)
2-1	I	A	T	A	C	G	G	T	G	C	A	A	T	A	A	G	C	G	A	C	G	A	A	C	A	AC	GT	C	4,9-15,17,20,22,24,27-29,31
	II	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	..	..	.	4
	III	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	..	..	.	2,6,8,12,15-19,21-23,25,26
	IV	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	G	..	..	.	20
	V	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	G	..	..	.	11,13,15,16
	VI	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	..	..	.	26
	VII	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	--	..	.	8,10
	VIII	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	..	--	.	8,10,14,16,19,25
	IX	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	..	..	.	1-3,5,7,13,24,27,29,30,32
	X	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	..	--	.	1,3,5,7,29-32
	XI	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	..	..	.	30
	XII	.	C	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	G	..	..	.	2
2-2	XIII	G	C	G	T	A	A	.	A	T	G	G	C	G	G	.	T	A	G	T	A	.	T	.	.	.T	..	A	17
	XIV	G	C	G	T	A	A	.	A	T	G	G	C	G	G	.	T	A	G	T	A	.	.	.	.	.T	..	.	17
Outgroup		.	C	G	T	A	A	.	A	T	G	G	C	G	G	.	T	A	.	T	A	.	T	.	.	.T	..	A	

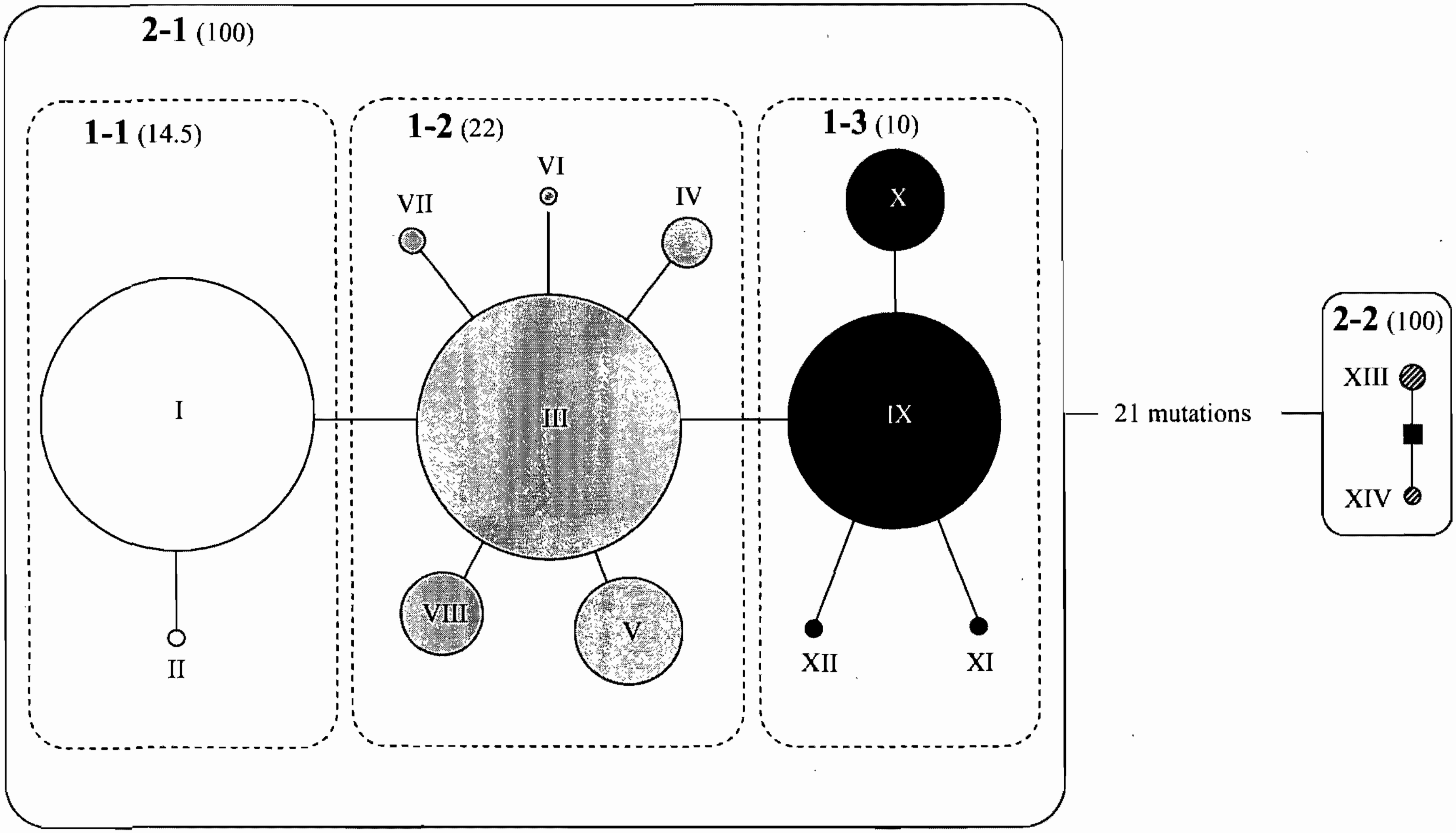


Figure 11: Minimum spanning network based on the mutational steps of mtDNA haplotype sequences observed on longnose dace (Rhynicthys cataractae). Nested clades design is also shown. Numbers in parentheses are the percentages of occurrences of the clades after 1000 bootstraps when a maximum parsimony analysis was performed. The diameter of each circle reflects the frequency of the given haplotype observed within all populations surveyed. Shading corresponds to that used in Fig. 10. The solid square represents an haplotype that was not observed in the survey.

Processes defining the population structure

Below the 1-step level, the nested clade analysis was not very informative because of the presence of unique haplotypes (II, VI, XI, and XII) and those found in a restricted number of populations (IV, V, and VII). Considering the geographic scale of this study, the sampling plan can be invoked to explain these inconclusive results. The analysis performed on the 1-step level reveals that the spatial genetic structure conformed to a contiguous range expansion by the clades of lineage 2-1 (Table 7). The 2-step level analysis suggests a past allopatric fragmentation as a possible cause for the divergence of lineages 2-1 and 2-2 (Table 7). This conclusion strengthens the hypothesis of distinct glacial refuge as the origin for those lineages.

Geographic organisation

The CCA revealed that 64% of the genetic variance among populations can be explained by the geographic variables. The overall significance test of the two canonical axes showed that the relationship between clade frequencies and geographic variables was highly significant (F = 4.307, p = 0.001). The correlation biplot (Fig. 12) underlines the specific geographic distribution of each clade. Clade 1-2 is closely related to populations sampled in lower latitudes of the Saint Lawrence and Outaouais drainages. An opposing signal can be observed for clade 1-3, exclusively associated with populations in higher latitudes of the James Bay and, but to a lower extent, of Saguenay drainages. Clade 1-1 is particular in that it shows a longitudinal association and is negatively correlated to altitude. Populations in which clade 1-1 was detected were almost exclusively in eastern regions (Fig. 10) of the Saint Lawrence and Saguenay drainages. On the other hand, no significant associations were found between geographical factors and both gene ($R^2 = 0.51$, $p = 0.20$) and nucleotide diversities ($R^2 = 0.25$, $p = 0.74$).

Structuring effects of postglacial seas

Interestingly, the genetic structure is not only related to geographic coordinates, because structuring effects of postglacial seas can also be inferred. As pointed out in the correlation biplot (Fig. 12), clade 1-1 is almost exclusively present in populations once covered by postglacial seas (t = 2.416, p = 0.025). In contrast, clade 1-2 is more abundant in populations outside the extent of the Champlain and Laflamme seas (t = –2.048, p = 0.046). Finally, no relation was detected for clade 1-3, which is equally frequent in locations within and outside the extent of the postglacial seas (t = 0.076, p = 0.924). These results, combined with the geographic organisation delineated in the previous section, suggest that the contiguous range expansion was not straightforward, but was constituted by three waves of colonists, each associated with a particular clade.

In contrast to geographic factors, the structuring effects of the postglacial seas can also be observed in the genetic diversity within populations. Populations sampled in locations once covered by postglacial seas showed on average twofold greater haplotype and nucleotide diversity (0.32 × 10–3 and 1.37 × 10–3, respectively) than populations living outside the maximum extent of the marine invasions (0.15 × 10–3 and 0.62 × 10–3, respectively). Both of these differences were significant (p = 0.025 and 0.033 for haplotype and nucleotide diversity, respectively).

Discussion

Genetic structure of longnose dace populations

Longnose dace populations sampled in Quebec showed a genetic diversity similar to those of other fish species found between 45°N and 50°N (Bernatchez and Wilson 1998). The three clades of lineage 2-1 observed on longnose dace differed by no more than 0.3%. This pattern is consistent with northern species that show a genetic structure dominated by

Table 7: Results of the nested clade analysis following the 2004 inference key by Posada and Templeton (RGF: restricted gene flow, RE: range expansion, FR: fragmentation, IBD: isolation by distances). Superscripts indicate both Dc and Dn statistics that were significantly small (S), large (L) or that could not be calculated (nc) because of a single occurrence. Interior haplotypes and clades are in grey.

Haplotypes			1-step			2-step		
No.	Dc	Dn	No.	Dc	Dn	No.	Dc	Dn
I	145^S	146						
II	0^{nc}	301						
IT	145^S	-156						
1-2-11Yes: RE			1-1	147^S	174^S			
III	210^L	204^L						
IV	0^S	22^S						
VI	0^{nc}	427						
VIII	122	163						
V	23^S	126						
VII	71	253						
IT	158^L	70^L						
1-2-3-4-9-10 No: FR or IBD?			1-2	147^S	188^S			
IX	276^L	275^L						
X	182^S	240^S						
XI	0^{nc}	151						
XII	0^{nc}	321						
IT	105^L	35^L						
1-2-3-4 No: IBD			1-3	267^L	307^L			
			IT	-54^S	-46^S			
			1-2-11-12 No: Contiguous RE			2-1	211	211
						2-2	0^S	97^S
						IT	210^L	114^L
						1-2-3-4 No: RGF with IBD		

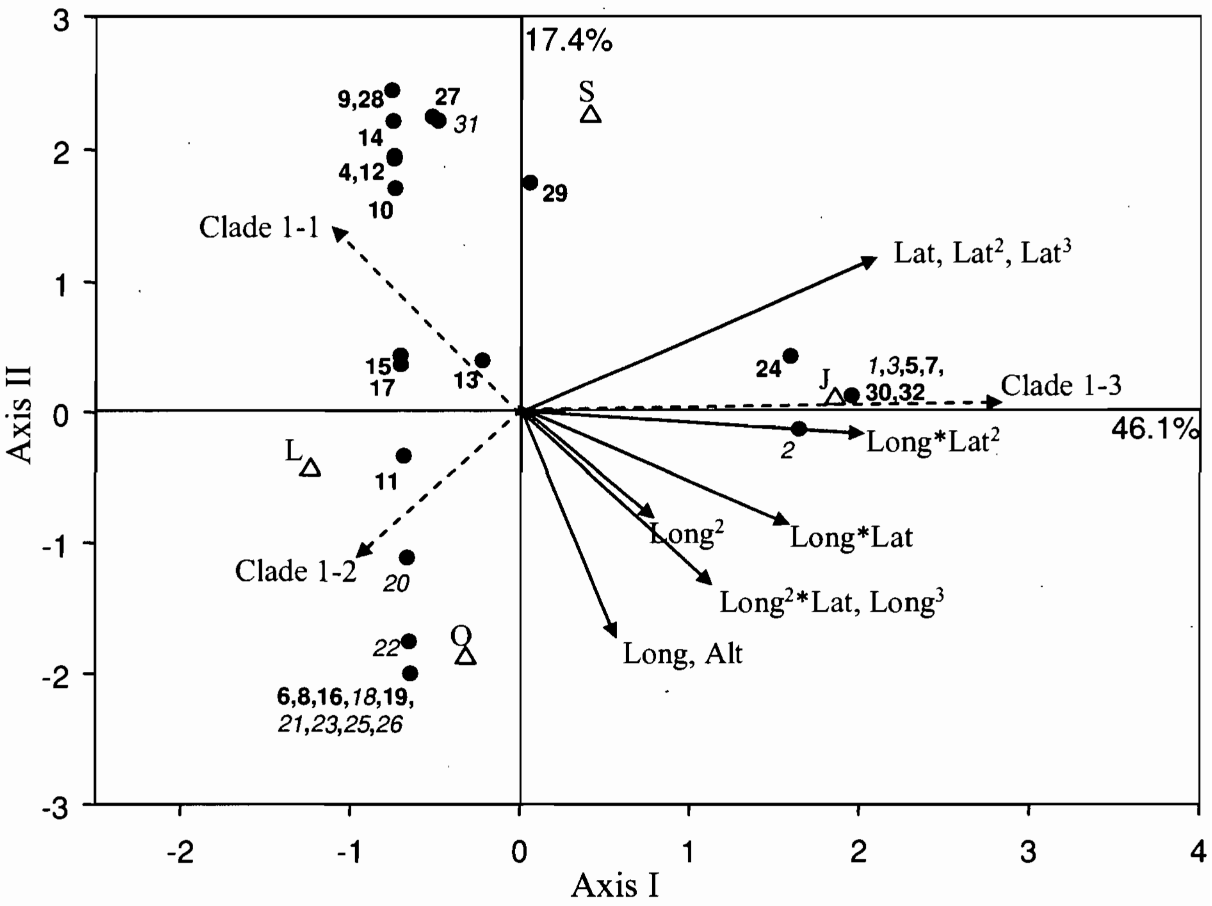


Figure 12: Canonical correspondence analysis (CCA) ordination biplot representing clades (broken arrows), sites (dots), and geographic variables (solid arrows for longitude (Long), latitude (Lat), their combinations and for altitude (Alt) and open triangles for binary variables representing Saint Lawrence (L), Outaouais (O), James Bay (J) and Saguenay (S) drainages. Site identifiers correspond respectively to numbers in Table 5. The identifiers in bold represent the sites that were putatively covered by Champlain or Laflamme postglacial seas. The proportion of total variance explained by each of the canonical axis is also represented.

a few widely dispersed haplotypes diverging from 0.5% to 2% (e.g., Turgeon and Bernatchez 2001).

Longnose dace populations are essentially composed of three closely related clades. Despite the low divergence level between them, a highly significant structure was observed. The geographic organization of this genetic structure is consistent with the idea that the contiguous range expansion was the result of more than one wave of colonists. The opposed distributions of clades 1-2 and 1-3 are consistent with parallel colonization from two distinct entries. Clade 1-2 probably entered the peninsula from the southwest connections via the Great Lakes, whereas a northern entry from the tributaries of James Bay or Ojibway–Barlow Proglacial Lake (8000 – 10 000 years ago) can be proposed for clade 1-3 (e.g., Legendre and Legendre 1984). The colonization of the northern part of the peninsula was made possible readily by headwater interconnections (Legendre and Legendre 1984) and a temporary postglacial drainage system (Gagnon and Angers 2006).

The genetic variation as well as the level of structure is of the same order as those observed on the walleye (*Stizostedion vitreum*) in the Great Lakes (Billington and Hebert 1988), on the rainbow melt (*Osmerus mordax*) in Saint Lawrence estuary (Bernatchez 1997) and on the lake whitefish (*Coregonus clupeaformis*) in the Gulf of Saint Lawrence (Bernatchez and Dodson 1990). However, it is higher than what is usually observed over the territory sampled in this work. For instance, most of the species for which data are available for the region surveyed in this study show an absence of lineage variation (Wilson and Herbert 1996), a nearly complete overlap among lineages (Bernatchez and Dodson 1991; Turgeon and Bernatchez 2001), or an unstructured mitochondrial variability (Lafontaine and Dodson 1997). This higher structure may be explained by the intrinsic characteristics of the Longnose dace, which couples large population sizes, small scale site fidelity (Thompson et al. 2001), and a geographic distribution almost unaltered by translocation. These demographic, ecological and behavioural characteristics likely contributed to its high genetic diversity and structure.

Effects of the postglacial seas

Interestingly, the geographic organization of the genetic diversity of longnose dace appeared closely associated with the maximum extents of the marine invasions of Laflamme (Saguenay) and Champlain (Saint Lawrence) seas. The dominance of clade 1-2 in the western regions and clade 1-3 on the north shore of Saguenay River likely resulted from an early colonization. However, the presence of the salted Champlain and Laflamme seas likely acted as temporary barriers to longnose dace dispersion to eastern parts of Quebec. After the retreat of these seas, new habitats appear to have been colonized by the clades already in place (1-2 in the south and 1-3 in the north) and by a new colonization wave composed essentially of clade 1-1. This hypothesis is supported by the nearly complete absence of clade 1-1 above the level of former Champlain and Laflamme seas.

The genetic diversity within populations is not related to geographical factors, including altitude. However, diversity is two times higher in populations located within the marine extents. Because these populations are not expected to have higher effective populations size, the contact zone between genetically distinct founder groups is likely responsible of this high diversity. These results are in accordance with previous studies in which the admixture of different waves of colonists resulted in higher diversity indices (Wilson and Herbert 1998; Turgeon and Bernatchez 2001; Austin et al. 2002).

The distribution of clade 1-1 throughout all regions submerged by postglacial seas and the higher intrapopulation diversity resulting from an admixture of clades are expected to affect the genetic structure. This hypothesis is consistent with the higher genetic structure observed in populations outside the postglacial seas extents (FST = 0.79) as compared with the one covered by postglacial seas (FST = 0.55). Furthermore, clade composition of populations once covered by a postglacial sea appeared more randomly distributed: populations fixed for a clade are next to those fixed for a different clade. The larger diversity associated with several founder groups made stochastic processes such as founder events and (or) genetic drift responsible for such a pattern.

Longnose dace origins

The wide distribution of longnose dace suggests the existence of various refuges that allowed the survival of this species during the Pleistocene glaciations (Radforth 1944; McPhail and Lindsey 1970). Among them, three can be proposed as sources of the colonization of the northeastern part of North America, namely the Atlantic, the Missourian, and the Mississipian (Crossman and McAllister 1986). However, as lineage 2-1 is clearly dominant in Quebec populations, the contribution of a single refuge may appear as a relevant hypothesis. Assuming 1%–2% divergence per million years (Brown et al. 1979; Wilson et al. 1985), a separation time within the last 75 000 or 150 000 years is expected among clades of this lineage and therefore likely occurred before or during the last glaciation event (75 000 to 10 000 BP). The presence of a few individuals that shared haplotypes from the very divergent lineage 2-2 in a single population seems to be anecdotal; unintentional translocations by anglers may have occurred.

The low sampling effort in populations located in regions covered by the glacial refuges makes any proposal of the refugial origin of lineage 2-1 rather speculative. However, evidence suggests that this lineage has a Mississippian origin. First, the level of divergence (4.5%) observed between this lineage and the outgroup from Eight-Mile River is expected to be the result of approximately 2 to 4 million years of restricted gene flow. This time lag largely predates the Pleistocene glaciations and suggests different refugial origins. According to the geographical location of Eight-Miles River, lineage 2-2 likely originated from the Atlantic refuge. Second, colonization entries via the Great Lakes or the tributaries of James Bay, as suggested by the spatial organization of mitochondrial diversity observed in this study, are known to be directly connected to the Mississipian refuge (Prest 1970). Third, the importance of this refuge for the postglacial colonization of the north eastern part of North America has been demonstrated in both biogeographical (Bailey and Smith 1981; Crossman and McAllister 1986; Underhill 1986) and phylogeographic (Lu et al. 2001; Gagnon and Angers 2006) studies.

If the effective contribution of only one refuge is accurate (despite the evidence of the presence of longnose dace in all Atlantic, Mississippian, and Missourian refuges), the phylogeographic signal of this species highly contrast with the one of other fish species. For instance, current Quebec lake trout (*Salvelinus namaycush*) populations were established with contributions from both the Atlantic and Mississipian refuges (Wilson and Hebert 1996). A similar pattern was observed for lake cisco (*Coregonus artedii*; Turgeon and Bernatchez 2001), brook char (*Salvelinus fontinalis*; Angers and Bernatchez 1998; Danzmann et al. 1998) and walleye (Billington and Hebert 1988) These comparisons lead to the conclusion that the dispersal connections allowing colonization of Quebec by salt-tolerant species have not been used by the longnose dace. As a result, the barrier caused by marine invasions can be proposed to explain the observed difference in colonization patterns. This hypothesis is especially relevant in the case of the Atlantic refuge, as southeast connections to the peninsula were temporary blocked by the Champlain Sea and the Gulf of Saint Lawrence.

Postglacial seas appeared to have played a major role in the spatial genetic structure of longnose dace populations. First, a lower genetic structure and a higher diversity within the marine extents were due to the presence of a different founder group restricted to this region. Second, acting as barriers to colonization, the marine extents are likely responsible for the single refuge origin of the longnose dace populations in Quebec. These results present a strong argument for integrating this phenomenon into further phylogeographic studies of freshwater organisms in northeastern North America. We also believe that an increased examination of Cyprinidae species would greatly improve our understanding of fish postglacial dispersal. Examinations of such species will no doubt increase the precision of a general hypothetical model describing the spatial genetic structure of fishes established after the Pleistocene glaciations.

CHAPITRE IV

POST (POpulation STructure): A new method to depict complex genetic organization among populations without *a priori* clustering hypothesis.

Résumé

La description de la structure des populations est d'une importance cruciale pour permettre une interprétation non biaisée de la diversité génétique des populations. Les méthodes existantes permettant de faire ces inférences nécessitent généralement des connaissances *a priori* du système sous étude et/ou ne tiennent pas compte de la possibilité qu'une population puisse avoir une origine multiple. Dans cette étude, nous proposons une procédure statistique appelée POST (pour STructure des POpulations) qui permet d'identifier les groupes de populations génétiquement similaires en plus d'estimer leur contribution dans le bagage génétique de populations où ces groupes se retrouvent mélangés. Cette méthode utilise le raisonnement à la base de l'analyse hiérarchique de la covariance génétique. Les performances de la méthode ont été évaluées par simulations ainsi qu'à l'aide d'un jeu de données dont la structure des populations était connue. Les résultats ont montré la capacité de la méthode pour correctement décrire des structures complexes, telles que celles provenant d'événements d'introgression, d'hybridation ou d'introduction, ainsi que les singletons. Combinant simplicité et objectivité, POST est une procédure avantageuse pouvant permettre l'identification des unités évolutives distinctes qui constituent la base de plusieurs plans de conservation.

Abstract

The description of population structure is of prime importance for unbiased interpretations of the population diversity. Most of the existing methods used for such inferences necessitate a priori hypotheses on the system under study and/or do not take into account populations with multiple origins. In this paper, we describe a new procedure called POST (for Population Structure) that aims at depicting the genetic organization among populations without a priori clustering hypotheses. The procedure identifies groups of closely related populations and estimates their contribution for populations of multiple origins. The procedure uses the basic ideas behind the hierarchic analysis of the genetic covariance. Its performances are evaluated using different simulated scenarios of basic and complex structure as well as a real dataset for which the population structure was previously inferred. These results showed the capacity of POST to correctly depicted complex population structure, including introgression, hybridization and introduction events, and singleton. Because it is simple, efficient and completely objective, our procedure appears to be suitable method to identify the evolutionary significant units which form the bases of many conservation procedures.

Introduction

Individuals of a species are generally organized in space (or time) following a hierarchical pattern. Individuals preferably mate with those which are geographically close, reproduce during the same period or live in similar habitats, leading to panmictic populations. Similarly, a population exchanges more migrants with spatially close populations than with those which are distant or geographically isolated. Such a pattern of gene flow leads to the formation of groups of populations with more or less different genetic backgrounds.

The description of these population structures is of prime importance to make unbiased interpretation of the processes responsible for population diversity. Indeed, all biological differences among populations will reflect at least partly this genetic structure (Excoffier 2001). The resulting lack of independence among observations will prevent the use of statistical analysis that necessitates the complete independence between objects. The structure component must then be taken into account in models describing the genetic variability among population

One possible approach allowing the analysis of genetic structure is the hierarchical analysis of the genetic covariance (Cockerham 1969, Cockerham 1973, Weir and Cockerham 1984). This framework consists in partitioning the total covariance of the allelic frequencies (F_{ST}) into among (F_{CT}) and within (F_{SC}) sub-unit components. This method can also be used to test *a priori* hypotheses about factors leading to population structure such as geographic barriers to migration (e.g. Lucchini et al. 2004, Ross and Markow 2006), pattern of colonization (e.g. Tsuruta and Goto 2006, Whiteley et al. 2006) and kin selection (e.g. Behrmann-Godel et al. 2006, Selkoe et al. 2006).

However, an abundant literature reported that population structure is often not straightforward, such as structure that results from historical processes rather than current factors (Bernatchez and Dodson 1991, Angers and Bernatchez 1998, Turgeon and Bernatchez 2001, Gagnon and Angers 2006). If populations are not clearly identified, individual-based approaches (Guillot et al. 2005, Pritchard et al. 2000) may be used to identify panmictic populations of genetically similar individuals. However, none of these methods have been specifically designed to identify groups of predefined panmictic populations. In this particular case, Dupanloup et al. (2002) proposed the method SAMOVA (for Spatial Analysis of Molecular Variance). For a predetermined number of groups, their approach uses a spatially constrained heuristic algorithm to find the solution that maximizes the genetic and the geographic homogeneity within population groups. Even though this procedure can be very effective to detect spatial structures, the authors acknowledge that the geographic constraint implemented in their algorithm decreases the performance of SAMOVA to identify complex genetic organization, such as fragmented distribution of genetically similar populations (Dupanloup et al. 2002).

Another major complication when inferring structure comes from the numerous processes leading to admixtures between evolutionary groups. For instance, contact zones between genetically distinct founder groups (Richardson et al. 2002, Walter and Epperson 2004, Girard and Angers 2006), hybridization (Alexandrino et al. 2006, Nittinger et al. 2007), coexistence of distinct groups in a sympatry (Arnegard et al. 2005, Drummond and Hamilton 2007), isolation by distance (Worley et al. 2004, Clauss and Mitchell-Olds 2006) or ecological clines (Berry and Kreitman 1993, Sotka et al. 2004, Toju and Sota 2006) result in populations associated to admixed genetic backgrounds.

In this paper, we propose a procedure called POST (for POpulation STructure) that identifies groups of closely related populations and populations of multiple origins. This procedure uses the idea behind the analysis of genetic variance initially defined by Cockerham (1969, 1973) to identify groups of genetically similar populations as well as group admixture. Below, we describe in detail this procedure and evaluate its performances on both simulated and real datasets.

Materials and Methods

The objectives of POST are twofold. It aims at i) defining the number of evolutionary distinct groups of populations and ii) identifying the populations for which genetic background resulted from admixture between them. The method follows an iterative step-by-step procedure, where each iteration aims at identifying one of the four basic structures:

1) the populations belong to a single group,

2) the populations belong to two distinct groups and no admixture is present,

3) the populations belong to two groups that are admixed in some populations,

4) the populations belong to more than two distinct groups.

The iterative process allows the progressive decomposition of any set of populations to these simple structures.

The procedure will be described in two parts. First, we will describe the reasoning of POST and we will explain how these four structures can be discriminated. Second, we will provide the details of the iterative procedure allowing the decomposition of a population structure.

POST rational

Partitioning a subset of objects consists in finding an optimal solution for which objects of a given cluster are more similar to each others than to objects of the other clusters. This problem can easily be transposed to a genetic structure perspective. Quantitatively, the genetic structure of populations corresponds to the partition solution that maximizes the genetic differentiation among clusters and minimizes the variation within clusters. This optimal solution can be found by evaluating different combinations of populations with a heuristic algorithm and by selecting the solution with the higher differentiation among clusters.

Let us start with a set of populations which are forced to be grouped in an optimal solution of three clusters. Except in the theoretical case for which the three clusters are equidistant, if the clustering is performed within a Euclidean space, the solution will always be composed of two distant clusters (hereafter referred as the reference clusters) and one intermediate cluster (Fig. 13a-d).

Using these three clusters, it is possible to identify the structure within a set of populations by comparing the populations of each cluster. This comparison is based upon the F_{CT} which represents the portion of the total genetic variability observed among clusters. We propose a statistic called maxF which represents the maximum F_{CT} value when a population is placed into one or the other reference clusters:

$$\text{maxF} = \max(F_{CT1}, F_{CT2})$$

$$FCT1 = \frac{\sigma_{a1}^2}{\sigma_T^2}$$

$$FCT2 = \frac{\sigma_{a2}^2}{\sigma_T^2}$$

where: $\sigma_{a1}{}^2$ = the covariance component due to differences among the two reference clusters if the population is placed in reference cluster 1.

$\sigma_{a2}{}^2$ = the covariance component due to differences among the two reference clusters if the population is placed in reference cluster 2.

$\sigma_T{}^2$ = the total genetic covariance

According to this statistic, the higher maxF is expected when a sample genetically close to samples within a given reference cluster (hereafter referred as a pure sample) is placed into this cluster. At the opposite, the lower maxF is expected when a sample is distinct from those present into any of the reference clusters (hereafter referred as an external sample), as such a sample will strongly inflate the within-group component of the genetic variability (and decrease F_{CT} as a consequence). An admixed sample is also expected to provide a low F_{CT} value but to a lesser extent than an external sample due to the contribution of reference groups into such a sample. As a result, maxF will be intermediate.

The low (external population), intermediate (admixed population) and high maxF (pure population) ranges are determined with maxF distributions obtained from pure and admixed allelic pool of reference clusters. Two simulated sets are obtained by random sampling of alleles from the reference clusters. A first set is composed of 100 simulated samples from "pure" populations. Each sample is constructed by drawing 2N alleles with replacement (N being the average sample size of the population set) from only one reference cluster (Figure 14). Half of these samples are thus sampled in the first reference cluster and the other half in the second reference cluster. The second set is composed of 100 simulated samples from 50/50 admixtures of reference clusters. Such samples were constructed by sampling 2N alleles with replacement in a pool composed of equal proportions of alleles from both reference clusters (Figure 14). If the reference clusters do not share the same effective (for instance 100 and 50 individuals in reference 1 and 2 respectively), the difference between them (50 individuals or 100 alleles) is filled by sampling alleles with replacement within the reference cluster that has the lowest effective (100 alleles within reference cluster 2).

For each of these simulated samples, the maxF is determined and used to construct the maxF distribution for samples from pure populations (referred hereafter as the Pure Distribution, PD) and from admixed populations (referred hereafter as the Mixed Distribution, MD) (Figure 14). According to a threshold set by the user, three critical values (e.g. 1%, 5%, etc) are computed: α_1 and α_2 are respectively the lower and upper limits of the MD, while α_3 is the lower limit of the PD. These thresholds are then used to determine which of the four basic structures is representative of this set of populations.

1) The populations belong to a single group. If the reference clusters are genetically similar and form a single group, the pure and admixed simulated samples from reference clusters are expected to display comparable maxF values. Accordingly the PD and the MD will overlap ($\alpha_2 > \alpha_3$; Fig 13e).

2) The populations belong to two distinct groups and no admixture is present. If two groups or more are present in the populations set (rules 2, 3 and 4), the PD and the MD are not expected to overlap ($\alpha_2 < \alpha_3$). MaxF is then computed for all the samples composing the intermediate cluster. Those that display maxF $\geq$ $\alpha3$ are expected to belong to a population from one or the other of the reference clusters (Fig.13f).

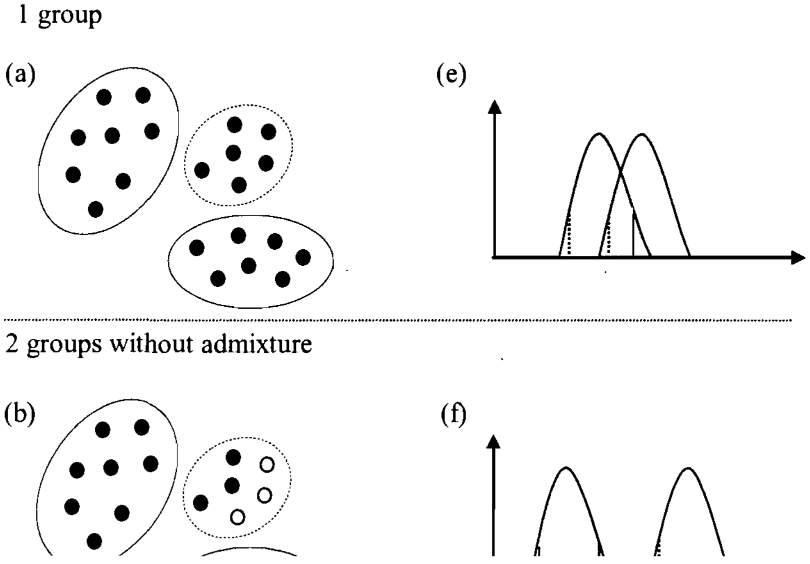
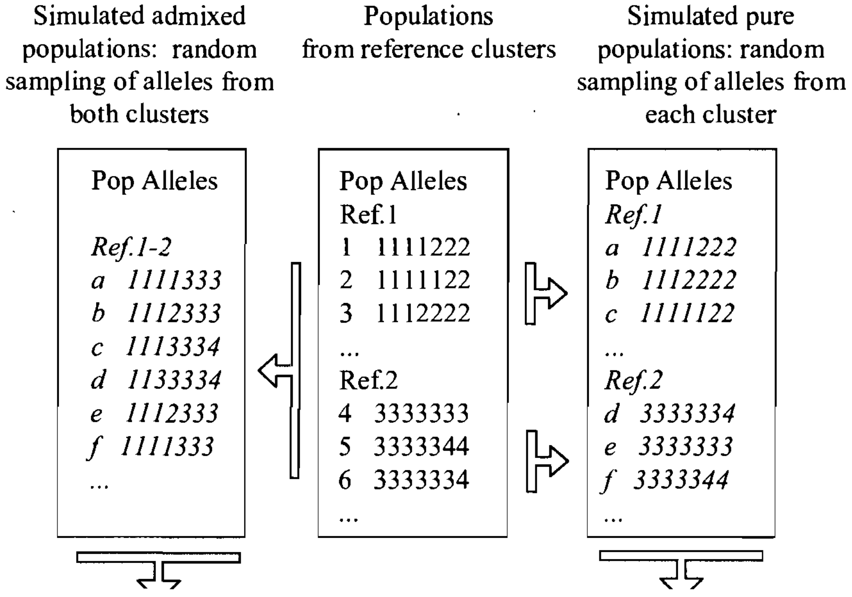


Figure 13: Expected K-means and maxF distributions for each basic structure. K-means partition of the populations into three clusters (a, b, c and d) and the maxF distributions (e, f, g and h) for each basic structure (see text). The mixed (MD: grey curve) and pure (PD: white curve) maxF distributions are presented as well as the three threshold that delineated the ranges associated to external, admixed and pure maxF. When more than a single group is present, the range of the maxF calculated for populations of the intermediate cluster is represented by a grey arrow.



Simulated admixed populations: random sampling of alleles from both clusters

Pop Alleles

Ref.1-2
a 1111333
b 1112333
c 1113334
d 1133334
e 1112333
f 1111333
...

Populations from reference clusters

Pop Alleles
Ref.1
1 1111222
2 1111122
3 1112222
...
Ref.2
4 3333333
5 3333344
6 3333334
...

Simulated pure populations: random sampling of alleles from each cluster

Pop Alleles
Ref.1
a 1111222
b 1112222
c 1111122
...
Ref.2
d 3333334
e 3333333
f 3333344
...

Figure 14: A schematic view of the random allelic sampling leading to the construction of the mixed (MD) and pure (PD) maxF distributions at each POST iteration. The thresholds that delimited external, mixed and pure maxF range are then calculated.

3) The populations belong to two groups and a subset of populations is admixed. The samples of the intermediate cluster that display a maxF located between α_1 and α_3 (Fig.13g) are expected to come from an admixture between the reference clusters.

4) The populations belong to three or more distinct groups. The samples of the intermediate group that display a maxF significantly below those obtained in the MD ($\leq\alpha_1$; Fig.13h) are expected to have no contribution from any of the reference clusters. These samples are then expected to belong to an external population.

Iterative procedure

Below, we provide the details of the iterative procedure and how to interpret the outputs describing the genetic structure of a subset of populations. The following details are also presented in Fig. 15.

Step 1: Clustering the populations

The populations are first separated in three clusters with a K-means procedure. The K-means is performed upon a pairwise distance matrix (the Chord distance D_{CE} from Cavalli-Sforza and Edwards 1967 was used in this work) using the distance-based approach proposed by Hathaway et al. (1989). The Chord distance is fully embeddable in Euclidean space. The final solution is defined as the one which maximizes the sum of squares of the distances among group centroids. The reference clusters are identified as those for which the distance between centroids is the higher (the higher genetic differentiation). The third cluster corresponds to the intermediate one.

Step 2: Distributions of maxF

The MD and the PD are constructed and α_1, α_2 and α_3 are evaluated. If the MD and the PD are overlapping ($\alpha_2>\alpha_3$), reference clusters are expected to belong to a single

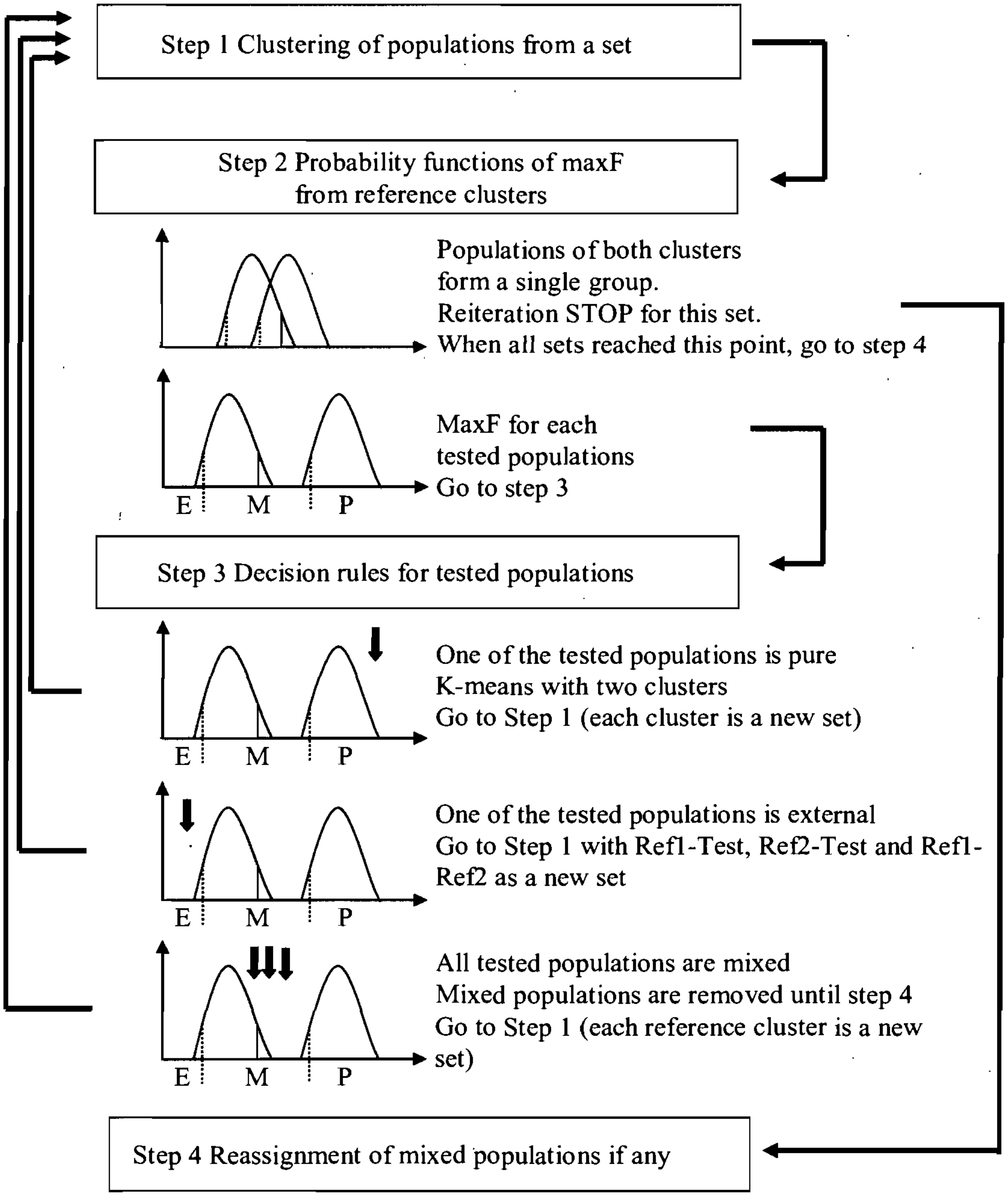

Step 1 Clustering of populations from a set
Step 2 Probability functions of maxF from reference clusters
Populations of both clusters form a single group.
Reiteration STOP for this set.
When all sets reached this point, go to step 4
MaxF for each tested populations
Go to step 3
E
M
P
Step 3 Decision rules for tested populations
One of the tested populations is pure
K-means with two clusters
Go to Step 1 (each cluster is a new set)
E
M
P
One of the tested populations is external
Go to Step 1 with Ref1-Test, Ref2-Test and Ref1-Ref2 as a new set
E
M
P
All tested populations are mixed
Mixed populations are removed until step 4
Go to Step 1 (each reference cluster is a new set)
E
M
P
Step 4 Reassignment of mixed populations if any

Figure 15: The details of the iterative procedure along with the alternative results that can be obtained at steps 2 and 3. For each expected result, the maxF distribution (MD: grey curve; PD: white curve), maxF ranges (external (E), mixed (M) and pure (P)) and the maxF computed from the samples of the intermediate cluster (black arrows) are shown. The dataset treatment associated to each result and the following steps to be performed are explained.

group and the procedure stops for this set of populations. Otherwise, the process continues on step 3.

Step 3: decision rules for samples of the intermediate cluster

If the two distributions do not overlap ($\alpha_2<\alpha_3$), the maxF is computed for every samples of the intermediate cluster. According to the critical values, they are considered as pure, admixed, or external. Three different outputs are then expected, each followed by a particular treatment.

1) At least one sample is pure. The simulation study (see results) shows that POST will infer presence of some admixed samples even in the absence of admixture. Consequently, the general rule is that if at least one sample of the intermediate cluster is identified as pure, then the entire set of populations represents two groups without admixture.

Treatment. A new K-means analysis is performed using two clusters instead of three. Each cluster is then considered as a new set of populations and the procedure is performed anew from step 1 upon each of these sets.

2) All samples are admixed. Because of their membership to more than one group, admixed populations may introduce errors in the number of group (they may be considered as a different group) and the genetic characteristics of each group (i.e. if an admixed sample is considered as pure).

Treatment. The admixed samples are removed from the dataset and the procedure is performed anew from step 1 upon each reference group as a new set of populations. The admixed samples are further analyzed at step 4

3) At least one sample is external. As was observed for pure samples, POST may infer presence of some admixed samples even in the absence of admixture. Consequently, the general rule is that if at least one sample of the intermediate cluster is identified as external, at least three groups are present in the set of populations.

Treatment. The procedure is performed anew upon all pairs of clusters as a new set of populations. The first pair to be re-analyzed combines the two reference clusters; the admixed samples (if any) are removed. The second pair combines the intermediate cluster with the most distant reference cluster (excluding the admixed samples identified previously); again, the identified admixed samples (if any) are removed. Finally, the last pair is analyzed and the admixed samples (if any) are removed. If there are three groups, each of these pairs is expected to provide two groups (with or without admixture); an additional iteration will be required before the procedure stops.

Steps 1 to 3 are performed until each set of populations is considered as a single group (when the PD and the MD are overlapping) or if the set is composed of fewer than three populations (and thus cannot be distributed among three clusters). When iterations are necessary to depict a population structure, the analysis scheme of POST is hierarchic and can be seen as a tree in which each branch ends with a single pure group (Fig. 16). In this case, we recommend to take one branch at a time and to reach its end before taking another branch. In other words, according to figure 16, the subsets should be analysed following the order 1, 1a, 1a', etc. instead of 1, 2, 1a, 1b, etc. This difference is important since admixed samples are removed throughout the analyses. Consequently, results at each step influence the dataset to be used during the next ones. The choice of the vertical way of analysis was made to reduce biases related to admixed samples and to use the purest group possible in the further steps of the population structure analysis.

Step 4: reassignment of admixed samples

The final step is to determine the contribution of each pure group in the admixed samples removed throughout the procedure. A population resulting from a historical admixture between two pure groups (contact zone, ecological cline) is composed of individuals with genotypes resulting from a random draw of alleles from the two parental

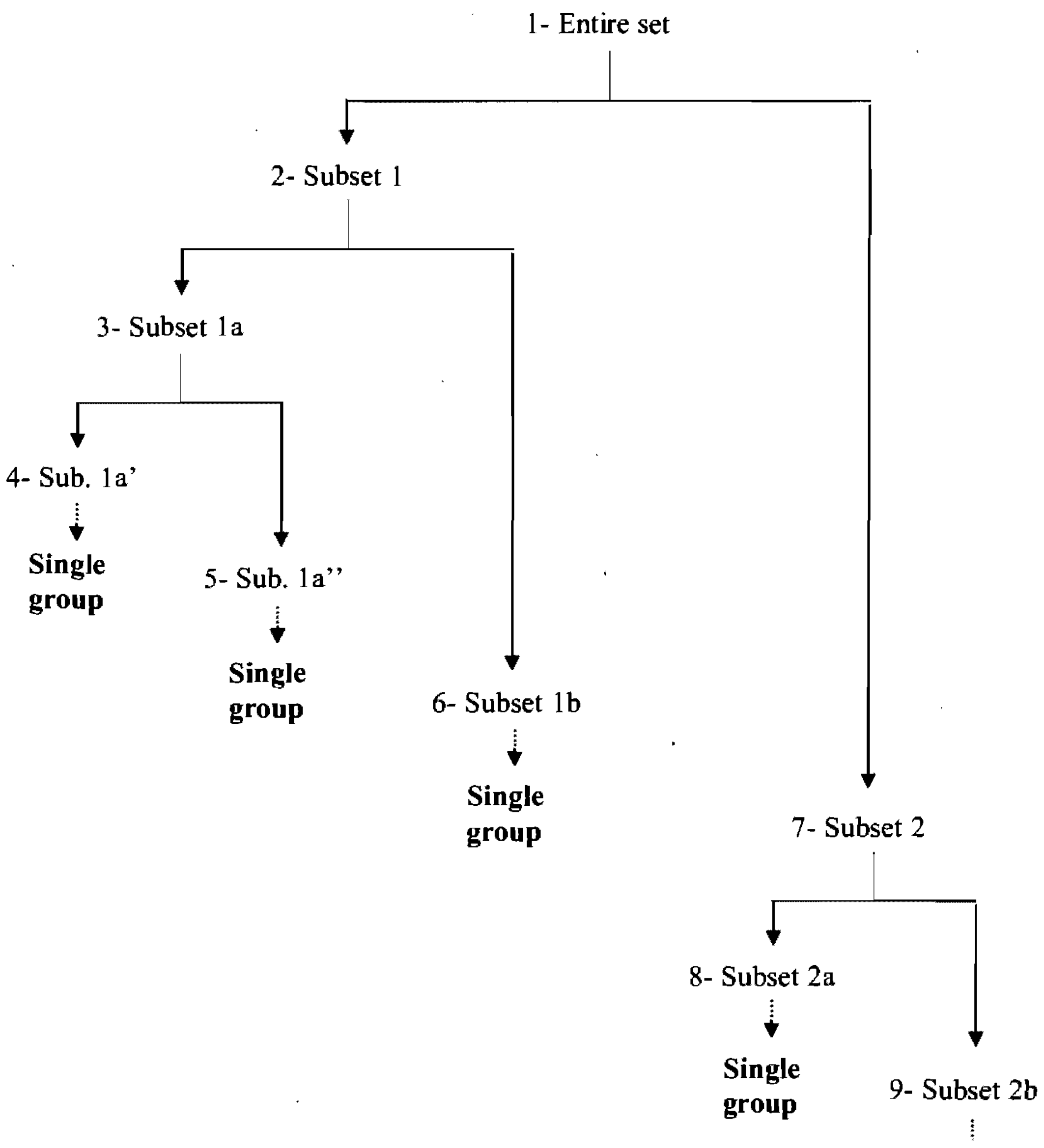


Figure 16: The hierarchic scheme of POST when several iterations are necessary to depict the structure of the population set. Starting from the entire population set, POST sequentially fragments and reduces the population set to be analyzed. The entire procedure takes the form of a tree in which each branch ends with a single group. In this example, the decomposition of the population structure necessitates nine iterations.

groups. However, an admixed sample might also result from a random sampling of individuals of two distinct populations living in sympatry. POST is not able to discriminate these two types of admixture. In this perspective, we propose two types of analyses for admixed sample reassignment. The first is the mixed-stock analysis. This kind of method appears to be more suitable in the context of historic admixture where individuals represent random allelic combination of the parental populations. The pure groups thus represent the parental stocks for which the relative contributions are computed for each admixed population.

The alternative might be the individual-based assignment methods. These methods allow the quantification of the probabilities of origin of a given individual to each parental group. To transform these individual probabilities into pure group contributions we propose the following simple computation. For a given admixed population, the individual assignment probabilities (>0.05) of each pure group are summed over all individuals. This sum is then divided by the number of individuals, thus given the average probability of each pure group. These averages are then interpreted as estimations of the contributions of the pure groups to the compositions of the admixed populations.

POST application

For both simulated and real datasets described below, a pairwise D_{CE} matrix was computed using the software POPULATION v.1.2.28 (Langella 1999). The K-means clustering and the evaluation of the distance between group centroids were performed with the GINKGO Multivariate Analysis System v.1.5.4 (Bouxin 2005) using 50000 starting random seeds. The K-means result and the allelic composition of each population were used as inputs within an R function that performed steps 2 and 3 of the procedure. This function (available from the corresponding author upon request) was coded using features included in the *hierfstat* (Goudet 2005) and *genetics* (Warnes and Leisch 2005) R libraries. PD and MD were constructed using 100 simulated populations. Values of 0.1, 0.99 and 0.01 were set respectively for $\alpha1$, $\alpha2$ and $\alpha3$. This combination of high ($\alpha1$) and low ($\alpha3$) thresholds was chosen to shorten the maxF interval associated to mixed populations and thus to increase the power to detect external and pure populations. Finally, the contribution of groups to admixed populations was assessed using both the individual reassignment and the admixed stock analysis methods. First the reassignment method proposed by Rannala and Mountain (1997) available in GENECLASS2 (Piry et al. 2004) was used. Using a Bayesian approach, their method evaluates the probabilities to obtain the genotype to be reassigned by creating random genotypes from the allele of each sample. Second, an admixed stock analysis using the coalescence-based admixture estimator developed by Dupanloup et Bertorelle (2001) was performed with the program ADMIX 2.0 (Dupanloup and Bertorelle 2001).

Simulated datasets

We performed preliminary simulations to attain two objectives. First, the capacity of POST to correctly identify each of the four basic structures was assessed using different conditions (number of individuals per population, differentiation between reference groups and number of loci). These simulations included the case where a single group of populations was present in the data; this situation allowed us to evaluate the type I error rate of the method. Second, the iterative procedure was evaluated using a simulated complex genetic structure composed of three pure groups and two distinct groups of admixed population. Below, we will describe the details of each of these two sets of simulations.

Simulations of the four basic structures

The basic structures were simulated using the simulation program EASYPOP v.2.0.1 (Balloux 2001). Populations were composed of 1000 diploid individuals. Genetic data were obtained from five or ten unlinked loci with a mutation rate of 10^{-5} following a strict stepwise mutation model as expected for an ideal microsatellite locus (Di Rienzo et al. 1994). The maximum of allelic states was fixed to 50. These states were randomly distributed among the genotypes of the first generation in each population. Population structures were simulated using a hierarchical island migration (HIM) model. HIM is an extension of the classical island model, for which exchanges between populations within and among islands have their own rate. In all simulations, the migration rate within each island was set to 0.1, while the rate was set to 0 among islands, resulting in completely discrete and homogenous groups of populations. Two successive HIM models were necessary to simulated overlaps between groups. Accordingly, a first HIM was set for either 9900 or 900 generations (Figure 17, step 1) and an additional island was defined in-between the previous ones during 100 generations using the same migration rates (Figure 17, step 2) resulting in islands composed of either pure, mixed with different proportions of the reference groups (50/50, 60/40, 70/30 and 80/20) and external populations. Based on different combinations of these islands, the four basic structures were simulated. A random sampling of 15 or 30 individuals was performed in every populations of a given simulation and POST analysis was performed. The single-group structure (F_{CT} between reference groups fixed to 0) was analysed with four different sets of conditions, (15 or 30 individuals, 5 or 10 loci). Other basic structures were analyzed with eight different sets of conditions (F_{CT} between reference groups of 0.2 or 0.02, 15 or 30 individuals, 5 or 10 loci).

Complex genetic structure

The complex genetic structure was simulated following the same baselines as for the basic structures. A dataset composed of 36 populations of 1000 diploid individuals per population and five loci was simulated using EASYPOP. The molecular parameters (mutation rate, mutation model and maximum number of allelic state) were the same as above. Three pure groups (A, B and C) and two admixed groups (AB and BC) were created using two sequential HIM (9900 and 100 generations for the first and second HIM respectively). Each admixed and pure group was composed of 6 and 8 populations respectively. Admixed populations were composed of a 70-30 proportion mix. Group B was underrepresented in both cases. 30 individuals were randomly sampled from each population.

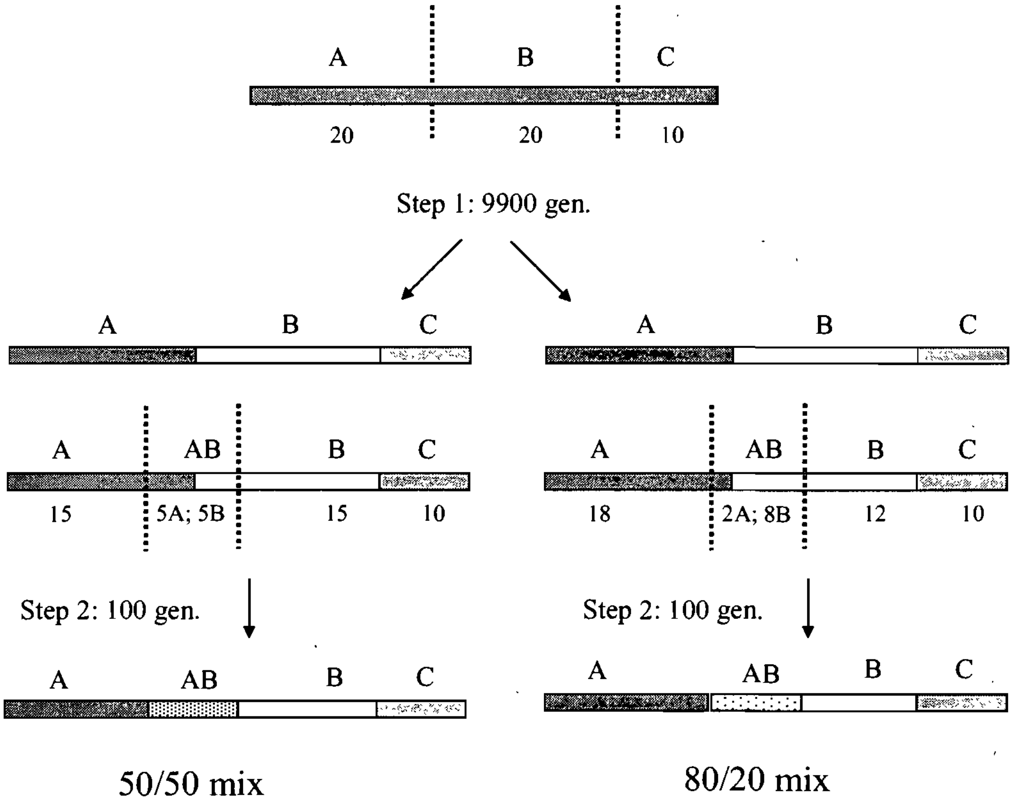


Figure 17: Simulation steps. Step 1: The populations set is divided in three islands and a hierarchical island model (HIM) is run for 9900 generations leading to three distinct evolutionary units (A: 20 pop., B: 20 pop. and C: 10 pop). Step 2: The resulting set is divided in four islands allowing exchange of migrants between some populations from A and B, and a second HIM is run for 100 generations. The final set is composed of pure (A and B), mixed (AB) or external (C) population groups. According to the boundaries of each island at step 2, mixed groups composed of different proportion of each reference group are created. Here, examples of 50/50 and 80/20 mixes are shown.

Real dataset

As an example of highly structured set of populations with overlapping between groups, the genetic data of brook charr (*Salvelinus fontinalis* Mitchill) populations from La Mauricie National Park (Quebec, Canada) reported by Angers and Bernatchez (1998) was used. The dataset includes a total of 779 individuals, collected from 26 lakes located in seven different drainages (Figure 20) and analyzed with five microsatellite loci. Further information can be found in Angers and Bernatchez (1998) and Angers et al. (1999).

Results

Simulations of the four basic structures

The efficiency of POST to detect a single group of populations was first assessed. The results of these analyses show the reliability of POST to discriminate a single group of populations from structures composed of at least two clusters (Table 8; 4 bottom lines). For simulations performed with a single group of populations, the MD and the PD are not significantly different and partially overlap ($\alpha_2 > \alpha_3$), while they are disjointed when at least two groups are present ($\alpha_2 < \alpha_3$). POST appears to be sufficiently powerful to detect the presence of two groups for F_{CT} values as low as 0.02 (Table 8; 4 middle lines).

For the basic structures composed of at least two groups, the proportion of populations within the intermediate cluster that were correctly identified (as pure, mixed or external depending on which structure is analyzed) is shown in Table 8. The results show that POST is a conservative procedure in that it infers the presence of some admixed populations even in the absence of admixture. For instance, when the structure is composed of three groups (external) or two groups without overlap (pure), a certain proportion (between 0.1 to up to 0.9) of the populations composing the intermediate

Table 8: Results obtained by POST following simulations of the four basic structures. The F_{CT} statistic (F_{CT}) between the reference groups, number of individuals in each sample (ind) and the number of loci used are presented for each simulation (loci). Eight simulations were performed for each of the following structure: more than two groups (external), two groups with admixture (mixed 50/50, 60/40, 70/30 and 80/20) and two groups without admixture (pure). The number represented the proportion of maxF computed on samples of the intermediate cluster that fell within the expected maxF range. Finally the basic structure described by a single group (FCT between reference group of 0) was analysed four times.

FCT	Ind	Locus	Single	External	Mixed				Pure
					50/50	60/40	70/30	80/20	
0.2	15	5	$\alpha_2<\alpha_3$ (no)	0.9	0.9	1.0	1.0	1.0	0.5
		10	$\alpha_2<\alpha_3$ (no)	0.9	0.9	1.0	1.0	1.0	0.8
	30	5	$\alpha_2<\alpha_3$ (no)	0.9	0.8	1.0	1.0	1.0	0.4
		10	$\alpha_2<\alpha_3$ (no)	0.9	0.9	1.0	1.0	1.0	0.5
0.02	15	5	$\alpha_2<\alpha_3$ (no)	0.4	0.7	1.0	1.0	1.0	0.1
		10	$\alpha_2<\alpha_3$ (no)	0.2	0.8	1.0	1.0	1.0	0.1
	30	5	$\alpha_2<\alpha_3$ (no)	0.2	0.9	1.0	1.0	1.0	0.1
		10	$\alpha_2<\alpha_3$ (no)	0.6	0.8	1.0	1.0	1.0	0.1
0	15	5	$\alpha_2>\alpha_3$ (yes)	--	--	--	--	--	--
		10	$\alpha_2>\alpha_3$ (yes)	--	--	--	--	--	--
	30	5	$\alpha_2>\alpha_3$ (yes)	--	--	--	--	--	--
		10	$\alpha_2>\alpha_3$ (yes)	--	--	--	--	--	--

groups have a maxF located within the admixed range. However, because only one pure or external population is necessary (according to the rule stated previously) to diagnose an absence of overlap or the presence of a third independent group, the good structure is recover in all conditions.

Surprisingly, POST does not appear to be influenced by either the sample size or the number of loci. On the other hand, the reliability is clearly related to the level of differentiation between reference groups. For instance, both external and pure populations are correctly diagnosed in a higher proportion when $F_{CT} = 0.2$ than when $F_{CT} = 0.02$ (Table 8). The simulations did not include variation in the number of populations per group; that number was 10 in all cases. The simulation results also reveal the high capacity of POST to correctly identify mixed populations when present. The procedure is especially reliable when mixed populations are composed from unequal fractions of the reference groups (Table 8). Indeed, in these circumstances, mixed populations are all correctly diagnosed. However, the results are more mitigated for a perfect 50-50 mixture, for which a given proportion (between 10 and 30%) of the populations are misidentified and consider as external. This result was clearly associated to the high α_l threshold. In these circumstances, according to the stated rules, these two groups with overlap will be misinterpreted as sets of populations with more than two groups.

Complex genetic structure

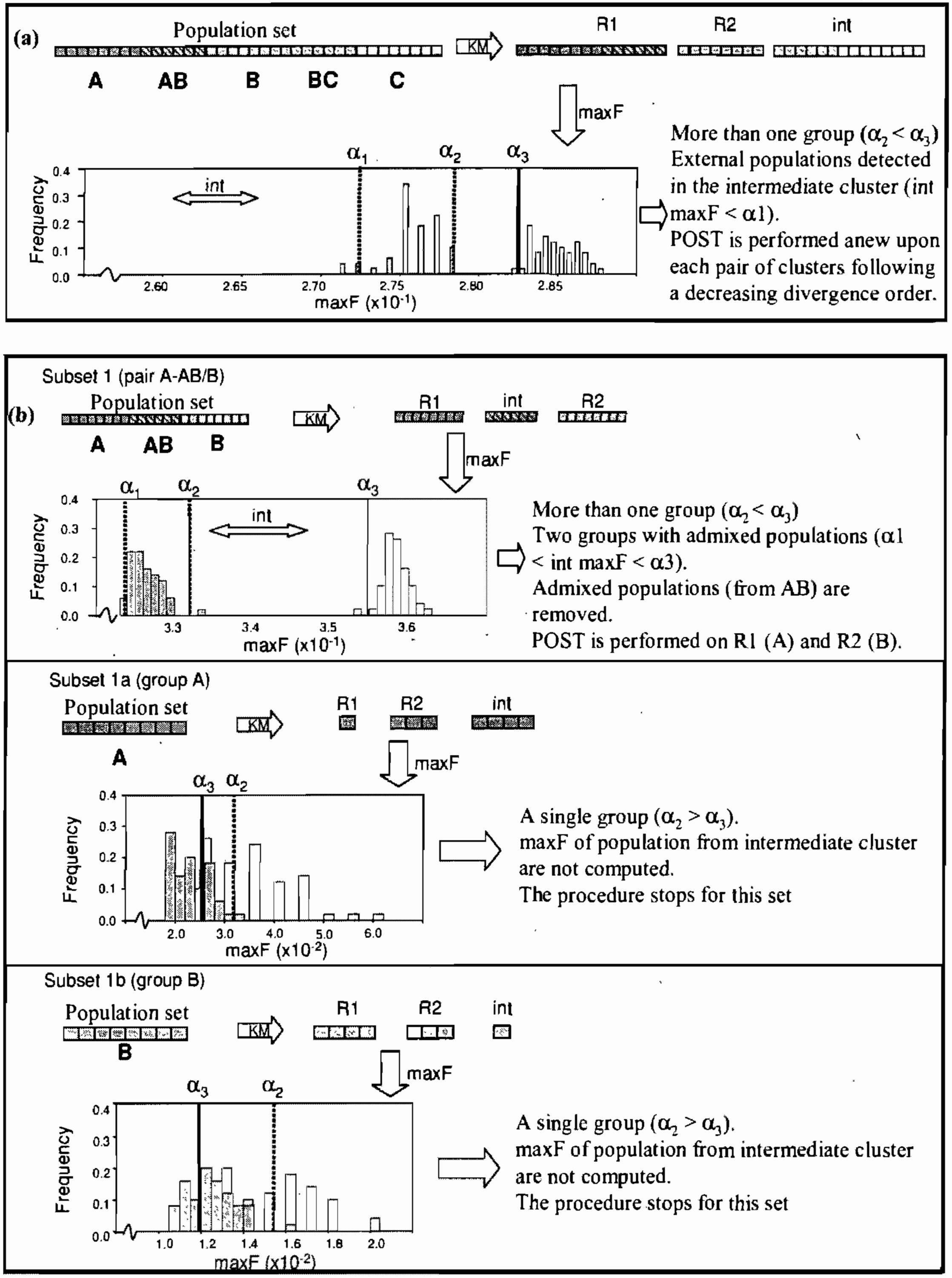
POST has necessitated seven iterations to completely decompose the complex structure. The complete description of the results at the first iteration is shown in Figure 18a. First, the whole set of populations was used. The three clusters delineated by the K-means procedure were composed of (A-AB), (B) and (BC-C) for the reference group 1 (R1), reference group 2 (R2) and the intermediate group (int) respectively. The MD and the PD distributions did not overlap indicating that more than one group are present. The maxF computed on populations of the intermediate cluster all felt within the external range ($<\alpha_l$). Accordingly, additional iterations were performed anew within subsets combining populations of each pair of reference and intermediate groups.

The first subset combined (A-AB) and (B). Following the K-means clustering (A) (AB) (B), the MD and the PD were not overlapping and the maxF values from intermediate clusters all fall within the admixed ranges ($\alpha_1 < \text{maxF} < \alpha_3$). Consequently, the populations composing the intermediate cluster (AB) were removed (Figure 18b; subset 1). Following the vertical way of analysis described above, groups A (subset 1a) and B (subset 1b) were analyzed before the second pair dictated by the result obtained with the entire set of populations. In both cases, the MD and the PD were overlapping; this indicated the presence of a single group.

The second subset was composed of (B) and (BC-C). A similar result for subset 1 (Figure 18c, subset 2) was obtained for this set; the populations composing the intermediate cluster (BC) were removed. The next iteration consisted in analyzing each reference group. Group B (subset 2a) was already analyzed with subset 1b. For group C, the MD and the PD overlapped, indicating the presence of a single group (Figure 18c, subset 2b).

As the admixed populations AB and BC were removed, the third subset was composed of groups A and C. (Figure 18d, subset 3). The MD and the PD did not overlap. maxF computed from the populations of the intermediate cluster overlapped the maxF range associated to pure populations (3 populations on 4 had a maxF > α_3). The K-means was thus performed anew but with two groups instead of three. Iterations are performed anew on each of these new clusters. However, the analyses on groups A (subset 1a) and C (subset 2b) were already performed. Consequently, the iterative procedure is complete. The final output is thus composed of three groups of populations (A, B and C, each composed of 8 populations) and 12 admixed populations removed throughout the procedure.

The last step consisted in assessing the contribution of each pure group in the mixed populations. These results are presented in Figure 19. Individual reassignment and admixed stock analyses gave similar results. The unequal 70/30 proportions mixes of A



More than one group ($\alpha_2 < \alpha_3$)
External populations detected in the intermediate cluster (int maxF < α1).
POST is performed anew upon each pair of clusters following a decreasing divergence order.

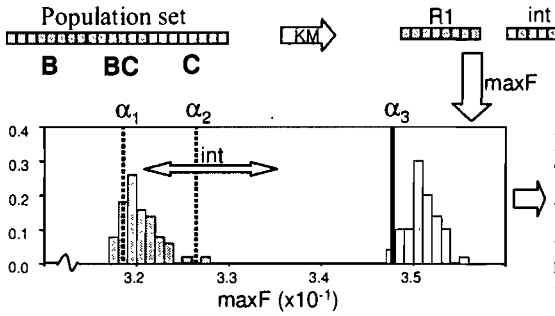
More than one group ($\alpha_2 < \alpha_3$)
Two groups with admixed populations (α1 < int maxF < α3).
Admixed populations (from AB) are removed.
POST is performed on R1 (A) and R2 (B).

A single group ($\alpha_2 > \alpha_3$).
maxF of population from intermediate cluster are not computed.
The procedure stops for this set

A single group ($\alpha_2 > \alpha_3$).
maxF of population from intermediate cluster are not computed.
The procedure stops for this set

Subset 2 (pair B/BC-C)

(c)

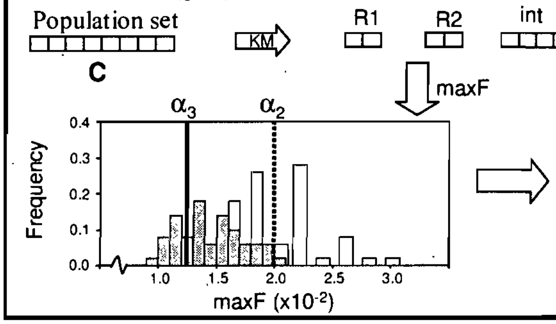


More than one group ($\alpha_2 < \alpha_3$)
Two groups with admixed populations ($\alpha 1$ < int maxF < $\alpha 3$).
Admixed populations (from AB) are removed.
POST is performed on R1 (B) and R2 (C).

Subset 2a (group B)

Already performed with subset 1b

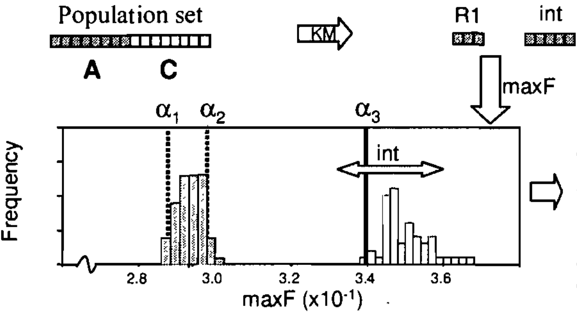
Subset 2b (group C)



A single group ($\alpha_2 > \alpha_3$).
maxF of population from intermediate cluster are not computed.
The procedure stops for this set

Subset 3 (pair A/C)

(d)



More than one group ($\alpha_2 < \alpha_3$)
Two groups without admixed populations ($\alpha 3$ < int maxF).
K-means is performed anew with two groups and POST is performed on each of these new clusters (A and C).

Subset 3a (group A)

Already performed with subset 1a

Subset 3b (group C)

Already performed with subset 2b

Figure 18: Decomposition steps of a complex genetic structure. For each POST iteration, the set of populations, the K-means (KM) result, the identification of the reference (R1 and R2) and the intermediate (int) clusters, MD (grey bars), the PD (white bars), a1 and a2 (dot lines), a3 (solid line) and the range covered by the maxF (when computed) on samples within intermediate clusters (grey arrow) are shown. For each iteration, the interpretation and the next step to be performed are described.

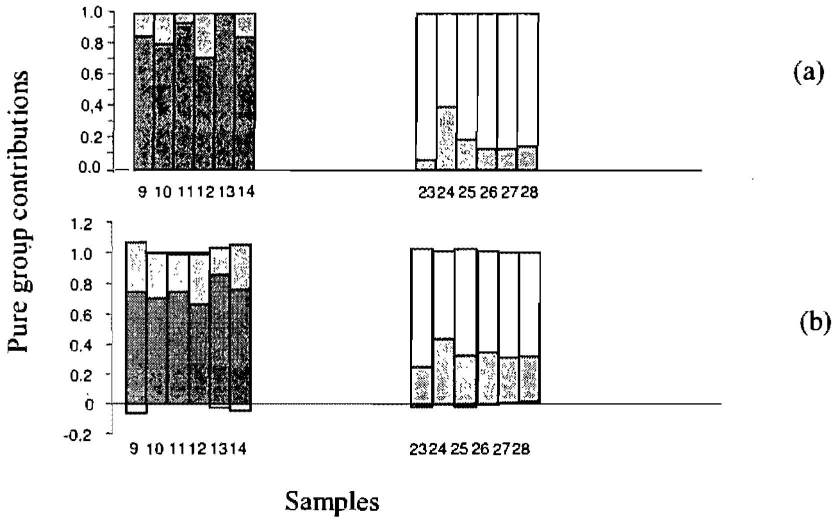


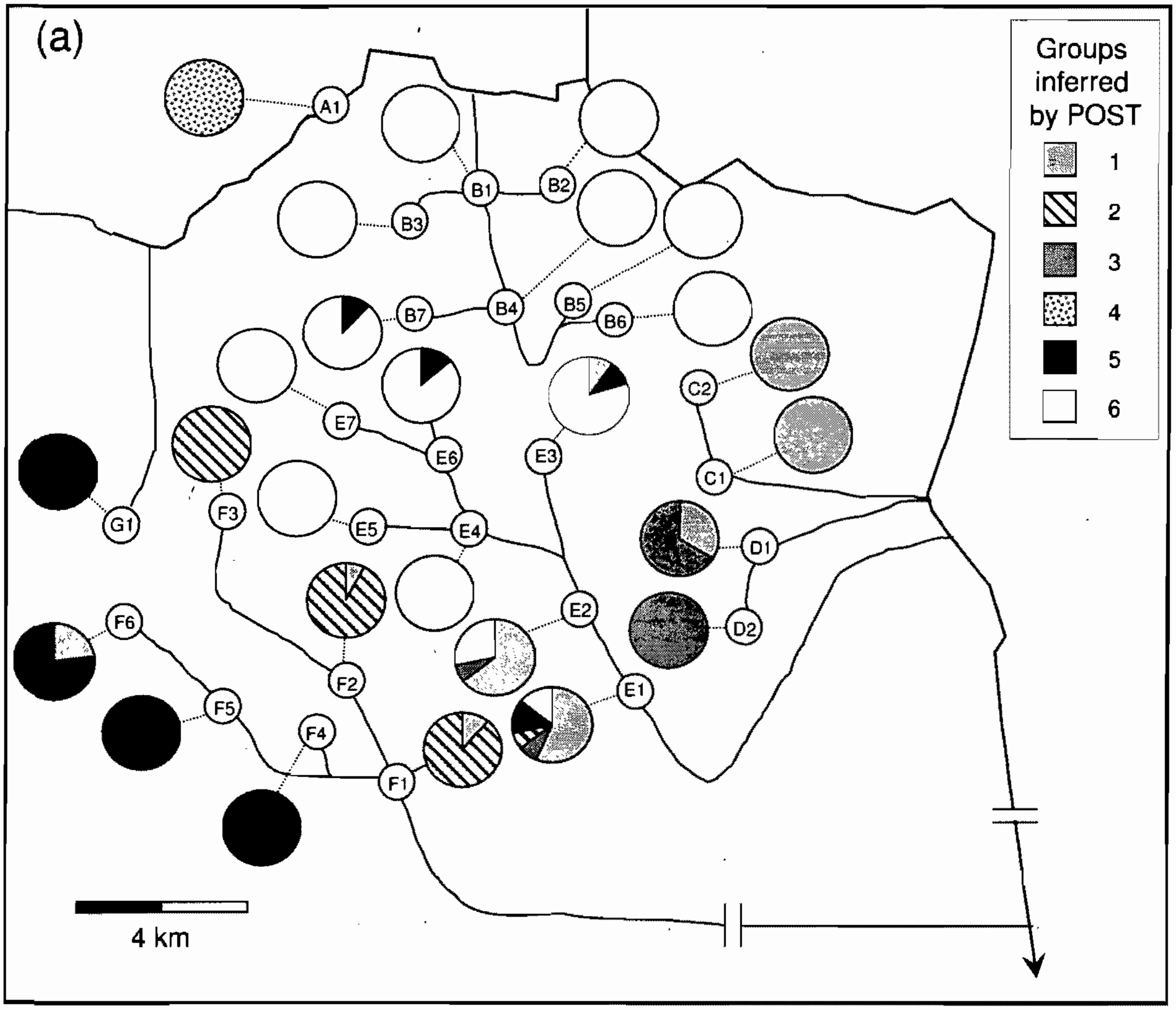
Figure 19: The reassignment results according to the individual reassignment analysis (a) and the admixed-stock analysis (b) of the admixed samples removed from the dataset throughout the decomposition of the simulated complex structure. The bars represented the relative contribution of group A (dark grey), B (light grey) and C (white) in each of the twelve admixed samples identified.

and B (pop 9 to 14) and of B and C (pop 23 to 28) were correctly recovered by both approaches. However, admixed stock analysis give small negative contributions for group A (23, 24, 25 and 26) and C (9, 13 and 14) in several populations (Figure 19b).

Application to Brook charr

The total decomposition resulted in six distinct groups, each corresponding to a single population (A1, B2, C2, D2, F3, and G1) and 20 populations identified as mixed and thus removed through the procedure. According to the individual reassignment approach, ten populations considered as mixed (B1, B3, B4, B5, B6, E4, E5, E7, F4 and F5) were assigned to a single group. The others appear to be composed of more than one group (Figure 20a). These results are congruent with those obtained previously by Angers and Bernatchez (1998). The six main lineages were recovered, including the singleton (A1). The three populations putatively associated to introgression or hybridization events (E1, E2 and F6; Angers and Bernatchez 1998) were correctly identified as mixed by the procedure.

The results obtained with the mixed-stock analysis are less structured as all populations considered as mixed appear to be composed of more than one group. Negative contributions by pure groups (from -0.01 to -0.22) were observed in 18 of 20 admixed populations. Furthermore, an average of 4 positive contributions (on a maximum of 6) was observed in each admixed population. The figure 20b show results in which all negative contributions and those below 10% were assumed to be null. The presence of group 5 in almost all populations of drainages B and E is inconsistent with the results of Angers and Bernatchez (1998). Furthermore, much discordance can be observed with the putative introgression events mentioned above. Even though the results of E1, E2 and D1 are in accordance with the previous analyses, the introgression signal within F1, F2, F6 and E3 is not observed.


(a)
Groups inferred by POST
1
2
3
4
5
6
A1
B1
B2
B3
B4
B5
B6
B7
C1
C2
D1
D2
E1
E2
E3
E4
E5
E6
E7
F1
F2
F3
F4
F5
F6
G1
4 km

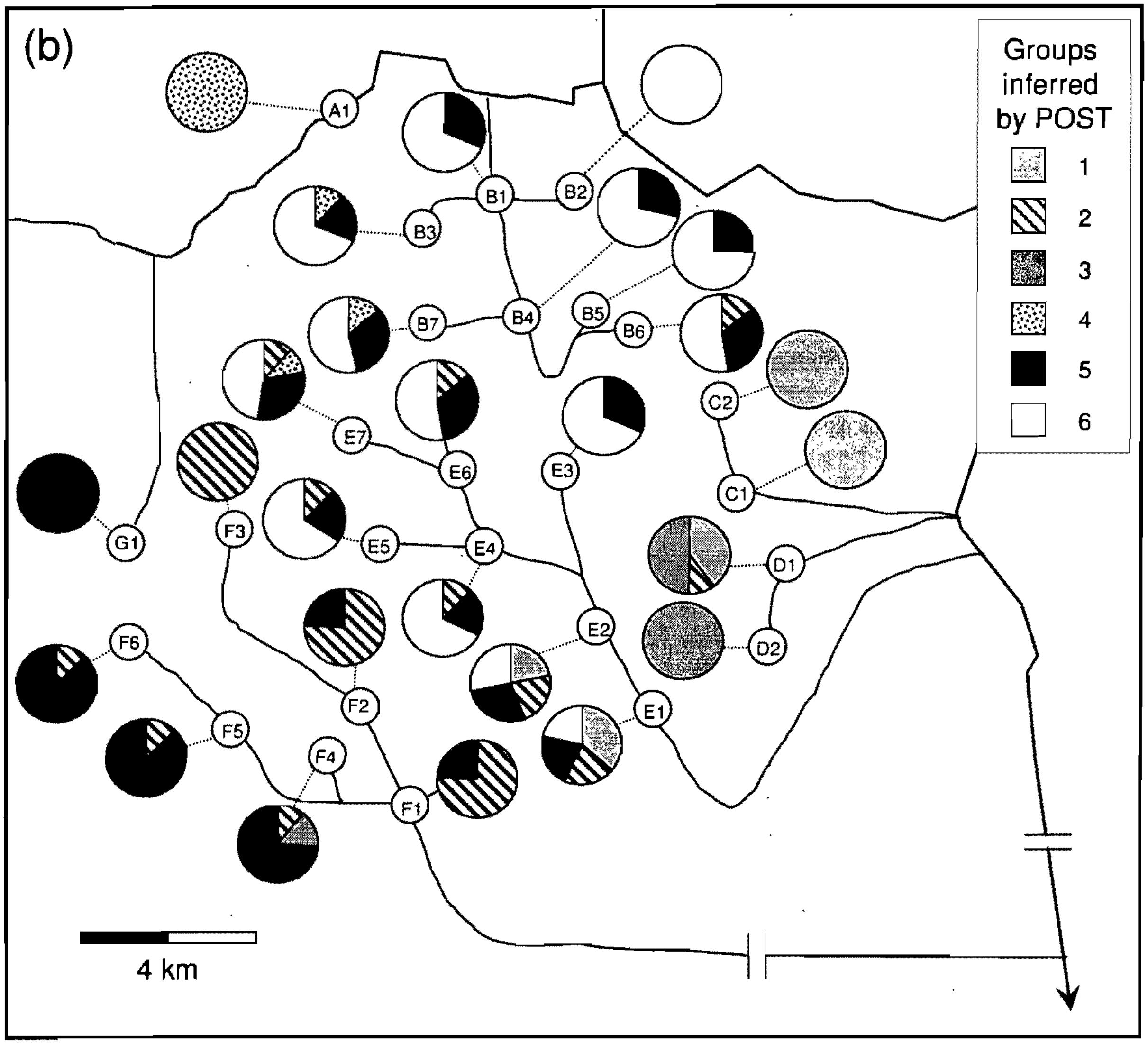

(b)
Groups inferred by POST
1
2
3
4
5
6
A1
B1
B2
B3
B4
B5
B6
B7
C1
C2
D1
D2
E1
E2
E3
E4
E5
E6
E7
F1
F2
F3
F4
F5
F6
G1
4 km

Figure 20: Population structure inferred by the POST procedure on the 26 populations of brook charr (*Salvelinus fontinalis* Mitchill) followed by either an individual reassignment approach (a) or a mixed-stock analysis (b). The procedure was performed using the information from five microsatellite markers (data from Angers and Bernatchez, 1998). Drainages are represented by a different letter (A to G).

Discussion

The absence of *a priori* knowledge about a set of populations may bring difficulties into the investigation of potential genetic structure. First, the number of groups being unknown, users will usually try many possibilities and keep the result that satisfies a given criterion. In the best case, this criterion has statistical bases (Pritchard et al. 2000). Second, finding the best partition among a given number of clusters implies trying numerous possible subdivision solutions. Even though this can be performed easily with heuristic algorithms (e.g. SAMOVA, K-means), forcing discrete partition of populations may be a false solution, as shown previously in quantity of works in which overlaps between genetic distinct groups has been detected.

In this perspective, POST offers numerous advantages. This new method follows a heuristic algorithm which allows the exploration of the genetic structure of a set of natural populations. It does not require *a priori* population clustering hypotheses and, by allowing populations to belong to more than one group, it increases the precision of the genetic structure description. POST also appears versatile and could be applied to numerous studies. For instance, it may have the potential to handle different types of markers and it should be possible to extend the method to other members of the family of F statistics such as the ρ_{ST} (Rousset 1996) and ϕ_{ST} (Excoffier et al. 1992).

Even though the simulation analyses performed in this work are only preliminary, the results obtained suggest that POST is quite robust to small sample sizes (15 individuals), number of loci (5 loci) and genetic differentiation between groups. For instance, POST was able to discriminate a single-group structure from one composed of numerous discrete groups in all simulation conditions, including those for which the genetic differentiation between reference groups was very low ($F_{CT} = 0.02$). Consequently, POST does not detect any group structure if one is not present. However, the capacity of POST to discriminate basic structures associated to two groups with or without admixture and to more than two groups was more mitigated. Except for the admixed populations composed of unequal proportions of each reference groups that were all correctly identified, erroneous conclusions were reached in the case of perfect 50/50 mixture, pure and external populations. However, two of these three errors (pure and external) can be minimized in two ways. First, the diagnosis rules were constructed to be very liberal (only one pure or external population from the intermediate cluster is necessary to conclude to two groups without admixture and more than two groups respectively). Second, the α_1 and α_3 thresholds can be set in order to reduce the maxF range associated to admixed populations. This last solution must however be used with caution since it may lead to wrong conclusions. As an example of this situation, the 50-50 mix populations can be misidentified as external populations. Such errors inevitably overestimate the number of groups. Potential error can also be associated to mixed population composed of largely unequal fractions of both reference groups. In such situations, the smaller fraction may be considered as negligible and the population may be identified as pure. Additional simulation runs will be necessary to specify and understand the weaknesses of POST. For instance, all simulations were performed by fixing the number of population within groups to 10. If this value is realistic for population genetics analyses in natural environment, it remains quite small in statistical analysis perspective. Consequently, the reliability of POST for different number of population within reference and intermediate groups need to be evaluated more properly. In addition, the potential unequal diversity among groups and the presence of mutation-drift disequilibrium populations are also parameters for which the effects on POST reliability are still not known.

We proposed two different approaches to assess the origin of admixed populations. A population resulting from an admixture between two pure groups is composed of individuals with genotype resulting of a random drawing of alleles from both parental groups. Consequently, an admixed stock analysis appears as the most appropriate approach. However, results of such a method could be highly biased and/or show a high variance when using a low number of loci (<10 loci) and individual (<50 individuals) and if the divergence time (and thus the genetic differentiation) between parental stocks is low (Bertorelle and Excoffier 1998). These biases are increased if the molecular information is not taken into account (and thus if loci follow an infinite allele model). On the other hand, even though the individual assignment approach might appear less appropriate (especially if admixture come from an historical process), its robustness to low numbers of individuals (>10) and loci (>5) even in low differentiation among parental stock (FST > 0.08) (Cornuet et al. 1999) is less sensible than the mixed-stock analysis. Thus, this approach is a potential alternative if the data set does not reach the conditions required for admixed-stock analyses or if the mix resulted from the sampling of individuals belonging to parental groups living in sympatry. The reliable results obtained with the Bayesian individual reassignment approach in the simulated (5 loci, 30 individuals per sample and 100 generations since admixture event) and on the brook charr datasets (5 loci, 30 individuals per sample) are in agreement with this idea. The choice between an admixed stock analysis and an individual-based assignment method should then be based on the characteristics of the data set.

The brook charr populations provided a good example of the difficulties encountered to describe population structure when no *a priori* knowledge is available. The genetic structure of this species largely reflected the colonization by different evolutionary groups that evolved in allopatry during the Pleistocene glaciations, leading to a generally high genetic structure among populations, but that does not always reflect the hydrographic network (Angers and Bernatchez 1998, Angers et al 1999, Castric and Bernatchez 2003, Fraser and Bernatchez 2005). The congruence obtained by POST with the results of Angers and Bernatchez (1998) appears as a strong demonstration of its capacity to correctly recover a complex population structure without *a priori* clustering hypotheses.

In conclusion, our heuristic procedure offers many interesting aspects for population genetics analysis, particularly within a conservation perspective. An important issue in the elaboration of conservation strategies is the identification of conservation units. Beside the theoretical debate about the definition of such units (see Fraser and Bernatchez 2001 and references therein), the identification and the interpretation of conservation units will largely depend, not only on the type of genetic marker, but also on the analytical method. Because it is simple, efficient and completely objective, POST appears as a good method to identify the evolutionary significant units which lied at the base of many conservation procedures.

CHAPITRE V

Phylogeographic inferences of the Longnose dace (*Rhinichthys cataractae*) populations of Quebec Peninsula using microsatellite markers.

Résumé

L'identification de groupes évolutifs distincts et de leurs zones de contact est nécessaire pour permettre une interprétation non biaisée de la diversité génétique entre les populations. Dans cette étude, la méthode POST est utilisée sur le naseux des rapides, une espèce pour laquelle la structure analysée au niveau mitochondrial a démontré la présence de deux grandes zones de contacts au niveau de la péninsule Québécoise. Les résultats ont montré que la richesse allélique et l'hétérozygosité sont négativement corrélées à l'altitude et à la latitude. De plus, malgré une bonne corrélation entre les groupes nucléaires définis par POST et les clades mitochondriaux préalablement identifiés, cette méthode a permis de mettre en lumière la fragmentation de la diversité génétique en plusieurs groupes fondateurs lors de la colonisation des parties nord et est du territoire québécois. Ces résultats sont une démonstration supplémentaire de l'importance de l'utilisation de plusieurs types de marqueurs génétiques pour permettre de raffiner la compréhension des processus historiques modelant les diversités intra- et inter-populations.

Mots clés: altitude, latitude, naseux des rapides, zone de contact, microsatellites, clade mitochondrial.

Abstract

Identification of distinct evolutionary groups and their potential admixture is required for unbiased interpretations of the genetic diversity among populations. In this study, the method POST was used on the Longnose dace, a species for which two major contact zones were previously identified on the Quebec Peninsula with a mitochondrial analysis. The results showed that allelic richness and heterozygosity are negatively correlated to altitude and latitude. Furthermore, while nuclear groups defined by POST were highly correlated with the mitochondrial clades previously identified, it highlighted the fragmentation of the genetic diversity in numerous founder groups during the northward and the eastward colonizations. These results appear to be a strong demonstration of the importance of using different genetic markers to insure a better comprehension of the historical processes that shaped neutral genetic diversity within and among populations.

Keywords: altitude, contact zone, latitude, Longnose dace, microsatellites, mitochondrial clade

Introduction

The Longose dace has a wide geographic repartition in North America, being located in most of the turbulent rivers from Atlantic to Pacific coasts, from Mexico to Yukon (Scott and Crossman 1973). Because its present-day distribution overlaps a large fraction of the territory covered by Pleistocene glaciations, the modifications that occurred to the hydrological network during the glacier retreat likely played a central role in the shaping of the actual population structure of this species (Rempel and Smith 1998). Indeed, the analysis of the mitochondrial genome revealed that the postglacial colonization of the Longnose dace on Quebec Peninsula has been profoundly affected by postglacial seas (Girard and Angers 2006b). According to their study, three distinct groups geographically well-structured overlapped within the areas once covered by the large marine invasions. Consequently, several populations resulted from historical and distinct admixtures between different evolutionary lineages.

The relevance of the mtDNA into detecting the main trends of the colonization sequence of a species is obvious (Avise 2004 and references therein). However, this genome reaches limitations when one wants to investigate a finer spatial or temporal scale (eg. Angers and Bernatchez 1998). The lower genetic diversity related to the effective population size of the mtDNA may result in a few large evolutionary units with a widespread distribution. While quite precise when comparing to other fish species (Brown et al. 1992; Ferguson et al. 1993), the phylogeographic model of the Longnose dace developed with mtDNA could possibly be refined by using markers showing a higher level of polymorphism.

In order to refine the resolution of this phylogeographic model, we aimed at analyzing this system using microsatellite markers. However, the genetic structure of Longnose dace populations from Quebec Peninsula appears as a text book example of a complex genetic organization. Because they do not consider group overlaps, usual population structure analyses such as AMOVA or phylogenetic algorithms are ineffective in the presence of admixture. On the other hand, this system provides an excellent opportunity to test the performance of the method POST, that aims at identifying groups of genetically similar populations as well as potential overlaps between them (Girard and

Angers 2007b). Using this method, the objective was to depict the population structure organization of the Longnose dace with seven microsatellites. This result was then compared to the mtDNA structure previously defined in order to increase the comprehension of the effects of the historical processes on the genetic structure of this species.

Material and Methods

Sampling and molecular analyses

A set of 27 populations of Longnose daces for which the sample size ranged between 9 and 24 individuals was analyzed (Table 9; Figure 21). These rivers are located within four drainages (James Bay, Saint Lawrence River, Outaouais River and Saguenay River). Further details on geographic location of each sample can be found in Girard and Angers (2006b). Populations were screened with seven microsatellite loci (Rhca15b, Rhca16, Rhca20, Rhca23, Rhca31, Rhca34, Rhca52) according to conditions described in Girard and Angers (2006a, 2008). The mitochondrial dataset used in this work was the one reported by Girard and Angers (2006b). Using short segments of the cytochrome b (232 bp) and the control region (236 bp), they were able to identify 12 haplotypes grouped in three main clades (I, II and III). In the following analyses using the mitochondrial information, the composition of each population was related to the frequency of each of these clades.

Statistical analyses

Population diversity

Genetic diversity of microsatellites within populations was calculated using the Nei gene diversity (H_E) and the allelic richness of El Mousadik and Petit (1996) with the

Table 9: Characteristics of the Longnose dace populations sampled in this study. The number of individuals (N), the microsatellite richness (R), the gene diversity (H_E) and the probability to be under the mutation-drift equilibrium (MD) are presented. Numbers next to each population name will be used to identify populations in text and figures.

Populations	N	R	H_E	MD
James' Bay bassin				
1-Martel	22	2.55	0.44	ns
2-Coigny	23	3.69	0.63	ns
3-Mégiscane	15	3.82	0.58	**<0.05**
Saint Lawrence River bassin				
6-Trois-Pistoles	20	1.91	0.31	**<0.05**
7-Sault-aux-Cochons	17	2.09	0.33	ns
8-Malbaie	16	4.00	0.60	ns
9-Chaudière	21	4.69	0.65	**<0.05**
10-Jacques-Cartier	18	3.42	0.57	ns
12-Des Envies	23	4.18	0.62	ns
13-Nicolet	19	4.62	0.70	ns
14-Mékinak du Nord	19	3.53	0.52	ns
15-Yamaska	22	4.52	0.69	ns
16-Bécancour	17	3.65	0.61	ns
18-Noire	22	3.44	0.49	ns
19-Châteauguay	18	4.08	0.62	ns
20-Assomption	9	3.52	0.57	ns
21-Du Loup	15	3.49	0.52	ns
Outaouais River bassin				
22-Diable	15	2.91	0.45	ns
23-Rouge	11	3.15	0.47	ns
24-Kazabazua	9	3.42	0.60	ns
25-Lavallée	24	2.82	0.42	ns
26-Loutre	22	3.17	0.51	ns
Saguenay River bassin				
27-Petit-Saguenay	20	2.99	0.49	ns
28-À Mars	22	2.91	0.50	ns
29-Shipshaw	20	3.03	0.49	ns
30-Péribonka	15	3.26	0.59	ns
31-Métabetchouane	21	3.25	0.54	ns

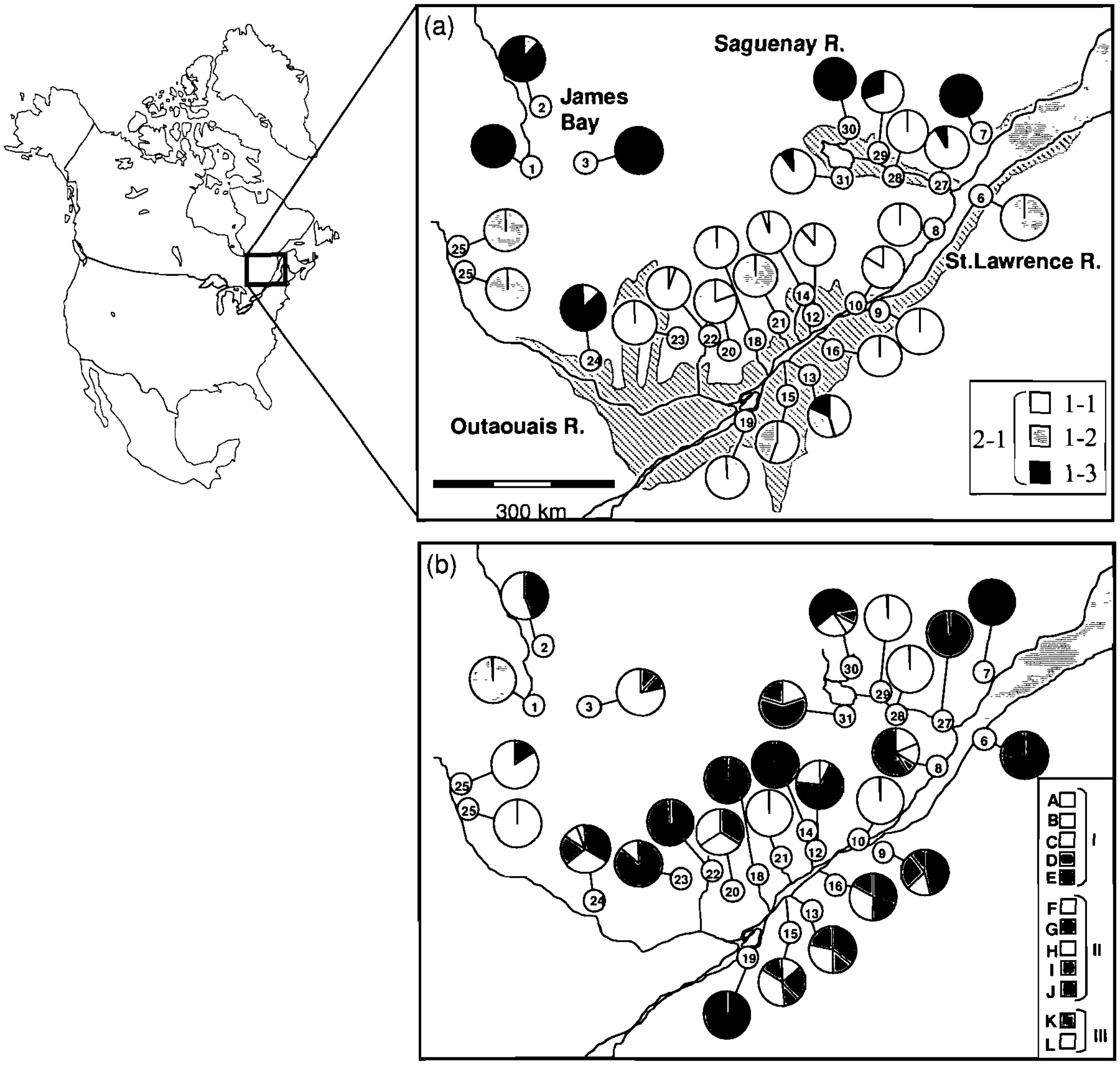


Figure 21: The geographic position of populations combined to their mitochondrial composition according to Girard and Angers (modified from Girard and Angers 2006b) (a) and their group composition according to POST (b). The hatched region in figure (a) corresponds to the extension of the Champlain and Laflamme seas (after Ochietti 1989). Particular colour ranges were associate to each sets delineated on the CCorA biplots presented on Figures 19. Refer to Table 9 for populations and drainages identification.

software FSTAT v.2.9.3 (Goudet 2001). Departure from drift/mutation equilibrium was evaluated for each locus in each population using the BOTTLENECK program (Cornuet and Luikart 1996). A two phase mutation model (TPM; DiRienzo et al. 1994) was assumed for all the microsatellites. The TPM was composed of one-step mutations with 10% of multistep changes with a variance of 10 repeats. The temporal stability of the effective size of each population was evaluated with the Wilcoxon sign-rank test (Luikart and Cornuet 1998). Finally, population pairwise F_{ST} and total F_{ST} (Weir and Cockerham 1984) were computed.

The importance of the geographic localization in shaping the diversity within population was also evaluated. The longitude, the latitude and the altitude of each sample were taken in Girard and Angers (2006b). Multiple regressions were performed using either microsatellites average richness and heterozygosity *versus* altitude and a cubic polynomial function of the geographic coordinates.

Structure description

POST was performed using both the parameters and recommendations of Girard and Angers (2007b). The pairwise Chord distance (D_{CE}; Cavalli-Sforza and Edwards 1967) matrix was computed using POPULATION v.1.2.28 (Langella 1999). The K-means algorithm and the evaluation of the distance among group centroids were performed with the GINKGO Multivariate Analysis System v.1.5.4 (Bouxin 2005) using 50000 starting random seeds. The maxF distributions were constructed using 100 random populations using the R function developed by Girard and Angers. Values of 0.1, 0.99 and 0.01 were set respectively for $\alpha 1$, $\alpha 2$ and $\alpha 3$. The reassignment of the mixed populations was performed using the method proposed by Rannala and Mountain (1997) available in the software GENECLASS2 (Piry et al. 2004). This method was chosen instead of the mixed-stock analysis because the number of loci (7) and of individuals per samples (average 20) of our dataset may make this later approach unreliable (Bertorelle and Excoffier 1998). Additional details about the method and settings are reported in Girard and Angers (2007b).

Comparison between nuclear and mitochondrial population structures

The microsatellite structure resulting from POST was compared to the mitochondrial composition with a canonical correlation analysis (CCorA), using an R function (P. Legendre, personal communication). A matrix **Y** representing the relative contribution of each group composing the nuclear structure was compared to a matrix **X** in which relative frequencies of mitochondrial clades were compiled.

Results

Within population diversity

Allelic richness of microsatellites ranges from 1.91 to 4.69 (average 3.41), while H_E ranges from 0.31 to 0.70 (average 0.54). Only three populations are not in accordance with mutation-drift equilibrium (3, 6 and 9; Table 9). The pairwise F_{ST} values ranges from 0.08 (12 and 14) and 0.57 (1 and 7) and are all significant, indicating the presence of completely differentiated populations. The F_{ST} computed over all populations is 0.34.

Both microsatellite richness ($R^2_{adj} = 0.78$; $p < 0.0001$) and heterozygosity ($R^2_{adj} = 0.65$; $p<0.0001$) are strongly associated to the geographic model combining altitude and the cubic polynomial function of the geographic coordinates. However, two parameters can be highlighted. Both richness and heterozygosity are significantly and negatively associated to altitude and latitude (Figure 22). These analyses were performed by removing the three populations which were not in accordance with mutation-drift equilibrium, because their genetic diversity is believed to be function of local demographic events.

Population structure

The total decomposition from POST results in twelve independent groups, each corresponding to a single population (1, 7, 10, 14, 18, 19, 21, 22, 25, 27, 28 and 29). The remaining populations are composed of admixture of 2 to 5 groups (average 2.1). The

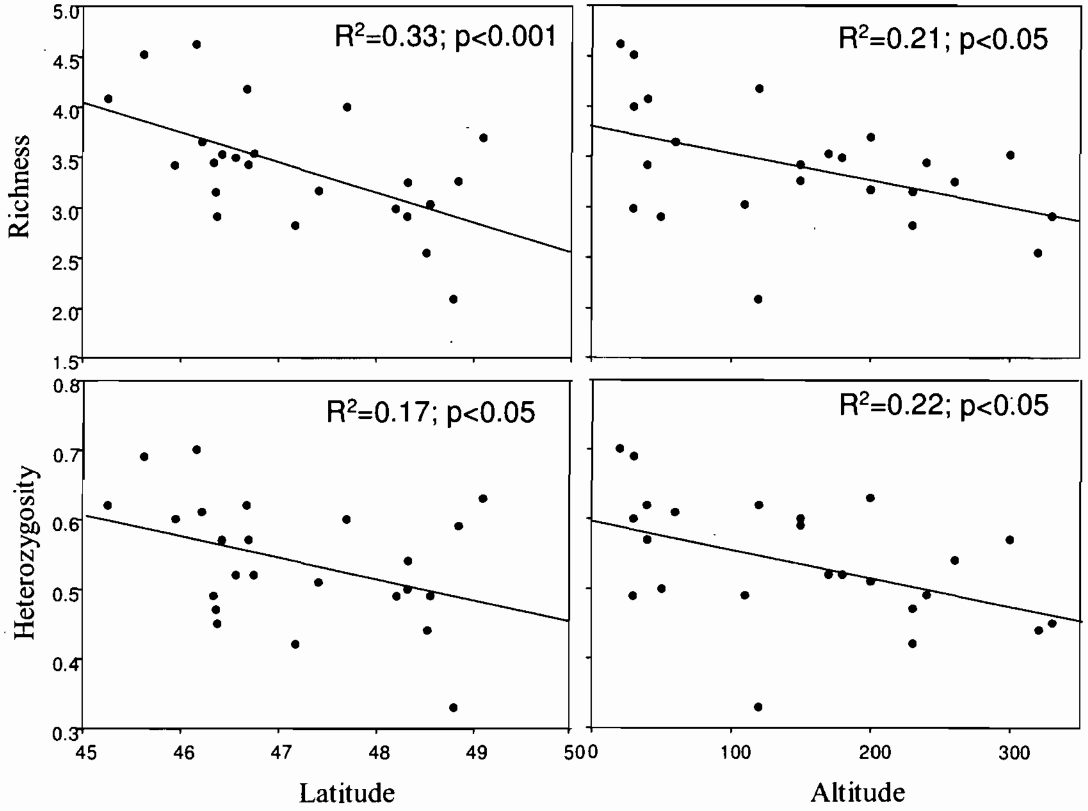


Figure 22: The average microsatellite richness and heterozygosity within samples in relation to either the latitude (°) or the altitude (m).

maximum frequency of a given group within mixed populations ranges between 0.20 and 0.87. Each group is present within 1 (A and G) to 10 (J) populations.

The population structure is geographically organized, being mainly related to drainages. For instance, Saint Lawrence (A, E, I and H), Saguenay (B, C, D and K), Outaouais (G and F) and James' Bay (L) are strongly associated to particular nuclear groups (Figure 21b). Otherwise, group J is not associated to any particular drainages, because it can be observed in St. Lawrence, in Outaouais and in James' Bay. Furthermore, admixed populations are generally adjacent or close to their associated pure populations. For instance, group H is the unique group of the population 21, and it was present in different proportions within nearby populations (12, 13, 15 and 16). Similarly, group D is found in pure population 27 and is admixed within adjacent populations 8, 30 and 31.

Comparison between nuclear and mitochondrial population structure

The CCorA indicates strong correlation between the mitochondrial clade compositions and the twelve groups obtained with POST (Figures 23a and b). Correlations of 0.97 and 0.91 are observed between the two matrices on first and second canonical axes respectively (Pillai's trace = 1.77). The correlation biplots underline the close relationship between groups and mitochondrial clades (Figures 23 a and b). More precisely, groups A to E (mtDNA clade 1-1), F to J (clade 1-2) and K and L (clade 1-3) are associated to a particular mitochondrial genetic background.

These associations highlight the similarities between the geographic organization of the nuclear and the mitochondrial structures. Even though discordances can be observed between mitochondrial and nuclear groups within populations (i.e. pop 15, 24 and 30), strong similarities in the main trends among populations can be observed. In addition to the drainage organization described above (which was also reported for mitochondrial markers), the major contact zones observed upon the areas once covered by post-glacial seas are retrieved with nuclear groups. For instance, the overlap between

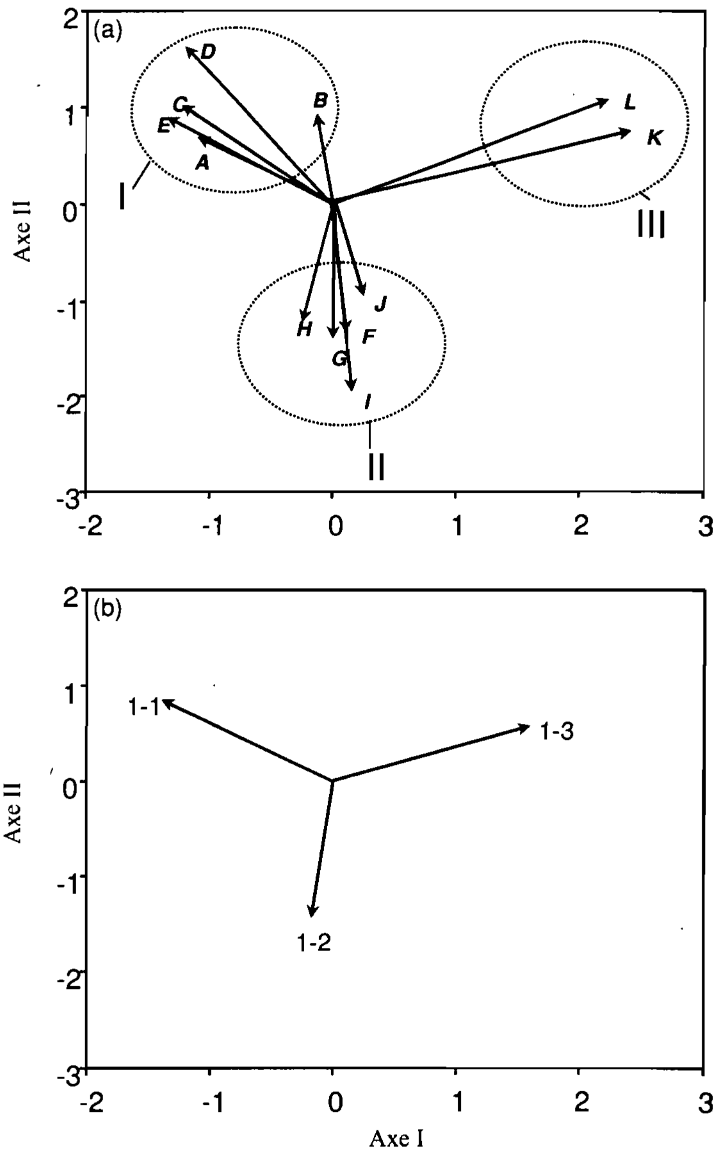


Figure 23: The correlations between mitochondrial and nuclear populations structures. The CCorA ordination biplots represent the nuclear groups according to POST (a) and the mitochondrial clades of Girard and Angers (2006b) (b). Three sets of nuclear groups (I, II and III), each associated to a particular mitochondrial clade, are defined.

A to E groups (clades 1-1) and H to J (clade 1-2) is clearly observed over Saint Lawrence drainage. The groups A to E are mainly observed in populations of the eastern part of the drainage (pop. 9, 10, 12, 14), while F to J show a more western distribution (pop. 16, 18, 19, 21). The contact zone over the Laflamme Sea between A to E groups (clade 1-1) and K and L groups (clade 1-3) is also consistent with the observations of Girard and Angers (2006b, figures 21).

The correlation between the mitochondrial and the nuclear structures is also perceptible in terms of richness and diversity. For instance, the number of nuclear groups within population is highly correlated to the number of mitochondrial clades previously observed by Girard and Angers (r = 0.73; p<0.001). A strong correlation is also observed between nuclear groups and mitochondrial clade diversities (r = 0.86; p<0.001). As observed previously, the two areas once covered by the postglacial seas appear as highly diversified. A total of eight (B, C, D, E, F, G, H and I) and five (B, C. D, I and K) groups are respectively observed upon the areas covered by Laflamme and Champlain Seas (Figure 21a). However, differences in the diversity within populations were observed in both of these areas. While the populations were highly mixed over Champlain Sea, most of the populations were pure (but for different nuclear groups) over the Laflamme Sea.

Discussion

The deglaciation processes are of prime importance to understand the colonization sequence of the organisms living on territories once covered by the Pleistocene ice sheet (Rempel and Smith 1998). The analyses performed on microsatellites diversity within and among populations of Longnose dace increased the comprehension of the effects of these historical processes on the genetic structure of this species.

While the genetic diversity within populations is highly geographically structured, it appears closely related to altitude. The negative relationship observed between both allelic richness and heterozygosity with latitude is in agreement with previous observations performed on other species which colonized new icefree territories of the north hemisphere (Bernatchez and Wilson 1998, Edmands, 2001). Over the Quebec territory, two main non-exclusive hypotheses can be proposed to explain this relationship for the Longnose dace.

First, these populations might suffer from more pronounced founder effects because of a reduction into the number of founders with latitude. Second, considering that the retreat of the glacier followed a South-North trajectory (Dyke and Prest 1987), it is conceivable that these founder waves faced important decrease of genetic diversity because of their presence at the edge of the glacier, where the environmental conditions were more unstable (Avise et al. 1984). The altitude appears to be the other geographical variable to which the genetic diversity of Longnose dace is related. The negative influence of altitude on genetic diversity has been previously observed in freshwater habitats (Angers et al. 1999, Castric et al 2001). High-altitude populations are expected to be more physically isolated, because of the presence of physical barriers to gene flow (Castric et al 2001). As for latitude, it is also possible that these populations suffered from more pronounced founder effects due to a negative effect of altitude on the number of individuals that colonized them. Furthermore, the presence of contact zones among several founding groups in the low-altitude areas may also explain this relationship.

The effects of founder events on the differentiation of populations following their establishment into northern and/or higher drainages are also highlighted by the population structure depicted by POST. Indeed, populations located on the south shore of the St. Lawrence River and in low altitude showed admixture of nuclear groups. In counterparts, these groups are not admixed in populations located in high altitude over the north shore and in the northern part of Outaouais drainage. This signal increases the likelihood of the hypothesis according to which the highly diversified contact zone over the Champlain Sea area was fragmented in numerous founder groups. After the isostatic rebound and the following retreat of the marine invasion, the different channels carved by the melt-water of the glacier were available for the colonization. Each of these channels was likely colonized by different subsets of founders. The further effects of drift within each founder group likely increase the level of differentiation between them (Nei et al 1975). These processes may potentially explain the higher number of nuclear groups in comparison to the number of mtDNA clades identified in Girard and Angers (2006b). This pattern of colonization can be applied to nearly all populations of the north shore of Saint Lawrence and Outaouais.

While the signal was not statistically significant for the within-population diversity parameters, the results obtained by POST highlight the longitudinal progression of the colonization of Longnose dace at least for the populations associated to the clade 1-3. While mtDNA gave a unique evolutionary group from West to East, POST distinguished the populations of the James' Bay drainage from those of the North-eastern part of the Peninsula. In this case, the isolation by distance differentiation following the colonization through the centre via the river interconnections is the most parsimonious explanation. According to the progression of the glacier and the Lake Ojibway-Barlow that took place from West to East (Dyke and Prest 1987), the colonization pattern of Longnose dace is in accordance with the geomorphological history of the region surveyed. Furthermore, these observations are in agreement with those performed on the northern pike (*Esox lucius)*, the lake whitefish (*Coregonus clupeaformis)* and the yellow perch (*Perca flavescens)* (Gagnon and Angers 2006).

In conclusion, even though both mtDNA and microsatellites analyses were largely consistent, their different precision conducted to an increase comprehension of the postglacial colonization of the Longnose dace. For instance, the nuclear groups defined by POST were highly correlated with the mitochondrial clades previously identified. Furthermore, the effects of latitude and altitude on population structure were highlighted by both marker types. However, noticeable dissimilarities were also pointed out. While mtDNA allowed the definition of three waves of colonists entering the peninsula at different time or from different accesses, the microsatellites markers highlighted the fragmentation of the genetic diversity in numerous founder groups during the northward and the eastward colonizations. These results are another demonstration of the importance of looking to different genetic markers to insure a better comprehension of the historical processes that shaped neutral genetic diversity within and among populations and to adequately identify unique genetic backgrounds into a conservation perspective.

CHAPITRE VI

Deterministic and stochastic evolutionary processes acting on MHC in Longnose dace (*Rhinichthys cataractae*) populations

Résumé

Les gènes du complexe majeur d'histocompatibilité (CMH) jouent un rôle central au niveau de la défense des organismes contre les pathogènes. Malgré de nombreuses études portant sur plusieurs taxons dans des milieux expérimentaux, semi-expérimentaux et naturels, les mécanismes évolutifs agissant sur ces gènes ne sont toujours pas clairement établis. L'objectif de ce travail est de décortiquer les effets à long et court terme de la sélection naturelle de ceux de la dérive génétique, des mutations et de la fragmentation allopatrique qui agissent sur la diversité du CMH IIβ. Les séquences et la diversité génétique du CMH à l'intérieur et entre les populations ont été analysées conjointement avec celles de régions nucléaires codantes (hormone de croissance et trypsine) et non codantes (microsatellites). Étonnamment, nos résultats suggèrent l'action de processus essentiellement neutres au niveau de l'évolution du CMH à long terme. Cependant, ces résultats pourraient être biaisés par la faible richesse allélique observée dans l'ensemble des populations. D'autre part, des mécanismes tant stochastiques que déterministes dont l'action varie d'une population à l'autre sont inférés à court terme. L'ensemble de ces résultats suggère donc un mélange complexe de différents processus évolutifs impliqués au niveau du polymorphisme des régions codantes du CMH IIβ dans les populations naturelles.

Mots clés: CMH, dérive, fragmentation allopatrique, mutation, démographie, sélection.

Abstract

Genes of the major histocompatibility complex (MHC) play a central role into the defence of the organism against pathogens. Despite numerous studies performed on widely different taxa in experimental, semi-experimental or natural environment, the evolutionary mechanisms acting on these genes remain unclear. The objective of this work was to disentangle the long- and short-term effects of natural selection on the diversity of the MHC IIβ from those of drift, mutation and allopatric fragmentation. MHC sequence and gene diversity were thus jointly analyzed with those of coding (growth hormone and trypsin) and noncoding (microsatellites) nuclear loci. Our results suggest that stochastic mechanisms drove long-term evolution of MHC. Nevertheless, these results may be biased by the surprisingly low MHC richness observed in the population set. The processes involved in the short-term evolution of MHC seem to vary from a population to another, being under selection or neutral. Altogether, these results suggest a complex mix of different evolutionary processes shaping the level of polymorphism of coding sequences of MHC IIβ in natural environment.

Keywords: allopatric fragmentration, demographic processes, drift, MHC, mutation, selection

Introduction

Functional genes potentially play a central role in local adaptation (Ford 2002). The identification of the processes that shaped the genetic diversity pattern is a major concern for prediction of evolutionary trajectories of populations and is thus an important issue in conservation biology. In this perspective, genes associated to long-term viability of populations received special attentions (Suprunova et al. 2004, Ingvarsson et al. 2006). Among gene candidates potentially having such a role are those of the major histocompatibility complex (Piertney and Oliver 2006). Lying at the base of the immune system of vertebrates, these genes encode several receptors that bind protein fragments of both viral and bacterial antigens. Those fragments are then transported to lymphocytes, which then initiate the immune response of the organism (Lodish et al. 1995). This active role in defence mechanisms highlights the influence of MHC on the occurrence of infection and disease (Bernatchez and Landry 2003).

Despite numerous studies performed on experimental, semi-experimental or natural environment (reviewed in Bernatchez and Landry 2003), generalizing mechanisms that drove MHC evolution remains difficult. As shown by both the excess of nonsynonymous substitutions (Hughes and Nei 1988, Schad et al. 2005, Froeschke and Sommer 2005) and the incomplete sorting of MHC lineages between species (Mayer et al. 1992, O'Brien and Yuhki 1999, Bryja et al. 2006), natural selection played an important role in shaping the MHC diversity through long evolutionary terms. However, the mechanisms implied in the maintenance of MHC alleles within and among contemporary populations appeared more ambiguous. Previous studies implied different types of balancing selection among which heterozygote advantage (Sauermann et al. 2001), geographically varying directional selection (Muirhead 2001, Miller et al. 2001, Heath et al. 2006) and negative frequency-dependant selection (Froeschke and Sommer 2005). By maintaining the MHC lineages within the genetic background of a species, all these types of selection are in accordance with the long term effects described above. However, they will affect the MHC diversity within and among populations differently. Heterozygote advantage and negative frequency-dependant selection maintain within population diversity and limit populations differentiation (Schierup et al. 2000), while geographically varying directional selection decrease within population diversity and favour population differentiation (Landry and Bernatchez 2001, Muirhead 2001). In addition to this complex selection portrait, a growing number of studies suggested that MHC diversity is driven by either strictly stochastic process (Hayashi et al. 2005, Langefors 2005) or by complex interactions of deterministic and stochastic processes (Slade 1992, Landry and Bernatchez 2001, Miller et al. 2001).

These contradictory observations on MHC are a clear demonstration of the difficulty to interpret the evolutionary trajectory of a single DNA segment. The genetic diversity of populations is the result of neutral and/or non-neutral processes acting at different evolutionary time scales. As a result, the effects of natural selection are entangled with the short term effects of drift and gene flow and the long term effects of mutation rate and allopatric fragmentation (Excoffier 2001, Otto 2000). The comprehension of the complete evolutionary portrait of a given gene requires the disentanglement of these numerous processes. In this perspective, a strategy using a joint analysis of sequence, frequency and genotype diversities observed on numerous genetic markers appears suitable (Lynch et al. 1999).

The objective is to disentangle the long and short-term effects of natural selection from those of drift, mutation and allopatric fragmentation on the MHC exon with both sequence and gene diversity analyses. In this context, both neutral (microsatellites) and non-neutral (growth hormone and trypsin) nuclear markers were analyzed. Combining these results to the predictions found in literature, we aim at bringing a better comprehension of the different evolutionary processes implied in the level of polymorphism of coding sequences of MHC IIβ in natural environment.

Material and Method

Model species and sampling

Longnose dace (*Rhinichthys cataractae*) is typical of turbulent rivers in almost all temperate habitats located between Atlantic and Pacific coasts, from Mexico to Yukon (Scott and Crossman 1973; Figure 24a). This fish has little interactions with humans and populations are thus expected to be almost undisturbed by stocking activities or translocations (Scott and Crossman 1973). A total of 542 individuals was sampled by electrofishing (LR-24, Smith-Root) from 27 rivers of the Quebec peninsula in Canada (Figure 24b-d, Table 10). These rivers are located within four drainage basins (James Bay, Saint Lawrence River, Outaouais River and Saguenay River). Sample sizes ranged between 10 and 24 individuals. For each individual, a piece of the caudal fin was removed and stored in 95% ethanol for molecular analyses. In addition, individuals belonging to a sister species, the Blacknose dace (*R. atratulus*), were also analyzed.

Molecular analyses

MHC polymorphism

Because of its key role in peptide bindings (Ono et al. 1993) and its known polymorphism (Garrigan and Edwards 1999, Landry et al. 2001), we focussed our analyses on the second exon of the MHC class IIβ. Primers were designed on the 5' end of this exon and the 3' end of the third exon according to the genomic sequence of *Danio rerio* available on Genbank (#U08870). The primers 5'-CCAGTGACTACAGTGATATGG-3' and 5'-TGGAGGTCACATCTGAGG-3' successfully amplified, by polymerase chain reaction (PCR), a segment of approximately 600 bp covering the 200 last bp of the second exon, the first 150 bp of the third exon and the entire intron between them. The following reaction conditions were used: a 12.5 μL reaction aliquot contained 1.5 mmol·L^{-1} of $MgCl_2$, 2.5 nmol·L^{-1} of each dNTP, 0.2 unit of *Taq* polymerase, 1.25 μl of 10x *Taq* polymerase buffer (Invitrogen) and approximately 20 ng of DNA. Reaction conditions included an initial denaturation of 10 s at 92 °C , followed by 45 cycles combining 15 s at 92 °C, 10 s at

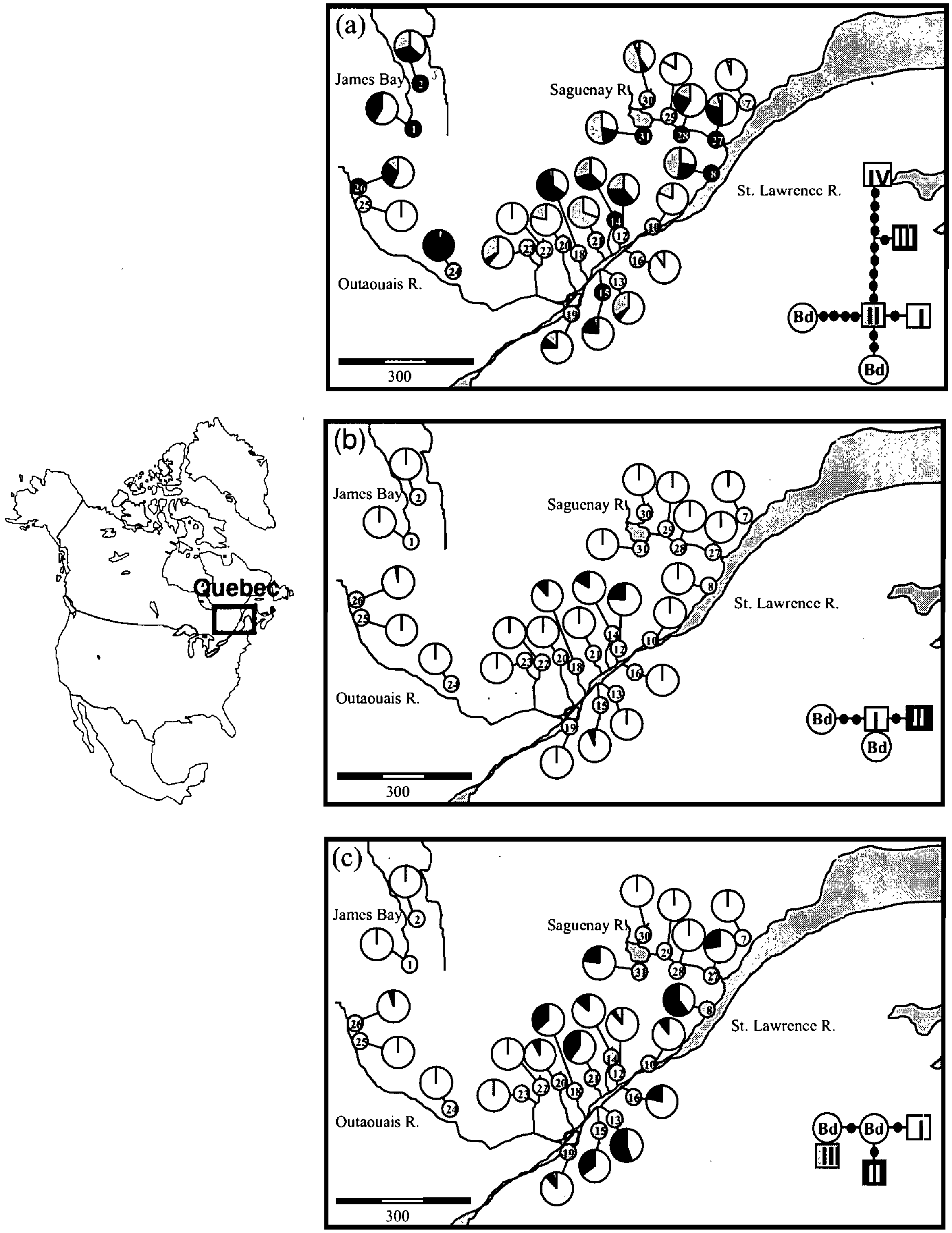

Quebec
(a)
James Bay
Saguenay R.
St. Lawrence R.
Outaouais R.
300
IV
III
II
I
Bd
Bd
(b)
James Bay
Saguenay R.
St. Lawrence R.
Outaouais R.
300
Bd
I
II
Bd
(c)
James Bay
Saguenay R.
St. Lawrence R.
Outaouais R.
300
Bd
Bd
I
III
II
1
2
7
8
10
12
13
14
15
16
18
19
20
21
22
23
24
25
26
27
28
29
30
31

Figure 24: Geographic position of populations combined to their MHC (a), GH (b) and TRY (c) composition. Colours of the identifiers of each population correspond to different Fu and Li test results. Black and grey identifiers represented respectively populations from F_{sign} and F_{ns}, while white identifiers are associated to populations for which the test was not performed (see text). Finally, the minimum spanning network between the Longnose dace and Blacknose dace (Bd) alleles is shown for each gene. Colours correspond to those presented in pie charts. The black circles correspond to mutation steps.

Tableau 10: Neutrality statistic tests computed on microsatellites and genes for each Longnose dace population. N is the number of individuals. Probabilities are shown for both Hardy-Weinberg (HW) and mutation-drift (MD) tests. Numbers in parentheses represented the number of loci highly deviating from HWE (p<0.001). The statistic value is shown for Tajima (D) and Fu and Li (F) analyses. Significant tests are presented in bold. Dashes (-) correspond to genes for which HW test could not be performed because of their low genetic diversity in the given population. Finally, the neutral structure according to Girard and Angers (2007c) is presented. The structure is composed of twelve microsatellite groups (A to L) divided in three clades according to their association to a given mitochondrial background (I, II and III). Three populations (20, 23 and 24) were not present in the Girard and Angers study. Their assignment to pure groups (*in italic*) was performed afterward.

		Microsatelllites					Genes											
		Tests		Structure			MHC				GH				TRY			
Populations	N	MD	HW	I	II	III	HW	MD	D	F	HW	MD	D	F	HW	MD	D	F
James' Bay bassin																		
1-Martel	22	ns	ns			L	0.40	0.08	3.17	**10.96**	fixed				fixed			
2-Coigny	23	ns	ns		J	L	**0.00**	**0.00**	2.74	**7.51**	fixed				fixed			
Saint Lawrence River bassin																		
7-Sault-aux-Cochons	17	ns	ns			K	-	0.27	**-2.33**	2.27	fixed				fixed			
8-Malbaie	16	ns	ns	B;C;D;E	I		0.36	**0.04**	1.90	**6.51**	fixed				0.64	0.11	2.07	3.61
10-Jacques-Cartier	18	ns	ns	A			1.00	0.40	0.41	0.84	fixed				1.00	0.51	-0.24	1.42
12-Des Envies	23	ns	**<0.01 (3)**	C;E	H													
13-Nicolet	19	ns	**<0.01 (3)**	D	H;I;J													
14-Mékinak du Nord	19	ns	ns	E			**0.03**	**0.01**	2.66	**7.12**	1.00	0.39	0.27	0.73	1.00	0.46	0.04	1.72
15-Yamaska	22	ns	ns	B;E	H;I;J		0.20	0.43	1.03	**5.11**	1.00	0.47	-0.60	-0.30	0.66	0.14	1.89	3.55
16-Bécancour	17	ns	ns	E	H;I;J		1.00	0.49	-0.31	0.13	fixed				1.00	0.29	0.96	2.69
18-Noire	22	ns	**<0.01 (3)**		I													
19-Châteauguay	18	ns	ns		J		0.32	0.49	-0.26	3.20	fixed				0.08	0.24	-0.78	0.03
20-Assomption	10	ns	ns		*G;H*	*L*	0.23	0.47	0.32	0.64	fixed				-	0.34	-1.45	0.43
21-Du Loup	15	ns	ns		H		0.12	0.21	1.19	1.48	fixed				0.35	0.11	2.03	3.55
Outaouais River bassin																		
22-Diable	15	ns	ns		G		fixed				fixed				fixed			
23-Rouge	13	ns	ns		*F;G*		0.64	0.45	-1.44	1.49	fixed				fixed			
24-Kazabazua	3.42	ns	ns		*F;G;H;J*	*L*	-	0.33	**-2.10**	2.22	fixed				fixed			
25-Lavallée	24	ns	ns		F		fixed				fixed				fixed			
26-Loutre	22	ns	**0.04 (0)**		F;J		0.08	0.11	2.44	**7.12**	-	0.27	-1.14	-1.26	-	0.26	**-1.48**	-0.74

Saguenay River bassin															
27-Petit-Saguenay	20	ns	**0.04 (0)**	D			**0.02**	0.22	0.95	**5.77**	fixed	0.62	0.23	1.41	3.09
28-À Mars	22	ns	ns	C			0.19	0.19	1.32	**4.65**	fixed	fixed			
29-Shipshaw	20	ns	ns	B			**0.00**	0.41	0.27	0.73	fixed	fixed			
30-Péribonka	15	ns	ns	B;D	H	K	0.15	0.28	-0.58	2.97	fixed	fixed			
31-Métabetchouane	21	ns	**<0.01 (1)**	C;D	I		0.06	0.06	1.16	**5.28**	fixed	1.00	0.28	1.04	2.75

54°C and 30 s at 68 °C, and a final extension of 120 s at 68 °C. Both strands were sequenced with a CEQ 2000XL DNA Analysis System (Beckman Coulter) on two individuals. According to these sequences, a primer closer to the second exon was designed (5'-TAATAACGGTAGCGCCTC-3') leading to the amplification of a segment of 295 bp (229 bp on the exon and 66 bp on the intron). PCR reaction and conditions that optimize this amplification were the same as above, but with an annealing temperature of 50 °C. The polymorphism of this locus was then screened using the single strand conformation polymorphism (SSCP, Orita et al. 1989). The locus was electrophoresed on a 6% nondenaturing gel for 10 hours at 20W in 0.5X TBE. To insure reliability of similar migration patterns among populations, one individual associated to each pattern observed within each population was sequenced. Alleles MHC-I, MHC-II, MHC-III and MHC-IV was thus sequences 24, 19, 16 and 4 times respectively.

Growth hormone and trypsin polymorphism

No clear information is available on the function and polymorphism of each exon of growth hormone (GH) and trypsin (TRY) in fish. Consequently, the choice of the regions to be analysed is not as straightforward as for the MHC. For each locus, we followed the next procedure to identify an appropriate DNA segment. Among the five exons that constituted each gene, the longer were targeted. This choice was based on a strict probabilistic assessment: a longer exon is most susceptible to accumulate mutations and thus to show polymorphism than a shorter one. Primers were thus designed according to homologous sequences available for other fish species in Genebank for the forth (~ 160 bp) and fifth (~ 201 bp) exons of GH, and for the second (~ 160 bp) and third (~ 260bp) ones of TRY. GH primers were designed according to conservation sequences found among very divergent species (*Cyprinus carpio* #X51969 and *Hypophthalmichthys molitrix* #M94348), while for TRY, primers were designed according to the genomic sequence #BX539313 of *Danio rerio*. Successful amplifications were performed on Longnose dace for all loci but fourth exon of GH, which was discarded. A SSCP screening was then performed on a total of 12 *R. cataractae* individuals from four distant populations (populations 1, 8, 19, 29) to evaluate the polymorphism potential of the three remaining loci. The fifth exon of GH and the third one of TRY showed more than one migration pattern. No variation was observed for the second exon of TRY and it was discarded. The PRC conditions of the two loci selected are presented below.

For GH, the primers 5'-CCATTTGTATCTGCACAG-3' and 5'-TACAGGCATTGACTAAC-3' successfully amplified a segment of 255bp covering the entire fifth exon plus 36bp of the following intron. In the case of TRY, primers pair 5'-CGTCTGGGTGAGCACAAC-3' and 5'-TCCCCATCCAGAGATCAGAC-3' amplified a segment of 222bp of the third exon. PCR reactions and conditions that optimize amplification were the same for both sets of primers. The following reaction was used: a 12.5 µL reaction aliquot contained 1.5 mmol·L^{-1} of $MgCl_2$, 2.5 nmol·L^{-1} of each dNTP, 0.2 unit of *Taq* polymerase, 1.25 µl of 10x *Taq* polymerase buffer (Invitrogen) and approximately 20 ng of DNA. Reaction conditions included an initial denaturation of 10 s at 92 °C , followed by 45 cycles combining 15 s at 92 °C, 10 s at 50 °C and 30 s at 68 °C , and a final extension of 120 s at 68 °C. The polymorphism of both genes was evaluated on all Longnose dace individuals following the same procedure as described above for MHC. Sequencing was performed for each individual showing variation in the migration pattern as explained above for MHC. GH-I, GH-II, TRY-I, TRY-II, TRY-III were sequenced 24, 5, 24, 14 and 2 times respectively.

Microsatellite polymorphism

Seven microsatellite loci, showing high to moderate mutation rates, were screened over all populations. Data were previously reported in Girard and Angers (2007c) and are summarized in Table 10.

Statistical analyses

Sequence analyses

Sequences of each exon were aligned using CLUSTAL W (Thompson et al. 1994). The substitutions neutrality was tested using the McDonald and Kreitman (MK) test (McDonald and Kreitman 1991), available in DNASP 4.10 (Rozas et al. 2003). Three and six homologous sequences of *Danio rerio* found on Genbank were used as the outgroup species for TRY and MHC respectively, while five from *Cyprinus carpio* were used for GH. These outgroup species were chosen because no fixed differences were observed between Longnose dace and Blacknose dace sequences, a condition required to compute MK test.

HKA tests (Hudson et al. 1987) were also performed between each pair of genes to explore the potential balancing selection effects. This test is based on the neutral theory of molecular evolution (Kimura 1983) which predicts that the rate of evolution of a DNA segment is correlated with the levels of polymorphism within a species. The test requires intra- and inter-specific polymorphism data. For this analysis, the genebank sequences of *Danio rerio* #NW001513985.1, NW001513092.1 and NW001511399.1 were used for MHC, GH and TRY respectively.

Gene diversity within populations

Genetic diversity within population was calculated using both allele frequencies and sequence information. The Nei's gene diversity (H_E) and the allelic richness (R) according to the rarefaction index of El Mousadik and Petit (1996) were computed with the program FSTAT v.2.9.3 (Goudet 2001). Molecular diversity indices within each sample were evaluated from the observed number of segregating sites (θ_S) and from the mean number of pairwise differences (θ_π). To evaluate the effect of demographic processes on the genetic diversity within population, the gene diversity indices computed on genes were correlated to R and H_E of microsatellite loci. The significance of this correlation was evaluated using a Pearson statistic. The null hypothesis ($r > 0$) was tested with 999 random permutations of the data.

Deviations from Hardy-Weinberg equilibrium (HWE) were tested in each population using the exact test of Guo and Thompson (1992) implemented in the software GENEPOP v3.4 (Raymond and Rousset 1995b). The accordance with neutral expectation was evaluated with the Ewen-Watterson neutrality test (Watterson 1978). Finally, the neutrality of mutations was evaluated with the D of Tajima (1989) and the F of Fu and Li (1993). These analyses were performed overall and within each sample using the program ARLEQUIN v3.1 (Schneider et al. 2000).

Gene diversity among populations

The effects of population structure on the allelic composition of genes were evaluated with redundancy analyses (RDA) using the *rda* function available in the *vegan* package of R. For each gene, a **Y** matrix containing the allele frequencies within each population was created. A **X** matrix describing the microsatellite structure among populations was then constructed based on the fractional contributions of the twelve nuclear groups defined by Girard and Angers (2007c) in each population. Data were transformed using the Hellinger correction (Legendre and Gallagher 2001). For each analysis, the adjusted R squared was evaluated. Test of significance was performed using 999 permutations with the *anova* function available in the *stats* package of R.

Overall F_{ST} was computed with microsatellites, MHC, GH and TRY using FSTAT. F_{ST} were chosen (instead of R_{ST} for example) following the same rationale as Landry and Bernatchez (2001), who had a similar dataset. The low number of microsatellite loci (Gaggiotti et al. 1999), the differences among sample sizes (Ruzzante 1998) and the recent (and likely independent from mutation processes) populations differentiation (Rousset 1996), are all conditions in which the use of R_{ST} is disfavoured. The neutral expectations of the overall population differentiation for MHC, GH, TRY and microsatellites were evaluated using the method proposed by Beaumont and Nichols (1996) and implemented in the software FDIST2. The average overall F_{ST} across markers (0.36) was used in the simulations. This value was targeted after an exploratory analysis based on the indications of Beaumont that can be found in the FDIST2 package. Accordingly, a robust starting F_{ST} value should yield to an output where the observations will be distributed almost equally below and above the average line describing the expected relationship between F_{ST} and H_E. Simulations were performed with either a stepwise mutation model (microsatellites) or an infinite allele model (MHC, GH and TRY). A total of 50000 simulations was performed to obtain a good compromise between precision and computation time. Each simulation was composed of 100 demes, which is the maximum value that can be computed by the software. Observed F_{ST} value for each marker was then compared to the simulation intervals according to their heterozygosity.

Results

Sequence analyses

Six distinct alleles were observed among the Longnose dace populations. Sequencing replicas, performed on at least two individuals from different populations that exhibit similar migration pattern, confirmed that SSCP and sequencing were reliable. Two of the alleles were highly divergent from the others and included at least one stop codon. These alleles were considered as potentially originating from a pseudo-gene and were discarded from the analyses. These alleles were present in low proportion and restricted to a single population (30). The four remaining MHC alleles (MHC-I to MHC-IV; Genbank #000000-000000) are characterized by a total of 11 polymorphic sites. Among them, six are singletons (65, 91, 95, 193, 214 and 218) and five are parsimony-informative (125, 126, 135, 139 and 144). Alleles differ by an average of 6.3 substitutions. Ten substitutions result in amino acid replacements, while only one was synonymous. The nucleotides diversity of the non-synonymous sites (0.032) is twice higher than the one of synonymous sites (0.014). In spite of an over representation of non-synonymous substitutions in polymorphic vs divergent sites, providing a NI (neutrality index) of 3.75, the MK test was not significant (G=1.46; p=0.23).

At the opposite of MHC, both GH and TRY show low polymorphism, mostly represented by synonymous substitutions. For instance, the GH exon was close to fixation among Longnose dace populations. Only two alleles (GH-I and GH-II; Genbank #000000-000000) differing by a single nonsynonymous mutation are observed, resulting in a nucleotide diversity of 0.004. The MK test was not significant (G=0.618; p=0.43). Three alleles are observed on the TRY exon (TRY-I to TRY-III; Genbank #000000-000000). These alleles diverge by a single synonymous (117) and two nonsynonymous (16 and 109) substitutions. The nucleotides diversity of the nonsynonymous sites (0.008) was slightly lower than the one of synonymous sites (0.013) and the MK test was not significant (G=0.387, p=0.533). In addition, none of the HKA tests provide significant result indicating that polymorphism of these genes is not under selection.

The minimum spanning network of MHC alleles revealed that their origin predates the speciation between the Longnose dace and the Blacknose dace (Figure 24a). MHC-I and MHC-II were closer to Blacknose dace alleles (average of 3.5 substitutions) than to MHC-III and MHC-IV (average of 8 substitutions; Figure 24a). Interestingly, the same pattern was observed for GH and TRY for which one allele of each gene showed the very same sequence in both species (Figure 24b-c).

Genetic diversity within populations

According to the microsatellites analyses, seven populations (9, 12, 13, 18, 26, 27 and 31) are not in accordance with HWE (Table 10). However, only three populations (12, 13 and 18) show a highly significant deviation ($p<0.001$) for more than one locus. Consequently, only these populations were considered as deviating from panmixia and were removed from the further analyses of genetic diversity within populations. Furthermore, three other populations were previously identified as being slightly in discordance with mutation-drift equilibrium (Table 10; see Girard and Angers 2007c for further details). These populations were also removed.

Genetic diversity within populations varies significantly among the three functional genes (Table 10). For instance, R computed on MHC, GH and TRY average of 2.4, 1.2 and 1.6 alleles by population respectively. This difference is highly significant ($F=21.19$; $p<0.0001$). Significant differences are also observed for H_E ($F=14.34$; $p<0.001$), θ_S ($F=41.27$; $p<0.0001$) and θ_π ($F=30.93$; $p<0.0001$). These differences are mainly explained by the MHC exon for which gene and nucleotide diversities are higher than for those of GH and TRY.

GH and TRY are fixed or in accordance with both HW and mutation-drift equilibriums. At the opposite, MHC exon shows an excess of homozygotes (27 and 29), an excess of diversity (8) or both (2 and 14). Results obtained with the D of Tajima are not significant for almost all populations except two for MHC (7 and 24) and one for TRY (26; Table 10). Furthermore, the overall D for MHC (1.34; $p=0.08$), GH (-0.40; $p=0.31$) and TRY (0.81; $p=0.19$) are not significant. The F of Fu and Li show high discordances between MHC and the two other genes. Nine populations (hereafter referred as the F_{sign}

subset) show a lower number of MHC alleles than expected under neutrality according to the observed molecular diversity. At the opposite, ten (the F_{ns} subset) followed the neutral expectation (Table 10). This result is mainly explained by the high number of non-synonymous substitution differences between each pair of alleles. Indeed, between 1.24 (pop. 15) to 1.75 (pop. 1) non-synonymous substitutions are observed within F_{sign} subset comparing to 0.27 (pop. 16) to 1.29 (pop. 30) in the F_{ns} one (Figure 25). The difference between the averages was highly significant ($t=7.591$; $p<0.0001$). These results suggest some sort of balancing selection. This signal (or the opposite associated to purifying selection) is not observed both within populations (Table 10) and overall, for either GH ($F=-0.42$; $p=0.21$) or TRY ($F=2.76$; $p=0.14$).

The MHC richness is significantly correlated to the one of microsatellites ($r=0.42$; $p<0.05$; Figure 26a). While the correlation remains significant when only the F_{ns} populations are taking into account ($r=0.53$; $p<0.05$; Figure 26a), a complete absence of correlation is observed when considering the F_{sign} subset ($r=-0.002$; $p=0.55$). Similar trends are observed for H_E. The correlation between microsatellites and MHC is significant when considering all the populations. Removing the F_{sign} subset highly increases the relationship between MHC and the diversity of neutral markers ($r = 0.53$; $p<0.01$). The MHC and neutral H_E are completely independent in the case of the F_{sign} populations ($r = -0.29$; $p = 0.81$; Figure 26b). Populations may thus be divided in two different groups according to the selective (F_{sign} subset) or the neutral (F_{ns} subset) evolution of MHC. This suggests a complex interaction of stochastic and deterministic processes that vary from a population to another.

While the results of GH are not very informative due to its low level of diversity (Figure 26c,d), it worth noticing that TRY indices of diversity display strong correlations with those of microsatellites. The results of TRY, similar to those observed on MHC on the F_{ns} subset (Figure 26e and f), suggest that this gene evolves mainly under stochasticprocesses, at least over a short term evolutionary scale.

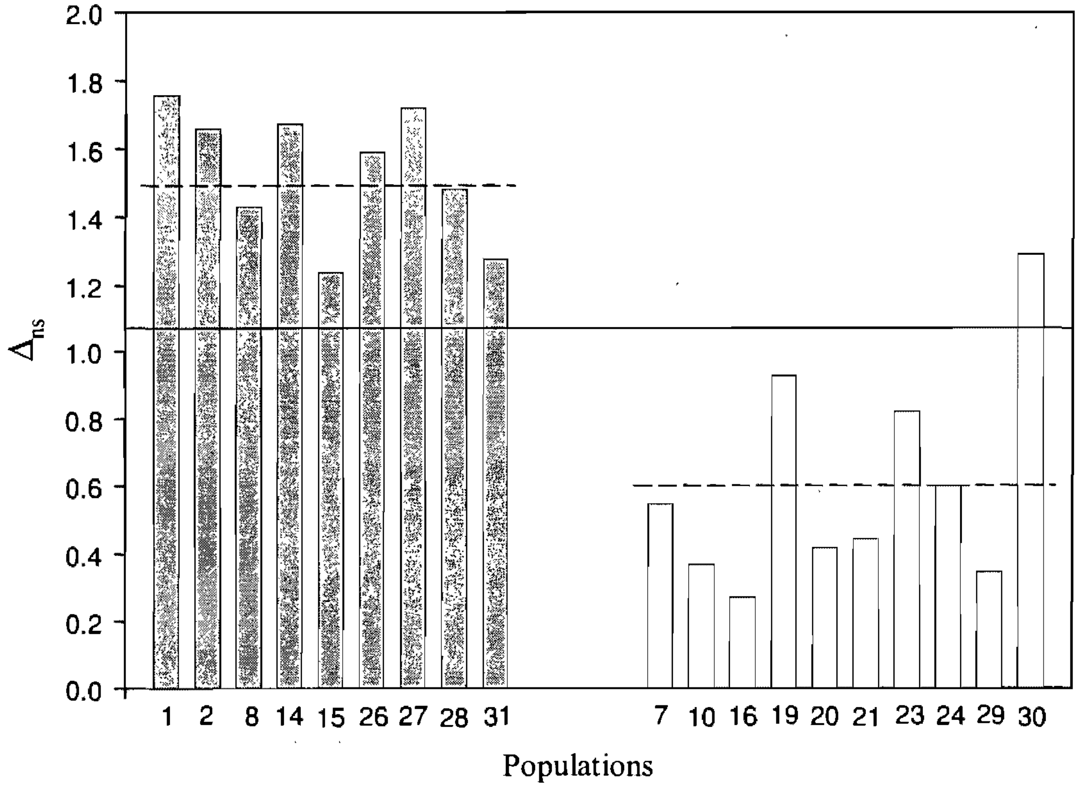


Figure 25: Average number of nonsynonymous substitutions between each pair of MHC alleles within population of F_{sign} (grey) F_{ns} (white) subsets. Within subsets (dash lines) and overall averages (solid line) are also presented.

Gene diversity among populations

The redundancy analyses reveals an absence of correlation between the MHC diversity among populations and microsatellite structure (F = 0.31, p = 0.29). Unfortunately, the analysis could not be performed on each F subset, because the number of explanatory variables equalled or exceed the number of populations (9 and 11 microsatellite groups compared to 9 and 10 populations for F_{sign} and F_{ns} respectively). However, the analysis can be performed using the three major nuclear groups defined by Girard and Angers (2007c) and defined as I, II and III in Table 10. Neither F_{sign} (F = 1.03, p = 0.37) nor F_{ns} (F = 0.29, p = 0.71) groups show significant relationships with microsatellite structure. Interestingly, the analyses performed on the two other genes are consistent with the results obtained with within population diversity. While the absence of relationship between GH and the neutral structure must be interpreted with caution because of its very low diversity (F = 0.08; p = 0.745), TRY diversity appear strongly associated to the genetic structure inferred from microsatellites (F = 0.49; $p<0.05$).

The simulation procedures performed using FDist2 show that two microsatellites (Rhca52 and Rhca34) are marginally outside of the 0.99 interval and thus slightly deviate from neutral expectations (Figure 27a). When taking into account all populations, MHC, GH and TRY all fall within the neutral interval (Figure 27b). The F_{ST} computed overall populations of the F_{ns} subset for MHC is in accordance with neutral expectations. However, the F_{ST} of the F_{sign} subset is clearly below the lower limit of the neutral confidence interval (Figure 27b). The lower population differentiation than expected is consistent with predictions of either heterozygotes advantage or negative frequency-dependant selection.

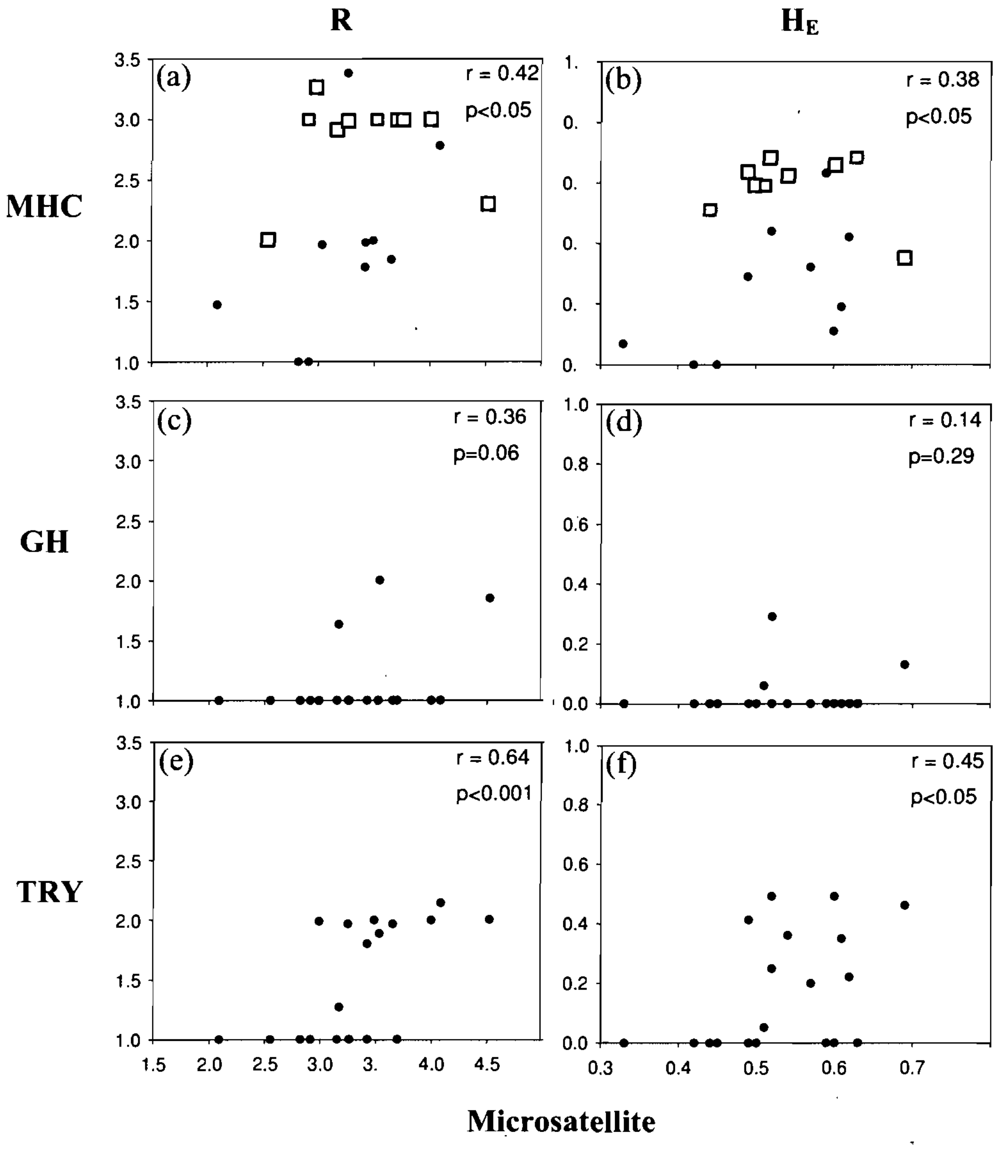


Figure 26: Relationships between genes and microsatellites R and H_E within each population. White squares and black circles represented populations within F_{sign} and F_{ns} subsets respectively. Correlation values were calculated using all the populations.

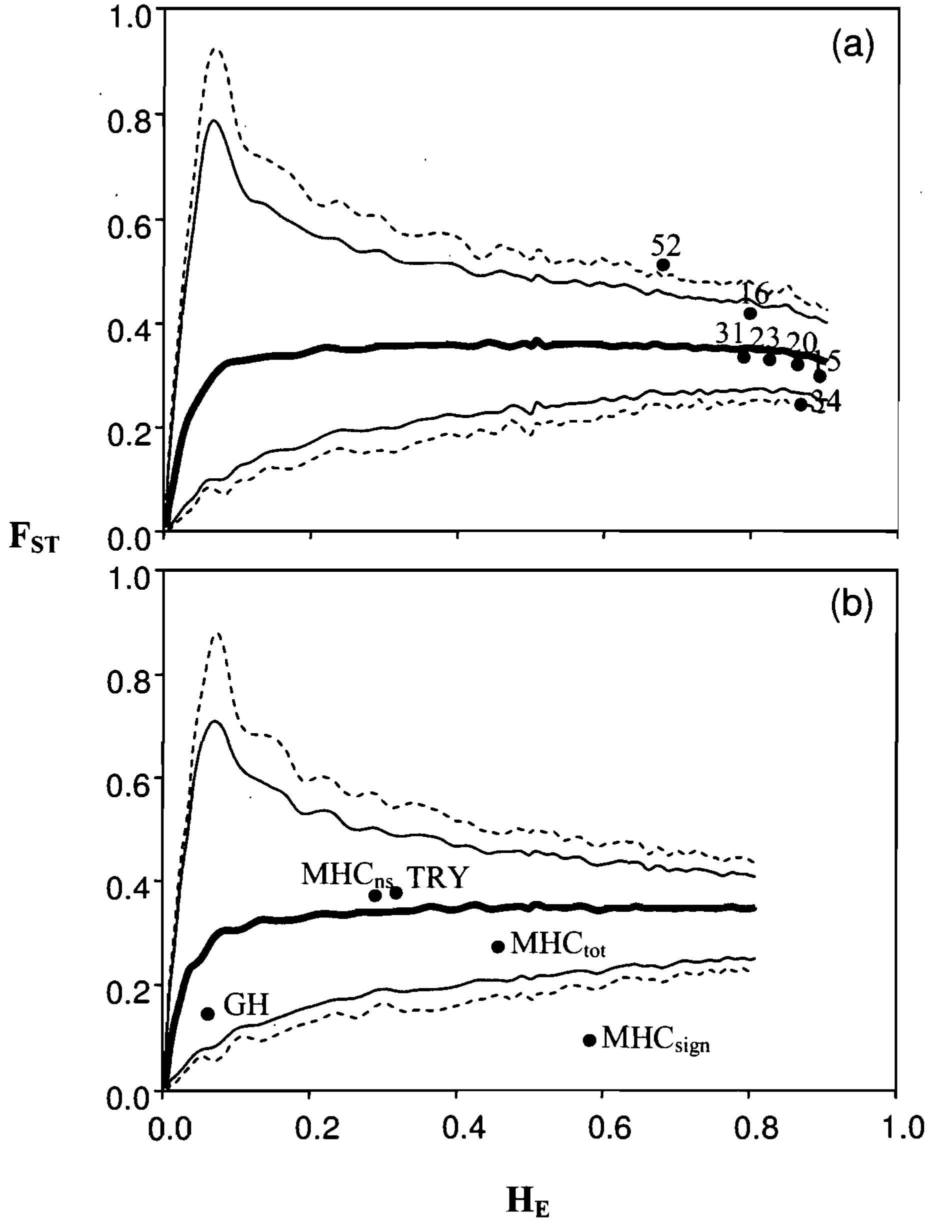


Figure 27: Average (bold line), and 95% (solid lines) and 99% (dash lines) confidence intervals of the relationships between F_{ST} and H_E according to FDIST2 simulations performed on SMM (a) and IAM (b) mutation models. Observed values of the seven microsatellites and the three genes are superimposed to these simulations results. In the case of MHC, values including all populations (MHC_{tot}), only F_{ns} (MHC_{ns}) and F_{sing} (MHC_{sing}) subsets are presented.

Discussion

Long-term processes

The sequence analyses performed in this work revealed a high MHC diversity, expressed as a high ratio of non-synonymous/synonymous substitutions. Such imbalance in favour of non-synonymous substitutions usually supports an historical effect of diversifying selection acting specifically on MHC alleles (Hughes and Nei 1988, Froeschke and Sommer 2005, Schad et al. 2005, Aguilar and Garza 2006). However, as suggested by the nonsignificant MK test, the effect of stochastic mutations cannot be totally rejected. Indeed, the ratio may be lowered by the absence of novel non-synonymous mutations and/or the accumulation of synonymous mutations since the divergence of MHC alleles (Van Oosterhout et al. 2006). HKA tests also failed to detect other forms of balancing selection (such as overdominance or negative frequency-dependant selection) in any of the genes studied, suggesting that neutral molecular evolution might play a more important role than expected.

The incomplete MHC lineage sorting observed between Longnose dace and Blacknose dace must be interpreted with caution. The maintenance of alleles (or allele lineages) through speciation events is often associated to the action of balancing selection (Mayer et al. 1988, O'brien and Yuhki 1999, Bryja et al. 2006). However, the same results were also observed on GH and TRY genes. In addition, conservation of primer sites and overlap of allelic size ranges between both species was observed on 10 microsatellites loci (Girard and Angers 2006a). Such similarity throughout the genome is not in accordance with selection and thus suggests other evolutionary processes to explain incomplete lineage sorting on MHC. To our knowledge, the hybridization between both *Rhinichthys* species has never been reported. Consequently, large effective population size and recent speciation appeared likelier to explain incomplete lineage sorting between both species (Nei 1987; Pamilo and Nei 1988; Takahata 1989). While no estimation of Longnose dace effective population size is available, field observations revealed higher densities for dace than the other species present in the same locations (Thompson et al. 2001, Reida et al. 2005). These observations, combined to the recent speciation between Longnose and Blacknose daces (approximately 6 million years assuming 1% divergence per million years on cytochrome b sequences #AF452078 and DQ990251.1), support this hypothesis.

Altogether, these results on Longnose dace populations suggest that demographic and stochastic events mainly drove the long term evolution of MHC. Such importance of neutral mechanisms is in discordance with previous observations in populations living in natural environment (Slade 1992, Landry and Bernatchez 2001, Miller et al. 2001). However, these results may be explained at least partly by the restrict number of alleles observed within our population set. Alleles from other regions should be sampled in order to reach further reliable conclusions about long-term neutral or deterministic evolution of MHC in Longnose dace.

Short term processes

The distribution of populations in different groups according to selection evolution or neutral expectations for the MHC diversity is one of the most striking results of this work. The genetic and molecular diversities appeared strongly related to the one observed on neutral markers in almost half of the sampled populations (F_{ns} populations). In these populations, we cannot reject the hypothesis that MHC evolution resulted mainly from random effects associated to demographic processes following the postglacial colonization. On the other hand, the MHC gene and molecular diversities of nine populations (F_{sign} populations) appears to depart from neutral expectations. Significant F tests, lower population differentiation than expected, and the absence of correlation between diversity of MHC and microsatellite, reflect the effects of a balancing selection process.

The causes of such balancing selection are unclear in these populations because of the absence of evaluations of the habitat conditions. However, according to previous works, two main hypotheses can be proposed. The first one is a pathogens-mediated selection that should favour heterozygotes (Doherty and Zinkernagel 1975, Hughes and Hughes 1995) or carriers of rare alleles (Parham and Ohta 1996, Gilbert et al. 1998) in habitats showing a high diversity of pathogens. A large amount of empirical evidences demonstrated the role of MHC genetic variation into the host capacity to defend against a large range of parasites (Paterson 2005 and references therein). This is in accordance with observations of specific associations between MHC alleles and pathogens resistance (Patterson et al. 1998, Langefors et al. 2001, Westerdahl 2004). According to this hypothesis, the results obtained with Longnose dace should be explained by the fact that i) populations showing balancing selection lived in highly pathogenic habitats and ii) they faced the same kind of pathogens since they shared mainly the same three alleles (MHC-I, MHC-II and MHC-III). The second main hypothesis is that diversity is naturally maintained through mating behaviour (Potts and Wakeland 1990, Jordan and Bruford 1998, Penn and Potts 1999). Even though this hypothesis remains associated to a defence mechanism against pathogens, the key lies in the fact that individuals choose in function of the available MHC diversity and not of the environmental conditions. Consequently, balancing selection will act even in habitats that do not show high selection pressures. However, this nonrandom mating hypothesis implies deviation from HWE, which is not supported by our data. This discordance between selectively-allelic and neutral-genotypic frequencies is unusual. It could suggest that selection processes are not constant through time within a given population. A similar signal observed on MHC by Hayashi et al. (2005) was interpreted as the result of natural selection on long-term evolution of MHC, while drift is the main process acting on within short-term periods. The effects of selection on allele frequency distribution would then persist for long time, while those on genotypic frequencies remain only during selection period.

Considering the importance of co-evolution in host-pathogen systems such as the one involving MHC, the allelic diversity is of prime importance for the viability of populations. In this perspective, the MHC richness (4 alleles) observed on Longnose dace is surprisingly low when compared to the one observed on other species (Froeschke and Sommer 2005, Aguilar and Garza 2006, Ottova et al. 2007). The explanation of this low diversity is not clear. It may be caused by a decrease during the Pleistocene glaciations and/or during the postglacial colonization (Bernatchez and Wilson 1998, Avise et al. 1984). Another hypothesis may be related to the fact that our populations are located on the northern part of North America. Previous works have reported low MHC polymorphism in higher latitude possibly as a result of low parasite diversity (Mainguy et al. 2007). However, additional analyses on the MHC composition on the southern populations of Longnose dace as well as an exhaustive description of the parasite fauna in both northern and southern populations are necessary to confirm or not this hypothesis.

In conclusion, the explanation of the among-population variation between stochastic and deterministic processes as well as the low MHC richness remained largely unexplained. However, even though Longnose dace MHC evolves following a complex evolutionary trajectory, its low variability may greatly affect the viability of populations in regards of introduced pathogens or northward shifts in the distribution of pathogens with global climate warming.

Conclusion

L'analyse conjointe de marqueurs nucléaires neutres ou sous sélection, ainsi que de régions du génome mitochondrial a permis d'identifier les mécanismes impliqués dans l'évolution à court et long termes du naseux des rapides. Les résultats présentés dans cette thèse ont confirmé l'intérêt de l'utilisation de différents types de marqueurs pour décortiquer l'action de ces processus évolutifs. Cette thèse a également démontré les avantages et les limites de plusieurs outils permettant d'analyser la diversité génétique à l'intérieur ainsi qu'entre les populations naturelles.

Développement des marqueurs microsatellites

Abondants chez la plupart des Eucaryotes (Jarne et Lagoda 1996) et très polymorphes (Weber et Wong 1993), les microsatellites sont aujourd'hui considérés incontournables dans les études de génétiques des populations (Avise 2004). Ainsi, l'absence de microsatellites disponibles chez une espèce limite la gamme d'études évolutives potentielles. Or, malgré ses qualités indéniables en tant que modèle, le naseux des rapides n'avait pas fait l'objet récemment d'étude de génétiques des populations (la seule étude étant de Merritt et al. (1978)). Le travail effectué dans le cadre de cette thèse au niveau du développement de microsatellites spécifiquement conçu pour le naseux des rapides a donc deux utilités majeures. Non seulement, a-t-il permis d'aborder les problématiques visées, mais l'amplification interspécifique effectuée chez 5 autres espèces sœurs représente un pas important permettant des analyses en génétique des populations chez d'autres espèces de Leuciscinae. Ces marqueurs ont d'ailleurs été utilisés depuis dans au moins une publication récente (Skalski et Grose 2006).

L'utilisation de tels marqueurs chez une espèce, et particulièrement s'ils ont été développés pour une espèce parente, peut s'avérer parfois hasardeuse. En effet, il n'est pas rare que l'accumulation des mutations ponctuelles dans les régions flanquantes des microsatellites prévienne l'amplification de certains allèles (Callen et al. 1993). Ces allèles dits "nuls" représentent un problème de taille. Leur présence cause une apparente hérédité non-Mendellienne qui introduit des biais importants dans l'interpértation des diversités intra- et inter-populationnelles (Pemberton et al. 1995, Bowling et al. 1997, Avise 2004, Chapuis and Estoup 2007). C'est dans ce cadre que s'insèrent nos travaux sur la puissance et la précision des outils statistiques conçus pour détecter ces allèles. Les résultats de ces travaux ont démontré le manque de puissance de ces analyses. Or, une rapide recherche dans la littérature démontre l'utilisation de plus en plus répandue de ces analyses (Karlsson et Mork 2005, Green et al. 2006, Hänfling et Weetman 2006, Yawson et al. 2007). La disponibilité de logiciels gratuits, tels Micro-Checker, qui effectuent en quelques clics de souris seulement, la détection et l'estimation des allèles nuls, ainsi que la correction des bases de données originales à l'aide de ces outils, explique sans doute cette situation. Notre travail sur les allèles nuls est d'autant plus pertinent que nous avons proposé une marche à suivre simple et efficace, permettant de corriger ces artefacts et ainsi minimiser l'introduction de biais lors de l'analyse du polymorphisme génétique.

Colonisation post-glaciaire du naseux des rapides

Le génome mitochondrial présente des séquences de choix permettant d'évaluer les impacts des processus historiques sur la structure actuelle des populations naturelles. C'est en effet par une analyse combinée de la région de contrôle, très variable, et d'un gène (cytochrome b) dont les pressions de sélection purificatrice rendent moins polymorphe, que les origines et la séquence de la colonisation postglaciaire du naseux des rapides ont pu être clairement établies. Or, la grande variabilité observée chez le naseux des rapides au niveau de ces régions mitochondriales, combinée à la grande surface échantillonnée, ont permis d'acquérir une précision inédite au niveau de la phylogéographie d'une espèce de poisson dans l'est de l'Amérique du Nord. Cette précision aura mené à une compréhension beaucoup plus large des composantes spatiales et temporelles de la colonisation postglaciaire des poissons d'eau douce au niveau du territoire échantillonné. Tout en confirmant l'importance du refuge mississippien en tant qu'origine, ainsi que des tributaires de la Baie James et des Grands Lacs comme portes d'entrée sur la péninsule québécoise, nos résultats représentent la première démonstration de l'importance des invasions marines comme barrière structurante sur une espèce de poissons d'eau douce.

Les effets de ces invasions marines sur la diversité génétique des populations sont nombreux. En effet, l'homogénéisation de plusieurs vagues de colonisateurs à l'intérieur des zones inondées a eu pour effet d'augmenter la diversité à l'intérieur des populations tout en minimisant la structure entre elles. Ces résultats contrastent de manière importante avec la composition génétique très structurée retrouvée à l'extérieur des marges des mers postglaciaires. Dans un cadre théorique, ces résultats démontrent l'importance d'intégrer ces invasions marines en tant que paramètres structurants dans les modèles décrivant la colonisation post-glaciaire des espèces d'eau douce. Au niveau pratique, ces résultats apparaissent primordiaux pour l'interprétation de la diversité génétique du génome nucléaire, sur lequel leurs effets se feront également sentir.

La méthode POST

Les microsatellites restent sans aucun doute des marqueurs de choix pour l'analyse du polymorphisme à court et moyen terme d'un groupe d'individus. Cependant, les outils disponibles pour gérer plusieurs loci à la fois ne permettent pas d'aborder certains problèmes au niveau de l'organisation des populations, tels la détection du nombre de groupes distincts ainsi que la présence de chevauchements entre eux. Or, la méthode que nous avons développée a démontré son efficacité sur l'omble de Fontaine du Parc National de la Mauricie pour identifier avec précision et sans *a priori* la structure génétique de populations. D'autre part, l'analyse des données génétiques amassées sur le naseux des rapides à l'aide de cette méthode a permis de mettre en lumière la fragmentation de la diversité génétique en plusieurs groupes fondateurs lors de la colonisation des parties nord et est du territoire québécois. Les résultats obtenus ont démontré la capacité de POST à décrire une structure complexe telle que celle du naseux et d'en inférer des processus dont les effets n'avaient pas été détectés au préalable à l'aide des marqueurs mitochondriaux.

La précision et la fiabilité des résultats de POST permettent d'entrevoir le développement d'un cadre analytique puissant permettant de décrire l'organisation géographique de la structure génétique des populations. L'utilisation subséquente des résultats de POST dans une analyse de redondance pourrait en effet permettre d'étendre la méthode locus-par-locus proposée par Angers et al. (1999) à une perspective multilocus et ainsi mettre en lumière les processus structurants impliqués dans le façonnement de la diversité génétique entre les populations naturelles.

Évolution du CMH chez le naseux des rapides

La comparaison entre le deuxième exon du CMH et ceux de l'hormone de croissance et de la trypsine d'un côté et les marqueurs neutres de l'autre ont confirmé la présence d'une mécanique évolutive spécifique et complexe agissant sur le CMH. D'une part, l'analyse des séquences ne suggére pas de rôle historique évident de la sélection. D'autre part, la comparaison des trois gènes fonctionnels démontre que les mécanismes évolutifs ayant agi sur le CMH diffèrent de ceux des deux autres.

Paradoxalement, les résultats présentés dans cette thèse suggèrent un effet plus prononcé des processus neutres sur une échelle de temps plus courte dans certaines populations. Les corrélations positives (ou à tout le moins tendancieuses) entre la diversité des gènes et ceux des marqueurs neutres, de même que l'accord avec les équilibres mutations-dérives et de Hardy-Weinberg, suggèrent une activité significative des processus aléatoires sur la composition génétique actuelle dans ces populations au niveau du CMH et des gènes fonctionnels en général. Cependant, le processus de sélection sur le CMH semble également apparent dans un certain nombre de populations dans lesquelles des allèles CMH très divergents sont maintenus. L'absence d'information sur les conditions environnementales dans lesquelles ces populations évoluent n'a malheureusement pas permis de déterminer les causes de cette évolution différentielle.

Implications

L'analyse du polymorphisme génétique actuel d'une population permet d'identifier les différents mécanismes qui en sont la source. Or, la diversité actuelle du naseux des rapides sur les trois gènes fonctionnels est très faible. Compte tenu de l'importance jouée entre autres par le CMH sur l'adaptation locale des populations (revue par Bernatchez et Landry 2003), une telle diversité pourrait mettre en péril la viabilité des populations de naseux des rapides.

L'implication de tels résultats reste à évaluer. Cependant, il est possible que la sélection agisse non pas sur les gènes mais sur leur expression. De récents travaux sur la sélection des promoteurs (Dunbar et al. 2007, Evgen'ev et al. 2007, Gregori et al. 2007) ou sur les processus épigénétiques (Balon 2002, Richards 2006) s'insèrent dans cette lignée. La question de l'adaptation locale des populations reste donc, pour l'instant, entière car les résultats de cette thèse ne permettent pas de jeter un éclairage total sur les mécanismes moléculaires d'adaptations agissant sur de très courtes périodes de temps.

Ceci dit, la viabilité d'une population dans un environnement changeant est probablement fonction de sa capacité adaptative. Il est également raisonnable de croire que celle-ci est une fonction de la diversité génétique (et/ou épigénétique). Selon ces deux prémisses, une population plus diversifiée génétiquement verrait ses probabilités de survie dans un environnement changeant augmenter. Or, à défaut de bien comprendre les relations entre l'environnement et la composition génétique des populations, une stratégie de conservation efficace pourrait mettre de l'avant les processus évolutifs neutres tels la dérive et le degré de migration. En effet, il est permis de croire qu'un habitat propice à la conservation n'est pas simplement fonction de caractéristiques environnementales mais favorise également la taille effective des populations et les interconnections entre elles. Ce genre de stratégie permettrait le maintien de la diversité génétique des populations et augmenterait leur viabilité face aux changements environnementaux.

Perspectives

Le développement de tels modèles de conservation nécessiterait cependant une meilleure compréhension des dynamiques populationnelles agissant à une échelle plus grande que celle de la population (Hanski and Gaggiotti 2004, Bohrer et al. 2005). Or, les modèles standards de dynamiques et génétiques des populations deviennent caducs dans un tel système, car ils assument que tous les individus ont une probabilité égale d'interagir les uns avec les autres. Cependant, en combinant des analyses génétiques et de structure du paysage dans une perspective métapopulationnelle, il serait sans doute possible de construire un cadre d'étude approprié permettant de comprendre les dynamiques d'un tel système. Les résultats de telles études pourraient ainsi mener au développement de modèle réaliste de conservation d'habitat favorisant le maintien de la diversité génétique et la capacité adaptative des populations naturelles.

Références

Alexandrino P, Faria R, Linhares D, Castro F, Le Corre M, Sabatie R, Bagliniere JL, Weiss S (2006) Interspecific differentiation and intraspecific substructure in two closely related clupeids with extensive hybridization, *Alosa alosa* and *Alosa fallax*. J. Fish Biol. 69(Suppl B):242-259

Allendorf F, Ryman N, Utter F (1987) Genetics and fishery management. Past, present and future. In: Ryman N, Utter F (eds) Population genetics and fishery management. Washington Sea Grant Program, Seattle, pp 1-19

Angers B, Bernatchez L (1998) Combined use of SMM and non-SMM methods to infer fine structure and evolutionary history of closely related Brook Charr (*Slavelinus fontinalis*, Salomidae) populations from microsatellites. Mol. Biol. Evol. 15:143-159

Angers B, Magnan P, Plante M, Bernatchez L (1999) Canonical correspondence analysis for estimating spatial and environmental effects on microsatellite gene diversity in brook charr (*Salvelinus fontinalis*). Mol. Ecol. 8:1043-1054

Ardren WR, Borer S, Thrower F, Joyce JE, Kapuscincki AR (1999) Inheritance of 12 microsatellites loci in *Oncorhynchus mykiss*. J. Hered. 90:529-536

Arnegard ME, Bogdanowicz SM, Hopkins CD (2005) Multiple cases of striking genetic similarity between alternate electric fish signal morphs in sympatry. Evolution 59:324-343

Austin JD, Lougheed SC, Neidrauer L, Chek AA, Boag PT (2002) Cryptic lineages in a small frog: the post-glacial history of the spring peeper, *Pseudacris crucifer* (Anura: Hylidae). Mol. Phylogenet. Evol. 25:316-329

Avise JC (2004) Molecular markers, natural history and evolution, 2nd edn. Sinauer Associates Inc., Sunderland

Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T, Neigel JE, Reeb CA, Saunders NC (1987) Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Annu. Rev. Ecol. Syst. 18:489-522

Avise JC, Neigel JE, Arnold J (1984) Demographic influences on mitochondrial DNA survivorship in animal populations. J. Mol. Evol. 20:99-103

Avise JC, Walker DJ, G.C. (1998) Speciation durations and Pleistocene effects on vertebrate phylogeography. Proc R Soc London, B 265:1707-1712

Bailey RM, Smith GR (1981) Origin and geography of fish fauna of the Laurentian Great Lakes basin. Can. J. Fish. Aquat. Sci. 38:1539-1561

Balloux F (2001) EASYPOP (version 1.7) A computer program for the simulation of population genetics. J. Hered. 92:301-302

Balon EK (1990) Epigenesis of an epigeneticist: the development of some alternative concepts on the early ontogeny and evolution of fishes. Guelph Ichthyology Reviews 1:1-48

Balon EK (2002) Epigenetic processes, when natura non facit saltum becomes a myth, and alternative ontogenies a mechanism of evolution. Environ. Biol. Fishes 65:1-35

Beaudry JR (1985) Génétique générale, Décarie Inc. edn, Mont-Royal

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. Proc. R. Soc. Biol. Sci. Ser. B 263:1619-1626

Behrmann-Godel J, Gerlach G, Eckmann R (2006) Kin and population recognition in sympatric Lake Constance perch (Perca fluviatilis L.): can assortative shoaling drive population divergence? Behav. Ecol. Sociobiol. 59:461-468

Bernatchez L (1997) Mitochondrial DNA analysis confirms the existence of two glacial races of rainbow smelt (*Osmerus mordax*) and their reproductive isolation in the Saint Lawrence R. Estuary (Québec, Canada). Mol. Ecol. 7:73-83

Bernatchez L, Dodson JJ (1990) Mitochondrial DNA variation among anadromous population of cisco (*Coregonus artedii*) as revealed by restriction analysis. Can. J. Fish. Aquat. Sci. 47:533-543

Bernatchez L, Dodson JJ (1991) Phylogenetic structure in mitochondrial DNA of the lake whitefish (*Coregonus clupeaformis*) and its relation to pleistocene glaciations. Evolution 45:1016-1035

Bernatchez L, Giroux M (2000) Guide des poissons d'eau douce du Québec et leur distribution dans l'Est du Canada., Broquet edn, St-Constant

Bernatchez L, Landry C (2003) MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? J. Evol. Biol. 16:363-377

Bernatchez L, Wilson C (1998) Comparative phylogeography of nearctic and palearctic fishes. Mol. Ecol. 7:431-452

Berry A, Kreitman M (1993) Molecular analysis of an allozyme cline: alcohol dehydrogenase in *Drosophila melanogaster* on the East coast of North America. Genetics 134:869-893

Bertness MD, Gaines SD (1993) Larval dispersal and local adaptation in acorn barnacles. Evolution 47:316-320

Bertorelle G, Excoffier L (1998) Inferring admixture proportions from molecular data. Mol. Biol. Evol. 15:1298-1311

Billiard S, Castric V, Vekemans X (2007) A general model to explore complex dominance patterns in plant sporophytic self-incompatibility systems. Genetics 175:1351-1369

Billington N, Hebert PDN (1988) Mitochondrial DNA variation in Great Lakeswalleye (*Stizostedion vitreum*) populations. Can. J. Fish. Aquat. Sci. 45:643-654

Bohrer G, Nathan R, Volis S (2005) Effects of long-distance dispersal for metapopulation survival and genetic structure at ecological time and spatial scales. J. Ecol. 93:1029-1040

Borghans JAM, Beltman JB, De Boer RJ (2004) MHC polymorphism under host-pathogen coevolution. Immunogenetics 55:732-739

Bouxin G (2005) Ginkgo, a multivariate analysis package. J. Veg. Sci. 16:355-359

Bowling AT, Eggleston-Stoot ML, Byrns G, Clark RS, Dileanis S, Wictum E (1997) Validation of microsatellite markers for routine horse parentage testing. Anim. Genet.:247-252

Bronikowski AM (2000) Experimental evidence for the adaptive evolution of growth rate in the garter snake *Thamnophis elegans*. Evolution 54:1760-1767

Brookfield JFY (1996) A simple new method for estimating FNULL from heterozygote deficiency. Mol. Ecol. 5:453-455

Brown JR, Beckenbach AT, Smith MJ (1992) Influence of Pleistocene glaciations and human intervention upon mitochondrial DNA diversity in white sturgeon (*Acipenser transmontanus*) populations. Can. J. Fish. Aquat. Sci. 49:358-367

Brown WM, George M, Jr., Wilson AC (1979) Rapid evolution of animal mitochondrial DNA. Proceedings of the National Academy of Sciences, USA 76:1967-1971

Bryja J, Galan M, Charbonnel N, Cosson JF (2006) Duplication, balancing selection and trans-species evolution explain the high levels of polymorphism of the DQA MHC class II gene in voles (Arvicolinae). Immunogenetics 58:191-202

Buri P (1956) Gene frequency in small populations of mutant *Drosophila*. Evolution 10:367-402

Callen DF, Thompson AD, Shen Y, Phillips HA, Richards RI, Mulley JC, Sutherland GR (1993) Incidence and origin of "null" alleles in the (AC)n microsatellite markers. Am. J. Hum. Genet. 52:922-927

Castric V, Bernatchez L (2003) The rise and fall of isolation by distance in the anadromous brook charr (Salvelinus fontinalis Mitchill). Genetics 163:983-996

Castric V, Bonney FB, Bernatchez L (2001) Landscape structure and hierarchical genetic diversity in *Salvelinus fontinalis*. Evolution 55:1016-1028

Cavalli-Sforza L, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. Evolution 32:550-570

Cereb N, Hughes AL, Yang SY (1997) Locus-specific conservation of the HLA class I introns by intra-locus homogenization. Immunogenetics 47:30-36

Chakraborty R, De Andrade M, Daiger SP, Budowle B (1992) Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications. Ann. Hum. Genet. 56:45-57

Chapuis MP, Estoup A (2007) Microsatellite Null Alleles and Estimation of Population Differentiation. Mol. Biol. Evol. 24:621-631

Clauss MJ, Mitchell-Olds T (2006) Population genetic structure of *Arabidopsis lyrata* in Europe. Mol. Ecol. 15:2753-2766

Clay Greem M, Waits JL, Avery ML, Tobin ME, Leberg PL (2006) Microsatellite Variation of Double-Crested Cormorant Populations in Eastern North America. J. Wildl. Manag. 70:579-583

Cockerham CC (1969) Variance of gene frequencies. Evolution 23:72-83

Cockerham CC (1973) Analysis of gene frequencies. Genetics 74:679-700

Cornuet JM, Luikart G (1996) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. Genetics 144:2001-2014

Cornuet JM, Piry S, Luikart G, Estoup A, Solignac M (1999) New methods employing multilocus genotypes to select or exclude populations as origins of individuals. Genetics 153:1989-2000

Costello AB, Down TE, Pollard SM, Pacas CJ, Taylor EB (2003) The influence of history and contemporary stream hydrology on the evolution of genetic diversity within species: An examination of microsatellite DNA variation in bull trout, *Salvelinus confluentus* (Pisces: Salmonidae). Evolution 57:328-344

Coyne JA, Barton NH, Turelli M (1997) Perspective: a critique of Sewall Wright's shifting balance theory of evolution. Evolution 143:353-364

Crossman EJ, McAllister DE (1986) Zoogeography of freshwater fishes of the Hudson Bay drainage, Ungava Bay and the Arctic Archipelago. In: Hocutt CH, Wiley EO (eds) The zoogeography of North American freshwater fishes. John Wiley and Sons, New York, pp 53-104

Dakin EE, Avise JC (2004) Microsatellite null alleles in parentage analysis. Heredity 93:504-509

Danzmann RG, Morgan II RP, Jones MW, Bernatchez L, Ihssen PE (1998) A major sextet of mitochondrial DNA phylogenetic assemblages extant in eastern North American brook trout (*Salvelinus fontinalis*): distribution and post-glacial dispersal patterns. Can. J. Zool. 76:1300-1318

De Sousa SN, Finkeldey R, Gailing O (2005) Experimental verification of microsatellite null alleles in Norway spruce (*Picea abies* [L.] Karst.): Implications for population genetic studies. Plant Molecular Biology Reporter 23:113-119

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B 39:1-38

Di Rienzo A, Peterson AC, Garza JC, Valdes AM, Slatkins M, Freimer NB (1994) Mutational processes of simple-sequence repeat loci in human populations. Proceedings of the National academy of Sciences. 91:3166-3170

Doherty PC, Zinkernagel RM (1975) Enhanced immunological surveillance in mice heterozugous at the H-2 gene complex. Nature 256:50-52

Drummond CS, Hamilton MB (2007) Hierarchical components of genetic variation at a species boundary: population structure in two sympatric varieties of *Lupinus microcarpus* (Leguminosae). Mol. Ecol. 16:753-769

Dunbar HE, Wilson ACC, Ferguson NR, Moran NA (2007) Aphid thermal tolerance is governed by a point mutation in bacterial symbionts - art. no. e96. Plos Biology 5:1006-1015

Dupanloup I, Bertorelle G (2001) Inferring admixture proportions from molecular data: extension to any number of parental populations. Mol. Biol. Evol. 18:672-675

Dupanloup I, Schneider S, Excoffier L (2002) A simulated annealing approach to define the genetic structure of populations. Mol. Ecol. 11:2571-2581

Dyke AS, Prest VK (1987) Late Wisconsinan and Holocene history of the Laurentide ice sheet. Geographie physique et Quaternaire 41:237-263

Edmands S (2001) Phylogeography of the intertidal copepod *Tigriopus californicus* reveals substantially reduced population differentiation at northern latitudes. Mol. Ecol. 10:1743-1750

El Mousadik A, Petit RJ (1996) High level of genetic differentiation for allelic richness among populations of the argan tree [*Argania spinosa* (L.) Skeels] endemic to Morocco. Theor. Appl. Genet. 92:832-839

Elson JA (1969) Late quaternary marine submergence of Quebec. Revue Géographique de Montréal 23:247-258

Estoup A, Jarne P, Cornuet JM (2002) Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. Mol. Ecol. 11:1591-1604

Evgen'ev MB, Garbuz DG, Shilova VY, Zatsepina OG (2007) Molecular mechanisms underlying thermal adaptation of xeric animals. J Biosci 32:489-499

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theor. Popul. Biol. 3:87-112

Excoffier L (2001) Analysis of population subdivision. In: Balding DJ, Bishop MJ, Cannings C (eds) Handbook of statistical genetics, pp 271-308

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. Genetics 131:479-491

Fay JC, Wu CI (1999) A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. Molecular and Biological Evolution 16:1003-1005

Felsenstein J (2004) PHYLIP (Phylogeny Inference Package) ver. 3.6. In. Department of Genome Sciences, University of Washington, Seattle

Fisher RA (1930) The genetical theory of natural selection. Clarendon Press, Oxford

Fisher RA (1954) Statistical Methods for Research Workers, 12th ed, Oliver and Boyd edn, Edinburgh

Ford MJ (2002) Applications of selective neutrality tests to molecular ecology. Mol. Ecol. 11:1245-1262

Fraser D, Bernatchez L (2005) Allopatric origins of sympatric brook charr populations: colonization history and admixture. Mol. Ecol. 14:1497-1510

Fraser DJ, Bernatchez L (2001) Adaptive evolutionary conservation: towards a unified concept for defining conservation units. Mol. Ecol. 10:2741-2752

Froeschke G, Sommer S (2005) MHC Class II DRB Variability and Parasite Load in the Striped Mouse (*Rhabdomys pumilio*) in the Southern Kalahari. Mol. Biol. Evol. 22:1254-1259

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133:693-709

Gaggiotti OE, Lange O, Rassmann Gliddon K (1999) A comparison of two indirect methods for estimating average levels of gene flow using microsatellite data. Mol. Ecol. 8:1513-1520

Gagnon MC, Angers B (2006) The determinant role of temporary proglacial drainages on the genetic structure of fishes. Mol. Ecol. 15:1051-1065

Garrigan D, Edwards SV (1999) Polymorphism across an exon-intron boundary in an avian MHC class II B gene. Molecular and Biological Evolution 16:1599-1606

Gilbert SC, Plebanski M, Gupta S, Morris J, Cox M, Aidoo M, Kwiatkowski D, Greenwood BM, Whittle HC (1998) Association of malaria parasite population structure, HLA, and immunological antagonism. Science 279:1173-1177

Girard P, Angers B (2006a) Characterization of microsatellite loci in Longnose dace (*Rhinichthys cataractae*) and interspecific amplification in five other Leuciscinae species. Mol. Ecol. Notes 6:69-71

Girard P, Angers B (2006b) The impact of post-glacial marine invasions on the genetic diversity of an obligate freshwater fish, the Longnose dace (*Rhinichthys cataractae*), on the Quebec peninsula. Can. J. Fish. Aquat. Sci. 63:1429-1438

Girard P, Angers B (2007a) Phylogeographic inferences of the Longnose dace (*Rhinichthys cataractae*) populations of Quebec Peninsula using microsatellite markers. In prep.

Girard P, Angers B (2007b) POST (POpulation STructure): A new method to depict complex genetic organization among populations without a priori clustering hypothesis. In prep.

Girard P, Angers B (2008) Assessment of power and accuracy of methods for detection and frequency-estimation of null alleles. Genetica (online)

Goudet J (2001) FSTAT, a program to estimate and test gene diversities and fixation indices (version 2.9.3).

Goudet J (2005) Hierfstat, a package for R to compute and test variance components and F-statistics. Mol. Ecol. Notes 5:184-186

Goudet J, Raymond M, de Meeüs T, Rousset F (1996) Testing differentiation in diploid populations. Genetics 144:1933-1940

Green MC, Waits JL, Avery ML, Tobin ME, Leberg PL (2006) Microsatellite Variation of Double-Crested Cormorant Populations in Eastern North America. J. Wildl. Manag. 70:579-583

Gregori C, Bauer B, Schwartz C, Kren A, Schuller C, Kuchler K (2007) A genetic screen identifies mutations in the yeast WAR1 gene, linking transcription factor phosphorylation to weak-acid stress adaptation. FEBS Journal 274:3094-3107

Guillot G, Mortier F, Estoup A (2005) Geneland: a computer package for landscape genetics. Mol. Ecol. Notes 5:712-715

Haldane JBS (1954) An exact test for randomness mating. J. Genet. 52:631-635

Halliburton R (2004) Introduction to population genetics., Prentice-Hall edn, Upper Saddle River

Hänfling B, Weetman D (2006) Concordant Genetic Estimators of Migration Reveal Anthropogenically Enhanced Source-Sink Population Structure in the River Sculpin, *Cottus gobio*. Genetics 173:1487-1501

Hanski I, Gaggiotti OE (2004) Metapopulation biology: past, present, and future. In: Hanski I, Gaggiotti OE (eds) Ecology, Genetics, and Evolution of Metapopulations. Elsevier Academic Press, Burlington, pp 3-22

Harpending HC, Batzer MA, Gurven M, Jorde LB, Rogers AR, Sherry ST (1998) Genetic traces of ancient demography. Proceedings of the National Academy of Sciences, USA 95:1961-1967

Hartl DL, Clark AG (1997) Principles of population genetics Third Edition. Sinauer Associates, Inc., Sunderland

Hathaway RJ, Davenport JW, Bezdek JC (1989) Relational duals of the c-means clustering algorithms. Pattern Recogn. 22:205-212

Hay J, Harris E (1999) Population bottlenecks and pattern of human polymorphism. Molecular and Biological Evolution 16:1423-1426

Hayashi K, Yoshida H, Nishida S, Goto M, Pastene LA, Kanda N, Baba Y, Koike H (2005) Genetic Variation of the MHC DQB Locus in the Finless Porpoise (*Neophocaena phocaenoides*). Zoological Science 23:147-153

Heath DD, Shrimpton JM, Hepburn RI, Jamieson SK, Brode SK, Docker MF (2006) Population structure and divergence using microsatellite and gene locus markers in Chinook salmon (*Oncorhynchus tshawytscha*) populations. Can. J. Fish. Aquat. Sci. 63:1370-1383

Heckel G, von Helversen O (2002) Male tactics and reproductive success in the harem polygynous bat *Saccopteryx bilineata*. Behav. Ecol. 13:750-756

Hernandez JL, Weir BS (1989) A desiquilibrium coefficient approach to Hardy-Weinberg testing. Biometrics 45:53-70

Hill J, Grossman GD (1987) Home range estimates for three North American stream fishes. Copea 1987:376-380

Holm LE, Loeschcke V, Bendixen C (2001) Elucidation of the molecular basis of a null allele in a rainbow trout microsatellite. Mar. Biotechnol. 3:555-560

Horton H, Moran L, Ochs R, Rawn J, Scrimgeour K (1993) Principles of biochemistry, Prentice-Hall, Inc. edn. Neil Patterson Publishers, Upper Saddle River

Hubbs CL, Lagler KF (1949) Fishes of Isle Royale, Lake Superior, Michigan. Papers of the Michigan Academy of Science, Arts, and Letters 33:73-133

Hudson RR, Kreitman M, Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116:153-159

Hughes AL, Hughes MK (1995) Natural selection on the peptide-binding regions of major histocompatibility complex molecules. Immunogenetics 42:233-243

Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167-170

Ingvarsson PK, Garcia MV, Hall D, Luquez V, Jansson S (2006) Clinal variation in phyB2, a candidate gene for day-length-induced growth cessation and bud set, across a latitudinal gradient in European aspen (*Populus tremula*). Genetics 172:1845-1853

Ioerger TR, Clark AG, Kao TH (1991) Polymorphism at the self-incompatibility locus in Solanaceae predates speciation. Proceedings of the National Academy of Sciences, USA 87:9732-9735

Jarne P, Lagoda PJL (1996) Microsatellites, from molecules to populations and back. Trends in Ecology and Evolution 11:424-429

Johnson MS, Black R (1996) Geographic cohesiveness versus associations with habitat: genetic subdivision of *Bembicium vittatum Philippi* (Gastropoda: Littorinidae) in the Houtman Abrolhos Islands. Biol. J. Linn. Soc. 58:57-74

Jordan WC, Bruford MW (1998) New perspectives on mate choice and the MHC. Heredity 81:239-245

Karlsson S, Mork J (2005) Deviation from Hardy–Weinberg equilibrium, and temporal instability in allele frequencies at microsatellite loci in a local population of Atlantic cod. ICES J. Mar. Sci. 62:1588-1596

Kim Y, Stephan W (2000) Joint effects of genetic hitchhicking and background selection on neutral variation. Genetics 155:1415-1427

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Kimura M, Crow JF (1964) The number of alleles that can be maintained in a finite population. Genetics 49:725-738

Kimura M, Ohta T (1969) The average number of generations until fixation of a mutant gene in a finite population. Genetics 61:763-771

Kimura M, Ohta T (1978) Stepwise mutation model and distribution of allelic frequencies in a finite population. Proceedings of the National Academy of Sciences, USA 75:2868-2872

Kreitman M (1983) Nucleotide polymorphism at the alcohol deshydrogenase locus in *Drosophila melanogaster*. Nature 304:412-417

Kreitman M, Hudson RR (1991) Inferring the evolutionary histories of the Adh and Adh-dup loci in Drosophila melanogaster from patterns of polymorphism and divergence. Genetics 127:565-582

Krimbas CB (1984) On adaptation, neo-Darwinian tautology, and population fitness. Evolutionary Biology 17:1-57

Lafontaine P, Dodson JJ (1997) Intraspecific genetic structure of white sucker (*Catostomus commersoni*) in north-eastern North America as revealed by mitochondrial DNA polymorphism. Can. J. Fish. Aquat. Sci. 54:555-565

Landry C, Bernatchez L (2001) Comparative analysis of population structure across environments and geographical scales at major histocompatibility complex and microsatellite loci in Atlantic salmon (*Salmo salar*). Mol. Ecol. 10:2525-2539

Langefors AH (2005) Adaptive and Neutral Genetic Variation and Colonization History of Atlantic Salmon. Environ. Biol. Fishes 74:297-308

Langella O (1999) Populations v.1.2.28, CNRS.

Lasalle P, Tremblay G (1978) Dépôts meubles Saguenay lac Saint-Jean. Rapport 191. In. Ministère des Richesses naturelles du Québec.

Legendre P (2002) Program for multiple linear regression (ordinary or through the origin) with permutation test – User's notes. In. Département de sciences biologiques, Université de Montréal.

Legendre P, Gallagher ED (2001) Ecologically meaningful transformations for ordination of species data. Oecologia 129:271-280

Legendre P, Legendre V (1984) Post-glacial dispersal of freshwater fishes in the Québec peninsula. Can. J. Fish. Aquat. Sci. 41:1781-1802

Levene H (1949) On amatching problem arising in genetics. Annals of mathematical statistics 20:91-94

Li WH (1997) Molecular evolution. Sinauer Associates, Inc., Sunderland

Lodish H, Baltimore D, Berk A, Zipursky SL, Matsudaira P, Darnell J (1995) Molecular Cell Biology, Third Edition. W.H. Freeman and Company, New York and Oxford

Lu G, Basley DJ, Bernatchez L (2001) Contrasting patterns of mitochondrial DNA and microsatellite introgressive hybridization between lineages of lake whitefish (*Coregonus clupeaformis*): relevance for speciation. Mol. Ecol. 10:965-985

Lucchini V, Galov A, Randi E (2004) Evidence of genetic distinction and long-term population decline in wolves (*Canis lupus*) in the Italian Apennines. Mol. Ecol. 13:523-536

Luikart G, Cornuet JM (1998) Empirical evaluation of a test for identifying recently bottlenecked populations from allele frequency data. Conserv. Biol. 12:228-237

Lynch M, Pfrender M, Spitze K, Lehman N, Hicks J, Allen D, Latta L, Ottene M, Bogue F, Colbourne J (1999) The quantitative and molecular genetic architecture of a subdivided species. Evolution 53:100-110

Mainguy J, Worley K, Côté S, Coltman D (2007) Low MHC DRB class II diversity in the mountain goat: past bottlenecks and possible role of pathogens and parasites. Conservation Genetics 8:885-891

Mayer WE, O'hUigin C, Zaleska-Rutcynska Z, Klein J (1992) Trans-species origin of Mhc-DRB polymorphism in the chimpanzee. Immunogenetics 37:12-23

Mayr E (1963) Animal species and evolution. Harvard University Press, Cambridge

McAllister DE, Harington CR, Cumba SL, Renaud CB (1988) Paleoenvironmental and biogeographic analyses of fossil fishes in the peri-Champlain Sea deposits in Eastern Canada. In: Gadd NR (ed) The Late Quaternary Development of the Champlain Sea Basin, vol Special Paper 35. Geological Association of Canada, pp 241-258

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in *Drosophila.* Nature 351:652-654

McMahon CR, Bradshaw CJA (2004) Harem choice and breeding experience of female southern elephant seals influence offspring survival. Behav. Ecol. Sociobiol. 55:349-362

McMillan WO, Palumbi SR (1997) Rapid rate of control-region evolution in Pacific butterflyfishes (Chaetodontidae). J. Mol. Evol. 45:473-484

McPhail JD, Lindsey CC (1970) Freshwater fishes of northwestern Canada and Alaska. Bulletin of the Fisheries Research Board of Canada 173

Merila J, Hemborg C (2000) Fitness and feather wear in the Collared Flycatcher *Ficedula albicollis.* J. Avian Biol. 31:504-510

Merritt RB, Rogers JF, Kurz BJ (1978) Genic variability in the longnose dace, *Rhinichthys cataractae.* Evolution 32:116-124

Miller CR, Joyce P, Waits LP (2002) Assessing allelic dropout and genotyping reliability using maximum likelihood. Genetics 160:357-366

Miller KM, Kaukinen KH, Beacham TD, Withler RE (2001) Geographic heterogeneity in natural selection on an MHC locus in sockeye salmon. Genetica 111:237-257

Miller MJ, Yuan BZ (1997) Semiautomated resolution of overlapping stutter patterns in genomic microsatellite analysis. Anal. Biochem. 251:50-56

Muirhead CA (2001) Consequences of population structure on genes under balancing selection. Evolution 55:1532-1541

Nei M (1987) Molecular Evolutionary Genetics. Columbia University Press, New York

Nei M, Maruyama T, Chakraborty R (1975) The bottleneck effect and genetic variability in populations. Evolution 29:1-10

Neuhauser C (2001) Mathematical models in population genetics. In: Balding DJ, Bishop MJ, Cannings C (eds) Handbook of statistical genetics, pp 153-178

Nittinger F, Gamauf A, Pinsker W, Wink M, Haring E (2007) Phylogeography and population structure of the saker falcon (Falco cherrug) and the influence of hybridization: mitochondrial and microsatellite data. Mol. Ecol. 16:1497-1517

Nordborg M (2001) Coalescent theory. In: Balding DJ, Bishop MJ, Cannings C (eds) Handbook of statistical genetics, pp 179-212

O'Brien SJ, Yuhki N (1999) Comparative genome organization of the major histocompatibility complex: Lessons from the Felidae. Immunol. Rev. 167:133-144

Occhietti S (1989) Quaternary geology of Saint-Lawrence Valley and adjacent Appalachian subregion. In: Fulton RJ (ed) Quaternary Geology of Canada and Greenland. Geological Survey of Canada, Geology of Canada 1., pp 350-389

Ohta T, Kimura M (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. 22:201-204

Ono H, O'Huigin C, Vincek V, Klein J (1993) Exon-intron organization of fish major histocompatibility complex class II beta genes. Immunogenetics 38:223-234

Orita M, Iwahana H, Kanazawa H, Hayashi K, Sekiya T (1989) Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphism. Proceedings of the National Academy of Sciences, USA 86:2766-2770

Otto SP (2000) Detecting the form of selection from DNA sequence data. Trends Genet. 16:526-529

Ottovà E, Imkovà AT, Morand S (2007) The role of major histocompatibility complex diversity in vigour of fish males (Abramis brama L.) and parasite selection. Biol. J. Linn. Soc. 90:525-538

Page LM, Burr BM (1991) A field guide to freshwater fishes of North America north of Mexico. Houghton Mifflin Company, Boston

Paland S, Schmid B (2003) Population size and the nature of genetic load in *Gentianella germanica*. Evolution 57:2242-2251

Pamilo P, Nei M (1988) Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568-583

Parham P, Ohta T (1996) Population Biology of Antigen Presentation by MHC Class I Molecules. Science 272:67-74

Paterson S (2005) No evidence for specificity between hosts and parasite genotypes in *Strongyloides ratti* (Nematoda) infections. Int. J. Parasitol. 35:1539-1545

Pemberton JMS, J., Bancroft DR, Barrett JA (1995) Non-amplifying alleles at microsatellite loci: a caution for praentage and population studies. Mol. Ecol. 4:249-252

Penn D, Potts WK (1999) The evolution of mating preference and major histocompatibility complex genes. Am. Nat. 153:145-164

Petit RJ, Grivet D (2002) Optimal randomization strategies when testing the existence of a phylogeographic structure. Genetics 161:469-471

Piertney SB, Oliver MK (2006) The evolutionary ecology of the major histocompatibility complex. Heredity 96:7-21

Piry S, Alapetite A, Cornuet JM, Paetkau D, Baudouin L, Estoup A (2004) GeneClass2: A Software for Genetic Assignment and First-Generation Migrant Detection. J. Hered. 95:536-539

Pitnick S, Markow TA (1994) Large-male advantages associated with costs of sperm production in Drosophila hydei, a species with giant sperm. Proceedings of the National Academy of Sciences, USA 91:9277-9281

Plourde S, Joly P, Runge JA, Zakardjian B, Dodson JJ (2001) Life cycle of *Calanus finmarchicus* in the lower St. Lawrence estuary; the imprint of circulation and late timing of phytoplankton bloom. Can. J. Fish. Aquat. Sci. 58:647-658

Posada D, Crandall KA, Templeton AR (2000) GeoDis: A program for the Cladistic Nested Analysis of the Geographical Distribution of Genetic Haplotypes. Mol. Ecol. 9:487-488

Posada D, Templeton AR (2004) GeoDis differentiating population structure from history. Version 2.2. Brigham Young University, Provo, UT

Potts WK, Wakeland EK (1990) Evolution of diversity at the major histocompatibility complex. Trends Ecol. Evol. 5:181-186

Prest VK (1970) Quaternary geology of Canada. In: Douglass RJW (ed) Geology and economic minerals of Canada. Geolical Survey Canada. Economic Geology Report 1., vol, Ottawa, pp 676-764

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945-959

Radforth I (1944) Some considerations on the distribution of fishes in Ontario. Contributions of the Royal Ontario Museum 25:1-116

Rannala B, Mountain JL (1997) Detecting immigration by using multilocus genotypes. Proc. Natl. Acad. Sci. U. S. A. 94:9197-9221

Raymond M, Rousset F (1995a) An exact test for population differentiation. Evolution 49:1280-1283

Raymond M, Rousset F (1995b) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. J. Hered. 86:248-249

Refseth UH, Fangan BM, Jakobsen KS (1997) Hybridization capture of microsatellites directly from genomic DNA. Electrophoresis 18:1519-1523

Reida SM, Carlb LM, Leana J (2005) Influence of riffle characteristics, surficial geology, and natural barriers on the distribution of the channel darter, *Percina copelandi*, in the Lake Ontario basin. Environ. Biol. Fishes 72:241-249

Rempel LL, Smith DG (1998) Post-glacial fish dispersal from the Mississippi refuge to Mackenzie River basin. Can. J. Fish. Aquat. Sci. 55:893-899

Richards EJ (2006) Inherited epigenetic variation - revisiting soft inheritance. Nat. Rev. Genet. 7:395-401

Richardson BA, Brunsfeld SJ, Klopfenstein NB (2002) DNA from bird-dispersed seed and wind-disseminated pollen provides insights into postglacial colonization and population genetic structure of whitebark pine (*Pinus albicaulis*). Mol. Ecol. 11:215-227

Roberts S, Romano C, Gerlach G (2005) Characterization of EST derived SSRs from the bay scallop, *Argopecten irradians*. Mol. Ecol. Notes 5:567-568

Ross CL, Markow TA (2006) Microsatellite variation among diverging populations of *Drosophila mojavensis*. J. Evol. Biol. 19:1691-1700

Rousset F (1996) Equilibrium values of measure of population subdivision for stepwise mutation processes. Genetics 142:1357-1362

Rousset F, Raymond M (1995) Testing heterozygote excess and deficiency. Genetics 140:1413-1419

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496-2497

Russell PJ (2002) Genetics, Benjamin Cummings edn, San Francisco

Ruzzante DE (1998) A comparison of several measures of genetic distance and population structure with microsatellite data: bias and sampling variance. Can. J. Fish. Aquat. Sci. 55:1-14

Salvane AGV, Balino BM (1998) Productivity and fitness in a fjord cod population: an ecological and evolutionary approach. Fisheries Research 37:143-161

Sauermann U, Nurnberg P, Bercovitch FB, Berard JD, Trefilov A, Widdig A, Kessler M, Schmidtke J, Krawczak M (2001) Increased reproductive success of MHC class II heterozygous males among free-ranging rhesus macaques. Hum. Genet. 108:249-254

Schad J, Ganzhorn JU, Sommer S (2005) MHC constitution and parasite burden in the Malagasy mouse lemur, *Microcebus murinus*. Conservation Genetics 5:299-309

Schibler L, Vaiman D, Cribiu EP (2000) Génétique moléculaire: principes et application aux populations animales. INRA Production Animales (hors série):37-43

Schierup MH, Vekemans X, Charlesworth D (2000) The effect of subdivision on variation at multi-allelic loci under balancing selection. Genet. Res. 76:51-62

Schneider S, Roessli D, Excoffier L (2000) Arlequin ver. 2.0: A Software for Population Genetics Data Analysis. In. Genetics and Biometry Laboratory, University of Geneva, Geneva

Schueler S, Tusch A, Scholz F (2006) Comparative analysis of the within-population genetic structure in wild cherry (Prunus avium L.) at the self incompatibility locus and nuclear microsatellites. Mol. Ecol. 15:3231-3243

Scott WB, Crossman EJ (1973) Freshwater fishes of Canada. Bulletin of the Fisheries Research Board of Canada 184

Selkoe KA, Gaines SD, Caselle JE, Warner RR (2006) Current shifts and kin aggregation explain genetic patchiness in fish recruits. Ecology 87:3082-3094

Senanan W, Kapuscinski AR (2000) Genetic relationships among populations of northern pike (*Esox lucius*). Can. J. Fish. Aquat. Sci. 57:391-404

Shipley P (2003) MICRO-CHECKER ver. 2.2.3. In. University of Hull, Hull

Simons AM, Berendzen PB, Mayden RL (2003) Molecular systematics of North American phoxinin genera (Actinopterygii: Cyprinidae) inferred from mitochondrial 12S and 16S ribosomal RNA sequences. Zool. J. Linn. Soc. 139:63-80

Skalski GT, Grose MJ (2006) Characterization of microsatellite loci in the creek chub (*Semotilus atromaculatus*). Mol. Ecol. Notes 6:1240-1242

Slade RW (1992) Limited MHC polymorphism in the southern elephant seal: implications for MHC evolution and marine mammal biology. Proc R Soc London, B 249:163-171

Sotka EE, Wares JP, Barth JA, Grosberg RK, Palumbi SR (2004) Strong genetic clines and geographical variation in gene flow in the rocky intertidal barnacle *Balanus glandula*. Mol. Ecol. 13:2143-2156

Stanton ML, Galen C, Shore J (1997) Population structure along a steep environmental gradient: consequences of flowering time and habitat variation in the snow buttercup, *Ranunculus adoneus*. Evolution 51:79-94

Stephens M (2001) Inference under the coalescent. In: Balding DJ, Bishop MJ, Cannings C (eds) Handbook of statistical genetics, pp 213-238

Suprunova T, Krugman T, Fahima T, Chen G, Shams I, Korol A, Nevo E (2004) Differential expression of dehydrin genes in wild barley, *Hordeum spontaneum*, associated with resistance to water deficit. Plant, Cell Environ 27:1297-1308

Taberlet P, Fumagalli L, Wust-Saucy AG, Cosson JF (1998) Comparative phylogeography and post-glacial colonization routes in Europe. Mol. Ecol. 7:453-464

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595

Takahata N (1989) Gene genealogy in three related populations: consistency probability between gene and population trees. Genetics 122:957-966

Takahata N, Satta Y (1998) Footprints of intragenic recombination at HLA loci. Immunogenetics 47:430-441

Taylor EB (1991) A review of local adaptation in Salmonidae, with particular reference to Pacific and Atlantic salmon. Aquaculture 98:185-207

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

Templeton AR (2002) "Optimal" Randomization Strategies When Testing the Existence of a Phylogeographic Structure: A Reply to Petit and Grivet. Genetics 161:473-475

Templeton AR, Crandall KA, Sing CF (1987) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and a analysis of alcohol dehydrogenase activiy in Drosophila. Genetics 117:343-351

Templeton AR, Routman E, Phillips C (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 140:767-782

Templeton AR, Sing CF (1993) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. Genetics 134:659-669

ter Braak CJF (1986) Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67:1167-1179

ter Braak CJF, Smilauer P (1999) Canoco for Windows ver. 4.02. In. Centre for Biometry Wageningen, Wageningen

Thompson AR, Petty JT, Grossman GD (2001) Multiscale effects of ressource patchiness on foraging behaviour and habitat use by longnose dace, *Rhynicthys cataractae*. Freshw. Biol. 46:145-160

Thuillet AC, Tenaillon MI, Anderson LK, Mitchell SE, Kresovich S, Stack SM, Gaut B, Doebley J (2007) A Weak Effect of Background Selection on Trinucleotide Microsatellites in Maize. J. Hered. online

Toju H, Sota T (2006) Phylogeography and the geographic cline in the armament of a seed-predatory weevil: effects of historical events vs. natural selection from the host plant. Mol. Ecol. 15:4161-4173

Travis J (1989) The role of optimizing selection in natural populations. Annu. Rev. Ecol. Syst. 20:279-296

Tsuruta T, Goto A (2006) Fine scale genetic population structure of the freshwater and Omono types of nine-spined stickleback *Pungitius pungitius* (L.) within the Omono River system, Japan. J. Fish Biol. 69(Suppl. B):155-176

Turgeon J, Bernatchez L (2001) Mitochondrial DNA phylogeography of lake cisco (*Coregonus artedi*): evidence supporting extensive secondary contacts between two glacial races. Mol. Ecol. 10:987-1001

Underhill JC (1986) The fish fauna of the Laurentian Great Lakes, the Saint-Lawrence Lowlands, Newfoundland and Labrador. In: Hocutt CH, Wiley EO (eds) The Zoogeography of North American Freshwater Fishes. John Wiley and Sons, New York, pp 105-136

Valière N (2002) GIMLET: a computer program for analysing genetic individual identification data. Mol. Ecol. 2:377-379

van Oosterhout C, Hutchison WF, Wills DPM, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. Mol. Ecol. Notes 4:535-538

van Oosterhout C, Joyce DA, Cummings SM (2006) Evolution of MHC class IIB in the genome of wild and ornamental guppies, *Poecilia reticulata.* Heredity 97:111-118

Volis S, Anikster Y, Olsvig-Whittaker L, Mendlinger S (2004) The influence of space in genetic-environmental relationships when environmental heterogeneity and seed dispersal occur at similar scale identifiers. Am. Nat. 163:312-327

Walker D, Avise JC (1998) Principles of phylogeography as illustrated by freshwater and terrestrial turtles in the southeastern United States. Annu. Rev. Ecol. Syst. 29:23-58

Walter R, Epperson BK (2004) Microsatellite analysis of spatial structure among seedlings in populations of *Pinus strobes* (Pinaceae). Am. J. Bot. 91:549-557

Warnes G, Leisch F (2005) genetics: Population Genetics. In: R package, 1.2.1 edn

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theoritical population biology 7:256-276

Watterson GA (1978) The homozygosity test of neutrality. Genetics 88:405-417

Weber JL, Wong C (1993) Mutation of human tandem repeats. Hum. Mol. Genet. 2:1123-1128

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38:1358-1370

Westerdahl H, Hansson B, Bensch S, Hasselquist D (2004) Between-year variation of MHC allele frequencies in great reed warblers: selection or drift? J. Evol. Biol. 17:485-492

Whiteley AR, Spruell P, Rieman BE, Allendorf FW (2006) Fine-scale genetic structure of bull trout at the southern limit of their distribution. Trans. Am. Fish. Soc. 135:1238-1253

Whitlock MC (2002) Selection, load and inbreeding depression in a large metapopulation. Genetics 160:1191-1202

Wiener P, Burton D, Ajmone-Marsan P, Dunner S, Mommens G, Nijman IJ, Rodellar C, Valentini A, Williams JL (2003) Signatures of selection? Patterns of microsatellite diversity on a chromosome containing a selected locus. Heredity 90:350-358

Wilson AC, Cann RL, Carr SM, George M, Jr., Gyllensten UB, Helm-Bychowski K, Higuchi RG, Palumbi SR, Prager EM, Sage RD, Stoneking M (1985) Mitochondrial DNA and two perspectives on evolutionary genetics. Biol. J. Linn. Soc. 26:375-400

Wilson CC, Hebert PDN (1996) Phylogeographic origins of lake trout (*Salvelinus namaycush*) in eastern North America. Can. J. Fish. Aquat. Sci. 53:2764-2775

Worley K, Strobeck C, Arthur S, Carey J, Schwantje H, Veitch A, Coltman DW (2004) Population genetic structure of North American thinhorn sheep (*Ovis dalli*). Mol. Ecol. 13:2545-2556

Wright S (1931) Evolution in mandelian populations. Genetics 16:97-159

Wyman A, White R (1980) A highly polymorphic locus in human DNA. Proceedings of the National Academy of Sciences, USA 77:6754-6758

Yang JY, Counterman BA, Eckert CG, Hodges SA (2005) Cross-species amplification of microsatellite loci in *Aquilegia* and *Semiaquilegia* (Ranunculaceae). Mol. Ecol. Notes 5:317-320

Yawson AE, Weetman D, Wilson MD, Donnelly MJ (2007) Ecological Zones Rather Than Molecular Forms Predict Genetic Differentiation in the Malaria Vector *Anopheles gambiae* s.s. in Ghana. Genetics 175:751-761